Quaestio facti. Revista Internacional sobre Razonamiento Probatorio
Quaestio facti. International Journal on Evidential Legal Reasoning
N. 3 | 2022 pp. 273-478
Madrid, 2022
Marcial Pons Ediciones Jurídicas y Sociales

ISSN: 2604-6202

PRESENTACIÓN DE LA TRADUCCIÓN AL CASTELLANO DEL INFORME DEL PCAST SOBRE LA CIENCIA FORENSE EN LOS TRIBUNALES PENALES

Carmen Vázquez
Universitat de Girona

En el ámbito de la prueba pericial es fundamental el análisis de la experiencia estadounidense, dados los desarrollos que han caracterizado ese sistema prácticamente desde 1923, con la resolución del caso *Frye*. No obstante, dicha experiencia muchas veces ha sido reducida al estudio del famoso caso *Daubert* y sus cuatro criterios establecidos por la Corte Suprema en 1993 para decidir si una prueba pericial era fiable y, con ello, si era admisible: que un método o técnica pueda ser (y haya sido) sometido a test o pruebas, que haya sido sometido a revisión por pares y publicado, que se indique su rango de error conocido o posible y se muestre que goza de una amplia aceptación en la comunidad experta de referencia[1].

Esos cuatro criterios han sido traducidos en muy diferentes idiomas para ser utilizados por la jurisprudencia[2], la dogmática procesal[3] o incluso por el propio

[1] Un análisis de los diversos criterios que conforman los casos *Frye* y *Daubert*, puede verse en Haack (2020) y en Vázquez (2015).

[2] Es el caso, por ejemplo, de México, donde la Suprema Corte de Justicia de la Nación los introdujo en la contradicción de tesis 154/2005-PS, diciendo que: «Para que un órgano jurisdiccional pueda apoyarse válidamente en una opinión de algún experto en una rama de la ciencia, es necesario que esa opinión tenga las siguientes características:

1. Que la evidencia científica sea relevante para el caso concreto de estudio, es decir, que a través de la misma pueda efectivamente conocerse la verdad de los hechos sujetos a prueba, y

Nota 3 en página siguiente

legislador[4] con el objetivo de establecer criterios de admisibilidad o, bien, para sugerirlos como criterios de valoración de este tipo de elementos de juicio. Sin embargo, la experiencia estadounidense va muchísimo más allá de la mera identificación de criterios para orientar las decisiones judiciales, pues sobre todo en los últimos años se han puesto en marcha esfuerzos institucionales diversos para obtener información relevante sobre la fiabilidad de métodos, técnicas, etc., *más allá de un proceso judicial concreto*. Dos rasgos destacan en estos esfuerzos:

i) Que muestran un interés en conocer la fiabilidad de los métodos, técnicas, teorías, etc., que usan los expertos (que luego fungen como peritos).Que dicha información es, en parte, independiente de un proceso judicial.

Aunque parezca una cuestión banal afirmar que estamos interesados en conocer la fiabilidad del conocimiento que emplea un experto cuando funge como perito en un proceso, en realidad no lo es y, de hecho, nuestros sistemas jurídicos han actuado como si no lo estuviéramos. ¿Por qué? Básicamente porque han prestado atención fundamentalmente a quién es el experto en turno para decidir si le creen o no[5] o

---

2. Que la evidencia científica sea fidedigna, esto es, que se haya arribado a ella a través del método científico, para lo cual se requiere, generalmente, que la teoría o técnica científica de que se trate:

a. Haya sido sujeta a pruebas empíricas, o sea, que la misma haya sido sujeta a pruebas de refutabilidad;

b. Haya sido sujeta a la opinión, revisión y aceptación de la comunidad científica;

c. Se conozca su margen de error potencial, y

d. Existan estándares que controlen su aplicación".

[3] En España, por ejemplo, Nieva Fenoll (2010: 294 ss.) sugiere que el dictamen debe seguir parámetros científicos de calidad, entre los cuales estaría: «que las técnicas y teorías científicas utilizadas para obtener datos y conclusiones han sido aplicadas previamente, son relevantes y están generalmente aceptadas por la comunidad científica internacional»; «que las técnicas utilizadas se han aplicado según los estándares y normas de calidad de vigentes»; o «que el dictamen contenga información sobre el posible grado o nivel de error y también el nivel o gradación de variabilidad e incertidumbre de los datos obtenidos por la citada técnica o teoría científica».

[4] En Colombia, el legislador en el Código procesal penal, Ley 906 de 2004, en el artículo 422, estableció: «Admisibilidad de publicaciones científicas y de prueba novel. Para que una opinión pericial referida a aspectos noveles del conocimiento sea admisible en el juicio, se exigiráá como requisito que la base científica o técnica satisfaga al menos uno de los siguientes criterios:

1. Que la teoría o técnica subyacente haya sido o pueda llegar a ser verificada.

2. Que la teoría o técnica subyacente haya sido publicada y haya recibido la crítica de la comunidad académica.

3. Que se haya acreditado el nivel de confiabilidad de la técnica científica utilizada en la base de la opinión pericial.

4. Que goce de aceptabilidad en la comunidad académica».

[5] Por supuesto, hay que atender, por ejemplo, a las credenciales de los expertos para asegurarse de que tienen el conocimiento relevante para analizar el caso y que, entonces, puedan ser nombrados como peritos. No obstante, la práctica de preocuparse por la titulación de un experto solo a efectos de atribuirle valor probatorio es ampliamente cuestionable, no solo porque la titulación no nos brinda gran información sobre los fundamentos que tiene un perito para afirmar lo que afirma, sino incluso porque la etapa en la que deberíamos atender a ese criterio es en la admisión. No debe admitirse a un sujeto que

han empleado criterios más bien formales relacionados con quién nombra al perito[6] o, en el peor de los casos, han atendido exclusivamente a ciertos comportamientos procesales del perito[7]. Ninguno de esos aspectos nos informa sobre la fiabilidad del conocimiento experto empleado y, por ello, solo podríamos decir que nos hemos interesado en la *credibilidad* que genera alguien que se presenta como experto o como un perito imparcial o que hace un peritaje que parece completo y coherente, etc. En este contexto, obviamente, credibilidad y fiabilidad son términos muy distintos: el primero es ampliamente subjetivo, dependiente de las creencias de cada uno sobre un sujeto; el segundo, en cambio, depende del mundo, de cómo de hecho funciona un método o una técnica con independencia de lo que cada uno de nosotros cree sobre cómo lo hace. Nótese que no solo se trata de parámetros más o menos objetivos, sino de qué es aquello que se está analizando, es decir, el sujeto o el conocimiento, método, técnica, etc. que emplea ese sujeto. Claramente la balanza de nuestros sistemas se ha inclinado vertiginosamente hacía el sujeto, descuidando los fundamentos que tiene para afirmar lo que afirma[8].

Vale la pena introducir un matiz en la distinción entre el sujeto y el conocimiento que emplea ese sujeto, pues hay métodos o técnicas que dependen en buena medida del criterio o juicio de una persona y, por ello, ciertas características suyas cobran suma relevancia para poder valorar mínimamente la calidad de sus afirmaciones. Un ejemplo de método que depende fuertemente de juicios humanos, de la percepción e interpretación del experto y que es empleado en algunas pericias antropológico-sociales es la llamada «observación participante» que, como su nombre indica, a grandes rasgos consiste en que el experto observa activamente a grupos o comunidades en su vida cotidiana para luego describir algún aspecto de ellas, por ejemplo, sobre sus costumbres en determinada actividad. Esto significa que el investigador, sus ojos, son en sí mismos el instrumento principal en este tipo de pruebas, pues su percepción, sus decisiones sobre qué observar, sus interpretaciones filtran todo el estudio de la

---

no es experto en la materia relevante para el caso, no podemos perder tiempo y esfuerzos en un proceso judicial en pruebas que son claramente irrelevantes.

[6] Sigue siendo común que se entienda que un experto seleccionado y pagado por una de las partes es necesariamente parcial y que no amerita valor probatorio; en cambio, por el contrario, que un perito oficial en sí mismo amerita mayor valor probatorio porque no ha sido seleccionado y pagado por las partes. No obstante, un perito de parte podría realizar su trabajo de manera totalmente correcta, usar métodos fiables, hacer un razonamiento impecable y, por ello, ameritar mayor valor probatorio que un perito oficial que podría usar métodos no tan fiables, cometer errores o incluso sufrir de ciertas parcialidades cognitivas. Precisamente por todo ello, la calidad de una prueba pericial no necesariamente depende de qué tipo de perito la haga.

[7] Abel Lluch (2017: 237 ss.) brinda como ejemplos «la ratificación y matización del dictamen en la vista o en el juicio» o «el dominio de la oralidad por el perito (silencios, dudas, comunicación no verbal, etc.)», etc.

[8] Esta situación es un problema compartido por las pruebas periciales y las pruebas testificales. En ambas, a efectos de su valoración, la atención se ha centrado en valorar quién es el sujeto que está afirmando X y no en las afirmaciones que hace.

comunidad que analiza. Así pues, podríamos decir que este método es más subjetivo que objetivo, en el sentido en que depende muy fuertemente de un sujeto.

No obstante, hay un buen conjunto de pruebas periciales con mucha mayor independencia del criterio del sujeto que las realiza y, en ese sentido, más objetivas. Entre esas encontramos las pruebas periciales que se basan en la comparación de ciertas características de personas u objetos que permitirían hacer una identificación relevante en un proceso judicial: por ejemplo, de un sujeto a través de su ADN o de un arma en función de determinados patrones. La validez y la fiabilidad de este tipo de pruebas constituye el objeto de análisis del informe del *President's Council of Advisors on Science and Technology* (PCAST) de Estados Unidos, realizado en septiembre de 2016, y titulado «Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature – Comparisions Methods», cuya traducción se publica aquí.

Y, precisamente respecto de la mencionada distinción entre métodos más o menos dependientes del criterio de un sujeto, el PCAST nos brinda un párrafo que no tiene desperdicio alguno:

> Queremos resaltar, finalmente, que ni la experiencia, ni el buen juicio, ni las buenas prácticas profesionales (tales como programas de certificación y acreditación, protocolos estandarizados, pruebas de aptitud técnica y códigos éticos) pueden sustituir las pruebas genuinas sobre la validez de sus fundamentos y su fiabilidad. La frecuencia con la que un patrón o un conjunto de características particulares se observa en diferentes muestras, que es un elemento esencial para inferir las conclusiones, no es un asunto de «juicio». Es un hecho empírico para el cual solo es relevante la evidencia empírica. De forma similar, la expresión de confianza utilizada por un analista, basada en su experiencia profesional, o las expresiones sobre un consenso entre analistas respecto a la precisión a la que llegan en su campo de trabajo, no pueden sustituir a las tasas de error estimadas a partir de estudios relevantes. Para los métodos forenses de comparación de características, el establecimiento de la validez de sus fundamentos a partir de evidencia empírica es entonces una condición *sine qua non*. Nada puede sustituirla.

El punto clave para entender por qué deberíamos estar interesados en conocer la validez y la fiabilidad de los métodos empleados para la identificación de objetos y personas que analiza el PCAST (y otros tantos empleados en los tribunales) es que la validez y la fiabilidad de un método es un hecho del mundo y que para evaluarlas solo nos sirve la evidencia empírica. Si se trata de un hecho que es susceptible de ser comprobado empíricamente, entonces se requiere hacer los experimentos necesarios para mostrar que los métodos o técnicas en comento tienen la capacidad de funcionar exitosamente. Quizá para el lector no familiarizado con la experimentación empírica, la siguiente narración de Luis Mochán (2010: 34), un físico de la Universidad Nacional Autónoma de México, le brinde cierta luz:

> Imagina que te lastimas un dedo jugando a voleibol. Un sanador te coloca un anillo imantado y una semana después tu dedo ha sanado y el dolor ha desaparecido. ¿Qué podrías concluir de ello?
> Que el magnetismo es terapéutico.
> No. El anillo magnético podría no haber jugado ningún papel. Todos los días hay gente que se lastima y sana sin ninguna ayuda. Para saber si el campo magnético sirve necesitamos repetir la experiencia muchas veces y hacer estadística con los resultados.
> Yo no estaría dispuesto a lastimarme repetidas veces para hacer estadística.

> Claro, pero un médico tiene acceso a muchos pacientes, y hay muchos médicos. Imagina que se consigue un grupo numeroso de pacientes con dedos lastimados y que la mitad de ellos se les coloca imanes y a la otra mitad no. A la semana han sanado un número determinado, que llamaremos P, de miembros del primer grupo y S miembros del segundo grupo. ¿Qué demostraría que P fuese mayor que S?
>
> Ahora sí, que los anillos magnéticos curan.
>
> No, todavía no. El primer grupo pudo haber sanado más rápido por sentir que su dolencia fue atendida y por su confianza en el tratamiento, mientras que la mejoría entre los pacientes del segundo grupo pudo haberse retrasado por la decepción de no haber recibido cuidado alguno. Quizás los miembros del primer grupo hubieran sanado igual con cualquier otro tratamiento o incluso con un tratamiento falso. Este es el conocido efecto placebo. Sería importante demostrar que el tratamiento es mejor que la administración de meros placebos. Para que el experimento sirva, se le tendría que poner un anillo similar a todos los pacientes, de los cuales solo la mitad estuvieran magnetizados y la otra mitad no. El paciente no debería saber qué clase de anillo le tocó. En este sentido, el experimento debe ser ciego.
>
> Si P es mayor que S, ahora sí sabríamos que el magnetismo es benéfico.
>
> Aún no del todo. Podría ser que el médico se mostrará más alegre y comunicara su confianza a los pacientes del primer grupo. Además, al registrar los resultados podrían manipularlos e ignorar pequeñas dolencias en los miembros del primer grupo y pequeñas mejorías en los del segundo grupo.
>
> ¡Eso sería deshonesto!
>
> No necesariamente. El médico podría favorecer los resultados del primer grupo inconscientemente, motivado por el deseo de que el tratamiento funcione. Para evitar esto, requeriríamos que el médico que vaya a distribuir los anillos y realizar las observaciones no sepa cuáles son magnéticos y cuáles no. La prueba debe ser doble ciego.
>
> Pero si nadie sabe nada sobre los anillos, ¿cómo podremos averiguar si los anillos magnéticos funcionaron?
>
> Claro, alguien debe saber algo de los anillos. Un participante en el experimento podría inscribir un número al azar en cada anillo y anotar qué números corresponden a qué tipo de anillo, pero sin informar al médico ni a los pacientes sino hasta después de haberse registrado los resultados.
>
> Ahora sí, ¿si P es mayor que S habremos demostrado las propiedades curativas de los imanes?
>
> Ya casi. Necesitaríamos que la diferencia sea suficientemente grande para ser estadísticamente significativa.
>
> Para ser ¿qué?
>
> Diferencias pequeñas podrían deberse al azar: por ejemplo, si tiras bolados honestos deberías obtener alrededor de 50 soles. No sería sorprendente que salieran 55, ¡pero si obtuvieses más de 80, yo podría sacar ciertas conclusiones estadísticamente significativas sobre tu poca honorabilidad!

Este tipo de experimentación científica, en dónde se pone en acción un cierto método o técnica (en el ejemplo, los anillos imantados) en una cantidad suficiente de varios casos en las mismas condiciones (*v. gr.,* dedos lastimados) para lograr un objetivo (en el relato, sanar el dedo), de ninguna manera puede ser sustituida por la mera experiencia de un sujeto o por acuerdos de grupos de expertos, etc. ¿Por qué la experiencia de un sujeto no contaría para ello? Porque se corren distintos riesgos importantes: como que el criterio de corrección de lo que dice un experto termine siendo el mismo experto o que el sujeto sufra de sesgos que le impidan ver sus errores o las pruebas contrarias a sus creencias o que no sea capaz de analizar un número suficiente de casos, entre muchos otros. Precisamente para evitar esos riesgos que

invalidarían el experimento, es que este tipo de investigación se debe llevar de forma controlada, es decir, estableciendo medidas que garanticen, en la medida de lo posible, que los resultados se deben exclusivamente al funcionamiento correcto de un método o técnica y no a otras cuestiones.

En todo lo anterior, la idea de «replicabilidad» es clave. Si un método o técnica empleado por un experto no puede ser puesto en marcha de igual manera por otros expertos para tratar el tipo de casos en que supuestamente funciona, entonces no hay manera de comprobar si efectivamente lo hace o no. Esto supone que la experimentación es sobre *un* método estandarizado, por ejemplo, que permite identificar las huellas dactilares, y no sobre distintos métodos que podrían llevar a identificar huellas dactilares; aunque, obviamente, se podrían someter a experimentaciones distintas esos diversos métodos. En todo caso, la idea es preguntarse si este método X *realmente* funciona o no para llegar a conocer algo sobre el mundo. Esa idea intuitiva sobre si "funciona" o no, al trasladarse a términos técnicos, corresponde a la «validez» y «fiabilidad» de un método o técnica. En términos muy llanos, como dijo la Corte Suprema estadounidense, un método es válido si mide lo que pretende medir y, en cambio, su grado de fiabilidad nos informará sobre cuán bien mide eso que efectivamente mide. Esto significa que primero habría que mostrar la validez y luego la fiabilidad; siendo la primera una cuestión de todo o nada, es válido o inválido, mientras que la segunda, por el contrario, es gradual, más o menos fiable. Si un método no es válido, poco sentido tiene entonces preguntarnos por su fiabilidad.

Pues bien, el PCAST en este informe, «Ciencia forense en los tribunales penales: asegurando la validez científica de los métodos forenses basados en la comparación de características», nos ofrece un conjunto de datos empíricos fundamentales sobre la validez y la fiabilidad de varios métodos forenses: análisis de ADN, marcas de mordeduras, huellas dactilares latentes, armas de fuego, huellas de calzado y cabellos. No lo hace mediante la realización de experimentos como los arriba mencionados, sino haciendo «meta-análisis» de los experimentos publicados. Esto es, analiza un conjunto de estudios publicados en los que se da cuenta de los resultados que arrojaron diversos experimentos que tuvieron como objetivo mostrar la validez y fiabilidad de los métodos o técnicas que se estudian.

Los meta-análisis, como su nombre indica, son análisis de los análisis, esto es, integran de forma estructurada y sistemática los resultados obtenidos en diferentes estudios empíricos. Para ello se recopila de forma exhaustiva un conjunto de estudios que se considera que fueron hechos en las condiciones adecuadas, ofreciendo entonces un estudio de más casos que cualquiera de los tomados en cuenta en los estudios individuales.

Y aquí es donde el PCAST se encuentra con una gran sorpresa: para la mayoría de los métodos o técnicas que analiza, no hay un número importante de estudios empíricos publicados que hayan pretendido demostrar que efectivamente funcionan y cómo lo hacen. Ello es en sí mismo un gran problema para el uso de esos métodos o técnicas en el ámbito jurídico, pues: ¿cómo es que seguimos usando métodos o

técnicas para decidir casos que impactan en los derechos y libertades de nuestros ciudadanos cuando ni siquiera sabemos si funcionan? Y esta pregunta, crucial en el ámbito jurídico, primero debería hacerse desde el ámbito de los expertos, en este caso, de las ciencias forenses.

Una de las explicaciones que se han dado es que, en ambos contextos, el jurídico[9] y el forense[10], existe una peligrosa *in*cultura de la investigación empírica y todo lo que esta supone. Por ello, los propios forenses no se cuestionan si está demostrado científicamente que los métodos o técnicas que emplean de hecho funcionan, simplemente actúan como si lo estuviese. Y peor aún, incluso siguen afirmando que sus métodos o técnicas son infalibles, lo que es imposible desde el punto de vista empírico y científico. No hay ningún método o técnica 100% fiable y es precisamente por ello que la investigación empírica debe mostrar cuáles son las condiciones en que mejor funciona. Se trata, entonces, de una labor colectiva, en la que las diferentes piezas del engranaje científico deberían operar para mostrar que determinado instrumento puede ser usado en un proceso judicial. En cambio, en muchas ocasiones, parece que todas esas piezas solo *suponen* que funciona y, con esa mera creencia, se termina privando de la libertad o restringiendo los derechos de las personas. Es un sinsentido y una grave irresponsabilidad colectiva.

Esa situación de ignorancia debe cambiar, no podemos seguir tomando decisiones públicas que se apoyan, aunque sea parcialmente, en métodos y técnicas que ni los propios expertos (o supuestos expertos) saben si funcionan o cuán bien lo hacen. En esta línea, la traducción al castellano de este informe del PCAST puede jugar un rol importante, no solo brindándonos información relevante sobre la validez y fiabilidad de los métodos que analiza, sino incluso formándonos en el tipo de información que los sistemas de justicia deberían exigir a los expertos. Hago referencia a los expertos, y no a los peritos, porque la investigación que se requiere para obtener la información relevante supone el trabajo de comunidades de expertos, no de una persona o ni siquiera de alguien que funge como perito en los procesos judiciales. Esto nos recuerda la importancia de las comunidades de expertos a las que aluden los criterios establecidos en los casos *Frye* y *Daubert*, pero no por su mera existencia o porque podamos consultarles si aceptan o no un método o técnica, sino porque son

---

[9] Edmond (2020a y 2020b), por ejemplo, al tratar el caso de las huellas dactilares latentes, nos muestra cómo la jurisprudencia y las prácticas de distintos países del *common law* han ignorado las diversas investigaciones empíricas que se han desarrollado sobre la fiabilidad de estos métodos de identificación, muchas veces argumentando que las herramientas jurídicas son suficientes para mostrar si son o no fiables.

[10] Mnookin, *et. al* (2011: 726) defienden explicita y contundentemente que las ciencias forenses en general «actualmente no poseen -y, sin lugar a duda, deberían desarrollar- una base científica bien establecida. Esto sólo puede lograrse mediante el desarrollo de una cultura de investigación que impregne todo el campo de la ciencia forense. Una cultura de investigación […] debe basarse en los valores del empirismo, la transparencia y el compromiso con un análisis crítico permanente. Las disciplinas de las ciencias forenses deben acrecentar sustancialmente su compromiso con las pruebas de la investigación empírica como base de sus conclusiones».

quienes pueden demostrar adecuadamente si aquellos funcionan y cuán bien funcionan. Por ello, para mejorar nuestras prácticas con las pruebas periciales y, con ello, nuestros sistemas de justicia, no basta trasladar los criterios *Daubert*, sino entender las implicaciones más profundas que tiene su establecimiento: generar conocimiento actualizado sobre la validez y fiabilidad de los métodos y técnicas que emplean los peritos.

Ahora bien, antes mencioné que la información sobre el funcionamiento de los métodos, técnicas, etc., es independiente de *un* proceso judicial. Ello quiere decir, por un lado, que esa información debe ser el resultado de investigaciones que no deberían estar relacionadas con un caso específico y, por el otro, que un si se atribuye validez y fiabilidad a un método o técnica, ello nada dice sobre la corrección de su *aplicación* a un caso concreto. Claro, un método válido y tan fiable como la identificación de ADN en casos de paternidad, podría ser aplicado de manera inadecuada en un caso concreto y llevarnos a una identificación errónea por causas diversas: contaminación en la muestra o en el laboratorio [11], incapacidad del analista que llevó a cabo la comparación, etc. Es precisamente en este punto donde tiene importancia analizar *quién* realiza la labor pericial, si tiene una adecuada formación e incluso experiencia; pero, como ya se mencionó antes, ello no puede substituir a los datos empíricos sobre la validez y fiabilidad de los métodos o técnicas empleados. Por eso, el lector de este informe erraría si asumiera que la razón por la que en él se omite la cuestión de quién es el experto o cómo hace su labor en un proceso judicial radica en que estos extremos no son relevantes: lo son, sin duda, pero en un nivel distinto y mucho más limitado al que habitualmente se le atribuye en nuestros sistemas.

Este informe del PCAST es claramente ejemplo de un esfuerzo institucional por indagar sobre la validez y la fiabilidad de determinados métodos, técnicas, etc., que utilizan los peritos. Fue precedido por otro, conocido como NAS Report, realizado por la National Academy of Science de Estados Unidos, titulado «Strengthening Forensic Science in the United States: A Path Forward» (2009) [12], donde se pusieron sobre la mesa un conjunto importante de serias debilidades de las ciencias forenses en Estados Unidos. Ambos informes fueron realizados en la administración de Barack Obama, que con ello mostró gran sensibilidad sobre el tema, quizá debido a un par de acontecimientos que impulsaron fuertemente su presencia en la agenda pública: escándalos por malas prácticas en varios laboratorios forenses oficiales y la identificación de un amplio conjunto de condenas erróneas basadas en pruebas periciales poco

[11] Evidentemente, no basta con que un área de conocimiento desarrolle métodos y técnicas válidos y fiables, los Estados deben invertir en tener todas las condiciones adecuadas para que esos métodos y técnicas puedan ser aplicados de manera fiable a los casos que se deben resolver. Así, por ejemplo, de poco sirve a un sistema de justicia que las pruebas de ADN realizadas en ciertas condiciones puedan alcanzar una fiabilidad del 99.999%, si ese Estado no tiene de hecho esas condiciones.

[12] Puede consultarse en: https://www.ojp.gov/pdffiles1/nij/grants/228091.pdf

sólidas[13]. Es de resaltar que en lugar de esconder el problema o dejarlo en el olvido mediante el paso del tiempo, se transparentaron e hicieron públicos varios problemas reales de las pruebas periciales provenientes de las ciencias forenses. Esa transparencia tuvo como efecto que varias de las ciencias forenses que salieron mal informadas en el 2009 mejoraran mucho mediante el trabajo arduo y conjunto de las comunidades expertas, entre ellas, la identificación de huellas dactilares. Este es claramente un rasgo de la ciencia, identificar los errores, aprender de ellos e intentar resolverlos.

Muchos de nuestros sistemas han dado el paso de copiar los criterios *Daubert*. Sin embargo, no se han hecho políticas públicas dirigidas a detectar las debilidades que pudieran tener los métodos y técnicas que emplean los peritos en un proceso judicial, si se están empleando solo aquellos que son válidos y, en todo caso, cuál es su nivel de fiabilidad. La protección de los derechos subjetivos en un Estado de Derecho exige evitar los errores en las decisiones judiciales. Para hacerlo, no basta con elaborar buenas leyes de procedimiento: es necesario también fomentar la investigación sobre la validez y la fiabilidad de los métodos y técnicas que se usan por parte de las ciencias forenses. Una vez detectadas sus debilidades, es imprescindible, por una parte, promover la mejora de esas disciplinas y, por otra, informar de esas debilidades a los operadores jurídicos para que no tomen decisiones basadas en pruebas que carecen de la validez o fiabilidad requeridas o que no las sobrevaloren.

Por eso, estimo de la mayor importancia que el contenido de este informe del PCAST sea conocido por todos los operadores de nuestros sistemas de justicia. Ojalá pronto veamos un estudio así en nuestras latitudes. Mientras tanto, la esperanza es que la traducción de este informe sirva para concienciar sobre esa necesidad, la de que nuestros sistemas de justicia usen solo métodos y técnicas que hayan mostrado funcionar adecuadamente.

BIBLIOGRAFÍA

Abel Lluch, X. (2017): «Criterios orientadores de la valoración de la prueba pericial», en Picó i Junoy, J.: *Peritaje y prueba pericial*, Barcelona: Bosch Editor.

Duce, M. (2020): «Prácticas probatorias y riesgos de condenas erróneas: una visión empírica», en Ferrer Beltrán, J. y Vázquez, C., *El razonamiento probatorio en el proceso judicial. Un encuentro entre diferentes tradiciones*, Madrid – Barcelona: Marcial Pons.

Edmond, G. (2020a): «Cuando el derecho es poco fiable. Respuestas jurídicas a la prueba de huellas dactilares latente. Parte I», en *Quaestio Facti. Revista Internacional sobre razonamiento probatorio*, no. 1.

- (2020b): «Cuando el derecho es poco fiable. Respuestas jurídicas a la prueba de huellas dactilares latente. Parte II», en Ferrer Beltrán, J. y Vázquez, C., *El razonamiento probatorio en el proceso judicial. Un encuentro entre diferentes tradiciones*, Madrid – Barcelona: Marcial Pons.

Haack, S. (2020): *Filosofía del derecho y de la prueba*, Madrid – Barcelona: Marcial Pons.

[13] Sobre el rol de las pruebas periciales en las decisiones judiciales erróneas, fundamentalmente en las condenas erradas, veáse Duce (2020).

Mnookin, J.; Cole, S.; Dror, I.; Fisher, B.; Houck, M.; Inman, K.; Kaye, D.; Koehler, J.; Langenburg, G.; Risinger, M.; Rudin, N.; Siegel, J.; y Stoney, D., (2011): «The Need for a Research Culture in the Forensic Sciences», en *UCLA Law Review*, no. 58.

Mochán, L., (2010): «Magia, Ciencia, Salud y Seguridad Nacional», *Academia de Ciencias de Morelos A.C.*, lunes 6 de septiembre de 2010.

Nieva Fenoll, J. (2010): *La valoración de la prueba pericial*, Madrid – Barcelona: Marcial Pons.

Vázquez, C. (2015): *De la prueba científica a la prueba pericial*, Madrid – Barcelona: Marcial Pons.

INFORME AL PRESIDENTE.

CIENCIA FORENSE EN LOS TRIBUNALES PENALES: ASEGURANDO LA VALIDEZ CIENTÍFICA DE LOS MÉTODOS FORENSES BASADOS EN COMPARACIÓN DE CARACTERÍSTICAS

OFICINA EJECUTIVA DEL PRESIDENTE
CONSEJO DE ASESORES DEL PRESIDENTE
EN CIENCIA Y TECNOLOGÍA

SEPTIEMBRE DE 2016

TRADUCCIÓN DE JOSÉ JUAN LUCENA MOLINA

Quaestio facti. Revista Internacional sobre Razonamiento Probatorio / International Journal on Evidential Legal Reasoning
Año 2022 N. 3 pp. 1-196 DOI: 10.33115/udg_bib/qf.i3. 22743

I. SOBRE EL CONSEJO DE ASESORES DEL PRESIDENTE EN CIENCIA Y TECNOLOGÍA

El Consejo de Asesores del Presidente en Ciencia y Tecnología (PCAST, por sus siglas en inglés) es un grupo consultivo de prestigiosos científicos e ingenieros de la nación, elegidos por el Presidente para mejorar el asesoramiento científico y tecnológico disponible a su persona tanto desde la Casa Blanca como desde los Departamentos del Gabinete y de otras Agencias federales. PCAST es consultado acerca de, y frecuentemente realiza recomendaciones sobre, una amplia variedad de asuntos en los que la comprensión de aspectos relacionados con la ciencia, la tecnología y la innovación poseen potencial influencia en las decisiones de carácter político que conciernen al Presidente.

Para más información sobre PCAST, consúltese www.whitehouse.gov/ostp/pcast.

II. EL CONSEJO DE ASESORES DEL PRESIDENTE EN CIENCIA Y TECNOLOGÍA

COPRESIDENTES

John P. Holdren: *Asistente del Presidente para la Ciencia y la Tecnología. Director, Oficina de Política en Ciencia y Tecnología*

STAFF

Ashley Predith; *Director Ejecutivo*

Diana E. Pankevich; *Becaria de Políticas de Ciencia y Tecnología de la AAAS*

Jennifer J. Michael; *Especialista de Apoyo al Programa*

III. GRUPO DE TRABAJO DEL PCAST

Los miembros del Grupo de Trabajo participaron en la preparación de este informe. Todos los miembros del PCAST lo revisaron y aprobaron.

GRUPO DE TRABAJO

Eric S. Lander (Director del Grupo de Trabajo): *Presidente. Instituto Broad de Harvard y el MIT*

S. James Gates Jr.: *Cátedra John S. Toll de Física. Director, Centro para la Teoría de Cuerdas y Partículas. Universidad de Maryland, College Park*

Susan L. Graham: *Catedrática Distinguida Emérita en Ingeniería Eléctrica y Ciencias de la Computación. Universidad de California, Berkeley*

Michael McQuade; *Vicepresidente senior para la Ciencia y la Tecnología. Corporación United Technologies*

William Press: *Cátedra Raymer en Ciencia Computacional y Biología Integrativa. Universidad de Texas en Austin*

Daniel Schrag: *Cátedra Sturgis Hooper de Geología. Catedrático, Ciencia e Ingeniería Ambiental. Director, Centro de la Universidad Harvard para el Medioambiente. Universidad de Harvard*

STAFF

Diana E. Pankevich: *Becaria de Políticas de Ciencia y Tecnología de la AAAS*

Kristen Zarrelli: *Consultor, Política Pública & Proyectos Especiales. Instituto Broad de Harvard y del MIT*

EDITORA

Tania Simoncelli: *Consejera senior del Director. Instituto Broad de Harvard y del MI*

IV. ASESORES SENIOR

El PCAST consultó a un grupo de expertos jurídicos para que proporcionaran orientación sobre aspectos prácticos relacionados con la interacción entre la ciencia

y el derecho. El PCAST también buscó orientación y la aportación de dos estadísticos con experiencia en este ámbito. Se dio a los consejeros senior la oportunidad de revisar los primeros borradores para asegurar la exactitud fáctica. El PCAST expresa su gratitud a los que aquí se listan. Su consentimiento para participar con el PCAST en puntos específicos no implica que respalden los puntos de vista expresados en este informe. La responsabilidad de las opiniones, hallazgos y recomendaciones en este informe, así como cualquier error de hecho o de interpretación, descansa únicamente sobre el PCAST.

COPRESIDENTES DE ASESORES SENIOR

Honorable Harry T. Edwards: *Juez. Corte de Apelaciones de los Estados Unidos. Distrito del Circuito de Columbia*

Jennifer L. Mnookin: *Decana, Cátedra de Derecho David G. Price y Dallas P. Price. Universidad de California, Facultad de Derecho de Los Ángeles*

ASESORES SENIOR

Honorable James E. Boasberg: *Juez de Distrito. Corte de Distrito de los Estados Unidos. Distrito de Columbia*

Honorable Andre M. Davis: *Magistrado Emérito. Corte de Apelaciones de los Estados Unidos. Cuarto Circuito*

David L. Faigman: *Rector interino y Decano. Facultad de Derecho Hastings de la Universidad de California*

Stephen Fienberg: *Catedrático de Estadística y Ciencias Sociales de la Universidad Maurice Falk (Emérito). Universidad Carnegie Mellon*

Honorable Pamela Harris: *Magistrada. Corte de Apelaciones de los Estados Unidos. Cuarto Circuito*

Karen Kafadar: *Rector interino y Decano. Facultad de Derecho Hastings de la Universidad de California*

Honorable Alex Kozinszi: *Magistrado Emérito. Corte de Apelaciones de los Estados Unidos. Noveno Distrito*

Honorable Cornelia T.L. Pillard: *Magistrada. Corte de Apelaciones de los Estados Unidos. Distrito del Circuito de Columbia*

Honorable Charles Fried: *Catedrático de Derecho. Facultad de Derecho de Harvard. Universidad de Harvard*

Honorable Nancy Gertner: *Profesora Titular de Derecho. Facultad de Derecho de Harvard. Universidad de Harvard*

Honorable Jed S. Rakoff: *Juez de Distrito. Corte de Distrito de los Estados Unidos. Distrito Sur de Nueva York*

Honorable Patti B. Saris: *Magistrada Presidente. Corte de Distrito de los Estados Unidos. Distrito de Massachusetts*

V. CARTA AL PRESIDENTE OBAMA

OFICINA EJECUTIVA DEL PRESIDENTE
CONSEJO DE ASESORES DEL PRESIDENTE
EN CIENCIA Y TECNOLOGÍA
WASHINGTON, D.C. 20502

Presidente Barack Obama
La Casa Blanca
Washington, DC 20502

Estimado Sr. Presidente:

Nos complace enviarle este informe del PCAST sobre *Ciencia forense en los Tribunales Penales de Justicia*: *Asegurando la validez científica de los métodos basados en comparaciones de características*. El estudio que condujo a este informe fue una respuesta a una pregunta suya al PCAST, en 2015, acerca de si hay pasos adicionales desde la vertiente científica que pudieran asegurar la validez de la prueba forense en el sistema jurídico nacional, más allá de los llevados a cabo por la Administración tras el informe extremadamente crítico del Consejo de Investigación Nacional del año 2009 sobre el estado de las ciencias forenses.

El PCAST concluye que hay dos importantes carencias: (1) la necesidad de clarificar los estándares científicos sobre la validez y fiabilidad de los métodos forenses y (2) la necesidad de evaluar métodos forenses específicos para determinar si han sido científicamente declarados como válidos y fiables. Nuestro estudio tiene como objetivo ayudar a resolver estas carencias para una serie de métodos forenses basados en «comparación de características» —específicamente, métodos de comparación de muestras de ADN, mordeduras, impresiones de huellas latentes, señales balísticas, huellas de calzado y cabello—.

Nuestro estudio, que incluye una extensa revisión de la literatura científica, tiene en cuenta también contribuciones de investigadores forenses del FBI (Oficina Federal de Investigación) y del NIST (Instituto Nacional de Estándares y Tecnología), así como de muchos otros científicos y analistas forenses, jueces, fiscales, abogados defensores, investigadores académicos, defensores de la reforma de la justicia penal y representantes de las Agencias federales. Los resultados y recomendaciones transmitidos en este informe son, desde luego, únicamente del PCAST.

Nuestro informe revisa estudios previos relacionados con la práctica forense y las acciones federales puestas recientemente en marcha para fortalecer la ciencia forense,

discute el papel de la validez científica dentro del sistema jurídico, explica los criterios que permiten juzgar la validez científica de los métodos forenses de comparación de características y aplica estos criterios a los métodos de comparación anteriormente mencionados.

Basándonos en nuestros resultados sobre la «validez de los fundamentos» de los métodos indicados, así como sobre la «validez de su aplicación» en la práctica ante los Tribunales, ofrecemos recomendaciones sobre las acciones que pudieran llevar a cabo el NIST, la Oficina sobre Política de Ciencia y Tecnología, y el Laboratorio del FBI, para fortalecer los fundamentos científicos de las disciplinas forenses, así como las acciones que pudieran realizar el Fiscal General y el Poder Judicial para promover el uso más riguroso de esas disciplinas en los Tribunales.

Atentamente,

John P. Holdren y Eric S. Lander, copresidentes.

VI. RESUMEN EJECUTIVO

«La ciencia forense» se ha definido como la aplicación de prácticas científicas o técnicas al reconocimiento, recogida, análisis e interpretación de evidencia relacionada con el derecho penal, civil o de cuestiones reglamentarias. Los acontecimientos de las últimas dos décadas —incluyendo: (i) exoneraciones de acusados que habían sido erróneamente condenados basados, en parte, en la prueba científico-forense, (ii) una serie de estudios sobre la base científica de las disciplinas forenses, (iii) revisiones de testimonios de analistas[*] basados en indicios forenses, y (iv) escándalos en laboratorios de criminalística estatales— han llamado cada vez más la atención sobre la validez y la fiabilidad de algunas importantes formas de prueba forense y el testimonio fundamentado en ellas[1].

---

[*] N. del T.: una de las primeras dificultades en la traducción está relacionada con lo que en el ámbito forense profesional español se entiende por «experto». Como ocurre en otros muchos países, quienes prestan testimonio experto de carácter forense en procesos penales como peritos pueden estar vinculados o no a laboratorios oficiales. En países como España, en los que rige un sistema procesal-penal continental, los «expertos» suelen ser peritos oficiales. Sin embargo, la formación académica de esos «expertos» es muy dispar y depende de la institución a la que pertenecen. Respecto a los cinco grandes laboratorios forenses españoles pertecientes a la red ENFSI (European Network of Forensic Science Institute), puede consultarse el Informe del Ministerio del Interior Español (2018). Debido a esa disparidad formativa, se ha preferido utilizar el término «analista» en la traducción, también utilizado en el ámbito cultural y jurídico anglosajón. Ese término intenta independizar al perito de su titulación académica, suponiéndole siempre capacitado para realizar las pericias de las que se responsabiliza oficialmente. En el ámbito cultural y jurídico mencionado se distingue entre *practitioners*, *experts*, *academics*, etc., de forma que esos términos se diferencian entre sí, fundamentalmente, por la preparación académica de los que ejercen la pericia ante los tribunales. Por todo ello, usaré pues el término «analista».

[1] Las citas de la bibliografía que sustentan aspectos mencionados en el Resumen Ejecutivo se encuentran en el cuerpo principal del informe.

Un estudio plurianual del *National Research Council* (Consejo de Investigación Nacional)[2] solicitado por el Congreso sobre este asunto, *Strengthening Forensic Science in the United States: A Path Forward* (Fortaleciendo la ciencia forense en los Estados Unidos: un paso hacia delante), y entregado en el año 2009, fue particularmente crítico en señalar la debilidad en la fundamentación científica de una serie de disciplinas utilizadas de manera habitual en el sistema judicial penal. Ese informe condujo a una amplia discusión sobre el camino a seguir dentro y fuera del gobierno federal y finalmente llevó al establecimiento de dos grupos: la *National Commission on Forensic Science* (Comisión Nacional sobre Ciencia Forense), integrada en el *Department of Justice* (Departamento de Justicia), y la *Organization for Scientific Area Committees for Forensic Science* (Organización de Comités de Áreas Científicas para la Ciencia Forense), del *National Institute of Standards and Technology* (Instituto Nacional de Estándares y Tecnología).

Tras la pregunta del Presidente Obama al *President's Council of Advisors on Science and Technology* (Consejo de Asesores del Presidente en Ciencia y Tecnología) (PCAST), en el año 2015, sobre la posibilidad de dar pasos adicionales que pudieran ser útiles desde el punto de vista científico para fortalecer las disciplinas forenses y así asegurar la validez de la prueba científica utilizada en el sistema jurídico nacional, el PCAST concluye que hay dos carencias importantes: (1) la necesidad de clarificar los estándares científicos sobre la validez y la fiabilidad de los métodos forenses y (2) la necesidad de evaluar métodos forenses específicos para determinar si han sido científicamente declarados como válidos y fiables.

Este informe tiene como objetivo ayudar a resolver estas carencias para el caso de métodos forenses basados en la «comparación de características» —es decir, métodos que intentan determinar si una muestra indiciaria (*e.g.* de la escena del crimen) está o no asociada con una muestra de una potencial «fuente» (*e.g.* de un sospechoso), fundamentándose en la presencia de patrones, impresiones u otras características que sean similares en la muestra y en la fuente—. Ejemplos de tales métodos son el análisis de ADN, cabellos, huellas latentes, armas y municiones disparadas, marcas de herramientas y marcas de mordeduras, huellas de calzado y de neumáticos y escritura manuscrita.

En el transcurso de su estudio, el PCAST ha recopilado y revisado más de 2000 artículos de diversas fuentes —incluyendo bibliografía preparada por el Subcomité sobre Ciencia Forense del *National Science and Technology Council* (Consejo de Ciencia y Tecnología Nacional) y por los Grupos de Trabajo relevantes organizados por el Instituto Nacional de Estándares y Tecnología (NIST); la información proporcionada por la comunidad interesada en la ciencia forense en respuesta a preguntas realizadas por PCAST; y las búsquedas en literatura propia del PCAST—.

---

[2] El National Research Council (Consejo de Investigación Nacional) es el órgano director de los estudios de las *National Academies of Science, Engineering, and Medicine* (Academias Nacionales de Ciencias, Ingeniería y Medicina).

Para informarse bien sobre problemas reales en la interacción entre la ciencia y el derecho, el PCAST ha consultado a un grupo de consejeros *senior* compuesto por nueve jueces federales (que ejercen o han ejercido como tales), un anterior procurador general de los Estados Unidos, un anterior juez de una Corte Suprema estatal, dos decanos de facultades de derecho y dos distinguidos estadísticos especializados en esta materia. Se obtuvo información adicional del Laboratorio del *Federal Bureau of Investigation* (Oficina Federal de Investigación) (FBI) y de científicos del NIST, así como de otros muchos científicos y prácticos forenses, jueces, fiscales, abogados defensores, académicos, defensores de la reforma de la justicia penal y representantes de agencias federales. La disposición de estos grupos e individuos para colaborar con el PCAST no implica que aprueben las opiniones expresadas en el informe. Las conclusiones y recomendaciones que se transmiten en el informe son de exclusiva responsabilidad del PCAST.

El informe resultante —resumido aquí sin las extensas elaboraciones técnicas y densas citas del texto principal que sigue a continuación— comienza con una revisión de estudios previos relacionados con la práctica forense y de las acciones federales que actualmente están en marcha para fortalecer la ciencia forense; reflexiona sobre el papel de la validez científica dentro del sistema jurídico; explica los criterios necesarios para valorar la validez científica de los métodos forenses basados en comparación de características; aplica estos criterios a seis de tales métodos en detalle y revisa una evaluación de un séptimo método que fue realizada por otros; finalmente, ofrece recomendaciones sobre acciones federales que pudieran adoptarse para fortalecer la ciencia forense y promover su más riguroso uso en los tribunales.

Creemos que las conclusiones y recomendaciones serán útiles tanto para el poder judicial como para aquellos que trabajan para fortalecer la ciencia forense.

1. Trabajos previos sobre validez científica de disciplinas de ciencia forense

Irónicamente, fue la aparición y proceso de madurez de una *nueva* ciencia forense, el análisis de ADN, en los años 90, el que condujo al cuestionamiento serio de la validez de la mayoría de las disciplinas forenses tradicionales. Cuando se introdujo por primera vez la prueba de ADN en los tribunales, a finales de los 80, se consideró inicialmente como una técnica infalible; pero los métodos utilizados en los primeros casos acabaron calificándose como poco fiables: los laboratorios carecían de procedimientos validados y consistentemente aplicados para definir los patrones de ADN de las muestras, para establecer cuándo dos patrones eran coincidentes dentro de una determinada tolerancia y para determinar la probabilidad de que tales coincidencias pudieran producirse por casualidad entre la población. Cuando a consecuencia de una decisión judicial sobre un caso acaecido en Nueva York, la prueba de ADN fue declarada inadmisible en el año 1989, los científicos involucrados en el análisis

de ADN, tanto en aplicaciones forenses como no forenses, aunaron esfuerzos para promover el desarrollo de principios y métodos fiables, lo que ha permitido que el análisis de ADN de muestras procedentes de un único individuo se califique como el «estándar de oro» de la ciencia forense tanto para la investigación como para el enjuiciamiento.

Una vez que el ADN llegó a ser una metodología fiable, el poder de la tecnología —que incluye su capacidad para analizar pequeñas muestras y distinguir entre individuos— hizo posible no sólo identificar y condenar a los verdaderos culpables, sino también contrarrestar falsas acusaciones antes del procesamiento y reexaminar condenas del pasado. Informes del *National Institute of Justice* (Instituto Nacional de Justicia) y otros institutos afirman que los tests de ADN durante el curso de las investigaciones han descartado a decenas de miles de sospechosos y que la revisión de condenas de casos cerrados mediante tests de ADN ha conducido a la exoneración de hasta 342 acusados. Informes independientes sobre esos casos han revelado que muchos se produjeron, en parte, por deficientes testimonios expertos de científicos forenses que comunicaron incorrectamente a los miembros del jurado que el hecho de encontrar similares características en un par de muestras (cabellos, muestras balísticas, marcas de mordedura, huellas de neumáticos o de calzado, u otros vestigios) tomadas, una del sospechoso y otra de la escena del crimen, implicaba al acusado en la comisión del crimen con un alto grado de certeza.

Los interrogantes que el análisis de ADN había suscitado sobre la validez científica de las disciplinas forenses tradicionales y el testimonio basado en ellas condujeron, naturalmente, a un incremento de los esfuerzos para comprobar empíricamente los métodos que tales disciplinas utilizaban. Los estudios relevantes que se produjeron son los siguientes:

— Un reexamen del FBI en 2002 de las comparaciones microscópicas de cabello que científicos de la agencia habían realizado en casos penales, en el que las pruebas de ADN revelaron que el 11% de las muestras de cabello que se encontraron con coincidencia microscópica en realidad provenían de individuos diferentes;

— Un informe realizado por el *National Research Council* (Consejo de Investigación Nacional) en 2004, a petición del FBI, sobre la prueba balística basada en la naturaleza del plomo, reveló que no existía ni suficiente investigación ni datos que permitieran concluir que hubiera una conexión definitiva entre dos proyectiles basándose en la similitud de la composición del plomo que contenían;

— Un informe del año 2005 realizado por un comité internacional, a solicitud del FBI, para revisar el uso de una prueba de una huella latente en el caso de un ataque terrorista con explosivos en España, constató que el «sesgo de confirmación» —es decir, la inclinación a confirmar una sospecha basada en otros fundamentos— contribuyó a una falsa identificación y a una detención incorrecta; y

— Estudios realizados en los años 2009 y 2010 sobre la prueba de mordedura, constataron que los actuales procedimientos para comparar marcas de mordedura no

eran capaces de excluir o incluir de modo fiable a un sospechoso como causante potencial de la mordedura.

Más allá de esta clase de deficiencias con respecto a «métodos fiables» en disciplinas forenses basadas en comparación de características, existen informes que concluyen que los testimonios expertos han sobrevalorado frecuentemente el valor probatorio de sus pruebas, yendo más allá de lo que la ciencia relevante pudiera justificar. A veces, los analistas han testificado que sus conclusiones tienen «una certeza del 100%»; o tienen una tasa de error igual a «cero», «esencialmente cero», o «insignificante». Como muchos estudios han resaltado —incluyendo el muy apreciado informe del Consejo de Investigación Nacional del año 2009—, sin embargo, tales expresiones no son científicamente defendibles: ninguno de los ensayos de los laboratorios y de los análisis de comparación de características tiene una tasa de error igual a cero.

Comenzando en el año 2012, el Departamento de Justicia (DOJ) y el FBI llevaron a cabo un estudio sin precedentes de las pruebas periciales en más de 3000 casos penales con análisis microscópico de cabellos. Los resultados iniciales, comunicados en el año 2015, mostraron que los analistas del FBI habían ofrecido un testimonio científicamente inválido en más del 95% de los casos en los que se utilizó para inculpar al acusado en el juicio oral. En marzo de 2016, el Departamento de Justicia anunció su intención de expandir a otros métodos de la ciencia forense la revisión de los testimonios forenses realizados por el laboratorio del FBI en casos penales cerrados. Esta revisión ayudará a valorar hasta qué punto se han dado similares testimonios sobreestimados en otras disciplinas forenses.

El informe del Consejo de Investigación Nacional del año 2009 es hasta la fecha la revisión más completa de las ciencias forenses en los Estados Unidos. El informe dejó claro que algunos tipos de problemas, irregularidades y errores judiciales, no pueden atribuirse simplemente a un puñado de analistas malintencionados o a laboratorios con bajo rendimiento, sino que son sistemáticos y generalizados —resultado de factores que tienen un alto nivel de fragmentación (incluyendo exigencias de entrenamiento y de formación, así como recursos y capacidades de los laboratorios dispares y frecuentemente inadecuados), carencia de estandarización de las disciplinas, insuficiente educación e investigación de alta calidad y escasez de estudios revisados por pares que establezcan la base científica y la validez de muchos métodos forenses rutinariamente utilizados—.

El informe del año 2009 resaltó que las deficiencias en las ciencias forenses eran especialmente prevalentes en las disciplinas de comparación de características, muchas de las cuales, decía el informe, carecían de sistemas bien definidos para determinar las tasas de error y en las que no se habían realizado estudios para establecer la unicidad, o la rareza o habitualidad de la frecuencia relativa de las marcas o características particulares examinadas. Además, las pruebas para evaluar la aptitud de los analistas, en los lugares en los que han sido realizadas, mostraron ejemplos de bajo rendimiento de analistas específicos. En resumen, el informe concluía que «muchas pruebas forenses —incluyendo, por ejemplo, las identificaciones por mordedura, por análisis de armas

de fuego y de marcas de herramientas— son introducidas en los procesos penales sin una significativa validación científica, una determinación de las tasas de error o una comprobación de su fiabilidad para explicar los límites de la disciplina.»

2. El contexto jurídico

Históricamente, la ciencia forense ha sido primordialmente utilizada en dos fases del procedimiento penal: (1) en la *investigación*, que busca identificar al presunto delincuente y (2) en la *acusación*, que busca probar la culpabilidad del acusado más allá de toda duda razonable. En los últimos años, la ciencia forense —particularmente el análisis de ADN— ha sido ampliamente utilizada para desafiar condenas pasadas.

Resulta importante ser conscientes de que las fases de investigación y procesamiento exigen diferentes estándares para el uso de la ciencia forense y de otras herramientas de investigación. En la investigación, las perspectivas y la información pueden proceder tanto de una ciencia bien establecida como de enfoques exploratorios. En la fase de procesamiento, la ciencia forense debe satisfacer un estándar más alto. Concretamente, las *Federal Rules of Evidence* (Reglas Federales de la Prueba) (Regla 702[c, d]) exigen para la admisibilidad del testimonio experto que esté fundamentado, entre otras cosas, sobre «principios y métodos fiables» que hayan sido «aplicados con fiabilidad» a los hechos del caso. Y la Corte Suprema ha establecido que los jueces deben determinar «si el razonamiento o la metodología subyacente al testimonio son científicamente válidos».

Aquí es donde los estándares jurídicos y los científicos se cruzan. Las decisiones de los jueces sobre la admisibilidad de la prueba científica descansan únicamente sobre estándares *jurídicos*; son un área exclusiva de los tribunales y el PCAST no opina sobre ellos. Pero esas decisiones requieren realizar determinaciones sobre la validez científica. Es facultad de la comunidad científica proporcionar una guía sobre los estándares científicos respecto a la validez científica y es sobre esos estándares *científicos* sobre los que se centra aquí el PCAST.

Distinguimos dos tipos de validez científica: la validez de los fundamentos y la validez en la aplicación.

(1) La *validez de los fundamentos* de un método científico-forense requiere que se haya demostrado, basándose en estudios empíricos, que es *repetible*, *reproducible* y *preciso* en los niveles que se hayan medido, y que sean apropiados para la aplicación prevista. La validez de los fundamentos, entonces, significa que un método puede, en principio, considerarse fiable. Pretendemos que el concepto *científico* se corresponda con el requisito jurídico previsto en la Regla 702(c), que alude a «principios y métodos fiables».

(2) La *validez en la aplicación* significa que el método ha sido fiablemente aplicado en *la práctica*. Pretendemos que el *concepto científico* pensamos se corresponda

con el requisito jurídico previsto en la Regla 702(d), que exige que un analista «haya aplicado fiablemente los principios y métodos a los hechos del caso».

3. Criterios científicos para la validez y la fiabilidad de los métodos forenses de comparación de características

La sección 4 del informe principal proporciona una detallada descripción de los criterios científicos para establecer la validez de los fundamentos y la fiabilidad de los métodos forenses de comparación de características, incluyendo tanto los métodos objetivos como los subjetivos[3].

Los métodos subjetivos requieren un escrutinio particularmente cuidadoso porque su fuerte dependencia de los juicios humanos conlleva que sean especialmente vulnerables a los errores humanos, a inconsistencias entre analistas y a sesgos cognitivos. En las disciplinas de comparación de características forenses, los sesgos cognitivos incluyen los fenómenos que, en ciertos entornos, los seres humanos podemos tender a fijarnos de forma natural en las similitudes entre las muestras y descartar sus diferencias, así como también podemos ser influenciados por información extraña al área de conocimiento y por presiones externas sobre el caso.

Las ideas esenciales sobre la validez de los fundamentos son las siguientes:

(1) La validez de los fundamentos requiere que un método haya estado sujeto a estudios *empíricos* realizados por múltiples grupos, en condiciones apropiadas para el uso pretendido. Los estudios deben (a) demostrar que el método es repetible y reproducible; y (b) proporcionar estimaciones válidas sobre la precisión del método (es decir, cuántas veces el método alcanza una conclusión incorrecta) que indiquen que es apropiado para la aplicación pretendida.

(2) Para métodos objetivos, la validez de los fundamentos del método puede establecerse estudiando las medidas de precisión, reproducibilidad y consistencia de cada una de sus etapas individuales.

(3) Para métodos subjetivos de comparación de características, dado que las etapas individuales no están objetivamente especificadas, el método debe evaluarse como si fuera una «caja negra» para la mente del analista. Las evaluaciones de validez y fiabilidad han de basarse en «estudios de caja negra», en los que muchos analistas

[3] Los métodos de comparación de características pueden clasificarse en métodos objetivos y subjetivos. Por métodos de comparación de características objetivos entendemos aquellos métodos consistentes en procedimientos definidos con suficiente detalle en cuanto a su estandarización y cuantificación de forma que puedan ser realizados por un sistema automatizado o por expertos humanos que no emitan juicios o apenas lo hagan. Por métodos subjetivos entendemos aquellos métodos que incluyen procedimientos esenciales que conllevan un significativo juicio humano —por ejemplo, sobre las características que deben seleccionarse dentro de un patrón, o cómo determinar si las características comparadas son lo suficientemente similares como para que pueda decirse que conforman una probable coincidencia—.

toman decisiones sobre una gran variedad de pruebas independientes (generalmente con muestras «dubitadas» y una o más muestras «indubitadas») y en los que se determinan las tasas de error.

(4) Si no existen estimaciones apropiadas de precisión, la aseveración de un analista consistente en que dos muestras son similares —o, incluso, indistinguibles— carece de significado científico: no tiene valor probatorio y, en cambio, sí tiene un considerable potencial para causar un impacto perjudicial.

Una vez que un método ha sido considerado válido en sus fundamentos como resultado de estudios empíricos apropiados, las afirmaciones acerca de la precisión del método y el valor probatorio de las identificaciones propuestas, para que sean válidas, deben fundamentarse en tales estudios empíricos. *Las declaraciones que afirmen o impliquen mayor certeza que la demostrada por la evidencia empírica son científicamente inválidas*. Los analistas forenses deben, por consiguiente, informar sobre los hallazgos de una identificación propuesta con claridad y con las restricciones debidas, explicando en cada caso que el hecho de que dos muestras satisfagan los criterios de un método para proponer un cotejo positivo (*match*) no significa que las muestras procedan de la misma fuente. Por ejemplo, si la tasa de falsos positivos de un método es de 1 en 50 comparaciones, los analistas no deben dar a entender que el método es capaz de producir resultados con una precisión mayor.

Para que pueda decirse que los criterios científicos de validez son aplicables, han de superarse dos pruebas:

(1) El analista debe haber demostrado que es *capaz* de aplicar de manera fiable el método y debe *realmente* haberlo hecho así. Demostrar que un analista es *capaz* de aplicar fiablemente el método es crucial —especialmente en métodos subjetivos, en los que el juicio humano desempeña un papel central—. Desde el punto de vista científico, la habilidad de aplicar fiablemente un método puede demostrarse únicamente mediante pruebas empíricas que midan cuántas veces el analista ofrece respuestas correctas. La determinación de si *realmente* un analista ha aplicado fiablemente el método, requiere que el procedimiento realmente utilizado en el caso, los resultados obtenidos y las notas de laboratorio, estén disponibles para que pueda llevarse a cabo una revisión científica por terceros.

(2) Las aseveraciones del analista sobre el valor probatorio de las identificaciones propuestas deben ser científicamente válidas. El analista debe informar sobre la tasa de falsos positivos y la sensibilidad del método, establecidas en los estudios sobre la validez de sus fundamentos y debe demostrar que las muestras utilizadas en los estudios sobre los fundamentos sean relevantes para los hechos del caso. Cuando sea aplicable, el analista debe informar sobre el valor probatorio del cotejo positivo (*match*) observado basándose en las características específicas observadas en el caso. Y el analista no debe realizar afirmaciones o implicaciones que vayan más allá de la evidencia empírica y de la aplicación de principios estadísticos válidos al tipo de prueba en juego.

Queremos resaltar, finalmente, que ni la experiencia, ni el buen juicio, ni las buenas prácticas profesionales (tales como programas de certificación y acreditación, protocolos estandarizados, pruebas de aptitud técnica y códigos éticos) pueden sustituir las pruebas genuinas con validez de los fundamentos y su fiabilidad. La frecuencia con la que un patrón o un conjunto de características particulares se observa en diferentes muestras, que es un elemento esencial para inferir las conclusiones, no es un asunto de «juicio». Es un hecho empírico para el cual solo es relevante la evidencia empírica. De forma similar, la expresión de *confianza* utilizada por un analista, basada en su experiencia profesional, o las expresiones sobre un *consenso* entre analistas respecto a la precisión a la que llegan en su campo de trabajo, no pueden sustituir a las tasas de error estimadas a partir de estudios relevantes. Para los métodos forenses de comparación de características, el establecimiento de la validez de sus fundamentos a partir de evidencia empírica es entonces una condición *sine qua non*. Nada puede sustituirla.

4. Evaluación de la validez científica de siete métodos de comparación de características

Para este estudio, el PCAST aplicó los criterios explicados más arriba a seis métodos de comparación de características forenses: (1) análisis de ADN de muestras procedentes de una única fuente y con una mezcla simple, (2) análisis de ADN de muestras con mezclas complejas, (3) marcas de mordedura, (4) huellas latentes, (5) identificación de armas de fuego y (6) análisis de huellas de calzado. Para cada método, la sección 5 del informe principal proporciona una breve explicación de la metodología, analiza estudios e información básicos, proporciona una evaluación sobre la validez científica y ofrece sugerencias de mejora para el futuro. Con respecto a un séptimo método de comparación de características —el análisis de cabellos— no disponemos de una evaluación completa de la validez científica, sino de una revisión de material de soporte recientemente entregado por el Departamento de Justicia para la realización de comentarios. Este resumen ejecutivo proporciona únicamente un breve resumen de algunos hallazgos clave concernientes a estos siete métodos.

4.1. Análisis de ADN de muestras procedentes de una única fuente y con una mezcla simple

La amplia mayoría de los análisis de ADN tienen que ver actualmente con muestras procedentes de un único individuo o de una mezcla simple de dos individuos (como el contenido en el llamado «kit de violación»[*]). El análisis de ADN en tales

[*] N. del T.: un *rape kit* es un paquete que contiene los instrumentos médicos necesarios para recoger y preservar evidencias físicas cuando se alega una violación sexual.

casos es un método objetivo en el que los protocolos de laboratorio están definidos con precisión y la interpretación conlleva muy poco o nada de juicio humano.

Para evaluar la validez de los fundamentos de un método objetivo, podemos examinar la fiabilidad de cada uno de los pasos individuales en lugar de tener que confiar en estudios de caja negra. En el caso del análisis de ADN de muestras procedentes de una única fuente o de mezclas simples, ha podido demostrarse que cada uno de esos pasos es «repetible, reproducible y preciso» con niveles que han sido medidos y que son «apropiados para la aplicación pretendida» (para citar el requisito de validez de los fundamentos expresado más arriba), y donde la probabilidad de que pueda encontrarse una coincidencia (*match*) por casualidad en la población ha sido directamente estimada a partir de bases de datos genéticas apropiadas y es extremadamente baja.

Con respecto a la validez en la aplicación de los análisis de ADN, al igual que sucede con todos los análisis forenses, no es infalible en la práctica. Puede haber errores y de hecho ocurren. Aunque la probabilidad de que dos muestras procedentes de fuentes distintas tengan el mismo perfil de ADN es pequeña, la posibilidad de un error humano es mucho mayor. Tales errores pueden derivarse de mezcla de muestras, contaminación, interpretación incorrecta y errores en la presentación de los informes.

Para minimizar el error humano, el FBI requiere, como una condición para participar en el *National DNA Index System* (Sistema nacional de base de datos de ADN) que los laboratorios sigan los *FBI's Quality Assurance Standards* (estándares de garantía de calidad del FBI). Estos requieren que el analista cumpla una serie de controles para comprobar una posible contaminación y asegurarse de que la prueba PCR transcurra apropiadamente. Los estándares también requieren evaluaciones semestrales respecto la aptitud de todos los analistas que realicen pruebas de ADN en casos penales. Hemos advertido, no obstante, la necesidad de mejorar las evaluaciones de aptitud.

4.2. Análisis de ADN de muestras con mezclas complejas

Algunas investigaciones conllevan realizar análisis de ADN con mezclas complejas de muestras biológicas procedentes de múltiples individuos desconocidos y en proporciones desconocidas (Tales muestras surgen, por ejemplo, de manchas de sangre mezcladas y, cada vez más, de múltiples individuos que tocan una superficie). La diferencia fundamental entre el análisis de ADN de muestras con mezclas complejas y el análisis de ADN de una única fuente y mezcla simple no estriba en el proceso del laboratorio sino en la interpretación del perfil de ADN resultante.

El análisis de ADN de mezclas complejas es intrínsecamente difícil. Tales muestras presentan un perfil de ADN que es la superposición de múltiples perfiles de

ADN individuales. Por muchas razones, la interpretación de un perfil mezcla es diferente y más complicada que la de un perfil simple. Con frecuencia es imposible decir con certeza qué variantes genéticas están presentes en la mezcla o cuántos individuos diferentes contribuyen a ella y mucho menos inferir con precisión el perfil de ADN de cada una de las personas.

Las preguntas que un analista debe hacerse, entonces, son: «¿podría estar el perfil de ADN de un sospechoso presente *dentro* del perfil de la mezcla? Y ¿cuál es la probabilidad de que eso pueda ocurrir por azar?». Dado que muchos perfiles diferentes de ADN pueden encajar dentro de algunos perfiles mezcla, la probabilidad de que un sospechoso «no pueda ser excluido» como posible contribuyente a un perfil mezcla puede ser *mucho más alta* (en algunos casos, millones de veces más alta) que las probabilidades encontradas para perfiles de ADN procedentes de una única fuente.

Las aproximaciones iniciales que se hicieron para la interpretación de mezclas complejas se apoyaron en juicios subjetivos de analistas y cálculos simplificados. Este enfoque es problemático porque las elecciones subjetivas hechas por los analistas pueden afectar drásticamente la respuesta y al valor probatorio estimado al incorporar un significativo riesgo tanto de error analítico como de sesgo de confirmación. El PCAST concluye que el análisis subjetivo de las mezclas de ADN complejas no ha sido establecido sobre la validez de sus fundamentos y no es una metodología fiable.

Dados los problemas encontrados en la interpretación subjetiva de mezclas complejas de ADN, una serie de grupos iniciaron esfuerzos para desarrollar programas informáticos que aplican varios algoritmos para interpretar mezclas complejas de manera objetiva. Estos programas representan claramente una mejora importante con respecto a la interpretación meramente subjetiva. Sin embargo, aún requieren escrutinio científico para determinar (1) si los métodos son científicamente válidos, incluyendo la delimitación de los márgenes de fiabilidad (es decir, las circunstancias en las cuales pueden ofrecer resultados no fiables) y (2) si el software implementa correctamente los métodos.

El PCAST encuentra que, hasta la fecha, los estudios han establecido la validez de los fundamentos de algunos métodos objetivos bajo determinadas circunstancias (específicamente, en una mezcla de tres personas en la que el contribuyente menor constituya, al menos, el 20% del ADN intacto en la mezcla), pero se considera que aún se necesita más evidencia sustancial para establecer la validez de los fundamentos en entornos más amplios.

4.3 Análisis de marcas de mordeduras

Los análisis de marcas de mordeduras normalmente conllevan examinar marcas ocasionadas sobre una víctima o sobre un objeto en la escena del crimen y compararlas con impresiones dentales tomadas de un sospechoso. La comparación de las marcas de mordeduras se basa en las premisas de que (1) las características dentales,

particularmente la disposición de los dientes delanteros, difieren sustancialmente entre las personas y (2) la piel (o alguna otra superficie marcada en la escena del crimen) puede recoger fiablemente esas características. El análisis de las marcas de mordedura comienza con el analista decidiendo si una herida es una marca causada por dientes humanos. Si ello es así, el analista hace fotografías o saca impresiones de la mordedura cuestionada y de la dentadura del sospechoso, compara la mordedura con la marca de la dentadura, y determina si la marca de la dentadura (1) no puede excluirse de haber producido la mordedura, (2) puede excluirse de haberla producido o (3) es inconclusa.

El análisis de marcas de mordeduras es un método subjetivo. Los actuales protocolos no proporcionan estándares bien definidos sobre la identificación de características o el grado de similitud que debe hallarse para sostener una conclusión fiable sobre si la marca pudo o no haber sido producida por la dentadura en cuestión. Estas preguntas quedan sometidas al juicio del analista.

Como se resaltó más arriba, la validez de los fundamentos de un método subjetivo solo puede ser establecida a través de estudios, múltiples y apropiadamente diseñados, de caja negra. Pocos estudios — y ningún estudio apropiado de caja negra— se han realizado para analizar la capacidad de los analistas para identificar con precisión la fuente de una marca de mordedura. En esos estudios, las tasas de falsos positivos observadas son muy altas —normalmente por encima del 10% o mucho más—. Además, algunos de estos estudios emplearon diseños de conjunto cerrado inapropiados que probablemente *sub*estiman la verdadera tasa de falsos positivos. En efecto, la evidencia científica disponible sugiere fuertemente que los analistas no sólo no pueden identificar la fuente de la marca de la mordedura con razonable precisión, sino que incluso ni siquiera pueden estar consistentemente de acuerdo sobre si la herida *es* una mordedura humana. Por estas razones, el PCAST encuentra que el análisis de marcas de mordeduras está lejos de cumplir los estándares científicos para la validez de sus fundamentos.

Observamos que algunos analistas forenses han expresado su preocupación por el hecho de que la exclusión de las marcas de mordedura por parte de los tribunales pudiera obstaculizar los esfuerzos para condenar a los acusados en algunos casos. Si así fuera, la solución correcta, desde una perspectiva científica, no sería la de admitir testimonio experto basado en métodos inválidos y no fiables sino, por el contrario, intentar desarrollar métodos científicamente válidos. Pero el PCAST considera que las perspectivas de desarrollo de análisis de marcas de mordedura para que sean métodos científicamente válidos son bajas. Desaconsejamos que se dediquen recursos significativos a tales esfuerzos.

4.4. Análisis de huellas latentes

El análisis de huellas latentes normalmente conlleva comparar (1) una «impresión latente» (una impresión completa o parcial de las crestas papilares de un sujeto desconocido) que ha sido revelada u observada sobre una superficie con (2) una o más «impresiones conocidas» (impresiones dactilares deliberadamente recogidas bajo una configuración controlada de sujetos conocidos; también referida como «reseña decadactilar»), para valorar si las dos pueden haber sido originadas por la misma fuente. Puede conllevar también la comparación de impresiones latentes entre sí. Al analista se le puede pedir que (1) compare una impresión latente con las impresiones decadactilares de un sospechoso conocido que ha sido identificado por otros medios («sospechoso identificado») o (2) que busque en una gran base de datos de impresiones decadactilares para identificar a un sospechoso («búsqueda en base de datos»).

El análisis de impresiones latentes fue propuesto para su uso en la identificación criminal en el siglo XIX y ha sido utilizado durante más de una centuria. El método fue considerado durante mucho tiempo como infalible, a pesar de la falta estudios empíricos apropiados para valorar su tasa de error. Como respuesta a las críticas sobre este particular vertidas por el informe del Consejo de Investigación Nacional en el año 2009, aquellos que trabajaban en el campo de los análisis de las impresiones latentes reconocieron la necesidad de realizar estudios empíricos para valorar la fundamentación de la validez y medir la fiabilidad, a partir de ahí ha habido progresos en esos menesteres. El mayor crédito se lo lleva el laboratorio del FBI, que guió el camino para realizar estudios de caja negra para valorar la validez y estimar la fiabilidad, así como los llamados estudios de «caja blanca» para comprender los factores que afectan a las decisiones de los analistas. El PCAST aplaude los esfuerzos del laboratorio del FBI. Existen también esfuerzos primigenios para comenzar a mover el campo desde un método puramente subjetivo hacia un método objetivo —aunque todavía falta un considerable camino para alcanzar esta importante meta—.

El PCAST encuentra que el análisis de impresiones latentes es una metodología subjetiva fundamentadamente válida —aunque con una tasa de falsos positivos sustancial y probablemente más alta que la esperada por muchos jurados debido a que tradicionalmente se han considerado infalibles los análisis de impresiones de huellas dactilares—. La tasa de falsos positivos pudiera llegar a ser tan alta como de 1 error cada 306 casos de acuerdo con el estudio del FBI y de 1 error cada 18 casos según un estudio realizado por otro laboratorio criminalístico[4]. A la hora de informar sobre

[4] El informe principal trata sobre los cálculos apropiados de las tasas de error, incluyendo los mejores estimadores (que son 1 de cada 604 y 1 de cada 24, respectivamente, para los dos casos citados) y los límites de confianza (los mencionados más arriba). También trata cuestiones con estudios específicos, incluyendo problemas con estudios que pueden conducir a diferencias en las tasas (como en los dos estudios citados).

resultados de exámenes de impresiones latentes, es importante establecer las tasas de falsos positivos basadas en estudios de validación apropiadamente diseñados.

Con respecto a la validez en la aplicación, se detectan, sin embargo, una serie de cuestiones abiertas, principalmente:

(1) Sesgo de confirmación. El trabajo de investigación de científicos del FBI ha demostrado que los analistas alteran con frecuencia las características que marcan inicialmente sobre una impresión latente cuando se compara con una huella conocida aparentemente coincidente. Tal razonamiento circular introduce un serio riesgo de sesgo de confirmación. Debe requerirse a los analistas que completen y documenten su análisis de huella dactilar latente *antes* de que observen cualquier impresión dactilar conocida y deberían documentar, separadamente, cualquier dato adicional utilizado durante sus comparaciones y evaluaciones.

(2) Sesgo contextual. El trabajo de académicos ha demostrado que los juicios de los analistas pueden verse influenciados por información irrelevante sobre los hechos del caso. Deberían realizarse esfuerzos para asegurar que los analistas no estén expuestos a información que pueda potencialmente producir sesgos.

(3) Pruebas de aptitud. Las pruebas de aptitud son esenciales para asegurar la capacidad y rendimiento del analista para emitir juicios precisos. Como se dijo en otro lugar de este informe, la prueba de aptitud necesita ser mejorada de forma que sea más rigurosa, incorporándola sistemáticamente en el flujo de trabajo de casos y divulgando pruebas para su evaluación por parte de la comunidad científica.

La validez científica en la aplicación requiere, por tanto, que un analista: (1) haya realizado pruebas de aptitud relevantes para probar su precisión y que informe de los resultados de la prueba de aptitud; (2) revele si documentó las características encontradas en la impresión latente por escrito antes de compararla con la impresión conocida; (3) proporcione un análisis escrito que explique la selección y la comparación de las características; (4) revele si, cuando realizó el examen, era consciente de algún hecho del caso que pudiera influirle en la conclusión; y (5) verifique que la impresión latente del caso en cuestión es similar, en calidad, dentro del rango de impresiones latentes consideradas en los estudios de fundamentación.

En lo referente al camino a seguir, se necesitan realizar esfuerzos continuos para mejorar el estado del arte de los análisis de impresiones latentes —y esos esfuerzos producirán claros beneficios en el sistema de justicia penal—. Una dirección posible es la de continuar mejorando el análisis de las impresiones latentes en cuanto método subjetivo. Hay necesidad de realizar estudios empíricos adicionales para estimar las tasas de error de los análisis de impresiones latentes de acuerdo con la variabilidad de su calidad y su integridad, utilizando mediciones bien definidas.

Una segunda —y más importante— dirección es la de convertir los análisis de impresiones latentes de un método subjetivo en un método objetivo. En la pasada década se han realizado extraordinarios avances en el análisis automatizado de las imágenes basado en *machine learning* y en otros enfoques —conduciendo a

mejoras notables en tareas tales como el reconocimiento facial y la interpretación de imágenes médicas—. Este progreso sostiene prometedoramente la posibilidad de que el análisis de impresiones latentes pueda hacerse completamente automático en un futuro próximo. Ha habido ya pasos en esta dirección tanto en el mundo académico como en la industria.

El recurso más importante para impulsar el desarrollo de métodos objetivos sería la creación de enormes bases de datos que contengan impresiones dactilares conocidas, cada una con muchas impresiones latentes «simuladas» de diferentes cualidades y grado de integridad, que podrían facilitarse a investigadores científicos capacitados en el ámbito académico y en la industria. Las impresiones latentes simuladas podrían crearse «modelando» («*morphing*») las impresiones conocidas, basándose en transformaciones derivadas de recolecciones de pares de impresión latente real – impresión rodada.

4.5. Análisis de armas de fuego

En el análisis de armas de fuego, los analistas intentan determinar si la munición está o no asociada a un arma de fuego específica basándose en «marcas de herramientas» producidas por las armas de fuego en la munición. La disciplina se basa en la idea de que las marcas de herramientas producidas por diferentes armas de fuego varían sustancialmente (debido a las variaciones en la fabricación y el uso) como para permitir que los componentes de cartuchos disparados se identifiquen con armas de fuego particulares. Por ejemplo, los analistas pueden comparar los casquillos «cuestionados» de un arma, recuperados de la escena del crimen, con disparos de prueba de un arma sospechosa. El examen comienza con una evaluación de las características de clase de los proyectiles y casquillos, que son características permanentes y predeterminadas anteriores a la fabricación. Si estas características de clase son diferentes, corresponde una conclusión de eliminación. Si las características de la clase son similares, el examen procede a identificar y comparar características individuales, como las marcas que surgen durante el disparo de un arma en particular.

Los analistas de armas de fuego han declarado durante mucho tiempo que su disciplina tiene una precisión casi perfecta; sin embargo, el estudio del Consejo de Investigación Nacional de 2009 sobre todas las disciplinas forenses concluyó acerca de los análisis de armas de fuego que «no se han realizado estudios suficientes para comprender la fiabilidad y reproducibilidad de los métodos», es decir, que la validez de los fundamentos de la disciplina no se ha establecido.

Nuestra extensa revisión de la literatura pertinente antes de 2009 es coherente con la conclusión del Consejo de Investigación Nacional. Encontramos que muchos de estos primeros estudios fueron inapropiadamente diseñados para evaluar la validez de sus fundamentos y estimar la fiabilidad. De hecho, hay evidencia interna entre

los propios estudios que indican que muchos estudios anteriores subestimaron al menos 100 veces la tasa de falsos positivos.

Identificamos un avance notable desde 2009: la finalización del primer estudio de caja negra apropiadamente diseñado sobre armas de fuego. El trabajo fue encargado y financiado por el *Defense Department's Forensic Science Center* (Centro de Ciencias Forenses del Departamento de Defensa) y fue llevado a cabo por un laboratorio de pruebas independiente (el Laboratorio Ames, un laboratorio nacional del *Department of Energy* (Departamento de Energía) afiliado a la Universidad Estatal de Iowa). La tasa de falsos positivos se estimó en 1 de cada 66, con un intervalo de confianza que indicaba que la tasa podría llegar hasta 1 de cada 46. Aunque el estudio está disponible como informe al gobierno federal, no se ha publicado en una revista científica.

Los criterios científicos para la validez de los fundamentos exigen que haya más de un estudio de este tipo para demostrar la reproducibilidad y que estos estudios deberían ser idealmente publicados en la literatura científica revisada por pares. Conforme a ello, las pruebas actuales aún no cumplen con los criterios científicos para la validez de los fundamentos.

Si el análisis de armas de fuego debe considerarse admisible sobre la base de las pruebas actuales es una decisión que corresponde a los tribunales. Si el análisis de armas de fuego *fuese* permitido en los tribunales, debería entenderse que los criterios científicos de validez en su aplicación exigen informar claramente de las tasas de error percibidas en el único estudio de caja negra apropiadamente diseñado. En la actualidad, no están justificadas científicamente las afirmaciones que asuman mayor precisión.

La validez en la aplicación también requeriría, desde un punto de vista científico, que un analista que testifique sobre el análisis de armas de fuego (1) se haya sometido a rigurosas pruebas de aptitud en un gran número de tests para medir su precisión y divulgue los resultados de las pruebas de aptitud y, además, (2) revele si al realizar el examen era consciente de cualquier otro hecho del caso que pudiera influir en la conclusión.

En cuanto al camino a seguir, en el análisis de armas de fuego al igual que con el análisis de huellas dactilares latentes, hay dos direcciones disponibles para fortalecer los fundamentos científicos de las disciplinas. La primera consiste en mejorar el análisis de armas de fuego como un método subjetivo, que requeriría estudios adicionales de caja negra para evaluar la validez científica y la fiabilidad, así como pruebas de aptitud más rigurosas a los analistas, utilizando problemas que sean apropiadamente desafiantes y que estos sean divulgados públicamente una vez hechas las pruebas.

La segunda dirección, al igual que con el análisis de impresiones latentes, es convertir el análisis de armas de fuego de un método subjetivo a un método objetivo. Esto implicaría desarrollar y probar algoritmos de análisis de imágenes para comparar la similitud de las marcas de herramientas en los proyectiles. Ha habido pasos alentadores hacia este objetivo. El mismo enorme progreso, llevado a cabo en la

última década en el análisis de imágenes, que aporta razones para esperar un logro temprano de un análisis de impresiones latentes totalmente automatizado, mueve al optimismo para que el análisis de armas de fuego totalmente automatizado pueda ser posible en un futuro próximo. Sin embargo, los esfuerzos en esta dirección se ven obstaculizados actualmente por la falta de acceso a bases de datos realmente grandes y complejas que puedan utilizarse para continuar el desarrollo de estos métodos y validar las propuestas iniciales.

El NIST, en coordinación con el Laboratorio del FBI, debe desempeñar un papel de liderazgo en la propulsión de la transformación necesaria mediante la creación y difusión de grandes conjuntos de datos apropiados. Estas agencias también deben proporcionar subvenciones y contratos para apoyar el trabajo y los procesos sistemáticos para evaluar los métodos. En particular, creemos que las competiciones «premiadas», que estén basadas en grandes colecciones de imágenes públicamente disponibles podrían atraer un interés significativo del ámbito académico y de la industria.

4.6. Análisis de huellas de calzado

El análisis de huellas de calzado es un proceso que ordinariamente conlleva comparar un objeto conocido, como un zapato, con una impresión completa o parcial hallada en la escena del crimen, para valorar si es probable que el objeto pudiera ser el origen de la impresión. El proceso se realiza secuencialmente, comenzando por una comparación de «características de clase» (tales como diseño, tamaño físico y desgaste general) y luego con una «identificación de características» o «características adquiridas aleatoriamente» (tales como marcas en el calzado causadas por cortes, incisiones y boquetes producidos en su uso).

El PCAST no ha abordado la cuestión de si los analistas pueden fiablemente determinar las características de clase —por ejemplo, si una impresión de calzado particular fue producida por un zapato de talla 12 de una específica manufactura—. Aunque es importante que se realicen estudios para estimar la fiabilidad de los análisis de huellas de calzado dirigidos a determinar las características de clase, el PCAST no quiso centrar su atención en este aspecto del examen de huellas de calzado porque la determinación de las características de clase, estimar la frecuencia de los tipos de calzado con una característica de clase particular o (a efectos de la tarea de los jurados) comprender la naturaleza de las propiedades en cuestión no es, *inherentemente*, un problema de medición desafiante.

En su lugar, el PCAST se centró en la fiabilidad de las conclusiones respecto la probabilidad de que una impresión proceda de un calzado *específico*. Se trata de un problema mucho más difícil porque requiere saber con qué precisión los analistas pueden identificar propiedades específicas compartidas entre un calzado y una impresión, cuánto fallan en identificar propiedades que los distinguirían y qué valor probatorio puede otorgarse a una particular «característica aleatoriamente adquirida».

El PCAST encuentra que no hay estudios de caja negra apropiados que soporten la validez de los fundamentos de los análisis de huellas de calzado consistentes en asociar impresiones con un calzado concreto mediante marcas identificativas específicas. Esas asociaciones no están fundamentadas en pruebas significativas o estimaciones de su precisión, por lo que no son científicamente válidas.

4.7. Análisis de cabellos

El análisis forense de cabellos es un proceso por el que los analistas comparan características microscópicas del cabello para determinar si una persona en particular puede ser el origen del cabello cuestionado. Mientras el PCAST completaba su informe, el Departamento de Justicia (DOJ) publicó, para que se hicieran comentarios, una propuesta de directrices sobre testimonios de exámenes de cabello, que incluía un documento de respaldo concerniente a la validez y fiabilidad de la disciplina. Aunque el PCAST no ha realizado una evaluación en profundidad de la disciplina de análisis de cabello como lo ha hecho en otras disciplinas basadas en comparación de características aquí tratadas, llevó a cabo una revisión del documento emanado del DOJ para arrojar más luz sobre los estándares para llevar a cabo una evaluación científica de una disciplina forense basada en comparación de características.

El documento establece que «la comparación microscópica de cabellos ha demostrado ser una metodología científica válida y fiable», mientras que subraya que «las comparaciones microscópicas de cabellos no pueden, por sí solas, conducir a una identificación personal y es crucial que esta limitación se transmita tanto en el informe escrito como en la comparecencia oral». Sin embargo, en apoyo a su conclusión de que el examen de cabellos es válido y fiable el documento sólo analiza un puñado de estudios sobre la comparación de cabello humano que abarcan los años 70 y 80. Los documentos de apoyo referidos fallan al no considerar que posteriores estudios encontraron defectos sustanciales en la metodología y en los resultados de los documentos clave. La propia revisión de los citados artículos realizada por el PCAST encuentra que esos estudios no establecen la validez de los fundamentos y fiabilidad de los análisis de cabellos.

El documento de apoyo del DOJ también cita un estudio del FBI realizado en el año 2002 que utilizó ADN mitocondrial para reexaminar 170 muestras de casos previos en los que el FBI había realizado análisis microscópicos de cabellos. Pero la conclusión principal de este estudio *no* respalda la conclusión de que el análisis de cabellos sea «una metodología científicamente válida y fiable». Los autores del FBI realmente encontraron que en 9 de 80 casos (11%) el Laboratorio del FBI había dictaminado que los cabellos eran microscópicamente indistinguibles, mientras que los análisis de ADN mostraron que los cabellos procedían de *diferentes* individuos.

Estas deficiencias ilustran tanto la dificultad de esas evaluaciones científicas como la razón por la que es mejor ellas sean llevadas a cabo por una agencia de carácter

científico que no esté involucrada en la aplicación de la ciencia forense dentro del sistema jurídico. Ellas también subrayan por qué es importante que la información *cuantitativa* sobre la fiabilidad de los métodos (por ejemplo, la frecuencia de falsas asociaciones en los análisis de cabellos) sea claramente establecida en el testimonio experto.

5. Observaciones finales de las siete evaluaciones

Aunque hemos realizado evaluaciones detalladas solo de seis métodos específicos —y una revisión de una evaluación realizada por otros de un séptimo— nuestro enfoque podría aplicarse para valorar tanto la validez de los fundamentos como la validez en la aplicación de cualquier método forense basado en la comparación de características, incluyendo tanto las disciplinas forenses tradicionales como también métodos aún pendientes de desarrollo (tales como el análisis microbiano o los patrones de navegación por Internet).

Hacemos notar, finalmente, que la evaluación de la validez científica ha de basarse, necesariamente, en la evidencia científica disponible en un momento dado. Algunos métodos que no han demostrado poseer validez en sus fundamentos podrían, al final, ser declarados fiables, aunque se requieran modificaciones significativas para que esos métodos alcancen esa meta. Algunos métodos puede que no la consigan, como puede ser el caso del análisis de la composición del plomo de los proyectiles y muy probablemente es el caso de las marcas de mordeduras. Otros pueden acabar siendo subsumidos por métodos distintos y más fiables, tal como el análisis de ADN ha reemplazado a otros métodos en algunos casos.

5.1. Recomendaciones al NIST y a la OSTP

1.ª Recomendación. Valoración de la validez de los fundamentos.

Es importante que las evaluaciones científicas de la validez de los fundamentos se realicen de forma continua para poder valorar la validez de los fundamentos de tecnologías de comparación de características forenses actuales y recientemente desarrolladas. Para asegurar que los juicios científicos no sean sesgados y sean independientes, tales evaluaciones deben estar dirigidas por un organismo científico que no tenga interés en el resultado.

(A) El Instituto Nacional de Estándares y Tecnología (NIST) debe realizar esas evaluaciones y debe emitir un informe público anual evaluando la validez de los fundamentos de los métodos clave de comparación de características.

(i) Las evaluaciones deben (a) valorar si cada método revisado ha sido adecuadamente definido, si ha sido adecuadamente establecida la validez de sus fundamentos

y si el nivel estimado de precisión se basa en evidencia empírica; (b) basarse en estudios publicados en la literatura científica de laboratorios y organismos de Estados Unidos y de otros países, así como cualquier trabajo dirigido por la propia plantilla y concesionarios del NIST; (c) como mínimo, elaborar evaluaciones sobre la línea del presente informe, actualizándolos según corresponda; (d) llevarse a cabo bajo los auspicios del NIST, con asesoramiento adicional realizado por analistas ajenos a la ciencia forense cuando se considere necesario.

(ii) El NIST debe establecer un comité asesor de científicos experimentales y estadísticos ajenos a la comunidad de ciencias forenses que proporcione asesoramiento sobre las evaluaciones y debe asegurarse de que sean rigurosos e independientes. Los miembros del comité consultivo deben seleccionarse conjuntamente por el NIST y la Oficina de Políticas Científicas y Tecnológicas.

(iii) El NIST debe priorizar aquellos métodos forenses de comparación de características que más necesiten de una evaluación, incluyendo los actualmente en uso y los que estén en desarrollo en etapas avanzadas, partiendo de las aportaciones del Departamento de Justicia y de la comunidad científica.

(iv) Cuando el NIST evalúe que un método ha sido establecido como válido en sus fundamentos, debe (a) aportar estimaciones apropiadas de tasas de error basadas en estudios sobre los fundamentos y (b) identificar cualquier problema relevante para la validez de su aplicación.

(v) Cuando el NIST evalúe que un método no ha sido establecido como válido en sus fundamentos, debe sugerir qué medidas, si las hubiere, podrían adoptarse para establecer la validez del método.

(vi) NIST no debe tener responsabilidades regulatorias con respecto a la ciencia forense.

(vii) NIST debe alentar a una o más revistas científicas líderes ajenas a la comunidad forense a que desarrolle mecanismos para promover la rigurosa revisión por pares y la publicación de artículos relacionados con la validez de los fundamentos de los métodos forenses de comparación de características.

(B) El Presidente debe solicitar y el Congreso debe proporcionar un aumento de financiamiento al NIST de (a) 4 millones de dólares para sustentar las actividades de evaluación anteriormente descritas y (b) de 10 millones de dólares para sustentar el aumento de las actividades de investigación en ciencia forense, incluyendo las mezclas complejas de ADN, las huellas dactilares latentes, el reconocimiento de voz y del hablante y la biometría facial y del iris.

2.ª Recomendación. Desarrollo de métodos objetivos para el análisis de ADN de muestras de mezclas complejas, análisis de huellas dactilares latentes y análisis de armas de fuego.

El Instituto Nacional de Estándares y Tecnología (NIST) debe desempeñar el papel de líder en la transformación de tres importantes métodos de comparación de características que son actualmente subjetivos —análisis de huellas dactilares latentes, análisis de armas de fuego y, en algunas circunstancias, análisis de ADN de mezclas complejas— en métodos objetivos.

(A) El NIST debe coordinar esos esfuerzos con el Laboratorio de la Oficina Federal de Investigaciones (FBI), el Centro de Ciencias Forenses de la Defensa, el Instituto Nacional de Justicia y otras agencias relevantes.

(B) Esos esfuerzos deben incluir (i) la creación y difusión de grandes bases de datos y materiales de pruebas (tales como mezclas de ADN complejas) que permitan el desarrollo y la prueba de los métodos, tanto por el sector privado como por investigadores académicos, (ii) subvenciones y contratos, y (iii) procesos de patrocinio, tales como competiciones premiadas, para evaluar métodos.

3.ª Recomendación. Mejora del proceso de organización para los Comités de Áreas Científicas.

(A) El Instituto Nacional de Estándares y Tecnología (NIST) debería mejorar la organización para los Comités de Áreas Científicas (OSAC), que fue establecida para desarrollar y promulgar estándares y directrices que permitan las mejores prácticas en la comunidad de ciencia forense.

(i) El NIST debe establecer un Comité de Recursos de Metrología, compuesto por metrólogos, estadísticos y otros científicos ajenos a la comunidad de ciencia forense. Un representante del Comité de Recursos de Metrología debe prestar servicio en cada uno de los Comités de Área Científica (SACs) para proporcionar orientación sobre la aplicación de los principios estadísticos y de medición de los estándares documentados en desarrollo[*].

(ii) El Comité de Recursos de Metrología, en su conjunto, debe revisar y aprobar o desaprobar públicamente todos los estándares propuestos por los Comités de Áreas Científicas antes de que sean transmitidos a la Junta Directiva de Estándares de Ciencia Forense.

(B) El NIST debe asegurar que el contenido de los estándares y directrices registrados por la OSAC estén disponibles libremente para cualquier parte que los desee

[*] N. del T.: los estándares documentados son los generados por los organismos de normalización de procedimientos técnicos.

en relación con un caso jurídico, o para evaluación o investigación, incluso alineándose con las políticas relacionadas con la disponibilidad razonable de estándares en la Oficina de Gestión y Circular presupuestaria A-119, en la Participación Federal en el Desarrollo y Uso de Estándares de Consenso Voluntario y en las Actividades de Evaluación de la Conformidad y la Oficina del Registro Federal, Manual IBR (incorporación por referencia).

4.ª Recomendación. Estrategia de I+D para la ciencia forense.

(A) La Oficina de Política de la Ciencia y de la Tecnología (OSTP) debe coordinar la creación de una estrategia nacional de investigación y desarrollo de la ciencia forense. La estrategia debe abordar los planes y las necesidades de financiación para:

(i) una gran expansión y fortalecimiento de la comunidad de investigación académica que trabaja en las ciencias forenses, incluyendo un aumento sustancial de la financiación tanto para la investigación como para la formación;

(ii) estudios de validez de los fundamentos de los métodos forenses de comparación de características;

(iii) mejora de los métodos forenses actuales, incluyendo la transformación de métodos subjetivos en métodos objetivos, y el desarrollo de nuevos métodos forenses;

(iv) el desarrollo de bases de datos de características forenses, con las adecuadas protecciones de la privacidad, que puedan utilizarse para la investigación;

(v) salvar la brecha entre los investigadores científicos y los profesionales forenses;

(vi) supervisión y revisión periódica de la investigación en ciencias forenses.

(B) Al preparar la estrategia, la OSTP debe solicitar la opinión de las agencias federales apropiadas, incluyendo especialmente el Departamento de Justicia, el Departamento de Defensa, la Fundación Nacional de la Ciencia y el Instituto Nacional de Estándares y Tecnología; a profesionales de la ciencia forense de ámbito federal y estatal; a investigadores científicos forenses y no forenses; y a otras partes interesadas.

5.2 Recomendación al Laboratorio del FBI

5.ª Recomendación. Agenda para la ampliación de las ciencias forenses en el Laboratorio del FBI.

(A) Programas de investigación. El Laboratorio de la Oficina Federal de Investigación (FBI) debería emprender un vigoroso programa de investigación para mejorar la ciencia forense, sobre la base de su reciente e importante trabajo en análisis de huellas dactilares latentes. El programa debe incluir:

(i) la realización de estudios sobre la fiabilidad de los métodos de comparación de características, junto a terceros independientes sin interés en el resultado;

(ii) el desarrollo de nuevos enfoques que mejoren la fiabilidad de los métodos de comparación de características;

(iii) la ampliación de los programas de colaboración con científicos externos; y

(iv) la garantía de que los científicos externos tengan acceso adecuado a los conjuntos de datos y colecciones de muestras para que puedan realizar estudios independientes.

(B) Estudios de caja negra. Basándose en su experiencia en la investigación científica forense, el Laboratorio del FBI debería ayudar en el diseño y ejecución de estudios adicionales de caja negra para métodos subjetivos, incluído los análisis de huellas dactilares latentes y de armas de fuego. Esos estudios deben ser realizados por o en conjunto con terceros independientes sin interés en el resultado.

(C) Desarrollo de métodos objetivos. El Laboratorio del FBI debe trabajar con el Instituto Nacional de Estándares y Tecnología para transformar tres importantes métodos de comparación de características que actualmente son subjetivos —análisis de huellas dactilares latentes, análisis de armas de fuego y, en algunas circunstancias, análisis de ADN de mezclas complejas—, en métodos objetivos. Estos esfuerzos deben incluir (i) la creación y difusión de grandes bases de datos que sustenten el desarrollo y las pruebas de métodos, tanto por parte de las empresas como de investigadores académicos, (ii) becas y apoyo contractual, y (iii) patrocinio de competiciones premiadas para evaluar los métodos.

(D) Pruebas de aptitud. El Laboratorio del FBI debe promover un mayor rigor en las pruebas de aptitud (i) dentro de los próximos cuatro años, instituyendo pruebas de aptitud ciegas de rutina dentro del flujo de casos en su propio laboratorio, (ii) ayudando a otros laboratorios federales, estatales y locales a hacerlo igualmente, y (iii) fomentar el acceso rutinario y la evaluación utilizadas en las pruebas comerciales de aptitud.

(E) Análisis de huellas dactilares latentes. El Laboratorio del FBI debe promover vigorosamente la adopción, por todos los laboratorios que realizan análisis de huellas dactilares latentes, de las normas que requieren un proceso de «Análisis, Comparación y Evaluación lineales» —en el que los analistas deben completar y documentar sus análisis de huellas dactilares latentes antes de mirar a una impresión dactilar conocida y deben, separadamente, documentar cualquier dato adicional utilizado en la evaluación y en la comparación—.

(F) Transparencia en materia de calidad en casos. El Laboratorio del FBI, así como otros laboratorios forenses federales, debe informar y regular públicamente sobre problemas de calidad en el análisis de casos (de forma similar a las prácticas empleadas por el Instituto Forense Holandés, descritas en la sección 5), como un medio para mejorar la calidad y promover la transparencia.

(G) Presupuesto. El Presidente debe solicitar y el Congreso debe proporcionar un incremento de la partida presupuestaria del FBI que restaure el presupuesto del Laboratorio del FBI para las actividades de investigación en ciencias forenses partiendo de su nivel actual con 30 millones de dólares adicionales, y debe evaluar la necesidad de aumentar la financiación para otras actividades de investigación en ciencias forenses en el Departamento de Justicia.

5.3. Recomendación al Fiscal General

6.ª Recomendación. Uso de los métodos de comparación de características en procesos judiciales federales.

(A) El Fiscal General debe ordenar a los fiscales que comparezcan en nombre del Departamento de Justicia (DOJ) se aseguren de que los testimonios ante los tribunales de los analistas sobre métodos de comparación de características cumplan los estándares científicos de validez científica.

Si bien las investigaciones previas al juicio pueden basarse en una gama más amplia de métodos, el testimonio de los analistas en el debate sobre métodos forenses de comparación de características en casos penales —que pueden ser muy influenciables y que ha conducido a muchas condenas injustas— debe cumplir un estándar más alto. Particularmente, los fiscales que comparezcan en nombre del Departamento de Justicia se deben asegurar de que:

(i) Los métodos forenses de comparación de características sobre los que se basa el testimonio hayan sido establecidos como válidos en sus fundamentos, como lo demuestran los estudios empíricos apropiados y la consistencia con las evaluaciones realizadas por el Instituto Nacional de Estándares y Tecnología (NIST), cuando estén disponibles; y

(ii) El testimonio sea científicamente válido, con las declaraciones del analista sobre la precisión de los métodos y el valor probatorio de las identificaciones propuestas constreñidas por la evidencia empírica que les sirve de apoyo, sin implicar un mayor grado de certeza.

(B) El Departamento de Justicia debe llevar a cabo una revisión inicial, con la asistencia del NIST, de los métodos de comparación de características subjetivos utilizados por el DOJ, con el fin de identificar qué métodos (más allá de los revisados en este informe) carecen de estudios apropiados de caja negra necesarios para evaluar la validez de los fundamentos. Como consecuencia de que se presume que tales métodos subjetivos no son válidos en sus fundamentos, el Departamento de Justicia debe evaluar si es apropiado que se presenten ante los Tribunales conclusiones basadas en tales métodos.

(C) Cuando existan métodos relevantes que aún no se hayan establecido como válidos en sus fundamentos, el DOJ debe fomentar y proporcionar apoyo para que se

realicen estudios apropiados de caja negra para evaluar la validez de los fundamentos y medir la fiabilidad. El diseño y la ejecución de estos estudios debe llevarse a cabo por o en unión de terceros independientes sin participación alguna en el resultado.

7.ª Recomendación. Directrices del Departamento de Justicia sobre el testimonio experto.

(A) El Fiscal General debe revisar y publicar nuevamente para recibir comentarios públicos, la propuesta de «Lenguaje Uniforme para Testimonios e Informes» y documentos de apoyo con el fin de alinearlos con los estándares científicos de validez científica.

(B) El Fiscal General debe emitir instrucciones que indiquen que:

(i) Cuando existan estudios empíricos y/o modelos estadísticos que arrojen luz sobre la precisión de un método de comparación de características, el analista debe proporcionar información cuantitativa sobre las tasas de error, de acuerdo con las directrices que establecerán el DOJ y el Instituto Nacional de Estándares y Tecnología, basada en el asesoramiento de la comunidad científica.

(ii) Cuando no existan estudios empíricos y/o modelos estadísticos apropiados para proporcionar información significativa sobre la precisión de un método forense de comparación de características, los fiscales del Departamento de Justicia y los analistas no deben ofrecer testimonio basado en ese método. Si es necesario dar testimonio en relación con el método, deben reconocer claramente ante los tribunales la falta de tales pruebas.

(iii) En el testimonio, los analistas deben siempre afirmar, claramente, que pueden ocurrir errores y que, de hecho, ocurren, debido tanto a la similitud entre las características como a errores humanos en el laboratorio.

5.4. Recomendación al Poder Judicial

8.ª Recomendación. La validez científica como fundamento del testimonio experto.

(A) Al decidir sobre la admisibilidad de testimonios de analistas, los jueces federales deben tener en cuenta los criterios científicos apropiados para evaluar la validez científica, que incluyen:

(i) La validez de los fundamentos, con respecto al requisito de la Regla 702(c) de que el testimonio sea el producto de principios y métodos fiables; y

(ii) La validez en la aplicación, con respecto al requisito de la Regla 702(d) de que el analista haya aplicado fiablemente los principios y métodos a los hechos del caso.

Estos criterios científicos se describen en el Hallazgo 1.

(B) Los jueces federales, cuando permitan que un analista testifique sobre un método de comparación de características válido en sus fundamentos, deben asegurarse de que el testimonio sobre la precisión del método y el valor probatorio de las identificaciones propuestas sea científicamente válido en el sentido de que esté limitado a lo que la evidencia empírica sustente. Las declaraciones que sugieran o impliquen una mayor certeza no son científicamente válidas y no deben permitirse. En particular, los tribunales nunca deben permitir aseveraciones científicamente indefendibles como: tasa de error «cero», «muy pequeña», «esencialmente cero», «despreciable», «mínima» o «microscópica»; «certeza del 100%» o prueba «hasta un grado razonable de certeza científica»; identificación «hasta la exclusión de todas las demás fuentes posibles»; o una probabilidad de error tan remota como para que sea una «imposibilidad práctica».

(C) Para ayudar a los jueces, la Conferencia Judicial de los Estados Unidos, a través de su Comité Consultivo Permanente sobre las Reglas Federales de la Prueba, debe preparar, con el asesoramiento de la comunidad científica, un manual de buenas prácticas y un documento del Comité Consultivo, proporcionando orientación a los jueces federales sobre la admisibilidad bajo la Regla 702 de testimonios de analistas basados en métodos forenses de comparación de características.

(D) Para ayudar a los jueces, el Centro Judicial Federal debe desarrollar programas sobre criterios científicos para la validez científica de los métodos forenses de comparación de características.

VII. INFORME

1. Introducción

La «ciencia forense» ha sido definida como la aplicación de prácticas científicas y técnicas al reconocimiento, recogida, análisis e interpretación de la evidencia en asuntos jurídicos penales o civiles o de otros ámbitos normativos[5]. La ciencia forense abarca un amplio rango de disciplinas, cada una con su propio conjunto de tecnologías y prácticas. El Instituto Nacional de Justicia (NIJ) divide esas disciplinas en doce categorías: toxicología general, balística y trazas instrumentales; documentos cuestionados; evidencia de traza (tales como análisis de cabellos y fibras); sustancias controladas; exploración biológica/serológica (incluyendo el análisis de ADN); análisis de incendios y de sus escombros; evidencia impresa (huellas de calzado, neumáticos, etc.); evidencia de patrones de sangre; inspección ocular; investigación médico-legal de la causa de la muerte; y la evidencia digital[6]. En los próximos años, la ciencia

[5] Definición de «ciencia forense» proporcionada por la Comisión Nacional de Ciencia Forense en su documento estratégico, «Defining forensic science and related terms». Publicado el 30 de mayo de 2015. www.justice.gov/ncfs/file/786571/download.

[6] Véase: National Institute of Justice, 2006. www.ojp.usdoj.gov/nij/pubs-sum/213420.htm.

y la tecnología probablemente nos ofrezcan poderosas herramientas en el dominio forense —quizá la capacidad de comparar poblaciones de bacterias en el intestino o patrones de búsqueda en Internet—.

Históricamente, la ciencia forense ha sido utilizada primordialmente en las dos fases del proceso de justicia penal: (1) *investigación*, que busca identificar al probable perpetrador de un delito, y (2) *procesamiento*, que busca probar la culpabilidad de un acusado más allá de toda duda razonable (En años recientes, la ciencia forense —particularmente el análisis de ADN— ha sido también ampliamente utilizada para cuestionar condenas del pasado). Es importante resaltar que las fases de investigación y de procesamiento conllevan diferentes estándares en el uso de la ciencia forense, así como otras herramientas de investigación. En las investigaciones, las perspectivas e informaciones pueden proceder de ciencia bien establecida o de aproximaciones exploratorias[7]. En la fase de procesamiento, la ciencia forense debe satisfacer un estándar más alto. Específicamente, las reglas de la prueba federales requieren que el testimonio experto esté basado, entre otras cosas, sobre «principios y métodos fiables» que hayan sido «fiablemente aplicados» a los hechos del caso[8]. Y la Corte Suprema ha establecido que los jueces deben determinar «si el razonamiento o metodología subyacente al testimonio experto es científicamente válido»[9].

Aquí es donde los estándares jurídicos y científicos se interseccionan. Las decisiones de los jueces sobre la admisibilidad de la prueba científica descansan únicamente en estándares *jurídicos*; ellas son exclusiva potestad de los tribunales. Pero la cuestión preponderante a analizar para tomar esa decisión judicial es la validez científica[10]. Y pertenece a la competencia de la comunidad científica proporcionar directrices sobre los estándares científicos de validez científica[11].

Resulta oportuna una mirada al lado científico de esta intersección porque ha quedado cada vez más claro en estos años que la falta de rigor en la valoración de la validez científica de la prueba forense no es un problema hipotético sino una real y significativa debilidad del sistema judicial. Como se relata en la sección 2, las revisiones que han hecho organismos competentes de los fundamentos científicos de las disciplinas forenses y el uso en los juzgados de la prueba basada en esas disciplinas han revelado una preocupante frecuencia en el uso de pruebas forenses que no superan una prueba objetiva de validez científica.

---

[7] Aunque los métodos de investigación no necesitan cumplir los estándares de fiabilidad requeridos bajo las Reglas de la Evidencia Federal, deben estar basados en principios y prácticas científicas sanas para evitar falsas acusaciones.

[8] Regla de la Evidencia Federal 702.

[9] *Daubert v. Merrell Dow Pharmaceuticals*, 509 U.S. 579 (1993) en 592.

[10] *Daubert*, en 594.

[11] En este informe, PCAST aborda únicamente los estándares *científicos* para la validez y fiabilidad científicas. No ofrecemos opinión sobre los criterios *jurídicos*.

La revisión más completa hecha hasta la fecha fue realizada por un comité del Consejo de Investigación Nacional (NRC) codirigido por el Juez Harry Edwards del Tribunal de Apelación para el Distrito del Circuito de Columbia y Constantine Gatsonis, director del Centro para las Ciencias Estadísticas de la Universidad de Brown. Con el mandato del Congreso en un proyecto de ley convertido en ley a finales del año 2005, el estudio empezó en otoño de 2006 y el comité emitió su informe en febrero de 2009[12].

El informe del NRC (2009) describió un perturbador patrón de deficiencias común a muchos de los métodos rutinariamente utilizados en el sistema de justicia penal, de forma particularmente importante la falta de rigor y de estudios apropiados que estableciesen su validez científica, concluyendo que «muchas pruebas forenses —incluyendo, por ejemplo, identificaciones por marcas de mordeduras, balística y de herramientas— se introducen en los procesos penales sin ninguna validación científica significativa, sin determinación de tasas de error o sin una comprobación de su fiabilidad que explique los límites de la disciplina[13].

En el año 2013, después de una prolongada discusión de los resultados y recomendaciones del informe NRC, dentro y fuera del gobierno federal, el Departamento de Justicia (DOJ) —en colaboración con el Instituto Nacional de Estándares y Tecnología (NIST)— estableció la Comisión Nacional de la Ciencia Forense (NCFS) como un cuerpo consultivo federal con el encargo de proporcionar directrices en ciencia forense y recomendaciones políticas a la Fiscalía General; codirigido por el Fiscal General Adjunto y el Director del NIST, los 32 miembros del NCFS, incluidos ocho científicos del ámbito académico y otros cinco científicos doctorados; los demás miembros incluyen jueces, abogados y analistas forenses. Para fortalecer la ciencia forense de forma más general, en el año 2014 el NIST estableció la Organización para los Comités de Áreas Científicas para la Ciencia Forense (OSAC) con el fin de «coordinar el desarrollo de los estándares y directrices… para mejorar la calidad y consistencia del trabajo en la comunidad de la ciencia forense»[14].

En septiembre de 2015, el Presidente Obama pidió a su Consejo de Asesores en Ciencia y Tecnología (PCAST) que explorara, a la luz del trabajo hecho por NCSF y OSAC, qué esfuerzos adicionales pudieran contribuir a fortalecer las disciplinas de la ciencia forense y asegurar la fiabilidad científica de la prueba forense utilizada en el sistema jurídico de la Nación. Después de revisar las actividades en marcha y la literatura jurídica y científica relevante —incluyendo, particularmente, las valoraciones científicas y jurídicas del informe del NRC (2009)— el PCAST concluyó que hay dos importantes lagunas: (1) la necesidad de clarificar el significado científico de los «principios y métodos fiables» y de la «validez científica» en el contexto de ciertas

[12] National Research Council, 2009.
[13] *Ibid*., 107-8.
[14] Véase: www.nist.gov/forensics/organization-scientific-area-committees-forensic-science

disciplinas forenses y (2) la necesidad de evaluar métodos forenses específicos para determinar si han sido científicamente establecidos como válidos y fiables.

Dentro de la amplia gama de disciplinas forenses, elegimos reducir nuestra atención hacia técnicas que referimos aquí como métodos forenses de «comparación de características» (véase el cuadro 1)[15]. La limitación del análisis a estos métodos, fue una motivación para hacer nuestra tarea manejable dentro de los límites de tiempo y recursos disponibles, no obstante elegimos esta particular clase de métodos porque: (1) son comúnmente utilizados en casos penales; (2) han atraído un alto grado de preocupación con respecto a su validez (*e.g*, el informe del NRC [2009]); y (3) todos pertenecen a la misma amplia disciplina científica, la *metrología*, que es la «ciencia de la medición y de su aplicación», en este caso midiendo y comparando características[16].

CUADRO 1.
MÉTODOS FORENSES DE COMPARACIÓN DE CARACTERÍSTICAS

El PCAST utiliza el término «métodos de comparación de características» para referirse a la amplia variedad de métodos que se dirigen a determinar si una muestra que constituye una evidencia (por ejemplo, recogida en la escena del crimen) está o no asociada con una muestra de su origen potencial (por ejemplo, del sospechoso) basada en la presencia de patrones, impresiones, propiedades, o características similares en la muestra y en la fuente. Como ejemplos señalamos los análisis de ADN, cabellos, huellas dactilares latentes, de armas y munición utilizada, herramientas y sus marcas, de huellas de calzado y de neumáticos, mordeduras y escritura a mano.

El PCAST comenzó este estudio formando un grupo de trabajo con seis de sus miembros recogiendo información para su consideración[17]. Para instruirse a sí mismo sobre aspectos prácticos relacionados con la interacción entre ciencia y derecho,

[15] Vale la pena señalar que hay cuestiones relacionadas con la validez científica de otros tipos de pruebas forenses que se escapan del ámbito de este informe, pero que requieren urgente atención —incluyendo principalmente la ciencia que estudia los incendios y el traumatismo craneoencefálico abusivo, comúnmente referido como «Síndrome del bebé sacudido»—. Además, un área de gran importancia no abordada por este informe es la de los métodos científicos para la evaluación de la causalidad —por ejemplo, si la exposición a determinada sustancia fue la causa que probablemente ocasionó un daño a un individuo—.

[16] *International Vocabulary of Metrology – Basic and General Concepts and Associated Terms*, 2012.

[17] Dos de los miembros han estado implicados en la ciencia forense. El codirector del PCAST, Eric Lander, ha trabajado en varios roles científicos (como testigo experto en *People v. Castro* 545 N.Y.S. 2d 985 (Corte Suprema, 1989), un caso relevante y novel sobre la calidad del análisis de ADN que se expone en la página 31; testigo del Tribunal en *U.S. v. Yee*, 134 F.R.D. 161 en 1991; miembro del panel del NRC sobre el análisis forense de ADN en 1992; coautor científico con un científico forense del Laboratorio del FBI en 1994; y miembro de la Junta de Directores del Proyecto Inocencia desde el año 2004 hasta el presente). Ninguno de estos roles ha sido remunerado. El miembro del PCAST S. James Gates Jr. ha sido miembro desde su inicio de la Comisión Nacional sobre Ciencia Forense.

el PCAST consultó a un grupo de asesores con experiencia (listados al comienzo del documento), compuesto por nueve jueces federales, actuales o anteriores, un ex Procurador General de los Estados Unidos y un juez de la Corte Suprema Estatal, dos decanos de facultades de derecho y dos estadísticos con experiencia en esta materia. El PCAST buscó información de un grupo adicional de expertos y partes interesadas, que incluyó científicos y analistas forenses, jueces, fiscales, abogados defensores, defensores de la reforma de la justicia penal, estadísticos, académicos y representantes de agencias federales (véase Apéndice B). La información se recabó a través de múltiples reuniones personales y conferencias telefónicas, incluyendo una sesión en una reunión del PCAST el 15 de enero de 2016. El PCAST también llevó a cabo el inusual paso de iniciar una convocatoria abierta en línea para ampliar las aportaciones, en particular de la comunidad de analistas en ciencia forense; se recibieron más de 70 respuestas[18].

El PCAST también compartió un borrador de este informe con el NIST y el DOJ, que proporcionaron comentarios detallados y útiles que fueron cuidadosamente considerados en la revisión de este informe.

El PCAST expresa su gratitud a todos los que han compartido sus opiniones. Su voluntad de colaborar con el PCAST no implica su apoyo a los puntos de vista expresados en este informe. La responsabilidad de las opiniones, resultados y recomendaciones expresados en este informe, así como como cualquier error de hecho o de interpretación, sólo es imputable al PCAST.

El resto de nuestro informe está organizado de la siguiente manera:

— La sección 2 proporciona una breve visión general de los resultados de otros estudios relacionados con la práctica forense y el testimonio basado en ella y revisa también las acciones federales recientemente llevadas a cabo para fortalecer la ciencia forense.

— La sección 3 revisa brevemente el papel de la validez científica dentro del sistema jurídico. Describe la importante distinción entre estándares jurídicos y estándares científicos.

— La sección 4 después describe los estándares científicos que estarían implicados cuando se habla de «principios y métodos fiables» y «validez científica» en la medida en que se aplican a métodos de comparación de características, también ofrece criterios claros para que puedan ser fácilmente aplicados por los tribunales.

— La sección 5 ilustra la aplicación de los criterios indicados para usarlos en la evaluación de la validez científica de seis importantes métodos de comparación de características: análisis de ADN de muestras de una única fuente y mezcla simple, análisis de ADN con mezclas complejas, análisis de mordeduras, análisis de impre-

[18] Véase: www.whitehouse.gov/sites/default/files/microsites/ostp/PCAST/pcast_forensics_request_for_information.pdf.

siones latentes, análisis balísticos y análisis de huellas de calzado. También analizamos una evaluación llevada a cabo por otros respecto de un séptimo método, el análisis de cabellos.

— En las Secciones 6 a 9 ofrecemos recomendaciones basadas en los resultados de las Secciones 4 y 5, concernientes a acciones federales que pudieran ponerse en marcha para fortalecer la ciencia forense y promover su más riguroso uso en las salas de audiencia.

2. Trabajo previo sobre validación de métodos de ciencia forense

Los desarrollos de las dos últimas décadas —incluyendo la exoneración de acusados que habían sido erróneamente condenados debido en parte a la evidencia en ciencia forense, a una serie de estudios sobre los fundamentos científicos de las disciplinas forenses, revisiones del testimonio de experto basado en hallazgos forenses y a escándalos en laboratorios criminalísticos estatales— han suscitado una atención creciente sobre la cuestión de la validez y fiabilidad de algunos métodos de prueba forense y el testimonio basados en ellos (Sobre definiciones de términos clave como validez y fiabilidad científica, véase Cuadro 2 en la página 57).

En esta sección revisamos brevemente esta historia para dar nuestra valoración del estado actual de los métodos de la ciencia forense y su validez, así como el camino a seguir[19].

2.1. Prueba de ADN y condenas erróneas

Irónicamente, fue la aparición y maduración de una nueva ciencia forense, el análisis de ADN, la primera en cuestionar seriamente la validez de muchas de las disciplinas forenses tradicionales. Cuando los acusados que fueron condenados con la ayuda de prueba forense de esas disciplinas tradicionales comenzaron a ser exonerados sobre la base de persuasivas comparaciones de ADN, se puso en marcha una investigación más profunda sobre su validez científica. La manera en la que esto llegó a suceder nos proporciona un contexto útil para nuestra actual investigación.

Cuando la prueba de ADN se introdujo por primera vez en los tribunales, comenzando a finales de los 80, fue inicialmente considerada como infalible. Pero los métodos utilizados en los primeros casos resultaron poco fiables: los laboratorios de pruebas carecían de procedimientos validados y aplicados consistentemente para definir los patrones de ADN a partir de las muestras, para declarar si dos patrones

[19] Para realizar este resumen, nos apoyamos particularmente en el informe del Consejo de Investigación Nacional (National Research Council, 2009) y en el informe de las Academias Nacionales de Ciencias, Ingeniería y Medicina (National Academies of Sciences, Engineering, and Medicine, 2015).

coincidían dentro de una cierta tolerancia y para determinar la probabilidad de que tales coincidencias pudieran ocurrir por mero azar en la población [20].

Cuando la prueba de ADN se consideró inadmisible en *People v. Castro*, un caso acaecido en Nueva York en 1989, los científicos —incluyendo los de la Academia Nacional de Ciencias y de la Oficina Federal de Investigación (FBI)— convinieron en promover conjuntamente el desarrollo de principios y métodos fiables, haciendo posible que el análisis de ADN de muestras procedentes de una única fuente haya llegado a ser el «estándar de oro» de la ciencia forense tanto para la investigación como para el procesamiento [21].

Tanto el reconocimiento inicial de serios problemas como el desarrollo subsiguiente de procedimientos fiables se vieron favorecidos por la existencia de una robusta comunidad de biólogos moleculares que utilizaban los análisis de ADN en aplicaciones no forenses, tales como las ciencias biomédicas y las agrícolas. También contaron con la ayuda de jueces que reconocieron que este poderoso método forense solo debía ser admitido en los tribunales una vez que fuera establecida apropiadamente su fiabilidad.

Una vez que el análisis de ADN llegó a ser un método fiable, el poder de la tecnología —incluyendo su capacidad para analizar pequeñas muestras y distinguir entre individuos— hizo posible no solo la identificación y condena de los verdaderos autores de los delitos sino también el descarte de sospechosos falsamente acusados antes del procesamiento y el reexamen de una serie de condenas del pasado. Algunos informes del Instituto Nacional de Justicia (NIJ) [22] y de otros organismos, han resaltado que los tests de ADN en el curso de las investigaciones han descartado a decenas de miles de sospechosos. El proceso de reexamen de casos pasados basado en el ADN, además, ha conducido hasta la fecha a la exoneración de 342 condendos, incluyendo 20 que habían sido sentenciados a pena de muerte, y a la identificación de 147 verdaderos culpables [23].

---

[20] Véanse: Lander, 1989: 501-5; Lander y Budowle, 1994: 735-8; Kaye, 1993: 101-72; Roberts, 1991: 1721-1723; Thompson y Ford, 1990: 38-43; Neufeld y Colman, 1991: 46-53.

[21] *People v. Castro* 545 N.Y.S.2d 985 (Corte Suprema 1989). Un caso en el que un conserje fue acusado del asesinato de una mujer en el Bronx fue uno de los primeros casos penales en los que se aplicó el análisis de ADN en los Estados Unidos. El Tribunal estuvo un total de quince semanas en audiencia previa al juicio debatiendo la admisibilidad de la prueba de ADN. Al término de la audiencia, los analistas independientes, tanto de la defensa como de la acusación, acordaron unánimemente que la prueba de ADN presentada no era científicamente fiable —y el Juez declaró la prueba inadmisible—. Véase: Lander, 1989: 501-505. Estos sucesos condujeron finalmente a dos informes del NRC sobre análisis de ADN forense, en 1992 y en 1996, y a la fundación del Proyecto Inocencia (www.innocenceproject.org).

[22] Los tests de ADN han excluido entre un 20-25% de los sospechosos iniciales en casos de agresiones sexuales. U.S. Department of Justice, Oficina de Programas de Justicia, National Institute of Justice, 1996.

[23] Innocence Proyect, «DNA Exonerations in the United States». Véase: www.innocenceproject.org/dna-exonerations-in-the-united-states.

Revisiones independientes de estos casos han revelado que muchos se basaron, en parte, en testimonios periciales defectuosos de científicos forenses que habían dicho a los miembros del Jurado que las características similares en un par de muestras, una tomada del sospechoso y otra encontrada en la escena del crimen (por ejemplo, cabellos, proyectiles, mordeduras, bandas de rodadura de neumáticos y huellas de calzado, u otros artículos), implicaban al acusado en el delito con un alto grado de certeza[24]. De acuerdo con las revisiones, estos errores no fueron simplemente un asunto consistente en una serie de analistas testificando conclusiones que se revelaron incorrectas; sino que, por el contrario, reflejaban un problema sistémico —el testimonio se basó en métodos e incluyó afirmaciones sobre su precisión que estaban encubiertas por una supuesta respetabilidad científica, pero realmente nunca habían estado sujetos a un escrutinio científico significativo[25]—.

2.2. Estudios de métodos de ciencia forense específicos y prácticas de laboratorio

Las preguntas que el análisis de ADN había suscitado sobre la validez científica de las tradicionales disciplinas forenses y el testimonio basado en ellas condujeron, naturalmente, a crecientes esfuerzos para comprobar empíricamente la fiabilidad de los métodos que empleaban esas disciplinas. Análogamente, se realizaron indagaciones dirigidas a examinar las prácticas de recogida, almacenamiento y análisis de la evidencia forense en los laboratorios de criminalística de todo el país. El laboratorio del FBI, ampliamente considerado como uno de los mejores del país, jugó un importante papel en las investigaciones recién mencionadas, reevaluando sus propias prácticas y las de los otros. En lo que sigue, resumimos algunos de los resultados clave de los estudios de los métodos y las prácticas que se produjeron en el caso de las disciplinas de «comparaciones» que constituyen el centro de atención de este informe.

2.2.1. Examen del plomo de los proyectiles

Desde los años 60 hasta el año 2005, el FBI utilizó el análisis de la composición del plomo de los proyectiles como una herramienta forense para identificar su origen. Sin embargo, un informe de NRC encargado por el FBI y emitido en el año 2004 puso en entredicho la validez de los fundamentos de las identificaciones basadas en esa disciplina. La técnica conllevaba comparar la cantidad de los distintos elementos de los proyectiles encontrados en la escena del crimen con la de proyectiles no utilizados, con el fin de determinar si los proyectiles provenían de la misma caja de

[24] Por ejemplo, véase: Gross y Shaffer (2012) disponible en: www.law.umich.edu/special/exoneration/Documents/exonerations_us_1989_2012_full_report.pdf. Veáse también: Saks, y Koehler, 2005: 892-5.

[25] Garret y Neufeld, 2009: 1-97; National Research Council, 2009, 42-43.

munición. El informe NRC 2004 dictaminó que no había base científica para realizar tal determinación[26]. Mientras que el método para determinar las concentraciones de los diferentes elementos dentro de un proyectil fue considerado fiable, el informe dictaminó que no había suficiente investigación y datos para sostener una conexión, basada en la similitud de composición, entre un proyectil particular y un lote determinado de munición, lo que comúnmente es lo relevante en un caso penal[27]. En el año 2005, el FBI anunció que abandonaba la práctica del examen del plomo de los proyectiles, señalando que mientras «apoyaba firmemente el fundamento científico del análisis del plomo de los proyectiles», la fabricación y la distribución de los proyectiles era demasiado variable como para hacer fiable el cotejo positivo de una comparación[28].

2.2.2. Huellas dactilares latentes

En el año 2005, un comité internacional establecido por el FBI dictó un informe sobre defectos en las prácticas del FBI en identificación dactilar que había conducido a una relevante falsa identificación. Basado casi enteramente en una huella latente recuperada en el atentado con bombas en el sistema de trenes de cercanías de Madrid en el año 2004, el FBI detuvo erróneamente a un estadounidense en Portland (Oregón) y lo mantuvo durante dos semanas como *testigo material*[29]. Un analista del FBI concluyó que las huellas coincidían con «un 100% de certeza», aunque las autoridades españolas no pudieron confirmar el cotejo positivo[30]. El comité de revisión concluyó que la falsa identificación del FBI había ocurrido, principalmente,

---

[26] National Research Council, 2004. El examen del plomo de los proyectiles, también conocido como Análisis Composicional del Plomo de los Proyectiles (CABL), conlleva la comparación de la composición elemental de los proyectiles encontrados en la escena del crimen con la de los cartuchos no utilizados en posesión del sospechoso. Esta técnica asume que (1) la fuente fundida utilizada para producir un mismo «lote» de proyectiles tiene la misma composición por todas partes, (2) que no es posible que dos fuentes fundidas distintas tengan la misma composición, y (3) que los proyectiles con distintas composiciones no se han mezclado durante el proceso de fabricación o de envío. Sin embargo, en la práctica esto no es así. El informe del NRC (2004) dictaminó que los volúmenes composicionalmente indistinguibles de plomo pueden producir desde pequeños números de lotes de proyectiles —del orden de 12.000— a lotes más grandes —con más de 35 millones—. El informe también dictaminó que no hay seguridad de que distintos volúmenes de plomo producidos en distinto tiempo y lugar sean indistinguibles. Ni los científicos ni los fabricantes de proyectiles son capaces de declarar de forma contundente el significado de una asociación entre proyectiles en el curso de un examen de plomo de proyectil. Lo más que se puede decir es que los proyectiles que son indistinguibles por CABL podrían haber procedido de la misma fuente.

[27] Faigman, Cheng, Mnookin, Murphy, Sander y Slobogin Eds., 2016.

[28] Oficina Federal de Investigación. *FBI Laboratory Announces Discontinuation of Bullet Lead Examinations.* (1 de septiembre de 2005, nota de prensa) www.fbi.gov./news/pressrel/press-releases/fbi-laboratory-announces-discontinuation-of-bullet-lead-examinations (última consulta el 6 de mayo de 2016).

[29] Stacey, 2005.

[30] Application for Material Witness Order and Warrant Regarding Witness: Brandon Bieri Mayfield, In re Federal Grand Jury Proceedings 03-01, 337 F. Supp. 2d 1218 (D. Or. 2004) (No. 04-MC-9071).

como resultado de un «sesgo de confirmación»[31]. De forma similar, un informe de la Oficina del Inspector General del DOJ subrayaba que fue «el razonamiento inverso», desde la huella conocida a la imagen de la huella latente, lo que condujo a centrar la atención de forma exagerada en las aparentes similitudes y a una inadecuada falta de atención a las diferencias entre las imágenes[32].

2.2.3. Análisis de cabellos

En el año 2002, los científicos del FBI utilizaron la secuenciación del ADN mitocondrial para reexaminar 170 comparaciones de cabellos microscópicas que los científicos de la agencia habían realizado en casos penales. El análisis de ADN demostró que en un 11% de los casos en los que los analistas del FBI habían dictaminado un cotejo positivo entre las muestras de cabellos comparadas microscópicamente el test de ADN de las muestras reveló que realmente procedían de individuos diferentes[33]. Estas falsas asociaciones pueden no haber sido fallos del analista a la hora de realizar los análisis; en lugar de ello, las características coincidentes pudieron haberse producido por casualidad. El estudio demostró que el poder de la comparación microscópica de cabellos entre muestras procedentes de diferentes fuentes era mucho menor que el previamente asumido. (Por ejemplo, anteriores estudios sugerían que la tasa de falsos positivos de los análisis de cabellos entraba en un rango de 1 en 40.000)[34].

2.2.4. Mordeduras

Un estudio llevado a cabo en 2010 sobre mordeduras realizadas por personas claramente identificadas en el experimento halló que la deformación de la piel distorsiona las mordeduras tan sustancial y variablemente que los procedimientos actuales para comparar mordeduras son incapaces de excluir o implicar fiablemente a un sospechoso como potencial mordedor. («Los datos derivados demostraron carencias de correlación y reproducibilidad, es decir, la misma dentición no podía crear una impresión medible que fuera consistente en todos los parámetros, sea cuales fueren las circunstancias del test»)[35]. Un estudio reciente de la *American Board of Foren-*

---

[31] Específicamente, la semejanza entre las dos impresiones, combinada con la presión inherente de trabajar en casos de alto perfil, influyó en el juicio inicial del analista: características ambiguas fueron interpretadas como puntos semejantes y las diferencias entre las impresiones fueron dejadas de lado. Un segundo analista, no protegido frente a las conclusiones del primero, simplemente confirmó sus resultados. Véase: Stacey, 2005.

[32] U.S. Department of Justice, Office of the Inspector General, 2006. oig.justice.special/s0601/final.pdf.

[33] Houck y Budowle, 2002: 964-967.

[34] Gaudette y Keeping, 1975: 599-606. Este estudio fue citado recientemente por el DOJ para sostener la afirmación de que el análisis de cabellos es una metodología científica válida y fiable. www.justice.gov/dag/file/877741/download. El tema del análisis de cabellos se trata en la sección 5.

[35] Bush, Cooper y Dorion, 2010: 976-83. Véase también: Bush, M.A., Miller, Bush, PJ., et al., 2009: 167-76.

sic Odontology (Junta Directiva Norteamericana de Odontología Forense) también mostraba una inquietante falta de consistencia en la forma en que los odontólogos forenses analizaban las marcas de mordeduras, que incluían incluso decisiones sobre si había prueba suficiente para determinar si una mordedura fotografiada era una mordedura humana[36]. En febrero de 2016, tras seis meses de investigación, la *Texas Forensic Science Commission* (Comisión de Ciencia Forense de Texas) recomendó una moratoria en el uso de las identificaciones de mordeduras en juicios penales, concluyendo que la validez de la técnica no había sido científicamente establecida[37].

Estos ejemplos ilustran cómo diversos métodos de comparación de características que han tenido extendido uso no han sido sometidos, sin embargo, a tests significativos sobre su validez científica o para medir su fiabilidad.

2.3. Testimonio sobre la prueba forense

Al revisarse transcripciones de juicios orales, se ha comprobado que los testimonios periciales han sobreestimado frecuentemente el valor probatorio de sus evidencias, yendo mucho más allá de lo que la ciencia relevante puede justificar. Por ejemplo, algunos analistas han testificado:

— Que sus conclusiones son «100% ciertas»; tienen tasas de error igual a «cero», «esencialmente cero», «vanamente apreciable», «despreciable», «mínima» o «microscópica»; o tienen una probabilidad de error tan remota que es «prácticamente imposible»[38]. Como muchas revisiones han señalado, sin embargo, tales expresiones no son científicamente defendibles. Ningún test de laboratorio o análisis de comparación de características tiene error cero, incluso si un analista recibiera la máxima puntuación en la ejecución de un test particular consistente en un número limitado de muestras[39]. Incluso los tests altamente automatizados no tienen tasa de error igual a cero[40] [41].

— Que pueden «individualizar» la evidencia —por ejemplo, utilizando marcas sobre un proyectil para atribuirlo a un arma específica «hasta la exclusión de cual-

---

[36] Balko, 2015. www.washingtonpost.com/news/the-watch/wp/2015/04/08/a-bite-mark-matching-advocacy-group-just-conducted-a-study-that-dicredits-bite-mark-evidence.; Freeman y Pretty, 2015. (los datos están disponibles por parte de los autores previa solicitud).

[37] Texas Forensic Science Commission, 2016. www.fsc.texas.gov/sites/default/files/FinalBiteMarkReport.pdf.

[38] Thompson, Taroni y Aitken, 2003: 1-8. Thompson, 2013; Cole, 2005: 985-1078; y Koehler, J.J. «Forensics or fauxrensics? Ascertaining accuracy in the forensic sciences» papers.ssrn.com/sol3/papers.cfm?abstract_id=2773255 (última consulta el 28 de junio de 2016).

[39] Cole, 2005: 985-1078; y Koehler, Forensics or fauxrensics? Ascertaining accuracy in the forensic sciences» papers.ssrn.com/sol3/papers.cfm?abstract_id=2773255 (última consulta el 28 de junio de 2016).

[40] Thompson, Taroni y Aitken, 2003: 1-8.

[41] Los resultados de falsos positivos pueden proceder de dos fuentes: (1) similitud casual entre dos características y (2) fallos humanos/técnicos. Véase discusión en la sección 4 páginas 60-62.

quier otra arma de fuego en el mundo»— es una aseveración que no es sostenible por la ciencia relevante[42].

— Que un resultado sea verdadero «hasta un razonable grado de certeza científica». Esta frase no tiene un significado generalmente aceptado en la ciencia y está abierta a amplias y diversas interpretaciones por parte de diferentes científicos[43]. Además, la expresión puede tomarse como si implicara certeza.

2.3.1. Revisión del DOJ del testimonio sobre el análisis de cabellos

En el año 2012, el DOJ y el FBI anunciaron que iniciarían una revisión formal del testimonio forense en 3000 casos penales relacionados con el análisis microscópico de cabellos. Los resultados iniciales de esta revisión sin precedentes, realizados en consulta con el *Innocence Project* (Proyecto Inocencia) y la *National Association of Criminal Defense Lawyers* (Asociación Nacional de Abogados Defensores en materia penal), mostraron que los analistas del FBI habían proporcionado testimonio científicamente inválido en el 95 % de los casos en los que esa prueba fue utilizada para inculpar al acusado en el juicio. Estos problemas fueron sistemáticos: 26 de 28 analistas de cabellos del FBI que testificaron en 328 casos proporcionaron testimonio científicamente inválido[44, 45].

La importancia de la revisión del análisis de cabellos realizada por el FBI fue puesta de relieve por la decisión de enero de 2016 del juez de la Corte Superior de Massachusetts, Robert Kane, de anular la condena de George Perrot basada en parte en el reconocimiento por parte del FBI de los errores en el análisis de cabellos[46].

---

[42] Véanse: National Research Council, 2008.; Saks y Koehler, 2008: 199-218.

[43] National Comission of Forensic Science, «Recommendations to the Attorney General Regarding Use of the Term "Reasonable Scientific Certainty"», 2016, disponible en: www.justice.gov/ncfs/file/839726/download. El NCFS dictamina que «las conclusiones de las disciplinas forenses a menudo se testifican como apoyadas hasta un razonable grado de certeza científica" o "hasta un cierto grado de certeza [de la disciplina]". Estos términos no tienen significado científico y pueden confundir a los investigadores de los hechos sobre el nivel de objetividad envuelto en el análisis, su fiabilidad científica y limitaciones, así como sobre la capacidad del análisis para alcanzar una conclusión».

[44] Federal Bureau of Investigation. *FBI Testimony on Microscopic Hair Analysis Contained Errors in at Least 90 Percent of Cases in Ongoing Review*, (20 de abril de 2015, nota de prensa). www.fbi.gov/news/pressrel/press-releases/fbi-testimony-on-microscopic-hair-analysis-contained-errors-in-at-least-90-percent-of-cases-in-ongoing-review.

[45] Las expresiones erróneas se clasificaron en tres categorías, en las que el analista: (1) declaró o insinuó que el cabello encontrado pudiera estar asociado con un específico individuo hasta la exclusión de todos los demás; (2) asignó hasta la asociación positiva una valoración estadística o una probabilidad de que el cabello probatorio procedía de una fuente en particular; o (3) citó el número de casos realizados en el laboratorio y el número de comparaciones positivas exitosas para sustentar una conclusión de que el cabello encontrado pertenecía a un específico individuo. Reimer, «The hair microscopy review project: An historic breakthrough for law enforcement and a daunting challenge for the defense bar». *The Champion*, (Julio 2013): 16. www.nacdl.org/champion.aspx?id=29488.

[46] *Commonwealth v. Perrot*, No. 85-5415, 2016 WL 380123 (Mass. Super. Man. 26, 2016).

2.3.2. Revisión expandida del DOJ

En marzo de 2016, el DOJ anunció su intención de expandir su revisión del testimonio pericial por el laboratorio del FBI en casos penales cerrados a métodos de ciencia forense adicionales. La revisión proporcionará oportunidades de valorar hasta qué punto hayan podido ocurrir similares testimonios sobrevalorados en otras disciplinas[47]. El DOJ planea establecer una estructura para auditar muestras de testimonios procedentes de Unidades del FBI que emplean tipos de pruebas adicionales basadas en características como el rastro de las impresiones que las armas dejan sobre los proyectiles, huellas de calzado, fibras, suelo y otras evidencias en la escena del crimen.

2.4. Sesgos cognitivos

Además de las cuestiones anteriormente descritas, los científicos han estudiado un problema más sutil pero igualmente importante que afecta a la fiabilidad de las conclusiones en muchos campos, incluyendo la ciencia forense: los sesgos cognitivos. Los sesgos cognitivos son los modos en que las percepciones y juicios humanos pueden alterarse por factores distintos a los relevantes para la decisión que se ha de tomar en un momento dado. Incluye el «sesgo de contexto» por el que los individuos se ven influenciados por información de contexto irrelevante; el «sesgo de confirmación» por el que los individuos interpretan información, o buscan nueva evidencia, de modo tal que se adecúe a las creencias o asunciones preexistentes; y «evitar la disonancia cognitiva» por la que los individuos son reacios a aceptar nueva información que sea inconsistente con sus primeras conclusiones. La comunidad científica biomédica, por ejemplo, hace todo lo posible para minimizar los sesgos cognitivos mediante el empleo de protocolos estrictos, tales como pruebas de doble ciego en ensayos clínicos.

Ha habido estudios que han demostrado que los sesgos cognitivos pueden ser un serio problema en la ciencia forense. Por ejemplo, un estudio realizado por Itiel Dror y sus colegas demostró que el juicio de los analistas en huellas dactilares latentes puede estar influenciado por el conocimiento de las decisiones de otros analistas forenses (una forma de sesgo de confirmación)[48]. Estos estudios se explican con mayor detalle en la Sección 5.4. Estudios similares han replicado estos resultados en otros campos forenses, incluyendo la interpretación de mezclas de perfiles de ADN, análisis microscópico de cabellos y en la investigación de incendios[49, 50].

---

[47] Véase: www.justice.gov/dag/file/870671/download.

[48] Dror, Charlton y Peron, 2006: 74-8.

[49] Véase, por ejemplo: Dror y Hampikian, 2011: 204-8; Miller, 1987: 157; y Bieber, 2014, disponible en onlinelibrary.wiley.com/doi/10.1002/9780470061589.fsa1119/abstract.

[50] Véase, generalmente, Dror, 2016: 121-127.

Se han propuesto diversas estrategias para mitigar los sesgos cognitivos en los laboratorios forenses, incluyendo la gestión del flujo de información en un laboratorio criminalístico para minimizar la exposición de los analistas forenses a información contextual irrelevante (tales como confesiones o identificaciones de testigos oculares) y asegurarse de que los analistas trabajan de modo lineal, documentando sus hallazgos sobre las evidencias procedentes de la escena del crimen *antes* de realizar comparaciones con muestras procedentes de un sospechoso [51].

2.5. Estado de la ciencia forense

El estudio del NRC (2009) concluyó que muchas de las dificultades de la ciencia forense posiblemente se deriven de la realidad histórica de que muchos métodos fueron ideados como heurística básica para ayudar en las investigaciones penales y no se fundamentaron en prácticas de validación de la investigación científica [52]. Aunque muchos laboratorios forenses requieren ahora que los analistas en ciencia forense recién contratados tengan un grado en ciencias, muchos analistas en laboratorios forenses no tienen grados avanzados en una disciplina científica [53]. Además, hasta el año 2015 no existían programas de doctorado específicos en ciencia forense en los Estados Unidos (aunque tales programas existían en Europa) [54]. Ha habido una financiación muy limitada para la investigación en ciencia forense, especialmente para estudiar la validez o fiabilidad de estas disciplinas. Resulta bastante reducido el número de revistas de ciencia forense con revisión por pares que sean especializadas en los campos de comparación de características.

Como señalaron el estudio del NRC 2009 y otros, fundamentalmente las ciencias forenses no han alcanzado todavía una «cultura investigadora» bien desarrollada [55]. Es importante destacar que una cultura investigadora incluye los principios: (1) debe presumirse que los métodos no son fiables hasta que se haya establecido la validez de los fundamentos basada en evidencia empírica y (2) incluso entonces, el cuestionamiento científico y la revisión de los métodos debe continuar de manera continua.

---

[51] Kassin, Dror y Kakucka, 2013: 42-52. Véase también: Krane, Ford, S., Gilder, et al., 2008: 1006-1007.

[52] National Research Council, 2009: 128.

[53] National Research Council, 2009: 223-230. Véase también: Conney, 2010: 223-58. («Las áreas en las que no hubo consenso incluían requerimientos de grado (hubo casi una proporción de 50/50 entre agencias que requerían una titulación de 4 años o más frente a las que requerían menos de 4 años o ninguna titulación»).

[54] National Research Council, 2009: 223. Mientras existen diversos programas de doctorado en justicia penal, psicología forense, antropología forense o programas en química o en disciplinas relacionadas que ofrecen una concentración en la ciencia forense, solo el *Sam Houston State University College of Criminal Justice* (Instituto Universitario Estatal de Justicia Criminal Sam Houston) ofrece un programa doctoral en ciencia forense. Véase: www.shsu.edu/programs/doctorate-of-philosophy-in-forensic-science.

[55] Mnookin, Cole, Dror, et al., 2011: 754-8.

En particular, algunos analistas forenses abrazan la noción de que una prolongada «experiencia» en resolver casos puede sustituir a los estudios empíricos de validez científica[56]. El ejercicio de la pericia no es una investigación científicamente válida y solo la experiencia no puede establecer la validez científica. En particular, no es posible que podamos estimar tasas de error a partir trabajo de casos porque ordinariamente no disponemos de conocimiento independiente sobre «fundamento de verdad» o «la respuesta correcta»[57].

Más allá de la cuestión fundamental de la validez científica, históricamente, la mayoría de las disciplinas de comparación de características no prestaron suficiente atención a la importancia de blindar a los analistas frente a información potencialmente sesgada; al desarrollo de medidas objetivas de evaluación e interpretación; a prestar una cuidadosa atención a las tasas de error y su medición; y a desarrollar evaluaciones objetivas del significado de una asociación entre una muestra y su fuente potencial[58].

El informe del NRC (2009) estimuló a algunos miembros de la comunidad científica forense a reconocer esos defectos. Algunos científicos forenses han abrazado la necesidad de colocar la ciencia forense sobre un fundamento científico sólido y han iniciado esfuerzos para lograrlo[59].

2.6. Estado de la práctica forense

Las investigaciones de la práctica forense han desenterrado igualmente problemas procedentes de la falta de una fuerte «cultura sobre la calidad». Específicamente, docenas de investigaciones de laboratorios criminalísticos —primordialmente de nivel estatal o local— han revelado fallos repetidos en el manejo y procesado de las evidencias, así como interpretación incorrecta de los resultados de los análisis forenses[60].

---

[56] Véase Sección 4.7.

[57] Véase Sección 4.7.

[58] National Research Council, 2009: 8, 124, 184-185, 188-191. Véase también: Koppl y Krane, 2016.

[59] Véase Sección 4.8.

[60] Unos cuantos ejemplos de esas investigaciones incluyen:

(1) una investigación independiente durante 2 años en el laboratorio de criminalística del Departamento de Policía de Houston que conllevó la revisión de 3500 casos (Informe Final del Investigador Independiente para el Laboratorio de Criminalística del Departamento de Policía de Houston, preparado por Michael R. Bromwich, 13 de junio de 2007, www.hpdlabinvestigation.org/reports/070613report.pdf);

(2) la investigación y cierre de la Unidad Balística del Laboratorio de Criminalística de la Policía de Detroit tras el descubrimiento de pruebas contaminadas y fallo en el debido mantenimiento de los equipos de pruebas (véase Bunkley, «Detroit police lab is closed after audit finds serious errors in many cases». *New York Times*, 25 de septiembre de 2008: www.nytimes.com/2008/09/26/us/26detroit.html?_r=0);

Varios comentaristas han resaltado una cuestión fundamental que puede subyacer a estos serios problemas: el hecho de que casi todos los laboratorios de criminalística están estrechamente vinculados a la acusación en los casos penales. Esta estructura socava la mayor objetividad que típicamente puede encontrarse en los laboratorios que realizan pruebas en otros campos y crea situaciones en las que el personal puede cometer errores debido a sutiles sesgos cognitivos o a una presión palmaria[61].

El informe del NRC (2009) recomendó que todos los laboratorios forenses públicos y sus instalaciones fueran desvinculados del control administrativo de las agencias encargadas de hacer cumplir las leyes o de las fiscalías[62]. Por ejemplo, Houston —después de disolver su laboratorio criminalístico dos veces en tres años— siguió esta recomendación y, a pesar de una importante oposición política, consiguió el éxito en transformar el laboratorio en un centro de ciencia forense independiente[63].

2.7. Informe del Consejo de Investigación Nacional

El informe del NRC (2009), *Fortaleciendo la Ciencia Forense en los Estados Unidos: un Paso Hacia Adelante*, ha sido la revisión más completa de las ciencias forenses en los Estados Unidos hasta la fecha. El informe dejó claro que los tipos de problemas, irregularidades y errores judiciales que resume no podían atribuirse simplemente a

---

(3) una investigación realizada en 2010 en el laboratorio de criminalística de la Oficina de Investigación del Estado de Carolina del Norte que descubrió que los agentes constantemente retenían pruebas exculpatorias o distorsionaban las pruebas en más de 230 casos en un periodo de 16 años (véase Swecker y Wolf, «An Independent Review of the SBI Forensic Laboratory»: images.bimedia.net/documents/SBI+Report.pdf); y

(4) una revisión en el año 2013 de la utilización de la prueba de ADN, realizada por la Oficina del Médico forense jefe del Hospital de la Ciudad de Nueva York, en más de 230 casos de violación (véase Estado de Nueva York, Oficina del Inspector General. Diciembre de 2013, www.ig.ny.gov/sites/default/files/pdfs/OCMEFinalReport.pdf). Un análisis estimó que, al menos, en 50 laboratorios principales, se emitieron informes falsos, se destruyeron pruebas, fallaron en tests de aptitud, tergiversaron resultados en el testimonio o manipularon drogas entre los años 2005 y 2011. 28 de esos laboratorios estaban acreditados a nivel nacional. Memorando de Marvin Schechter a la Comisión de Ciencia Forense del Estado de Nueva York (25 de marzo de 2011): 243-4. Véase: www.americanbar.org/content/dam/aba/administrative/legal_aid_indigent_defendants/ls_sclaid_def_train_memo_schechter.authcheckdam.pdf.

[61] El informe del National Research Council, 2009: 24-25 dice: «La mejor ciencia se lleva a cabo en un entorno científico y no en un organismo encargado de hacer cumplir la ley. Como consecuencia de que los científicos forenses se conducen en su trabajo por la necesidad de responder a una pregunta particular relacionada con las cuestiones de un caso específico, a veces se enfrentan a la presión de sacrificar su propia metodología en aras de la conveniencia». Véase también: Giannelli, 2010: 247-66 y Thompson, 2015.

[62] National Research Council, 2009: 24.

[63] El Centro de Ciencia Forense de Houston abrió sus puertas en abril de 2014, reemplazando al Laboratorio de Criminalística del Departamento de Policía de Houston. El Centro opera como una «corporación gubernamental local» con sus propios directores, funcionarios y empleados. La estructura fue diseñada a propósito para aislar al Centro de la influencia de policías, fiscales, oficiales o grupos con algún interés especial. Véase: Thompson, 2015: 214.

un conjunto de analistas deshonestos o a laboratorios con bajo rendimiento. En su lugar, el informe descubrió que los problemas que asolan a la comunidad de la ciencia forense son sistémicos y generalizados —resultado de factores entre los que se incluyen un alto grado de fragmentación (incluyendo requerimientos de instrucción y formación, recursos y capacidades de los laboratorios dispares e inadecuados); una falta de estandarización de las disciplinas e insuficiente investigación y formación de alto nivel; y escasez de estudios revisados por pares que establezcan las bases científicas y la validez de muchos métodos forenses rutinariamente utilizados—.

Fueron especialmente prevalentes en las ciencias forenses las deficiencias en las disciplinas de comparación de características. El informe 2009 descubrió que muchas de esas disciplinas carecían de sistemas bien definidos para determinar las tasas de error y que no se habían realizado estudios para establecer la unicidad o rareza relativa o lo que resulta común en las marcas o características particulares examinadas. Además, las pruebas de aptitud, cuando se habían realizado, mostraban casos de bajo rendimiento por parte de determinados analistas. En resumen, el informe concluye que se introduce en los juicios penales «abundante prueba forense sin una seria validación científica, sin determinación de tasas de error, o sin pruebas de fiabilidad que expliquen los límites de la disciplina —incluyéndose, por ejemplo, las identificaciones por mordeduras, balísticas y de trazas instrumentales—» [64].

El informe del NRC (2009) descubrió que los problemas que asolan las ciencias forenses eran tan graves que solo podían ser abordados mediante «un compromiso nacional para revisar la actual estructura que sostiene la comunidad de la ciencia forense en este país» [65]. Subyacente a las 13 recomendaciones básicas del informe yacía una llamada al liderazgo al más alto nivel de los gobiernos federales y estatales y a la promoción y adopción de una agenda a largo plazo para impulsar el desarrollo de la ciencia forense desde su actual situación de debilidad.

El informe del NRC (2009) requirió la realización de estudios para comprobar si una serie de métodos forenses eran válidos en sus fundamentos, incluyendo la realización de tests empíricos sobre la precisión de los resultados. También solicitó la creación de una nueva agencia federal independiente que proporcionara la necesaria supervisión al sistema de ciencia forense; la estandarización de la terminología utilizada en los dictámenes y en el testimonio rendido sobre los resultados de las ciencias forenses; la desvinculación de los laboratorios públicos forenses del control administrativo de las agencias encargadas de hacer cumplir las leyes; la implementación de requisitos de certificación de analistas y programas de acreditación de laboratorios obligatorios; la investigación sobre el sesgo del observador y las fuentes de error humano en los exámenes forenses; el desarrollo de herramientas de medición avanzada, validación, fiabilidad y pruebas de aptitud en la ciencia forense; y el fortalecimiento y desarrollo de grados y programas de formación e instrucción continua.

---

[64] National Research Council, 2009: 107-108.
[65] National Research Council, 2009.

2.8. Progresos recientes

En respuesta al informe del NRC (2009), la Administración Obama inició una serie de reformas dirigidas a fortalecer las ciencias forenses, comenzando con la creación en 2009 de un Subcomité en Ciencia Forense dentro del Comité de Ciencia del Consejo Nacional de Ciencia y Tecnología que recibió el encargo de considerar cómo conseguir del mejor modo posible los objetivos del informe NRC. Debajo se describen con algún detalle las actividades resultantes.

2.8.1. Comisión Nacional de la Ciencia Forense

En el año 2013, el DOJ y el NIST, con el apoyo de la Casa Blanca, firmaron un Memorando de Entendimiento que perfilaba un marco de cooperación y colaboración entre las dos agencias juntando esfuerzos para fortalecer la ciencia forense.

En el año 2013, el DOJ estableció una Comisión Nacional de Ciencia Forense (NCFS), que es un comité consultivo federal para informar al Fiscal General, codirigida por el Fiscal General Adjunto y el Director del NIST; entre los 32 miembros del NCFS se incluyen siete científicos del ámbito académico y otros cinco científicos doctorados; los demás miembros incluyen jueces, abogados y analistas forenses. La Comisión está encargada de proporcionar las recomendaciones políticas al Fiscal General[66]. La NCFS emite recomendaciones formales al Fiscal General, así como «documentos de opiniones» que reflejan el punto de vista de la mayoría de dos tercios del NCFS y que no requieren acciones específicas por parte del Fiscal General. Hasta la fecha, el NCFS ha emitido diez recomendaciones, entre otras cosas, sobre la acreditación de los laboratorios forenses y a la certificación de los analistas forenses, el avance de la interoperatividad de sistemas de información de huellas dactilares impresas, el desarrollo de protocolos de análisis de causa raíz para los proveedores de servicios forenses y la mejora de las comunicaciones entre los médicos examinadores y las oficinas forenses[67]. Hasta la fecha, el Fiscal General ha adoptado formalmente el primer conjunto de recomendaciones sobre acreditación[68] y ha ordenado al Departamento que comience a dar pasos dirigidos hacia alguna de las otras recomendaciones puestas de manifiesto hasta el presente[69].

---

[66] Véase: www.justice.gov/ncfs.

[67] Para una lista completa de documentos aprobados por el NCFS, véase www.justice.gov/ncfs/work-products-adopted-commission.

[68] Departamento de Justicia. «El Departamento de Justicia anuncia nuevas políticas de acreditación para hacer progresar la ciencia forense» (7 de diciembre de 2015, nota de prensa). www.justice.gov/opa/pr/justice-department-announces-new-accreditation-policies-advance-forensic-science.

[69] Memorando del Fiscal General a los jefes de departamento con respecto a las recomendaciones de la Comisión Nacional de Ciencia Forense, 17 de marzo de 2016. www.justice.gov/ncfs/file/841861/download.

En el año 2014, el NIST estableció la Organización de Comités de Áreas Científicas (OSAC), un cuerpo colaborador de más de 600 miembros voluntarios en gran parte provenientes de la comunidad de la ciencia forense[70]. La OSAC se estableció para fomentar el desarrollo de estándares y directrices voluntarias para la consideración de la comunidad de analistas forenses[71]. Su estructura consta de seis Comités de Área Científica (SACs) y 25 Subcomités que trabajan para desarrollar estándares, directrices y códigos deontológicos para cada una de las disciplinas y metodologías forenses[72]. Tres Comités de Recursos Globales proporcionan orientación sobre cuestiones jurídicas, factores humanos y de garantía de la calidad. Todos los documentos desarrollados por las SACs se aprueban por la Junta Directiva de Estándares de Ciencia Forense (FSSB), que es un componente de la estructura de OSAC, para su incorporación al listado del Registro de OSAC de Estándares Aprobados. La OSAC no es un comité consultivo federal.

2.8.2. Fondos federales para la investigación

El gobierno federal ha dado también pasos para abordar un factor que contribuye a los problemas que se constatan en la ciencia forense —la falta de una comunidad de investigación científica robusta y rigurosa en muchas disciplinas de la ciencia forense—. Aunque existen múltiples razones que explican la ausencia de tal comunidad científica, una de ellas es que, a diferencia de la mayoría de las disciplinas científicas, ha habido muy pocos fondos que atraigan y sostengan a un grupo de científicos de excelencia centrados en la *investigación básica* en la ciencia forense.

La Fundación Nacional de la Ciencia (NSF) ha iniciado recientemente esfuerzos para ayudar a reparar esta deficiencia fundacional de la ciencia forense. En el año 2013, la NSF resaltó su interés en esta área y animó a los investigadores a presentar propuestas de investigación dirigidas a las cuestiones fundamentales que puedan hacer avanzar el conocimiento y la formación en las ciencias forenses[73]. Como resultado de un proceso de interagencias conducido por la OSTP y la NSF, en colaboración con el Instituto Nacional de Justicia (NIJ), se invitó a la presentación de propuestas para la creación de nuevos centros de investigación multidisciplinar para su financiación con los presupuestos de 2014[74]. Sobre la base de nuestra revisión de

---

[70] Entre los miembros se incluyen analistas en ciencia forense y otros analistas que representan a agencias federales, estatales y locales, así como al mundo académico y a la industria.

[71] Para más información, véase: www.nist.gov/forensics/osac.cfm.

[72] Los seis Comités del Área Científica de OSAC son: Biología/ADN; Química/Análisis Instrumental; Escena del Crimen/Investigación Médico-forense, Digital/Multimedia; Física/Interpretación de patrones (www.nist.gov/forensics/upload/OSAC-Block-Org-Chart-3-17-2015.pdf).

[73] Véase: Dear Colleague Letter: Forensic Science – Opportunity for Breakthroughs in Fundamental and Basic Research and Education. www.nsf.gov/pubs/2013/nsf13120/nsf13120.jsp.

[74] Los centros que NSF está proponiendo crear son centros de cooperación en investigación industria/universidad (I/UCRCs). I/UCRCs son colaborativos por diseño y pueden ser eficaces para ayudar a tender puentes entre las brechas culturales y científicas de los académicos que trabajan en áreas rele-

los resúmenes de subvenciones, PCAST estima que la NSF compromete un total de 4,5 millones de dólares anuales para fomentar proyectos de investigación extracurriculares sobre ciencia forense fundacional.

El NIST ha dado pasos también dirigidos a este tema creando un nuevo Centro de Excelencia de Ciencia Forense que se denomina Centro para la Estadística y Aplicaciones en Pruebas Forenses (CSAFE), que centrará sus esfuerzos investigadores en mejorar la fundamentación estadística de los análisis de las huellas latentes, balísticas, neumáticos, escritura manuscrita, patrones de manchas de sangre, trazas instrumentales y análisis de comparación de patrones, así como en análisis de sistemas de computación e información, dispositivos móviles, tráfico en la red, redes sociales y en evidencia digital GPS[75]. El CSAFE está financiado bajo un acuerdo de cooperación con la Universidad del Estado de Iowa, para establecer un centro en asociación con investigadores de la Universidad Carnegie Mellon, la Universidad de Virginia y la Universidad de California, Irvine; con una dotación total de veinte millones de dólares en cinco años. El PCAST estima que el NIST compromete un total de cinco millones de dólares al año para financiar proyectos de investigación básica extracurricular sobre ciencia forense, asignando, aproximadamente, 4 millones de dólares al CSAFE y 1 millón aproximadamente a otros proyectos.

El NIJ no tiene presupuesto específicamente asignado para la investigación en ciencia forense. Para que pueda apoyar actividades de investigación, el NIJ debe extraer de su presupuesto base financiación para esas actividades utilizando los programas de asistencia de la Oficina de Programas de Justicia para investigación y estadísticas o los programas de reducción de atrasos en informes de ADN[76]. La mayoría de la ayuda a la investigación está dirigida a la investigación aplicada. Aunque es difícil clasificar los proyectos de investigación del NIJ, estimamos que destina un total de aproximadamente cuatro millones de dólares al año para fomentar proyectos de investigación básica extramuros sobre ciencia forense[77].

---

vantes de la ciencia para el ámbito forense y los analistas forenses. www.nsf.gov/pubs/2014/nsf14066/nsf14066.pdf.

[75] National Institute of Standards and Technology, 2015. www.nist.gov/forensics/center-excellence-forensic052615.cfm.

[76] National Academies of Sciences, Engineering, and Medicine, 2015. De acuerdo con el informe «La financiación otorgada por el Congreso para fomentar los programas de investigación del NIJ disminuyeron desde inicios hasta mediadados de la década del 2000 y permanecen insuficientes, especialmente a la luz de los crecientes retos a los que se enfrenta la comunidad científica forense… Con financiación de partida limitada, el NIJ la financia y desarrolla con los créditos de los programas de reducción de atrasos en informes de ADN y otros programas asistenciales. Estos fondos para salir del paso están apoyando esencialmente la cartera actual de las ciencias forenses del NIJ, pero hay presiones para reducir la cantidad asignada a la investigación para estos programas. En los tres últimos años, la financiación que asiste a estos programas ha disminuido; por tanto, la financiación disponible para la investigación también se ha reducido».

[77] U.S. Department of Justice, National Institute of Justice, 2016.

Incluso con los incrementos recientes, es probable que la financiación extramuros total para la investigación básica en ciencia forense entre la NSF, el NIST y el NIJ esté en el rango de solo 13,5 millones de dólares anuales. El informe del NRC (2009) dijo que:

> La investigación en ciencia forense no está [en general] bien sustentada… En comparación a otras áreas de la ciencia, las disciplinas de la ciencia forense tienen oportunidades extremadamente limitadas para financiar la investigación. Aunque el FBI y el NIJ han sostenido algunas investigaciones en las disciplinas de la ciencia forense, el nivel de soporte ha sido muy inferior con respecto a lo necesario para que la comunidad de ciencia forense establezca lazos fuertes con una amplia base de universidades dirigidas hacia la investigación y la comunidad de investigadores a nivel nacional. Además, la financiación de la investigación en la academia es limitada…, lo que puede inhibir la búsqueda de soluciones a cuestiones científicas fundamentales, esenciales para establecer la base de la ciencia forense. Finalmente, la comunidad de investigadores, en un sentido más amplio, no está generalmente involucrada en llevar a cabo investigación relevante para desarrollar las disciplinas de la ciencia forense[78].

Un informe del NRC del año 2015, *Support for Forensic Science Research: Improving the Scientific Role of the National Institute of Justice* (Apoyo a la Investigación en Ciencia Forense: Mejorando el Rol Científico del Instituto Nacional de Justicia), encontró que la situación de la financiación para la investigación en ciencia forense no había mejorado mucho desde el informe del NRC (2009)[79].

Además, el Centro de Ciencias Forenses de la Defensa ha comenzado recientemente a financiar investigación externa que abarca las disciplinas de la ciencia forense como parte de su misión de proporcionar capacidades especializadas de investigación forense y biométrica y apoyo al Departamento de Defensa. Rediseñado como DFSC en el año 2013, el Centro fue primeramente el Laboratorio de Investigación Criminalística del ejército estadounidense, originalmente encargado de apoyar las investigaciones penales dentro del ámbito militar, pero adicionalmente encargado en el año 2007 de proporcionar «duraderas capacidades forenses expedicionarias», en respuesta en parte a las necesidades de investigación y enjuiciamiento por ataques con explosivos en Irak y Afganistán. Mientras que la mayor parte del soporte de DFSC ha sido tradicionalmente la investigación en análisis de ADN y en bioquímica, el Centro ha destinado recientemente recursos hacia proyectos que abordan carencias básicas críticas en otras disciplinas, incluyendo los análisis de balística y de huellas dactilares latentes.

De forma destacada, el DFSC ha ayudado a estimular la investigación en la comunidad de la ciencia forense. Fruto de las conversaciones entre DFSC y la Sociedad estadounidense de directores de laboratorios de criminalística (ASCLD), esta última organizó una reunión en el año 2011 para identificar las prioridades de investigación para la comunidad de la ciencia forense. El DFSC acordó financiar dos estudios fundamentales para abordar las necesidades de investigación de mayor prioridad identi-

[78] National Research Council, 2009: 78.
[79] National Academies of Sciences, Engineering, and Medicine, 2015: 15.

ficadas por el Comité de Investigación Forense de ASCLD: el primer estudio independiente de análisis de «caja negra» sobre análisis de armas de fuego y un estudio de interpretación de mezclas de perfiles de ADN (véase Sección 5). En el año Fiscal 2015, DFSC asignó aproximadamente 9,2 millones de dólares a la investigación externa en ciencia forense. El 75% del crédito de DFSC sustentó proyectos relacionados con el ADN y bioquímica; el 9% sobre la prueba digital; el 8% sobre pruebas de comparación de patrones distintos al ADN; y el 8% a la química[80]. Como en el caso de NIJ, no existe línea de créditos dedicados a la investigación en ciencia forense en los presupuestos del DFSC; el DFSC, en su lugar, debe solicitar créditos de múltiples fuentes dentro del Departamento de Defensa para sostener esta investigación.

2.8.3. Una carencia crítica: la validez científica

La Administración ha dado pasos importantes y muy necesarios creando mecanismos para discutir la política pública, desarrollar mejores prácticas entre los analistas de métodos específicos y sostener la investigación científica. Al mismo tiempo, el trabajo a fecha de hoy no ha abordado la demanda del informe del NRC (2009) a examinar la validez científica fundamental y la fiabilidad de muchos métodos forenses utilizados cada día en los tribunales. El resto de nuestro informe se centra en ese tema.

3. El papel de la validez científica en los tribunales

El tema central de este informe es la validez científica de las pruebas de ciencia forense —más específicamente, la prueba de métodos científicos de comparación de características (por ejemplo, en muestras de ADN, huellas latentes, marcas balísticas y otros vestigios)—. La fiabilidad de los métodos para interpretar las pruebas es una consideración fundamental en toda ciencia. En consecuencia, cada especialidad científica dispone de un dominio específico de comprensión de lo que implica la validez científica de los métodos bien desarrollado.

El concepto de validez científica también juega un importante papel en el sistema jurídico. Particularmente, como se dijo en la sección 1, las *Federal Rules of Evidence* (FRE) (Reglas Federales de la Prueba) exigen que el testimonio pericial sobre ciencia forense sea resultado de «principios y métodos fiables» y de que estos hayan sido «fiablemente aplicados … a los hechos del caso».

Este informe explica los criterios científicos para la validez científica en el caso de métodos forenses de comparación de características, tanto para su uso dentro del sistema jurídico como en el ámbito de los que trabajan para fortalecer los fundamen-

[80] Centro de Ciencia Forense de la Defensa, oficina del jefe científico, informe anual dla sección de investigación, 5 de enero de 2016.

tos científicos de esas disciplinas. Antes de profundizar en esa explicación científica, proporcionamos en esta sección un resumen muy breve, dirigido principalmente a científicos y a lectores no especializados, sobre cuestiones jurídicas y términos relevantes en el ámbito estadounidense, así como sobre la naturaleza de esta intersección entre el derecho y la ciencia.

3.1. Evolución de los criterios de admisibilidad

A lo largo del siglo xx, el enfoque para decidir la admisibilidad de las pruebas científicas evolucionó en respuesta a los avances de la ciencia. En 1923, en el caso *Frye v. United States,*[81] el Tribunal de Apelación del Distrito de Columbia consideró la admisibilidad del testimonio concerniente a los resultados de un supuesto «detector de mentiras», un test basado en la presión sanguínea sistólica que fue precursor del polígrafo. Después de describir el dispositivo y su operación, el Tribunal rechazó el testimonio, estableciendo:

> Si bien los Tribunales recorren un largo camino para admitir el testimonio pericial deducido de un principio o descubrimiento científico bien reconocido, aquello de lo que parte la deducción debe estar lo suficientemente establecido para haber ganado la aceptación general en el área del conocimiento específica a la que pertenece[82].

El Tribunal estimó que el test sistólico «no había aún conseguido tal nivel de establecimiento y reconocimiento científico entre las autoridades fisiológicas y psicológicas» y, por consiguiente, fue declarado inadmisible.

Más de medio siglo después, las *Federal Rules of Evidence* (FRE) fueron promulgadas en 1975 para regular los litigios civiles y penales en los tribunales federales. La Regla 702, en su forma original, estableció que:

> Si el conocimiento científico, técnico o cualquier otro conocimiento especializado ayudará al Juez a comprender la prueba o a determinar el hecho en disputa, un testigo cualificado por su conocimiento, habilidad, experiencia, entrenamiento o educación puede testificar en forma de una opinión o de otro modo[83].

Hubo un considerable debate entre los abogados litigantes, los jueces y los académicos sobre si la regla adoptaba el criterio *Frye* o establecía un nuevo criterio[84]. En 1993, la Corte Suprema de los Estados Unidos trató de resolver estas cuestiones en su histórico fallo del caso *Daubert v. Merrell Dow Pharmaceuticals.* Interpretando la Regla 702, la Corte Suprema mantuvo que las FRE reemplazaban a *Frye* como criterios de admisibilidad de la prueba pericial en los tribunales federales. El Tribunal rechazó

---

[81] *Frye v. United States*, 293 F. 1013 (D.C. Cir. 1923).

[82] *Ibid.*, 1014.

[83] Act of January 2, 1975, Pub. Law No. 93-595, 88 Stat. 1926 (1975). Véase: federalevidence.com/pdf/FRE_Amendments/1975_Orig_Enact/1975-Pub.L._93-595_FRE.pdf

[84] Véanse: Giannelli, 1980; McCabe, 1996: 455-82; Page, Taylor y Blenkin, 2011: 913-7.

«la aceptación general» como el criterio de admisibilidad y en su lugar mantuvo que la admisibilidad del testimonio científico dependía de su fiabilidad científica.

Mientras que *Frye* establecía a los jueces deferir al juicio de la comunidad científica relevante, *Daubert* asignó a los jueces el papel de «guardianes» encargados de asegurar que el testimonio pericial «tuviera un fundamento fiable» [85].

El Tribunal dictaminó que para su admisión «el juez debe determinar … si el razonamiento o la metodología subyacente al testimonio es científicamente válida» [86]. Identificó cinco factores que un juez debe, entre otros, considerar ordinariamente a la hora de valorar la validez de la metodología subyacente. Estos factores son: (1) si la teoría o la técnica puede comprobarse (y ha sido comprobada); (2) si la teoría o la técnica ha estado sujeta a revisión por pares y publicación; (3) la tasa de error conocida o potencial de la técnica científica específica; (4) la existencia y mantenimiento de estándares que controlen la operación de la técnica; y (5) el grado de aceptación de la técnica en la comunidad científica relevante.

El Tribunal Supremo en el caso *Daubert* también señaló que los jueces para admitir los ofrecimientos de testimonio científico pericial debían considerar otras reglas aplicables, incluyendo la:

— Regla 403, que permite la exclusión de una prueba relevante «si su valor probatorio es sustancialmente superado por el peligro de un prejuicio injusto, de una confusión en los asuntos a tratar, o lleve a engaño al jurado» (señalando que la prueba pericial puede ser «tan contundente como engañosa a causa de la dificultad en valorarla»); y

— Regla 706, que permite al tribunal, a su propia discreción, solicitar la asistencia de un analista de su elección [87].

El Congreso modificó la Regla 702 en el año 2000 para hacerla más precisa, y realizó ulteriores cambios de estilo en el 2011. En su actual versión, la regla 702 impone cuatro requisitos:

> Un testigo cualificado como experto por su conocimiento, habilidad, experiencia, entrenamiento o educación, puede testificar en forma de opinión o de cualquier otro modo si:
> *(a)* el conocimiento científico, técnico, o cualquier otro especializado que tenga el experto ayuda al juez a comprender las pruebas o a determinar los hechos en cuestión;
> *(b)* el testimonio se fundamenta en suficientes hechos o datos;
> *(c)* el testimonio es resultado de principios y métodos fiables;
> *(d)* el analista ha aplicado fiablemente los principios y métodos a los hechos del caso.

El llamado «*Advisory Committee*»* de las FRE realizó una Nota a la Regla 702 especificando también una serie de factores de fiabilidad que completaba los cinco

---

[85] *Daubert*, en 587.

[86] *Daubert*, en 580. Véase también, Nota 9 («En un caso en el que haya una prueba científica, *la fiabilidad de la prueba* estará fundamentada en la *validez científica*». [énfasis en el original]).

[87] *Daubert*, en 595, citando a Weinstein, 138 F.R.D., en 632.

factores enumerados en Daubert. Entre esos factores se encuentra «si el área experta invocada por el perito es conocida por alcanzar resultados fiables»[88], [89].

Muchos estados han adoptado reglas de prueba que siguen en sus aspectos clave a las reglas federales. Esas reglas son actualmente el derecho en aproximadamente la mitad de los estados, mientras que otros continúan rigiéndose por el criterio *Frye* o variaciones de este[90].

3.2. Validez de los fundamentos y validez en su aplicación

Como se describió en *Daubert*, el sistema jurídico prevé un diálogo importante entre el derecho y la ciencia:

> El análisis judicial previsto en la Regla 702 es, lo enfatizamos, flexible. Su tema principal es la validez científica —y, por tanto, la relevancia y fiabilidad probatoria— de los principios que subyacen a la propuesta ofrecida[91].

Entonces, tanto consideraciones de carácter jurídico como científico juegan roles importantes:

---

[*] N. del T.: la Corte Suprema de los Estados Unidos estableció por primera vez un comité asesor de normas en junio de 1935 para ayudar a redactar las Reglas Federales de Procedimiento Civil, que entraron en vigor en 1938. Los Comités Asesores sobre el Reglamento de Apelación, Quiebra, Civil, Procedimiento Penal y las Reglas sobre la prueba realizan un estudio continuo de las normas que les competen y recomiendan cambios en la denominada *Judicial Conference,* a través de un Comité Permanente de Práctica y Procedimiento. Cfr. Committee Membership Selection | United States Courts (uscourts.gov)

[88] Véase: Nota del *Advisory Committee* a la FRE 702 (2000). Los siguientes factores pueden ser relevantes bajo la Regla 702: si la investigación subyacente se realizó independientemente del litigio; si el experto, a partir de una premisa aceptada, infirió injustificadamente una conclusión infundada; si el experto ha dado cuenta adecuadamente de explicaciones alternativas obvias; si el experto fue tan cuidadoso como lo hubiera sido en su trabajo profesional, fuera de su actividad pagada en un litigio; y *si el área de especialización a la que el experto considera pertenecer es conocida por alcanzar resultados fiables* [énfasis añadido].

[89] Esta nota ha sido señalada como un apoyo a los esfuerzos para desafiar campos enteros de la ciencia forense, incluidas las huellas dactilares y las comparaciones de cabellos. Véase: Gianelli, 2003: 1096.

[90] Incluso bajo *Frye*, lo que piensan los científicos sobre el significado de la fiabilidad es relevante. *Frye* exige que una técnica o método científico «tenga aceptación general» en la comunidad científica relevante para ser admisible. Como una cuestión científica, la comunidad científica relevante para asesorar sobre la fiabilidad de las ciencias de comparación de características incluye a metrólogos (incluidos estadísticos), así como otros científicos de la física o de la vida desde disciplinas sobre las que se fundamentan los métodos específicos. Es importante señalar que la comunidad no está constreñida a los científicos forenses que practican el método específico. Por ejemplo, el tribunal del caso *Frye* evaluó si el detector de mentiras ofrecido había conseguido un «reconocimiento científico estable entre las autoridades médicas y psicológicas» y no entre analistas en detectores de mentiras. *Frye v. United States*, 293 F. 1013 (D. C. Cir. 1923).

[91] *Daubert,* p. 594.

(1) La admisibilidad del testimonio pericial depende de un criterio gradual como, entre otras cosas, de si cumple ciertos requisitos *jurídicos* previstos en la Regla 702. Estas decisiones sobre admisibilidad son de competencia exclusiva de los tribunales.

(2) Ahora bien, como se advirtió más arriba, el tema general de análisis de los jueces bajo la regla 702 es «la validez científica.» Está en las competencias de la comunidad científica proporcionar guías sobre los estándares *científicos* para la validez científica.

El PCAST no opina aquí sobre los criterios jurídicos, sino que busca solo clarificar los estándares científicos subyacentes. Para lograr una completa claridad sobre lo que nos proponemos, hemos adoptado términos específicos para referirnos a los estándares científicos para dos tipos clave de validez científica, que queremos que se correspondan, como estándares científicos, con los estándares jurídicos previstos en la Regla 702 (c, d):

(1) por «validez de los fundamentos» queremos decir que el *estándar* científico se corresponde con el criterio jurídico de que la prueba se base en «principios y métodos fiables», y

(2) por «validez en la aplicación» queremos decir que el *estándar* científico se corresponde con el criterio jurídico de que el experto o analista «aplique fiablemente los principios y métodos».

En el siguiente apartado discurriremos los estándares científicos de estos conceptos. Cerramos esta sección advirtiendo que la respuesta a la cuestión de la validez científica de las disciplinas forenses es importante no solo para los tribunales sino también porque establece estándares de calidad que se extienden a todas ellas, afectando a la práctica y definiendo la investigación necesaria.

4. Criterios científicos para la validez y fiabilidad de los métodos forenses de comparación de características

— En este informe, el PCAST ha elegido centrarse en la validez y fiabilidad de un área específica dentro de la ciencia forense: los métodos forenses de comparación de características. Hemos hecho eso porque es posible e importante hacerlo para esta específica clase de métodos.

— Es *posible* porque la comparación de características es una actividad científica común y la ciencia posee estándares claros para determinar si tales métodos son fiables. En particular, los métodos de comparación de características pertenecen a la disciplina de la metrología —la ciencia de la medida y su aplicación—[92, 93].

---

[92] International Vocabulary of Metrology – Basic and General Concepts and Associated Terms, 2012.

[93] Que los métodos de comparación de características pertenecen a la disciplina de la metrología es claro partiendo del hecho de que el NIST —cuya misión es asistir a la Nación mediante «ciencia de me-

— Es *importante* porque en la pasada década se ha evidenciado que una deficiente comparación de características forense ha conducido a numerosos errores judiciales[94]. También se ha revelado que los problemas no son debidos simplemente a una pobre ejecución de unos pocos analistas sino más bien al hecho de que la fiabilidad de muchos métodos de comparación de características forenses nunca ha sido significativamente evaluada[95].

Comparado con muchos tipos de testimonio pericial, el basado en métodos de comparación de características posee exclusivos peligros de llevar al error a los miembros de un jurado por dos razones:

— La gran mayoría de los miembros de un jurado no tienen una habilidad independiente para interpretar el valor probatorio de los resultados basados en la detección, comparación y frecuencia de la evidencia científica. Si dos mitades de una

---

dición avanzada, estándares y tecnología», y que es el laboratorio metrológico líder mundial— es la sede del Gobierno Federal que se ocupa de los esfuerzos de investigación en ciencia forense. Los programas del NIST incluyen investigación interna, financiación externa de la investigación externa, conferencias y preparación de los materiales y estándares de referencia. Véanse: www.nist.gov/public_affairs/mission.cfm y www.nist.gov/forensics/index.cfm. Los métodos de comparación de características conllevan determinar si dos conjuntos de características concuerdan dentro de una cierta tolerancia en la medida.

[94] La reexaminación mediante ADN de casos cerrados de informes ha conducido a la exoneración de 342 condenados hasta la fecha, incluyendo 20 que habían sido sentenciados a muerte, y a la identificación de 147 verdaderos culpables. Véase: Innocence Project, «DNA Exonerations in the United States», www.innocenceproject.org/dna-exonerations-in-the-united-states. La revisión de estos casos ha revelado que, aproximadamente la mitad, se deben en parte al testimonio de analistas, basados en métodos que no habían sido sometidos a un escrutinio científico significativo o que incluían conclusiones científicas afirmadas con niveles de seguridad inválidos. Véanse: Gross y Shaffer, 2012. Disponible en: www.law.umich.edu/special/exoneration/Documents/exonerations_us_1989_2012_full_report.pdf; Garrett y Neufeld, 2009: 1-97; National Research Council, 2009, 42-43. La naturaleza de los asuntos se ilustra en los documentos citados mediante ejemplos específicos: Levon Brooks y Kennedy Brewer, cada uno condenado por asesinato de dos niños, separadamente, en los años 90, casi enteramente sobre la base de un testimonio experto de análisis de marcas de mordedura, que conllevó que pasaran más de 13 años en prisión antes de que el análisis de ADN identificara al auténtico cupable, que confesó ambos crímenes; Santae Tribble, condenado por asesinato después de que el analista del FBI testificara que los cabellos encontrados en una máscara de media implicaba a Tribble en el crimen al coincidir (con su cabello) «en todas las características microscópicas», pasó más de 20 años en prisión antes de que el análisis de ADN revelara que ninguno de los 13 cabellos examinados pertenecían a Tribble y que uno procedía de un perro; Jimmy Ray Bromgard de Montana pasó 15 años de prisión por violación antes de que el análisis de ADN mostrase que los cabellos recogidos de la cama de la víctima y considerados coincidentes con los de Bromgard no podían ser suyos; Stephan Cowans, condenado por asesinato por disparar a un oficial de policía de Boston después de que dos analistas en huellas dactilares testificaran que la impresión del pulgar dejada por el autor del crimen era «única e idéntica» (a la de Bromgard), pasó más de 5 años en prision antes de que el análisis de ADN sobre múltiples muestras le excluyera como autor del crimen; y Steven Barnes, del norte del estado de New York, pasó 20 años en prisión por un delito de violación y asesinato que no cometió después de que un analista en criminalística testificara que una superposición fotográfica entre la tela de los jeans de la víctima y una impresión del camión de Barnes mostraba patrones que eran «similares» y que los cabellos recogidos del camión eran similares a los cabellos de la víctima.

[95] Véase: Sección 5.

nota de rescate que fueran coincidentes se encontrasen en la escena del crimen y en el domicilio del acusado, los miembros del jurado podrían confiar en su propia experiencia para valorar la probabilidad de que dos trozos rotos coincidan si no procediesen de un solo original. Si un testigo describiera a un criminal como «alto y de cabello tupido», los miembros del jurado podrían realizar un juicio razonable sobre cuánta gente se ajusta a esa descripción. Pero, si un testigo experto dijera que en dos muestras de ADN el tercer exón del gen *DYNC1H1* tiene precisamente 174 nucleótidos de longitud, la mayoría de los miembros del jurado no tendrían manera de saber si deberían estar impresionados por la coincidencia; dependerían completamente de las declaraciones revestidas con el manto de la ciencia hechas por los analistas (No deberían sentirse impresionados por la aseveración precedente: en el marcador de la cadena de ADN citada, más del 99.9 por ciento de la gente tiene un fragmento del tamaño indicado)[96].

— El potencial impacto perjudicial es inusualmente alto porque los miembros del jurado suelen sobreestimar el valor probatorio de un cotejo positivo entre muestras. De hecho, el mismo DOJ históricamente sobreestimó el valor probatorio de cotejos positivos en su prolongada posición, reconocida ahora como inapropiada, sobre que los cotejos de huellas latentes eran «infalibles»[97]. De forma similar, un exjefe de la unidad de huellas dactilares del FBI testificó que el FBI tenía «una tasa de error de 1 por cada 11 millones de casos»[98]. En un experimento online, los investigadores preguntaron a miembros del jurado ficticios que estimaran la frecuencia de que un científico forense experimentado y cualificado pudiera erróneamente concluir que dos muestras de tipos específicos provinieran de la misma persona cuando realmente lo eran de personas distintas. Los miembros del jurado ficticios creyeron que tales errores era probable que ocurrieran en 1 de cada 5.5 millones de comparaciones de huellas dactilares; 1 de cada millón de comparaciones de marcas de mordedura; y 1 de cada 100.000 comparaciones de escritura manuscrita[99]. Mientras que se desconocen las tasas de error de la mayoría de esas técnicas, todos los indicios apuntan a que las tasas de error son de mayores órdenes de magnitud. Por ejemplo, los propios estudios del FBI sobre análisis de huellas latentes indican tasas de error en el rango de 1 en varios cientos de cotejos[100] (Como consecuencia de que el término «match» o coincidencia es probable que implique un valor probatorio inapropiadamente alto, debiera utilizarse un término más neutral para expresar la creencia del analista de que las dos muestras proceden de la misma fuente. Sugerimos el término «*identificación* propuesta» para transmitir adecuadamente la conclusión del analista y, a la vez, la posibilidad de que quizá sea errónea. Usaremos este término a lo largo de este informe).

---

[96] Véase: Base de datos ExAC: http://exac.broadinstitute.org/gene/ENSG0000019710.

[97] Véase: www.justice.gov/olp/file/861906/download.

[98] *U.S. v. Baines* 573 F.3d 979 (2009) en 984.

[99] Koehler, 2016. Disponible en papers.ssrn.com/sol3/papers.cfm?abstract_id=2817443

[100] Véase: Sección 5.4.

Este apartado recoge las conclusiones del PCAST sobre los criterios científicos de validez científica. Las conclusiones se basan en el principio fundamental del «método científico» —aplicable a lo largo de toda la ciencia— respecto a que el conocimiento científico válido *sólo* puede obtenerse a través de la comprobación *empírica* de las proposiciones específicas[101]. Las conclusiones del PCAST de esta sección pueden resumirse brevemente como sigue:

> La validez y fiabilidad científicas requieren que un método haya sido sometido a comprobación empírica, bajo condiciones apropiadas a su pretendido uso, que proporcione estimaciones válidas de la frecuencia con que el método llega a conclusiones incorrectas. Para los métodos de comparación subjetiva de características se requieren estudios de caja negra apropiadamente diseñados, en los que muchos analistas tomen decisiones sobre muchas pruebas independientes (típicamente incluyendo muestras «cuestionadas» y una o más muestras «conocidas») y en los que las tasas de error se determinen. Sin las apropiadas estimaciones de acierto en las comparaciones, una proposición del analista que diga que dos muestras comparadas son similares —o, incluso, indistinguibles— carece de significado científico: carece de valor probatorio y tiene potencial considerable de causar un impacto perjudicial. Nada —ni personal formado y con experiencia, ni prácticas profesionales— pueden sustituir a una demostración empírica adecuada sobre el nivel de acierto.

La sección se organiza como sigue:

— La primera sección describe la distinción entre dos tipos fundamentalmente diferentes de métodos de comparación de características: objetivos y subjetivos.

— Las cinco secciones siguientes tratan sobre los criterios científicos para los dos tipos de validez científica: la validez de los fundamentos y la validez en la aplicación.

— Las dos secciones finales reflejan puntos de vista mantenidos en la comunidad forense.

4.1. Métodos de comparación de características: métodos objetivos y subjetivos

Un método de comparación de características es un procedimiento mediante el cual un analista busca determinar si un vestigio probatorio (por ejemplo, de la escena del crimen) está o no asociado con una muestra de origen conocido (por ejemplo, del sospechoso)[102] basándose en características similares.

Los métodos de comparación de características pueden clasificarse en objetivos o subjetivos. Por métodos de comparación de características objetivos entendemos

---

[101] Por ejemplo, el diccionario online de Oxford define *scientific method* como «un método o procedimiento que ha caracterizado a las ciencias naturales desde el siglo XVII, consistente en la observación, medida y experimentación sistemáticas, y en la formulación, comprobación y modificación de hipótesis». «Scientific Method» *Oxford Dictionaries Online*. Oxford University Press (última consulta el 19 de agosto de 2016).

[102] Una «muestra de origen conocido» se refiere a un individuo u objeto específico (por ejemplo, un neumático o una pistola).

aquellos consistentes en procedimientos definidos con un detalle suficientemente estandarizado y cuantificable que pueden ejecutarse bien mediante un sistema automático, bien mediante analistas que emiten escasos o ningún juicio. Por métodos subjetivos entendemos métodos que incluyen procedimientos clave que conllevan juicios humanos significativos —por ejemplo, sobre qué características han de seleccionarse o cómo determinamos si las características son suficientemente similares para considerarlas para proponer una identificación—.

Los métodos objetivos son, en general, preferibles a los subjetivos. Los análisis que dependen de juicios humanos (en lugar de una medida cuantitativa de similitud) son obviamente más susceptibles al error humano, los sesgos y la variabilidad de ejecución entre los analistas [103]. Por contra, los métodos objetivos y cuantificados tienden a producir una mayor precisión, repetibilidad y fiabilidad, incluyendo una reducción en la variación de los resultados entre los analistas. Los métodos subjetivos pueden evolucionar hacia métodos objetivos o ser reemplazados por ellos [104].

4.2. Validez de los fundamentos: requerimiento de estudios empíricos

Para que un método metrológico sea científicamente válido y fiable, debe mostrar se el procedimiento en el que consiste, fundamentado en estudios empíricos, ser *repetible*, *reproducible* y *preciso*, en los niveles que ha sido medido, y ser apropiado para la aplicación requerida [105], [106].

El método no necesita ser perfecto, pero es claramente *esencial* que su precisión se haya medido mediante comprobaciones empíricas apropiadas y sea suficientemente alta para ser adecuada a la aplicación. Sin una apropiada estimación de su precisión, un método metrológico es inútil —porque no hay forma de interpretar sus resultados—. La importancia de conocer la precisión del método fue resaltada en el informe del NRC (2009) sobre la ciencia forense y por un informe NRC del año 2010 sobre tecnologías biométricas [107].

---

[103] Dror, 2016: 121-127.

[104] Por ejemplo, antes del desarrollo de la prueba de alcoholemia, los tribunales tenían que confiar exclusivamente en el testimonio de los oficiales de policía y de otros que, a su vez, se basaban en indicios conductuales de embriaguez y en la presencia de alcohol en la respiración. El desarrollo de tests químicos objetivos condujo a un cambio de estándares subjetivos por objetivos.

[105] National Physical Laboratory, 2010. Disponible en: www.npl.co.uk/upload/pdf/NPL-Beginners-Guide-to-Measurement.pdf; Pavese, 2009.

[106] Los métodos de comparación de características que dan resultados erróneos con demasiada frecuencia tienen, por definición, bajo valor probatorio. Como se dijo más arriba, el impacto perjudicial es probable que de este modo supere al valor probatorio.

[107] «El reconocimiento biométrico es una tarea inherentemente probabilística … En consecuencia, incluso aunque la tecnología y el sistema incrustado en ella estén funcionando como se diseñaron, hay inevitablemente incertidumbre y existe riesgo de error», National Research Council, 2010: viii-ix.

CUADRO 2.
DEFINICIÓN DE TÉRMINOS CLAVE

Por «repetible», entendemos que, con probabilidad conocida, un perito obtiene el mismo resultado cuando analiza muestras procedentes de las mismas fuentes.

Por «reproducible» entendemos que, con probabilidad conocida, diferentes peritos obtienen el mismo resultado cuando analizan las mismas muestras.

Por «preciso», entendemos que, con probabilidades conocidas, un perito obtiene resultados correctos tanto para (1) muestras procedentes de la misma fuente (verdaderos positivos) como para (2) muestras procedentes de distintas fuentes (verdaderos negativos).

Por «fiabilidad», entendemos repetibilidad, reproducibilidad y precisión [108].

Por «científicamente válido», entendemos que un método ha demostrado, basado en estudios empíricos, ser fiable con niveles de repetibilidad, reproducibilidad y precisión apropiados a la aplicación pretendida.

Por un «estudio empírico» entendemos una prueba en la que se usa un método para analizar un gran número de conjuntos independientes de muestras, similares en aspectos relevantes a las encontradas en los casos en los que hemos de trabajar, para estimar la repetibilidad, reproducibilidad y precisión del método.

Por «estudio de caja negra», entendemos un estudio empírico que valora un método subjetivo mediante examinadores que analizan muestras y opinan sobre el origen o las similitudes entre las muestras.

Para cumplir los criterios científicos de validez en los fundamentos se requieren dos elementos:

(1) un procedimiento reproducible y consistente para (a) identificar características en las muestras que son aportadas como pruebas; (b) comparar las características de dos muestras; y (c) determinar, basándose en la similitud entre las características de dos muestras, si las muestras deben considerarse aptas para una propuesta de identificación («regla de coincidencia»).

(2) medidas empíricas, resultantes de múltiples estudios independientes, de (a) la tasa de falsos positivos del método —es decir, la probabilidad de que se declare una propuesta de identificación entre muestras que realmente proceden de *diferentes* fuentes— y (b) la sensibilidad del método —es decir, la probabilidad de que declare una propuesta de identificación entre muestras que realmente proceden de la *misma* fuente—.

Trataremos sobre estas cuestiones en su momento.

---

[108] Queremos resaltar que «fiabilidad» tiene también un significado próximo a «consistencia», usado en estadística» —es decir, hasta qué punto un método produce el mismo resultado, con independencia de si el resultado es preciso—. Ese no es el sentido en que «fiabilidad» es utilizado en este informe, o en la ley.

4.2.1. Procedimientos reproducibles y consistentes

Para que un método sea objetivo, deben definirse *cada uno* de los tres pasos (identificación de características, comparación de características y reglas de coincidencia) de forma precisa, reproducible y consistente. Los analistas forenses deben identificar las características relevantes de la misma manera y alcanzar el mismo resultado. Deben comparar las características de la misma forma cuantitativa. Para plantear una propuesta de identificación deben calcular si las características de una muestra ofrecida como prueba y las de una muestra procedente de un sospechoso se encuentran dentro de una pre-especificada tolerancia de medición (regla de coincidencia)[109]. De un método objetivo podemos establecer la validez del fundamento de cada uno de los pasos individuales midiendo su precisión, reproducibilidad y consistencia.

Para los métodos subjetivos, los procedimientos han de ser cuidadosamente definidos —pues conllevan un sustancial juicio humano—. Por ejemplo, diferentes analistas pueden reconocer o centrarse en diferentes características, pueden otorgar diferente importancia a las mismas características y pueden tener diferentes criterios a la hora de hacer propuestas de identificación. Como consecuencia de que los procedimientos de identificación de características, las reglas de coincidencia y las determinaciones frecuenciales de las características no están objetivamente especificados, el procedimiento en su conjunto ha de tratarse como una especie de «caja negra» dentro de la cabeza del analista.

Los métodos subjetivos requieren cuidadoso escrutinio, normalmente más, pues su fuerte dependencia de los juicios humanos lleva consigo que sean especialmente vulnerables al error humano, a la inconsistencia entre analistas y a los sesgos cognitivos. En las disciplinas forenses de comparación de características, los sesgos cognitivos incluyen el fenómeno de que, en ciertos entornos, los seres humanos (1) pueden tender de modo natural a centrarse sobre las similitudes entre las muestras y a descartar las diferencias y (2) pueden también ser influenciados por información no pertinente y presiones externas sobre el caso[110] (Estos últimos temas se ilustran con la equivocada identificación de una huella latente por el FBI relacionada con los atentados con bombas en los trenes de Madrid, que se aborda en la página 34).

Puesto que la caja negra en la cabeza del analista no puede examinarse directamente con respecto a su fundamentación en la ciencia, la validez de los fundamentos

---

[109] Si se declara que una fuente *no* comparte las mismas características, el test la excluye. La regla de coincidencia debe elegirse cuidadosamente. Si se elige una «regla de coincidencia» demasiado estricta, muestras que realmente proceden de la misma fuente se declararán como no coincidentes (falsos negativos). Si es demasiado laxa, entonces el método no tendrá un alto poder discriminatorio porque la probabilidad de coincidencia aleatoria será demasiado alta (falsos positivos).

[110] Véanse, por ejemplo: Boroditsky, 2007: 118- 128; Hassin, 2001: 728–731; Medin, Goldstone y Gentner, 1993: 254–278; Tversky, 1977: 327–352; Kim, Novemsky y Dhar, 2012: 225–229; Larkey y Markman, 2005: 1061–1076; Medin, Goldstone y Markman, 1995: 1–19; Goldstone, 1994: 125–157; Nosofsky, 1986: 39–57.

de los métodos subjetivos *solo* puede establecerse a través de estudios empíricos sobre las acciones del analista para determinar si pueden proporcionar respuestas precisas; a tales estudios se les llama «estudios de caja negra» (Cuadro 2). En los estudios de caja negra, a muchos analistas se les presentan muchos problemas de comparación independientes —normalmente, incluyendo muestras «cuestionadas» y una o más muestras «conocidas»— y se les pregunta que declaren si las muestras cuestionadas proceden de la misma fuente que la de las muestras conocidas[111]. Los investigadores, posteriormente, determinan cuántas veces los analistas llegan a conclusiones erróneas.

Como ejemplo sobresaliente, el FBI condujo recientemente un estudio de caja negra en análisis de huellas latentes, implicando a 169 analistas que examinaron 744 pares de huellas dactilares, y publicó los resultados del estudio en una revista científica líder[112]. (Algunos científicos forenses han advertido que prestar demasiada atención a aspectos subjetivos de los métodos forenses —tales como estudios de sesgos cognitivos y cajas negras— puede distraer del fin de mejorar el conocimiento sobre las características objetivas de la evidencia forense y el desarrollo de métodos verdaderamente objetivos[113]. Otros han resaltado que eso no es actualmente un problema, porque los esfuerzos y los medios económicos para afrontar los retos asociados a los métodos forenses subjetivos son muy limitados[114]).

4.2.2. Medidas empíricas de precisión

Es necesario que tengamos medidas empíricas apropiadas de la tasa de falsos positivos y de la sensibilidad del método. Como se explicó en el Apéndice A, resulta necesario conocer esas dos medidas para estimar el valor probatorio de un método. La tasa de falsos positivos es la probabilidad de que el método declare una propuesta de identificación entre muestras que realmente procedan de fuentes *diferentes*. Por ejemplo, una tasa de falsos positivos del 5% significa que dos muestras de fuentes *diferentes* serán (debido a las limitaciones del método) incorrectamente declaradas como procedentes de la misma fuente un 5% de las veces. (La cantidad igual a 1 menos la tasa de falsos positivos —95% en el ejemplo— es referida como la especificidad).

La sensibilidad del método es la probabilidad de que el método declare una propuesta de identificación entre muestras que realmente procedan de una *misma* fuente. Por ejemplo, una sensibilidad del 90% significa que dos muestras procedentes de una misma fuente se declararán como tales el 90% de las veces y declaradas como procedentes de fuentes diferentes el 10% de las veces. (La última cantidad se conoce como tasa de falsos negativos).

---

[111] Las respuestas pueden expresarse en términos de «cotejo positivo / cotejo negativo / inconclusión» o «identificación / exclusión / inconclusión».

[112] Ulery, Hicklin, Buscaglia y Roberts, 2011: 7733-7738.

[113] Champod, 2014: 107–109.

[114] Risinger, Thompson, Jamieson, et al., 2014: 508-509.

La tasa de falsos positivos es especialmente importante porque los resultados de falsos positivos pueden conducir directamente a condenas erróneas[115]. En algunas circunstancias es posible estimar una tasa de falsos positivos relacionada con características específicas de las pruebas del caso (Por ejemplo, la probabilidad de coincidencia aleatoria calculada en los análisis de ADN depende en parte del genotipo específico observado en una muestra de la prueba. La tasa de falsos positivos de los análisis de huellas latentes depende de la calidad de la huella latente). Para otros métodos de comparación de características, solo es posible tener una estimación media de la tasa de falsos positivos en las muestras.

Para métodos objetivos, la tasa de falsos positivos se compone de dos fuentes distinguibles —cotejos positivos por azar (en los que muestras de diferentes fuentes tienen, sin embargo, *características* que caen dentro de las tolerancias de las reglas de coincidencia objetivas)— y fallos humano-técnicos (en los que las muestras presentan características fuera de las reglas de coincidencia, pero en los que existió, sin embargo, una declaración de propuesta de identificación debido a un fallo humano o técnico).

Para métodos objetivos en los que la probabilidad de coincidencia aleatoria sea muy baja (como sucede en los análisis de ADN), la tasa de falsos positivos aplicada a un caso real estará dominada por la tasa de fallos humano-técnicos —que puede ser cientos de veces mayor—. En cuanto a los métodos subjetivos, los tipos de errores descritos —coincidencia aleatoria y fallos humano-técnicos— ocurren igualmente, pero sin una «regla de coincidencia» objetiva, no podrán distinguirse esas dos fuentes de error. Para el establecimiento de la validez de los fundamentos es esencial realizar estudios de caja negra que midan empíricamente la tasa de error global de muchos analistas. (Véase el Cuadro 3 sobre el término «error»).

CUADRO 3.
LOS SIGNIFICADOS DE «ERROR»

El término «error» tiene diferentes significados en las ciencias y en el derecho, que pueden conducir a confusión. En sentido jurídico, el término »error» a menudo implica una falta —por ejemplo, que una persona haya cometido un error cuando pudo haber sido evitado si él o ella hubiera seguido un procedimiento adecuado o que una máquina haya dado un resultado erróneo que pudo haber sido evitado si hubiera sido apropiadamente calibrada—. En ciencias, el término «error» también incluye la situación en la que el propio procedimiento, siendo propiamente aplicado, no da la respuesta correcta debido a una circunstancia imprevista.

Cuando aplicamos un método forense de comparación de características con el fin de valorar si dos muestras procedieron o no de la misma fuente, los cotejos positivos por

[115] Véase nota 94. Bajo algunas circunstancias, los resultados de falsos negativos pueden contribuir a también a condenas erróneas.

azar y los fallos humano-técnicos se consideran, desde el punto de vista estadístico, como «errores» porque pueden inducir a conclusiones incorrectas.

Los estudios que se diseñan para estimar la tasa de falsos positivos y la sensibilidad de un método son necesariamente realizados utilizando solo un limitado número de muestras. Como consecuencia, no pueden proporcionar valores «exactos» para esas cantidades (y no deben sostener que lo hacen), sino solo «intervalos de confianza», cuyos límites reflejan, respectivamente, el rango de valores que es razonablemente compatible con los resultados. Cuando se informa de un falso positivo a un jurado, es científicamente importante establecer que por encima del límite de confianza superior unilateral del 95% se refleja el hecho de que la tasa de falsos positivos real pudiera ser razonablemente tan alta como ese valor. (Para más información, véase el Apéndice) [116].

Con frecuencia los estudios categorizan los resultados como concluyentes (por ejemplo, identificación o exclusión) o no concluyentes (sin una determinación realizada) [117]. Cuando se informa de la tasa de falsos positivos a un jurado, es científicamente importante calcular la tasa fundamentándose en la proporción en los dictámenes concluyentes, en lugar de en la proporción en todos los dictámenes. Esto es apropiado porque la evidencia utilizada contra un acusado estará basada generalmente en dictámenes concluyentes en lugar de dictámenes inconclusos. Para ilustrar este punto, consideremos un caso extremo en el que un método se ha testeado 1000 veces y se han obtenido 990 resultados inconclusos, 10 falsos positivos y ningún resultado correcto. Sería engañoso decir que la tasa de falsos positivos fue del 1% (10/1000 dictámenes). En su lugar, deberíamos informar que en el 100% de los dictámenes concluyentes se obtuvieron falsos positivos (10/10 dictámenes).

Mientras que los estudios científicos exploratorios pueden ser de muchas formas, los estudios de validación científica —dirigidos a valorar la validez y la fiabilidad de un método metrológico para una aplicación de comparación de características forense particular— deben satisfacer una serie de criterios, los cuales se describen en el Cuadro 4.

---

[116] El límite superior de confianza incorpora propiamente la precisión de la estimación basada en el tamaño muestral. Por ejemplo, si un estudio no encontró errores en 100 pruebas, sería engañoso informar al jurado que la tasa de error es cero. De hecho, si las pruebas son independientes, el límite superior de confianza del 95% de la verdadera tasa de error es el 3%. De acuerdo con eso, al jurado se le debe decir que la tasa de error pudiera ser tan alta como el 3% (es decir, 1 por cada 33). La verdadera tasa pudiera ser más alta, pero con bastante poca probabilidad (menor al 5%). Si el estudio fuera más pequeño, el límite superior de confianza del 95% será más alto. Para un estudio sin errores en 10 pruebas, el límite superior de confianza del 95% es el 26% —es decir, la tasa de falsos positivos real pudiera ser aproximadamente de 1 por cada 4— (véase Apéndice A).

[117] Véase: Sección 5.

CUADRO 4.
CRITERIOS CLAVE PARA ESTUDIOS DE VALIDACIÓN QUE ESTABLEZCAN LA VALIDEZ DE LOS FUNDAMENTOS

Los estudios de validación científica —dirigidos a valorar la validez y la fiabilidad de un método metrológico para una particular aplicación de comparación de características forense— han de satisfacer una serie de criterios.

(1) Los estudios deben tener en cuenta un número suficientemente grande de examinadores y han de basarse en conjuntos de muestras *conocidas* y *representativas* suficientemente *grandes* de poblaciones *relevantes* que reflejen el rango de características o combinaciones de características que ocurran en la aplicación. En particular, los conjuntos de muestras deben ser:

(a) representativos de la calidad de las muestras de pruebas vistas en casos reales. (Por ejemplo, si un método va a ser utilizado sobre huellas dactilares latentes parciales y distorsionadas, debemos determinar la *probabilidad de coincidencia aleatoria* —es decir, la probabilidad de que la coincidencia ocurra por casualidad— para huellas dactilares latentes parciales y distorsionadas; la probabilidad de coincidencia aleatoria para huellas dactilares completas escaneadas, o incluso de huellas latentes de muy alta calidad, no sería relevante.

(b) elegidos de poblaciones relevantes a casos reales. Por ejemplo, para características de muestras biológicas, la tasa de falsos positivos ha de determinarse sobre la totalidad de la población estadounidense y para los principales grupos étnicos, como se hace con los análisis de ADN.

(c) suficientemente grande para proporcionar estimaciones apropiadas de las tasas de error.

(2) Los estudios empíricos deben conducirse de forma que ni el examinador ni aquellos con los que el examinador interactúa tengan información alguna sobre la respuesta correcta.

(3) El diseño del estudio y el marco de referencia del análisis se deben especificar de antemano. En los estudios de validación es inapropiado modificar el protocolo a posteriori basándose en los resultados[118].

(4) Los estudios empíricos deben ser dirigidos o supervisados por individuos u organizaciones que no tengan beneficio en la entrega de los estudios[119].

[118] La situación análoga en medicina es un test clínico para asegurarnos de la seguridad y eficacia de un fármaco para una aplicación particular. En el diseño de los tests clínicos, la FDA requiere que los criterios para los análisis sean previamente especificados y las notas que se añaden a posteriori a los análisis comprometen la validez del estudio. Véase: FDA Guidance, 2016. Disponible en: www.fda.gov/downloads/medicaldevices/deviceregulationandguidance/guidancedocuments/ucm446729.pdf; Alosh, Fritsch, Huque, et al., 2015: 286-303; FDA Guidance, 2013. Disponible en: www.fda.gov/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/ucm373750.htm); FDA Guidance for Industry, 1998. Disponible en: www.fda.gov/downloads/drugs/guidancecomplianceregulatoryinformation/guidances/ucm073137.pdf); Pocock, 1983.

[119] Cuando se inician ensayos clínicos, quien patrocina la prueba (una farmacéutica, una empresa que fabrica dispositivos o una empresa biotecnológica o, en algunos casos, una institución académica)

(5) Los datos, el software y los resultados de los estudios de validación deben estar disponibles para permitir que otros científicos revisen las conclusiones.

(6) Para asegurarse que las conclusiones son reproducibles y robustas, debe haber múltiples estudios de grupos distintos que lleguen a conclusiones similares.

Una medida empírica de las tasas de error no es solo una característica deseable, es esencial para determinar si un método es válido en sus fundamentos. En la ciencia, un procedimiento de prueba —como el test de embarazo o la comprobación de si el agua pudiera estar contaminada— no se considera válido hasta que no se haya medido su fiabilidad. Por ejemplo, necesitamos conocer la frecuencia con la que el test de embarazo indica que hay un embarazo cuando no es cierto y viceversa. Los mismos principios científicos se aplican, en no menor grado, a las pruebas forenses, que pueden contribuir a que un acusado pierda su vida o su libertad.

Es particularmente importante no inferir las tasas de error desde la casuística forense, sino que debe determinarse basándose en muestras de las que se conozca la respuesta correcta. Por ejemplo, el exjefe de la Unidad de Huellas Dactilares del FBI testificó que el FBI tenía «una tasa de error de 1 cada 11 millones de casos» basándose en el hecho de que la agencia era conocida por haber cometido solo un error en los últimos 11 años, en los que habían realizado 11 millones de identificaciones»[120]. La falacia es obvia: el analista simplemente *asumió sin pruebas* que habían salido a la luz todos los errores en casos reales.

¿Por qué es esencial conocer la tasa de falsos positivos y la sensibilidad de un método? Porque sin la apropiada medida empírica de la precisión de un método, el hecho de que dos muestras, en un caso particular, muestren similares características, *no tiene valor probatorio* —y, como se dijo anteriormente, puede tener un impacto perjudicial considerable, porque los miembros del jurado adjudicarán un significado equivocado a la observación—[121].

Desde una perspectiva científica, la absoluta necesidad de datos empíricos es elegantemente expresada en una analogía del juez de distrito norteamericano John Potteren en su decisión del caso *U.S. v. Yee (1991)*, un caso temprano sobre el uso del análisis de ADN:

---

financia e inicia el estudio, pero la prueba es dirigida por individuos independientes del patrocinador (frecuentemente, médicos académicos), para asegurar la fiabilidad de los datos generados en el estudio y minimizar la posibilidad de que haya sesgos. Véanse, por ejemplo, 21 C.F.R. § 312.3 y 21 C.F.R. § 54.4(a).

[120] *U.S. v. Baines* 573 F.3d 979 (2009) en 984.

[121] Bajo la regla federal de la prueba (Fed. R. Evid.) nº 403, la prueba debe ser excluida si «su valor probatorio es sustancialmente sobrevalorado por el temor de un perjuicio injusto».

> Sin la valoración de la probabilidad, el jurado no sabe qué hacer ante el hecho de que dos patrones comparados coincidan: el jurado no sabe si los patrones son tan comunes como las imágenes con dos ojos o tan únicos como la Mona Lisa[122, 123].

4.3. Validez de los fundamentos: requerimiento para el testimonio científicamente válido

Debe considerarse obvio —pero merece que se enfatice— que una vez que un método ha sido establecido como válido en sus fundamentos por estar basado en estudios empíricos apropiados, la forma de asegurar que las afirmaciones sobre la precisión de un método y el valor probatorio de las identificaciones propuestas sean válidas es que estén basadas en tales estudios empíricos. *Los enunciados que aseveren o impliquen mayor certeza que la demostrada mediante la evidencia empírica, son científicamente inválidos*. Los analistas forenses deben, por tanto, informar de sus hallazgos sobre una propuesta de identificación con claridad y mesura, explicando en cada caso que el hecho de que dos muestras satisfagan los criterios de un método para una propuesta de identificación no implica necesariamente que las muestra procedan de la misma fuente. Si la tasa de falsos positivos de un método se cifra en 1 cada 50, los analistas no pueden deducir que el método sea capaz de producir resultados con un mayor grado de precisión.

Es preocupante que los peritos en ocasiones vayan más allá de la evidencia empírica sobre la frecuencia de las características —incluso hasta el extremo de sostener o deducir que una muestra procedió de una específica fuente con casi certeza absoluta o con ella, a pesar de carecer de base para tales opiniones—[124]. Desde el punto de vista de la validez científica, a los analistas no se les debe permitir establecer ante los tribunales que puedan emitir conclusiones con certeza o con cercanía a la certeza (tales como tasas de error igual a «cero», «imperceptiblemente pequeña», «esencialmente cero», «despreciable», «mínima» o «microscópica»; «certeza del 100%» o «un razonable grado de certeza científica»; o una identificación «hasta la exclusión de todas las fuentes»[125].

---

[122] *U.S. v. Yee*, 134 F.R.D. 161 (N.D. Ohio 1991).

[123] Algunos tribunales han dictaminado que no hay peligro en admitir pruebas basadas en comparación de características sobre la base de que los jurados pueden ver las características con sus propios ojos y decidir por sí mismos si comparten características. *U.S. v. Yee* muestra por qué ese razonamiento es falaz: los jurados no tienen forma de conocer con qué frecuencia dos muestras diferentes comparten sus características, y con qué nivel de especificidad.

[124] Como se dijo con anterioridad, la larga historia de afirmaciones exageradas sobre la precisión de los métodos forenses incluye el propio enunciado inicial de DOJ de que los análisis de huellas latentes son «infalibles», que DOJ ha juzgado después como inapropiado. www.justice.gov/olp/file/861906/download.

[125] Cole, 2004. Véase también: National Research Council, 2009: 87, 104, y 143.

Que esos testimonios sean inapropiados es certeramente advertido en una analogía que la juez del Tribunal de Apelación del Distrito de Columbia Catharine Easterly emite en su opinión concurrente en el caso *Williams v. United States*, un caso en el que un analista testificó que las marcas sobre ciertos proyectiles eran únicas y procedentes de una pistola recogida en el apartamento del acusado:

> Tal y como están las cosas actualmente, una declaración de certeza con respecto a la coincidencia de patrones de marcas de herramientas tiene el mismo valor probatorio que la visión de un vidente: no refleja nada más que la fe individual sin fundamento en lo que cree que es verdad. Esa no es la clase de prueba en la que podamos confiar en buena conciencia, particularmente en casos penales, donde exigimos que esté probado—prueba real—más allá de la duda razonable, precisamente porque lo que se juega vale mucho[126].

En ciencias, la afirmación de que un método metrológico es más preciso que lo que ha sido demostrado empíricamente se considera con razón como una mera especulación, no una conclusión válida que merezca credibilidad.

4.4. Ni la experiencia ni las prácticas profesionales pueden sustituir la validez de los fundamentos

En algunas situaciones, un científico puede ser científicamente capaz de emitir juicios basándose principalmente en su «experiencia» y en su «juicio personal». Basándose en su experiencia, un cirujano podría estar cualificado para ofrecer una opinión sobre si otro doctor actuó apropiadamente o no en la sala de operaciones o un psiquiatra podría estar científicamente cualificado para ofrecer una opinión sobre si un acusado está mentalmente bien para defenderse.

Por contraste, ni la «experiencia» ni el «juicio personal» pueden utilizarse para establecer la validez científica y fiabilidad de un método metrológico, como lo es un método de comparación de características forense. La frecuencia con la que un patrón particular o un conjunto de características puede observarse en diferentes muestras, que es un elemento esencial para emitir conclusiones, no es un asunto de «juicio personal». Es una cuestión empírica para la que solo es relevante información empírica. Más aún, la «experiencia» de un analista forense a partir de muchos casos reales no es informativa —porque las «respuestas correctas» no se conocen ordinariamente en los casos reales y por eso los analistas no conocen con precisión con qué frecuencia declaran cotejos positivos erróneos y no es fácil que perfeccionen su precisión aprendiendo de sus errores mientras hacen informes periciales—.

Es importante destacar que las buenas prácticas profesionales —tales como la existencia de sociedades profesionales, programas de certificación, programas de acreditación, artículos revisados por pares, protocolos estandarizados, tests de apti-

[126] *Williams v. United States*, DC Court of Appeals, dictaminado el 16 de enero de 2016 (voto concurrente de Easterly).

tud y códigos éticos— no pueden ser sustitutos de una prueba real sobre la validez y fiabilidad científicas[127].

Análogamente, la manifestación de *confianza* que haga un analista basado en su experiencia profesional o afirmaciones sobre el *consenso* entre quienes practican una técnica respecto la fiabilidad de sus procedimientos no sustituye a las estimaciones de las tasas de error a partir de estudios relevantes. Para que un método sea *fiable*, se requiere una prueba empírica de su validación, como se mencionó anteriormente.

Finalmente, lo que se ha dicho subraya que la validez científica de un método debe valorarse dentro de un marco científico más amplio del que aquel forma parte (por ejemplo, la ciencia de la medida en el caso de métodos de comparación de características). El hecho de que los analistas de marcas de mordedura defiendan la validez de los cotejos de marcas de mordedura significa poco.

4.5. Validez en la aplicación: elementos clave

La validez de los fundamentos significa que un método, *en principio*, puede ser fiable. Validez en la aplicación significa que el método ha sido fiablemente aplicado *en la práctica*. Se trata del concepto científico que se corresponde con el requerimiento jurídico previsto en la Regla 702(d): que un experto «haya aplicado fiablemente los principios y métodos a los hechos del caso».

Desde un punto de vista científico, existen ciertos criterios esenciales para establecer que un analista forense ha aplicado fiablemente un método a los hechos de un caso. Esos elementos se describen en el Cuadro 5.

CUADRO 5. CRITERIOS CLAVE PARA LA VALIDEZ EN LA APLICACIÓN

El analista forense debe haber demostrado que es capaz de aplicar fiablemente el método y debe realmente haberlo hecho.

La demostración de que un analista es *capaz* de aplicar fiablemente el método es crucial —especialmente para métodos subjetivos, en los que los juicios humanos juegan un papel central. Desde el punto de vista científico, la capacidad para aplicar fiablemente un método puede demostrarse solo mediante comprobación empírica que mida con qué frecuencia el analista consigue el resultado correcto. (Los tests de aptitud se tratan más extensamente en las páginas 68-71). La determinación de si un analista ha *realmente* aplicado fiablemente el método requiere que el procedimiento que ha sido de hecho utilizado en el caso, los resultados obtenidos y las notas de laboratorio estén disponibles para una revisión científica llevada a cabo por terceros.

Las aserciones sobre la probabilidad de que las características observadas ocurran por casualidad han de ser científicamente válidas.

[127] Por ejemplo, tanto las disciplinas científicas como las pseudocientíficas emplean tales prácticas.

El analista forense debe informar de la tasa global de falsos positivos y de la sensibilidad del método establecida en los estudios de validez de los fundamentos y debe demostrar que las muestras utilizadas en los estudios sobre su fundamentación son relevantes para los hechos del caso[128].

Cuando sea aplicable, el analista debe informar la probabilidad de coincidencia aleatoria basada en las características específicas observadas en el caso.

Un analista no debe hacer aseveraciones o realizar implicaciones que vayan más allá de la evidencia empírica y de la aplicación de los principios estadísticos válidos para esa prueba.

4.6. Validez en la aplicación: pruebas de aptitud

Incluso aunque un método sea válido en sus fundamentos, hay muchas razones que explican por qué los analistas no siempre consiguen resultados correctos[129]. Como se dijo anteriormente, el único modo de establecer científicamente que un analista es capaz de aplicar un método válido en sus fundamentos es a través de pruebas empíricas apropiadas que midan la frecuencia con la que el analista consigue resultados correctos. A esas pruebas empíricas se les conoce frecuentemente como «pruebas de aptitud». Queremos advertir que el término «prueba de aptitud» es a veces utilizado para referirse a otros tipos de pruebas —tales como (1) tests que determinan si un analista de laboratorio sigue los pasos establecidos en un protocolo, sin que se *valore* la precisión de sus conclusiones, y (2) práctica de ejercicios que ayuden a los analistas de laboratorio a mejorar sus habilidades poniendo de manifiesto sus errores, sin que reflejen con precisión las circunstancias de los casos reales.

---

[128] Por ejemplo, para el análisis de ADN, se conoce que la frecuencia de las variantes genéticas varía entre los grupos étnicos; de este modo, es importante que la recogida de muestras refleje los grupos étnicos relevantes para el caso en cuestión. Para huellas latentes, el riesgo de declarar falsamente una identificación es más alto cuando poseen escasa calidad; así, para ser relevantes, los conjuntos de muestras utilizados para estimar la precisión deben basarse en huellas latentes impresas comparables en calidad y completitud al caso en cuestión.

[129] J.J. Koehler ha enumerado una serie de problemas que podrían producirse, en principio: las características pueden medirse de forma deficiente; las muestras pueden intercambiarse, etiquetarse o codificarse mal, alterarse o contaminarse; el equipamiento puede descalibrarse; fallos técnicos y fallos en general pueden ocurrir sin previo aviso y sin que claramente se manifiesten; y los resultados pueden leerse equivocadamente, malinterpretarse, registrarse deficientemente, mal etiquetarse, mezclarse, descolocarse o descartarse.

Koehler, «Forensics or fauxrensics? Ascertaining accuracy in the forensic sciences», papers.ssrn.com/sol3/papers.cfm?abstract_id=2773255 (última consulta el 28 de junio de 2016).

En este informe, utilizamos el término prueba de aptitud para referirnos a pruebas empíricas en curso que «evalúen la capacidad y la ejecución del trabajo de los analistas» [130, 131, 132].

Los tests de aptitud deben realizarse en condiciones que sean representativas de casos reales y sobre muestras de las que se conozca la verdadera respuesta, y que sean representativas del entero rango de tipos de muestras y calidad que es probable encontrar en un caso real en la aplicación pretendida. (Por ejemplo, el hecho de que un analista supere una prueba de aptitud que contenga un análisis de muestras simples y procedentes de una única fuente de ADN, no demuestra que sea capaz de hacer análisis de ADN de mezclas complejas del tipo que se encuentra en casos reales; véanse pp. 91-100).

Para garantizar su integridad, los tests de aptitud deben ser supervisados por una tercera parte desinteresada que carezca de incentivo institucional o económico que sesgue su realización. Resaltamos que los llamados *Testing Services* han declarado que la comunidad forense prefiere que estas pruebas no sean demasiado difíciles [133].

Como se advirtió previamente, la tasa de falsos positivos consiste tanto en la tasa de coincidencias aleatorias como en fallos técnicos o humanos. Para algunas tecnologías (como el ADN), los últimos pueden ser cientos de veces mayores que los primeros.

La prueba de aptitud es especialmente importante en los métodos subjetivos: como consecuencia de que el procedimiento no está basado únicamente en criterios objetivos, sino que se apoya en el juicio humano, es inherentemente vulnerable al error y a la variabilidad inter-analista. Cada analista debe pasar esos tests, dado que en estudios empíricos se han constatado considerables diferencias en cuanto a la precisión entre los analistas [134, 135].

---

[130] ASCLD/LAB Supplemental Requirements for Accreditation of Forensic Testing Laboratories. des.wa.gov/SiteCollectionDocuments/About/1063/RFP/Add7_Item4ASCLD.pdf.

[131] Queremos resaltar que las pruebas de aptitud no están dirigidas a estimar las tasas de error inherentes a un método; esas tasas deben valorarse mediante estudios de validez de los fundamentos.

[132] El término en inglés «*proficiency testing*» debe también distinguirse de «*competency testing*», que es «la evaluación del conocimiento y capacidad de una persona antes de realizar informes de forma independiente en casos forenses reales». des.wa.gov/SiteCollectionDocuments/About/1063/RFP/Add7_Item4ASCLD.pdf.

[133] Christopher Czyryca, presidente de *Collaborative Testing Services, Inc.*, la empresa líder en Estados Unidos de tests de aptitud, ha declarado públicamente que «la comunidad prefiere los tests sencillos». Reunión de agosto de 2015 de la *National Commission on Forensic Science*, una presentación en el *Accreditation and Proficiency Testing Subcommittee* (Subcomité de Acreditación y Pruebas de Aptitud). www.justice.gov/ncfs/file/761061/download.

[134] Por ejemplo, un estudio sobre decisiones de cotejos de huellas latentes en 2011 observó que los analistas frecuentemente diferían sobre si las huellas eran adecuadas para alcanzar una conclusión. Ulery, Hicklin, Buscaglia, et al., 2011: 7733-7738.

[135] No es suficiente hacer las pruebas de aptitud con voluntarios en un laboratorio, porque los analistas que mejor realizan su tarea son más proclives a participar. Koehler, «Forensics or fauxrensics?

Los problemas contenidos en los test utilizados en las pruebas de aptitud deben hacerse públicos una vez que la prueba haya terminado, lo que permitiría a los científicos valorar si el test ha sido apropiado y adecuado para el propósito perseguido.

Por último, las pruebas de aptitud deben *idealmente* conducirse a ciegas —es decir, con muestras insertadas dentro del flujo de trabajo de forma que los analistas desconozcan que están siendo sometidos a una prueba— (Por ejemplo, la *Transportation Security Administration* (Administración de Seguridad en el Transporte) lleva a cabo pruebas blindadas de envío de armas y explosivos dentro de maletas en los puntos de control para ver la frecuencia con que los controladores los detectan). Es un hecho comprobado en muchos campos de trabajo (incluyendo el de los análisis de huellas latentes) que cuando los individuos son conscientes de que están siendo inspeccionados, realizan sus tareas de forma distinta que cuando los hacen en el transcurso de un día de trabajo ordinario (conocido como «efecto Hawthorne») [136], [137].

Mientras que un test de aptitud ciego es ideal, hay desacuerdo en la comunidad forense sobre su viabilidad en todos los entornos. Por una parte, los laboratorios varían considerablemente con respecto al tipo de casos que reciben, cómo se gestiona y procesa la prueba y qué información se proporciona a un analista sobre la prueba o el caso en cuestión. En consecuencia, los tests de aptitud ciegos interlaboratorios pueden ser difíciles de diseñar y dirigir a gran escala [138]. Por otra parte, los tests de aptitud ciegos se han utilizado en análisis de ADN [139] y algunos selectos laboratorios han comenzado a implementar este tipo de pruebas internamente, como parte de sus programas de aseguramiento de la calidad [140]. Resaltamos que los tests de aptitud ciegos son mucho más fáciles de adoptar en laboratorios que hayan adoptado «procedimientos de gestión de contextos» para reducir sesgos contextuales [141].

El PCAST cree que los tests de aptitud ciegos de los analistas forenses deben fomentarse vigorosamente, con la esperanza de que tengan amplio uso, al menos en los laboratorios grandes, en los próximos cinco años. Sin embargo, el PCAST cree que

---

Ascertaining accuracy in the forensic sciences.» papers.ssrn.com/sol3/papers.cfm?abstract_id=2773255 (última consulta el 28 de junio de 2016).

[136] Sobre el efecto Hawthorne, véanse, por ejemplo: Bracht y Glass, 1968: 437-474; Weech, y Goldhor, 1982: 305-324; Bouchet, Guillemin y Braincon, 1996: 15-20; Mangione-Smith, Elliott, McDonald, et al., 2002: 1603-1623; Mujis, 2006: 53–74; McCarney, Warner, Iliffe, et al., 2007.

[137] Para demostraciones de que los analistas forenses cambian su comportamiento cuando conocen que están siendo inspeccionados de una determinada forma, véase Langenburg, 2009.

[138] Algunos de los retos asociados con tests ciegos de aptitud interlaboratorio podrían abordarse si los laboratorios forenses avanzaran hacia un sistema donde el conocimiento de un caso por parte de los analistas fuera limitado al dominio de la información relevante.

[139] Véase: Peterson, Lin, Ho, et al., 2003: 32-40.

[140] Por ejemplo, el *Houston Forensic Science Center* ha implementado pruebas ciegas de aptitud rutinarias para sus balísticos y la unidad de análisis químico y está planeando llevar a cabo pruebas similares para sus analistas en ADN y en huellas latentes.

[141] Para más detalle, véase www.justice.gov/ncfs/file/888586/download.

no es aún realista exigir esos tests de aptitud ciegos porque los procedimientos aún no han sido diseñados y evaluados.

Mientras no se utilicen otros tests de aptitud, más allá de los no ciegos, para garantizar la validez en la aplicación, resulta científicamente importante informar de esta limitación, incluyendo a los jurados —porque como se dijo más arriba, los tests de aptitud no ciegos tienden a sobrevalorar la precisión porque los analistas conocían que estaban siendo inspeccionados—.

4.7. Puntos de vista no empíricos en la comunidad forense

Mientras que la validez científica de los métodos metrológicos requiere demostración empírica de la precisión, históricamente han existido esfuerzos en la comunidad forense para justificar aproximaciones no empíricas. Esto requiere atención particular porque tales puntos de vista son en ocasiones erróneamente codificados en políticas o en prácticas. Estos modos heterodoxos de ver las cosas repetidamente involucran cuatro temas que revisamos a continuación.

4.7.1. Teorías» de identificación

Un argumento común es que las prácticas forenses deben considerarse válidas porque se asientan sobre «teorías» científicas similares a las leyes fundamentales de la física, que deben ser aceptadas porque han sido comprobadas y no «falsadas»[142].

Un ejemplo es la «Teoría de la Identificación en su relación con las marcas de herramientas», emitida en 2011 por la Asociación de analistas de Marcas de Armas y Herramientas[143, 144]. Afirma rotundamente:

1. La teoría de la identificación, en lo que se refiere a la comparación de las marcas de herramientas, permite opiniones de procedencia que deben hacerse cuando la superficie única de dos marcas de herramientas se halla en «concordancia suficiente».

2. Esta «concordancia suficiente» se relaciona con la duplicación significativa de marcas de herramientas aleatorias, como lo demuestra la correspondencia de un patrón o una combinación de patrones de contornos de superficies. La significación está determinada por el examen comparativo de dos o más conjuntos de patrones de contorno de superficies, compuesto por picos individuales, crestas y surcos. Específicamente, la altura o profundidad relativa, la anchura, la curvatura y la relación espacial entre los picos, crestas y surcos individuales dentro de un conjunto de contornos

---

[142] Véanse: www.swggun.org/index.php?option=com_content&view=article&id=66:the-foundations-of-firearm-and-toolmarkidentification&catid=13:other&Itemid=43 y www.justice.gov/ncfs/file/888586/download.

[143] Association of Firearm and Tool Mark Examiners, 2011: 287.

[144] El análisis de armas de fuego es tratado en detalle en la sección 5.

de superficie se definen y comparan con las correspondientes características del segundo conjunto de contornos de superficie. La concordancia es significativa cuando la hallada entre las características individuales supera la mejor concordancia demostrada entre marcas de herramientas que se sabe que han sido producidas por diferentes herramientas y es consistente con la concordancia demostrada por marcas de herramientas que se sabe que han sido producidas por la misma herramienta. La declaración de «concordancia suficiente» entre dos marcas de herramientas significa que la concordancia entre las características individuales es de una cantidad y calidad tal que la probabilidad de que otra herramienta pudiera haber hecho la marca es tan remota como para considerarla una imposibilidad práctica.

3. Actualmente, la interpretación de identificación/individualización es de naturaleza subjetiva, fundamentada en principios científicos y basada en la formación y experiencia del analista.

La declaración claramente no es una teoría científica, definida por la Academia Nacional de Ciencias como «una comprehensiva explicación de algún aspecto de la naturaleza sustentada por un vasto cuerpo de evidencia»[145]. En su lugar, es una declaración de que los analistas que aplican aproximaciones subjetivas pueden individualizar el origen de una marca de herramienta con precisión. Sin embargo, no es una «teoría» lo que se necesita. Lo que se necesita son pruebas empíricas que permitan ver cuán bien funciona el método.

Más importante aún, el método mencionado es circular. Declara que un analista puede establecer que dos marcas de herramientas tienen un «origen común» cuando sus características están en «concordancia suficiente». Entonces, define «concordancia suficiente» como el que se da cuando el analista considera la «imposibilidad práctica» de que las marcas de herramienta tengan distintos orígenes. (En respuesta a la objeción del PCAST sobre esta circularidad, el laboratorio del FBI replicó que: «la imposibilidad práctica es la certeza que existe cuando hay suficiente concordancia en la cantidad y calidad de características individuales[146]. Esta respuesta no resolvió la circularidad).

4.7.2. Apoyándose en «entrenamiento y experiencia» en vez de en la demostración empírica de la precisión

Muchos analistas creen honestamente que son capaces de hacer juicios precisos sobre identificación basándose en su capacitación y experiencia. Esta noción es explícita en la Teoría de la Identificación de AFTE, que resalta que la interpretación es subjetiva por naturaleza, «basada en la capacitación y experiencia del analista». Análogamente, el libro de texto líder en el análisis de huellas de calzado afirma,

---

[145] Véase: www.nas.edu/evolution/TheoryOrFact.html

[146] Comunicación del Laboratorio del FBI a PCAST (6 de junio de 2016).

> Las identificaciones positivas pueden hacerse con tan solo una característica identificativa aleatoria, pero solo si esa característica es confirmable; tiene suficiente definición, claridad y propiedades; está en la misma ubicación y orientación en la suela del zapato; y *en opinión del analista experimentado, no se presentaría en otro zapato*[147] [énfasis añadido].

En efecto, dice el libro, la identificación positiva depende de que el analista diga que hay una identificación *positiva*.

La «experiencia» es un fundamento inadecuado para emitir dictámenes sobre si dos conjuntos de propiedades pudieran haber sido producidos por (o encontrados en) fuentes diferentes. Incluso si los analistas pudieran recordar con suficiente detalle todos los patrones o conjunto de características que hubieran visto no tendrían modo de saber con precisión en qué casos dos patrones procedieron de fuentes diferentes, porque las respuestas correctas rara vez se conocen en casos reales.

La falacia de confiar en la «experiencia» fue evidente en el testimonio de un exjefe de la Unidad de Huellas Dactilares del FBI (explicado anteriormente) al declarar que el FBI tenía «una tasa de error de 1 cada 11 millones de casos», basándose en el hecho de que la agencia era únicamente consciente de haber cometido un único error[148]. Por contraste, recientes estudios empíricos realizados por el laboratorio del FBI (explicados en la sección 5) indican tasas de error de aproximadamente uno en varios cientos.

La «capacitación» es incluso un fundamento más débil. El mero hecho de que un individuo haya sido capacitado en un método no significa que el método, en sí mismo, sea científicamente válido, ni que el individuo sea capaz de producir respuestas fiables cuando aplique el método.

4.7.3. Centrarse en la «unicidad» en lugar de hacerlo en la «precisión»

Muchas disciplinas forenses de comparación de características se basan en la premisa de que diversos conjuntos de características (por ejemplo, impresiones dactilares, marcas de herramienta sobre proyectiles, la dentición humana, etc.) son «únicas»[149].

---

[147] Bodziak, 2000.

[148] *U.S. v. Baines* 573 F.3d 979 (2009) en 984

[149] Para huellas dactilares, véanse, por ejemplo: Wertheim, 2002: 669 («La ley de unicidad biológica afirma que no puede ocurrir en la naturaleza una replicación exacta (la naturaleza nunca se repite a sí misma) y, por consiguiente, nunca una entidad biológica será exactamente igual que otra») y Budowle, Buscaglia y Perlman 2006. («El uso de las comparaciones de crestas papilares de la piel como medio de identificación se fundamenta en que asume que el patrón de crestas papilares de la piel es único y permanente»). Para armas de fuego, véanse, por ejemplo, Riva y Christope, 2014: 637 («La capacidad para identificar un arma de fuego como fuente de un casquillo o proyectil de origen desconocido se fundamenta en dos premisas que constituyen el fundamento científico de la disciplina. La primera asume la unicidad de las impresiones dejadas por las armas») y SWGGUN Admis Kit (ARK): Foundational Overview of Firearm/Toolmark Identification. disponible en: afte.org/resources/swggun-ark («La base para la identificación en la Identificación de Marcas de Herramientas está fundamentada

La literatura en ciencia forense contiene muchos estudios sobre la «unicidad» que hacen todo lo posible para establecer la exactitud de esa premisa [150]. Por ejemplo, un artículo de 2012 que estudió 39 zapatillas para correr de la marca Adidas Supernova Classic (talla 12), calzadas por un único corredor durante 8 años, corriendo un mismo tiempo diariamente y sobre los mismos tipos de superficie [151]. Después de aplicar crema de zapatos negra en las suelas de los zapatos, el autor pidió al corredor que produjera cuidadosamente marcas de pisadas en hojas de papel legal sobre un suelo de madera. El autor demostró que era posible distinguir pequeñas diferencias identificativas entre las marcas de pisada producidas por los diferentes pares de zapatillas.

No obstante, los estudios de unicidad olvidan la cuestión fundamental. El tema no es si los *objetos* o las *propiedades* difieren; seguramente lo hacen si miramos con suficiente nivel de detalle. El tema es en qué medida y bajo qué circunstancias aplicando un método metrológico *los analistas* pueden *detectar* fiablemente diferencias significativas en las propiedades para identificar fiablemente si comparten un origen común. Los estudios de unicidad que se centran en las propiedades de las características en sí mismas no pueden nunca, por tanto, establecer si un *método* particular para medir y comparar propiedades es válido en sus fundamentos. Solo lo pueden hacer los estudios empíricos.

Más aún, no es *necesario* que las propiedades sean únicas para ser útiles porque reducen las fuentes posibles de una característica. Más bien, es esencial que haya evidencia empírica sobre la frecuencia con que un método atribuye incorrectamente la fuente de una característica.

4.7.4. Disociando conclusiones sobre identificación desde estimaciones sobre precisión

Por último, algunos mantienen la opinión de que cuando la aplicación de un método científico conduce a una conclusión de propuesta de asociación o identificación, es *innecesario* informar al tribunal sobre la fiabilidad del método [152]. A modo

---

en el principio de unicidad . . . por el cual, todos los objetos son únicos por sí mismos, y, de este modo, pueden ser diferenciados unos de otros»). Para marcas de mordedura, véanse, por ejemplo, Kieser, Bernal, Waddell, et al., 2007: 671-677 («Hay dos postulados que subyacen en todos los análisis de marcas de mordedura: primero, que las características de la dentadura anterior actuante en una mordedura son únicas y, segundo, que esta unicidad es registrada con precisión en el material mordido.») y Pretty, 2011 («El análisis de marcas de mordedura está basado en dos postulados: (a) las características dentales de los dientes anteriores actuantes en una mordedura son únicas entre individuos y (b) esta unicidad aseverada es transferida y grabada en la lesión.»).

[150] Algunos autores han criticado los intentos de afirmar la proposición de unicidad basada en observaciones, resaltando que se apoyan en el puro razonamiento inductivo, un método para la investigación científica que «cayó en desgracia durante la época de Sir Francis Bacon en el siglo xvi» Page, Taylor y Blenkin, 2011: 12-18.

[151] Wilson, 2012: 194-204.

[152] Véase: www.justice.gov/olp/file/861936/download.

de fundamento racional, se ha argumentado en ocasiones que no es posible medir las tasas de error perfectamente o que es imposible conocer la tasa de error en el *específico* caso que llevamos entre manos.

Esta noción es contraria al principio fundamental de validez científica en metrología —concretamente, la afirmación de que dos objetos tras ser comparados presenten las mismas propiedades (longitud, peso o patrón dactiloscópico) carece de sentido si no existe información cuantitativa sobre la fiabilidad del proceso de comparación—.

Es una práctica estándar en medicina estudiar e informar sobre tasas de error —tanto para establecer en principio la fiabilidad de un método como para valorar su implementación en la práctica—. Nadie argumenta que medir o informar sobre tasas de error clínicas sea inapropiado porque quizá no reflejen perfectamente la situación de un paciente *específico*. Si la transparencia sobre las tasas de error es apropiada para el acierto en la detección del tipo de sangre antes de una transfusión, es también apropiada para el acierto en el cotejo positivo entre muestras forenses —donde los errores pueden tener consecuencias similares a las amenazas contra la vida—.

Volveremos sobre este tema en la sección 8, donde veremos que la guía sobre testimonio experto recientemente propuesta por el DOJ está basada, en parte, sobre este modo de pensar científicamente inapropiado.

4.8. Puntos de vista empíricos en la comunidad forense

Aunque algunos de los integrantes de la comunidad forense mantengan modos de pensar como el descrito en la anterior sección, un creciente segmento de la comunidad de ciencia forense ha respondido al informe NRC del año 2009 con un mayor reconocimiento sobre la necesidad de realizar estudios empíricos y con esfuerzos iniciales para llevarlos a cabo. Como ejemplos podemos citar algunos estudios de investigación publicados que han sido realizados por investigadores forenses, evaluaciones de necesidades de investigación por Grupos de Trabajo Científicos y Comités de la OSAC y declaraciones del NCFS.

A continuación, resaltamos algunos ejemplos de artículos recientes realizados por científicos forenses: Investigadores de la Academia Nacional de Ciencias y de otros lugares (por ejemplo, Saks y Koeler, 2005; Spinney, 2010) han argumentado que hay una urgente necesidad de desarrollar medidas objetivas de precisión en la identificación mediante huellas dactilares. Aquí presentamos esos datos[153].

— La prueba de impresión de marcas de herramientas, por ejemplo, se ha presentado ante los tribunales con éxito durante décadas, pero su examen ha carecido de prueba científica, estadística, que de forma independiente corroboraría conclusiones basadas en características morfológicas (2-7). En nuestro estudio, aplicaremos méto-

[153] Tangen, Thompson y McCarthy, 2011: 995-997.

dos de reconocimiento de patrones estadístico (por ejemplo, «machine learning») a los análisis de impresiones de marcas de herramienta[154].

— El informe NAS llamó a incrementar las investigaciones en el área de las marcas de mordedura para demostrar que existe un grado de valor probatorio y para una posible restricción del uso de los análisis a la exclusión de individuos. Esta llamada ha de ser escuchada si la prueba de marcas de mordedura quiere ser defendida en adelante como disciplina[155].

— El National Research Council of the National Academies (Consejo de Investigación Nacional de las Academias Nacionales) y las comunidades jurídica y de ciencias forenses han convocado a investigar con el objetivo de medir la precisión y fiabilidad de las decisiones de los analistas de huellas latentes, un problema complejo que constituye un reto que necesita un análisis sistemático. Nuestra investigación está dirigida al desarrollo de aproximaciones empíricas para estudiar este problema[156].

— Pensamos que este informe debe impulsar a la comunidad jurídica a requerir que el emergente campo de la neuroimagen forense, incluyendo el detector de mentiras fundamentado en tecnología FMRI, tenga la apropiada fundamentación científica antes de que sea admitido por los tribunales[157].

— Es preferible una solución empírica que trata al sistema [referido a las huellas acústicas de las voces] como una caja negra y sus salidas como valores puntuales[158].

Análogamente, la OSAC y otros grupos han reconocido grietas críticas en investigación respecto a la evidencia que sostiene a diversas disciplinas de ciencia forense y ha comenzado a desarrollar planes para cerrar alguna de esas grietas. Vale la pena resaltar algunos ejemplos:

— Aunque se han desarrollado algunos estudios de validación de programas de análisis de armas de fuego y marcas de herramienta, la mayoría se han hecho con conjuntos de datos relativamente pequeños. Si se hiciera un gran estudio bien diseñado y tuviera suficiente participación, anticipamos que podrían aprenderse lecciones similares para las disciplinas de armas de fuego y marcas de herramienta[159].

— No conocemos estudio alguno que valore la capacidad de las disciplinas de armas de fuego y marcas de herramientas para categorizar la evidencia correcta/con-

---

154 Petraco, Shenkin, Speir, et al., 2012: 900-911.

155 Pretty y Sweet, 2010: 38-44.

156 Ulery, Hicklin, Buscaglia, et al., 2011: 7733-7738.

157 Langleben y Moriarty, 2013: 222–234.

158 Morrison, Zhang y Rose, 2011: 59–65.

159 OSAC Research Needs Assessment Form (Formulario de evaluación de necesidades de investigación de la OSAC), «Study to Assess the Accuracy and Reliability of Firearm and Toolmark.» Emitido en octubre de 2015 (Aprobado en enero de 2016). Disponible en: www.nist.gov/forensics/osac/upload/FATM-Research-NeedsAssessment_Blackbox.pdf.

sistentemente mediante características de clase, identificar marcas de subclase y descartar vestigios utilizando características individuales [160].

— Actualmente no existe una valoración fiable de la fuerza discriminatoria de los tipos específicos de puntos característicos de las crestas de fricción [161].

— Hasta la fecha, hay pocos datos científicos que cuantifiquen el riesgo global de cotejos negativos dudosos en las bases de datos AFIS. Es difícil crear estándares relativos a la suficiencia para realizar un examen o para realizar búsquedas en AFIS sin este tipo de investigación [162].

— Se precisa investigación para averiguar si el desenmascaramiento secuencial reduce los efectos negativos de sesgos durante el análisis de huellas latentes [163].

— El IAI ha buscado, durante muchos años, apoyo para realizar investigaciones que validasen muchos de los análisis comparativos conducidos por sus miembros como analistas forenses. Aunque existe una gran cantidad de evidencia empírica que apoya a esos análisis, no se ha hecho una validación independiente [164].

La Comisión Nacional de Ciencia Forense ha reconocido análogamente la necesidad de realizar una rigurosa evaluación empírica de los métodos forenses en un documento que compila diversas opiniones al respecto aprobado por la comisión:

> Todas las metodologías de la ciencia forense deben ser evaluadas por un cuerpo científico independiente que defina sus capacidades y limitaciones para responder con precisión y fiabilidad a una específica y claramente definida cuestión forense [165].

El PCAST aplaude este creciente esfuerzo por centrarse en la evidencia empírica. Resaltamos que será necesaria una mayor inversión en investigación para alcanzar estos objetivos críticos (véase Sección 6).

---

[160] OSAC Research Needs Assessment Form (Formulario de evaluación de necesidades de investigación de la OSAC). «Assessment of Examiners' Toolmark Categorization Accuracy.» Emitido en octubre de 2015 (Aprobado en enero de 2016). Disponible en: www.nist.gov/forensics/osac/upload/FATM-Research-NeedsAssessment_Class-and-individual-marks.pdf.

[161] OSAC Research Needs Assessment Form (Formulario de evaluación de necesidades de investigación de la OSAC). «Assessing the Sufficiency and Strength of Friction Ridge Features.» Emitido en octubre de 2015. Disponible en: www.nist.gov/forensics/osac/upload/FRS-Research-Need-Assessment-of-Features.pdf.

[162] OSAC Research Needs Assessment Form (Formulario de evaluación de necesidades de investigación de la OSAC). «Close Non-Match Assessment» , Emitido en octubre de 2015. Disponible en: www.nist.gov/forensics/osac/upload/FRS-Research-Need-Close-Non-Match-Assessment.pdf.

[163] OSAC Research Needs Assessment Form (Formulario de evaluación de necesidades de investigación de la OSAC). «ACE-V Bias.» Emitido en octubre de 2015. Disponible en: www.nist.gov/forensics/osac/upload/FRS-Research-Need-ACE-V-Bias.pdf.

[164] International Association for Identification (Asociación Internacional para le Identificación). Letter to Patrick J. Leahy, Chairman, Senate Committee on the Judiciary, March 18, 2009. Disponible en: www.theiai.org/current_affairs/nas_response_leahy_20090318.pdf.

[165] National Commission on Forensic Science, (Comisión Nacional sobre Ciencia Forense): «Views of the Commission Technical Merit Evaluation of Forensic Science Methods and Practices». Disponible en: www.justice.gov/ncfs/file/881796/download.

4.9. Resumen de hallazgos científicos

HALLAZGO 1:
CRITERIOS CIENTÍFICOS PARA LA VALIDEZ CIENTÍFICA DE UN MÉTODO DE COMPARACIÓN DE CARACTERÍSTICAS FORENSE

(1) Validez de los fundamentos. Para establecer la validez de los fundamentos de un método de comparación de características, se requieren los siguientes elementos:

(a) un procedimiento reproducible y consistente para (i) la identificación de las propiedades en muestras que son aportadas como pruebas; (ii) la comparación de las propiedades entre dos muestras; (iii) la determinación, basándose en la similitud de las propiedades entre dos conjuntos de características, sobre si se debe declarar que las muestras proceden probablemente del mismo origen (regla de coincidencia"); y

(b) estimaciones empíricas, tras estudios apropiadamente diseñados a partir de múltiples grupos, que establezcan: (i) la tasa de falsos positivos del método —es decir, la probabilidad de una propuesta de identificación entre muestras que realmente proceden de diferentes fuentes— y (ii) la sensibilidad del método —es decir, la probabilidad de una propuesta de identificación entre muestras que realmente proceden de la misma fuente—.

Como se describió en el Cuadro 4, la validación científica debe satisfacer una serie de criterios: (a) debe basarse en conjuntos suficientemente grandes de muestras conocidas y representativas de poblaciones relevantes; (b) debe conducirse de tal modo que los analistas no tengan información sobre las respuestas correctas; el diseño del estudio y el plan de análisis deben especificarse por adelantado y no modificarse posteriormente después de alcanzar los resultados; (d) el estudio debe ser conducido o supervisado por individuos u organizaciones sin beneficio por los resultados; (e) los datos, el software y los resultados deben estar disponibles para permitir que otros científicos puedan revisar las conclusiones; y (f) para asegurar que los resultados sean robustos y reproducibles, debe haber diversos estudios independientes realizados por grupos separados que alcance resultados similares.

Una vez que un método ha sido establecido como válido en sus fundamentos mediante adecuados estudios empíricos, las afirmaciones sobre la precisión y el valor probatorio de las propuestas de identificación, para que sean válidas, deben basarse sobre tales estudios empíricos.

Para métodos objetivos, la validez de los fundamentos puede establecerse demostrando la fiabilidad de cada uno de los pasos individuales (identificación de características, comparación de propiedades, regla de coincidencia, probabilidad de coincidencia falsa y sensibilidad).

Para métodos subjetivos, la validez de los fundamentos *solo* puede establecerse mediante estudios de caja negra que midan la frecuencia en la que muchos examinadores alcancen conclusiones precisas en muchos problemas de comparación de características que contengan muestras representativas del uso pretendido. En ausencia de tales estudios, un método subjetivo de comparación de características no puede considerarse científicamente válido.

La validez de los fundamentos es un *sine qua non* y solo puede demostrarse mediante estudios empíricos. Es muy importante resaltar que las buenas prácticas profesionales —tales como la existencia de asociaciones profesionales, programas de certificación, programas de acreditación, artículos revisados por pares, protocolos estandarizados, pruebas de aptitud y códigos éticos— no pueden sustituir a la evidencia empírica de la validez y fiabilidad científica.

(2) Validez en la aplicación. Una vez que un método de comparación de características forense ha sido considerado como válido en sus fundamentos, es necesario establecer su validez en la aplicación en un caso dado.

Como se describió en el Cuadro 5, la validez en la aplicación requiere que: (a) el examinador forense debe haber demostrado que es *capaz* de aplicar el método fiablemente, lo que se hace sometiéndose a tests de aptitud apropiados (véase Sección 4.6) y debe *de hecho* haberlo llevado a cabo, como se ha demostrado en el procedimiento *de hecho* utilizado en el caso, los resultados obtenidos y las notas de laboratorio, que deben estar disponibles para que otros puedan revisarlas científicamente; y (b) las aserciones sobre el valor probatorio de las propuestas de identificación deben ser científicamente válidas —incluyendo que los examinadores deben informar sobre la tasa completa de falsos positivos y la sensibilidad del método establecidas en los estudios de validez de los fundamentos—; se demuestre que las muestras utilizadas en los estudios de los fundamentos son relevantes para los hechos del caso; y cuando sea aplicable, que se demuestre el valor probatorio de la coincidencia observada a partir de las características específicas observadas en el caso; además de no realizar afirmaciones o implicaciones que vayan más allá de la evidencia empírica.

5. Evaluación de la validez científica de siete métodos de comparación de características

En la sección precedente, describimos los criterios científicos que un método de comparación de características debe cumplir para ser considerado científicamente válido y fiable, y subrayamos la necesidad de contar con evidencia empírica sobre su precisión y fiabilidad.

En esta sección, ilustraremos el significado de esos criterios aplicándolos a seis métodos de comparación de características forense específicos: (5.1) análisis de ADN de un único origen y de muestras con mezclas simples; (5.2) análisis de ADN de muestras con mezclas complejas; (5.3) marcas de mordedura; (5.4) huellas dactilares latentes; (5.5) identificación de armas de fuego y (5.6) análisis de calzado[166]. Con

[166] La American Association for the Advancement of Science (AAAS) (Asociación estadounidense para el avance de la ciencia) está llevando a cabo un análisis de las bases científicas subyacentes de las herramientas y métodos forenses utilizados actualmente en el sistema de justicia penal. A fecha 1 de septiembre de 2016, todavía no han emitido informe alguno. Véase: www.aaas.org/page/forensic-science-assessments-quality-and-gap-analysis.

respecto al séptimo método de comparación de características, el análisis de cabellos, no realizamos una evaluación completa, pero revisamos una evaluación reciente del DOJ.

En adelante, se evaluará si esos métodos han sido considerados como válidos en sus fundamentos y fiables y, si así fuera el caso, qué estimación de precisión debería acompañar al testimonio pericial concerniente a una propuesta de identificación que está fundamentada en estudios científicos actuales. También discutiremos brevemente algunos asuntos relacionados con la validez en la aplicación.

El PCAST compiló una lista de 2019 artículos procedentes de diversas fuentes —incluyendo bibliografías preparadas por el Subcomité sobre Ciencia Forense del Consejo Nacional de Ciencia y Tecnología, los Grupos de Trabajo Científicos relevantes (predecesores del actual OSAC) [167] y los comités de OSAC relevantes; documentos que se presentaron en respuesta a la solicitud del PCAST para obtener información de los miembros interesados de la comunidad de la ciencia forense; y nuestras propias búsquedas de literatura especializada [168]—. Miembros y personal del PCAST identificaron y revisaron esos artículos relevantes para establecer la validez científica. Después de alcanzar un conjunto de conclusiones iniciales, se recibió información del laboratorio del FBI y de científicos individuales del NIST, así como de otros analistas —incluyendo la petición de identificación de artículos adicionales que sustentaran la validez científica y que quizá habíamos omitido—.

En cada uno de los métodos proporcionaremos una breve visión general de la metodología, explicaremos sus antecedentes y estudios y revisamos la evidencia sobre la validez científica.

Como se trató en la sección 4, los métodos objetivos tienen procedimientos bien definidos para (1) identificar las características en las muestras, (2) medir las propiedades, (3) determinar si las propiedades de dos muestras coinciden dentro de una medida establecida de tolerancia (regla de coincidencia) y (4) estimar la probabilidad de que dos muestras de diferentes fuentes coincidan (probabilidad de coincidencia falsa). Es posible examinar la validez y fiabilidad de cada uno de estos pasos diferentes. De los seis métodos considerados en esta sección, solo los dos primeros métodos (que conllevan análisis de ADN) emplean métodos objetivos. Los restantes cuatro métodos son subjetivos.

En los métodos subjetivos los procedimientos no están definidos con precisión, por lo que se requiere sustancialmente del juicio humano experto. Los analistas pueden centrarse en ciertas propiedades mientras ignoran otras, pueden compararlas de distintos modos y pueden tener distintos estándares para declarar una propuesta de identificación entre muestras. Como se describió en la sección 4, la única forma

[167] Véase: www.nist.gov/forensics/workgroups.cfm.

[168] Véase: www.whitehouse.gov/sites/default/files/microsites/ostp/PCAST/pcast_forensics_references.pdf

de establecer la validez de los fundamentos es a través de múltiples e independientes estudios de «caja negra» que midan la frecuencia en que los analistas alcanzan conclusiones precisas en muchos problemas de comparación de características que contengan muestras representativas para el uso pretendido. En ausencia de tales estudios, un método de comparación de características no puede considerarse como científicamente válido.

El PCAST encontró pocos estudios de caja negra apropiadamente diseñados para valorar la validez científica de métodos subjetivos. Dos notables excepciones, tratadas en esta sección, fueron un estudio sobre huellas dactilares latentes dirigido por el Laboratorio del FBI y un estudio sobre identificación de armas de fuego patrocinado por el Ministerio de Defensa y dirigido por el Laboratorio AMES del Departamento de Energía.

Nos hemos pronunciado sobre si las pruebas de aptitud, dirigidas por organizaciones comerciales en algunas disciplinas, pudieran ser útiles para establecer la validez de los fundamentos. A la fecha, concluímos que no por diversas razones. En primer lugar, las pruebas de aptitud no están dirigidas a establecer la validez de los fundamentos. En segundo lugar, los problemas de los tests o el conjunto de tests utilizados en pruebas de aptitud comerciales ordinariamente no se hacen públicos —haciendo imposible determinar si los tests valoran apropiadamente el método en el rango de aplicaciones en el que es utilizado—. La publicación y la revisión crítica de los métodos y datos es un componente esencial en el establecimiento de la validez científica. En tercer lugar, la empresa dominante en el mercado, *Collaborative Testing Services, Inc. (CTS)*, afirma explícitamente que sus pruebas de aptitud no son apropiadas para estimar las tasas de error de una disciplina porque (a) los resultados de los tests, que están abiertos a cualquiera, puede que no reflejen las habilidades de los analistas forenses y (b) «los resultados de los que se informa no reflejan respuestas "correctas" o "incorrectas", sino más bien respuestas que están de acuerdo o no con las conclusiones consensuadas entre la población participante[169]». En cuarto lugar, los tests para métodos de comparación de características forenses constan normalmente solo de uno o dos problemas al año. Y, en quinto lugar, «la comunidad favorece los tests fáciles», pues tests que sean demasiado difíciles podría poner en peligro la repetición del negocio para un proveedor comercial[170].

---

[169] Véase: www.ctsforensics.com/assets/news/CTSErrorRateStatement.pdf

[170] El PCAST agradece al Presidente de Collaborative Testing Services, Inc. *(CTS)*, Christopher Czyryca, las útiles conversaciones con él mantenidas sobre los tests de aptitud. Czyryca explicó que (1) CTS define que hay consenso cuando se alcanza acuerdo, al menos, entre el 80% entre quienes responden y (2) las pruebas de aptitud para huellas latentes solo ocasionalmente incluyen un problema en el cual la impresión cuestionada *no coincide con alguna* de las posibles respuestas. Czyryca resaltó que la comunidad forense no es favorable a tests más retadores —y que las empresas de tests temen que podrían ver mermado su negocio si sus tests fueron vistos como demasiado difíciles—. Un ejemplo de test «retador» es el importantísimo escenario en el que ninguna de las muestras cuestionadas coincide con algunas de las muestras conocidas: a causa de que los analistas pueden esperar encontrar alguna coincidencia, tales escenarios proporcionan una oportunidad para valorar con qué frecuencia los analistas

Las observaciones y conclusiones del PCAST que a continuación se recogen son en gran medida coherentes con las conclusiones de informes anteriores de la NRC[171].

5.1. Análisis de ADN de muestras procedentes de una única fuente y de mezclas simples

El análisis de ADN de una muestra procedente de una única fuente o de una mezcla simple incluye excelentes ejemplos de métodos objetivos en los que la validez de sus fundamentos ha sido apropiadamente establecida[172].

5.1.1. Metodología

El análisis de ADN implica comparar perfiles genéticos de diferentes muestras para ver si una muestra conocida pudo haber sido el origen de una muestra cuestionada (habiendo sido aportada como prueba).

Para generar un perfil genético, se *extrae* el ADN mediante un proceso químico de una muestra que contiene material biológico, como sangre, semen, cabello o células de la piel. Después, se *amplifica* un conjunto predeterminado de segmentos de ADN («marcadores») que contienen pequeñas secuencias repetidas[173], utilizando la *Polymerase Chain Reation* (PCR) (Reacción en Cadena de la Polimerasa): un proceso enzimático que replica un segmento de ADN dirigido una y otra vez a producir millones de copias. Después de la amplificación, las longitudes de los fragmentos de ADN resultantes se miden utilizando una técnica llamada electroforesis capilar, basada en el hecho de que los fragmentos más largos se mueven más lentamente que los fragmentos más cortos a través de una solución de polímero. Los datos originales recogidos de este proceso se analizan mediante un software que produce una imagen gráfica (un electroferograma) y una lista de números (el perfil de ADN) que se corresponden con los tamaños de cada uno de los fragmentos («mediante la comparación con tamaños de moléculas conocidos y estandarizados»).

Como práctica actual, el método utiliza 13 marcadores específicos y el proceso de amplificación se diseña para que los fragmentos de ADN correspondientes a los diferentes marcadores ocupen diferentes rangos de tamaño —facilitando reconocer

---

declaran coincidencias que son falsos positivos. (Véase también la presentación a la Comisión Nacional de la Ciencia Forense por el Presidente de CTS, el Sr. Czyryca, resaltando que «La comunidad favorece los tests fáciles». www.justice.gov/ncfs/file/761061/download.)

[171] National Research Council, 2009; National Research Council, 2008.

[172] El análisis forense de ADN pertenece a dos disciplinas anteriores de las que depende —metrología y genética molecular— y se ha beneficiado de la extensa aplicación de la tecnología de ADN en investigación biomédica y en aplicaciones médicas.

[173] Las repeticiones, denominadas repeticiones cortas en tandem (STRs), consisten en copias consecutivas repetidas de segmentos de 2 a 6 pares de bases.

qué fragmentos vienen de cada marcador[174]—. En cada marcador, cada ser humano lleva dos variantes (denominadas 'alelos'), uno heredado de la madre y otro del padre —que pueden ser de la misma o diferente longitud—[175].

5.1.2. Análisis de muestras procedentes de una única fuente

El análisis de ADN de una muestra procedente de un único individuo es un método objetivo. Además de que los protocolos del laboratorio están definidos con precisión, la interpretación también conlleva poco o ningún juicio humano.

Un analista puede valorar si una muestra procedió de una única fuente basándose en si el perfil de ADN contiene, ordinariamente y para cada marcador, exactamente un fragmento de cada cromosoma que contiene el marcador, lo que produce una o dos longitudes de fragmentos distintos de cada marcador[176]. El perfil de ADN puede compararse luego con el perfil de ADN de un sospechoso conocido. Puede también ser introducido en el *FBI's National DNA Index System (NDIS)* (Sistema Indexado de Perfiles de ADN Nacional del FBI) y buscado en una base de datos de perfiles de ADN de condenados (y detenidos, en más de la mitad de los estados) o de crímenes aún no resueltos.

Se declara que dos perfiles de ADN coinciden si las listas de alelos son iguales[177]. La probabilidad de que dos perfiles de ADN de *diferentes* fuentes tengan el mismo perfil (la probabilidad de coincidencia aleatoria) se calcula después fundamentándose

---

[174] El actual kit utilizado por el FBI (Identifiler Plus) tiene 16 marcadores en total: 15 marcadores STR y el marcador «Amelogenina». Más adelante, durante este año, se implementará un kit que tiene 24 marcadores.

[175] El FBI anunció en 2015 que planea expandir los marcadores que conforman su núcleo añadiendo los 7 que normalmente más se utilizan en las bases de datos de otros países. (Los datos poblacionales han sido publicados para el conjunto expandido, incluyendo frecuencias en 11 poblaciones étnicas www.fbi.gov/about-us/lab/biometric-analysis/codis/expanded-fbi-str-2015-final-6-16-15.pdf.) A principios de 2017, se requerirán esos marcadores para introducir y buscar perfiles de ADN en el sistema nacional. Se espera que los datos expandidos en cada perfil provean de mayor capacidad de potencial discriminante para la identificación, especialmente en muestras coincidentes, aunque con solo perfiles parciales de ADN, investigaciones de personas desaparecidas y en casos de aplicación de leyes internacionales y contra el terrorismo.

[176] El analista revisa el electroferograma para determinar si cada uno de los picos es un verdadero alelo o un artefacto (por ejemplo, ruido de fondo en forma de pequeños picos antes de un pico alélico, agujas y otros fenómenos) y para determinar si más de un individuo pudiera haber contribuido al perfil. En casos raros, un individuo puede tener más de dos fragmentos en un marcador debido a una extraña variación del número de copias en el genoma humano.

[177] Cuando solo es posible conseguir un perfil parcial en la muestra que constituye la prueba (por ejemplo, en casos con limitadas cantidades de ADN, degradación de la muestra o la presencia de inhibidores de PCR), el analista puede también informar de una «inclusión» si el perfil parcial es *consistente* con el perfil de ADN obtenido de una muestra de referencia. El analista puede también informar de una inclusión cuando los resultados de ADN de una muestra de referencia estén presentes en una mezcla. Estos casos generalmente requieren significativamente mayor análisis e interpretación humanos que las muestras de una única fuente.

en las frecuencias medidas empíricamente de cada alelo y en principios establecidos de la genética poblacional (véase p. 92) [178].

5.1.3. Análisis de mezclas simples

Muchos casos de agresiones sexuales conllevan mezclas de ADN de dos individuos, donde uno de ellos (por ejemplo, la víctima) es conocido. El análisis de ADN de esas mezclas simples es también relativamente sencillo. Se han utilizado métodos durante 30 años para extraer, diferenciadamente, ADN de células de esperma y de células epiteliales vaginales, haciendo posible generar perfiles de ADN de las dos fuentes. Cuando los dos tipos de células son iguales y uno de los contribuyentes es conocido, los alelos del individuo conocido pueden sustraerse del conjunto de alelos identificados en la mezcla [179].

5.1.4. Validez de los fundamentos

Para evaluar la validez de los fundamentos de un método objetivo (tales como los análisis de una única fuente y de mezclas simples), podemos examinar la fiabilidad de cada uno de los pasos individuales en lugar de tener que confiar en estudios de caja negra.

5.1.5. Muestras de una única fuente

Cada paso en el análisis es objetivo y conlleva poco o ningún juicio humano.

(1) Identificación de características. En contraste con otros métodos tratados en este informe, las características utilizadas en los análisis de ADN (las longitudes de los marcadores) están definidas *de antemano*.

(2) Medida de las características y comparación. La amplificación PCR, inventada en 1983, se usa en decenas de miles de laboratorios de biología molecular, incluyéndose muchas aplicaciones médicas para las cuales ha sido rigurosamente validada. Los kits PCR multiplex diseñados por casas comerciales para su uso en laboratorios forenses deben validarse tanto externamente (a través de estudios de validación de su desarrollo, publicados en revistas revisadas por pares) como internamente (por cada

[178] Las probabilidades de coincidencia aleatoria pueden también expresarse en términos de relaciones de verosimilitud (LR), que es la relación de (1) la probabilidad de observar el perfil de ADN si el individuo en cuestión es la fuente de la muestra de ADN y (2) la probabilidad de observar el perfil de ADN si el individuo en cuestión *no* es la fuente de la muestra de ADN. En la situación de una muestra procedente de una única fuente, el LR debe ser simplemente el recíproco de la probabilidad de coincidencia aleatoria (porque la primera probabilidad en el LR es 1 y la segunda es la probabilidad de coincidencia aleatoria).

[179] En muchos casos, el ADN estará presente en la mezcla en diferentes pero suficientes cantidades de forma que las alturas de los picos en el electroferograma procedente de las dos fuentes puedan distinguirse, permitiendo al analista separar más fácilmente las fuentes.

laboratorio que desee utilizar el kit), antes de que puedan utilizarse[180]. Los tamaños de los fragmentos se miden mediante un procedimiento automatizado cuya variabilidad está bien caracterizada y es pequeña; la desviación estándar es de, aproximadamente, 0.05 pares de bases, lo cual proporciona mediciones altamente fiables[181], [182]. Se han llevado a cabo estudios de validación del desarrollo —incluyendo al FBI— para verificar la precisión, repetibilidad y reproducibilidad del procedimiento[183, 184].

(3) Comparación de características. En el caso de muestras procedentes de una única fuente, existen «reglas de coincidencia» claras y bien especificadas para declarar si los perfiles de ADN coinciden. Cuando se buscan en NDIS perfiles de ADN completos con criterios de «elevada rigurosidad», se encuentran «coincidencias» únicamente cuando cada alelo del perfil desconocido coincide con un alelo del perfil conocido, y *viceversa*. Cuando se buscan perfiles parciales obtenidos de muestras parcialmente degradadas o contaminadas con criterios de «moderada rigurosidad», se muestran candidatos si uno de los alelos del perfil desconocido coincide con un alelo del perfil conocido[185, 186].

---

[180] Se requiere que los laboratorios que realizan análisis de ADN con fines forenses sigan los Estándares de Aseguramiento de la Calidad del FBI para Laboratorios de Pruebas de ADN como condición para participar en el Sistema Indexado de ADN Nacional: www.fbi.gov/about-us/lab/biometricanalysis/codis/qas-standards-for-forensic-dna-testing-laboratories-effective-9-1-2011. El FBI's Working Group on DNA Analysis Methods (SWGDAM) (Grupo de Trabajo Científico del FBI sobre Métodos de Análisis de ADN) ha publicado directrices para laboratorios relativas a los procesos de validación de acuerdo con los Estándares de Aseguramiento de la Calidad del FBI (QAS). SWGDAM Validation Guidelines for DNA Analysis Methods (Directrices para la Validación de Métodos de Análisis de ADN del Grupo de Trabajo SWGDAM), diciembre de 2012. Véase: www.media.wix.com/ugd/4344b0_cbc27d16dcb64fd88cb36ab2a2a25e4c.pdf.

[181] Los laboratorios forenses generalmente utilizan analizadores genéticos desarrollados por el grupo Applied Biosystems de Thermo Fisher Scientific (ABI 310, 3130 o 3500).

[182] Para estimar incorrectamente la longitud de un fragmento por un par de bases (la mínima diferencia en tamaño), se requiere un error de medida de 0.5 pares de bases, que se corresponde con 10 desviaciones estándar. Además, ordinariamente los alelos se diferencian por, al menos, 4 pares de bases (aunque algunos marcadores STR tengan alelos bastante comunes que difieren en uno o dos nucleótidos).

[183] Como ejemplos de esos estudios véanse: Budowle, Moretti, Keys, Koons y Smerick, 1997: 701-707; Kimpton, Oldroyd, Watson, et al., 1996: 1283-1293; Lygo, Johnson, Holdaway, et al., 1994: 77-89; Fregeau, Bowen y Fourney, 1999: 133-166.

[184] Por ejemplo, en un estudio de 2001 se compararon las características de funcionamiento de distintos kits de STR disponibles comercialmente, comprobándose la consistencia y la reproducibilidad de los resultados utilizando muestras de casos previamente tipados, muestras influenciadas ambientalmente y muestras de fluidos corporales depositadas sobre varios sustratos. El estudio halló que todos los kits se podían utilizar para amplificar y tipar marcadores STR satisfactoriamente, así como que todos los procedimientos utilizados para cada uno de los kits eran robustos y válidos. No se hallaron evidencias de resultados de falsos positivos o falsos negativos, así como tampoco hubo evidencia sustancial de amplificación preferente dentro de un marcador en ninguno de los kits para las pruebas. Moretti, Baumstark, Defenbaugh, et al., 2001: 647-660.

[185] Véase: FBI's Frequently Asked Questions (FAQs) on the CODIS Program and the National DNA Index System. www.fbi.gov/about-us/lab/biometric-analysis/codis/codis-and-ndis-fact-sheet.

[186] Las muestras contaminadas no se introducen en NDIS.

(4) Estimación de la probabilidad de coincidencia aleatoria. El proceso de cálculo de la probabilidad de coincidencia aleatoria (es decir, la probabilidad de una coincidencia por azar) se fundamenta en principios bien establecidos en la genética poblacional y estadística. El FBI ha obtenido las frecuencias de los alelos individuales a partir de perfiles de aproximadamente 200 individuos no emparentados entre sí, pertenecientes a 6 grupos poblacionales y fueron evaluadas antes de que fueran utilizadas [187]. La frecuencia de un patrón completo de alelos —es decir, la probabilidad de coincidencia aleatoria— se estima generalmente multiplicando las frecuencias de los marcadores individuales, asumiendo que los alelos son independientes entre sí— [188]. La probabilidad resultante suele ser ordinariamente menor a 1 en 10 billones[*], excluyendo la posibilidad de parientes cercanos [189]. (Nota: La multiplicación de la frecuencia de los alelos puede sobreestimar la rareza de un patrón porque los alelos pueden no ser completamente independientes debido a la subestructura de población. Un informe del NRC de 1996 concluyó que el efecto de la subestructura poblacional sobre el valor calculado venía a estar probablemente dentro de un factor de 10 (por ejemplo, para una probabilidad de frecuencia aleatoria de 1 entre 10 millones, es altamente probable que la verdadera probabilidad esté entre 1 en 1 millón y 100 millones) [190]. Sin embargo, un estudio reciente realizado por científicos del NIST sugiere que la variación puede ser sustancialmente más grande que 10 veces más [191].

---

[187] Los datos de población iniciales generados por el FBI incluían datos de 6 poblaciones étnicas distintas, con tamaños muestrales de 200 individuos. Véase: Budowle, Moretti, Baumstark, et al., 1999: 1277-1286; Budowle, Shea, Niezgoda, et al., 2001: 453-89. En julio de 2015 se informó de errores encontrados en la base de datos original. (Erratum, *Journal of Forensic Sciences*, Vol. 60, No. 4 (2015): 1114-6, se valoró el impacto de estas discrepancias sobre los cálculos de probabilidad de los perfiles (y se encontró ser menor que un factor de 2 en un perfil completo) y las frecuencias alélicas estimadas fueron correspondientemente corregidas. Al mismo tiempo que se corrigió el conjunto de datos original, el Laboratorio del FBI también publicó conjuntos de datos adicionales tras retipar las muestras originales para incluir marcadores adicionales. Además, las muestras poblacionales que habían sido estudiadas originalmente en otros laboratorios fueron tipadas con marcadores adicionales, de tal forma que la base de datos completa incluye nueve poblaciones. Estos conjuntos de datos «expandidos» están en uso en el Laboratorio del FBI y pueden encontrase en www.fbi.gov/about-us/lab/biometric-analysis/codis/expanded-fbi-str-final-6-16-15.pdf.

[188] De forma más precisa, primero se calcula la frecuencia genotípica de cada marcador. Si el marcador tiene dos copias del mismo alelo con frecuencia p, la frecuencia se calcula como p2. Si el marcador tiene dos alelos diferentes con frecuencias respectivas p y q, la frecuencia genotípica se calcula como 2pq. La frecuencia del patrón completo se calcula multiplicando los valores obtenidos para cada marcador individual.

[*] N. del T.: el billón anglosajón equivale a 1000 millones o 109 en el sistema decimal; el billón continental como millón de millones es también representado como 1012).

[189] La probabilidad de coincidencia aleatoria es más elevada entre parientes cercanos. Se espera que los perfiles de ADN de dos gemelos idénticos coincidan perfectamente. Para parientes en primer grado, la probabilidad de coincidencia aleatoria es del orden de 1 entre 100,000 cuando se analizan los 13 marcadores STR del núcleo de CODIS. Véase: Butler, 2015.

[190] National Research Council, 1996; Goode, 2002: 45-77.

[191] Gittelson y Buckleton, 2016. Véase: www.cstl.nist.gov/strbase/pub_pres/Gittelson-AAFS2016- Factor-of-10.pdf

La probabilidad de coincidencia aleatoria debe calcularse utilizando una fórmula estadística adecuada que tenga en cuenta la subestructura de población[192].

5.1.6. Mezclas simples

Las etapas para analizar mezclas simples son las mismas que para analizar muestras de una única fuente, hasta incluso la interpretación. Los perfiles de ADN que contienen una mezcla de dos contribuyentes, donde uno de ellos es conocido, pueden interpretarse del mismo modo que las muestras procedentes de una única fuente. Esto ocurre frecuentemente en casos de agresiones sexuales, donde un perfil genético contiene una mezcla de ADN de la víctima y del agresor. Los métodos que se utilizan para extraer ADN diferenciadamente de las células de esperma y de las células epiteliales vaginales en agresiones sexuales están bien establecidos[193]. Cuando los dos tipos celulares son el mismo, uno de los contribuyentes a la mezcla de ADN puede ser dominante, produciéndose un distintivo contraste entre las alturas de los picos de los dos contribuyentes; en estos casos, los alelos del contribuyente mayoritario (correspondiente al de los picos alélicos más grandes) y los del contribuyente minoritario pueden ser, por lo general, fiablemente interpretados, cuando la proporción del contribuyente minoritario no sea demasiado baja[194].

5.1.7. Validez en la aplicación

Aunque el análisis de ADN de muestras de una única fuente y el de mezclas simples es un método válido en sus fundamentos y fiable, no es infalible en la práctica. Pueden ocurrir y ocurren errores en las pruebas de ADN. Aunque la probabilidad de que dos muestras de diferentes fuentes tengan el mismo perfil de ADN es pequeña, la probabilidad de que ocurra un error humano es mucho más alta. Tales errores pueden derivarse de mezclas de muestras, contaminación, interpretación incorrecta y errores en los informes[195].

Para minimizar el error humano, el FBI requiere, como una condición para participar en NDIS, que los laboratorios sigan los Estándares de Aseguramiento de la

---

[192] Balding y Nichols, 1994: 125-140.

[193] Gill, Jeffreys y Werrett, 1985: 577-579.

[194] Clayton, Whitaker, Sparkes, et al., 1998: 55-70.

[195] Krimsky y Simoncelli, 2011. Quizá, el más espectacular error humano hasta la fecha está relacionado con la investigación del gobierno alemán del caso «*Phantom of Heilbronn*», una mujer cuyo ADN apareció en la escena del crimen en más de 40 delitos de tres países, incluyendo 6 asesinatos, varios atracos y docenas de allanamiento de morada en el transcurso de más de una década. Después de un esfuerzo que incluyó el análisis de muestras de ADN de más de 3000 mujeres de cuatro países y que costó 18 millones de dólares, las autoridades descubrieron que la mujer de interés era una trabajadora de una fábrica austríaca que fabricaba hisopos utilizados para la recogida de ADN. La mujer había contaminado inadvertidamente un gran número de hisopos con su propio ADN y, por eso, se encontró en muchos análisis de ADN.

Calidad del FBI (QAS)[196]. Antes de que los resultados de los análisis de ADN puedan ser comparados, el analista tiene que someterse a una serie de controles que comprueban si ha podido producirse una posible contaminación y que aseguran que el proceso PCR ha transcurrido apropiadamente. El QAS también requiere que todos los analistas que realizan pruebas de ADN en casos forenses se sometan a una prueba de aptitud cada semestre. Los resultados de las pruebas no tienen por qué publicarse, pero el laboratorio debe guardarlos, así como las discrepancias, errores cometidos y acciones correctivas efectuadas[197].

Los analistas que practican la ciencia forense en los Estados Unidos generalmente no informan de cuestiones de calidad que surjan durante los análisis forenses. Por contraste, las tasas de error en pruebas médicas generalmente se miden y se informa sobre ellas[198]. Merece destacarse un artículo del año 2014 dimanante del Instituto Forense Holandés (NFI), una agencia gubernamental, que realizó un amplio análisis de todas las «notificaciones sobre cuestiones de calidad» halladas en casos forenses, categorizadas por tipología, fuente e impacto[199, 200]. Los autores piden mayor «transparencia» y un «cambio cultural» escribiendo que:

> El análisis de ADN con fines forenses se realiza en un gran número de laboratorios de todo el mundo, tanto pertenecientes a empresas privadas como a institutos gubernamentales. Los procedimientos de calidad están implantados en todos los laboratorios, pero la naturaleza del sistema de calidad varía mucho entre ellos. En particular, hay muchos laboratorios de ADN forense que operan sin un sistema de notificación de incidencias de calidad como el descrito en este artículo. Desde nuestra experiencia, ese sistema es extremadamente importante para la detección y la gestión apropiada de los errores. Esto es crucial en casos forenses que puedan tener un mayor impacto en la vida de las personas. Por consiguiente, proponemos que la implementación del sistema de notificación de incidencias de calidad es necesario para cualquier laboratorio que esté involucrado en resolver casos de ADN forenses.
>
> Un sistema así solo puede trabajar en óptimas condiciones, cuando exista una cultura en el laboratorio no centrada en la culpabilización y que se extienda a la policía y al sistema judicial. La gente tiene tendencia natural a ocultar sus errores y es esencial crear una atmósfera donde no haya consecuencias personales adversas cuando se cometan errores en los informes. La dirección debe tomar la iniciativa en este cambio cultural…
>
> Hasta lo que sabemos, el NFI es el primer laboratorio forense de ADN del mundo en revelar detalladamente tales datos e informes. Demuestra que eso es posible sin que ocurra ningún desastre o abuso, y no existen razones para la no divulgación. Como se mencionó en la introducción, la pub-

---

[196] FBI, 2011. Véase: www.fbi.gov/aboutus/lab/biometric-analysis/codis/qas-standards-for-forensic-dna-testing-laboratories-effective-9-1-2011.

[197] *Ibid.*, Secciones 12, 13, y 14.

[198] Véase, por ejemplo: Plebani y Carroro, 1997: 1348-1351; Stahl, Lund y Brandslund, 1998: 2195-2197; Hofgartner y Tait, 1999: 14-21; Carroro y Plebani, 2007: 1338-1342.

[199] Kloosterman, Sjerps y Quak, 2014: 77-85 y Butler 2015. www.cstl.nist.gov/strbase/pub_pres/Butler-ErrorManagement-DNA-Error.pdf.

[200] Holanda tiene un sistema de justicia penal «inquisitorial» no un sistema adversarial como el utilizado en los Estados Unidos. La preocupación por tener que explicar cuestiones de calidad ante los Tribunales puede explicar en parte por qué los laboratorios en los Estados Unidos no informan rutinariamente sobre ellas.

licación de datos sobre tasas de error en el uso de medicamentos de laboratorio se ha convertido en una práctica estándar. Las tasas de fallos de calidad en este dominio son comparables a las nuestras.

Finalmente, resaltamos que es necesario mejorar las pruebas de aptitud. No hay actualmente requerimientos sobre el nivel de dificultad que deben tener estos tests. Las pruebas deben ser representativas del rango completo de situaciones que es probable encontrar en los casos reales.

HALLAZGO 2:
ANÁLISIS DE ADN

Validez de los fundamentos. El PCAST encuentra que el análisis de muestras procedentes de una única fuente o de mezclas simples de dos individuos, tales como las que dan en casos de violación, es un método objetivo que ha sido establecido como válido en sus fundamentos.

Validez en la aplicación. Como los errores por fallos humanos son dominantes en la probabilidad de coincidencias aleatorias, los criterios científicos para la validez en la aplicación requieren que un analista (1) deba haber realizado una prueba de aptitud rigurosa y relevante para demostrar su capacidad para aplicar fiablemente el método, (2) deba revelar rutinariamente en informes y testimonios si, cuando llevó a cabo el examen, era consciente de algunos hechos del caso que quizá influyeran en la conclusión y (3) deba revelar, previa solicitud, toda la información sobre las pruebas de calidad y las incidencias de calidad en su laboratorio.

5.2. Análisis de ADN en muestras con mezclas complejas

Algunas investigaciones conllevan análisis de ADN de mezclas complejas de muestras biológicas procedentes de múltiples individuos desconocidos en proporciones desconocidas. Tales muestras quizá se produzcan, por ejemplo, en manchas de sangre mezcladas. A medida que los kits de pruebas de ADN han adquirido mayor sensibilidad, ha habido un creciente interés en el «ADN por contacto» —por ejemplo, pequeñas cantidades de ADN dejadas por múltiples individuos en el volante de un coche—.

5.2.1. Metodología

La diferencia fundamental entre el análisis de ADN de muestras con mezclas complejas y el análisis de ADN de una única fuente o de una mezcla simple no está en el proceso de laboratorio sino en la interpretación del perfil de ADN resultante.

El análisis de ADN de mezclas complejas —definido como mezclas con más de dos contribuyentes— es inherentemente difícil e incluso más para pequeñas cantida-

des de ADN [201]. Esas muestras producen un perfil de ADN que superpone múltiples perfiles individuales. La interpretación de un perfil mezcla es diferente por múltiples motivos: cada individuo puede contribuir con dos, uno o cero alelos en cada marcador; los alelos pueden superponerse entre sí; las alturas de los picos pueden diferir considerablemente, debido a las diferencias en la cantidad y en el estado de preservación del ADN de cada fuente; y los «pequeños picos» que rodean a los alelos (comunes artefactos del proceso de amplificación del ADN) pueden dificultar la detección de los alelos presentes o sugerir alelos que no están [202]. Es frecuentemente imposible decir con certeza qué alelos están presentes en la mezcla o cómo distintos individuos contribuyen a la mezcla, por no hablar de la precisión para inferir el perfil de ADN de cada individuo [203].

En cambio, los analistas se deben preguntar: «¿podría el perfil de ADN del sospechoso estar presente dentro del perfil mezcla? Y, ¿cuál es la probabilidad de que esa observación pueda ocurrir por casualidad?». Esas preguntas son difíciles por las razones anteriormente dadas. Como consecuencia de que muchos perfiles genéticos diferentes pueden encajar dentro de algunos perfiles mezcla, la probabilidad de que un sospechoso «no pueda ser excluido» como posible contribuyente a una mezcla compleja puede ser *mucho más alta* (en algunos casos, millones de veces más alta) que las probabilidades halladas en coincidencias de perfiles procedentes de una única fuente. Como resultado, el cálculo apropiado del peso estadístico es crítico a la hora de presentar información precisa al tribunal.

5.2.2. Interpretación subjetiva de mezclas complejas

Los planteamientos iniciales para interpretar mezclas complejas se apoyan en juicios subjetivos de los analistas, así como en el uso de métodos estadísticos simplificados como la *«Combined Probability of Inclusion»* (CPI) (Probabilidad Combinada de Inclusión). Estos planteamientos son problemáticos porque las elecciones subjetivas hechas por los analistas, tales como qué alelos se incluyen en el cálculo, pueden alterar sustancialmente el resultado y conducir a respuestas imprecisas.

El problema existente con el análisis subjetivo de mezclas complejas se ilustra bien con el caso de doble homicidio *Winston v. Commonwealth* en 2003 [204]. Un perito de la acusación informó que el acusado no podía ser excluido como posible contribuyente al ADN encontrado en un guante desechado que contenía perfiles de ADN mezclados de al menos tres contribuyentes; el acusado fue condenado y sentenciado

---

[201] Véase, por ejemplo, el documento SWGDAM sobre interpretación de mezclas de ADN. www.swgdam.org/#!public-comments/c1t82.

[202] Los retos en el análisis de ADN «low-template» (con baja cantidad de ADN molde) se describen en este reciente artículo, Butler, 2015.

[203] Véanse: Buckleton, Curran y Gill, 2007: 20-28; Coble, Bright, Buckleton, et al., 2015: 207-11.

[204] *Winston v. Commonwealth*, 604 S.E.2d 21 (Va. 2004).

a muerte. El fiscal dijo al jurado que la probabilidad de que sucediera el cotejo positivo por azar era de de 1 entre 1.1 billones (anglosajones). Un artículo publicado en el año 2009, sin embargo, muestra un caso científicamente razonable en el que la probabilidad es próxima a 1 entre 2 —es decir, el 50% de la población relevante no podía ser excluido—[205]. Esa enorme discrepancia no es aceptable, especialmente en casos en los que el acusado fue sentenciado a muerte.

Hay dos artículos que claramente demuestran que las aproximaciones utilizadas generalmente en el análisis de ADN de mezclas complejas pueden ser problemáticas. En un estudio de 2011, Dror y Hampikian probaron si información de contexto irrelevante pudiera sesgar las conclusiones de los analistas, para ello utilizaron una prueba de ADN de un caso penal ya juzgado (un caso de violación en grupo en Georgia)[206]. En este caso, uno de los sospechosos implicó a otro mediante un acuerdo de culpabilidad. Los dos analistas que examinaron la prueba de la escena del crimen eran conscientes de este testimonio contra el sospechoso y sabían que el testimonio basado en el acuerdo de culpabilidad podía ser utilizado ante los tribunales sólo si la evidencia de ADN lo corroboraba. Debido a la compleja naturaleza de la mezcla de ADN recogida en la escena del crimen, el análisis de esta prueba requirió juicio e interpretación por parte de los analistas. Ambos analistas concluyeron que el sospechoso no podía ser excluido como contribuyente.

Dror y Hampikian presentaron la prueba de ADN original de este delito a 17 analistas en ADN, pero sin facilitarles ninguna información de contexto irrelevante. Encontraron que sólo 1 de los 17 analistas estuvo de acuerdo con los analistas originales que fueron expuestos a la información sesgadora (de hecho, 12 de los analistas *excluyeron* al sospechoso como posible contribuyente).

En otro artículo, De Keijser y sus colegas, se presentó a 19 analistas en ADN un caso simulado de un supuesto robo con violencia fuera de un bar:

> Hay un sospechoso varón que niega haber hecho nada malo. Las piezas de convicción que fueron muestreadas para el análisis de ADN son una camisa de una (supuesta) víctima que es mujer (que sostiene haber sido sujetada por su asaltante), una colilla recogida por la policía y cuyo cigarrillo fue supuestamente fumado por la víctima /o el sospechoso y pedazos de uñas de la víctima que dice haber arañado al agresor[207].

Aunque a todos los analistas se les proporcionaron los mismos perfiles de ADN (preparados a partir de las tres muestras y las dos personas mencionadas), sus conclusiones variaron radicalmente. Un analista excluyó al sospechoso como posible contribuyente, mientras que otro declaró una coincidencia entre el perfil del sospechoso y algunos picos menores en el perfil mezcla de las uñas —informando de una probabilidad de coincidencia aleatoria de aproximadamente 1 en 209 millones—. Otros analistas incluso declararon la prueba como no concluyente.

---

[205] Thompson, 2009: 257-276.
[206] Dror y Hampikian, 2011: 204-208.
[207] de Keijser, Malsch, Luining, et al., 2016: 71-82.

En el verano de 2015, una notable cadena de sucesos en Texas reveló que los problemas con los análisis subjetivos de mezclas complejas de ADN no se limitaban a unos pocos casos: eran sistemáticos[208]. El Departamento de Seguridad Pública de Texas (TX-DPS) emitió una declaración pública el 30 de junio de 2015 a la comunidad de justicia penal de Texas observando que (1) el FBI había informado recientemente de que había identificado y corregido errores menores en sus bases de datos poblacionales utilizadas para cálculos estadísticos de casos de ADN; (2) se espera que los errores no tengan un efecto significativo en los resultados; y, (3) que el Sistema de Laboratorios de Criminalística de Texas, previa solicitud, recalcularía estadísticas previamente notificadas en casos individuales.

Cuando algunos fiscales presentaron solicitudes de recálculo a TX-DPS y otros laboratorios, se quedaron sorprendidos al descubrir que algunos resultados estadísticos habían cambiado considerablemente —*por ejemplo, de 1 en 1.4 billones (anglosajones) a 1 en 36, en un caso; de 1 en 4000 a no concluyente, en otro*—. Estos fiscales buscaron la asistencia de la Comisión de Ciencia Forense de Texas (TFSC) para entender las razones del cambio y el alcance de los casos potencialmente afectados.

Tras consulta con analistas en ADN forense, el TFSC determinó que los grandes cambios observados en algunos casos no estaban relacionados con las correcciones menores en la base de datos poblacional del FBI, sino que más bien eran debidos al hecho de que los laboratorios forenses habían cambiado la forma en que calculaban la probabilidad combinada de inclusión (CPI estadística) —especialmente en cómo habían tratado fenómenos como la falta de alelos (drop-out) en determinados marcadores de ADN—.

El TFSC puso en marcha un *DNA Mixture Notification Subcommittee* (Subcomité estatal de Notificación de Mezclas de perfiles de ADN), que incluyó representantes de unidades de integridad de condenas, fiscales de distrito y condado, abogados defensores, representantes de ciertos Proyectos Inocencia, el fiscal general del estado y el gobernador de Texas. Por septiembre de 2015, el TX-DSP había generado una lista, condado por condado, de más de 24.000 casos de mezclas de perfiles de ADN analizados entre 1999 y 2015. Debido a que TX-DPS es responsable de aproximadamente la mitad de los casos en el estado, el número total de casos de ADN de Texas que requieren revisión puede exceder de los 50.000 (Aunque no se han realizado esfuerzos comparables en otros estados, es probable que el problema tenga alcance nacional y no serlo específicamente solo en los laboratorios forenses en Texas).

TFSC convocó también un panel internacional de analistas científicos —de la *Harvard Medical School* (facultad de medicina de Harvard), el *University of North Texas Health Science Center* (Centro de ciencias de la salud de la Universidad del Norte de Texas), la *New Zeland's Forensic Research Unit* (Unidad de Investigación

[208] Pueden encontrarse documentos relevantes y más detalles en: www.fsc.texas.gov/texas-dna-mixture-interpretation-casereview. Lynn Garcia, Consejera General de la Comisión de Ciencia Forense de Texas, también proporcionó un resumen útil a PCAST.

Forense de Nueva Zelanda) y el NIST— para clarificar el uso apropiado de la CPI. Estos científicos presentaron observaciones en un congreso público, en el que muchos abogados aprendieron por primera vez hasta qué punto los análisis de mezclas de ADN conllevaba interpretaciones subjetivas. Salieron a colación muchos de los problemas de la CPI, porque las directrices existentes no especificaron de manera clara, adecuada o correcta el uso apropiado o las limitaciones de esa aproximación estadística.

En resumen, la interpretación de las mezclas complejas de perfiles de ADN con la CPI estadística ha resultado ser un método inadecuadamente especificado —y de este modo, inapropiadamente subjetivo—. Como tal, el método carece claramente de validez en sus fundamentos.

En un intento de colmar este vacío, los analistas convocados por TFSC escribieron un documento científico conjunto, publicado online el 31 de agosto de 2016[209]. El documento subraya la «necesidad perentoria … de una estandarización del procedimiento, formación y de pruebas (de aptitud) continuas de los analistas de ADN». Los autores proponen un conjunto de normas específicas para el uso de la CPI estadística.

Las normas propuestas son claramente *necesarias* para un método científicamente válido que aplique la CPI. Como el artículo vio luz cuando este informe se estaba terminando de redactar, el PCAST no ha dispuesto del tiempo adecuado para valorar si las normas son también suficientes para definir un método válido, objetiva y científicamente, para la aplicación de la CPI.

5.2.3. Esfuerzos actuales para desarrollar métodos objetivos

Dada la existencia de esos problemas, algunos grupos han realizado esfuerzos para desarrollar programas computerizados de «genotipado probabilístico» que aplican diversos algoritmos para interpretar las mezclas complejas. Hasta marzo de 2014, se habían desarrollado al menos 8 programas de genotipado probabilístico (con el nombre de LRmix, Lab Retriever, likeLTD, FST, Armed Xpert, TrueAllele y DNA View Mixture Solution), siendo algunos de ellos de código abierto y otros productos comerciales[210]. El Laboratorio del FBI empezó a utilizar el software STRmix hace menos de un año, diciembre de 2015, y aún sigue en proceso de publicación su propia validación interna de desarrollo.

Estos programas informáticos de software genotipado probabilístico representan claramente una mejora importante sobre la interpretación subjetiva pura. Sin embargo, aún se requiere un cuidadoso escrutinio para determinar (1) si los métodos son

---

[209] Bieber, Buckleton, Budowle, et al., «Evaluation of forensic DNA mixture evidence: protocol for evaluation, interpretation, and statistical calculations using the combined probability of inclusion». *BMC Genetics*. www.bmcgenet.biomedcentral.com/articles/10.1186/s12863-016-0429-7.

[210] El tema se trata en Butler, 2015: 333-48.

científicamente válidos, incluyendo la definición de las limitaciones de su fiabilidad (es decir, las circunstancias en las cuales puedan ofrecer resultados poco fiables) y (2) si el software implementa correctamente los métodos. Esto es particularmente importante porque los programas emplean diferentes algoritmos matemáticos y pueden dar diferentes resultados para el mismo perfil mezcla[211].

Cualquier evaluación apropiada de los métodos propuestos ha de consistir en estudios realizados por múltiples grupos, *no asociados con los desarrolladores del software*, que investiguen el funcionamiento y definan las limitaciones de los programas, probándolos sobre un amplio rango de mezclas con diferentes propiedades. En particular, es importante abordar los siguientes problemas:

(1) ¿Cuán bien funciona el método en función del número de contribuyentes a la mezcla? ¿Qué nivel de funcionamiento tiene el método cuando el número de contribuyentes a la mezcla es *desconocido*?

(2) ¿Cuán bien funciona el método en función del número de alelos compartidos entre individuos en la mezcla? Y con relación a esto, ¿cómo funciona cuando las mezclas incluyen individuos emparentados?

(3) ¿Cuán bien funciona el método —y cómo se degrada la precisión— en función de cantidades absolutas y relativas de ADN procedente de varios contribuyentes? Por ejemplo, puede resultar difícil determinar si un pequeño pico en el perfil de la mezcla representa un verdadero alelo procedente de un contribuyente minoritario o tratarse de un pico secundario de un alelo procedente de otro contribuyente diferente. (En particular, esta cuestión está en el fondo de un caso actual que ha recibido una atención considerable[212]).

(4) ¿En qué circunstancias —y por qué— el método produce resultados (probabilidades de inclusión aleatorias) que difieren sustancialmente de las producidas por otros métodos?

---

[211] Algunos programas utilizan métodos discretos (semicontinuos), que usan solo información alélica junto con probabilidades de pérdida y ganancia de alelos, mientras que otros programas utilizan métodos continuos, que también incorporan información sobre altura de picos y de otro tipo. Dentro de esas dos clases, los programas difieren con respecto a cómo utilizan la información. Algunos de los métodos hacen asunciones sobre el número de individuos que contribuyen a un perfil de ADN y utilizan esa información para eliminar ruido (como los picos «stutter» en un perfil de ADN).

[212] En este caso, los analistas utilizaron dos programas informáticos distintos (STRMix y TrueAllele) y obtuvieron conclusiones diferentes sobre si el ADN del acusado pudiera estar incluido dentro de un perfil mezcla de baja cantidad de ADN obtenido de una muestra recogida en una de las uñas de la víctima. El juez dictaminó que la muestra de ADN que implicaba al acusado era inadmisible. McKinley, 2016. Véase: www.nytimes.com/2016/07/25/nyregion/potsdam-boys-murder-case-mayhinge-on-statistical-analysis.html (última consulta el 22 de agosto de 2016). Sommerstein, «DNA results will not be allowed in Hillary murder trail.» North Country Public Radio (última consulta el 1 de septiembre de 2016). La decisión puede encontrarse aquí: www.northcountrypublicradio.org/assets/files/08-26-16DecisionandOrder-DNAAnalysisAdmissibility.pdf

Se han publicado una serie de artículos dirigidos a abordar algunas de estas cuestiones que analizan mezclas conocidas[213]. Hay que tener en cuenta dos cosas sobre estos estudios. La primera cuestión es que la mayoría de los estudios que evalúan paquetes de software han sido realizados por los propios desarrolladores del software. Mientras que resulta plenamente adecuado que los propios desarrolladores de un método lo evalúen, el establecimiento de la validez científica también requiere que otros grupos de científicos que no desarrollaron el método lo hagan. Y la segunda cuestión es que ha habido pocos estudios comparativos entre los métodos para evaluar las diferencias entre ellos —y hasta donde llega nuestro conocimiento, no se han realizado estudios comparativos dirigidos por grupos independientes—[214].

Y lo más importante, los actuales estudios han explorado adecuadamente sólo un rango limitado de tipos de mezclas (con respecto al número de contribuyentes, la proporción de contribuyentes minoritarios, y la cantidad total de ADN). Los dos métodos más usados (STRMix y TrueAllele) aparecen como fiables dentro de un cierto rango, fundamentándose en la evidencia disponible y en la dificultad inherente al problema[215]. Específicamente, estos métodos aparecen como fiables en mezclas de tres personas, en las cuales, el contribuyente minoritario constituye al menos el 20% del ADN intacto en la mezcla y en el cual, la cantidad de ADN excede el mínimo nivel requerido por el método[216].

Para mezclas más complejas (por ejemplo, mayor número de contribuyentes o proporciones menores), hay relativamente poca evidencia publicada[217]. En genética

---

[213] Por ejemplo: Perlin, Hornyak, Sugimoto y Miller, 2015: 857-868; Greenspoon, Schiermeier-Wood, et al., 2015:1263–1276; Bright, Taylor, McGovern, et al., 2016: 226-239; Bright, Taylor, Curran, et. al., 2014:102-110; Taylor, Buckleton y Evett, 2015: 165-171; Taylor y Buckleton, 2015: 13-16.

[214] Bille, Weitz, Coble, et.al., 2014: 3125–3133.

[215] La interpretación de las mezclas de ADN incrementa su dificultad a medida que crece el número de contribuyentes. Véanse, por ejemplo: Taylor, Buckleton, Evett, 2015: 165-171; Bright, Taylor, McGovern, et al., 2016: 226-39; Bright, Taylor, Curran y Buckleton, 2014:102-110; Bieber, Buckleton, Budowle, et. al., «Evaluation of forensic DNA mixture evidence: protocol for evaluation, interpretation, and statistical calculations using the combined probability of inclusion». *BMC Genetics*. bmcgenet.biomedcentral.com/articles/10.1186/s12863-016-0429-7.

[216] Las muestras mezcladas de tres personas que tienen proporciones similares son más fáciles de interpretar debido al número limitado de alelos y a la relativa similitud de las alturas de los picos. Los métodos pueden ser aplicados fiablemente a muestras procedentes de una única fuente y a mezclas simples, con tal que, en los casos en que dos contribuyentes no puedan separarse por extracción diferencial, la proporción del contribuyente minoritario no sea demasiado baja (por ejemplo, al menos un 10%).

[217] Para mezclas de cuatro personas, por ejemplo, hay artículos que describen validaciones experimentales con mezclas conocidas utilizando TrueAllele que incluyen 7 y 17 mezclas diferentes, respectivamente, con relativamente grandes cantidades de ADN (al menos 200 picogramos), mientras que aquellos artículos en los que se usa STRMix incluyen 2 y 3 mezclas diferentes, respectivamente, aunque usan mucha menos cantidad de ADN (en el rango de 10 picogramos). Greenspoon, Schiermeier-Wood y Jenkins, 2015:1263–1276; Perlin, Hornyak, Sugimoto y Miller, 2015: 857-868; Taylor, 2014: 144-153; Taylor., Buckleton y Evett, 2015: 165-171; Taylor y Buckleton, 2015: 13-16; Bright, Taylor, McGovern, Cooper, Russell, Abarno y Buckleton, 2016: 226-239.

molecular humana, la validación experimental de un método diagnóstico importante necesitaría generalmente cientos de muestras distintas[218]. Un científico forense dijo al PCAST que han sido analizadas, de hecho, muchas más muestras diferentes, pero que los datos no se han recopilado y publicado[219]. Como consecuencia de que la evidencia empírica es esencial para establecer la validez de los fundamentos de un método, el PCAST urge a los científicos forenses a que presenten y publiquen en revistas científicas estudios de validación de alta calidad que propiamente establezcan el rango de fiabilidad de los métodos para el análisis de mezclas de ADN complejas.

Una vez que se publiquen más estudios, es probable que sea posible ampliar el rango en el que la validez científica ha sido establecida para incluir muestras más desafiantes. Como se señaló anteriormente, dichos estudios deben realizarse por grupos de investigación independientes que no estén relacionados con los desarrolladores de los métodos y sin intereses por el resultado.

5.2.4. Conclusión

El PCAST, fundamentándose en su evaluación de la literatura publicada hasta la fecha, llegó a una serie de conclusiones sobre la validez de los fundamentos de los métodos para el análisis de mezclas de ADN complejas. Resaltamos que la validez de los fundamentos debe establecerse con respecto a un método específico que se aplica en un determinado rango. Además de formar su propio juicio, el PCAST también realizó una consulta al Sr. Jonn Butler, asistente especial del Director de Ciencias Forenses en el NIST y Vicepresidente del NCFS. Butler estuvo de acuerdo con el resultado de la investigación del PCAST.

HALLAZGO 3:
ANÁLISIS DE ADN DE MUESTRAS DE MEZCLAS COMPLEJAS

Validez de los fundamentos. El PCAST ha averiguado que:

(1) Métodos basados en la Probabilidad Combinada de Inclusión (CPI). El análisis de ADN de mezclas complejas basado en aproximaciones fundamentadas en la CPI ha sido un método inadecuadamente especificado y subjetivo que tiene potencial para conducir a resultados erróneos. Como tal, no es válido en sus fundamentos.

Un artículo reciente ha propuesto normas específicas que abordan una serie de problemas con el uso de la CPI. Esas normas son claramente *necesarias*. Sin embargo, el PCAST no ha tenido tiempo suficiente para valorar si son también suficientes para definir un método científicamente válido y objetivo. Si, por un tiempo limitado, los tribunales deciden admitir resultados basados en la aplicación de la CPI, la validez de su

[218] La preparación y la realización de la amplificación PCR de cientos de mezclas de ADN es sencilla; puede conseguirse en unas pocas semanas o en menos tiempo.

[219] Entrevista del PCAST a John Buckleton, científico principal del Instituto de Ciencia Medioambiental y de Investigación en Nueva Zelanda y codesarrollador de STRMix.

aplicación requeriría que, como mínimo, fuera consistente con las normas especificadas en el artículo.

El análisis de ADN de mezclas complejas debe dirigirse sin demora hacia métodos más apropiados basados en genotipado probabilístico.

(2) Genotipado probabilístico. El análisis objetivo de mezclas complejas de ADN con software de genotipado probabilístico es un enfoque relativamente nuevo y prometedor. Se requiere evidencia empírica para establecer la validez de los fundamentos de cada método dentro de rangos específicos. Hasta el momento, hay evidencia publicada que apoya la validez de los fundamentos del análisis, con algunos programas, de mezclas de ADN de tres individuos, en las que el menor contribuyente constituye, al menos, el 20% del ADN intacto en la mezcla y en las que la cantidad de ADN excede del mínimo requerido por el método. El rango en el que ha quedado establecida la validez de los fundamentos es probable que crezca en la medida en que se obtenga evidencia adecuada para mezclas más complejas y se publique.

Validez en la aplicación. Para métodos que son válidos en sus fundamentos, la validez en la aplicación conlleva consideraciones similares a las realizadas para análisis de ADN procedente de una única fuente y para mezclas simples, con un énfasis especial en asegurar que el método se aplique correctamente y dentro de su rango empíricamente establecido.

5.2.5. El camino a seguir

Existe un camino claro para ampliar el rango de métodos objetivos establecidos como válidos en sus fundamentos —específicamente, a través de la publicación de los estudios científicos pertinentes—.

Esos esfuerzos se verán favorecidos por la creación y difusión (bajo restricciones relacionadas con el uso apropiado de los datos y su privacidad) de grandes colecciones de cientos de perfiles de ADN a partir de mezclas conocidas —que representen la amplia variedad de la complejidad con respecto (1) al número de contribuyentes, (2) a las relaciones entre los contribuyentes, (3) las cantidades absolutas y relativas de materiales y (4) al estado de conservación de los materiales —que pueden utilizarse por grupos independientes para evaluar y comparar los métodos—. De forma destacada, el denominado *PROVEDIt Initiative (Project Research Openness for Validation with Experimental Data)* (Proyecto de tranparencia de la investigación para la validación con datos experimentales) de la Universidad de Boston ha logrado que pueda disponerse de 25.000 perfiles de mezclas de ADN para la investigación[220],[221].

[220] Véase: www.bu.edu/dnamixtures.

[221] La colección contiene muestras con mezclas de ADN de 1 a 5 personas, amplificadas con ADN molde (*targets ranging*) en el rango de 1 a 0.007 nanogramos. En las mezclas multi-personas, la relación de los contribuyentes varía de 1:1 a 1:19. Adicionalmente, los perfiles se generaron utilizando una variedad de condiciones de laboratorio a partir de muestras que contenían ADN prístino; ADN dañado por rayos UV; ADN degradado enzimática o sónicamente; y ADN con inhibidores.

Además de estudios científicos sobre conjuntos comunes de muestras con el propósito de evaluar la validez de los fundamentos, los laboratorios forenses querrán conducir sus propios estudios internos de validación del desarrollo de modo individual, para valorar la validez del método por sí mismos[222].

El NIST debe ser lider en este proceso, asegurando la creación y diseminación de los materiales y estimulando estudios por grupos independientes mediante subvenciones, contratos y premios; y evaluando los resultados de esos estudios.

5.3. Análisis de marcas mordeduras

5.3.1. Metodología

El análisis de marcas de mordedura es un método subjetivo. Generalmente conlleva examinar marcas dejadas sobre la víctima o un objeto en la escena del crimen y comparar esas marcas con impresiones dentales tomadas al sospechoso[223]. La comparación de marcas de mordedura se fundamenta en las premisas: (1) las características dentales, particularmente la disposición de los dientes frontales, difieren sustancialmente entre las personas y (2) la piel (o alguna otra superficie marcada en la escena del crimen) puede capturar fiablemente esas características distintivas.

El análisis de marcas de mordedura comienza con la decisión del analista sobre si una herida es una marca causada por dientes humanos[224]. Si es así, el analista toma fotografías o impresiones de la mordedura en cuestión y de la dentición del sospechoso, compara la marca de mordedura con la dentición y determina si la dentición: (1) no puede excluirse como la que ha dejado la marca, (2) cabe excluirla de haber dejado la marca o (3) es no concluyente. Los estándares de marcas de mordedura no están bien definidos respecto el grado de similitud que debe identificarse para apoyar una conclusión fiable sobre si la marca pudo producirse o no por la dentición en cuestión. Las conclusiones sobre estos asuntos se dejan a juicio del analista.

5.3.2- Estudios sobre cuestiones de fondo

Antes de adentrarnos en la cuestión de la validez de los fundamentos, abordamos algunos estudios sobre cuestiones de fondo (concernientes a temas como la unicidad y la consistencia) que arrojan algo de luz sobre el asunto. Estos estudios ponen en serias dudas las premisas de esta disciplina.

---

[222] El Laboratorio del FBI ha completado recientemente un estudio de validación de desarrollo y lo está preparando para su publicación.

[223] Menos frecuentemente se encuentran marcas sobre el sospechoso que pueden proceder de la víctima.

[224] AFBO Bitemark Methodology Standards and Guidelines (Estándares y Guías de la Metología de Marcas de Mordedura de AFBO, el Consejo Americano de Odontología Forense), abfo.org/wp-content/uploads/2016/03/ABFO-BitemarkStandards-03162016.pdf (última consulta el 6 de julio de 2016).

Un artículo ampliamente citado de 1984 afirmaba que «la dentición humana era única más allá de toda duda razonable»[225]. El estudio examinó cuidadosamente 397 marcas de mordedura en una oblea de cera, midió 12 parámetros en cada una, y asumiendo, sin evidencia alguna, que los parámetros no estaban correlacionados —sugirió que la probabilidad de que dos marcas de mordedura tuvieran los mismos parámetros era menor a 1 entre 6 trillones (anglosajones)—. El artículo fue teórico en vez de empírico: no intentó realmente comparar las marcas de mordedura entre sí.

Un artículo en 2010 desacreditó esas afirmaciones[226]. Tras estudiar 344 moldes dentales humanos y medirlos con un escáner láser tridimensional, estos autores demostraron que las coincidencias ocurrían mucho más frecuentemente que lo esperado por el modelo teórico. Por ejemplo, el modelo teórico predijo que la probabilidad de encontrar *incluso una simple coincidencia entre cinco dientes* en la colección de marcas de mordedura era menor a 1 entre 1 millón; aunque la comparación empírica reveló que se encontraron 32.

En particular, estos estudios examinaron patrones de dentición humana medidos en condiciones idealizadas. Por contra, se ha demostrado que la piel es un medio poco fiable para registrar patrones precisos de los dientes. Los estudios de marcas de mordedura realizados en cerdos vivos[227] (utilizados como un modelo de piel humana), o en cadáveres humanos[228], han demostrado una significativa distorsión en todas las direcciones. Un estudio de 2010 sobre marcas de mordedura creadas experimentalmente y producidas por personas identificadas concluyó que la deformación de la piel distorsiona las marcas de mordedura de forma tan sustancial y con tal variabilidad que los actuales procedimientos para comparar marcas de mordedura son incapaces de excluir o incluir fiablemente a un sospechoso como potencial mordedor («Los datos derivados no mostraban correlación alguna y no eran reproducibles, es decir, la misma dentición podría no producir una impresión medible que fuera consistente en todos los parámetros en cualquiera de las circunstancias de la prueba»)[229]. Tal distorsión es aún más complicada en el contexto de los casos penales, en los que la mordedura tiene lugar con frecuencia durante forcejeos y en los que la piel puede estirarse y contorsionarse en el momento en que se produce la marca de la mordedura.

La investigación empírica sugiere que los odontólogos forenses no están de acuerdo consistentemente incluso en la determinación de si una herida es una mordedura humana. Un estudio del Consejo Nortemericano de Odontología Forense (AFBO)[230]

---

[225] Rawson, Ommen, Kinard, et al., 1984: 245-53.

[226] Bush, M.A., Bush, P.J. y Sheets, 2011: 118-23.

[227] Dorion, ed., 2011.

[228] Sheets, Bush, P.J. y Bush M.A., 2012: 202-207; Bush, M.A., Miller, Bush, P.J, et al., 2009: 167-76.

[229] Bush M.A., Cooper y Dorion, 2010: 976-983.

[230] Freeman y Pretty, «Construct validity of bitemark assessments using the ABFO decision tree», presentación en el Congreso Anual de la Academia Americana de Ciencias Forenses. Véase: online.wsj.com/public/resources/documents/ConstructValidBMdecisiontreePRETTYFREEMAN.pdf.

supuso la observación de 100 fotos de patrones de lesiones por parte de analistas de marcas de mordedura certificados por la AFBO y se les pidió que contestaran a tres cuestiones básicas sobre (1) la existencia de pruebas suficientes para emitir una opinión sobre si el patrón de la lesión era el de una mordedura humana; (2) si la marca era una mordedura humana, si era posible sugerir esa posibilidad o si no lo era; y (3) si las características distintivas (arcos y marcas dentales) eran identificables[231]. De los 38 analistas que completaron el estudio, se informó que hubo acuerdo unánime en la primera pregunta en solo 4 de los 100 casos y acuerdo en, al menos, el 90% en solo 20 de los 100 casos. En las tres preguntas hubo un acuerdo de al menos el 90% en solo 8 de los 100 casos.

En un estudio similar en Australia, a 15 odontólogos se les mostró una serie de seis marcas de mordedura de casos contemporáneos, cinco de los cuales habían sido reconocidas como producidas por dientes por víctimas sobrevivientes, y se les pidió que explicaran de forma narrativa si las heridas eran realmente marcas de mordedura[232]. El estudio mostró amplia variabilidad entre los analistas en sus conclusiones sobre el origen, circunstancias y características de los patrones de las heridas en las seis imágenes. Sorprendentemente, los más experimentados (21 o más años) tendían a tener el más amplio rango de opiniones sobre si la marca era de origen dental humano o no[233]. Las opiniones de los analistas sobre si pensaban que una marca era apta para su análisis variaron considerablemente y los profesionales demostraron poca consistencia en sus enfoques en el análisis de una marca desde una mordedura a otra. El estudio concluyó que «esta inconsistencia indica un defecto fundamental en la metodología del análisis de las marcas de mordedura y debe conducir a preguntarnos, con preocupación, sobre la fiabilidad de las conclusiones alcanzadas sobre las coincidencias entre una marca de mordedura y una dentición»[234].

5.3.3. Estudios de validez científica y fiabilidad

Como se dijo más arriba, la validez de los fundamentos de un método subjetivo solo puede establecerse mediante múltiples estudios de caja negra independientes.

---

[231] Los datos están a disposición de los interesados previa petición a los autores. Fueron revisados por la Profesora Karen Kafadar, miembro del panel de consejeros *senior* para este estudio.

[232] Page, Taylor y Blenkin, 2013: 664-672.

[233] Por ejemplo, un analista manifestó que con certeza una de las imágenes era de una marca de mordedura, diciendo: «Conozco, desde la experiencia, que es de dientes porque tuve un caso al comienzo del año que cuando miré por primera vez las imágenes, a causa de la severidad de las heridas, no pensé que fueran de dientes. Pero cuando vi los modelos y los observé atentamente, me parecieron que lo eran». Otro expresó dudas de que la misma imagen fuera una marca de mordedura, también basándose en su experiencia: «Honestamente, no creo que sea una marca de mordedura… hay muchas cosas que podrían haber causado eso. Dudo de que haya aquí marcas de dientes individuales. Nunca he visto nada como esto». *Ibid.*, 666.

[234] *Ibid.*, 670.

El informe del NRC (2009) aseveró que la validez científica de los análisis de marcas de mordedura no había sido establecida[235]. El PCAST, en su propia revisión de la literatura, encontró pocos estudios empíricos que intentasen estudiar la validez y fiabilidad de los métodos para identificar los orígenes de una marca de mordedura.

En un artículo de 1975, a dos analistas les pidieron que emparejasen fotografías de marcas de mordedura realizadas por 24 voluntarios en la piel de cerdos recién sacrificados con modelos dentales de esos mismos voluntarios[236]. Las fotografías se tomaron entre 0,1 y 24 horas después de que se produjeran las marcas de mordedura. El rendimiento de los analistas fue pobre, deteriorándose conforme el transcurrir del tiempo. La proporción de fotografías incorrectamente atribuidas fue del 28%, 65% y 84% tras 0, 1 y 24 horas.

En un artículo de 1999, a 29 forenses analistas dentales —y a 80 personas más, incluyendo dentistas en general, estudiantes de odontología y otros participantes legos en la materia— les mostraron impresiones en color de mordeduras humanas de 50 casos judiciales y les pidieron que decidiesen si cada marca de mordedura era de un adulto o de un niño[237]. Las decisiones fueron comparadas con los veredictos de todos los casos. Todos los grupos realizaron su función deficientemente[238].

En un artículo del año 2001, a 32 diplomados certificados por AFBO se les pidió que informaran sobre su certeza respecto si 4 marcas de mordedura específicas procedían de 7 modelos dentales, consistentes en 4 fuentes correctas y 3 de muestras no relacionadas[239, 240]. Tal diseño de «conjunto cerrado» (donde la fuente correcta está presente para cada una de las muestras cuestionadas) es inapropiado para valorar la fiabilidad, porque tenderá a subestimar la tasa de falsos positivos[241]. Incluso con el

---

[235] «Hay una discusión continua sobre el valor y validez científica de la comparación e identificación de las marcas de mordedura». National Research Council, 2009: 151.

[236] Whittaker, 1975: 166–71.

[237] Whittaker, Brickley y Evans, 1998: 11-20.

[238] Los autores preguntaron a los observadores que indicasen hasta qué punto tenían certeza de que la marca de mordedura había sido realizada por un adulto utilizando una escala de 6 puntos. Se calcularon y representaron curvas ROC *(Receiver-Operator Characteristic)* a partir de los datos. El área bajo la curva (AUC) se calculó para cada grupo (donde AUC = 1 representa una clasificación perfecta y AUC = 0.5 es equivalente a una decisión aleatoria). El área bajo la curva (AUC) quedó en el intervalo 0.62-0.69, que representa una ejecución pobre.

[239] Arheart y Pretty, 2001: 104-111.

[240] De las cuatro marcas de mordedura, tres eran de casos penales y una producida por un individuo que deliberadamente mordió un trozo de queso. Los siete modelos dentales procedían de tres acusados en casos penales en los que fueron declarados culpables (presumiblemente fueron los que hicieron las mordeduras), el individuo que mordió el queso y tres individuos no relacionados.

[241] En tests de conjunto cerrado, los analistas realizarán bien su cometido siempre y cuando elijan el modelo dental más parecido. En un diseño de conjunto abierto, en el que puede que ninguno de los modelos sea correcto, la probabilidad de falsos positivos es más alta. El diseño de conjunto abierto se parece a la aplicación en casos reales. Véase la extensa discusión sobre diseños en conjunto cerrado en análisis de armas de fuego (Sección 5.5).

diseño de conjunto cerrado, el 11% de las comparaciones con fuente incorrecta fueron declaradas coincidencias «probables», «posibles», o de «razonable certeza médica».

En otro artículo de 2001, a 10 diplomados con certificado AFBO les dieron 10 tests independientes, cada uno consistente en evidencia de marca de mordedura y dos fuentes posibles. La evidencia se produjo aplicando un modelo dental a cerdos recién sacrificados, confirmando subjetivamente que «fueron registrados suficientes detalles» y fotografiando la marca de la mordedura. La fuente correcta estuvo presente en todos menos en dos de los tests (un diseño en conjunto cerrado mayor). La tasa media de falsos positivos fue del 15.9% —es decir, aproximadamente 1 de cada 6—.

En un artículo del año 2010, a 29 analistas con distintos niveles de entrenamiento (incluidos 9 diplomados con certificado AFBO) se les proporcionaron 18 fotografías de marcas de mordedura y la dentición humana de tres individuos (A, B y C) y se les pidió que decidieran si las marcas de mordedura procedían de A, B o C, o de ninguno de los tres. Las marcas de mordedura habían sido producidas en cerdos vivos, utilizando una máquina de morder con dentición de los individuos A, B y D (del cual no se proporcionó su dentición a los analistas). De las marcas de mordedura producidas por D, los diplomados declararon erróneamente una coincidencia con A, B o C en el 17% de los casos —de nuevo, aproximadamente 1 de cada 6—.

5.3.4. Conclusión

Se han realizado pocos estudios empíricos para estudiar la capacidad de los analistas en identificar con precisión la fuente de una marca de mordedura. Entre esos estudios que se han llevado a cabo, las tasas de falsos positivos observadas fueron tan altas que el método claramente no es científicamente fiable en la actualidad. (Más aún, algunos de esos estudios emplean diseños de conjunto cerrado inapropiados que probablemente subestiman la tasa de falsos positivos).

HALLAZGO 4:
ANÁLISIS DE MARCAS DE MORDEDURA

Validez de los fundamentos. El PCAST encuentra que el análisis de marcas de mordedura no cumple los estándares científicos para la validez de los fundamentos, y está lejos de cumplir tales estándares. Al contrario, la evidencia científica disponible sugiere fuertemente que los analistas no pueden estar de acuerdo consistentemente sobre si una herida es de mordedura humana y no pueden identificar la fuente de una marca de mordedura con precisión razonable.

5.3.5. El camino a seguir

Algunos profesionales han expresado su preocupación por el hecho de que la exclusión de las marcas de mordedura en los tribunales pueda obstaculizar los esfuerzos

para conseguir la condena de los acusados en algunos casos[242]. Si así fuera el caso, la solución correcta, desde una perspectiva científica, no sería la de admitir testimonio pericial basado en métodos inválidos y no fiables, sino más bien en intentar desarrollar métodos fiables desde un punto de vista científico.

Sin embargo, el PCAST considera que las perspectivas de desarrollar el análisis de marcas de mordedura como un método científicamente válido son bajas. Aconsejamos no dedicar recursos significativos a tales esfuerzos.

5.4. Análisis de huellas dactilares latentes

Los análisis de huellas latentes se propusieron primeramente para su uso en identificación criminal en el siglo XIX y han sido utilizados durante más de un siglo. El método fue aclamado durante mucho tiempo como infalible, a pesar de la falta de estudios apropiados para valorar su tasa de error. Como se dijo anteriormente, esta escasez de pruebas empíricas manifestaba una grave debilidad en la cultura científica de la ciencia forense —donde la validez fue asumida en lugar de probada—. Citando anteriores guías que ahora se reconoce que han sido inapropiadas[243], el DOJ recientemente señaló,

> [H]istóricamente, fue práctica común que un analista testificara que cuando la … metodología era correctamente aplicada, siempre produciría la conclusión correcta. De este modo, cualquier error que ocurriera sería un error humano y la tasa de error resultante de la metodología sería cero. Este enfoque fue descrito por el Departamento de Justicia en 1984 en la publicación titulada The Science of Fingerprints (La Ciencia de las Huellas Dactilares), donde se afirma que «de todos los métodos de identificación, sólo el de las huellas dactilares ha resultado ser infalible y factible[244].

En respuesta al informe del NRC (2009), la disciplina del análisis de las huellas latentes ha hecho progresos en el reconocimiento de la necesidad de realizar estudios empíricos para valorar la validez de los fundamentos y medir la fiabilidad. Merece especial reconocimiento el laboratorio del FBI, que ha liderado el camino en la realización tanto de estudios de caja negra, diseñados para medir la fiabilidad, como de «estudios de caja blanca», diseñados para comprender los factores que afectan a las decisiones de los analistas[245]. El PCAST aplaude los esfuerzos del FBI. También existen esfuerzos incipientes para empezar a transformar la disciplina desde un método puramente subjetivo hacia un método objetivo, aunque existe aún un considerable camino que recorrer para lograr ese importante objetivo.

---

[242] La proporción precisa de casos en los que las marcas de mordedura juegan un papel clave no está clara, aunque es claramente pequeña.

[243] Federal Bureau of Investigation, 1984: iv.

[244] Véase: www.justice.gov/olp/file/861906/download.

[245] Véanse: Hicklin, Buscaglia, Roberts, et al., 2011: 385-419; Hicklin. Buscaglia y Roberts, 2013: 106-17; Ulery, Hicklin, Kiebuzinski, et al., 2013: 99-106; Ulery, Hicklin y Buscaglia, 2012; Ulery, Hicklin, Roberts, et al., 2015: 54-61.

5.4.1. Metodología

El análisis de huellas dactilares conlleva generalmente (1) comparar una «huella latente» (una impresión con crestas papilares completas o parciales de un sujeto desconocido) que ha sido revelada u observada sobre un vestigio) con (2) una o más «impresiones conocidas» (huellas dactilares deliberadamente coleccionadas en un entorno controlado de sujetos conocidos; también referidas como «decadactilares»), para valorar si las dos pueden haberse originado a partir de la misma fuente. (También puede conllevar la comparación de huellas latentes entre sí).

Es importante distinguir las huellas latentes de las huellas conocidas. Una impresión de huellas conocidas contiene imágenes de huellas dactilares hasta un total de diez, que se toman en un entorno controlado, como en una detención o en una verificación de antecedentes[246]. Como las impresiones conocidas tienden a ser de alta calidad, pueden buscarse automática y fiablemente en grandes bases de datos. Por contra, las huellas latentes reveladas en casos penales a menudo son incompletas y de calidad variable (manchadas o distorsionadas de otro modo), con calidad y claridad dependiente de factores como la superficie tocada y la mecánica del tacto.

Puede requerirse un analista para (1) comparar una huella latente con las huellas de una impresión decadactilar de un sospechoso conocido que ha sido identificado por otros medios («sospechoso identificado») o para (2) buscar en una gran base de datos de huellas dactilares para identificar a un sospechoso («búsqueda en base de datos»).

Los analistas siguen ordinariamente un proceso denominado «ACE» o «ACE-V», para el Análisis, Comparación, Evaluación y Verificación[247, 248]. El proceso pide a los analistas que realicen una serie de valoraciones subjetivas. Un analista utiliza juicio subjetivo para seleccionar regiones particulares de una huella latente para su análi-

---

[246] Véase: Committee on Science, Subcommittee on Forensic Science of the National Science and Technology Council, 2014. www.whitehouse.gov/sites/default/files/microsites/ostp/NSTC/afis_10-20-2014_draftforcomment.pdf.

[247] «Un examen de una huella latente utilizando el proceso ACE-V consiste en lo siguiente: *Análisis* se refiere a una fase de recogida de información inicial en la que el analista estudia la impresión desconocida para valorar la calidad y cantidad de detalles discriminantes presentes. El analista considera aspectos como el sustrato, el método de desarrollo, diversos niveles de detalle de las crestas y distorsiones debida a la presión. Luego realiza un análisis separado de la huella impresa de autoría conocida. La *comparación* consiste en la observación simultánea del detalle de las crestas de fricción en las dos impresiones para determinar el acuerdo o desacuerdo en los detalles. En la fase de *Evaluación*, el analista valora el acuerdo o desacuerdo de la información observada durante el Análisis y la Comparación y redacta una conclusión. La *Verificación* en algunas unidades policiales o instituciones forenses es una revisión de las conclusiones del analista con conocimiento de esas conclusiones; en otras entidades es un nuevo examen independiente realizado por un segundo analista que no conoce el resultado del primer examen». National Institute of Standards and Technology, 2012. Disponible en: www.nist.gov/oles/upload/latent.pdf.

[248] Reznicek, Ruth y Schilens, 2010: 87-103.

sis. Si no hay personas identificadas en las que se pueda estar interesado, el analista someterá la huella latente a una búsqueda en el *Automated Fingerprint Identification System (AFIS)* (Sistema Automático de Identificación Dactilar)[249], que contiene grandes números de huellas conocidas, que utiliza algoritmos de reconocimiento de imágenes patentadas (no públicas)[250] para generar una lista de potenciales candidatos que comparten características dactilares similares[251]. El analista, después, compara manualmente la huella latente con las huellas dactilares de la persona específica de interés o con las huellas de los candidatos con mayores coincidencias generados por el ordenador, mediante el estudio de las características seleccionadas[252], y luego toma una decisión de forma subjetiva sobre si son lo suficientemente similares como para hacer una propuesta de identificación.

ACE-V añade una etapa de verificación. Con respecto a la verificación, la implementación varía ampliamente[253]. En muchos laboratorios solo se verifican las identificaciones, porque se considera demasiado gravoso, en términos de tiempo y coste, realizar exámenes independientes en todos los casos (por ejemplo, con exclusiones). Este procedimiento es problemático porque no es ciego: el segundo analista conoce que el primero hizo una propuesta de identificación, ello genera un potencial sesgo de confirmación. Tras el caso de la identificación errónea en el atentado con bombas en los trenes de Madrid (véase más adelante), el Laboratorio del FBI adoptó medidas para llevar a cabo, en ciertos casos, «la aplicación independiente de ACE a una impre-

---

[249] El estado y las jurisdicciones locales comenzaron a comprar sistemas *AFIS* en las décadas de los 70 y 80 a proveedores privados, cada uno con su propio software y algoritmos de búsqueda patentado. En 1999, el FBI puso en marcha el Sistema Automático de Identificación de Huellas Dactilares Integrado *(IAFIS)*, una base de datos nacional que contiene huellas dactilares y antecedentes policiales de más de 70 millones de individuos introducidos por agencias policiales estatales, locales y federales (recientemente reemplazado por el *Next Generation Identification (NGI) System* (Sistema de Identificación de la Próxima Generación). Algunas agencias de justicia penal tienen la capacidad de buscar huellas latentes no solo en su propia base de datos dactilar, sino también en una jerarquía de bases de datos locales, estatales y federales. Sin embargo, todavía no se ha alcanzado la total interoperabilidad del sistema. Véase: Committee on Science, Subcommittee on Forensic Science of the National Science and Technology Council, 2014. www.whitehouse.gov/sites/default/files/microsites/ostp/NSTC/afis_10-20-2014_draftforcomment.pdf.

[250] Los algoritmos utilizados para generar candidatos de potenciales coincidencias están patentados y no están públicamente disponibles.

[251] El Laboratorio del FBI requiere a los analistas que completen y documenten sus análisis de las huellas dactilares latentes antes de revisar cualquier huella conocida o moverse a la fase de comparación y evaluación. Este requisito no es compartido por todos los laboratorios.

[252] Las características dactilares se comparan en tres niveles de detalle: —nivel 1 («flujo de las crestas»), nivel 2 («camino de las crestas») y nivel 3 («características de las crestas» o «formas»). «El flujo de las crestas» se refiere a las clases de patrones compartidos por muchos individuos como formaciones en bucle o espiral; este nivel es solo suficiente para realizar exclusiones, no para declarar identificaciones. «El camino de las crestas» se refiere a los puntos característicos que pueden ser utilizados para declarar identificaciones como las bifurcaciones o los puntos. «Las formas de las crestas» incluyen las terminaciones de las crestas y la localización de poros. Véase: National Institute of Standards and Technology, 2012. Disponible en: www.nist.gov/oles/upload/latent.pdf.

[253] Black, 2012: 80-100.

sión de crestas de fricción por otro analista cualificado, que no conociera la conclusión del primer analista»[254]. En particular, el Laboratorio del FBI utiliza verificación ciega en los casos considerados hasta el presente como los que tienen mayor riesgo de error, tal como cuando una única huella es identificada, excluida o se considera no concluyente[255].

Como se resaltó en la sección 2, las anteriores preocupaciones sobre la fiabilidad de los análisis de huellas latentes[256], se incrementaron sustancialmente tras el destacado error de identificación de una huella latente recuperada en el atentado con explosivos en el año 2004 en el sistema de trenes de cercanías de Madrid. Un analista del FBI concluyó que con «una certeza del 100%» la huella coincidía con la de Brandon Mayfield, un norteamericano de Portland (Oregón), incluso aunque las autoridades españolas no pudieron confirmar la identificación. Los revisores creen que la identificación errónea se debió, en parte, al «sesgo de confirmación» o «razonamiento inverso» —es decir, el ir de la impresión conocida (la del sospechoso) a la imagen latente, de forma que llevó a confiar exacerbadamente en las aparentes similitudes y prestar inadecuada atención a las diferencias—[257]. Como se describió en un reciente artículo de científicos del Laboratorio del FBI,

> [U]n notable ejemplo del problema del sesgo a partir de una huella de autoría conocida que da lugar a un razonamiento circular ocurrió en el error de identificación del caso del atentado en Madrid, en el que el primer analista reinterpretó cinco de los originales siete puntos de análisis como más consistentes con la huella conocida (incorrecta). «Habiendo encontrado hasta diez puntos característicos de inusual similitud, los analistas del FBI comenzaron a "encontrar" adicionales características en LPF 17 [la huella latente] que realmente no estaban allí, sino que fueron sugeridas a los analistas por las características en las impresiones de Mayfield[258].

A diferencia de los análisis de ADN, las reglas para declarar una identificación que históricamente se utilizaron en los análisis de huellas dactilares no fueron establecidas de antemano ni fueron uniformes entre los analistas. Como se describió en un informe realizado por un Grupo de Trabajo de Analistas fechado en febrero de 2012, encargado por el NIST y el NIJ:

> Los umbrales para estas decisiones pueden variar entre los analistas y entre los proveedores de servicios forenses. Algunos analistas afirman que declaran una identificación si encuentran un

---

[254] U.S. Department of Justice, Office of the Inspector General, 2011. www.oig.justice.gov/special/s1105.pdf. Véase también: Federal Bureau of Investigation. Laboratory Division, 2007 (modificada el 24 de mayo de 2011).

[255] Federal Bureau of Investigation. Laboratory Division, 2007 (modificado el 24 de mayo de 2011).

[256] Faigman, Kaye, Saks, et al. (Eds.), 2016; Saks, 1994.

[257] Una revisión del tratamiento del FBI del caso Brandon Mayfield. U.S. Department of Justice, Office of the Inspector General, 2006. oig.justice.gov/special/s0601/final.pdf.

[258] Ulery, Hicklin, Roberts, et. al., 2015: 54-61. La cita interna es del U.S. Department of Justice, Office of the Inspector General. Una revisión del tratamiento del FBI del caso Brandon Mayfield (marzo 2006), www.justice.gov/oig/special/s0601/PDF_list.htm. US Department of Justice Office of the Inspector General: A review of the FBI's handling of the Brandon Mayfield case (March 2006), www.justice.gov/oig/special/s0601/PDF_list.htm.

> número particular de relativamente raras concurrentes características, por ejemplo, ocho o doce. Otros no utilizan ningún número fijo estándar. Algunos analistas descartan detalles aparentemente diferentes siempre que encuentren suficientes similitudes entre las dos impresiones. Otros analistas practican la regla de una disimilitud, excluyendo una impresión si existe una única disimilitud no atribuible a una distorsión perceptible. Si el analista decide que el grado de similitud se queda corto para satisfacer el estándar, puede declarar un resultado no concluyente. Si la conclusión es que el grado de similitud satisface el estándar, el analista declara una identificación [259].

En septiembre de 2011, el *Scientific Working Group on Friction Ridge Analysis, Study and Technology (SWGFAST)* (Grupo de Trabajo Científico sobre Análisis, Estudio y Tecnologías de Crestas de Fricción) emitió el documento *«Standards for Examining Friction Ridge Impressions and Resulting Conclusions (Latent/Tenprint)»* (Estándares para el Examen de Impresiones de Crestas de Fricción y Conclusiones Resultantes (Latentes/Decadactilares)) que comienza a transformar el análisis de las huellas latentes en la dirección de un marco objetivo. En particular, sugiere criterios sobre qué combinación de calidad de imagen y cantidad de características (por ejemplo, número de puntos característicos compartidos entre dos huellas dactilares) serían suficientes para declarar una identificación. Los criterios no son todavía plenamente objetivos, pero suponen un paso en la dirección correcta. El Subcomité de Crestas de Fricción de OSAC ha reconocido la necesidad de contar con criterios objetivos al identificar las «Necesidades de Investigación» [260]. Observamos que los estudios de caja negra que describimos a continuación no se propusieron probar estos criterios específicos, por lo que aún no han sido científicamente validados.

5.4.2. Estudios de validez científica y fiabilidad

Como se dijo previamente, la validez de los fundamentos de los métodos subjetivos puede *solo* establecerse a través de múltiples estudios independientes de caja negra apropiadamente diseñados para evaluar la validez y la fiabilidad.

Más abajo explicamos diversos estudios de análisis de huellas latentes. Los primeros cinco no se idearon como estudios de validación, aunque proporcionan alguna información incidental sobre el rendimiento. Notablemente, solo ha habido dos estudios de caja negra que fueron intencionada y apropiadamente diseñados para valorar la validez y la fiabilidad —el primero fue publicado por el Laboratorio del FBI en 2011; el segundo se terminó en 2014, pero no se ha publicado todavía—. Las conclusiones sobre la validez de los fundamentos deben apoyarse sobre esos dos estudios recientes.

Para resumir esos estudios, aplicamos la guía escrita anteriormente en este informe (véase Sección 4 y Apéndice A). Primero, aunque observamos tanto (1) la tasa estimada de falsos positivos como (2) el límite superior de confianza del 95% de falsos positivos, nos centramos en el segundo al considerarla como la tasa más apropiada

---

[259] Véase: NIST, 2012. Disponible en: www.nist.gov/oles/upload/latent.pdf.

[260] Véase: workspace.forensicosac.org/kws/groups/fric_ridge/documents.

desde el punto de vista científico para informar a un jurado —porque su principal preocupación debería ser subestimar la tasa de falsos positivos, pudiendo llegar a ser la tasa real razonablemente tan alta como este valor—[261]. Segundo, aunque notamos tanto (1) la tasa de falsos positivos en las examinaciones *concluyentes* (identificaciones o exclusiones) como (2) *todas* las examinaciones (incluyendo las no concluyentes) que sean relevantes, nos centramos en la primera al ser la tasa desde el punto de vista científico más apropiada para informar a un jurado —porque la prueba de huellas dactilares utilizada ante un tribunal contra un acusado será generalmente el resultado de un examen concluyente—.

5.4.2.1. Evett y Williams (1996)

Este artículo es un ensayo de revisión histórica discursiva que contiene una breve descripción de un pequeño «estudio de colaboración» relevante para la precisión de los análisis de huellas dactilares[262]. En este estudio, a 130 analistas de Inglaterra y Gales muy experimentados, cada uno con al menos diez años de experiencia en análisis de huellas dactilares forenses, se les presentaron diez parejas de huellas latentes conocidas. Nueve de las parejas procedían de casos realizados en el pasado por la Scotland Yard, suponiéndose «correctamente emparejadas» (es decir, procedentes de la misma fuente). La décima pareja no estaba realmente emparejada (procedían de fuentes diferentes), incluyendo una latente deliberadamente producida en un vaso de cerveza con hoyuelos. Sobre esta única pareja dispar, los analistas no hicieron identificaciones falsas. Como el artículo no distinguió entre exclusiones y examinaciones no concluyentes (y los autores no disponen ya de los datos)[263], no es posible inferir el límite superior del 95% de confianza[264].

5.4.2.2. Langeburg (2009a)

En un pequeño estudio piloto, el autor examinó la ejecutoria de seis analistas sobre 60 tests cada uno[265]. Solo hubo 15 examinaciones concluyentes relacionados con parejas dispares en origen (véase la Tabla 1 del artículo). Hubo un falso positivo que el autor excluyó porque pareció ser un error administrativo que no se repitió en una nueva prueba posterior. Incluso si ese error se excluye, el pequeño tamaño de la muestra da lugar a un enorme intervalo de confianza (límite superior del 95% de confianza del 19%), con ese límite superior correspondiendo a 1 error cada 5 casos.

---

[261] Por convención, el límite superior del nivel de confianza del 95% es el más utilizado en estadística para reflejar el rango de valores posibles (véase Apéndice A).

[262] Evett y Williams, 1996: 49-73.

[263] I.W. Evett, información dada en persona.

[264] Por ejemplo, el límite superior del 95% de confianza sería 1 cada 44 si los 130 exámenes fueran concluyentes y 1 cada 22 si lo fueran la mitad.

[265] Langenburg, 2009: 219–257.

5.4.2.3. Langeburg (2009b)

En este pequeño estudio piloto para el siguiente artículo, el autor comprobó la destreza de los analistas en una sala de conferencias dentro de una convención de analistas en identificación forense[266]. Los analistas fueron divididos en tres grupos: alto-sesgo (n=16), bajo-sesgo (n=12), y control (n=15). A cada grupo se le presentaron 6 parejas de huellas latentes conocidas, 3 emparejadas realmente y otras tres no. Los dos primeros grupos recibieron información diseñada para sesgar su juicio al aumentar su atención, mientras que el grupo de control recibió una descripción genérica. Con respecto a las no emparejadas, el grupo de control tuvo 1 falso positivo entre 43 examinaciones concluyentes. La tasa de falsos positivos fue del 2.3% (límite superior del 95% de confianza en el 11%), con un límite superior correspondiente a 1 error cada 9 casos[267, 268].

5.4.2.4. Langeburg, Champod y Genessay (2012)

Este estudio no fue diseñado para valorar la precisión de los análisis de huellas dactilares latentes, sino más bien para explorar cómo los analistas en huellas dactilares incorporarían información proveniente de nuevas herramientas desarrolladas (tales como una herramienta de calidad para ayudar a la evaluación de la claridad de los detalles de las crestas de fricción; una herramienta estadística para proporcionar relaciones de verosimilitudes que representen la fuerza de las correspondientes características entre las huellas comparadas; e información de consenso de un grupo de analistas en huellas entrenados) en sus procesos de toma de decisiones[269]. Sin embargo, el estudio provee alguna información sobre la precisión del análisis de huella dactilares latentes. Brevemente, se les pidió a 158 analistas (y a algunos en capacitación) que cotejaran 12 parejas de huellas latentes indubitadas, 7 emparejadas y 5 no emparejadas realmente. Con las no emparejadas hubo 17 falsos positivos entre 711 examinaciones concluyentes de los analistas[270]. La tasa de falsos positivos fue del 2.4% (límite superior del 95% de confianza del 3.5%). El error estimado se corresponde con 1 error cada 42 casos, con un límite superior correspondiente a 1 error cada 28 casos)[271].

---

[266] Langenburg, Champod y Wertheim, 2009: 571-82.

[267] Si se incluyen los dos exámenes inconclusos, los valores son solo ligeramente diferentes: 2.2 % (límite superior del 95% de confianza del 10.1%), con una apuesta de 1 a 10.

[268] Los grupos sesgados no cometieron errores entre los 69 exámenes concluyentes.

[269] Langenburg, Champod y Genessay, 2012: 183-98.

[270] Agradecemos a G. Langenburg por proporcionarnos los datos de los analistas individualizados.

[271] Si se incluyen los 79 exámenes inconclusos, la tasa de falsos positivos fue 2.15% (límite superior del 95% de confianza en el 3.2%). La tasa de falsos positivos estimada se corresponde con 1 error cada 47 casos, con un límite superior de confianza correspondiente a 1 cada 31.

5.4.2.5. Tangen et al. (2011)

Este estudio australiano fue diseñado para estudiar la fiabilidad de los análisis de huellas latentes por parte de los analistas en huellas dactilares[272]. Los autores pidieron a 37 analistas en huellas dactilares, así como a 37 novatos, que examinaran 36 parejas de huellas latentes conocidas —consistentes en 12 pares emparejados, 12 no emparejadas, elegidas por su «similitud» (las huellas indubitadas procedentes de fuentes diferentes mejor situadas en la clasificación en el Sistema Automático Nacional de Identificación Dactilar de Australia) y 12 no emparejadas y «no similares» (elegidas aleatoriamente a partir de otras impresiones)—. A los analistas se les pidió que midieran la probabilidad de que vinieran de la misma fuente en una escala de 1 a 12. Los autores eligieron definir las puntuaciones de 1-6 como identificaciones y de 7-12 como exclusiones[273]. Esta aproximación no se corresponde con los procedimientos utilizados en el examen de huellas dactilares convencional.

Con las parejas «similares» pero no emparejadas, los analistas cometieron 3 errores en 444 comparaciones; la tasa de falsos positivos fue del 0.68% (límite superior del 95% de confianza en el 1.7%), con el límite superior correspondiendo a 1 error cada 58 casos. Con los pares «no similares» no emparejados, los analistas no cometieron error alguno en las 444 comparaciones; la tasa de falsos positivos fue, de este modo, del 0% (límite superior del 95% de confianza en el 0.62%), con el límite superior correspondiendo a 1 error cada 148 casos. Los analistas superaron sustancialmente a los novatos.

Aunque es interesante, el estudio no constituye un estudio de validación de caja negra de análisis de huellas latentes porque su diseño no se asemeja a los procedimientos utilizados en la práctica forense (en particular, el proceso de asignación de calificación en una escala de 12 puntos que los autores posteriormente convirtieron en identificaciones y exclusiones).

5.4.3. Estudios del FBI

El primer estudio diseñado para comprobar la validez de los fundamentos y medir la fiabilidad de los análisis de huellas dactilares latentes fue un estudio de caja negra realizado por científicos y colaboradores del FBI. Emprendido como respuesta al informe del NRC (2009), el estudio se publicó en 2011 en una revista científica líder internacional, *Proceedings of the National Academy of Sciences 274*. Los autores recogieron una selección de 744 parejas de huellas latentes conocidas, consistente en 520 emparejamientos y 224 no emparejadas. Para intentar asegurar que los pares no

---

[272] Tangen, Thompson y McCarthy, 2011: 995-7.

[273] No hubo, de este modo, resultados inconclusos en este estudio.

[274] Ulery, Hicklin, Buscaglia , et al., 2011: 7733-7338.

emparejados fuesen representativos del tipo de coincidencias que pueden presentarse cuando la policía identifica a un sospechoso buscado en la base de datos de huellas dactilares, las huellas conocidas fueron seleccionadas buscando las huellas latentes en 58 millones de huellas dactilares de la base de datos AFIS y seleccionando uno de los resultados coincidentes más cercanos. A cada uno de los 169 analistas de huellas dactilares se les mostraron 100 parejas y se les pidió que las clasificaran como una identificación, una exclusión o una inconclusión. El estudio informó que se habían producido 6 falsas identificaciones entre los 3.628 pares de huellas no emparejadas y que los analistas juzgaron que tenían «valor para la identificación». La tasa de falsos positivos fue, de este modo, del 0.17% (límite superior del 95% de confianza del 0.33%). La tasa estimada se corresponde con 1 error cada 604 casos, con el límite superior indicando que la tasa podría ser tan alta como 1 error cada 306 casos[275, 276].

En 2012, los mismos autores informaron de un estudio de seguimiento que probaba la repetibilidad y la reproducibilidad. Después de un periodo de 7 meses, 75 de los analistas del estudio anterior reexaminaron un subconjunto de comparaciones de huellas latentes conocidas de ese estudio precedente. Entre los 476 pares de huellas latentes no emparejadas que condujeron a examinaciones concluyentes (incluyendo 4 de las parejas que condujeron a falsos positivos en el estudio inicial y que fueron reasignadas al analista que había cometido la decisión errónea), no hubo falsos positivos. Esos resultados (límite superior del 95% de confianza en el 0.63%), correspondiente a 1 error cada 160) son ampliamente consistentes con la tasa de falsos positivos medida en el estudio previo[277].

5.4.3.1. Estudio Miami-Dade (Pacheco et al.)

La Oficina de Servicios Forenses del Departamento de Policía de Miami-Dade, con financiación del NIJ, llevó a cabo un estudio de caja negra diseñado para valorar la validez de los fundamentos y medir la fiabilidad; los resultados fueron entregados al patrocinador y publicado en internet, pero aún no se han publicado en una revista científica con revisión por pares[278]. El estudio difiere significativamente del estudio de caja negra del FBI en 2011 en importantes aspectos, incluido el hecho de que las huellas conocidas no fueron seleccionadas mediante una búsqueda en una gran base de datos para que fueran similares a las huellas latentes (que debía, en principio, haber hecho más fácil declarar exclusiones en los pares no emparejados). El estudio

---

[275] Si incluimos los 455 resultados inconclusos dentro de las huellas latentes juzgadas como que tienen «valor para la identificación», la tasa de falsos positivos es del 0.15% (límite superior del 95% de confianza del 0.29%). La tasa de falsos positivos estimada se corresponde con 1 error cada 681 casos, con límite superior correspondiente de 1 cada 344.

[276] La sensibilidad (proporción de muestras emparejadas que fueron correctamente declaradas como coincidentes) fue del 92.5%.

[277] En general, entre el 85-90% de los resultados concluyentes no se modificaron, con aproximadamente el 30% de falsas exclusiones repetidas.

[278] Pacheco, Cerchiai y Stoiloff, 2014. www.ncjrs.gov/pdffiles1/nij/grants/248534.pdf.

informó de 42 falsos positivos entre 995 examinaciones concluyentes. La tasa de falsos positivos fue del 4.2% (límite superior del 95% de confianza en el 5.4%). La tasa estimada se corresponde con 1 error cada 24 casos, con un límite superior indicando que la tasa podría llegar a ser tan alta como 1 error en 18 casos[279]. (Observación: El artículo dice que «en 35 de las identificaciones erróneas, los participantes parecían haber cometido un error administrativo, pero los autores no pudieron determinarlo con certeza»). En estudios de validación es inapropiado excluir errores de manera *post hoc* (véase Cuadro 4). Sin embargo, si esos errores fueran excluidos, la tasa de falsos positivos sería del 0.7% (intervalo de confianza del 1.4%), con límite superior correspondiente a 1 error cada 73 casos.

5.4.4. Conclusiones de los estudios

Si bien es descorazonador que los estudios significativos para valorar la validez de los fundamentos y la fiabilidad no comenzaran hasta época reciente, nos alienta que ahora se estén haciendo serios esfuerzos para tratar de dotar a la disciplina de un fundamento científico sólido —incluso midiendo la precisión, definiendo la calidad de las huellas latentes, estudiando las razones de los errores, etc.—. Gran parte del crédito pertenece al Laboratorio del FBI, así como a académicos que habían presionado acerca de la necesidad de la investigación. De manera significativa, el FBI es responsable del único estudio de caja negra hasta la fecha que ha sido publicado en una revista científica con revisión por pares.

Los estudios mencionados no pueden compararse directamente entre sí por muchas razones —incluyendo diferencias en el diseño experimental, la selección y el nivel de dificultad de los pares de huellas latentes, así como el grado en el cual representan las circunstancias, procedimientos y presiones que se encuentran en los casos reales—. Sin embargo, pueden emitirse ciertas conclusiones a partir de los resultados de los estudios (resumidos en la Tabla 1 más abajo):

(1) Los estudios colectivamente demuestran que muchos analistas pueden, bajo *algunas* circunstancias, producir respuestas correctas con *algún* nivel de precisión.

(2) Las tasas de falsos positivos empíricamente estimadas son mucho más altas de lo que el público en general (y, por extensión, la mayoría de los miembros del jurado) probablemente creerían, basándose en las afirmaciones largamente sostenidas en el tiempo sobre la seguridad de los análisis de huellas dactilares[280, 281].

---

[279] Si se incluyen los 403 exámenes inconclusos, la tasa de falsos positivos fue del 3% (límite superior del 95% de confianza del 3.9%). La tasa de falsos positivos estimada se corresponde con 1 error cada 33 casos, con límite superior correspondiente a 1 cada 26.

[280] La conclusión se mantiene con independencia de si las tasas están basadas en estimaciones puntuales o en límites de confianza del 95%, así como sobre exámenes concluyentes o sobre todos los exámenes.

[281] Esas afirmaciones incluyen la afirmación anterior del Departamento de Justicia, durante mucho tiempo mantenida, de que el análisis de huellas dactilares es «infalible». (www.justice.gov/olp/

(3) De los dos estudios de caja negra apropiadamente diseñados, el mayor estudio (el del FBI de 2011) arrojó una tasa de falsos positivos que es improbable que exceda de 1 cada 306 exámenes concluyentes, mientras que el otro (el de Miami-Dade de 2014) arrojó una tasa de falsos positivos considerablemente más alta, 1 cada 18 [282]. (Los estudios anteriores, que no fueron diseñados como estudios de validación, arrojaron también altas tasas de falsos positivos).

En general, sería apropiado informar a los miembros del jurado que (1) solo se han realizado dos estudios correctamente diseñados sobre la precisión de los análisis de huellas latentes y que (2) esos estudios informaron de tasas de falsos positivos que pudieran llegar a ser tan altas como de 1 cada 306 casos en un estudio y 1 cada 18 en el otro. Esto informaría apropiadamente a los miembros del jurado que los errores ocurren con una determinada frecuencia, permitiéndoles sopesar el valor probatorio de la evidencia.

Es probable que un programa de verificación sistemática y ciega apropiadamente diseñado disminuyera la tasa de falsos positivos, porque los analistas en los estudios tienden a cometer errores *diferentes 283*. Sin embargo, no se han realizado pruebas empíricas para obtener una estimación cuantitativa de la tasa de falsos positivos que se hubieran producido con un programa como el mencionado [284]. Y no sería apropiado simplemente *inferir* el impacto de la verificación independiente basándose en el supuesto teórico de que los errores de los analistas no están correlacionados [285].

---

file/861906/download); el testimonio de un anterior jefe de la unidad de huellas dactilares del FBI era que «el FBI tenía una tasa de error de 1 cada 11 millones de casos» (véase p. 68); y un estudio con miembros del jurado simulados que estimaban que la tasa de falsos positivos en análisis de huellas latentes era de 1 cada 5.5 millones (véase p. 60). Koehler, 2016. Disponible en: www.papers.ssrn.com/sol3/papers.cfm?abstract_id=2817443.

[282] Como se observó anteriormente, la tasa es de 1 cada 73 si ignoramos los errores supuestamente administrativos —aunque este ajuste *post hoc* no es apropiado en estudios de validación—.

[283] Los autores del estudio de caja negra del FBI observan que cinco de los falsos positivos ocurrieron en problemas de prueba en los que una gran mayoría de los analistas declararon correctamente una exclusión, mientras que uno ocurrió en un problema de prueba en el que la mayoría de los analistas tomaron decisiones de inconclusión. Afirman que «esto sugiere que estos errores de individualización hubieran sido detectados si se hubiese realizado una verificación ciega de manera rutinaria». Ulery, Hicklin, Buscaglia, et al., 2011: 7733-7738.

[284] El estudio Miami-Dade incluyó un pequeño test sobre el paso de la verificación, lo que supuso la verificación de 15 de los 42 falsos positivos. De esos 15 casos, el segundo analista declaró 13 casos de exclusión y 2 de inconclusión. El tamaño muestral es demasiado pequeño como para proporcionar una conclusión significativa. Y el artículo no informa sobre resultados de verificación sobre los otros 27 casos positivos.

[285] El DOJ ha propuesto al PCAST que «la teoría básica de la probabilidad establece que, dada una tasa de error para un analista, la probabilidad de que un segundo analista cometa exactamente el mismo error (verificación/verificación ciega) conlleva que las tasas deberían multiplicarse». Sin embargo, tal modelo teórico asume que los errores de los diferentes analistas no están correlacionados; pero ellos pueden depender de la dificultad del problema y, por tanto, estar correlacionados. Se necesitan estudios empíricos que estimen las tasas de error bajo verificación ciega.

Es importante observar que para que un programa de verificación sea completamente ciego y, por consiguiente, evite el sesgo cognitivo, los analistas no pueden solo verificar individualizaciones. Como proponen los autores del estudio de caja negra del FBI, «esto puede asegurarse realizando verificaciones sobre una mezcla de tipos de conclusiones, no meramente sobre individualizaciones» —es decir, una mezcla que asegure que quienes verifican no puedan hacer inferencias sobre las conclusiones que están siendo verificadas[286]—. No tenemos conocimiento de ningún programa de verificación ciega que siga actualmente esta práctica.

Hasta el presente, el testimonio que asevere cualquier nivel específico de mayor precisión (más allá de lo medido en los estudios) debido a la verificación ciega independiente, sería científicamente inapropiado, en tanto que especulación no está apoyado por evidencia empírica.

Resaltamos que el DOJ opina que la alta tasa de falsos positivos en el estudio Miami-Dade (1 cada 24, con límite superior de confianza de 1 cada 18) es improbable que sea aplicable a los casos del Laboratorio del FBI, porque cree que tan alta tasa de error hubiera sido detectada por los procedimientos de verificación del laboratorio. Una evaluación independiente de los protocolos de verificación podría arrojar luz sobre hasta qué punto tales inferencias podrían realizarse partiendo procedimientos de verificación actuales del laboratorio.

También resaltamos que es posible que la tasa de falsos positivos en casos reales pueda ser más alta que la observada en los estudios experimentales, debido a la exposición a información potencialmente susceptible de producir sesgos en el transcurso de un caso. La introducción de muestras de prueba ciega en el flujo de casos podría proporcionar información valiosa sobre las tasas de error reales en los casos en curso.

En conclusión, el estudio de caja negra del Laboratorio del FBI ha contribuido a un significativo avance en la disciplina. Es necesario el estudio continuo sobre la fiabilidad de los análisis de huellas latentes, basándose en el estudio de su diseño. Los estudios deben, idealmente, estimar las tasas de error en huellas latentes de diversos niveles de «calidad», utilizando medidas bien definidas (idealmente, medidas objetivas implementadas mediante software automático[287]). Como se señaló anteriormente, los estudios deben ser diseñados y dirigidos junto con terceras partes que no

---

[286] Ulery, Hicklin, Buscaglia, et. al., 2011: 7733-7738.

[287] Un ejemplo es *Latent Quality Assessment (LQAS)* (Valoración de la calidad de la latente), diseñado como una herramienta de prueba de un concepto para evaluar la claridad de las impresiones. Hay estudios que demuestran que las tasas de error están correlacionadas con la calidad de las impresiones. El software proporciona un manual y definiciones automatizadas de mapas de claridad, funciones para procesar mapas de claridad y la anotación de los puntos correspondientes que proporcionan un método para superponer áreas de impresión. Hicklin, Buscaglia y Roberts, 2013: 106-117. Otro ejemplo es *Picture Annotation System (PiAnoS)*, desarrollado por la Universidad de Lausana, que está siendo probado como una métrica de calidad y una herramienta de valoración estadística para analistas. Esta plataforma utiliza herramientas que (1) valoran la calidad de los detalles de las crestas de fricción, (2) proporciona relaciones de verosimilitud que representan la fuerza de las características coincidentes

posean interés en el resultado. Este importante rasgo no estuvo presente en el estudio del FBI.

TABLA 1: TASAS DE ERROR EN ESTUDIOS DE ANÁLISIS DE HUELLAS LATENTES[◊]

ESTUDIO	FALSOS POSITIVOS			
	Datos originales	Frecuencia (límite de confianza)	Tasa estimada	Límite en tasa estimada
Estudios anteriores				
Langeburg (2009a)	0/14	0% (19%)	1 en ∞	1 en 5
Langeburg (2009b)	1/43	2.3% (11%)	1 en 43	1 en 9
Langeburg et al. (2012)	17/711	2.4% (3.5%)	1 en 42	1 en 28
Tangen et al. (2011) («pares similares»)	3/444	0.68% (1.7%)	1 en 148	1 en 58
Tangen et al. (2011) («pares diferentes»)	0/444	0% (0.67%)	1 en ∞	1 en 148
Estudios de caja negra				
Ulery et al. 2011 (FBI)[◊◊]	6/3628	0.17% (0.33%)	1 en 604	1 en 306
Pacheco et al. 2014 (Miami-Dade)	42/995	4.2% (5.4%)	1 en 24	1 en 18
Pacheco et al. 2014 (Miami-Dade) (excluyendo errores administrativos)	7/960	0.7% (1.4%)	1 en 137	1 en 73

◊ «Datos originales»: Número de falsos positivos dividido por el número de examinaciones concluyentes incluyendo pares no coincidentes. «Frecuencia (límite de confianza)»: Estimación puntual de la frecuencia de falsos positivos, y límite superior del 95% de confianza. «Tasa estimada»: la apuesta de que ocurra un falso positivo, basada en la proporción observada de falsos positivos. «Límite de la tasa estimada»: la apuesta de ocurrencia de un falso positivo, basado en el límite superior del 95% de confianza—es decir, la tasa podría ser razonablemente tan alta como este valor—.

◊◊ Si los exámenes inconclusos se incluyen en el estudio del FBI, las tasas son 1 cada 681 y 1 cada 344, respectivamente.

5.4.5. Estudios científicos sobre cómo los analistas de huellas latentes alcanzan sus conclusiones

Complementando los estudios de caja negra, algunos estudios han arrojado luz de forma importante sobre cómo los analistas de huellas latentes alcanzan sus conclusiones y cómo esas conclusiones pueden verse influenciadas por factores externos. Esos estudios subrayan los serios riesgos que pueden suscitarse con métodos subjetivos.

---

entre las huellas y (3) brinda información de consenso procedente de un grupo de analistas en huellas dactilares con experiencia. *PiAnoS* es un software de fuente abierta disponible en: ipslabs.unil.ch/pianos.

5.4.5.1. Estudios sobre sesgos cognitivos

Itiel Dror y sus colegas han realizado un trabajo pionero sobre el potencial papel de los sesgos cognitivos en los análisis de huellas latentes [288]. En un estudio exploratorio de 2006, demostraron que el juicio de los analistas podía verse influenciado por las decisiones de otros analistas forenses (una forma de «sesgo de confirmación») [289]. A cinco analistas de huellas dactilares se les facilitó pares de huellas dactilares que ellos habían estudiado cinco años antes en casos reales y que habían declarado «coincidentes». Se les pidió que reexaminaran las impresiones, pero se les indujo a creer que ellas eran un par de impresiones que habían sido declaradas erróneamente coincidentes por el FBI en un caso con notoriedad. Aunque fueron instruidos para que ignorasen esta información, cuatro de los cinco analistas dejaron de considerar a las huellas como «coincidentes». Aunque estos estudios son demasiado reducidos como para proporcionar estimaciones precisas del impacto de los sesgos cognitivos, han sido fundamentales para llamar la atención sobre la cuestión.

Se han propuesto diversas estrategias para mitigar los sesgos cognitivos en los laboratorios forenses, incluyendo la gestión del flujo de información en un laboratorio de criminalística para minimizar la exposición del analista forense a información de contexto irrelevante (tales como confesiones o identificaciones de testigos oculares) y asegurar que los analistas trabajen de un modo lineal, documentando sus hallazgos sobre la evidencia recogida en la escena del crimen *antes* de que realicen comparaciones con muestras de un sospechoso [290, 291].

5.4.5.2. Estudios de caja blanca del FBI

En los últimos años, científicos del FBI y sus colaboradores también han llevado a cabo una serie de estudios de «caja blanca» para comprender los factores subyacentes en el proceso de análisis de huellas dactilares latentes. Esos estudios incluyen análisis de la calidad de las huellas [292, 293], los procesos de los analistas para determinar el valor de una huella latente para su identificación o su exclusión [294], la suficiencia de

---

[288] Dror, Charlton y Peron, 2006: 74-878; Dror y Charlton, 2006: 600-616.

[289] Dror, Charlton y Peron, 2006: 74-878.

[290] Kassin, Dror y Kakucka, 2013: 42-52. Véase también: Krane, Ford, Gilder, et al., 2008: 1006-7.

[291] La información de contexto irrelevante podría, dependiendo de su naturaleza, sesgar la opinión de un analista conduciéndole hacia una incorrecta identificación o a una incorrecta exclusión. Ambos resultados son indeseables.

[292] Hicklin, Buscaglia, Roberts, et al., 2011: 385-419.

[293] Hicklin, Buscaglia y Roberts, 2013: 106-17.

[294] Ulery, Hicklin, Kiebuzinski, et al., 2013: 99-106.

información para las identificaciones[295] y cómo cambia la valoración de los analistas sobre una huella latente cuando la comparan con una posible huella coincidente[296].

Entre el trabajo realizado sobre métodos subjetivos de comparación de características, esta serie de artículos es única en su amplitud, rigor y voluntad de explorar temas desafiantes. No pudimos encontrar análisis autorreflexivos similares para otras disciplinas subjetivas.

Los dos artículos más recientes son particularmente notables porque abordan el grave asunto del sesgo de confirmación.

En un artículo de 2014, los científicos del FBI escribieron:

> ACE distingue entre la fase de Comparación (valoración de las características) y la fase de Evaluación (determinación), implicando que las determinaciones están basadas en la valoración de las características. Sin embargo, nuestros resultados sugieren que no se trata de una simple relación causal: las marcas de los analistas están también influenciadas por sus determinaciones. No es obvio el modo en que se produce esta influencia inversa. Los analistas pueden alcanzar subconscientemente una determinación preliminar rápidamente y esto influye en su comportamiento durante la Comparación (por ejemplo, en el nivel de esfuerzo realizado, la manera de tratar las características ambiguas). Después de tomar una decisión, los analistas pueden luego revisar sus anotaciones como ayuda para documentar esa decisión y pueden estar más motivados a proporcionar un marcado metículoso y cuidadoso en apoyo de las individualizaciones más que otras determinaciones. Como evidencia en apoyo de nuestra conjetura, observamos en particular las distribuciones de los recuentos de puntos característicos, que muestran un aumento escalonado asociado con los umbrales de decisión: este paso ocurrió en torno a los siete puntos característicos para la mayoría de los analistas, aunque en torno a 12 para aquellos que siguen el estándar de los 12 puntos[297].

Dror *et al.* han hecho observaciones similares al notar que el número de puntos característicos marcados en una huella latente era mayor cuando estaba presente una huella indubitada[298]. Además, Evett y Williams describieron cómo los analistas británicos, que utilizaban el estándar de 16 puntos para declarar una identificación, utilizaban una huella indubitada (con la que se compara la latente) para «descartar los puntos discordantes» de la huella latente después de que hubieran alcanzado la «íntima convicción» de que las huellas eran coincidentes[299].

En un informe complementario del año 2015, los científicos del FBI estudiaron cuidadosamente cómo los analistas analizaban las impresiones y confirmaron que, en la gran mayoría de las decisiones de identificación (> 90%), los analistas modificaron los puntos característicos marcados en la huella latente como respuesta a una aparente coincidencia con una huella dactilar conocida (más a menudo añadiendo que

---

295 Ulery, Hicklin y Buscaglia, 2012.
296 Ulery, Hicklin, Roberts, et al., 2015: 54-61.
297 Ulery, Hicklin, Roberts y Buscaglia, 2014.
298 Dror, Champod, Langenburg, et. al., 2011: 10-17.
299 Evett y Williams, 1996: 49–73.

sustrayendo puntos característicos)[300]. (El único falso positivo en su estudio fue un caso extremo en el que la conclusión estuvo casi enteramente basada en el marcado subsiguiente de puntos característicos que no habían sido inicialmente encontrados y en el borrado de puntos catacterísticos que habían sido previamente marcados).

Los autores concluyeron que «hay necesidad de que los analistas tengan algún medio de documentar sin ambigüedad lo que ven durante el análisis y la comparación (en el proceso ACE-V)» y que «métodos rigurosamente definidos y consistentemente aplicados para realizar y docuentar ACE-V mejoraría la transparencia del proceso de examen de las huellas latentes».

El PCAST felicita a los científicos del FBI por llamar la atención sobre el riesgo de sesgo de confirmación derivado de un razonamiento circular. Como una cuestión de validez científica, a los analistas se les debe requerir que «completen y documenten sus análisis de una huella latente antes de mirar cualquier huella conocida» y «deben documentar por separado cualquier dato en el que se basaron durante la comparación o la evaluación que difiera de la información en la que se basaron durante el análisis»[301]. El FBI adoptó estas reglas tras el error de identificación relacionado con el atentado con explosivos de los trenes de Madrid; tales reglas deben ser universalmente adoptadas por todos los laboratorios.

5.4.6. Validez en la aplicación

La validez de los fundamentos significa que un gran grupo de analistas examinando un específico tipo de muestra puede, bajo las condiciones de la prueba, generar respuestas correctas con una frecuencia conocida y útil. Eso no significa que un analista particular tenga la capacidad de aplicar fiablemente el método, que las muestras en los estudios de los fundamentos sean representativas de la evidencia real en los casos o que las circunstancias del estudio de los fundamentos representen una aproximación razonable a las circunstancias del caso.

Para abordar estos asuntos, los tribunales deben tener en cuenta algunas consideraciones clave:

(1) Como consecuencia de que el análisis de huellas latentes, tal y como se practica actualmente, depende de un juicio subjetivo, no está científicamente justificado concluir que un analista particular es capaz de aplicar fiablemente el método a menos que el analista haya sido sometido a pruebas de aptitud regulares y rigurosas. Desafortunadamente, no es posible asegurar que las actuales pruebas de aptitud sean apropiadas porque los problemas que se evalúan en esas pruebas no están disponibles públicamente (Como se enfatizó previamente, ni el entrenamiento ni la experiencia

---

[300] Ulery, Hicklin, Roberts, et. al., 2015: 54-61.

[301] U.S. Department of Justice, Office of the Inspector General, 2011: 5, 27. www.oig.justice.gov/special/s1105.pdf.

son sustitutos, porque ninguno proporciona garantía alguna de que el analista pueda aplicar el método fiablemente).

(2) En cualquier caso, debe establecerse que las huellas latentes poseen la calidad y completitud representada en los estudios de validez de los fundamentos.

(3) Como consecuencia de que el sesgo contextual puede tener un impacto sobre las decisiones de los analistas, los tribunales deben valorar las medidas tomadas para mitigar los sesgos durante la realización de un caso —por ejemplo, asegurándose de que los analistas no están expuestos a información potencialmente sesgadora y asegurarse de que los analistas documenten las características de las crestas de una huella desconocida antes de referirse a la impresión conocida (un procedimiento conocido como «ACE-V lineal»)—[302].

HALLAZGO 5: ANÁLISIS DE HUELLAS DACTILARES LATENTES

Validez de los fundamentos. Basándose principalmente en dos estudios recientes de caja negra apropiadamente diseñados, el PCAST concluye que el análisis de huellas dactilares latentes es una metodología subjetiva válida en sus fundamentos —aunque con una tasa de falsos positivos importante y probablemente más alta que la esperada por muchos miembros del jurado considerando afirmaciones que han sido mantenidas durante mucho tiempo sobre la infalibilidad de los análisis de huellas dactilares.

Las conclusiones de una propuesta de identificación pueden ser científicamente válidas si están acompañadas de información precisa sobre las limitaciones de la fiabilidad de la conclusión —específicamente, que (1) solo se han llevado a cabo dos estudios apropiadamente diseñados sobre la validez de los fundamentos y la precisión de los análisis de huellas latentes; (2) estos estudios encontraron tasas de falsos positivos que podrían ser tan altas como de 1 error cada 306 casos, en uno de ellos, y de 1 error cada 18 en el otro; y (3) como consecuencia de que los analistas eran conscientes de que estaban siendo evaluados, la tasa real de falsos positivos en casos reales puede ser más alta. Hasta el presente, afirmaciones sobre una mayor precisión no están garantizadas o científicamente justificadas. Se requieren estudios de caja negra adicionales para aclarar la fiabilidad del método.

Validez en la aplicación. Aunque concluímos que el método es válido en sus fundamentos, hay una serie de cuestiones importantes relacionadas con la validez de su aplicación.

(1) Sesgo de confirmación. El trabajo realizado por científicos del FBI ha mostrado que los analistas suelen alterar las características que marcan inicialmente en una huella latente basándose en las comparaciones realizadas con una huella indubitada aparentemente coincidente. Ese razonamiento circular introduce un serio riesgo de sesgo de confirmación. A los analistas se les debe requerir que completen y documenten sus análisis de una huella latente *antes* de mirar una huella conocida y deben documentar,

[302] U.S. Department of Justice, Office of the Inspector General, 2011: 27. www.oig.justice.gov/special/s1105.pdf.

separadamente, cualquier dato adicional utilizado durante su comparación y evaluación.

(2) Sesgo contextual. El trabajo realizado por miembros de la comunidad académica ha demostrado que el juicio de los analistas puede estar influenciado por información irrelevante sobre los hechos de un caso. Deben realizarse esfuerzos para asegurar que los analistas no están expuestos a información potencialmente sesgadora.

(3) Pruebas de aptitud. La prueba de aptitud es esencial para valorar la capacidad de un analista y su rendimiento en la emisión de juicios precisos. Como se trató sobre el particular en otro lugar de este informe, las pruebas de aptitud necesitan mejorarse haciéndolas más rigurosas, incorporándolas dentro del flujo de trabajo, y divulgando las pruebas para que ellas sean evaluadas por la comunidad científica.

Desde el punto de vista científico, la validez en la aplicación requiere que un analista: (1) haya realizado pruebas de aptitud apropiadas para asegurar que es capaz de analizar todo el rango de huellas latentes que es posible encontrar en los casos reales y que informe de los resultados de las pruebas de aptitud; (2) revele si documentó las características halladas en la huella latente por escrito antes de compararla con la impresión conocida; (3) proporcione un análisis escrito explicando la selección y comparación de las características; (4) revele si, al realizar la examinación, tenía conocimiento de otros hechos del caso que pudieran haber influido en su conclusión; y (5) verifique que la huella latente en el caso en cuestión es similar en calidad al rango de huellas latentes consideradas en los estudios de los fundamentos.

5.4.7. El camino a seguir

Hacen falta continuados esfuerzos para mejorar el estado de los análisis de huellas latentes —y esos esfuerzos darán claramente dividendos para el sistema de justicia penal—.

Una primera dirección es continuar mejorando el análisis de huellas latentes como método subjetivo. Con solo dos estudios de caja negra hasta la fecha (con tasas de error muy diferentes), hay necesidad de que se realicen estudios de caja negra adicionales sobre la base del diseño del estudio de caja negra del FBI. Los estudios deben estimar las tasas de error de las huellas latentes en función de su variabilidad en calidad y completitud, utilizando mediciones bien definidas. Como se señaló anteriormente, los estudios deben diseñarse y dirigirse junto a terceros sin interés en los resultados.

Una segunda dirección —y más importante— es la de convertir los análisis de huellas latentes de un método subjetivo a un método objetivo. En las décadas pasadas se han visto avances extraordinarios en análisis de imagen automatizados basados en *machine learning* y otras aproximaciones— que han conducido a mejoras espectacu-

lares como las conseguidas en reconocimiento facial[303], [304]. En medicina, por ejemplo, se espera que el análisis de imágenes automatizado conducirá hacia un estándar de oro para muchas aplicaciones que precisan de una interpretación de imágenes en radiografías, resonancias magnéticas, oftalmoscopías y dermatológicas[305].

Los métodos objetivos basados en análisis de imagen automatizados podrían producir grandes beneficios —incluida una mayor eficiencia y menores tasas de error—; podrían permitir también la estimación de las tasas de error de millones de comparaciones por parejas. Los esfuerzos iniciales para desarrollar sistemas automatizados no pudieron superar a los seres humanos[306]. Sin embargo, dado el ritmo de los progresos en análisis de imágenes y de *machine learning*, creemos que es probable que el análisis de huellas latentes totalmente automatizado sea posible en un futuro próximo. Se han dado ya pasos iniciales en esa dirección, tanto en el ámbito académico como en la industria[307].

El recurso más importante para impulsar el desarrollo de los métodos objetivos sería la creación de enormes bases de datos que contengan impresiones conocidas, cada una con muchas impresiones latentes «simuladas» de calidad e integridad muy diversa, que pudieran ponerse a disposición de muchos investigadores científicamente capacitados del ámbito académico y de la industria. Las impresiones latentes simuladas podrían crearse «modelando» (*morphing*) las impresiones conocidas, basándose en transformaciones derivadas de la recolección de pares de impresión latente real–impresión registrada[308].

---

[303] Véase: cs.stanford.edu/people/karpathy/cvpr2015.pdf.

[304] Lu y Tang, «Surpassing human-level face verification performance on LFW with GaussianFace.» arxiv.org/abs/1404.3840 (última consulta el 2 de julio de 2016). Taigman, Yang, Ranzato y Wolf, «Deepface: Closing the gap to human-level performance in face verification» .www.cs.toronto.edu/~ranzato/publications/taigman_cvpr14.pdf (última consulta el 2 de julio de 2016); Schroff, Kalenichenko y Philbin, «FaceNet: A unified embedding for face recognition and clustering». arxiv.org/abs/1503.03832 (última consulta el 2 de julio de 2016).

[305] Doi, 2007: 198-211; Shiraishi, Li, Appelbaum, et al. 2011: 449-462.

[306] Por ejemplo, un estudio de 2010 concluyó que los seres humanos mejoraban los resultados de los programas automatizados de comparación de marcas de herramientas. Véase: Chumbley, Morris, et al., 2010: 953-961.

[307] Arunalatha, Tejaswi, Shaila, et al., 2015: 482-490; Srihari, 2013. www.crime-sceneinvestigator.net/QuantitativeMeasuresinSupportofLatentPrint.pdf. Además, el grupo de Christophe Champod en la Universidad de Lausana mantiene un programa activo en esta área.

[308] Por razones de privacidad, podrían utilizarse huellas dactilares de fallecidos.

5.5. Análisis de armas de fuego

5.5.1. Metodología

En los análisis de armas de fuego, los analistas intentan determinar si la munición está o no asociada con una *específica* arma de fuego basándose en las marcas de herramientas producidas por las armas sobre la munición[309], [310]. (Brevemente, los cañones de las armas usualmente están estriados para mejorar la precisión, lo que conlleva que se corten ranuras en espiral en el interior del cañón para dar un giro sobre el proyectil). Las imperfecciones individuales aleatorias que se producen durante el proceso del corte de la herramienta y mediante el «desgaste» por el uso del arma dejan marcas de herramienta sobre proyectiles y casquillos cuando salen del arma de fuego. Las partes del arma de fuego que entran en contacto con los casquillos tienen otros métodos mecánicos de funcionamiento).

Esta disciplina se fundamenta en la idea de que las marcas de herramienta producidas por diferentes armas de fuego varían sustancialmente lo suficiente (debido a las variaciones en su fabricación y uso) como para permitir que los componentes de los cartuchos disparados sean identificados como provenientes de armas de fuego específicas. Por ejemplo, los analistas pueden comparar los cartuchos «en cuestión» de un arma recuperados en una escena del crimen con los disparos de prueba de un arma sospechosa.

Resumidamente, el examen comienza con una evaluación de características de clase de los proyectiles y casquillos, que son características permanentes y predeterminadas anteriores a su fabricación. Si esas características de clase son diferentes, se emite una conclusión de exclusión. Si son similares, el examen procede a identificar y comparar características individuales, tales como las estrías que surgen durante el disparo de un arma en particular. De acuerdo con la *Association of Firearm and Tool Mark Examiners (AFTE)* (Asociación de Analistas de Armas de Fuego y de Marcas de Herramientas), «el método más ampliamente aceptado que se utiliza para llevar a cabo un examen de marcas de herramientas es el de una comparación microscópica conjunta entre las marcas sobre una pieza de material cuestionada y las marcas de origen conocido realizadas por una herramienta»[311].

---

[309] Los analistas pueden también emprender otras clases de análisis, tales como determinaciones de distancias, operatividad de las armas y restauraciones de números de serie, así como el análisis de residuos de disparo para determinar si alguien utilizó el arma recientemente.

[310] Para descripciones más completas, véase, por ejemplo, National Research Council, 2009; y www.archives.fbi.gov/archives/aboutus/lab/forensic-science-communications/fsc/july2009/review/2009_07_review01.htm.

[311] Véase: Foundational Overview of Firearm/Toolmark Identification t*ab* (descripción general de la pestaña sobre los fundamentos de la identificación de armas de fuego/marcas de herramientas) en afte.org/resources/swggun-ark (última consulta el 12 de mayo de 2016).

5.5.2. Antecedentes

En la sección anterior, el PCAST expresó su preocupación sobre ciertos documentos fundamentales que subyacen a la disciplina científica sobre el análisis de armas de fuego y de marcas de herramienta. En particular, observamos que la «Teoría de la identificación en cuanto relacionada con las marcas de herramienta» de AFTE —que define los criterios para declarar una identificación— es circular[312]. La «teoría» establece que un analista puede concluir que dos vestigios tienen un origen común si sus marcas son «suficientemente concordes», donde «suficientemente concordes» se define cuando el analista está convencido de que es extremadamente improbable que los dos vestigios tengan diferente origen. Además, la «teoría» establece, explícitamente, que las conclusiones son subjetivas.

En esta disciplina científica se ha prestado mucha atención en tratar de probar la noción de que cada arma produce marcas de herramienta «únicas». En el año 2004, el NIJ pidió a NRC que estudiase la viabilidad, precisión, fiabilidad y conveniencia de desarrollar una base de datos nacional de balística integral de imágenes de proyectiles disparados por todas o casi todas las armas de nueva fabricación o importadas, con el propósito de permitir la comparación balística de una escena del crimen con un arma y con la información de su propietario inicial.

En su informe de 2008, un comité del NRC, respondiendo la solicitud del NIJ, dijo que «la validez de los presupuestos fundamentales de unicidad y reproducibilidad de las marcas de herramientas relacionadas con armas de fuego» aún no se habían demostrado y que, dados los métodos de comparación actuales, una búsqueda en una base de datos «probablemente produciría un subconjunto de candidatos a coincidencias demasiado grande como para ser útil a efectos prácticos con fines de investigación»[313].

Desde luego, no es necesario que las marcas de herramientas sean únicas para que provean de información útil sobre si un proyectil ha sido disparado por un arma en particular. Sin embargo, es *esencial* que la precisión del método para compararlos se conozca mediante estudios empíricos.

Los analistas de armas de fuego llevan mucho tiempo sosteniendo que su disciplina es de una precisión cercana a la perfección. En un artículo del año 2009, el jefe de la unidad de armas de fuego-marcas de herramientas del laboratorio del FBI afirmó que «un analista cualificado, raramente, si es que ocurre alguna vez, cometerá un falso positivo (falsa identificación)», para ello cita su revisión, haciendo una declaración jurada, de estudios empíricos que mostraban virtualmente ausencia de errores[314].

---

[312] Association of Firearm and Tool Mark Examiners, 2011: 287.

[313] National Research Council, 2008, 3-4.

[314] Véase: www.fbi.gov/about-us/lab/forensic-science-communications/fsc/july2009/review/2009_07_review01.htm

Con respecto al análisis de armas de fuego, el informe del NRC (2009) concluyó que «no se han realizado suficientes estudios para comprender la fiabilidad y reproducibilidad de los métodos» —es decir, que la validez de los fundamentos de la disciplina no había sido establecida—[315].

El *Scientific Working Group on Firearms Analysis (SWGGUN)* (grupo de trabajo científico sobre análisis de armas de fuego) respondió a las críticas del informe del NRC (2009) diciendo que:

> SWGGUN desde hace tiempo ha sido consciente de las cuestiones científicas y sistémicas identificadas en este informe y ha estado trabajando diligentemente para abordarlas… [el informe del NRC] identifica las áreas en las que fundamentalmente debemos mejorar nuestros procedimientos para mejorar la calidad y fiabilidad de nuestros resultados científicos, así como para articular mejor la base de nuestra ciencia[316].

5.5.3. Estudios de análisis de armas de fuego distintos a los de caja negra: análisis basados en conjuntos

Como consecuencia de que el análisis de armas de fuego es, hasta el presente, un método subjetivo de comparación de características, la validez de sus fundamentos, tal como se explicó anteriormente, *solo* puede ser establecida mediante múltiples e independientes estudios de caja negra.

Aunque el análisis de armas de fuego se ha utilizado durante muchas décadas, solo recientemente su validez ha sido sometida a pruebas empíricas significativas. En los últimos 15 años, la disciplina ha emprendido una serie de estudios con el objetivo de estimar la precisión de las conclusiones de los analistas. Aunque los resultados demuestran que los analistas pueden, en ciertas circunstancias, identificar la fuente de la munición disparada, muchos de los estudios no fueron apropiados para valorar la validez científica y estimar la fiabilidad, porque emplearon diseños artificiales que difieren en aspectos importantes de los problemas a los que se enfrentan los analistas en casos reales.

Específicamente, muchos de los estudios emplean análisis «basados en conjuntos», en los que se pide a los analistas que realicen todas las comparaciones por pares dentro de un pequeño conjunto de muestras o entre pequeños conjuntos de muestras. Por ejemplo, un análisis «dentro del conjunto» que involucra n objetos pide a los analistas que rellenen una matriz $n \times n$ que indique cuál de los $n(n\text{-}1)/2$ posibles

---

[315] El informe afirma que «los análisis de marcas de herramientas y de armas de fuego padecen las mismas limitaciones discutidas anteriormente que la prueba de impresiones. A causa de que no resulta suficientemente conocida la variabilidad entre las herramientas y armas individuales, no es posible especificar cuántos puntos de similitud son necesarios para alcanzar un determinado nivel de confianza en el resultado. No se han realizado suficientes estudios para comprender la fiabilidad y repetibilidad de los métodos. El comité está de acuerdo en que las características de clase son útiles para estrechar la lista de herramientas que pueden haber dejado una marca distintiva». National Research Council, 2009: 154.

[316] Véase: www.swggun.org/index.php?option=com_content&view=article&id=37&Itemid=22.

pares coinciden. Algunos científicos forenses han favorecido los diseños basados en conjuntos porque un pequeño número de objetos da lugar a un gran número de comparaciones. Sin embargo, el diseño del estudio tiene un grave defecto: las comparaciones no son *independientes* entre sí. Más bien, implican dependencias internas que (1) restringen y, por consiguiente, informan las respuestas de los analistas y (2), en algunos casos, permite a los analistas hacer inferencias sobre el diseño del estudio. (El primer punto se ilustra mediante la observación de que, si A y B se juzgan coincidentes, entonces cada elemento adicional C ha de coincidir con *ambos* o con *ninguno* de ellos, reduciendo el espacio de posibles respuestas a la mitad. Si A y B coinciden entre sí, pero no coinciden con C, esto crea dependencias adicionales. Y así sucesivamente. El segundo punto se ilustra mediante diseños de «conjunto cerrado», que se describen más abajo).

Como consecuencia de las complejas dependencias entre las respuestas, los estudios basados en conjuntos no son estudios de caja negra de diseño apropiado a partir de los cuales podamos obtener estimaciones adecuadas de precisión. Es más, el análisis de los resultados empíricos de, al menos, algunos estudios basados en el diseño de conjunto (los diseños de «conjunto cerrado»), sugiere que pueden subestimar la tasa de falsos positivos.

El director del Centro de Ciencias Forenses de la Defensa (DFSC en inglés) hizo una analogía de los estudios de conjunto cerrado con la resolución de un «sudoku», donde las respuestas iniciales se pueden utilizar para ayudar a completar las respuestas subsiguientes[317]. Como se explica a continuación, la disconformidad de DFSC con los estudios basados en conjuntos le condujo a financiar el primer estudio (y hasta la fecha, el único) de caja negra apropiadamente diseñado para el análisis de armas de fuego.

Seguidamente analizamos los estudios más citados basados en conjuntos. Adoptamos el mismo marco de referencia que con las huellas latentes, tomando en consideración como medidas apropiadas para mencionarlas en los informes (véase p.115) principalmente: (1) el límite superior del 95% de confianza de la tasa de falsos positivos y (2) la tasa de falsos positivos basada en la proporción de examinaciones concluyentes.

5.5.4. Comparación dentro de un conjunto

Algunos estudios incluyen comparaciones dentro de un conjunto, en el que se presenta a los analistas, por ejemplo, una selección de muestras y se les pide que determinen qué muestras fueron disparadas con la misma arma de fuego. Revisa-

[317] Entrevista del PCAST a Jeff Salyards, director del DFSC.

mos dos estudios frecuentemente citados con este diseño[318], [319]. En estos estudios, la mayoría de las muestras procedían de distintas fuentes, con solo 2 o 3 muestras procedentes del mismo origen. A lo largo de los dos estudios, los analistas identificaron 55 de las 61 coincidencias y no cometieron falsos positivos. En el primer estudio, la gran mayoría de las muestras procedentes de fuentes diferentes (97%) se declararon inconclusas; solo hubo 18 examinaciones concluyentes de casquillos procedentes de fuentes diferentes y ninguna examinación concluyente de proyectiles procedentes de fuentes diferentes[320]. En el segundo estudio, los resultados se describen solamente en breves párrafos y el número de examinaciones concluyentes para muestras procedentes de fuentes diferentes no fue reflejado. De este modo, resulta imposible estimar la tasa de falsos positivos entre las examinaciones concluyentes, que es la medida clave que hay que tener en cuenta (como se explicó anteriormente).

5.5.4.1. Comparación del tipo conjunto a conjunto/conjunto cerrado

Otro diseño común ha sido la comparación entre conjuntos implicando un «conjunto cerrado». En este caso, a los analistas se les entrega un conjunto de muestras cuestionadas y se les pide que las comparen con un conjunto de estándares conocidos, que representan las posibles armas de las cuales se dispararon las municiones cuestionadas. En un diseño de «conjunto cerrado», el arma fuente siempre está presente. Analizamos cuatro de esos estudios en detalle[321, 322,323,324]. En estos estudios, a los analistas

---

[318] Smith, E., 2005: 130-135. En este estudio del FBI, se dispararon casquillos y proyectiles de 9 pistolas Ruger P89 incautadas de casos reales. A los analistas se les dio cajas (de casquillos o proyectiles) con muestras disparadas por cada una de las 9 pistolas y una muestra adicional disparada con una de las pistolas; se les pidió que determinaran qué muestras habían sido disparadas con la misma arma. De las 16 comparaciones procedentes de la misma fuente, 13 (muestras) fueron identificadas y 3 se declararon inconclusas. De las 704 comparaciones procedentes de distintas fuentes, el 97% se declararon inconclusas, un 2.5% exclusiones y un 0% falsos positivos.

[319] DeFrance y Van Arsdale, 2003: 35-37. En este estudio del FBI, se dispararon 5 proyectiles de 5 cañones de barril de revólveres Smith & Wesson, de calibre 357 Magnum, fabricados consecutivamente. Cada uno de los 9 analistas recibió dos cajas para las pruebas, cada una conteniendo un proyectil de cada uno de los 5 revólveres y dos proyectiles adicionales (de diferentes revólveres en una de las cajas y del mismo revólver en la otra); se les pidió que realizaran las 42 posibles comparaciones por pares, que incluían 37 comparaciones con muestras de diferentes fuentes. Del total de 45 comparaciones de muestras procedentes de la misma fuente, hubo 42 identificaciones y 3 examinaciones inconclusas. Del total de 333 comparaciones de muestras de diferentes fuentes, el artículo afirma que no hubo falsos positivos, pero no informa sobre el número de examinaciones inconclusas.

[320] Algunas políticas de laboratorio exigen un listón muy alto para declarar exclusiones.

[321] Stroman, 2014: 157-175. En este estudio, los proyectiles fueron disparados por tres revólveres Smith & Wesson. Cada uno de los 25 analistas recibió un conjunto de prueba con 3 casquillos cuestionados y 3 casquillos conocidos procedentes de cada revólver. De las 75 respuestas, hubo 74 asignaciones correctas y un examen inconcluso.

[322] Brundage, 1998: 438-444. En este estudio, los proyectiles fueron disparados desde 10 cañones de pistola semiautomática de 9 milímetros, de la marca Ruger P-85, consecutivamente fabricados. Cada uno de los 30 analistas recibió un conjunto de prueba de 20 proyectiles cuestionados para su compa-

Nota 323 y 324 en página siguiente

se les entregó una colección de proyectiles cuestionados y/o casquillos disparados desde un pequeño número de armas consecutivamente fabricadas de la misma marca (3, 10, 10 y 10 armas, respectivamente) y una colección de proyectiles (o casquillos) conocidos por haber sido disparados por esas mismas armas. Se les pidió que realizaran un ejercicio comparativo —asignando los proyectiles (o casquillos) de un conjunto a los proyectiles (o casquillos) del otro—.

Este diseño de «conjunto cerrado» es más simple que los problemas que se encuentran en los casos reales, porque la respuesta correcta está siempre presente en la colección. En tales estudios, los analistas pueden ejercer perfectamente su función con tal que simplemente emparejen cada proyectil con el estándar que sea *más próximo*. Por contraste, en un estudio de conjunto abierto (como en un caso real), no hay garantía de que la respuesta correcta esté presente —y, de este modo, no hay garantía de que la coincidencia más próxima sea la correcta—. De las comparaciones de conjunto cerrado cabe esperar, por tanto, que subestimen la tasa de falsos positivos.

Es importante reseñar que no es necesario que a los analistas se les informe, explícitamente, que el diseño del estudio es de conjunto cerrado. Como se observó en uno de los estudios:

> A los participantes no se les dijo si los casquillos constituían un conjunto abierto o cerrado. Sin embargo, a partir del cuestionario/ hoja de respuestas, los participantes podrían haber asumido que se trataba de un conjunto cerrado y que cada casquillo cuestionado debía asociarse con una de las diez correderas[325].

Además, a medida que los participantes encuentren que muchos de los casquillos cuestionados tienen fuertes similitudes con los casquillos conocidos, su suposición de que los conocidos coincidentes siempre están presentes tenderá a confirmarse.

El problema que se suscita con este diseño de estudio no es solo una posibilidad teórica: es evidente en los propios resultados. Específicamente, los estudios de conjunto cerrado tienen tasas de examinaciones inconclusas y de falsos positivos sustan-

---

ración con un conjunto de 15 proyectiles estándares, conteniendo al menos un proyectil disparado con cada una de las diez pistolas. De las 300 respuestas dadas, no hubo asignaciones incorrectas y una examinación inconclusa.

[323] Fadul, Hernandez, Stoiloff, et al., 2013: 376-393. Un estudio empírico para mejorar la fundamentación científica de la identificación de armas de fuego y marcas de herramienta utilizando 10 pistolas con corredera fabricadas consecutivamente. En este estudio, los proyectiles fueron disparados desde 10 cañones de pistola semiautomática de 9 milímetros con corredera, de la marca Ruger P-85, consecutivamente fabricados. Cada uno de los 217 analistas recibió un conjunto de prueba de 15 casquillos cuestionados y 2 conocidos de cada una de las 10 pistolas. De las 3255 respuestas recibidas, hubo 3239 asignaciones correctas, 14 examinaciones inconclusas y 2 falsos positivos.

[324] Hamby, Brundage y Thorpe, 2009: 99-110. En este estudio, se dispararon proyectiles de 10 cañones consecutivamente estriados de la marca Ruger P-85. Cada uno de los 440 analistas recibió un conjunto de prueba consistente en 15 proyectiles cuestionados y 2 conocidos estandarizados de cada una de las 10 pistolas. De las 6600 respuestas recibidas, hubo 6593 asignaciones correctas, siete examinaciones inconclusas y ningún falso positivo.

[325] Fadul, Hernandez, Stoiloff, et al., 2013: 376-93.

cialmente más bajas (en más de 100 veces) que los estudios de diseño parcialmente abierto (estudio Miami-Dade) o totalmente abierto, los diseños de caja negra (Laboratorio Ames) descritos a continuación (Tabla 2)[326].

En resumen, el diseño en conjunto cerrado es problemático por principio y parece subestimar la tasa de falsos positivos en la práctica[327]. El diseño no es apropiado para valorar la validez científica y medir la fiabilidad.

5.5.4.2. Comparación del tipo conjunto a conjunto/conjunto parcialmente abierto ('Estudio Miami-Dade')

Un estudio involucró una comparación, conjunto a conjunto, en el que unas pocas de las muestras cuestionadas carecían de un estándar conocido coincidente[328]. A los 165 analistas en el estudio se les pidió que asignaran una colección de 15 muestras cuestionadas, disparadas desde 10 pistolas, a una colección de estándares conocidos; dos de las 15 muestras cuestionadas procedían de un arma de la que no se proporcionaron estándares conocidos. Respecto a estas dos muestras, hubo 188 eliminaciones, 138 examinaciones inconclusas y 4 falsos positivos. La tasa de inconclusión fue del 41.8% y la tasa de falsos positivos entre los exámenes concluyentes fue del 2.1% (intervalo de confianza del 0.6% al 5.25%). La tasa de falsos positivos se corresponde con una tasa estimada de 1 error cada 48 casos, con un límite superior de 1 error cada 19.

Como se indicó anteriormente, los resultados del estudio Miami-Dade son marcadamente diferentes de los estudios de conjunto cerrado: (1) la proporción de resultados no concluyentes fue 200 veces mayor y (2) la tasa de falsos positivos 100 veces más alta, aproximadamente.

5.5.5. Estudios recientes de análisis caja negra sobre armas de fuego

En 2011, el Comité de Investigación Forense de la *American Society of Crime Lab Directors* (Sociedad estadounidense de directores de laboratorios de criminalística) identificó, entre las necesidades de mayor rango en la ciencia forense, la importancia

---

[326] De las 10.230 respuestas dadas entre los tres estudios, hubo 10.205 asignaciones correctas, 23 examinaciones inconclusas y 2 falsos positivos.

[327] Stroman (2014) reconoce que, aunque las instrucciones de la prueba no indicaron explícitamente que el estudio era cerrado, su estudio podría mejorarse si «se utilizaran armas de fuego adicionales y se conociera que solo una parte de esas armas fueron usadas en los kits de prueba, presentando así un conjunto abierto de (armas) desconocidas para los participantes. Si bien esto podría aumentar la probabilidad de resultados no concluyentes, sería un reflejo más preciso de los tipos de evidencia que se reciben en los casos reales».

[328] Fadul, Hernandez, Stoiloff, et al., 2013, www.ncjrs.gov/pdffiles1/nij/grants/244232.pdf.

de acometer un estudio de caja negra en el análisis de armas de fuego análogo al estudio de caja negra del FBI sobre huellas dactilares latentes. El DFSC, insatisfecho con el diseño de estudios previos de análisis de armas de fuego, concluyó que era necesario un estudio de caja negra y que debía ser dirigido por un laboratorio de prueba independiente no afiliado a las fuerzas de seguridad pública que involucraría a los analistas forenses como participantes en el estudio. El DFSC y la *Defense Forensics and Biometrics Agency* (Agencia de ciencias forenses y biométricas de la defensa), conjuntamente financiaron un estudio realizado por el Laboratorio Ames, un laboratorio nacional del Departamento de energía afiliado a la Universidad estatal de Iowa[329].

5.5.6. Pruebas independientes/abiertas ('Estudio del laboratorio Ames')

El estudio empleó un diseño similar al estudio de caja negra del FBI sobre huellas dactilares latentes, con muchos analistas tomando una serie de decisiones de comparaciones *independientes* entre una muestra cuestionada y una o más muestras conocidas que pueden o no contener la fuente. Todas las muestras procedieron de 25 pistolas Ruger, de calibre 9 milímetros, compradas nuevas[330]. A cada uno de los 218 analistas[331] se les presentó 15 problemas de comparación *separados* —cada uno consistente en una muestra cuestionada y tres disparos de prueba conocidos realizados con una misma arma conocida, que podría o no haber sido la fuente—[332]. Sin que los analistas lo supieran, había cinco comparaciones de la misma fuente y diez de fuentes diferentes. (En un diseño ideal, la proporción de las comparaciones de la misma fuente y de distintas fuentes diferiría entre los analistas).

En las 2.178 comparaciones de diferentes fuentes hubo 1.421 eliminaciones, 735 no concluyentes y 22 falsos positivos. La tasa de no concluyentes fue del 33.7% y la de falsos positivos entre eliminaciones concluyentes fue del 1.5% (límite superior del 95% de confianza del 2.2%). La tasa de falsos positivos se corresponde con una tasa estimada de 1 error cada 22 casos, con límite superior de 1 cada 46. (Debe advertirse que 20 de los 22 falsos positivos fueron realizados por solo 5 de los 218 analistas —sugiriendo fuertemente que la tasa de falsos positivos es altamente heterogénea entre los analistas—).

---

329 Baldwin, Bajic, Morris, et al., 2014.

330 Hubo una crítica, realizada por un científico forense, consistente en decir que el estudio no se llevó a cabo con armas *consecutivamente fabricadas*.

331 Los participantes fueron miembros de AFTE que eran analistas en ejercicio, empleados o retirados, de una agencia nacional o internacional de seguridad pública, con adecuado entrenamiento.

332 Los casos reales pueden ser situaciones más complejas (por ejemplo, muchos proyectiles diferentes en la escena del crimen). Sin embargo, una apropiada valoración de la validez de los fundamentos debe *comenzar* con la pregunta sobre con cuánta frecuencia puede un analista determinar si un proyectil cuestionado procede de una fuente específica conocida.

Los resultados de los distintos estudios se muestran en la Tabla 2. Las tablas muestran una diferencia llamativa entre los estudios de conjunto cerrado (donde por diseño siempre existe un estándar de coincidencia) y los estudios que no lo son (donde no hay garantía de que alguno de los estándares conocidos coincida). Específicamente, los estudios de conjunto cerrado muestran una tasa sustancialmente menor de examinaciones no concluyentes y de falsos positivos. Con este diseño inusual, los analistas logran tener éxito en responder a todas las preguntas y alcanzan esencialmente puntuaciones perfectas. En los diseños abiertos, más realistas, esas tasas son mucho más altas.

TABLA 2: RESULTADOS DE LOS ESTUDIOS DE ARMAS DE FUEGO[◊]

TIPO DE ESTUDIO	RESULTADOS DE COMPARACIONES CON DIFERENTES FUENTES				
	Datos originales	No concluyentes	Falsos positivos entre exámenes concluyentes[333]		
	Exclusiones / Inconclusiones / Falsos positivos		Frecuencia (Límite de confianza)	Tasa estimada	Límite sobre tasa estimada
Conjunto a conjunto/cerrado (4 estudios)	10.205/23/2	0.2%	0.02% (0.06%)	1 en 5.103	1 en 1.612
Conjunto a conjunto/parcialmente abierto (estudio Miami-Dade)	188/134/4	41.8%	2% (2.7%)	1 en 49	1 en 21
Estudio de caja negra (Estudio del Laboratorio Ames)	1.421/735/22	33.7%	1.5% (2.2%)	1 en 66	1 en 46

◊ «No concluyentes»: proporción del total de examinaciones que se calificaron como inconclusas. «Datos originales»: Número de falsos positivos dividido por el número de examinaciones concluyentes, incluyendo vestigios cuestionados sin un correspondiente conocido (para conjunto a conjunto/abierto ligeramente) o parejas no realmente emparejadas (para independiente/abierto). «Frecuencia (límite de confianza)»: Estimación puntual de frecuencia de falsos positivos, y límite superior del 95% de confianza. «Tasa estimada»: la apuesta de ocurrencia de un falso positivo, basada en la proporción observada de falsos positivos. «Límite en tasa estimada»: la apuesta de ocurrencia de un falso positivo, basada en el límite superior del intervalo de confianza del 95% —es decir, la tasa podría ser razonablemente tan alta como este valor—.

5.5.7. Conclusiones

Los primeros estudios indican que los analistas pueden, bajo ciertas circunstancias, asociar la munición con el arma con la que fue disparada. Sin embargo, como

[333] Las tasas para *todos* los exámenes son, leyendo entre filas: 1 en 5.115; 1 en 1.416; 1 en 83; 1 en 33; 1 en 99; y 1 en 66.

se describió más arriba, la mayoría de esos estudios se realizaron con diseños que no son apropiados para valorar la validez científica o estimar la fiabilidad del método en la práctica. En efecto, la comparación de los estudios sugiere que, debido a su diseño, muchos estudios citados con frecuencia subestiman seriamente la tasa de falsos positivos.

Hasta la fecha, hay solo un estudio que fue apropiadamente diseñado para evaluar la validez de los fundamentos y estimar la fiabilidad (el estudio del Laboratorio Ames). Es importante señalar que el estudio fue dirigido por un grupo independiente, no afiliado a un laboratorio de criminalística. Aunque el informe está disponible en la web, no ha sido todavía sometido a revisión por pares y a publicación.

Los criterios científicos para la validez de los fundamentos requieren estudios apropiadamente diseñados *por más de un grupo* para asegurar la reproducibilidad. Dado que solo ha habido un estudio apropiadamente diseñado, la evidencia actual no cumple los criterios científicos para la validez de los fundamentos[334]. Hay, de este

[334] El DOJ pidió al PCAST que revisara un reciente artículo, publicado en julio de 2016, y juzgara si constituye un estudio de caja negra adicional apropiadamente diseñado sobre el análisis de armas de fuego (es decir, la capacidad de asociar munición con un arma *en particular*). El PCAST revisó cuidadosamente el artículo, incluyendo una entrevista con los tres autores sobre el diseño del estudio. Smith, T.P., Smith, G.A. y J.B. Snipes, 2016: 939-946.

El artículo incluye un novedoso y complejo diseño que es diferente a cualquier estudio previo. Brevemente, el diseño del estudio fue el siguiente: (1) se dispararon seis tipos diferentes de munición desde 8 pistolas del calibre 40 de cuatro fabricantes (2 Taurus, 2 Sig Sauer, 2 Smith and Wesson y 2 Glock) que habían estado en uso en la población en general y obtenido por el Departamento de policía de San Francisco; (2) se crearon kits de pruebas seleccionando aleatoriamente 12 muestras (proyectiles o casquillos); (3) a 31 analistas se les dijo que la munición fue recuperada de una única escena del crimen y se les pidió que preparasen notas describiendo sus conclusiones sobre qué conjunto de muestras habían sido disparadas con la misma arma; y (4) basándose en las notas de cada analista, los autores buscaron recrear la ruta lógica de las comparaciones seguida por cada analista y calcular las estadísticas basadas en este número inferido de comparaciones realizadas por cada analista.

Aunque es interesante, el artículo no es claramente un estudio de caja negra para valorar la fiabilidad de los análisis de armas de fuego para asociar munición con un arma en particular y sus resultados no pueden compararse con estudios previos. Específicamente: (1) El estudio emplea un diseño de *comparación dentro de un conjunto* (comparaciones interdependientes dentro de un conjunto) en vez de un diseño de *caja negra* (muchas comparaciones independientes); (2) el estudio involucra a un pequeño número de analistas; (3) la pregunta central con respecto al análisis de armas de fuego es si los analistas pueden asociar munición consumida con un arma *en particular*, no simplemente con una particular *marca* de arma. Para responder a este interrogante, los estudios han de valorar el desempeño del analista con munición disparada de diferentes armas de la *misma marca* («comparaciones dentro de la misma clase») en lugar de con armas de *diferentes marcas* («comparación entre clases»); la última comparación es mucho más simple porque las armas de distintas marcas producen marcas con distintas características «de clase» (debido al diseño del arma), mientras que las armas de la misma marca deben distinguirse por características «adquiridas aleatoriamente» por cada arma (adquiridas durante el estriado o por el uso). De acuerdo con esto, los estudios previos utilizaron solo comparaciones intra-clase. Por contraste, el reciente estudio consiste en una mezcla de comparaciones intra-clase contra entre-clases, siendo una sustancial mayoría las comparaciones más simples entre-clases. Para estimar la tasa de falsos positivos para comparaciones *intra-clase* (la cantidad relevante), necesitamos saber el número de pruebas inde-

modo, necesidad de estudios de caja negra adicionales, apropiadamente diseñados, para proporcionar estimaciones de su fiabilidad.

HALLAZGO 6:
ANÁLISIS DE ARMAS DE FUEGO

Validez de los fundamentos. El PCAST encuentra que el análisis de armas de fuego no cumple en la actualidad los criterios para la validez de los fundamentos, porque solo hay un único estudio apropiadamente diseñado para medir la validez y estimar la fiabilidad. Los criterios científicos para la validez de los fundamentos requieren más de un estudio de ese tipo para demostrar la reproducibilidad.

Pertenece a los tribunales decidir si el análisis de armas de fuego debe considerarse admisible basándose en la actual evidencia.

Si el análisis de armas de fuego se permite en los tribunales, los criterios científicos para la validez de su aplicación requieren que se informe claramente sobre las tasas de error vistas en estudios de caja negra apropiadamente diseñados (estimadas en 1 error cada 66 comparaciones, con un límite superior del 95% de confianza que se corresponde con 1 error cada 46 comparaciones, en el único estudio realizado hasta la fecha).

Validez en la aplicación. Si el análisis de armas de fuego es permitido en los tribunales, la validez en su aplicación requiere, desde el punto de vista científico, que el analista:

(1) haya superado una prueba de aptitud rigurosa sobre un gran número de problemas de prueba que permitan evaluar su capacidad y rendimiento, y que revele los resultados de las pruebas de aptitud; y

(2) revele si, cuando llevó a cabo la examinación, era consciente de algunos otros hechos del caso que pudieran influir en la conclusión.

5.5.8. El camino a seguir

Se necesitan continuos esfuerzos para mejorar el estado de los análisis de armas de fuego —y esos esfuerzos aportarán claramente dividendos para el sistema de justicia penal—.

---

pendientes que conllevan comparaciones intra-clase de distintas fuentes que den como resultado examinaciones concluyentes (identificación o eliminación). El artículo no distingue entre comparaciones intra-clase de entre-clases y los autores reconocieron que no hicieron esos análisis.

Los comentarios del PCAST no tratan de hacer una crítica del artículo reciente, que es un proyecto de investigación novedoso y valioso. Simplemente responden a la pregunta específica del DOJ: el reciente artículo no representa un estudio de caja negra adecuado para valorar la validez científica o estimar la precisión de los analistas para asociar munición con un arma *en particular*.

Una dirección es la de continuar mejorando el análisis de armas de fuego como método subjetivo. Con solo un estudio de caja negra hasta la fecha, hay necesidad de estudios de caja negra adicionales basados en el diseño de estudio de caja negra del Laboratorio Ames. Como se señaló anteriormente, los estudios deben ser diseñados y dirigidos en conjunto con terceras partes que no tengan interés en los resultados (tales como el Laboratorio Ames o centros de investigación como el *Center for Statistics and Applications in Forensic Evidence* [*CSAFE]* [Centro de estadística y aplicaciones en pruebas forenses]). Hay necesidad de pruebas de aptitud más rigurosas, utilizando problemas que sean apropiadamente difíciles y que puedan ser públicamente accesibles después de las pruebas.

Una segunda —y más importante— dirección es (al igual que con los análisis de huellas latentes) convertir los análisis de armas de fuego de un método subjetivo a un método objetivo.

Esto implicaría el desarrollo y la comprobación de funcionamiento de los algoritmos de análisis de imagen para comparar la similitud de las marcas de herramienta sobre los proyectiles. Ya ha habido pasos alentadores hacia ese objetivo[335]. Recientes esfuerzos para caracterizar imágenes 3D de proyectiles han utilizado métodos estadísticos y de *machine learning* para construir una «firma» cuantitativa de cada proyectil que pueda ser utilizada en comparaciones entre muestras. Un artículo reciente analiza la posibilidad de utilizar métodos topográficos de superficie en balística y se sugieren enfoques para el uso de esos métodos en los exámenes de armas de fuego[336]. Los autores señalan que el desarrollo de métodos ópticos ha mejorado la velocidad y la precisión en captar la topografía de superficie, conduciendo a una cuantificación mejorada de los grados de similitud.

En un estudio reciente, los investigadores utilizaron imágenes de un estudio anterior para desarrollar un sistema asistido por ordenador para comparar proyectiles que minimiza la intervención humana[337]. El algoritmo del grupo genera una firma cuantitativa de la imagen 3D del proyectil, compara la firma con dos o más muestras y produce una «puntuación de coincidencias» que refleja la fuerza de la coincidencia. En un conjunto de datos de prueba reducido, el algoritmo tuvo una tasa de error muy baja.

Existen esfuerzos adicionales en el sector privado que se centran en desarrollar representaciones precisas de casquillos en alta resolución para mejorar la precisión y

---

[335] Por ejemplo, un estudio reciente utilizó datos de microscopía confocal 3-D de munición para desarrollar una métrica de similitud para comparar imágenes. Tras realizar todas las comparaciones por pares entre un total de 90 casquillos disparados desde 10 pistolas con corredera, los autores encontraron que la distribución de la métrica para los pares de las mismas armas no se solapaba con la distribución de la métrica para los pares de distintas armas. Aunque es un estudio reducido, es alentador. Weller, Zheng, Thompson, et al., 2012: 912-917.

[336] Vorburger, Song y Petraco, 2016, 013002.

[337] Hare, Hofmann y Carriquiry, «Automatic matching of bullet lands.» Unpublished paper, available at: arxiv.org/pdf/1601.05788v2.pdf.

permitir funciones de similitud de mayor calidad que mejoren y asignen niveles de confianza de las coincidencias en las búsquedas en bases de datos. La actual base de datos NIBIN utiliza tecnología más antigua (no 3D) y no proporciona una función de puntuación o una asignación de confianza a cada candidato a una coincidencia. Se ha sugerido que podría utilizarse una función de puntuación para una verificación ciega por parte de analistas humanos.

Dado el enorme progreso del análisis de imagen en otros campos en la pasada década, creemos que es probable que el análisis de las armas de fuego completamente automático sea posible en un futuro próximo. Sin embargo, los esfuerzos se ven actualmente obstaculizados por la falta de acceso a bases de datos realmente grandes y complejas que puedan utilizarse para continuar con el desarrollo de estos métodos y validar las propuestas iniciales.

El NIST, en coordinación con el laboratorio del FBI, debería desempeñar un papel de liderazgo en impulsar esta transformación creando y difundiendo bases de datos apropiadamente grandes. Estos organismos también deberían proporcionar subvenciones y contratos para sustentar el trabajo y los procesos sistemáticos de evaluación de los métodos. En particular, creemos que los concursos «premiados» basados en grandes colecciones de imágenes accesibles al público[338] podrían atraer un notable interés del mundo académico y de la industria.

5.6. Análisis de huellas de calzado: identificando características

5.6.1. Metodología

El análisis de huella de calzado es un proceso que generalmente implica la comparación de un objeto conocido, tal como un zapato, con una impresión completa o parcial encontrada en la escena del crimen, para valorar la probabilidad de que el objeto sea la fuente de la impresión. El proceso se desarrolla de forma escalonada, comenzando con una comparación de «características de clase» (tales como diseño, tamaño físico y desgaste, en general) y pasando después a «identificar características» o «características aleatoriamente adquiridas (*RACs*, por las siglas en inglés de «*randomly acquired characteristics*») (tales como las marcas de un zapato causadas por cortes, muescas y roturas en el transcurso de su uso)[339].

---

[338] El 7 de julio de 2016, el NIST publicó la base de datos de investigación de código abierto *NIST Ballistics Toolmark Research Database (NBTRD)* como una base de datos de investigación de acceso libre de datos de marcas de herramientas de proyectiles y casquillos (tsapps.nist.gov/NRBTD). La base de datos contiene imágenes de microscopía de reflectancia y datos topográficos de superficie tridimensionales adquiridos por el NIST o aportados por los usuarios.

[339] Véanse: Rango de conclusiones estándares para las pruebas de examen de calzado e impresiones de neumáticos de SWGTREAD, 2013. Guía para las pruebas de examen de calzado e impresiones de

En este informe, no afrontamos la cuestión de si los analistas pueden fiablemente determinar las características de clase —por ejemplo, si una impresión de calzado concreta fue dejada por un calzado talla 12 de una marca en particular—. Aunque es importante que haya estudios que estimen la fiabilidad de los análisis de calzado dirigidos a determinar las características de clase, el PCAST no eligió poner su atención en este aspecto del examen del calzado porque la determinación de las características de clase la estimación de la frecuencia con que una característica de clase particular se presenta en el calzado o (para miembros del jurado) la comprensión de la naturaleza de las características en cuestión no son *inherentemente* un problema que suponga un reto de medición,.

En su lugar, el PCAST centró su atención en la fiabilidad de las conclusiones, basadas en *RACs*, sobre que una impresión provenga probablemente de un calzado específico. Esto es un problema mucho más difícil, porque requiere conocer con qué precisión los analistas identifican las características específicas compartidas entre un calzado y una impresión, con qué frecuencia fallan al identificar las características que los distinguirían y qué valor probatorio puede otorgarse a un *RAC* particular.

A pesar de la ausencia de estudios empíricos que midan la precisión de los analistas, las autoridades en la disciplina del examen de huella de calzado expresan su confianza en que pueden identificar la fuente de una impresión basándose en un único *RAC*.

Como se describió en un artículo de 2009 escrito por un analista del FBI y publicado en la revista *Forensic Science Communications* del FBI:

> Un analista primero determina si existe una correspondencia de características de clase entre la impresión del calzado cuestionada y el calzado conocido. Si el analista considera que no existen inconsistencias entre las características de clase, el examen progresa hacia cualquier característica identificativa en la impresión cuestionada. El analista compara estas características con cualquier característica identificativa observada en el calzado conocido. Aunque sean impredecibles en cuanto a su ocurrencia, el tamaño, la forma y la posición de esas características tienen una baja probabilidad de ocurrencia de la misma manera en diferentes calzados. De este modo, combinadas con las características de clase, incluso una sola característica identificativa constituye una prueba extremadamente poderosa en apoyo de la conclusión de la identificación[340].

Como apoyo a esta tesis, el artículo cita un manual líder en identificación de huellas de calzado:

> De acuerdo con William J. Bodziak (2000), «pueden hacerse identificaciones positivas con tan solo una característica identificativa aleatoria, pero solo si esa característica es confirmable, tiene suficiente definición, claridad y propiedades, está en la misma posición y orientación sobre la suela del calzado y, en opinión del analista experimentado, no se haría de nuevo por otra suela de calzado»[341].

---

neumáticos de SWGTREAD, 2006 y Bodziak, 2000: p 347.

[340] Smith, M.B., *The Forensic Analysis of Footwear Impression Evidence*. www.fbi.gov/about-us/lab/forensic-sciencecommunications/fsc/july2009/review/2009_07_review02.htm.

[341] Bodziak, 2000.

El artículo se refiere al modelo matemático de Stone que afirma que la probabilidad de que dos calzados compartan una misma característica identificativa es de 1 en 16.000 y la probabilidad de que compartan tres características es de 1 en 683 billones (anglosajones)[342].

Tales afirmaciones sobre «identificación» basadas en el análisis de calzado son impresionantes, pero carecen de fundamento científico.

La afirmación de Bodziak tiene dos componentes: (1) que el analista observe, consistentemente, un *RAC* demostrable en un conjunto de impresiones y (2) que el analista esté convencido de que el *RAC* no ocurriría en otro calzado. La primera parte es razonable, pero la segunda es profundamente problemática: requiere que el analista confíe en recuerdos y conjeturas sobre la frecuencia de las características.

El modelo de Stone es completamente teórico: realiza muchas suposiciones no respaldadas (sobre la frecuencia y la independencia estadística de las marcas) que no pone a prueba de ninguna manera.

Todo el proceso —desde la elección de características que deben considerarse (o ignorarse) y la determinación de la rareza— se basa enteramente en el juicio subjetivo de un analista. En tales circunstancias, es esencial que la validez del método y las estimaciones de su fiabilidad se establezcan mediante múltiples y apropiados estudios de caja negra[343].

5.6.2. Antecedentes

El informe del NRC (2009) mencionado citó algunos artículos que siembran dudas sobre si los analistas de huellas de calzado alcanzan conclusiones consistentes cuando se presentan con la misma prueba. Por ejemplo, el informe contiene una detallada explicación sobre un trabajo europeo de 1996 que presentó a los analistas frente a seis casos simulados —dos con calzado desgastado utilizado en la escena del crimen, cuatro con calzado nuevo cuyas características identificativas habían sido añadidas deliberadamente; el artículo informó de una considerable variabilidad en las respuestas[344]. El PCAST también destaca un estudio israelí que trata de dos casos en la escena del crimen y que llega a similares conclusiones[345].

Como respuesta al informe del NRC (2009), un artículo de 2013 afirmó demostrar que analistas de huellas de calzado estadounidenses y canadienses presentaban una mayor consistencia que la observada en el estudio europeo de 1996[346]. Sin embargo, este estudio difirió sustancialmente del primero porque los analistas no reali-

---

[342] Stone, 2006: 577-99.

[343] Además de los estudios de caja negra, también son valiosos los estudios de caja blanca para identificar las fuentes de los errores.

[344] Majamma e Ytti, 1996: 109-20.

[345] Shor y Weisner, 999: 380-4384.

[346] Hammer, Duffy, Fraser, et al., 2013: 205-18.

zaron los exámenes por sí mismos. Por ejemplo, las fotografías fueron previamente anotadas para señalar todas las características relevantes para la comparación —es decir, a los analistas no se les pidió que identificaran las características—[347]. De este modo, el estudio, en virtud de su diseño, no puede abordar la consistencia en el proceso de examinación.

Además, el tema fundamental no es la *consistencia* (si los analistas dieron la *misma* respuesta) sino la *precisión* (si dieron la respuesta *correcta*). La precisión puede ser evaluada solamente con grandes estudios de caja negra apropiadamente diseñados.

5.6.3. Estudios de validez científica y fiabilidad

El PCAST no pudo encontrar estudios de caja negra apropiadamente diseñados para establecer la validez de los fundamentos de la identificación basada en análisis de huellas de calzado.

HALLAZGO 7:
ANÁLISIS DE HUELLAS DE CALZADO

Validez de los fundamentos. El PCAST encuentra que no existen estudios empíricos apropiados que sostengan la validez de los fundamentos de los análisis de huellas de calzado para asociar impresiones de esas huellas con un calzado en particular con base en marcas identificativas específicas (en ocasiones denominadas "características adquiridas aleatoriamente"). Tales conclusiones no están respaldadas por ninguna evidencia significativa o estimaciones de su precisión y, por ello, no son científicamente válidas.

El PCAST no ha evaluado la validez de los fundamentos de los análisis de las huellas de calzado para identificar características de clase (por ejemplo, la talla del calzado o la marca).

Mostrando acuerdo con nuestra conclusión, el Subcomité de huellas de calzado y de neumáticos de la OSAC, recientemente identificó la necesidad de realizar estudios de caja negra y de caja blanca sobre la fiabilidad de los analistas —citándola como una «brecha importante en el conocimiento actual» en la que «no se están llevando a cabo una investigación en la actualidad o se hace limitadamente—»[348].

[347] El artículo afirma que «Todas las características y observaciones que fueron consideradas por los analistas durante la comparación fueron claramente identificadas y etiquetadas en cada impresión».

[348] Véase: www.nist.gov/forensics/osac/upload/SAC-Phy-Footwear-Tire-Sub-R-D-001-Examiner-ReliabilityStudy_Revision_Feb_2016.pdf (éúltimo acceso 12 de mayo de 2016).

5.6.4. El camino a seguir

En contraste con los análisis de huellas latentes y de armas de fuego, existe escasa investigación sobre la que fundamentar las conclusiones que buscan asociar una impresión de calzado a un calzado en particular (conclusiones de identificación).

Se necesitan nuevos enfoques para desarrollar paradigmas. Como paso inicial, el laboratorio del FBI está involucrado en un estudio en el que se examinan un conjunto de 700 botas similares que fueron utilizadas por los cadetes de agentes especiales del FBI durante las 16 semanas de su programa de formación. El estudio tiene como objetivo evaluar si los *RACs* se observan en calzados de diferentes individuos. Si bien tales estudios sobre «unicidad» (es decir, demostraciones de que muchos objetos tienen características distintivas) no pueden establecer la validez de los fundamentos (véase p. 51), las impresiones generadas a partir del calzado podrían proporcionar un conjunto de datos iniciales para (1) un estudio de caja negra piloto y (2) una base de datos piloto de frecuencias de características. Es importante destacar que el NIST está comenzando un estudio para determinar si es posible cuantificar el proceso de examinación de las huellas de calzado o, como mínimo, algún aspecto del proceso, en un esfuerzo por incrementar la objetividad de los análisis de huellas de calzado.

Separadamente, deben llevarse a cabo evaluaciones sobre la precisión y fiabilidad de la determinación de características de clase, un tema que no se ha abordado en este informe.

5.7. Análisis de cabellos

El análisis de cabellos es un proceso mediante el cual los analistas comparan propiedades microscópicas del cabello para determinar si una persona en particular pudiera ser la fuente de un cabello determinado. Mientras el PCAST estaba completando este informe, el DOJ publicó, para que pudieran hacerse comentarios, directrices sobre el testimonio en exámenes de cabello que incluían documentos de respaldo que abordan la validez y la fiabilidad de la disciplina[349]. Aunque el PCAST no ha llevado a cabo una revisión integral de la disciplina, realizó una revisión del documento de apoyo con el fin de arrojar más luz sobre los estándares para conducir una evaluación científica de una disciplina forense de comparación de características.

El documento de apoyo afirma que «la comparación de cabellos microscópica ha demostrado ser una metodología científica válida y fiable», aunque resalta que «la

[349] Véanse: Propuesta del Departamento de Justicia para uniformizar el lenguaje en el testimonio y en los informes de la disciplina del examen forense de cabellos, disponible en: www.justice.gov/dag/file/877736/download y documentación en apoyo de la propuesta del Departamento de Justicia para uniformizar el lenguaje en el testimonio y en los tnformes de la disciplina del examen forense de cabellos, disponible en: www.justice.gov/dag/file/877741/download.

comparación de cabellos microscópica por sí sola no puede conducir a una identificación personal y es crucial que esta limitación se transmita, tanto en el informe escrito como en el testimonio en juicio oral».

5.7.1. Estudios de los fundamentos del examen microscópico de cabellos

En apoyo a su conclusión de que el examen de cabellos es válido y fiable, el documento de apoyo del DOJ diserta sobre cinco estudios de comparación de cabello humano. El principal apoyo es una serie de tres estudios realizados por Gaudette en 1974, 1976 y 1978[350]. En 1974 y 1976, los estudios se centraron, respectivamente, sobre cabello de la cabeza y vello púbico. Dado que el diseño y los resultados son similares, nos centramos en el estudio del cabello de la cabeza.

El documento de apoyo del DOJ afirma que «del total de 370.230 intercomparaciones realizadas en los estudios del cabello de la cabeza, solo nueve pares de cabellos no pudieron distinguirse» —correspondiéndose con una tasa de falsos positivos de 1 en 40.000—. Más específicamente, el diseño de este estudio de 1974 fue como sigue: un único analista (1) puntuó entre 6 y 11 cabellos de cabeza de cada uno de 100 individuos (con un total de 861 cabellos) con respecto de 23 categorías distintas (con un total de 96 valores posibles); (2) comparó los cabellos de *diferentes* individuos para identificar pares de cabellos con menos de 4 diferencias; y (3) comparó estos pares de cabellos microscópicamente para ver si podían distinguirse.

El documento de apoyo del DOJ no aclara que estos estudios fueron fuertemente criticados por otros científicos por metodología defectuosa[351]. La crítica más seria fue que Gaudette solo comparó cabellos de *diferentes* individuos, y no se fijó en los cabellos de un *mismo* individuo. Como señalan en un artículo de 1990 dos autores de la *Hair and Fibre Unit of the Royal Canadian Mounted Police Forensic Laboratory* (Unidad de cabello y fibras del laboratorio forense de la policía montada del Canadá) (así como en otros artículos), la aparentemente baja tasa de falsos positivos podría ser el resultado de un sesgo del analista, es decir, el analista explícitamente conocía que todos los cabellos examinados procedían de *diferentes* individuos y, por tanto, pudo estar inclinado, consciente o inconscientemente, a buscar diferencias[352]. En resumen, no se puede valorar apropiadamente la tasa de falsos positivos de un método sin, simultáneamente, valorar su tasa de *verdaderos* positivos (sensibilidad). En el artículo de 1990, los autores utilizaron un diseño de estudio similar, pero se empleó a *dos* analistas que examinaron *todos* los pares de cabellos. Encontraron no repetibilidad en

---

[350] Gaudette y Keeping, 1974: 599-606; Gaudette, 1976: 514-517; Gaudette, 1978: 758–763.

[351] Wickenheiser y Hepworth, 1990: 1323-29. Véanse también Barnett y Ogle, 1982: 272–278; Gaudette, 1982: 279-289.

[352] Además, la inconsistencia en las características de puntuación añadiría ruido aleatorio a cualquier estructura de los datos (por ejemplo, correlaciones entre las características) y, por tanto, disminuiría la frecuencia de las coincidencias producidas por casualidad.

cada uno de los analistas («cada analista tuvo una considerable variación día a día en la clasificación de las características de los cabellos») y no reproducibiliad entre los analistas («en muchos casos, los analistas clasificaron los mismos cabellos de forma diferente»). En particular, observaron que, si bien los analistas no encontraron coincidencias entre los cabellos de diferentes individuos, tampoco encontraron casi ninguna coincidencia entre los cabellos de una *misma* persona. De 15 parejas de cabellos de una misma fuente que los autores determinaron que *deberían* haber sido declaradas coincidentes, solo dos fueron correctamente clasificados por ambos analistas.

En el estudio de Gaudette de 1978, el autor dio un cabello diferente a cada uno de tres aspirantes a analista en formación, que habían completado un año de entrenamiento, y se les pidió que identificaran cualquier muestra coincidente en un conjunto de referencia de 100 cabellos (que, sin que los aspirantes a analista en formación lo supieran, procedían de 100 personas diferentes, incluida las fuentes de los cabellos). Los tres aspirantes a analista en formación informaron de 1, 1 y 4 coincidencias, lo cual supone 3 respuestas correctas y 3 incorrectas. De las coincidencias declaradas, el 50% eran falsas asociaciones positivas. Entre las 300 comparaciones en total, la tasa general de falsos positivos fue del 1%, que es 400 veces más alta que la tasa estimada en el estudio de 1974.

Curiosamente, observamos que el documento de apoyo del DOJ informa erróneamente de los resultados del estudio, porque dice que el tercer analista en formación cometió solo un error en lugar de 3 errores. La explicación de esta discrepancia se encuentra en un pasaje notablemente franco del texto, que ilustra sobre la necesidad de emplear protocolos rigurosos en la evaluación de los resultados de los experimentos:

> Dos aspirantes a analista en formación identificaron correctamente un cabello y solo uno como similar al estándar. Al principio, el tercer aspirante a analista en formación concluyó que había cuatro cabellos similares al estándar. Tras un examen más detenido y una consulta con los otros aspirantes a analista, fue capaz de identificar fácilmente una de sus elecciones como incorrecta. Sin embargo, aún estaba convencido de que había tres cabellos similares al estándar, el correcto y otros dos. El examen del autor trajo la opinión de que uno de esos dos podía ser eliminado, pero que el restante era indistinguible de los cabellos del estándar. Otro analista experimentado estudió luego los cabellos y también llegó a la conclusión de que uno de los otros dos podía ser eliminado. Esta vez, sin embargo, ¡fue el contrario del escogido por el autor![353]

La reclasificación *post-facto* de los errores generalmente no es aconsejable en estudios relativos a la validez y la fiabilidad.

Los otros dos estudios de cabellos humanos que se tratan en el documento de apoyo del DOJ son también problemáticos. Un artículo de 1983 se refirió a muestras de cabello de 100 individuos, clasificados en tres grupos raciales[354]. Después de que la autora hubiera estudiado exhaustivamente los cabellos, pidió a una tercera parte

[353] Gaudette, 1978: 758–763.
[354] Strauss, 1983: 15-29.

neutral que configurase siete problemas difíciles y «ciegos» para ella, seleccionando 10 cabellos cuestionados y 10 conocidos (entre-grupos en 3 casos e intra-grupo en 4 casos)[355]. Los resultados consisten en una sola frase en la que la autora simplemente afirma que actuó con una «precisión del 100%». El rendimiento autoinformado en una prueba generalmente no se considera como una metodología científica apropiada.

Un artículo de 1984 estudió cabellos de 17 pares de gemelos (9 mellizos, 6 gemelos univitelinos y 2 de cigosidad desconocida) y un conjunto de trillizos idénticos[356]. Curiosamente, los cabellos de gemelos idénticos no mostraban mayor similitud que los de gemelos fraternales. En la única prueba diseñada para simular casos reales, a dos analistas se les proporcionó siete problemas difíciles, cada uno consistente en comparar un cabello cuestionado con entre cinco y diez cabellos conocidos. La tasa de falsos positivos fue de 1 cada 12, que es aproximadamente 3.300 veces más alta que la del estudio de cabellos de individuos no emparentados de Gaudette en 1974[357].

El PCAST, tomando en cuenta su metodología y sus resultados, encuentra que los artículos descritos en el documento de apoyo del DOJ no proporcionan una base científica para concluir que el examen microscópico del cabello sea un proceso válido y fiable.

Después de describir los artículos científicos, el documento del DOJ pasa a discutir las conclusiones que pueden aducirse sobre la comparación de cabellos:

> Estos estudios han demostrado también que la comparación microscópica de los cabellos por sí sola no puede conducir a una identificación personal y que es crucial que esta limitación sea transmitida tanto en el informe escrito como en el testimonio.
>
> La ciencia de la comparación de cabellos por microscopía reconoce que las características microscópicas exhibidas en un cabello cuestionado pueden ser abarcadas por la gama de características exhibidas por muestras de cabello conocidas de más de una persona. Si un cabello cuestionado se asocia con una muestra de cabello conocido que realmente no es la fuente, eso no significa que la asociación microscópica del cabello sea errónea. Más bien, resalta la limitación de la ciencia en el sentido de que existe un grupo desconocido de personas que podrían haber contribuido como fuentes del cabello cuestionado. Sin embargo, los estudios no han determinado el número de personas que comparten cabellos con las mismas o similares características.

Este pasaje viola principios científicos fundamentales de dos maneras importantes. El primer problema es que utiliza el hecho de que la precisión del método no sea *perfecta* para descartar la necesidad de conocer la precisión del método *por completo*. De acuerdo con el documento de respaldo, no es un «error» sino simplemente una «limitación de la ciencia» que un analista asocie un cabello con un individuo que

---

[355] El documento de apoyo del DOJ, erróneamente informa que la prueba de microscopía comparativa consistió en comparar 100 cabellos cuestionados con 100 cabellos conocidos.

[356] Bisbing yWolner, 1984: 780-786.

[357] El documento de apoyo del DOJ describe los resultados en términos positivos: «En las siete pruebas, un analista excluyó, correctamente, 47 de 52 muestras, y un segundo analista excluyó correctamente 49 de 52 muestras». Eso no especifica si los resultados restantes son inconclusos o falsos positivos.

no era realmente la fuente del cabello. Esto es falso. Cuando un perito dice a los miembros de un jurado que el cabello encontrado en la escena del crimen es microscópicamente indistinguible del cabello del acusado, el analista y la acusación pretenden que la declaración tenga peso. Sin embargo, el documento continúa diciendo que no existe información disponible sobre la proporción de individuos que tienen similares características. Como en la sección 4 se deja claro, esto es científicamente inaceptable. Sin estimaciones apropiadas de precisión, la afirmación de un analista de que dos muestras son similares —o, incluso, indistinguibles— carece de significado científico: no tiene valor probatorio y sí posee un considerable potencial impacto perjudicial. En resumen, si el análisis científico del cabello *significa* algo, debe haber *evidencia empírica* real sobre su significado.

El segundo problema del pasaje es su implicación de que no existen pruebas empíricas relevantes sobre la precisión del análisis del cabello. De hecho, sin embargo, tales pruebas fueron realizadas por el laboratorio del FBI. Tratamos sobre este punto seguidamente.

5.7.2. Estudio del FBI comparando exámenes microscópicos de cabello y análisis de ADN

Un aspecto particularmente preocupante del documento de apoyo del DOJ es cómo trata el estudio del FBI sobre el examen de cabello que se aborda en la sección 2. En ese estudio del año 2002, miembros del FBI utilizaron análisis de ADN mitocondrial para reexaminar 170 muestras de casos anteriores en los que el laboratorio del FBI había realizado un examen microscópico del cabello. Los autores encontraron que en 9 de 80 casos (11%) en los que el Laboratorio del FBI había hallado que los cabellos eran microscópicamente indistinguibles, el análisis de ADN mostraba que los cabellos procedían en realidad de individuos *diferentes*.

El estudio del FBI del año 2002 es un hito en la historia de la ciencia forense porque fue el primer estudio que analizó sistemática y completamente una gran colección de casos anteriores para medir la frecuencia de las asociaciones de falsos positivos. Su conclusión es de enorme importancia para la ciencia forense, para la policía, para los tribunales y para los miembros del jurado: *Cuando los analistas del cabello concluyen que dos muestras de cabello son microscópicamente indistinguibles, los cabellos proceden a menudo de distintas fuentes (1 de cada 9 veces)*.

Sorprendentemente, el documento del DOJ ignora completamente este hallazgo clave. En su lugar, hace referencia al estudio del FBI para apoyar la proposición de que el análisis de ADN «puede realizarse conjuntamente con el análisis microscópico del cabello», citando «un estudio del año 2002 que indicó que, de 80 asociaciones microscópicas, aproximadamente el 88% también fue incluido mediante prueba adicional de ADN mitocondrial». El documento no reconoce que el resto de asociaciones fueron asociaciones falsas, es decir, resultados que, si se presentaran como

pruebas en un juicio contra un acusado, confundirían a los miembros del jurado sobre los orígenes de los cabellos[358].

5.7.3. Conclusión

Nuestra breve revisión pretende simplemente ilustrar sobre los potenciales escollos en las evaluaciones de la validez de los fundamentos y de la fiabilidad de un método. El PCAST es consciente de las limitaciones a las que se enfrenta el DOJ para llevar a cabo evaluaciones científicas sobre la validez y fiabilidad de los métodos forenses, ya que las evaluaciones críticas del Departamento de Justicia podrían interpretarse como admisiones que podrían utilizarse para impugnar condenas pasadas o procesamientos actuales.

Estas cuestiones ponen de relieve por qué es importante que las evaluaciones sobre la validez científica y la fiabilidad sean realizadas por una agencia científica que no participe, por sí misma, en la aplicación de la ciencia forense dentro del sistema jurídico (véase Sección 6.1).

También subrayan por qué es importante que la información cuantitativa sobre la fiabilidad de los métodos (por ejemplo, la frecuencia de falsas asociaciones en los análisis de cabellos) se indique claramente en el testimonio de los analistas. Volveremos a esta cuestión en la sección 8, en el que consideraremos las directrices propuestas por el Departamento de Justicia que impedirían a los analistas dar información sobre el peso estadístico o la probabilidad de una conclusión sobre si un cabello cuestionado procede de una fuente en particular.

5.8. Aplicación a métodos adicionales

Aunque hemos emprendido detalladas evaluaciones de sólo seis métodos específicos e incluido una discusión sobre el séptimo método, el análisis básico puede aplicarse para valorar la validez de los fundamentos de cualquier método forense de comparación de características —incluyendo disciplinas forenses tradicionales (tales como el examen de documentos) así como métodos pendientes de desarrollo (tales como el análisis microbiano o de patrones de navegación por internet)—.

Queremos resaltar que la evaluación de la validez científica se fundamenta en la evidencia científica disponible en un momento dado. Algunos métodos cuya validez en sus fundamentos no se haya demostrado, pueden serlo en última instancia, aun-

---

[358] En una nota al pie de página, el documento también se esforzó en resaltar que el artículo no podía considerarse como el que proporcionaba una estimación de la *tasa de falsos positivos* en comparación microscópica de cabellos, porque no contenía datos sobre el número de comparaciones de fuentes diferentes que los analistas excluyeron correctamente. Aunque esta afirmación es correcta, es engañosa —porque el artículo proporciona una estimación de una cantidad mucho más importante— concretamente la frecuencia de falsas asociaciones acaecidas en casos reales.

que es posible que requieran modificaciones significativas para lograr ese objetivo. Otros métodos quizá no sean recuperables, como sucedió con el caso del análisis composicional del plomo de los proyectiles o es probable que tampoco lo sean las marcas de mordedura. Sin embargo, otros métodos pueden ser subsumidos por métodos diferentes, pero más fiables, al igual que ocurre con el análisis de ADN que ha reemplazado a otros métodos en muchos casos.

5.9. Conclusión

Como se deja claro en la sección anterior, históricamente muchos métodos forenses de comparación de características se han *asumido* como válidos en sus fundamentos, en lugar de haber sido *establecidos* como tales basándose en evidencia empírica apropiada. Solo dentro de la pasada década, la comunidad de ciencias forenses empezó a reconocer la necesidad de probar empíricamente si sus métodos específicos cumplían los criterios de validez científica. Solo en los últimos cinco años, por ejemplo, se han realizado estudios apropiados para establecer la validez de los fundamentos y medir la fiabilidad de los análisis de huellas dactilares latentes. Para la mayoría de los métodos subjetivos no hay estudios apropiados de caja negra, con el resultado de que no hay evidencia apropiada de la validez de sus fundamentos o estimaciones de su fiabilidad.

El análisis científico y los hallazgos de las Secciones 4 y 5 van dirigidos a ayudar a los actores relevantes sobre cómo garantizar la validez científica, tanto para las tecnologías existentes como para las pendientes de desarrollo.

El PCAST espera que algunos métodos forenses de comparación de características puedan ser rechazados por los tribunales como inadmisibles porque carecen de pruebas adecuadas de validez científica. Resaltamos que las decisiones de excluír métodos no fiables históricamente ha ayudado a impulsar mejoras importantes en la ciencia forense —como sucedió en los primeros días de la prueba de ADN— con el resultado de que algunos métodos llegan a establecerse como científicamente válidos (posiblemente tras una revisión) y otros son descartados.

En los Seccións restantes, ofrecemos recomendaciones sobre acciones específicas que podrían llevarse a cabo por el Gobierno Federal —incluidas las agencias científicas (NIST y OSTP), el Laboratorio del FBI, el Fiscal General y el Poder Judicial Federal— para garantizar la validez científica y la fiabilidad de los métodos forenses de comparación de características y promover su uso más riguroso en las salas de los tribunales.

6. Acciones para asegurar la validez científica en la ciencia forense: recomendaciones al NIST y a la OSTP

Basados en los hallazgos científicos de las Secciones 4 y 5, el PCAST ha identificado acciones que creemos deberían ser tomadas por las agencias científicas federales,

específicamente el NIST y la OSTP, para asegurar la validez científica de los métodos de comparación de características.

6.1. El papel del NIST en la evaluación continuada de la validez de los fundamentos

Existe la necesidad urgente de una evaluación continua de la validez de los fundamentos de métodos importantes, para proporcionar orientación a los tribunales, al DOJ y a la comunidad forense. Deben acometerse evaluaciones tanto de metodologías existentes que aún no hayan cumplido los estándares científicos para la validez de sus fundamentos como de nuevas metodologías que estén siendo desarrolladas o lo estarán en años venideros. Para asegurar que las evaluaciones científicas no estén sesgadas y sean independientes, tales evaluaciones deben estar claramente dirigidas por una agencia científica sin interés por los resultados[359].

Esta responsabilidad debe recaer en el NIST. El NIST es el laboratorio metrológico líder mundial, con una larga y distinguida historia en la ciencia y tecnología de la medición. Tiene enorme experiencia en el diseño y puesta en práctica de estudios de validación, así como en valorar la validez de los fundamentos y la fiabilidad de técnicas y prácticas de laboratorio. La misión del NIST de promover la ciencia, la tecnología y los estándares de la medición se ha ampliado, a partir de los estándares tradicionales de la medición física, para dar respuesta a otras muchas necesidades sociales importantes, incluidas las de las ciencias forenses, en la que el NIST tiene programas de calado[360]. Como se dijo anteriormente, el NIST ha comenzado a liderar una serie de esfuerzos importantes para fortalecer las ciencias forenses, incluyendo sus funciones con el NCFS y OSAC.

El PCAST recomienda que al NIST se le asigne la responsabilidad de preparar un informe anual que evalúe la validez de los fundamentos de los métodos forenses de comparación de características clave, basados en estudios empíricos publicados y disponibles. Estas evaluaciones deben llevarse a cabo bajo los auspicios del NIST, con aportaciones de analistas adicionales, ajenos a la ciencia forense, que se consideren necesarios, y supervisados por un panel de revisión apropiado. Los informes deben seguir, como mínimo, las líneas sobre los métodos expuestas en este informe, actualizándolas cuando sea apropiado. Nuestra intención no es que el NIST tenga un papel de regulador formal con respecto a la ciencia forense, sino más bien que las evaluaciones del NIST ayuden para informar a los tribunales, el DOJ y a la comunidad de ciencia forense.

---

[359] Por ejemplo, las agencias que apliquen métodos de comparación de características dentro del sistema jurídico tienen un claro interés en los resultados de tales evaluaciones.

[360] Véase: www.nist.gov/forensics.

No esperamos que el NIST asuma la responsabilidad de *realizar* los estudios de validación necesarios. Sin embargo, el NIST debería asesorar sobre el diseño y la ejecución de dichos estudios. El NIST podría realizar algunos estudios mediante su propio programa de investigación interno y a través del CSAFE. Sin embargo, la mayoría de los estudios probablemente se llevarán a cabo por otros grupos —como los Centros de investigación de cooperación entre la industria y la Universidad previstos por la NSF, el laboratorio del FBI, los laboratorios nacionales de los Estados Unidos, otras agencias federales, los laboratorios estatales y los académicos—.

Resaltamos que el NCFS ha respaldado recientemente la necesidad de una revisión científica independiente de los métodos de la ciencia forense. Un documento aprobado abrumadoramente por la comisión en junio de 2016, afirma que «todas las metodologías de la ciencia forense deben ser evaluadas por un organismo científico independiente para caracterizar sus capacidades y limitaciones, con el fin de responder de una manera precisa y fiable a una clara y bien definida cuestión forense» y que el «Instituto Nacional de Estándares y Tecnología (NIST), debe asumir el papel de evaluador científico independiente dentro del sistema de justicia para este propósito»[361].

Por último, creemos que el estado de la ciencia forense mejoraría si se publicasen artículos sobre la validez de los fundamentos de los métodos de comparación de características forenses en revistas científicas líderes en lugar de en revistas de ciencias forenses, donde, debido a las debilidades en la cultura de la investigación de la comunidad de ciencia forense abordadas en este informe, los estándares de revisión por pares son menos rigurosos. Es encomiable que los científicos del FBI publicaran su estudio de caja negra de huellas dactilares latentes en *Proceedings of the National Academy of Sciences.* Sugerimos que el NIST explore, con una o más revistas científicas líderes, la posibilidad de crear un proceso de revisión rigurosa y publicación online de estudios importantes sobre la validez de los fundamentos en las ciencias forenses. Las revistas apropiadas podrían incluir la revista *Metrología*, una revista internacional líder en metrología pura y aplicada, y *Proceedings of the National Academy of Sciences*.

6.2. Acelerar el desarrollo de los métodos objetivos

Como se describe a lo largo del informe, los métodos objetivos son generalmente preferibles a los métodos subjetivos. Las razones incluidas en esa preferencia son una mayor precisión, mayor eficiencia, menor riesgo de error humano, menor riesgo de sesgos cognitivos y mayor facilidad para establecer la validez de los fundamentos y

[361] Opiniones de la Comisión: Technical Merit Evaluation of Forensic Science Methods and Practices (Evaluación de la Calidad Técnica de los Métodos y Prácticas de la Ciencia Forense). www.justice.gov/ncfs/file/881796/download.

estimar la fiabilidad. Donde sea posible, deberían realizarse vigorosos esfuerzos para transformar los métodos subjetivos en métodos objetivos.

Hay dos métodos de comparación de características forenses —los análisis de huellas dactilares latentes y los de armas de fuego— que están listos para esa transformación. Como se trató en el anterior Sección, existen fuertes razones para creer que tales métodos pueden convertirse en objetivos a través del análisis de imagen automatizado. Además, el análisis de mezclas complejas de ADN se ha convertido recientemente en un método objetivo válido en sus fundamentos para un rango limitado de mezclas, aunque será necesario trabajo adicional para expandir los límites del rango.

El NIST, junto al Laboratorio del FBI, debe desempeñar un papel de liderazgo en impulsar esa transformación, mediante (1) la creación y difusión de grandes bases de datos que sustenten el desarrollo y la prueba de los métodos tanto por parte de las empresas como por académicos, (2) subvenciones y ayudas a la contratación, y (3) procesos de patrocinio, como concursos mediando algún premio, para la evaluación de los métodos.

6.3. Mejorar la organización de los Comités de Áreas Científicas

La creación por el NIST del OSAC fue un paso importante para el fortalecimiento de la práctica de la ciencia forense. El diseño organizativo —que alberga a todas las comunidades de un área temática bajo una única estructura y alienta la comunicación y coordinación interdisciplinar— es una mejora significativa respecto a los anteriores *Scientific Working Groups (SWGs)* (Grupos de trabajo científicos), que funcionaban menos formalmente como comités independientes.

Sin embargo, las lecciones iniciales tras los primeros años de su funcionamiento han revelado algunas deficiencias importantes. Entre los miembros de OSAC hay relativamente pocos científicos independientes: está dominado por profesionales forenses que suman más de dos tercios de sus miembros. Del mismo modo, tiene pocos estadísticos independientes: mientras que prácticamente todos los estándares y directrices evaluados por este organismo necesitan la consideración de principios estadísticos, solo hay 14 estadísticos repartidos por los cuatro *Scientific Area Committees* (Comités de áreas científicas) y 23 subcomités entre los 600 miembros de OSAC.

6.3.1. Reestructurando

El PCAST concluye que el OSAC carece de suficiente experiencia científica y supervisora independientes para superar los graves defectos de la ciencia forense. Se necesita cierta reestructuración para garantizar que científicos y estadísticos independientes tengan más voz en el proceso de desarrollo de los estándares, un requisito

para la validez científica significativa. Lo más importante es que el OSAC debe tener un comité formal —un Comité de recursos de metrología al nivel de los otros tres Comités de recursos (el de recursos jurídicos, el de factores humanos y el de infraestructuras de calidad)—. Este Comité debería estar compuesto por científicos de laboratorio y estadísticos ajenos a la comunidad de las ciencias forenses, encargándose de revisar cada estándar y directriz que los Comités de áreas científicas recomienden para la aprobación del registro, antes de su envío para una revisión final a la *Forensic Science Standards Board* (FSSB) (Junta de estándares para la ciencia forense).

6.3.2. Disponibilidad de los estándares de OSAC

El OSAC no es un organismo formal de normalización. Revisa y evalúa los estándares relevantes para la ciencia forense desarrollados por organizaciones de desarrollo de estándares como *ASTM International*, la *National Fire Protection Association (NFPA)* (Asociación nacional de protección contra incendios) y la *International Organization for Standardization* (ISO) (Organización internacional para la estandarización), para su inclusión en los *OSAC Registries of Standards and Guidelines* (Registros de estándares y directrices de OSAC). El proceso de evaluación de OSAC incluye un periodo de comentarios públicos. OSAC, trabajando con los desarrolladores de estándares, ha dispuesto que el contenido de los estándares bajo consideración sea accesible al público durante el periodo de comentarios públicos. Una vez aprobado por OSAC, un estándar se enumera listándolo por título en un registro público mantenido por NIST. Es habitual que algunas organizaciones de desarrollo de estándares, incluida *ASTM International*, cobre una cuota por una copia con licencia de cada estándar protegido por derechos de autor y restrinja a los usuarios la distribución de esos estándares[362, 363].

El NIST recientemente negoció un acuerdo de licencia con *ASTM International* por el que mediante una tarifa se permite a los empleados del gobierno federal, estatal y local el acceso online al Comité de estándares *ASTM E30*[364]. Sin embargo, esta lista no incluye a los acusados indigentes, abogados defensores privados o el enorme número de potenciales interesados de la comunidad académica. Hasta el presente, se han negociado contratos con otras *Standards Developing Organizations (SDOs)* (Organizaciones de desarrollo de estándares) que tienen, en la actualidad, estándares

---

[362] Para una lista de estándares de ciencia forense de *ASTM*, consúltese: www.astm.org/DIGITAL_LIBRARY/COMMIT/PAGES/E30.htm,

[363] La American Academy of Forensic Sciences (AAFD*)* (Academia Norteamericana de Ciencias Forenses) llegará también a ser una Organización de Desarrollo de Estándares acreditada (SDO) y podría, en el futuro, desarrollar estándares para su revisión e inclusión en la lista por OSAC.

[364] De acuerdo con el contrato revisado, *ASTM* proporcionará acceso ilimitado a su web de todos los Estándares de Ciencia Forense del Comité E30 de *ASTM*: a miembros y afiliados de OSAC; a Laboratorios de criminalística federales, estatales y locales y de NIST; a Oficinas de Abogados Defensores públicos; Oficinas de la Fiscalía; Agencias de Seguridad Públicas; y a Médicos forenses.

bajo revisión por la OSAC. El PCAST cree que es importante que los estándares destinados a ser usados en la administración de justicia penal estén ampliamente disponibles para todos los que puedan necesitar su acceso. Es importante que los estándares estén fácilmente accesibles tanto a los acusados como a los observadores externos, que tienen un papel muy importante que desempeñar para garantizar la calidad de la justicia penal[365].

El NIST debe garantizar que el contenido de los estándares y directrices registradas por la OSAC esté libremente accesible a cualquier parte que pueda necesitarlas en relación con un caso jurídico o para su evaluación e investigación, incluyendo su alineación con las políticas relacionadas con la razonable disponibilidad de los estándares de la *Office of Management and Budget Circular A-119* (Oficina de gestión y circular presupuestaria A-119), en la *Federal Paticipation in the Development and Use of Voluntary Consensus Standards and Conformity Assessment Activities* (Participación federal en el desarrollo y uso de estándares de consenso voluntario y actividades de evaluación de conformidad), y la *Office of Federal Register* (Oficina del registro federal), *IBR Handbook* (Manual IBR) (incorporación por referencia).

6.4. Necesidad de una estrategia de I+D para la ciencia forense

El informe del NRC (2009) expuso que existe una necesidad urgente de fortalecer la ciencia forense, resaltando que «la investigación en ciencia forense no está bien sustentada y no existe una estrategia unificada para desarrollar un plan de investigación en ciencia forense entre las agencias federales»[366].

Es especialmente importante crear y sustentar una comunidad de investigación académica vibrante enraizada en la cultura científica de las universidades. Esto requerirá una financiación significativa para apoyar a los grupos de investigación académica, pero redundará en obtener dividendos para impulsar la calidad y la innovación, tanto en métodos existentes como en otros completamente nuevos.

Tanto el NIST como NSF han tomado medidas iniciales, recientemente, para ayudar a cerrar las brechas de calado existentes entre los analistas forenses y las comunidades académicas a través de centros de investigación multidisciplinares. Estos centros prometen involucrar a la más amplia comunidad investigadora en el avance de la ciencia forense, así como crear vínculos necesarios entre la comunidad de ciencias forenses y una amplia base de universidades que pueda ayudar a impulsar la investigación básica crítica.

---

[365] El PCAST no expresa opinión alguna sobre la idoneidad de las tasas de los estándares en áreas diferentes a la justicia penal.

[366] National Research Council, 2009: 78.

No obstante, como se señala en la sección 2, el nivel total de fondos federales del NIJ, NIST y NSF destinado a la comunidad académica para la investigación básica en ciencias forenses es extremadamente reducido. Se precisa una financiación sustancialmente mayor para desarrollar una comunidad de investigación sólida y sustentar el desarrollo y la evaluación de nuevas tecnologías prometedoras.

Se necesita una mejor coordinación de los esfuerzos federales en I+D en ciencias forenses, tanto dentro como fuera de su ámbito natural. Ninguna agencia, individualmente, posee la responsabilidad de asegurar que las ciencias forenses estén adecuadamente sustentadas. Se necesita una mayor coordinación entre las agencias federales y los laboratorios relevantes para garantizar que la financiación se destine a las principales prioridades y que el trabajo sea de la máxima calidad.

OSTP debe convocar a las agencias federales, laboratorios y partes interesadas relevantes para desarrollar una estrategia de investigación nacional y un plan quinquenal para asegurar que la investigación básica en apoyo de las ciencias forenses esté bien coordinada, consolidar los compromisos contraídos hasta la fecha por las agencias federales, e impulsar nuevas acciones y fondos que pudieran adoptarse para fomentar la investigación básica adicional, mejorar los métodos forenses actuales, apoyar la creación de nuevas bases de datos de investigación y supervisar la revisión y priorización periódicas de la investigación.

6.5. Recomendaciones

RECOMENDACIÓN 1.
VALORACIÓN DE LA VALIDEZ DE LOS FUNDAMENTOS

Es importante que las evaluaciones científicas de la validez de los fundamentos se realicen de forma continua para poder valorar la validez de los fundamentos de tecnologías de comparación de características forenses actuales y recientemente desarrolladas. Para asegurar que los juicios científicos no sean sesgados y sean independientes, tales evaluaciones deben estar dirigidas por un organismo científico que no tenga interés en el resultado.

(A) El Instituto Nacional de Estándares y Tecnología (NIST) debe realizar esas evaluaciones y debe emitir un informe público anual evaluando la validez de los fundamentos de los métodos clave de comparación de características.

(i) Las evaluaciones deben (a) valorar si cada método revisado ha sido adecuadamente definido, si ha sido adecuadamente establecida la validez de sus fundamentos y si el nivel estimado de precisión se basa en evidencia empírica; (b) basarse en estudios publicados en la literatura científica de laboratorios y organismos de Estados Unidos y de otros países, así como cualquier trabajo dirigido por la propia plantilla y concesionarios del NIST; (c) como mínimo, elaborar evaluaciones sobre la línea del presente informe, actualizándolas según corresponda; (d) y que se lleven a cabo bajo los auspicios del

NIST, con asesoramiento adicional de analistas ajenos a la ciencia forense cuando se considere necesario.

(ii) El NIST debe establecer un comité asesor de científicos experimentales y estadísticos ajenos a la comunidad de ciencias forenses que proporcione asesoramiento sobre las evaluaciones y debe asegurarse de que sean rigurosos e independientes. Los miembros del comité consultivo deben seleccionarse conjuntamente por el NIST y la Oficina de políticas científicas y tecnológicas.

(iii) El NIST debe priorizar aquellos métodos forenses de comparación de características que más necesiten de una evaluación, incluyendo los que están actualmente en uso y los que estén en desarrollo en etapas avanzadas, partiendo de las aportaciones del Departamento de Justicia y de la comunidad científica.

Sobre la base de sus hallazgos científicos, el PCAST hace las siguientes recomendaciones.

(iv) Cuando el NIST evalúe que un método ha sido establecido como válido en sus fundamentos, debe (a) aportar estimaciones apropiadas de tasas de error basadas en estudios sobre los fundamentos y (b) identificar cualquier problema relevante para la validez de su aplicación.

(v) Cuando el NIST evalúe que un método no ha sido establecido como válido en sus fundamentos, debe sugerir qué medidas, si las hubiere, podrían adoptarse para establecer la validez del método.

(vi) El NIST no debe tener responsabilidades regulatorias con respecto a la ciencia forense.

(vii) El NIST debe alentar a una o más revistas científicas líderes ajenas a la comunidad forense a que desarrolle mecanismos para promover la rigurosa revisión por pares y la publicación de artículos relacionados con la validez de los fundamentos de los métodos forenses de comparación de características.

(B) El Presidente debe solicitar y el Congreso debe proporcionar un aumento de crédito al NIST de (a) 4 millones de dólares para sustentar las actividades de evaluación anteriormente descritas y (b) de 10 millones de dólares para sustentar el aumento de las actividades de investigación en ciencia forense, incluyendo las mezclas complejas de ADN, las huellas dactilares latentes, el reconocimiento de voz y del hablante y la biometría facial y del iris.

RECOMENDACIÓN 2.
DESARROLLO DE MÉTODOS OBJETIVOS PARA EL ANÁLISIS DE ADN DE MUESTRAS DE MEZCLAS COMPLEJAS, ANÁLISIS DE HUELLAS DACTILARES LATENTES Y ANÁLISIS DE ARMAS DE FUEGO.

El Instituto Nacional de Estándares y Tecnología (NIST) debe desempeñar un papel líder en la transformación de tres importantes métodos de comparación de características que son actualmente subjetivos —análisis de huellas dactilares latentes, análisis de

armas de fuego y, en algunas circunstancias, análisis de ADN de mezclas complejas— en métodos objetivos.

(A) El NIST debe coordinar esos esfuerzos con el Laboratorio de la Oficina Federal de Investigaciones (FBI), el Centro de Ciencias Forenses de la Defensa, el Instituto Nacional de Justicia y otras agencias relevantes.

(B) Esos esfuerzos deben incluir (i) la creación y difusión de grandes bases de datos y materiales de pruebas (tales como mezclas de ADN complejas) que permitan el desarrollo y la prueba de los métodos, tanto por el sector privado como por académicos, (ii) subvenciones y contratos y (iii) procesos de patrocinio, tales como competiciones premiadas, para evaluar métodos.

RECOMENDACIÓN 3. MEJORA DEL PROCESO DE ORGANIZACIÓN PARA LOS COMITÉS DE ÁREAS CIENTÍFICAS

(A) El Instituto Nacional de Estándares y Tecnología (NIST) debería mejorar la organización para los Comités de Áreas Científicas (OSAC), que fue establecida para desarrollar y promulgar estándares y directrices que permitan las mejores prácticas en la comunidad de ciencia forense.

(i) El NIST debe establecer un Comité de recursos de metrología, compuesto por metrólogos, estadísticos y otros científicos ajenos a la comunidad de ciencia forense. Un representante del Comité de recursos de metrología debe prestar servicio en cada uno de los Comités de áreas científicas (SACs) para proporcionar orientación sobre la aplicación de los principios estadísticos y de medición de los estándares documentados en desarrollo.

(ii) El Comité de recursos de metrología, en su conjunto, debe revisar y aprobar o desaprobar públicamente todos los estándares propuestos por los Comités de áreas científicas antes de que sean transmitidos a la Junta Directiva de estándares de ciencia forense.

(B) El NIST debe asegurar que el contenido de los estándares y directrices registrados por la OSAC estén disponibles libremente para cualquier parte que los desee en relación con un caso jurídico, o para evaluación o investigación, incluso alineándose con las políticas relacionadas con la disponibilidad razonable de estándares en la Oficina de gestión y circular presupuestaria A-119, en la Participación federal en el desarrollo y uso de estándares de consenso voluntario y en las actividades de evaluación de conformidad y la Oficina del Registro Federal, Manual IBR (incorporación por referencia).

RECOMENDACIÓN 4.
STRATEGIA DE I+D PARA LA CIENCIA FORENSE

(A) La Oficina de política de la ciencia y de la tecnología (OSTP) debe coordinar la creación de una estrategia nacional de investigación y desarrollo de la ciencia forense. La estrategia debe abordar los planes y las necesidades de financiación para:

(i) una gran expansión y fortalecimiento de la comunidad académica que investiga en las ciencias forenses, incluyendo un aumento sustancial de la financiación tanto para la investigación como para la formación;

(ii) estudios de validez de los fundamentos de los métodos forenses de comparación de características;

(iii) mejora de los métodos forenses actuales, incluyendo la transformación de métodos subjetivos en métodos objetivos y el desarrollo de nuevos métodos forenses;

(iv) el desarrollo de bases de datos de características forenses, con las adecuadas protecciones de la privacidad, que puedan utilizarse para la investigación;

(v) salvar la brecha entre los investigadores científicos y los profesionales forenses;

(vi) supervisión y revisión periódica de la investigación en ciencias forenses.

(B) Al preparar la estrategia, OSTP debe solicitar la opinión de las agencias federales apropiadas, incluyendo especialmente el Departamento de Justicia, el Departamento de Defensa, a la Fundación Nacional de la Ciencia, y el Instituto Nacional de Estándares y Tecnología; a profesionales de la ciencia forense de ámbito federal y estatal; a investigadores científicos forenses y no forenses; y a otras partes interesadas.

7. Acciones para garantizar la validez científica en la ciencia forense: recomendaciones para el Laboratorio del FBI

Sobre la base de los hallazgos científicos de las Secciones 4 y 5, el PCAST ha identificado acciones que creemos que han de adoptarse en el Laboratorio del FBI para garantizar la validez científica de los métodos de comparación de características.

Resaltamos que el Laboratorio del FBI ha desempeñado un importante papel en los últimos años en la realización de estudios científicos de alta calidad en análisis de huellas dactilares latentes. El PCAST aplaude esos esfuerzos e insta al Laboratorio del FBI a ampliarlos.

7.1. Papel del Laboratorio del FBI

El Laboratorio del FBI es un centro de excelencia en el estado del arte de última generación que trabaja para aplicar la ciencia de vanguardia con el fin de resolver

casos y prevenir el delito. Su misión consiste en aplicar capacidades científicas y servicios técnicos para la recogida, procesado y explotación de pruebas para el Laboratorio y otras agencias de seguridad y de inteligencia debidamente constituidas en apoyo de las prioridades de investigación e inteligencia. Actualmente, el Laboratorio tiene, aproximadamente, 750 empleados y más de 300 proveedores para cumplir con el amplio alcance de esta misión.

7.1.1. Capacidades y servicios del Laboratorio

El FBI dispone de capacidades y personal especializado para responder a incidentes, recoger pruebas en su campo, llevar a cabo análisis forenses y proporcionar testimonios de analistas en el proceso judicial. El FBI apoya a los equipos de recolección de pruebas en las 56 oficinas existentes en todo EUA y Puerto Rico y cuenta con personal especializado en pruebas de naturaleza peligrosa[367*] y documentación y recopilación de datos en la escena del crimen. El Laboratorio es responsable de capacitar y proporcionar estas actividades de respuestas para el personal del FBI en los Estados Unidos[368]. El Laboratorio también gestiona el Centro de análisis de dispositivos explosivos terroristas (TEDAC), que recibió cerca de 1000 solicitudes de pruebas en el ejercicio fiscal 2015 y comunicó unos 2000 informes de inteligencia.

El laboratorio del FBI dispone de analistas forenses que realizan análisis en una serie de disciplinas que incluyen la química, criptoanálisis, ADN, análisis de armas de fuego y de marcas de herramientas, huellas latentes, documentos cuestionados y rastreo de pruebas. El Laboratorio del FBI recibió más de 3.875 solicitudes de informes y confeccionó más de 4.850 informes en el año fiscal 2015.

Además de llevar a cabo el análisis de casos en el ámbito federal, el Laboratorio proporciona apoyo a laboratorios estatales y locales y realiza pruebas en casos de ámbito estatal y local en algunas disciplinas.

7.1.2. Actividades de investigación y desarrollo

Además de sus servicios, el Laboratorio del FBI desarrolla importantes actividades de investigación y desarrollo. Esas actividades son fundamentales para proporcionar al Laboratorio las herramientas más punteras para avanzar en su misión. Un programa y una cultura de investigación sólida son también importantes para que el Laboratorio pueda mantener la excelencia y atraer a personal altamente cualificado.

Debido al ámbito expansivo y a los muchos requerimientos de sus operaciones, solo alrededor del 5% de los 100 millones de dólares del presupuesto anual del Labo-

---

367 * N.del T.: por ejemplo, los incendios y entornos NBQ.

368 El Laboratorio del FBI apoyó 162 despliegues y 168 intervenciones de respuesta, así como la impartición de 239 cursos de formación en el ejercicio fiscal 2015.

ratorio del FBI está disponible para actividades de I+D[369]. El presupuesto en I+D se extiende a través de una serie de actividades de investigación aplicada, incluyéndose los estudios de validación (para nuevos métodos o productos comerciales tales como nuevos analizadores de ADN). Para sus actividades de investigación interna, el Laboratorio depende, en gran medida, de su Programa de científicos visitantes, que atrae a aproximadamente 25 estudiantes postdoctorales, de maestrías y de licenciatura cada año. El Laboratorio se ha asociado con otros organismos gubernamentales para proporcionar más recursos a sus prioridades de investigación mediante iniciativa compartida y también ha sido capaz de ampliar los presupuestos disponibles con estudios de investigación en cuestiones cruciales de forma creciente con el paso de los años.

La serie de estudios del Laboratorio del FBI en el examen de huellas latentes es un ejemplo de la importante investigación básica que fue capaz de desarrollar de forma creciente en un periodo de cinco años. El trabajo incluye estudios de «caja negra» que evalúan la precisión y la fiabilidad de las conclusiones de los analistas con huellas latentes, así como estudios de «caja blanca» que evalúan cómo la cantidad y calidad de las características se relacionan con las decisiones de los analistas en huellas latentes. Estos estudios han dado lugar a una serie de publicaciones importantes que han ayudado a cuantificar las tasas de error en la práctica de esas pericias y evalúan la repetibilidad y reproducibilidad de las decisiones de los analistas en huellas dactilares latentes. De hecho, la conclusión del PCAST de que el análisis de huellas dactilares latentes es válido en sus fundamentos descansa fuertemente en el estudio de caja negra del FBI. En algunas otras disciplinas se están llevando a cabo similares líneas de investigación, como en el examen de armas de fuego y documentos cuestionados.

Desafortunadamente, la limitada financiación disponible para esos estudios —y para el programa de investigación intramuros en general— ha obstaculizado el progreso en la prueba de la validez de los fundamentos de los métodos de las ciencias forenses y su fortalecimiento. PCAST cree que el presupuesto para el Laboratorio del FBI debe aumentarse significativamente y que debe orientarse a fin de permitir que el presupuesto de I+D se incremente en un total de 20 millones de dólares.

7.1.3. Accesos a bases de datos

El FBI tiene también un importante papel que jugar en fomentar la investigación entre científicos externos, facilitandoles el acceso, bajo apropiadas condiciones, a grandes bases de datos forenses. La mayoría de las bases de datos que se utilizan habitualmente en los análisis forenses no son accesibles a los investigadores y la falta de acceso obstaculiza el progreso en la mejora de la ciencia forense. Por ejemplo,

[369] En 2014, el Laboratorio del FBI invirtió 10.9 millones de dólares en I+D en ciencia forense, aproximadamente la mitad con su propio presupuesto y la otra mitad con subvenciones de NIST y el Departamento de Seguridad Nacional. Véase: National Academies of Sciences, Engineering, and Medicine, 2015: 31.

sistemas de bases de datos balísticas, como el Sistema Nacional Integrado de Información Balística de la Oficina de alcohol, tabaco, armas de fuego y explosivos (NIBIN), donde los analistas en armas de fuego buscan identificar un arma de fuego o un cartucho, no pueden ser evaluados para estudiar su completitud, relevancia o calidad y el algoritmo de búsqueda utilizado para identificar potenciales coincidencias tampoco puede ser evaluado.

El sistema NGI (formalmente IAFIS)[370] que actualmente alberga más de 70 millones de entradas de huellas dactilares ampliaría enormemente los datos disponibles para su estudio; en la actualidad, solo existe una base de datos de huellas dactilares disponible al público con 258 pares de huellas latentes e impresiones decadactilares[371]. Y el sistema NDIS del FBI, que actualmente contiene más de 14 millones de perfiles de ADN de delincuentes y detenidos. El NIST ha desarrollado un inventario de todas las bases de datos forenses que son más utilizadas por las fuerzas de seguridad y científicos forenses con información sobre su accesibilidad.

Se necesitan esfuerzos sustanciales para lograr que las bases de datos forenses existentes sean más accesibles a la comunidad investigadora, garantizando la apropiada protección de la privacidad como la eliminación de información de identificación personal y las restricciones al uso de los datos.

Para algunas disciplinas como los análisis de armas de fuego y las marcas de rodadura no existen preocupaciones significativas de privacidad.

Con respecto a las huellas latentes, podrían mejorarse las preocupaciones referentes a la privacidad de diversas formas. Por ejemplo, podría evitarse esa preocupación (1) generando grandes colecciones de pares de impresiones latentes conocidas con calidad y cantidad de información variables a través del tacto y del manejo de objetos naturales en una amplia variedad de circunstancias (superficies, presión, distorsión, etc.); (2) utilizando software para generar automáticamente «*morphing transformation*» de las impresiones conocidas y las impresiones latentes; y (3) aplicar esas transformaciones a impresiones de fallecidos para crear millones de pares de impresiones latentes conocidas[372].

En ADN se han desarrollado protocolos en investigación genómica humana, lo que plantea similares o mayores preocupaciones de privacidad, para permitir el acceso a información a los investigadores de buena fe[373]. Estas políticas también deberían

---

[370] NGI estandariza la «Identificación en la Siguiente Generación» y combina múltiples sistemas de información biométrica, incluyendo IAFIS, sistemas de reconocimiento de iris y rostro, y otros.

[371] Base de Datos Especial de NIST 27A, accesible en: www.nist.gov/itl/iad/image-group/nist-special-database-27a-sd-27a.

[372] Las oficinas de medicina forense toman rutinariamente las huellas dactilares a los individuos fallecidos como parte del proceso de autopsia; estas huellas dactilares podrían coleccionarse y utilizarse para crear grandes bases de datos con fines de investigación.

[373] Podrían aplicarse aquí una serie de modelos que han sido desarrollados en el contexto de la investigación biomédica para permitir el acceso a datos sensibles, al tiempo que proporcionan una

ser viables para las bases de datos forenses de ADN. Observamos que la ley que autoriza al FBI a mantener una base de datos nacional forense de ADN contempla, explícitamente, que se permite el acceso a las muestras de ADN y a los análisis de ADN si «se elimina la información de identificación personal ... para fines de investigación, de identificación y desarrollo de protocolos»[374]. Aunque la ley no contiene una declaración explícita sobre este particular, el DOJ interpreta que la ley permite el uso para este propósito solo por parte de los organismos de justicia penal. Es reacio, a falta de aclaraciones jurídicas, incluso a un acceso controlado a otros investigadores. Este tema merece atención.

El PCAST cree que la disponibilidad de datos acelerará el desarrollo de métodos, herramientas y software que mejorarán la ciencia forense. Para las bases de datos bajo su control, el Laboratorio del FBI debe desarrollar programas para hacer que las bases de datos forenses (o subconjuntos de esas bases) sean accesibles a investigadores en condiciones que protejan la privacidad. Con respecto a bases de datos propiedad de terceros, el Laboratorio del FBI y el NIST deben trabajar con las agencias y empresas que controlan esas bases de datos para desarrollar programas que les permitan un acceso apropiado.

7.2. Recomendaciones

Sobre la base de sus hallazgos científicos, el PCAST emite la siguiente recomendación:

RECOMENDACIÓN 5.
AGENDA PARA LA AMPLIACIÓN DE LAS CIENCIAS FORENSES EN EL LABORATORIO DEL FBI

(A) Programas de investigación. El Laboratorio de la Oficina Federal de Investigación (FBI) debería emprender un vigoroso programa de investigación para mejorar la ciencia forense, sobre la base de su reciente e importante trabajo en análisis de huellas dactilares latentes. El programa debe incluir:

(i) la realización de estudios sobre la fiabilidad de los métodos de comparación de características, junto a terceros independientes sin interés en el resultado;

(ii) el desarrollo de nuevos enfoques que mejoren la fiabilidad de los métodos de comparación de características;

(iii) la ampliación de los programas de colaboración con científicos externos; y

adecuada protección de la privacidad. Podría exigirse a los investigadores la firma de un acuerdo de no divulgación (NDA) o que accedan bajo un acuerdo de uso limitado. Podría exigirse a los investigadores que accedan a los datos en un determinado modo y lugar, de forma que los datos no puedan descargarse o compartirse, o podría permitirse la descarga de forma agregada o resumida.

[374] Federal DNA Identification Act, 42 U.S.C. §14132(b)(3)(D)).

(iv) la garantía de que los científicos externos tengan acceso adecuado a los conjuntos de datos y colecciones de muestras para que puedan realizar estudios independientes.

(B) Estudios de caja negra. Basándose en su experiencia en la investigación científica forense, el Laboratorio del FBI debería ayudar en el diseño y ejecución de estudios adicionales de caja negra para métodos subjetivos, incluído los análisis de huellas dactilares latentes y de armas de fuego. Esos estudios deben ser realizados por o en conjunto con terceros independientes sin interés en el resultado.

(C) Desarrollo de métodos objetivos. El Laboratorio del FBI debe trabajar con el Instituto Nacional de Estándares y Tecnología para transformar tres importantes métodos de comparación de características que actualmente son subjetivos —análisis de huellas dactilares latentes, análisis de armas de fuego y, en algunas circunstancias, análisis de ADN de mezclas complejas—, en métodos objetivos. Estos esfuerzos deben incluir (i) la creación y difusión de grandes bases de datos que sustenten el desarrollo y las pruebas de métodos, tanto por parte de las empresas como de académicos, (ii) becas y apoyo contractual y (iii) patrocinio de competiciones premiadas para evaluar los métodos.

(D) Pruebas de aptitud. El Laboratorio del FBI debe promover un mayor rigor en las pruebas de aptitud (i) dentro de los próximos cuatro años, instituyendo pruebas de aptitud ciegas de rutina dentro del flujo de casos en su propio laboratorio; (ii) ayudando a otros laboratorios federales, estatales y locales a hacerlo igualmente; y (iii) fomentar el acceso rutinario y la evaluación utilizadas en las pruebas comerciales de aptitud.

Recomendación 5. Agenda para la ampliación de las ciencias forenses en el Laboratorio del FBI (continuación)

(E) Análisis de huellas dactilares latentes. El Laboratorio del FBI debe promover vigorosamente la adopción, por todos los laboratorios que realizan análisis de huellas dactilares latentes, de las normas que requieren un proceso de "Análisis, comparación y evaluación lineales" —en el que los analistas deben completar y documentar sus análisis de huellas dactilares latentes antes de mirar a una impresión dactilar conocida y deben, separadamente, documentar cualquier dato adicional utilizado en la evaluación y en la comparación—.

(F) Transparencia en materia de calidad en casos. El Laboratorio del FBI, así como otros laboratorios forenses federales, debe informar públicamente y regular problemas de calidad en el análisis de casos (de forma similar a las prácticas empleadas por el Instituto forense holandés, descritas en la sección 5), como un medio para mejorar la calidad y promover la transparencia.

(G) Presupuesto. El Presidente debe solicitar y el Congreso debe proporcionar un incremento de la partida presupuestaria del FBI que restaure el presupuesto del Laboratorio del FBI para las actividades de investigación en ciencias forenses partiendo de su nivel actual con 30 millones de dólares adicionales y debe evaluar la necesidad de aumentar la financiación para otras actividades de investigación en ciencias forenses en el Departamento de Justicia.

8. Acciones para garantizar la validez científica en la ciencia forense: recomendaciones al Fiscal General

Sobre la base de los hallazgos científicos referidos en las Secciones 4 y 5, el PCAST ha identificado acciones que creemos que deberían ser llevadas a efecto por el Fiscal General para asegurar la validez científica de los métodos de comparación de características y promover su uso más riguroso en las salas de audiencias.

8.1. Garantizar el uso de métodos científicamente válidos en procesos judiciales

El Gobierno Federal tiene el profundo compromiso de garantizar que los procesos judiciales penales no solo sean justos en su procedimiento, sino correctos en sus resultados —es decir, que los culpables sean condenados, mientras que los inocentes no lo sean—.

Con este fin, el Departamento de Justicia debe asegurarse que el testimonio sobre las pruebas forenses presentadas ante un tribunal sea científicamente válido. Este informe proporciona orientación al Departamento de Justicia acerca de los criterios científicos tanto para la validez de los fundamentos como para la validez de su aplicación, así como evaluaciones de seis métodos forenses específicos y una discusión sobre un séptimo. A largo plazo, el DOJ debe inspeccionar las evaluaciones en curso de los métodos forenses que deben ser realizadas por el NIST (como se describe en la sección 6).

Mientras tanto, el DOJ debería llevar a cabo una revisión de los métodos de comparación de características (más allá de los revisados en el este informe) para identificar qué métodos utilizados por el DOJ carecen de los apropiados estudios de caja negra necesarios para evaluar la validez de los fundamentos. Como consecuencia de que de tales métodos subjetivos no están presuntamente establecidos como válidos en sus fundamentos, el DOJ debe evaluar (1) si debe presentar ante los tribunales conclusiones basadas en tales métodos y (2) si deben iniciarse estudios de caja negra para evaluar esos métodos.

8.2. Revisión de las directrices propuestas recientemente por el Departamento de Justicia sobre el testimonio de los analistas

El 3 de junio de 2016, el Departamento de Justicia publicó, para recepción de comentarios, un primer conjunto de directrices propuestas, junto con documentos de apoyo, sobre un «Lenguaje Uniforme Propuesto para Testimonio e Informes» en varias ciencias forenses, incluyendo el análisis de huellas dactilares latentes y el análisis de

impresiones de calzado y neumáticos forense[375]. El 21 de julio de 2016, el DOJ publicó, para recepción de comentarios, un segundo conjunto de directrices propuestas y documentos de apoyo para diversas ciencias forenses adicionales, incluyendo análisis microscópico del cabello, ciertos tipos de análisis de ADN y otras disciplinas.

Las directrices representan un importante paso hacia adelante porque instruyen a los analistas del Departamento de Justicia a no realizar afirmaciones categóricas como que pueden identificar la fuente de una huella dactilar o de una impresión de calzado excluyendo cualquier otra fuente posible. El PCAST aplaude la intención y los esfuerzos del DOJ para lograr uniformidad y evitar testimonios inexactos en comparaciones de características.

Algunos aspectos de las directrices, sin embargo, no son científicamente apropiadas y encarnan puntos de vista heterodoxos del tipo de aquellos discutidos en la Sección 4.7. A modo de ilustración, nos centramos en las directrices para el análisis de impresiones de calzado y neumáticos y en las de análisis del cabello.

8.2.1. Análisis de impresiones de calzado y neumáticos

En el Cuadro 6 se muestran partes relevantes de las directrices para el testimonio y la redacción de informes sobre impresiones de calzado y de ruedas de neumáticos.

CUADRO 6.
EXTRACTO DE LA PROPUESTA DEL DOJ SOBRE UN LENGUAJE UNIFORME PARA TESTIMONIOS E INFORMES EN LAS DISCIPLINAS DE IMPRESIONES DE CALZADO Y DE NEUMÁTICOS FORENSES[376].

Afirmaciones aprobadas para su uso en informes de los laboratorios y en las declaraciones de los analistas con respecto al examen forense de la evidencia de huellas de calzado y de neumáticos.

Identificación

El analista puede afirmar que, en su opinión, el calzado o el neumático es la fuente de la impresión porque hay suficiente cantidad y calidad de características coincidentes como para esperar que no pudiera encontrarse la misma combinación de características repetida por esa fuente. Se trata del mayor grado de asociación entre una impresión cuestionada y una fuente conocida. La opinión requiere que la impresión cuestionada y la conocida coincidan en las características de clase y que también compartan una o más características aleatoriamente adquiridas. Esta opinión reconoce

[375] Véase: www.justice.gov/dag/proposed-language-regarding-expert-testimony-and-lab-reports-forensic-science.

[El] 21 de julio de 2016 se liberó un segundo conjunto de directrices propuestas incluyendo análisis de cabello, ADN mitocondrial y tipado de cromosoma Y. (www.justice.gov/dag/proposed-uniform-language-documents-anthropology-explosive-chemistry-explosive-devicesgeology).

[376] Véase: www.justice.gov/olp/file/861936/download

que una identificación que excluya todas las demás fuentes posibles no puede nunca probarse empíricamente.

Afirmaciones no aprobadas para su uso en informes de los laboratorios y en las declaraciones de los peritos con respecto al examen forense de la evidencia de huellas de calzado y de neumáticos.

Exclusión de todas las demás fuentes potenciales

El analista no puede afirmar que un calzado o un neumático sea la fuente de una impresión cuestionada excluyendo todos los demás calzados/neumáticos porque los demás calzados/neumáticos no han sido examinados. Examinar todos los calzados/neumáticos existentes en el mundo es imposible en la práctica.

Tasa de error

El analista no puede establecer un valor numérico o un porcentaje con respecto a la tasa de error asociada, tanto con respecto a la metodología utilizada para realizar el examen pericial como con respecto a la ejecución del analista que realiza los análisis.

Valoración estadística

El analista no puede establecer un valor numérico o una probabilidad asociada a su opinión. No existen en la actualidad datos precisos y fiables o modelos estadísticos que permitan realizar una determinación cuantitativa en un examen forense de la evidencia de una impresión de calzado o rueda de neumático.

Estas directrices propuestas presentan problemas serios.

Un analista puede opinar que un calzado es la fuente de una impresión, pero no que el calzado es la fuente de la impresión *excluyendo todos los otros calzados posibles*. Pero, como una cuestión de lógica, no hay diferencia entre esas dos afirmaciones. Si un analista cree que X es la fuente de Y, cree necesariamente que nada más es la fuente de Y. Cualquier miembro de un jurado que sea sensato debe comprender esta equivalencia.

Entonces, ¿cuál es la finalidad de las directrices? Parece ser la de reconocer la posibilidad de error. En efecto, los analistas deben decir, «Yo creo que X es la fuente de Y, aunque podría equivocarme sobre ello».

Esto es apropiado. Pero la cuestión crucial es entonces: ¿con qué probabilidad se equivoca el analista?

He aquí el problema: las directrices impiden al perito disertar sobre la probabilidad de error porque no hay información precisa o fiable sobre la precisión (del método). En efecto, se instruye a los analistas para que digan: «Creo que X es la fuente de Y, aunque podría estar equivocado al respecto. Pero no sé con qué frecuencia puedo equivocarme al no disponer de información fiable sobre eso».

Tal afirmación no supera prueba plausible alguna sobre su validez científica. Como escribió el Juez Easterly en *Williams v. los Estados Unidos*, una afirmación de identificación en tales circunstancias:

tiene el mismo valor probatorio que la visión de un vidente: no refleja nada más que la infundada creencia individual de creer estar en lo cierto. Esta es no es la evidencia en la que podemos confiar en buena conciencia, particularmente en casos penales, donde demandamos pruebas -pruebas reales- más allá de toda duda razonable, precisamente porque lo que está en juego vale mucho[377].

8.2.2. Análisis de cabellos

En el Cuadro 7 se muestran partes relevantes de las directrices para el testimonio y la redacción de informes sobre el examen forense del cabello.

CUADRO 7.
EXTRACTO DE LA PROPUESTA DEL DOJ SOBRE UN LENGUAJE UNIFORME PARA TESTIMONIOS E INFORMES EN LA DISCIPLINA DEL EXAMEN FORENSE DEL CABELLO[378]

Afirmaciones no aprobadas para su uso en informes de los laboratorios y en las declaraciones de los analistas con respecto al examen forense del cabello.

Comparaciones de cabello humano

El analista puede afirmar o implicar que el cabello humano cuestionado es microscópicamente consistente con la muestra de cabello conocida y, consecuentemente, la fuente de la muestra de cabello conocida pudiera incluirse en las posibles fuentes del cabello cuestionado.

Afirmaciones no aprobadas para su uso en informes de los laboratorios y en las declaraciones de los analistas con respecto al examen forense del cabello.

Individualización

El analista no puede afirmar o implicar que el cabello humano cuestionado proceda de una fuente en particular excluyendo todas las demás.

Valoración estadística

El analista no puede afirmar o implicar un peso estadístico o una probabilidad a una conclusión, o una verosimilitud sobre si el cabello cuestionado procedió de una fuente en particular.

Tasa de error

El analista no puede afirmar o implicar que el método utilizado en realizar el examen microscópico del cabello tenga una tasa de error igual a cero o que sea infalible.

---

[377] *Williams v. United States*, DC Tribunal de Apelación, resuelto el 21 de junio de 2016, (voto concurrente de Easterly). Citamos la analogía por su valor expositivo sobre la cuestión científica; no expresamos posición alguna sobre el caso en cuanto autoridad jurídica.

[378] Department of Justice Proposed Uniform Language for Testimony and Reports for the Forensic Hair Examination Discipline (Lenguaje Uniforme Propuesto por el Departamento de Justicia para el Testimonio y los Informes en la Disciplina Forense de Examinación de Cabellos), disponible en: www.justice.gov/dag/file/877736/download

Las directrices afirman apropiadamente que los analistas no pueden defender que pueden individualizar la fuente del cabello o que tengan una tasa de error igual a cero. Sin embargo, mientras que los analistas pueden «afirmar o implicar que el cabello humano cuestionado es microscópicamente consistente con la muestra de cabello conocida y, consecuentemente, la fuente de la muestra de cabello conocida pudiera incluirse en las posibles fuentes del cabello cuestionado», se les impide proporcionar información precisa sobre la fiabilidad de tales conclusiones. Esto es contrario al requerimiento científico de que los métodos de comparación de características deben sustentarse y acompañarse de estimaciones empíricas apropiadas de su fiabilidad.

En particular, como se discutió en la Sección 5.7, un estudio emblemático realizado por científicos del Laboratorio del FBI en el año 2002 demostró que de las 80 ocasiones con casos reales en los que los analistas concluyeron que un cabello cuestionado era microscópicamente consistente con la muestra de cabello conocida, el análisis de ADN demostró que el cabello procedía de una fuente diferente en el 11% de los casos. El hecho de que una proporción tan significativa de conclusiones fueran falsas asociaciones es de enorme importancia en la interpretación de las conclusiones de los analistas del cabello.

En casos de análisis de cabellos no acompañados por análisis de ADN, se debe solicitar a los peritos que revelen la alta frecuencia de falsas asociaciones observadas en el estudio del FBI, de forma que los miembros del jurado puedan apropiadamente sopesar las conclusiones.

8.2.3. Conclusión

El DOJ debería revisar las directrices propuestas a fin de armonizarlas con los estándares científicos de validez científica. También debe revisarse la documentación justificativa, como se expone en la Sección 5.7.

8.3. Recomendaciones

Sobre la base de sus hallazgos científicos, el PCAST hace las siguientes recomendaciones.

RECOMENDACIÓN 6.
USO DE LOS MÉTODOS DE COMPARACIÓN DE CARACTERÍSTICAS EN PROCESOS JUDICIALES FEDERALES

(A) El Fiscal General debe ordenar a los fiscales que comparezcan en nombre del Departamento de Justicia (DOJ) que aseguren que el testimonio de los analistas ante los tribunales sobre los métodos de comparación de características cumpla los estándares científicos de validez científica.

Si bien las investigaciones previas al juicio pueden basarse en una gama más amplia de métodos, el testimonio de los analistas ante los tribunales sobre métodos forenses de comparación de características en casos penales —que pueden ser muy influenciables y que ha conducido a muchas condenas injustas— debe cumplir un estándar más alto. Particularmente, los fiscales que comparezcan en nombre del Departamento de Justicia deben asegurarse de que:

(i) los métodos forenses de comparación de características sobre los que se basa el testimonio hayan sido establecidos como válidos en sus fundamentos, como lo demuestran los estudios empíricos apropiados y la consistencia con las evaluaciones realizadas por el Instituto Nacional de Estándares y Tecnología (NIST), cuando estén disponibles; y

(ii) el testimonio sea científicamente válido, con las declaraciones del analista sobre la precisión de los métodos y el valor probatorio de las identificaciones propuestas constreñidas por la evidencia empírica que les sirve de apoyo, sin implicar un mayor grado de certeza.

(B) El Departamento de Justicia debe llevar a cabo una revisión inicial, con la asistencia de NIST, de los métodos de comparación de características subjetivos utilizados por el DOJ, con el fin de identificar qué métodos (más allá de los revisados en este informe) carecen de estudios apropiados de caja negra necesarios para evaluar la validez de los fundamentos. Como consecuencia de que se presume que tales métodos subjetivos no son válidos en sus fundamentos, el Departamento de Justicia debe evaluar si es apropiado que se presenten ante los tribunales conclusiones basadas en tales métodos.

(C) Cuando existan métodos relevantes que aún no se hayan establecido como válidos en sus fundamentos, el DOJ debe fomentar y proporcionar apoyo para que se realicen estudios apropiados de caja negra para evaluar la validez de los fundamentos y medir la fiabilidad. El diseño y ejecución de estos estudios debe llevarse a cabo por o en unión de terceros independientes sin participación alguna en el resultado.

9. Acciones para garantizar la validez científica en la ciencia forense: recomendaciones al poder judicial

Sobre la base de los hallazgos referidos en las Secciones 4 y 5, el PCAST ha identificado acciones que creemos que deberían ser tenidas en cuenta por el Poder Judicial para garantizar la validez científica de la evidencia fundamentada en métodos de comparación de características y promover su uso más riguroso en las salas de audiencias.

9.1. La validez científica como fundamento del testimonio experto

En los tribunales federales, a los jueces se les asigna el papel crítico de «porteros» encargados de garantizar que el testimonio experto «descanse sobre una base

fiable»[379]. Específicamente, la regla 702 (c,d) de las Reglas Federales de la Prueba requiere que (1) el testimonio experto debe ser el producto de «principios y métodos fiables» y (2) los expertos deben haber «aplicado fiablemente» los métodos a los hechos del caso[380]. La Corte Suprema ha dictaminado que los jueces deben determinar «si el razonamiento o la metodología subyacente al testimonio es científicamente válida»[381].

Como se explica en la sección 3, este entorno de trabajo establece un importante diálogo entre el poder judicial y la comunidad científica. La admisibilidad del testimonio experto depende de un umbral de cumplimiento de ciertos criterios *jurídicos* de fiabilidad probatoria, que son exclusiva potestad del poder judicial. Sin embargo, en los casos que entrañen pruebas científicas, esos estándares jurídicos deben «basarse en la validez científica»[382].

El PCAST no opina sobre los criterios jurídicos, pero tiene como objetivo en este informe esclarecer los estándares *científicos* que subyacen a ellos. Para garantizar que la distinción entre conceptos científicos y jurídicos sea clara, hemos adoptado términos específicos para referirnos a los conceptos *científicos* (*validez de los fundamentos* y *validez en la aplicación*) con la intención de establecer un paralelismo con los conceptos jurídicos expresados en la Regla 702 (c,d).

Como ha señalado la Corte Suprema, la investigación del juez en virtud de la Regla 702 es flexible: no existe una simple prueba única que pueda aplicarse uniformemente a todas las disciplinas científicas[383]. Más bien, la evaluación de la validez científica debe basarse en los criterios científicos adecuados para el campo científico. Más aún, el campo científico apropiado debe ser la disciplina científica más amplia a la que pertenece[384].

En este informe, el PCAST se ha centrado sobre métodos forenses de comparación de características —que pertenece al campo de la metrología, la ciencia de la medición y de su aplicación—[385]. Hemos buscado —en una forma utilizable por los tribunales, así como por científicos y otros que buscan mejorar la ciencia forense— establecer criterios científicos para la validez de los fundamentos y la validez en la

---

379 *Daubert v. Merrell Dow Pharmaceuticals*, 509 U.S. 579 (1993) en 597.

380 Véase: www.uscourts.gov/file/rules-evidence.

381 *Daubert v. Merrell Dow Pharmaceuticals*, 509 U.S. 579 (1993) en 592.

382 *Daubert*, en FN9 («en un caso que conlleve prueba científica, la *fiabilidad de la prueba* estará basada en la validez científica.» [énfasis en el texto original]).

383 *Daubert*, en 594.

384 Por ejemplo, en el caso *Frye*, el tribunal valoró si un detector de mentiras propuesto había conseguido «reconocimiento permanente y científico entre las autoridades fisiológicas y psicológicas» en lugar de entre analistas en detectores de mentiras. *Frye v. United States*, 293 F. 1013 (D.C. Cir. 1923). De forma similar, el hecho de que los analistas en mordeduras crean que el examen de las mordeduras es válido, tiene escaso valor.

385 Véase nota 93.

aplicación (Sección 4) e ilustrar su aplicación a métodos de comparación de características específicos (Sección 5).

Los criterios científicos se describen en el Hallazgo 1. Las conclusiones del PCAST se pueden resumir del siguiente modo:

> La validez y fiabilidad científicas requieren que un método haya sido sometido a prueba empírica, bajo condiciones apropiadas a su pretendido uso, que proporcione estimaciones válidas sobre la frecuencia con la que el método alcanza una conclusión incorrecta. Para los métodos subjetivos de comparación de características, se precisan estudios de caja negra apropiadamente diseñados, en los que muchos analistas toman decisiones sobre muchas pruebas independientes (generalmente involucrando muestras «cuestionadas» y una o más muestras «conocidas») y en los que se determinan las tasas de error. Sin estimaciones apropiadas de la precisión, una declaración de un perito que consista en decir que dos muestras son similares —o, incluso, indistinguibles— carece de significado científico: no tiene valor probatorio, y sí un considerable potencial de producir un impacto perjudicial. Nada —ni la experiencia personal ni las prácticas profesionales— puede sustituir una adecuada demostración empírica de la precisión.

En los hallazgos 2-7 se describen las aplicaciones a métodos de comparación de características específicos. El conjunto completo de hallazgos científicos se recoge en la sección 10.

Por último, resaltamos que la Corte Suprema en *Daubert* sugirió que los jueces deberían tener en cuenta la Regla 706, que permite a un tribunal, bajo su propia discreción, obtener la asistencia de un experto de su elección[386]. Tales expertos pueden proporcionar valoraciones independientes concernientes, entre otras cosas, a la validez de los métodos científicos y sus aplicaciones.

9.2. El papel de precedentes

Un tema importante que surgió en nuestras deliberaciones fue el del papel de los precedentes.

Como es explicó en la sección 5, nuestra revisión científica encontró que la mayoría de los métodos forenses de comparación de características (con la notable excepción del análisis de ADN de muestras procedentes de una sola fuente y de mezclas simples) históricamente se han *asumido* antes que *establecido* como válidos en sus fundamentos. Solo después de que se hiciera evidente en los últimos años (basándose en el ADN y en otros análisis) la existencia de problemas en la fundamentación de la fiabilidad de algunos de estos métodos, la comunidad de ciencias forenses comenzó a reconocer la necesidad de probar *empíricamente* si ciertos métodos específicos cumplían los criterios de validez científica.

---

386 *Daubert*, en 595.

Esto crea una tensión obvia, porque muchos tribunales con fundamento en precedentes que fueron establecidos antes de que estos problemas fundamentales se descubrieran siguen admitiendo métodos de comparación de características.

Desde un punto de vista puramente *científico*, la resolución es clara. Cuando hechos nuevos falsean antiguas suposiciones, los tribunales no deben verse obligados a citar los precedentes del pasado: deben mirar con nuevos ojos las cuestiones científicas. ¿Cómo se resuelven esas tensiones desde el punto de vista jurídico? La Corte Suprema ha dejado claro que un tribunal puede anular un precedente si constata que un caso anterior fue «erróneamente resuelto y que los acontecimientos posteriores han socavado su validez»[387].

El PCAST no expresa opinión sobre la cuestión jurídica de si algunos de los casos del pasado fueron «erróneamente decididos». Sin embargo, el PCAST advierte que, desde un punto de vista *científico*, acontecimientos posteriores han socavado la validez de las conclusiones que no se basaron en pruebas empíricas apropiadas. Estos acontecimientos incluyen (1) el reconocimiento de problemas sistémicos con algunos métodos forenses de comparación de características, que se muestran mediante los estudios de las causas de cientos de condenas erróneas descubiertas a través del ADN y otros análisis; (2) el informe del NRC (2009) de la Academia Nacional de Ciencias, el principal órgano consultivo científico establecido por el poder legislativo[388], que encontró que algunos métodos forenses de comparación de características carecen de fundamento científico; y (3) la revisión científica realizada por el PCAST en el presente informe, el principal órgano consultivo científico establecido por el poder ejecutivo[389], encuentra que algunos métodos forenses de comparación de características carecen de validez en sus fundamentos.

9.3. Medios para los jueces

Otra cuestión importante que surgió frecuentemente en nuestras conversaciones con analistas fue la necesidad de ofrecer a los jueces mejores medios para la evaluación de los métodos forenses de comparación de características para su uso en los tribunales.

---

[387] *Boys Markets, Inc. v. Retails Clerks Union*, 398 U.S. 235, 238 (1970). Véase también: *Patterson v. McLean* Credit Union, 485 U.S. 617, 618 (1988) (resaltando que el Tribunal ha «anulado los precedentes legales en una serie de casos»). El PCAST solicitó asesoramiento sobre esta cuestión a su panel de asesores superiores.

[388] La Academia Nacional de Ciencias fue creada por el Congreso en 1863 para asesorar al gobierno federal en asuntos de ciencia (U.S. Code, Section 36, Title 1503).

[389] El Presidente estableció, formalmente, un consejo científico permanente consultivo poco después del lanzamiento del Sputnik en 1957. Actualmente se denomina Consejo de Asesores de Ciencia y Tecnología del Presidente (que opera bajo la Orden Ejecutiva 13539, modificada por la Orden Ejecutiva 13596).

Los órganos más apropiados para proporcionar esos recursos son la Conferencia Judicial de los Estados Unidos y el Centro Judicial Federal.

La Conferencia Judicial de los Estados Unidos es el órgano nacional de formulación de políticas para los tribunales federales[390]. Su responsabilidad estatutaria incluye el estudio del funcionamiento y el efecto de las normas generales que regulan la práctica y el procedimiento en los tribunales federales. La Conferencia Judicial desarrolla manuales de buenas prácticas y emite notas del Comité Consultivo para ayudar a los jueces respecto a temas específicos, incluso a través de su Comité Consultivo Permanente sobre las Reglas Federales de la Prueba.

El Centro Judicial Federal es la agencia de investigación y educación del sistema judicial federal[391]. Sus deberes estatutuarios incluyen (1) dirigir y promover la investigación sobre procedimientos judiciales federales y el funcionamiento de los tribunales y (2) dirigir y promover la formación, la educación y la capacitación continua de los jueces federales, empleados judiciales y otros.

El PCAST recomienda que la Conferencia Judicial de los Estados Unidos, a través de su Subcomité de las Reglas Federales de la Prueba, desarrolle manuales de buenas prácticas y una nota sobre la interpretación adecuada de las reglas relevantes por parte del Comité Consultivo y que el Centro Judicial Federal desarrolle programas educativos relacionados con los procedimientos para evaluar la validez científica de los métodos forenses de comparación de características.

9.4. Recomendación

Sobre la base de sus hallazgos científicos, el PCAST realiza la siguiente recomendación.

RECOMENDACIÓN 8.
LA VALIDEZ CIENTÍFICA COMO FUNDAMENTO DEL TESTIMONIO EXPERTO

(A) Al decidir sobre la admisibilidad de testimonios de analistas, los jueces federales deben tener en cuenta los criterios científicos apropiados para evaluar la validez científica, que incluyen:

(i) la validez de los fundamentos, con respecto al requisito de la Regla 702(c) de que el testimonio sea el producto de principios y métodos fiables; y

(ii) la validez en la aplicación, con respecto al requisito de la Regla 702(d) de que el analista haya aplicado fiablemente los principios y métodos a los hechos del caso.

[390] Creado en 1922 bajo la denominación de Conferencia de Jueces de Circuito Superior, la Conferencia Judicial de los Estados Unidos está actualmente establecida mediante la 28 U.S.C. § 331.

[391] El Centro Judicial Federal fue creado por el Congreso en 1967 (28 U.S.C. §§ 620-629), tras la recomendación de la Conferencia Judicial de los Estados Unidos.

Estos criterios científicos se describen en el Hallazgo 1.

(B) Los jueces federales, cuando permitan que un analista testifique sobre un método de comparación de características válido en sus fundamentos, deben asegurarse de que el testimonio sobre la precisión del método y el valor probatorio de las identificaciones propuestas sea científicamente válido en el sentido de que esté limitado a lo que la evidencia empírica sustente. Las declaraciones que sugieran o impliquen una mayor certeza no son científicamente válidas y no deben permitirse. En particular, los tribunales nunca deben permitir aseveraciones científicamente indefendibles como: tasa de error "cero", "muy pequeña", "esencialmente cero", "despreciable", "mínima" o "microscópica"; "certeza del 100%" o prueba "hasta un grado razonable de certeza científica"; identificación "hasta la exclusión de todas las demás fuentes posibles"; o una probabilidad de error tan remota como para que sea una "imposibilidad práctica".

(C) Para ayudar a los jueces, la Conferencia Judicial de los Estados Unidos a través de su Comité Consultivo Permanente sobre las Reglas Federales de la Prueba, debe preparar, con el asesoramiento de la comunidad científica, un manual de buenas prácticas y una nota sobre la interpretación adecuada del Comité Consultivo, proporcionando orientación a los jueces federales sobre la admisibilidad bajo la Regla 702 de testimonios expertos basados en métodos forenses de comparación de características.

(D) Para ayudar a los jueces, el Centro Judicial Federal debe desarrollar programas sobre criterios científicos para la validez científica de los métodos forenses de comparación de características.

10. Hallazgos científicos

Los hallazgos científicos del PCAST en este informe se recogen a continuación. El Hallazgo 1, que se refiere a los criterios científicos para la validez científica, se fundamenta en las explicaciones de la sección 4. Los hallazgos 2-6, sobre la validez de los fundamentos de 6 métodos forenses de comparación de características, se basan en las evaluaciones de la sección 5.

HALLAZGO 1:
CRITERIOS CIENTÍFICOS PARA LA VALIDEZ CIENTÍFICA DE UN MÉTODO DE COMPARACIÓN DE CARACTERÍSTICAS FORENSE

(1) Validez de los fundamentos. Para establecer la validez de los fundamentos de un método de comparación de características, se requieren los siguientes elementos:

(a) un procedimiento reproducible y consistente para (i) la identificación de las propiedades en muestras que son ofrecidas como prueba; (ii) la comparación de las propiedades entre dos muestras; (iii) la determinación, basándose en la similitud de las

propiedades entre dos conjuntos de características, sobre si se debe declarar que las muestras proceden probablemente del mismo origen ("regla de coincidencia"); y

(b) estimaciones empíricas, tras estudios apropiadamente diseñados a partir de múltiples grupos, que establezcan: (i) la tasa de falsos positivos del método —es decir, la probabilidad de una propuesta de identificación entre muestras que realmente proceden de diferentes fuentes— y (ii) la sensibilidad del método —es decir, la probabilidad de una propuesta de identificación entre muestras que realmente proceden de la misma fuente—.

Como se describió en el Cuadro 4, la validación científica debe satisfacer una serie de criterios: (a) debe basarse en conjuntos suficientemente grandes de muestras conocidas y representativas de poblaciones relevantes; (b) debe conducirse de tal modo que los analistas no tengan información sobre las respuestas correctas; el diseño del estudio y el plan de análisis deben especificarse por adelantado y no modificarse posteriormente después de alcanzar los resultados; (d) el estudio debe ser conducido o supervisado por individuos u organizaciones sin beneficio en los resultados; (e) los datos, el software y los resultados deben estar disponibles para permitir que otros científicos puedan revisar las conclusiones; y (f) para asegurar que los resultados sean robustos y reproducibles, debe haber diversos estudios independientes realizados por grupos separados que alcance resultados similares.

Una vez que un método ha sido establecido como válido en sus fundamentos mediante adecuados estudios empíricos, las afirmaciones sobre la precisión y el valor probatorio de las propuestas de identificación, para que sean válidas, deben basarse sobre tales estudios empíricos.

Para métodos objetivos, la validez de los fundamentos puede establecerse demostrando la fiabilidad de cada uno de los pasos individuales (identificación de características, comparación de propiedades, regla de coincidencia, probabilidad de coincidencia falsa y sensibilidad).

Para métodos subjetivos, la validez de los fundamentos *solo* puede establecerse mediante estudios de caja negra que midan la frecuencia en que muchos examinadores alcancen conclusiones precisas en muchos problemas de comparación de características que contengan muestras representativas del uso pretendido. En ausencia de tales estudios, un método subjetivo de comparación de características no puede considerarse científicamente válido.

La validez de los fundamentos es un *sine qua non*, que solo puede demostrarse mediante estudios empíricos. Es muy importante resaltar que las buenas prácticas profesionales —tales como la existencia de sociedades profesionales, programas de certificación, programas de acreditación, artículos revisados por pares, protocolos estandarizados, pruebas de aptitud y códigos éticos— no pueden sustituir a la evidencia empírica de la validez y fiabilidad científica.

(2) Validez en la aplicación. Una vez que un método de comparación de características forense ha sido considerado como válido en sus fundamentos, es necesario establecer su validez en la aplicación en un caso dado.

Como se describió en el Cuadro 5, la validez en la aplicación requiere que: (a) el examinador forense debe haber demostrado que es *capaz* de aplicar el método fiablemente, lo que se hace sometiéndose a tests de aptitud apropiados (véase Sección 4.6) y debe

de hecho haberlo llevado a cabo, como se ha demostrado en el procedimiento *de hecho* utilizado en el caso, los resultados obtenidos y las notas de laboratorio, que deben estar disponibles para que otros puedan revisarlas científicamente; y (b) las aserciones sobre el valor probatorio de las propuestas de identificación deben ser científicamente válidas —incluyendo que los examinadores deben informar sobre la tasa completa de falsos positivos y la sensibilidad del método establecidas en los estudios de validez de los fundamentos—; demuestrar que las muestras utilizadas en los estudios de los fundamentos son relevantes para los hechos del caso; y cuando sea aplicable, demuestrar el valor probatorio de la coincidencia observada a partir de las características específicas observadas en el caso; además de no realizar afirmaciones o implicaciones que vayan más allá de la evidencia empírica.

HALLAZGO 2:
ANÁLISIS DE ADN

Validez de los fundamentos. El PCAST encuentra que el análisis de muestras procedentes de una única fuente o de mezclas simples de dos individuos, tales como las que dan en casos de violación, es un método objetivo que ha sido establecido como válido en sus fundamentos.

Validez en la aplicación. Como los errores por fallos humanos son dominantes en la probabilidad de coincidencias aleatorias, los criterios científicos para la validez en la aplicación requieren que un analista (1) deba haber realizado una prueba de aptitud rigurosa y relevante para demostrar su capacidad para aplicar fiablemente el método, (2) deba revelar rutinariamente en informes y testimonios si cuando llevó a cabo el examen era consciente de algunos hechos del caso que podrían influenciar en la conclusión y (3) deba revelar, previa solicitud, toda la información sobre las pruebas de calidad y las incidencias de calidad en su laboratorio.

HALLAZGO 3:
ANÁLISIS DE ADN DE MUESTRAS DE MEZCLAS COMPLEJAS

Validez de los fundamentos. El PCAST encuentra que:

(1) Métodos basados en la Probabilidad Combinada de Inclusión (CPI). El análisis de ADN de mezclas complejas basado en aproximaciones fundamentadas en la CPI ha sido un método inadecuadamente especificado y subjetivo que tiene potencial para conducir a resultados erróneos. Como tal, no es válido en sus fundamentos.

Un artículo reciente ha propuesto normas específicas que abordan una serie de problemas con el uso de la CPI. Esas normas son claramente *necesarias*. Sin embargo, el PCAST no ha tenido tiempo necesario para valorar si son también suficientes para definir un método científicamente válido y objetivo. Si, por un tiempo limitado, los tribunales deciden admitir resultados basados en la aplicación de la CPI, la validez de su

aplicación requeriría que como mínimo fuera consistente con las normas especificadas en el artículo.

El análisis de ADN de mezclas complejas debe dirigirse sin demora hacia métodos más apropiados basados en genotipado probabilístico.

(2) Genotipado probabilístico. El análisis objetivo de mezclas complejas de ADN con software de genotipado probabilístico es un enfoque relativamente nuevo y prometedor. Se requiere evidencia empírica para establecer la validez de los fundamentos de cada método dentro de rangos específicos. Hasta el momento, hay evidencia publicada que apoya la validez de los fundamentos del análisis, con algunos programas, de mezclas de ADN de tres individuos, en las que el menor contribuyente constituye, al menos, el 20% del ADN intacto en la mezcla, y en las que la cantidad de ADN excede del mínimo requerido por el método. El rango en el que ha quedado establecida la validez de los fundamentos es probable que crezca en la medida en que se obtenga evidencia adecuada para mezclas más complejas y se publique.

Validez en la aplicación. Para métodos que son válidos en sus fundamentos, la validez en la aplicación conlleva consideraciones similares a las realizadas para análisis de ADN procedente de una única fuente y para mezclas simples, con un énfasis especial en asegurar que el método se aplique correctamente y dentro de su rango empíricamente establecido.

HALLAZGO 4:
ANÁLISIS DE MARCAS DE MORDEDURA

Validez de los fundamentos. El PCAST encuentra que el análisis de marcas de mordedura no cumple los estándares científicos para la validez de los fundamentos y está lejos de cumplir tales estándares. Al contrario, la evidencia científica disponible sugiere fuertemente que los analistas no pueden estar de acuerdo consistentemente sobre si una herida es de mordedura humana y no pueden identificar la fuente de una marca de mordedura con precisión razonable.

HALLAZGO 5:
ANÁLISIS DE HUELLAS DACTILARES LATENTES

Validez de los fundamentos. Basándose principalmente en dos estudios recientes de caja negra apropiadamente diseñados, el PCAST encuentra que el análisis de huellas dactilares latentes es una metodología subjetiva válida en sus fundamentos —aunque con una tasa de falsos positivos que es sustancial y probablemente más alta que la esperada por muchos miembros del jurado que está basada en afirmaciones largamente mantenidas en el tiempo sobre la infalibilidad de los análisis de huellas dactilares.

Las conclusiones de una propuesta de identificación pueden ser científicamente válidas con tal que estén acompañadas de información precisa sobre las limitaciones de

la fiabilidad de la conclusión —específicamente, que (1) solo se han llevado a cabo dos estudios apropiadamente diseñados sobre la validez de los fundamentos y la precisión de los análisis de huellas latentes; (2) estos estudios encontraron tasas de falsos positivos que podrían ser tan altas como de 1 error cada 306 casos, en uno de ellos, y de 1 error cada 18 en el otro; y (3) como consecuencia de que los analistas eran conscientes de que estaban siendo evaluados, la tasa real de falsos positivos en casos reales puede ser más alta. Hasta el presente, afirmaciones sobre una mayor precisión no están garantizadas o científicamente justificadas. Se requieren estudios de caja negra adicionales para aclarar la fiabilidad del método.

Validez en la aplicación. Aunque concluímos que el método es válido en sus fundamentos, hay una serie de cuestiones importantes relacionadas con la validez en su aplicación.

(1) Sesgo de confirmación. El trabajo realizado por científicos del FBI ha mostrado que los analistas suelen alterar las características que marcan inicialmente en una huella latente basándose en las comparaciones realizadas con una huella indubitada aparentemente coincidente. Ese razonamiento circular introduce un serio riesgo de sesgo de confirmación. A los analistas se les debe requerir que completen y documenten sus análisis de una huella latente *antes* de mirar a una huella conocida y deben documentar, separadamente, cualquier dato adicional utilizado durante su comparación y evaluación.

(2) Sesgo contextual. El trabajo realizado por miembros de la comunidad académica ha demostrado que el juicio de los analistas puede estar influenciado por información irrelevante sobre los hechos de un caso. Deben realizarse esfuerzos para asegurar que los analistas no están expuestos a información potencialmente sesgadora.

(3) Pruebas de aptitud. La prueba de aptitud es esencial para valorar la capacidad de un analista y su rendimiento en la emisión de juicios precisos. Como se trató sobre el particular en otro lugar de este informe, las pruebas de aptitud necesitan mejorarse haciéndolas más rigurosas, incorporándolas dentro del flujo de trabajo, y divulgando las pruebas para que ellas sean evaluadas por la comunidad científica.

HALLAZGO 6:
ANÁLISIS DE ARMAS DE FUEGO

Validez de los fundamentos. El PCAST encuentra que el análisis de armas de fuego no cumple en la actualidad los criterios para la validez de los fundamentos, porque solo hay un único estudio apropiadamente diseñado para medir la validez y estimar la fiabilidad. Los criterios científicos para la validez de los fundamentos requieren más de un estudio de ese tipo para demostrar la reproducibilidad.

Es competencia de los tribunales decidir si el análisis de armas de fuego debe considerarse admisible basándose en la evidencia actual.

Si el análisis de armas de fuego se permite en los tribunales, los criterios científicos para la validez de su aplicación requieren que se informe claramente sobre las tasas

de error vistas en estudios de caja negra apropiadamente diseñados (estimadas en 1 error cada 66 comparaciones, con un límite superior del 95% de confianza que se corresponde con 1 error cada 46 comparaciones, en el único estudio realizado hasta la fecha).

Validez en la aplicación. Si el análisis de armas de fuego es permitido en los tribunales, la validez de su aplicación requiere, desde el punto de vista científico, que el analista:

(1) haya superado una prueba de aptitud rigurosa sobre un gran número de problemas de prueba que permitan evaluar su capacidad y rendimiento y que revele los resultados de las pruebas de aptitud; y

(2) revele si, cuando llevó a cabo la examinación, era consciente de algunos otros hechos del caso que pudieran influir en la conclusión.

HALLAZGO 7:
ANÁLISIS DE HUELLAS DE CALZADO

Validez de los fundamentos. El PCAST encuentra que no existen estudios empíricos apropiados que sostengan la validez de los fundamentos de los análisis de huellas de calzado para asociar impresiones de esas huellas con un calzado en particular basándonos en marcas identificativas específicas (en ocasiones denominadas "características adquiridas aleatoriamente"). Tales conclusiones no están respaldadas por ninguna evidencia significativa o estimaciones de su precisión y, por ello, no son científicamente válidas.

El PCAST no ha evaluado la validez de los fundamentos de los análisis de las huellas de calzado para identificar características de clase (por ejemplo, la talla del calzado o la marca).

APÉNDICE A: CUESTIONES ESTADÍSTICAS

Para mejorar su accesibilidad a un público amplio, el texto principal de este informe evita, en la medida de lo posible, el uso de terminología matemática y estadística. Sin embargo, para la implementación efectiva de alguno de los principios reflejados en el informe, se necesitan descripciones algo más precisas. Este apéndice resume los conceptos relevantes desde la estadística básica[392].

[392] Véanse, por ejemplo: Amitage, Berry y Matthews, 2002; Snedecor y Cochran, 1989; van Belle, Fisher, Heagerty, et al., 2004; Agresti y Coull, 1998: 119-126; Hogg, Tanis y Zimmerman, 2015; Freedman, Pisani y Purves, 2007; Moses, 1986; Moore, McCabe y Craig, 2009.

1. Sensibilidad y tasa de falsos positivos

Los métodos forenses de comparación de características generalmente tratan de determinar con qué probabilidad dos muestras proceden de una misma fuente, dado el resultado de una prueba forense sobre las muestras. Se consideran dos posibilidades: la hipótesis nula (H0) consistente en que las muestras proceden de fuentes diferentes y la hipótesis alternativa (H1) de que proceden de una misma fuente. El resultado de la prueba puede resumirse como una declaración de coincidencia (M) o de no coincidencia (O).

Existen dos caracterizaciones necesarias de la precisión de un método: sensibilidad (abreviado por SEN) y la tasa de falsos positivos (FPR).

La sensibilidad se define como la probabilidad de que el método declare una coincidencia entre dos muestras cuando se conoce que proceden de una misma fuente (extraída de una población apropiada), es decir, SEN = P(M|H1). Por ejemplo, un valor de SEN = 0.95 indicaría que las dos muestras procedentes de una misma fuente serían declaradas como coincidentes el 95% de las veces. En la literatura estadística, SEN es a veces llamada «tasa de verdaderos positivos», «TPR» o «el complementario del error tipo II»[393].

La tasa de falsos positivos (abreviado FPR) se define como la probabilidad de que el método declare una coincidencia entre dos muestras procedentes de fuentes diferentes (de nuevo, en una población apropiada), es decir, FPR = P(M|H0). Por ejemplo, un valor de FPR = 0.01 indicaría que dos muestras procedentes de fuentes diferentes serían (erróneamente) denominadas coincidentes un 1% de las veces[394]. Los métodos con alta FPR no son científicamente fiables como para emitir juicios importantes sobre la fuente de una muestra ante un tribunal. Para ser considerados fiables, el FPR debería ser menor al 5% y pudiera ser apropiado que sea considerablemente menor dependiendo de la finalidad de la aplicación.

Los resultados de un estudio empírico dado pueden resumirse en cuatro valores: el número de ocurrencias en el estudio de verdaderos positivos (TP), de falsos positivos (FP), de falsos negativos (FN) y de verdaderos negativos (TN). (La matriz de estos valores se conoce, curiosamente, como la «matriz de confusión»).

---

[393] El término «tasa de falso negativo» se utiliza, a veces, como el complementario de SEN, es decir, FNR = 1 – SEN.

[394] Los estadísticos pueden referirse a la especificidad de un método (SPC) en vez de a su tasa de falsos positivos (FPR). Los dos están relacionados mediante la fórmula FPR = 1 – SPC. En el ejemplo dado, FPR = 0.01 (1%) y SPC = 0.99 (99%).

	Resultado de la prueba	
	Coincidencia	No coincidencia
H1: verdadera procedencia de la misma fuente	TP	FN
H0: verdadera procedencia de fuentes diferentes	FP	TN

En esta terminología estándar, aunque confusa, «verdadero» y «falso» se refieren al acuerdo o no con la verdad (ya sea H0 o H1), mientras que «positivo» y «negativo» se refieren a los resultados de la prueba (es decir, resultan M y O, respectivamente).

Una estimación ampliamente utilizada de SEN, denominada estimación de *máxima verosimilitud*, viene dado por TP/(TP+FN), la fracción de sucesos H1 verdaderos (misma fuente) que se declaran correctamente como coincidentes M (match).

Puesto que la tasa de falsos positivos será frecuentemente el factor matemático determinante para el valor probatorio de un método en un caso particular (se explica más adelante), resulta particularmente importante que FPR sea medido empíricamente.

Además, las pruebas con muy baja sensibilidad deben ponerse bajo sospecha porque los escasos resultados positivos de las pruebas pueden ser coincidentes o ser superados por la ocurrencia de falsos positivos[395].

2. Intervalos de confianza

Como se explicó en el texto principal, para ser validas las medidas empíricas de SEN y FPR han de basarse en grandes colecciones de muestras conocidas y representativas de cada población relevante, de manera que reflejen la frecuencia con la que ocurre una determinada característica o una combinación de características. (En el texto principal también se explican otros requisitos para la validez).

Como las medidas empíricas se basan en un limitado número de muestras, SEN y FPR no pueden ser medidos con exactitud, sino solamente estimados. Dado los tamaños finitos de las muestras, las estimaciones de máxima verosimilitud no cuentan la historia completa. Por tanto, es necesario y apropiado citar los límites de confianza dentro de los cuales es muy probable que se encuentren SEN y FPR.

Debido a que uno debería estar preocupado principalmente de sobreestimar SEN o de subestimar FPR, resulta apropiado utilizar un límite de confianza *unilateral*. Por convención, un nivel de confianza del 95% es el que más se utiliza —lo que sig-

[395] El argumento a favor de una prueba que dice algo como «esta prueba tiene éxito solo ocasionalmente, pero en este caso tuvo éxito» es por tanto falaz.

nifica que hay un 5% de probabilidad de que el valor verdadero supere el límite—. Se deben utilizar, por tanto, límites de confianza unilaterales superiores al 95% para estimar las tasas de error y las cantidades asociadas que caracterizan los métodos forenses de comparación de características (El uso de valores más bajos puede verse, con razón, como un intento de ofuscación).

El límite de confianza para proporciones depende del tamaño muestral en el estudio empírico. Cuando es pequeño, las estimaciones pueden quedar lejos del valor verdadero. Por ejemplo, si un estudio empírico no encontró falsos positivos en 25 pruebas individuales existe todavía una probabilidad razonable (de menos del 5%) de que la verdadera tasa de error pueda llegar a ser tan alta como de 1 entre 9, aproximadamente.

Por razones técnicas, no existe un método único, universalmente aceptado, para calcular estos intervalos de confianza (un problema conocido como «intervalo de confianza de la proporción binomial»). Sin embargo, los diversos métodos generalmente dan resultados muy similares y todos deben considerarse aceptables: el método Clopper-Pearson/Binomial exacto, el intervalo de puntuación Wilson, el intervalo Agresti-Coull (Wald ajustado) y el intervalo de Jeffreys[396]. De todos estos métodos hay calculadoras disponibles en Internet[397]. Por ejemplo, si un estudio encuentra cero falsos positivos en 100 pruebas, los cuatro métodos mencionados dan, respectivamente, los valores 0.030, 0.026, 0.032 y 0.019 para el límite superior de confianza del 95%. Desde el punto de vista científico, cualquiera de ellos podría ser apropiadamente referido en un informe dirigido a un jurado significando que «la tasa de falsos positivos podría ser tan alta como» (En este informe utilizamos el método Clopper-Pearson/Binomial exacto).

3. Calculando resultados para pruebas concluyentes

En muchas pruebas forenses, los peritos pueden alcanzar una conclusión (por ejemplo, se ha producido una coincidencia o no se ha producido) o declarar que la prueba es inconclusa. Por tanto, SEN y FPR pueden calcularse sobre la base de pruebas *concluyentes* o sobre *todas* las pruebas realizadas. Aunque ambas tasas tienen interés desde el punto de vista científico, la primera ha de utilizarse para informar de FPR al jurado. Esto resulta apropiado porque las pruebas utilizadas contra un acusado se basarán generalmente en pruebas *concluyentes* y no en pruebas inconclusas. Para ilustrar este punto, consideremos el caso extremo en el que un método ha sido probado 1000 veces y se encontró 990 resultados inconclusos, 10 falsos positivos y ningún resultado correcto. Sería engañoso informar que la tasa de falsos positivos

[396] Brown, 2001: 101-133.

[397] Por ejemplo, véase: epitools.ausvet.com.au/content.php?page=CIProportion.

fue del 1% (10 sobre 1000 pruebas). Por el contrario, se debería informar de que el 100% de los resultados concluyentes fueron falsos positivos (10 sobre 10 pruebas).

4. Análisis bayesiano

En este apéndice, hemos centrado la atención en la sensibilidad y en la tasa de falsos positivos (SEN = P(M|H1) y FPR = P(M|H0)). La cantidad de mayor interés en un juicio penal es P(H1|M), es decir, «la probabilidad de que las muestras procedan de la misma fuente *dado* que se ha declarado la coincidencia»[*]. Esta cantidad es frecuentemente denominada *valor predictivo positivo* (PPV) de la prueba.

El cálculo de PPV depende de dos cantidades: del «factor de Bayes» BF = SEN/FPR y de una segunda cantidad denominada «apuesta *a priori*» (POR). Esta última cantidad se define, matemáticamente, como POR = P(H0)/P(H1), donde P(H0) y P(H1) son las probabilidades a priori (es decir, antes de realizar la prueba) de las hipótesis H0 y H1[398]. La fórmula de PPV en términos de BF y POR es: PPV = BF / (BF + POR), una fórmula que se deriva del principio estadístico conocido como el Teorema de Bayes[399].

El Teorema de Bayes ofrece una forma matemática de combinar el resultado de la prueba con información independiente —tales como (1) la propia probabilidad a priori de que dos muestras procedan de una misma fuente y (2) el número de muestras buscadas—. Algunos estadísticos bayesianos elegirían un *a priori* POR = 1 en el supuesto de una coincidencia con una única muestra (implicando que es igualmente probable *a priori* que las muestras proceden de una misma fuente que de fuentes diferentes) y una apuesta *a priori* (POR) = 100.000 para una coincidencia tras una búsqueda en una base de datos de 100.000 registros. Otros establecerían POR = (1-p)/p, donde p es la probabilidad *a priori* de la identidad de la misma fuente en la población relevante, dados los otros hechos del caso.

El enfoque bayesiano es matemáticamente elegante. Sin embargo, plantea retos para su uso en los tribunales: (1) diferentes personas pueden tener diferentes creencias sobre POR y (2) muchos miembros del jurado podrían no entender cómo las creencias sobre POR afectan al cálculo matemático de PPV. (Además, como se ha señalado anteriormente, las estimaciones empíricas de SEN y FPR son inciertas, por lo que el factor de Bayes estimado, BF= SEN/FPR, también lo es).

---

[*] N. del T.: la coincidencia se expresa con el término «match» en inglés en la literatura especializada.

[398] Es decir, si p es la probabilidad *a priori* de la identidad de la misma fuente en la población bajo examen, entonces POR = (1-p)/p.

[399] En el texto principal, la frase «apropiadamente correcto para el tamaño del grupo que se exploró para la identificación de un sospechoso» se refiere al uso de esta fórmula con un apropiado valor de POR.

Por tanto, algunos analistas, por consiguiente, prefieren simplemente informar de las cantidades medidas empíricamente (la sensibilidad, la tasa de falsos positivos de la prueba y la probabilidad de que una coincidencia sea falso positivo dado el número de muestras contra las que la muestra buscada ha sido comparada), permitiendo al jurado que las incorpore a sus propios juicios bayesianos intuitivos. (Por ejemplo, «*Sí, la prueba tiene una tasa de falsos positivos de solo 1 cada 100, pero dos testigos ubican al acusado a 1.000 millas de la escena del crimen, por lo que el resultado de la prueba fue probablemente uno de esos 1 de cada 100 falsos positivos*»).

APÉNDICE B: EXPERTOS ADICIONALES QUE REALIZAN APORTACIONES

El PCAST buscó aportaciones de un grupo adicional de expertos y sectores interesados. El PCAST expresa su gratitud a los que aquí se listan por compartir su experiencia. No tuvieron la oportunidad de revisar los borradores del informe, y su consentimiento para participar con el PCAST en puntos específicos no implica que respalden los puntos de vista expresados en este informe. La responsabilidad de las opiniones, hallazgos y recomendaciones en este informe, así como cualquier error de hecho o de interpretación, descansa únicamente sobre el PCAST.

Richard Alpert
Asistente del Fiscal de Distrito Penal Tarrant
Oficina del Fiscal de Distrito Criminal del Condado

Kareem Belt
Analista de Política Forense
Proyecto Inocencia

William Bodziak
Consultor
Bodziak Forensics

John Buckleton
Director Científico
Instituto de Medio Ambiente e Investigación Científica
Nueva Zelanda

Bruce Budowle
Catedrático, Director Ejecutivo del Instituto de Genética Aplicada
Centro de Ciencias de la Salud de la Universidad del Norte de Texas

Mary A. Bush
Profesora Asociada
Departamento de Odontología Restaurativa
Facultad de Medicina Dental de la Universidad de Búfalo

Peter Bush
Profesor de Investigación
Director del Centro de Instrumentos del Campus Sur
Facultad de Medicina Dental de la Universidad de Búfalo

John Butler
Adjunto Especial del Director de Ciencias Forenses
Oficina de Programas Especiales
Instituto Nacional de Estándares y Tecnología

Arturo Casadevall
Catedrático
Departamento de Microbiología e Inmunología y Departamento de Medicina
Facultad de Medicina Albert Einstein

Alicia Carriquiry
Catedrática Distinguida en el Estado de Iowa y Directora
Centro de Estadística y Aplicaciones en Pruebas Forenses
Universidad del Estado de Iowa

Richard Cavanagh
Director
Oficina de Programas Especiales
Instituto Nacional de Estándares y Tecnología

Eleanor Celeste
Analista Política
Ciencias Médicas y Forenses
Oficina de Política Científica y Tecnológica

Christophe Champod
Catedrático de Derecho, Ciencias Penales y Administración Pública
Universidad de Lausana

Sarah Chu
Abogada Senior de Política Forense
Proyecto Inocencia

Simon A. Cole
Catedrático de Criminología, Derecho y Sociedad
Facultad de Ecología Social
Universidad de California en Irvine

Kelsey Cook
Director de Programa
Medición e Imagen Químicas
National Science Foundation

Jennifer Friedman
Coordinadora de Ciencia Forense
Defensora Pública del Condado de Los Ángeles

Lynn Garcia
Departamento Jurídico
Comisión de Ciencia Forense de Texas

Daniel Garner
Director Ejecutivo y Presidente
Centro de Ciencia Forense de Houston

Constantine A. Gatsonis
Cátedra Henry Ledyard Goddard de Bioestadística
Catedrático de Bioestadística
Director del Centro de Ciencias Estadísticas
Brown University

Eric Gilkerson
Analista Forense
Laboratorio del FBI

Brandon Giroux
Presidente
Giroux Forensics, L.C.C.
Presidente
Forensic Assurance

Catherine Grgicak
Profesora Asociada
Anatomía y Neurobiología
Facultad de Medicina de la Universidad de Boston

Austin Hicklin
Miembro
Fellow

Cindy Homer
Científico Forense
Laboratorio de Criminalística de la Policía del Estado de Maine

Alice Isenberg
Subdirectora Adjunta
Laboratorio del FBI

Matt Johnson
Especialista Forense Senior
Departamento del Sheriff del Condado de Orange

Jeff Salyards
Director
Centro de Ciencias Forenses de la Defensa
Agencia Forense y Biométrica de la Defensa

Rodney Schenck
Centro de Ciencias Forenses de la Defensa
Antropología Biológica
Agencia Forense y Biométrica de la Defensa

David Senn
Director
Centro para la Educación e Investigación en Ciencias Forenses
Simposio del Suroeste en Odontología Forense
Centro de Ciencias de la Salud en San Antonio de la Universidad de Texas

Stephen Shaw
Analista de Trazas
Laboratorio del FBI

Andrew Smith
Supervisor de la Unidad de Armas de Fuego y Marcas de Herramienta
Departamento de Policía de San Francisco

Erich Smith
Científico en Física
Unidad de Armas de Fuego y Marcas de Herramienta
Laboratorio del FBI

Tasha Smith
Unidad de Armas de Fuego y Marcas de Herramienta
Laboratorio de Criminalística
Departamento de Policía de San Francisco

Jeffrey Snipes
Profesor Adjunto
Estudios de Justicia Penal
Universidad del Estado de San Francisco

Jill Spriggs
Director de Laboratorio
Oficina del Fiscal de Distrito del Condado de Sacramento

Harry Swofford
Jefe, Área de Impresión de Latentes
Centro de Ciencias Forenses de la Defensa
Agencia Forense y Biométrica de la Defensa

Robert Thompson
Director de Programas de Sistemas de Datos Forenses
Oficina de Estándares para el Cumplimiento de la Ley
Instituto Nacional de Estándares y Tecnología

William Thompson
Catedrático de Criminología, Derecho y Sociedad, y de Psicología y Comportamiento Social
Facultad de Derecho de Ecología Social
Universidad de California en Irvine

Rick Tontarski
Científico Jefe
Centro de Ciencias Forenses de la Defensa

Jeremy Triplett
Supervisor de Laboratorio
Laboratorio Forense Central de la Policía del Estado de Kentucky

Richard Vorder Bruegge
Técnico fotográfico senior
FBI

Victor Weedn
Cátedra de Ciencias Forenses
Departamento de Ciencias Forenses
Universidad George Whashington

Robert Wood
Profesor Adjunto y Jefe
Departamento de Oncología Dental
Prótesis Dentales, Oculares y Maxilofaciales
Centro (de investigación) sobre el Cáncer Princess Margaret
Universidad de Toronto

Xiaoyu Alan Zheng
Ingeniero Mecánico
Instituto Nacional de Estándares y Tecnología

BIBLIOGRAFÍA

1. Informes

FDA Guidance (2016), *Adaptive Designs for Medical Device Clinical Studies.* Disponible en: www.fda.gov/downloads/medicaldevices/deviceregulationandguidance/guidancedocuments/ucm446729.pdf.

Forensic Science Commission (2016), *Forensic bitemark comparison complaint filed by National Innocence Project on behalf of Steven Mark Chaney – Final Report.* www.fsc.texas.gov/sites/default/files/FinalBiteMarkReport.pdf.

Innocence Project, *DNA Exonerations in the United States*, www.innocenceproject.org/dna-exonerations-in-the-united-states.

Ministerio del Interior Español (2018*), La relevancia del título oficial del perito criminalístico nombrado por el juez en la jurisdicción penal española*

National Academies of Sciences, Engineering, and Medicine (2015), *Support for Forensic Science Research: Improving the Scientific Role of the National Institute of Justice.* The National Academies Press. Washington DC.

National Commission on Forensic Science (2015), *Defining forensic science and related terms*. www.justice.gov/ncfs/file/786571/download.

National Commission on Forensic Science (2016), «Recommendations to the Attorney General Regarding Use of the Term 'Reasonable Scientific Certainty'». www.justice.gov/ncfs/file/839726/download.

National Institute of Justice (2006), *Status and Needs of Forensic Science Service Providers: A Report to Congress.* www.ojp.usdoj.gov/nij/pubs-sum/213420.htm.

National Physical Laboratory (2010), *A Beginner's Guide to Measurement*. Disponible en: www.npl.co.uk/upload/pdf/NPL-Beginners-Guide-to-Measurement.pdf.

National Research Council (1996), *The Evaluation of Forensic DNA Evidence*, Washington DC, The National Academies Press.

National Research Council (2004), *Forensic Analysis: Weighing Bullet Lead Evidence,* Washington DC, The National Academies Press.

— (2008), *Ballistic Imaging,* Washington DC, The National Academies Press.

— (2009), *Strengthening Forensic Science in the United States: A Path Forward. The National Academies Press.* Washington DC, The National Academies Press.

— (2010), *Biometric Recognition: Challenges and Opportunities,* Washington DC, The National Academies Press.

Organization for Scientific Area Committees (OSAC) (2015), *Research Needs Assessment Form*, «Study to Assess the Accuracy and Reliability of Firearm and Toolmark». Disponible en: www.nist.gov/forensics/osac/upload/FATM-Research-NeedsAssessment_Blackbox.pdf.

U.S. Department of Justice, Office of Justice Programs, National Institute of Justice (1996), *Convicted by Juries, Exonerated by Science: Case Studies in the Use of DNA Evidence to Establish Innocence after Trial.* xxviii.

U.S. Department of Justice (2006), *Office of the Inspector General* (Oficina del Inspector General). *A review of the FBI's handling of the Brandon Mayfield case*, oig.justice.special/s0601/final.pdf.

— (2015), *New NIST Center of Excellence to Improve Statistical Analysis of Forensic Evidence*, www.nist.gov/forensics/center-excellence-forensic052615.cfm.

U.S. Department of Justice. National Institute of Justice (2016), *Report Forensic Science: Fiscal Year 2015 Funding for DNA Analysis, Capacity Enhancement and Other Forensic Activities*.

2. Libros/Artículos

Alosh, M., Fritsch, K., Huque, M., Mahjoob, K., Pennello, G., Rothmann, M., Russek-Cohen, E., Smith, F., Wilson, S. y Yue, L., 2015: «Statistical considerations on subgroup analysis in clinical trials», *Statistics in Biopharmaceutical Research*, vol. 7: 286-303.

Association of Firearm and Tool Mark Examiners, 2011: «Theory of Identification as it Relates to Tool Marks: Revised», *AFTE Journal*, vol. 43(4).

Balding, D.J. y Nichols, R.A., 1994: «DNA profile match probability calculation: how to allow for population stratification, relatedness, database selection and single bands», *Forensic Science International*, vol. 64: 125-140.

Bieber, P., 2014: «Fire investigation and cognitive bias», *Wiley Encyclopedia of Forensic Science*. Disponible en onlinelibrary.wiley.com/doi/10.1002/9780470061589.fsa1119/abstract.

Bieber, F.R., Buckleton, J.S., Budowle, B., Butler, J.M. y Coble M.D., «Evaluation of forensic DNA mixture evidence: protocol for evaluation, interpretation, and statistical calculations using the combined probability of inclusion», *BMC Genetics*. www.bmcgenet.biomedcentral.com/articles/10.1186/s12863-016-0429-7.

Bouchet, C., Guillemin, F. y Braincon, S., 1996: «Nonspecific effects in longitudinal studies: impact on quality of life measures», *Journal of Clinical Epidemiology*, vol. 49(1).

Boroditsky, L., 2007: «Comparison and the development of knowledge», *Cognition*, vol. 102.

Bodziak, W. J., 2000: *Footwear Impression Evidence: Detection, Recovery, and Examination*, 2nd ed., Boca Raton, Florida: CRC Press-Taylor & Francis.

Budowle, B., Buscaglia, J. y Perlman, R.S., 2006: «Review of the scientific basis for friction ridge comparisons as a means of identification: committee findings and recommendations», *Forensic Science Communications*, vol. 8.

Budowle, B., Moretti, T.R., Keys, K.M., Koons, B.W. y Smerick, J.B. 1997: «Validation studies of the CTT STR multiplex system», *Journal of Forensic Sciences*, vol. 42, No. 4.

Budowle, B., Moretti, T.R., Baumstark, A.L., Defenbaugh, D.A. y Keys, K.M., 1999: «Population data on the thirteen CODIS core short tandem repeat loci in African Americans, U.S. Caucasians, Hispanics, Bahamians, Jamaicans, and Trinidadians», *Journal of Forensic Sciences*, vol. 44(6).

Budowle, B., Shea, B., Niezgoda, S. y Chakraborty, R., 2001: «CODIS STR loci data from 41 sample populations», *Journal of Forensic Sciences*, vol. 46 (3): 453-489. En julio de 2015 se informó de errores encontrados en la base de datos original. (Erratum, *Journal of Forensic Sciences*, vol. 60, No. 4 (2015): 1114-6, pueden encontrase en www.fbi.gov/about-us/lab/biometric-analysis/codis/expanded-fbi-str-final-6-16-15.pdf.

Buckleton, J.S., Curran, J.M. y Gill, P., 2007: «Towards understanding the effect of uncertainty in the number of contributors to DNA stains» ,*Forensic Science International Genetics*, vol. 1(1).

Butler, J.M., 2015: *Advanced Topics in Forensic DNA Typing: Interpretation*, Waltham, MA: Elsevier/Academic.

Butler, J.M., 2015: «The future of forensic DNA analysis», *Philosophical Transactions of the Royal Society B*, 370: 20140252.

Butler, J.M., 2015: «DNA Error Rates», presentation at the International Forensics Symposium, Washington, D.C. (2015). www.cstl.nist.gov/strbase/pub_pres/Butler-ErrorManagement-DNA-Error.pdf.

Bracht, G.H. y Glass, G.V., 1968: «The external validity of experiments», *American Educational Research Journa*l, vol. 5(4): 437-474.

Bush, M.A., Cooper, H.I. y Dorion, R.B, 2010: «Inquiry into the scientific basis for bitemark profiling and arbitrary distortion compensation», *Journal of Forensic Sciences*, vol. 55(4).

Bush, M.A., Miller, R.G., Bush, PJ. y Dorion, RB., 2009: «Biomechanical factors in human dermal bitemarks in a cadaver model», *Journal of Forensic Sciences*, vol. 54(1).

Carroro, P. y Plebani M., 2007: «Errors in a stat laboratory: types and frequencies 10 years later», *Clinical Chemistry*, vol. 53.

Champod, C., 2014: «Research focused mainly on bias will paralyse forensic science», *Science & Justice*, vol. 54.

Clayton, T.M., Whitaker, J.P., Sparkes, R. y Gill, P., 1998: «Analysis and interpretation of mixed forensic stains using DNA STR profiling», *Forensic Science International*, vol. 91, No. 1: 55-70.

Coble, M.D., Bright, J.A., Buckleton, J.S. y Curran, J.M., 2015: «Uncertainty in the number of contributors in the proposed new CODIS set», *Forensic Science International Genetics*, vol. 19.

Cole, S.A., 2004: «Grandfathering evidence: Fingerprint admissibility rulings from Jennings to Llera Plaza and back again», *American Criminal Law Review,* 41(1189).

Cole, S.A., 2005: «More than zero: Accounting for error in latent fingerprint identification», *Journal of Criminal Law and Criminology*, vol. 95(3): 985-1078.

Conney, L., 2010: «Latent Print Training to Competency: Is it Time for a Universal Training Program?», *Journal of Forensic Identification*, vol. 60: 223-58.

Dror, I.E., Charlton, D. y Peron A.E., 2006: «Contextual information renders experts vulnerable to making erroneous identifications», *Forensic Science International*, vol. 156.

Dror, I.E. y Hampikian G., 2011: «Subjectivity and bias in forensic DNA mixture interpretation», *Science & Justice*, Vol. 51(4).

Dror, I.E., 2016: «A hierarchy of expert performance», *Journal of Applied Research in Memory and Cognition*, Vol. 5: 121-127.

Faigman, D.L., Cheng, E.K., Mnookin, J.L., Murphy, E.E., Sander, J. y Slobogin, C. (Eds.), 2016: *Modern Scientific Evidence: The Law and Science of Expert Testimony, 2015-2016,* Thomson/ West Publishing.

Fregeau, C.J., Bowen, K.L. y Fourney, R.M., 1999: «Validation of highly polymorphic fluorescent multiplex short tandem repeat systems using two generations of DNA sequencers», *Journal of Forensic Sciences*, vol. 44(1): 133-166.

Gaudette, B.D. y Keeping, 1975: «An attempt at determining probabilities in human scalp hair comparisons», *Journal of Forensic Sciences*, vol. 19.

Garrett, B.L. y Neufeld, P.J., 2009: «Invalid forensic science testimony and wrongful convictions», *Virginia Law Review*, vol. 91(1): 1-97.

Giannelli, P.C., 1980: «The admissibility of novel scientific evidence: Frye v. United States, a half-century later», *Columbus Law Review*, vol. 80(6).

Gianelli, P.C. 2003: «The Supreme Court's 'Criminal' Daubert Cases», *Seton Hall Law Review*, vol. 33(1096).

Giannelli, P.G., 2010: «Independent crime laboratories: The problem of motivational and cognitive bias», *Utah Law Review*, 247-266.

Gill, P., Jeffreys, A.J. y Werrett, D.J., 1985: «Forensic application of DNA 'fingerprints'», *Nature*, vol. 318(6046).

Goode, M., 2002: «Some observations on evidence of DNA frequency», *Adelaide Law Review*, vol. 23: 45-77.

Goldstone, R. L., 1994: «The role of similarity in categorization: Providing a groundwork», *Cognition*, vol. 52: 125–157.

Gross S.R., y Shaffer, M., 2012: «Exonerations in the United States, 1989-2012», National Registry of Exonerations. Disponible en: www.law.umich.edu/special/exoneration/Documents/exonerations_us_1989_2012_full_report.pdf.

Hassin, R., 2001: «Making features similar: comparison processes affect perception», *Psychonomic Bulletin & Review*, vol. 8.

Hofgartner, W.T. y Tait, J.F., 1999: «Frequency of problems during clinical molecular-genetic testing», American Journal of Clinical Pathology, vol. 112.

Houck, M.M.,y B. Budowle, 2002: «Crrelation of microscopic and mitochondrial DNA hair comparisons», *Journal of Forensic Sciences*, vol. 47(5).

Kassin, S.M., Dror I.E. y Kakucka, J., 2013: «The forensic confirmation bias: Problems, perspectives, and proposed solutions», *Journal of Applied Research in Memory and Cognition*, vol. 2 (1): 42-52.

Kaye D.H., 1993: «DNA Evidence: Probability, Population Genetics, and the Courts», *Harv. J.L. & Tech*, vol 7: 101-172.

Keijser, J.W., Malsch, M., Luining, E.T., Kranenbarg, M.W. y Lenssen, D.J.H.M., 2016: «Differential reporting of mixed DNA profiles and its impact on jurists' evaluation of evidence: An international analysis», *Forensic Science International: Genetics*, vol. 23.

Kim, J., Novemsky, N. y Dhar, R.2012: «Adding small differences can increase similarity and choice», *Psychological Science*, vol. 24.

Kieser, J.A., Bernal, V., Neil Waddell, J. y S. Raju, 2007: «The uniqueness of the human anterior dentition: a geometric morphometric analysis», *Journal of Forensic Sciences*, vol. 52.

Koehler, J.J., 2016: «Intuitive error rate estimates for the forensic sciences». Disponible en: papers.ssrn.com/sol3/papers.cfm?abstract_id=2817443

Koehler, J.J., 2016: «Forensics or fauxrensics? Ascertaining accuracy in the forensic sciences», papers.ssrn.com/sol3/papers.cfm?abstract_id=2773255

Kimpton, C.P., Oldroyd, N.J., Watson, S.K., Frazier, R.R., Johnson, P.E., Millican, E.S., Urguhart, A., Sparkes, B.L. y Gill, P., 1996: «Validation of highly discriminating multiplex short tandem repeat amplification systems for individual identification», *Electrophoresis*, vol. 17(8).

Kloosterman, A., Sjerps, M. y Quak, A., 2014: «Error rates in forensic DNA analysis: Definition, numbers, impact and communication», Forensic Science International: Genetics, vol. 12.

Koppl, R. y Krane, D., 2016: «Minimizing and leveraging bias in forensic science.» En Robertson C.T., y A.S. Kesselheim (Eds.), *Blinding as a solution to bias: Strengthening biomedical science, forensic science, and law.* Atlanta, GA: Elsevier.

Krane, D.E., Ford, S., Gilder, J., Iman, K., Jamieson, A., Taylor, M.S. y Thompson, W.C., 2008: «Sequential unmasking: A means of minimizing observer effects in forensic DNA interpretation», *Journal of Forensic Sciences*, vol. 53(4).

Krimsky, S. y T. Simoncelli, 2011: *Genetic Justice: DNA Data Banks, Criminal Investigations, and Civil Liberties,* Columbia University Press.

Lander, E.S., 1994: «DNA fingerprinting on trial», *Nature*, vol. 339: 501-505.

Lander, E.S. y B. Budowle, 1994: «DNA fingerprinting dispute laid to rest», *Nature*, vol. 371: 735-738.

Langleben, D.D. y Moriarty, J.C., 2013: «Using brain imaging for lie detection: Where science, law, and policy collide», *Psychology, Public Policy, and Law,* vol. 19(2).

Langenburg, G., 2009: «A performance study of the ACE-V process: A pilot study to measure the accuracy, precision, reproducibility, repeatability, and biasability of conclusions resulting from the ACE-V process», *Journal of Forensic Identification*, vol. 59(2).

Larkey, L.B., y Markman, A.B., 2005: «Processes of similarity judgment», *Cognitive Science*, vol. 29.

Lygo, J.E., Johnson, P.E., Holdaway, D.J., Woodroffe, S., Whitaker, J.P., Clayton, T.M., Kimpton, C.P. y Gill, P.: 1994: «The validation of short tandem repeat (STR) loci for use in forensic casework», *International Journal of Legal Medicine*, vol. 107(2).

Mangione-Smith, R., Elliott, M.N., McDonald, L. y McGlynn, E.A., 2002: «An observational study of antibiotic prescribing behavior and the Hawthorne Effect», *Health Services Research*, vol. 37(6): 1603-1623.

McCabe, J., 1996: «DNA fingerprinting: The failings of Frye», *Northern Illinois University Law Review*, vol. 16: 455-82.

McCarney, R., Warner, J., Iliffe, S., van Haselen, R., Griffin, M. y Fisher, P., 2007: «The Hawthorne Effect: a randomized, controlled trial», *BMC Medical Research Methodology*, vol. 7(30).

Medin, D.L., Goldstone, R.L. y Gentner, D.,1993: «Respects for similarity», *Psychological Review*, vol. 100.

Medin, D.L., Goldstone, R.L. y Markman, A.B., 1995: «Comparison and choice: Relations between similarity processes and decision processes», *Psychonomic Bulletin and Review*, vol. 2: 1–19.

Miller, L.S., 1987: «Procedural bias in forensic examinations of human hair», *Law and Human Behavior*, vol. 11.

Moretti, T.R., Baumstark, A.L., Defenbaugh, D.A., Keys, K.M., Smerick, J.B. y Budowle, B., 2001: «Validation of Short Tandem Repeats (STRs) for forensic usage: performance testing of fluorescent multiplex STR systems and analysis of authentic and simulated forensic samples», *Journal of Forensic Sciences*, vol. 46(3): 647-660.

Morrison, G.S., Zhang, C. y Rose, P., 2011: «An empirical estimate of the precision of likelihood ratios from a forensic-voicecomparison system», *Forensic Science International*, vol. 208: 59–65.

Mujis, D., 2006: «Measuring teacher effectiveness: Some methodological reflections», *Educational Research and Evaluation*, vol. 12(1): 53–74.

Nosofsky, R. M., 1986: «Attention, similarity, and the identification categorization relation», *Journal of Experimental Psychology*, General, vol. 115: 39–57.

Mnookin, J.L., Cole, S.A., Dror, I.E., Fisher, B.A.J., Houck, M.M., Inman, K., Kaye, D.H., Koehler, J.J., Langenburg, G., Risinger, D.M., Rudin, N., Siegel, J. y Stoney D.A., 2011: «The need for a research culture in the forensic sciences», *UCLA Law Review*, vol. 725.

Neufeld, P.J. y Colman, N., 1991: «When science takes the witness stand», *Scientific American*, vol. 262: 46-53.

Page, M., Taylor, J. y Blenkin, M., 2011: «Forensic identification science evidence since Daubert: Part II—judicial reasoning in decisions to exclude forensic identification evidence on grounds of reliability», *Journal of Forensic Sciences*, vol. 56(4).

Pavese, F., 2009: «An Introduction to Data Modelling Principles in Metrology and Testing», en Pavese, F. y Forbes A.B. (Eds.), *Data Modelling for Metrology and Testing in Measurement Science*, Birkhäuser.

Perlin, M.W., Hornyak, J.M., Sugimoto, G. y Miller, K.W.P., 2015: «TrueAllele genotype identification on DNA mixtures containing up to five unknown contributors», *Journal of Forensic Sciences*, vol. 60(4).

Peterson, J.L., Lin, G., Ho, M., Chen, Y. y Gaensslen, R.E, 2003: «The feasibility of external blind DNA proficiency testing. II. Experience with actual blind tests», *Journal of Forensic Science*, vol. 48(1).

Petraco, N.D., Shenkin, P., Speir, J., Diaczuk, P., Pizzola, P.A., Gambino, C. y Petraco, N., 2012: «Addressing the National Academy of Sciences' Challenge: A Method for Statistical Pattern Comparison of Striated Tool Marks», *Journal of Forensic Science*s, vol. 57.

Plebani, M. y Carroro, P., 1997: «Mistakes in a stat laboratory: types and frequency», *Clinical Chemistry*, vol. 43.

Pocock, S.J., 1983: *Clinical trials: a practical approach.* Wiley, Chichester.

Pretty, I.A., 2011: «Resolving Issues in Bitemark Analysis», en *Bitemark Evidence: A Color Atlas,* R.B.J Dorian, Chicago: CRC Press.

Pretty, I.A. y Sweet, D., 2010: «A paradigm shift in the analysis of bitemarks», *Forensic Science International*, vol. 201.

Risinger, D.M., Thompson, W.C., Jamieson, A., Koppl, R., Kornfield, I., Krane, D., Mnookin, J.L., Rosenthal, R., Saks, M.J. y Zabell, S.L., 2014: «Regarding Champod, editorial: «Research focused mainly on bias will paralyse forensic science», *Science and Justice,* vol. 54.

Riva, F. y C. Christope, 2014: «Automatic comparison and evaluation of impressions left by a firearm on fired cartridge cases», *Journal of Forensic Sciences*, vol. 59.

Roberts, L., 1991: «Fight erupts over DNA fingerprinting», *Science*, vol. 254: 1721-1723.

Saks, M.J. y Koehler, J.J., 2005: «The coming paradigm shift in forensic identification science», *Science*, vol. 309 (5736): 892-895.

Saks, M.J. y Koehler J.J., 2008: «The individualization fallacy in forensic science evidence», *Vanderbilt Law Review*, vol. 61(1): 199-218.

Stacey, R.B., 2005: «Report on the erroneous fingerprint individualization in the Madrid train bombing case», *Forensic Science Communications*, vol. 7.

Stahl, M., Lund, E.D. y Brandslund, I., 1998: «Reasons for a laboratory's inability to report results for requested analytical tests», *Clinical Chemistry*, vol. 44.

Tangen, J.M., Thompson, M.B. y McCarthy, D.J., 2011: «Identifying fingerprint expertise», *Psychological Science*, vol. 22(8).

Thompson, W.C. y Ford, S., 1990: «Is DNA fingerprinting ready for the courts?», *New Scientist*, vol. 125: 38-43.

Thompson, W.C., Taroni F. y Aitken, C.G.G., 2003: «How the Probability of a False Positive Affects the Value of DNA Evidence», *J Forensic Sci*, vol. 48(1).

Thompson, W.C., 2009: «Painting the target around the matching profile: the Texas sharpshooter fallacy in forensic DNA interpretation», *Law, Probability and Risk*, vol. 8(3).

Thompson, W.C., 2013: «The Myth of Infallibility», en Sheldon Krimsky & Jeremy Gruber (Eds.), *Genetic Explanations: Sense and Nonsense,* Harvard University Press.

Thompson, S.G., 2015: *Cops in Lab Coats: Curbing Wrongful Convictions through Independent Forensic Laboratories.* Carolina Academic Press.

Tversky, A., 1977: «Features of similarity», *Psychological Review*, vol. 84: 327–52.

Ulery, B.T., Hicklin, R.A., Buscaglia, J. y Roberts, M.A., 2011: «Accuracy and reliability of forensic latent fingerprint decisions», *Proceedings of the National Academy of Sciences*, vol. 108(19).

Weech, T.L. y Goldhor, H.,1982: «Obtrusive versus unobtrusive evaluation of reference service in five Illinois public libraries: A pilot study», *Library Quarterly: Information, Community, Policy*, vol. 52, No. 4: 305-324.

Wertheim, Kasey, 2002: «Letter re: ACE-V: Is it scientifically reliable and accurate?», *Journal of Forensic Identification*, vol. 52.

Wilson, H.D, 2012: «Comparison of the individual characteristics in the outsoles of thirty-nine pairs of Adidas Supernova Classic shoes», *Journal of Forensic Identification*, vol. 62(3).

FRECUENTES SIGLAS EN INGLÉS TRADUCIDAS AL ESPAÑOL

ABFO	*American Board of Forensic Odontology* (Junta Directiva Norteamericana de Odontología Forense).
DENL	*Department of Energy National Laboratories* (Laboratorios nacionales del Departamento de Energía).
DOD	*Department of Defense* (Departamento de Defensa).
DOJ	*Department of Justice* (Departamento de Justicia).
FBIL	*Federal Bureau of Investigation Laboratory* (Laboratorio de la Oficina Federal de Investigación).
FJC	*Federal Judicial Center* (Centro Judicial Federal).
FSC	*Defense Department's Forensic Science Center* (Centro de Ciencias Forenses del Departamento de Defensa).
FSSB	*Forensic Science Standards Board* (Junta Directiva de Estándares de Ciencia Forense).
FRE	*Federal Rules of Evidence* (Reglas Federales de la Prueba).
JCUS	*Judicial Conference of the United States* (Conferencia Judicial de los Estados Unidos).
MRC	*Metrology Resource Committee* (Comité de Recursos de Metrología).
NASEM	*National Academies of Science, Engineering, and Medicine* (Academias Nacionales de Ciencias, Ingeniería y Medicina).
NDIS	*National DNA Index System* (Sistema de Reseña de ADN Nacional).
NCFS	*National Commission on Forensic Science* (Comisión Nacional sobre Ciencia Forense).
NIJ	*National Institute of Justice* (Instituto Nacional de Justicia).
NIST	*National Institute of Standards and Technology* (Instituto Nacional de Estándares y Tecnología).
NSF	*National Science Foundation* (Fundación de la Ciencia Nacional).
NRC	*National Research Council* (Consejo de Investigación Nacional).
NSTC	*National Science and Technology Council* (Consejo de Ciencia y Tecnología Nacional).
OSAC	*Organization for Scientific Area Committees* (Organización de Comités de Áreas Científicas).
OSTP	*Office of Science and Technology Policy* (Oficina de Políticas Científicas y Tecnológicas).
PCAST	*President's Council of Advisors on Science and Technology* (Consejo de Asesores del Presidente en Ciencia y Tecnología).
QAS	*Quality Assurance Standards* (Estándares de Garantía de la Calidad).