InDret
REVISTA PARA EL ANÁLISIS DEL DERECHO
WWW. INDRET.COM

El uso de valoraciones del riesgo de violencia en Derecho Penal: algunas cautelas necesarias

Lucía Martínez Garay
Departamento de Derecho Penal, Universitat de València

Francisco Montes Suay
Departamento de Estadística e Investigación Operativa, Universitat de València

Valencia, abril 2018

Abstract[1]

La valoración del riesgo de violencia o de reincidencia con herramientas estructuradas, que cada vez está más presente en el ámbito de la justicia penal, es un procedimiento complejo, y para comprender correctamente sus resultados es necesario un conocimiento mínimo de algunos conceptos estadísticos. La forma en que usualmente se presenta en la literatura criminológica especializada la información sobre la capacidad de estas herramientas para predecir con acierto el riesgo de violencia no favorece una correcta interpretación de su alcance, pues se tiende a destacar aquellos aspectos en que estas herramientas funcionan mejor, y a subrayar mucho menos aquéllos otros en que sus resultados son más pobres, haciendo hincapié además en indicadores que pueden tener escasa relevancia para la práctica judicial. Este trabajo pretende ilustrar sobre la complejidad de esta problemática con el ejemplo del área bajo la curva ROC (AUC), y advertir sobre el peligro de que se extienda un optimismo injustificado sobre el rendimiento de estas herramientas, que en el ámbito forense podría favorecer la toma de decisiones limitativas de derechos fundamentales de los acusados o condenados carentes de suficiente fundamento, y en el ámbito de la política criminal puede alentar la proliferación de instituciones jurídicas cuyo presupuesto sea el riesgo de reincidencia en la creencia de que dicho riesgo puede comprobarse empíricamente con facilidad, cuando ello no es así.

Violence or recidivism risk assessment using actuarial tools, which is gaining importance in the field of criminal justice, is a complex procedure, and some basic knowledge of statistics is required in order to correctly understand its results. The way in which the information about the predictive accuracy of structured risk assessment tools is presented in the specific criminological literature does not contribute to a correct interpretation of its scope, because the published studies tend to emphasize the accuracy indicators that obtain satisfactory values, and less so those that achieve lower scores, focusing, in addition, on indicators that may be of little or no relevance in the criminal judicial practice. This paper tries to illustrate the complexity of this issue with the example of the area under the ROC curve (AUC), and warns against unjustified optimism around the predictive accuracy of these tools that in the criminal process could lead to adopting decisions that restrict fundamental rights of the accused without enough justification, and in the wider context of public policy could contribute to the spread of penalties or other punitive measures that are based on the probability of future recidivism, relying on the assumption that this probability can be estimated easily, which is not the case.

Title: Using risk assessment tools in criminal law: some necessary cautions

Key words: dangerousness, criminal law, risk assessment, sensitivity, specificity, predictive value, area under the ROC curve (AUC)

Palabras clave: peligrosidad, Derecho penal, valoración del riesgo, sensibilidad, especificidad, valor predictivo, área bajo la curva ROC (AUC)

[1] La investigación de Lucia Martínez Garay ha sido financiada por los Proyectos de I+D+I DER2013-47859-R (Ministerio de Economía y Competitividad) y GV/2016/085 (Generalitat Valenciana). Se benefició también de una estancia de investigación en el *Max Planck Institut für ausländisches und internationales Strafrecht* (Freiburg, Alemania) realizada entre enero y febrero de 2015, financiada con una Beca de la Consellería de Educación, Cultura y Deporte de la Generalitat Valenciana.

Sumario:

> *"While everyone will agree that risk assessment in mental health should be evidence based, there are problems with what constitutes the 'evidence' and the way it is reported. […] We aim to clarify for a non-specialist audience what the evidence means."* (SZMUKLER / EVERITT / LEESE, 2012)

1. Introducción

La legislación penal y penitenciaria ha considerado desde siempre la peligrosidad del reo o del condenado como uno de los criterios a valorar a la hora de tomar determinadas decisiones, tanto en el seno del proceso penal (por ejemplo, la imposición de medidas de seguridad, o la decisión sobre la suspensión de la pena) como en el ámbito penitenciario (por ejemplo, la progresión de grado, la concesión de permisos, o la concesión de la libertad condicional). Para tomar estas decisiones los operadores jurídicos se han servido generalmente de informes o valoraciones emitidos por expertos: principalmente médicos psiquiatras o forenses en el ámbito de la imposición de medidas de seguridad, por un lado, y por otro lado los diferentes especialistas que conforman los equipos de valoración y las juntas de tratamiento en el ámbito penitenciario para las decisiones que corresponden a la fase de ejecución de las penas. Estos informes tradicionalmente se han elaborado según lo que se conoce como método clínico puro o no estructurado, esto es, conforme al conocimiento y experiencia profesional individual de cada especialista, sin sujeción a reglas fijas o protocolos estrictos que estandaricen la metodología a emplear ni los factores a tener en cuenta.

El ámbito de la valoración de la peligrosidad ha constituido por tanto desde siempre un punto de encuentro entre la Criminología y el Derecho Penal, en el que creemos poder afirmar que hasta los años 80 del siglo XX dominaba en ambas disciplinas un ambiente de cautela. Se asumía que era necesario analizar la peligrosidad en algunos casos porque las leyes obligaban a hacerlo, pero siendo conscientes de que no era posible pronunciarse sobre ello con certeza. La peligrosidad se consideraba algo difícil de definir, los factores que la condicionan múltiples, variables y complejos, y el conocimiento que sobre ello puede alcanzarse sólo muy aproximado (VIVES ANTÓN, 1974; STEADMAN / COCOZZA, 1978). Cuando además en los años 70 algunos estudios constataron que los juicios sobre la elevada peligrosidad de ciertos grupos de sujetos quedaban desmentidos cuando se tenía la oportunidad de confrontarlos con su conducta posterior, se extendió el escepticismo sobre las posibilidades de efectuar juicios rigurosos sobre la peligrosidad de los delincuentes. Son conocidas las lapidarias frases de MONAHAN o DIAMOND señalando que los peritos psiquiatras y psicólogos se equivocan en una de cada tres predicciones de peligrosidad que realizan, y que no existe ningún estudio que demuestre que sea posible predecir el comportamiento realmente violento con algún grado de aceptabilidad[2].

[2] "The best clinical research currently in existence indicates that psychiatrists and psychologists are accurate in no more than one out of three predictions of violent behavior over a several-year period" (MONAHAN 1981: 77); "There seems to be no convincing study to show that we can predict really dangerous behavior with any amount of acceptability" (DIAMOND, 1974:451). En este último pasaje DIAMOND reproduce palabras atribuidas a N. CHRISTIE. Y unas pocas líneas después añade su propia opinión, coincidente: "Neither psychiatrists nor other behavioral scientists are able to predict the occurrence of violent behavior with sufficient reliability to justify the restriction of freedom of persons on the basis of the label of potential dangerousness. Accordingly, it is

Pero es importante subrayar que en ese momento histórico el escepticismo se proyectaba especialmente sobre la capacidad *de los peritos clínicos* para hacer juicios fiables de peligrosidad. Es decir, lo que se cuestionaba no era tanto la posibilidad en sí de efectuar estimaciones de peligrosidad, sino el que los médicos, psiquiatras y psicólogos estuviesen capacitados para emitir sobre ello juicios basados en los conocimientos científicos propios de su disciplina. El *amicus brief* que la ASOCIACIÓN AMERICANA DE PSIQUIATRÍA (en adelante, APA) remitió en 1982 al Tribunal Supremo de EEUU en el conocido caso *Barefoot vs Estelle*, desautorizando las opiniones que dos peritos habían emitido sobre la peligrosidad del condenado a muerte[3], refleja esta diferencia con claridad. Tras afirmar que la falta de fiabilidad de las predicciones psiquiátricas de peligrosidad a largo plazo "es a día de hoy un hecho reconocido en la profesión", la APA añadió que:

> "El hecho de que los psiquiatras sean incapaces de predecir el comportamiento futuro violento no significa que estas predicciones no se puedan hacer nunca. […] Lo que sostenemos, por el contrario, es que la predicción a largo plazo de la peligrosidad es esencialmente una determinación no especializada que no debería estar fundamentada en los diagnósticos u opiniones de expertos médicos, sino sobre la base de información prospectiva estadística o actuarial que es sustancialmente de naturaleza no médica." (APA, 1982:5)

Pues bien, la situación desde entonces ha evolucionado de manera muy significativa. En la Criminología la idea de "peligrosidad"[4] ha sido sustituida por el enfoque de la "valoración del riesgo" (*violence risk assessment*), según el cual de lo que se trata no es de averiguar si el individuo posee o no la cualidad subjetiva de peligroso (como si se le estuviera diagnosticando una enfermedad), sino de valorar un conjunto de factores, personales pero también ambientales, y cambiantes en el tiempo, que favorecen en mayor o menor grado la comisión de nuevos delitos y que permiten efectuar pronósticos sobre la reincidencia futura en términos de probabilidad: muy alta, alta, moderaba, baja, muy baja, etc (STEADMAN, 2000; ANDRÉS PUEYO / REDONDO ILLESCAS, 2007; ANDRÉS PUEYO / ECHEBURÚA, 2010; ANDRÉS PUEYO, 2013; LOINAZ, 2017:73 y s.). Y el cambio en la conceptuación de aquello que se valora – de la peligrosidad al riesgo – ha ido acompañado de un cambio en la forma de medir ese riesgo. El tradicional método clínico no estructurado, que se considera poco fiable, sesgado, y carente de suficiente base empírica, ha ido siendo sustituido por métodos estructurados de estimación del riesgo[5], ya se trate de métodos estrictamente actuariales, o bien de instrumentos de juicio clínico estructurado. Ambos tipos de herramientas

---

recommended that courts no longer ask such experts to give their opinion of the potential dangerousness of any person" (DIAMOND, 1974:452).

[3] Sobre las circunstancias del caso *Barefoot vs Estelle* (1983) cfr. LOINAZ, 2017:65 y s. y MARTÍNEZ GARAY, 2014:44 y ss.

[4] Concepto que sin embargo sigue manejando en muchos casos la legislación vigente, cfr. arts. 6, 36.3, 83.3 y 4, 90.5, 91, 92.3 y 97 (entre otros) del Código Penal español; el parágrafo 66c(1)1 del Código Penal alemán (*Gefährlichkeit*); los arts. 64c, 75a, 139 y 140 del Código Penal suizo (*Gefährlichkeit*, *Gemeingefährlichkeit*) o los arts. 203 y 108 del Código Penal italiano (*pericolosità*). Aunque las leyes penales también utilizan otras expresiones, como "pronóstico de reinserción social" (por ej., arts. 36.1 o 92.1 CP), en relación con las cuales se ha discutido si significan o no lo mismo que "peligrosidad", o incluso si en todos los casos en los cuales el Código penal habla de "peligrosidad" lo hace en el mismo sentido (cfr. CERVELLÓ DONDERIS, 2014)

[5] Conviene advertir que esta evolución se ha producido principalmente en el ámbito de la investigación criminológica; en la práctica forense es aún el método clínico el que continua siendo utilizado por la mayor parte de los profesionales cuando emiten informes sobre la valoración de casos concretos, al menos en España (ARBACH-LUCIONI et al, 2015:359; ANDRÉS-PUEYO / ECHEBURÚA, 2010:404 y s.). Y también en EEUU parece que los tribunales siguen admitiendo sin problemas el juicio clínico como prueba sobre la peligrosidad, e incluso con mayor facilidad que periciales basadas en la aplicación de herramientas estructuradas de valoración del riesgo (KRAUSS / SCURICH, 2013; SCURICH, 2016).

de predicción del riesgo están basados en la observación empírica de grupos de sujetos y en la cuantificación y combinación estadística de los factores de riesgo (y, en algunos casos, de factores protectores) que concurren en ellos, y que demuestran estar significativamente asociados a la aparición de conducta violenta y/o delictiva.

> La principal diferencia entre los instrumentos puramente actuariales y los de juicio clínico estructurado es que los primeros proporcionan una estimación de la peligrosidad automática, calculada con un algoritmo, a partir de la puntuación que el sujeto haya obtenido en los diferentes factores de riesgo que incluya el instrumento. Las herramientas de juicio clínico estructurado, por el contrario, son guías que indican cuáles son los factores que deben ser tenidos en cuenta en la valoración del nivel de riesgo, y cómo deben ser apreciados, pero permiten que el profesional pondere su peso relativo con mayor libertad, e incluso añada otros factores que le parezcan decisivos en el caso concreto aunque la herramienta no los contemple (cfr. sobre ello, con más detalles, HEILBRUN, 2009:53 y ss.; LOINAZ, 2017:201 y ss.)

Con el cambio de enfoque y de metodología la investigación sobre la valoración del riesgo de violencia ha experimentado un crecimiento espectacular. Aunque el empleo de tablas estadísticas y procedimientos actuariales en la valoración del riesgo en el ámbito penal tenía importantes precedentes desde al menos los años 30 del siglo pasado (HARCOURT, 2007:39 y ss; LOINAZ, 2017:41 y ss.) ha sido desde mitad de los años 80 a nivel internacional – especialmente en Canadá y EEUU –, y desde hace 10 o 15 años también en España, cuando se han desarrollado decenas de herramientas estructuradas de valoración de diferentes tipos de riesgo (de reincidencia en general, reincidencia violenta, violencia sexual, violencia contra la pareja, riesgo de reincidencia en jóvenes, en pacientes psiquiátricos, etc.), y se han publicado centenares de artículos de investigación sobre el tema.

Pero tan significativo o más que el aumento del interés científico sobre esta cuestión es el notable cambio de opinión que ha tenido lugar en cuanto a la valoración de los resultados de toda esta investigación. En efecto, en la Criminología actual aquel pesimismo al que nos hemos referido líneas arriba ha dado paso a un moderado (pero firme) optimismo sobre la capacidad predictiva de estos instrumentos, y sobre su utilidad en el ámbito forense (ANDRÉS PUEYO / REDONDO ILLESCAS, 2007:169; ANDRÉS PUEYO / ECHEBURÚA, 2010:408). STEADMAN, uno de los autores que más se significó por sus estudios críticos sobre el estado de las predicciones de peligrosidad hace cuarenta años, afirmaba sin embargo en el año 2000 que "incluso teniendo en cuenta las diversas limitaciones del estado actual del conocimiento, existe espacio [hoy en día] para un optimismo que habría resultado inapropiado en 1970" (STEADMAN, 2000:270).

Pues bien, si volvemos la mirada desde el campo de la Criminología al del Derecho penal, observaremos que también en este ámbito el riesgo de comisión de futuros delitos es una circunstancia que en los últimos tiempos ha ido adquiriendo cada vez mayor relevancia. Desde los años 90 ha aumentado en varios países el número o el ámbito de aplicación de consecuencias jurídicas restrictivas de derechos que dependen de la peligrosidad del sujeto (son ejemplos muy conocidos los *civil commitment* para *sexual violent predators* en EEUU, o las sucesivas reformas de la custodia de seguridad en Alemania). En España, a las cuestiones que desde siempre han dependido de la valoración de la peligrosidad (la suspensión de la condena, la concesión de permisos penitenciarios, la libertad condicional y las medidas de seguridad para inimputables) se han añadido en los últimos años otras nuevas. Por una parte, la libertad vigilada como medida de seguridad post-condena para delincuentes imputables cuya peligrosidad subsista tras el

cumplimiento de la pena de prisión[6]. Por otro lado, la peligrosidad es el criterio esencial del que depende la revisión de la pena de prisión permanente revisable, desde que esta pena se ha introducido en nuestro ordenamiento jurídico en el año 2015 (art. 92.1 CP). Y también la inclusión de identificadores de ADN en la base de datos policial prevista en el nuevo art. 129 bis CP depende de la constatación de la peligrosidad del sujeto.

Parecería natural entonces que ya que tantas decisiones en la justicia penal dependen de una estimación del riesgo de reincidencia del sujeto, y que sobre esta cuestión ha avanzado tanto la investigación criminológica en los últimos años, se incorporen dichos avances al ámbito forense y se utilicen los nuevos instrumentos estructurados como ayuda para tomar estas decisiones. Qué mejor, podríamos pensar, que en lugar de tomar estas decisiones sobre la base de estimaciones de la peligrosidad subjetivas, seguramente arbitrarias, y sesgadas por prejuicios, lo hagamos apoyándonos en estimaciones científicas sobre el riesgo de reincidencia, sustentadas en un sólido acervo de conocimientos empíricos sobre el tema, y objeto de múltiples investigaciones internacionales publicadas en revistas de prestigio.

Sin embargo, el propósito que persigue este trabajo es advertir que las cosas no son, por desgracia, tan sencillas. De un lado, porque el conocimiento científico sobre el riesgo de reincidencia no es tan exacto y preciso como podría parecer. Y de otro, porque es además muy fácil de malinterpretar por quien no es experto en la materia. En relación con esto último es interesante recordar de nuevo las advertencias que la APA hacía en su informe de 1982 sobre la falta de competencia de los psiquiatras y psicólogos para emitir dictámenes clínicos sobre peligrosidad. En aquel informe, al que nos hemos referido al inicio de este apartado, además de subrayar el elevado número de errores de las estimaciones clínicas sobre peligrosidad y su falta de fiabilidad, la APA señalaba otra razón para desaconsejarlas: el peligro de que pudieran influir indebidamente en los jurados, porque el prestigio y la autoridad de que está revestido un profesional de la medicina o de la psiquiatría favorecería que los ciudadanos que conforman el jurado den a sus opiniones una credibilidad excesiva, a pesar de que éstas no estén correctamente fundamentadas:

> "es probable que un jurado dé un peso decisivo al testimonio de un psiquiatra simplemente porque es, o pretende ser, la declaración de una opinión profesional. El psiquiatra entra en la sala rodeado de un aura de experto que inevitablemente aumenta la credibilidad, y con ello el impacto, de su testimonio. Como se afirma en una reciente resolución de un tribunal de distrito relacionada precisamente con esta cuestión, cuando un pronóstico de peligrosidad 'es emitido por un perito que posee el título de 'doctor', su impacto sobre el jurado es mucho mayor que si no estuviera disfrazado como algo que no es'" (APA, 1982:6)

La APA ha reiterado esa misma posición en *amici briefs* emitidos en fechas más recientes, igualmente relacionados con supuestos de pena capital – casos *US v. Fields*, 2007, y *Coble v. Texas*, 2011 – y sigue insistiendo en recomendar el uso de instrumentos estructurados de valoración del riesgo, que no sólo tendrían una base científica más sólida que los juicios clínicos sobre

[6] Introducida por la LO 5/2010 sólo para delitos contra la libertad e indemnidad sexuales (art. 192.1 CP) y de terrorismo (art. 579 bis.2 CP), se ha ampliado en virtud de la LO 1/2015 a otras figuras como los delitos contra la vida (art. 140 CP), y algunos supuestos de violencia de género y doméstica (art. 173.2 in fine, art. 156 ter CP). En esta última reforma el Proyecto presentado por el Gobierno en 2013 preveía y ampliar la libertad vigilada a muchos más delitos y también eliminar el límite máximo de cumplimiento para la medida de seguridad de internamiento en centro psiquiátrico, aunque finalmente en el trámite parlamentario estas modificaciones decayeron.

peligrosidad, sino que además evitarían estos peligros de condicionamiento indebido de los jurados (APA, 2007, 2011).

Pues bien, en nuestra opinión esos peligros de predisposición o condicionamiento indebido existen igualmente en las herramientas estructuradas de estimación del riesgo. Y ello porque también los números, las cifras, sobre todo según cómo sean presentadas, tienen un enorme poder para sugestionar no sólo a los jurados sino también a los operadores jurídicos y a los juristas en general. Poder que es tanto mayor cuanto menores sean los conocimientos especializados en matemáticas o en estadística del público al cual se dirija la información. En las páginas que siguen trataremos de demostrar que esto puede provocar en los operadores jurídicos una confianza exagerada e injustificada en la fiabilidad de las estimaciones de riesgo hechas con herramientas estructuradas, haciéndoles creer que son mejores y más precisas de lo que en realidad son. Y esto es peligroso porque un optimismo excesivo sobre el rendimiento de estas estimaciones de riesgo podría facilitar el que se adopten decisiones limitativas de derechos fundamentales de los acusados o condenados pensando que gozan de suficiente soporte empírico, cuando en realidad puede que no sea así.

Por último este trabajo también pretende advertir frente al peligro de que la sofisticación estadística y la consideración como "científicas" de las herramientas de valoración del riesgo favorezca el que se extendiera la ilusión de que su uso convierte en meras decisiones técnicas los seculares problemas valorativos que siempre han estado implicados, y lo siguen estando, en toda decisión que suponga aplicar consecuencias penales a un individuo basadas en un juicio acerca de su comportamiento en el futuro: nos referimos a la ponderación entre el respeto a los derechos fundamentales del reo, por un lado, y las necesidades de prevención por el otro. El uso de los protocolos de valoración del riesgo en Derecho penal ha de estar rodeado de enormes cautelas para no poner en peligro las garantías penales.

2. Capacidad predictiva de los instrumentos estructurados de valoración del riesgo: una cuestión compleja

Supongamos que un juez tiene que tomar una decisión sobre un sujeto, en la que uno de los criterios a considerar es la estimación del riesgo; por ejemplo, si impone o no una medida de seguridad. Y recibe un informe que le indica lo siguiente: "Al señor S. se le ha aplicado el instrumento de valoración del riesgo X, y la puntuación total que ha obtenido lo clasifica dentro del grupo de sujetos con un riesgo alto de reincidencia sexual". Probablemente una vez conocida esa información a este hipotético juez le interesaría saber dos cosas: la primera, cómo de alto es exactamente el "riesgo alto" de reincidencia, y la segunda, cómo de buena (fiable, precisa) es esta estimación.

Por desgracia, la segunda pregunta no es fácil de contestar. Y no lo es por varias razones. En primer lugar, porque la capacidad predictiva de los instrumentos estructurados de valoración del riesgo se puede expresar con muchos indicadores diferentes, cada uno de los cuales mide una dimensión distinta de esa capacidad predictiva (SINGH, 2013; LOINAZ, 2017:87 y ss.; MUÑOZ VICENTE / LÓPEZ-OSSORIO, 2016). Algunos expresan cómo de bien detecta el instrumento la reincidencia, y otros cómo de bien la predice; algunos indican cómo de bien se estima el riesgo

alto, y otros cómo de bien se estima el riesgo bajo; algunos son medidas de riesgo relativo, y otros de riesgo absoluto, etc.

En segundo lugar, puesto que cada indicador mide una dimensión diferente de la capacidad predictiva, cada uno puede adoptar valores muy distintos para una misma herramienta de estimación del riesgo. Por ejemplo, un mismo instrumento de valoración del riesgo puede tener una sensibilidad muy elevada pero una especificidad baja, o un área bajo la curva ROC aceptable y sin embargo un valor predictivo positivo muy débil. Es muy importante tener esto presente porque significa que la capacidad predictiva de un instrumento de valoración del riesgo podría ser calificada a veces como 'buena' y 'mala' "a la vez", si algunos de los indicadores alcanzan valores muy satisfactorios, y sin embargo otros se quedan en niveles mucho más modestos.

Y, por último, el hecho de que el funcionamiento de una herramienta de valoración del riesgo pueda ser descrito con muchos indicadores diferentes hace que la información que se proporciona sobre su rendimiento pueda ser fácilmente malinterpretada, especialmente por quien no conozca con precisión el significado de dichos parámetros, en función de qué datos se le faciliten y cómo se le presenten. Veamos todo ello con un ejemplo.

La Tabla 1 muestra los resultados obtenidos en un metaanálisis en el que se valoró el funcionamiento real de diferentes herramientas de estimación del riesgo aplicadas a distintas muestras de sujetos, comparando el valor que arrojaron diversos indicadores de su capacidad predictiva[7].

[7] Es importante advertir que los que aparecen en la Tabla 1 no son ni mucho menos los únicos indicadores de capacidad predictiva de las herramientas de valoración del riesgo que existen: por el contrario, y sólo por mencionar algunos más, cabría referirse a la *d* de Cohen, la *odds ratio*, la *r* de Pearson, etc. (cfr. sobre sus características y significado LOINAZ 2017 o SINGH 2013).

Tabla 1: Indicadores de capacidad predictiva en tres clases de instrumentos de valoración del riesgo (valores medios obtenidos en un metaanálisis de 68 estudios de validación sobre 73 muestras que abarcan 24.827 sujetos)

	Delincuencia violenta (HCR-20; SARA, SAVRY y VRAG)	Delincuencia sexual (SORAG, Static-99 y SVR-20)	Delincuencia en general (LSI-R y PCL-R)
Sensibilidad (IC 95%)	0.92 (0.88 – 0.94)	0.88 (0.83 – 0.92)	0.41 (0.28 – 0.56)
Especificidad (IC 95%)	0.36 (0.28 – 0.44)	0.34 (0.20 – 0.51)	0.80 (0.67 – 0.8)
Area bajo la curva ROC (mediana (intervalo intercuartílico))	0,72 (0,68 - 0,78)	0,74 (0,66 - 0,77)	0,66 (0,58 – 0,67)
Valor predictivo positivo (mediana (intervalo intercuartílico))	0.41 (0.27 – 0.60)	0.23 (0.09 – 0.41)	0.52 (0.32 – 0.59)
Valor predictivo negativo (mediana (intervalo intercuartílico))	0.91 (0.81 – 0.95)	0.93 (0.82 – 0.98)	0.76 (0.61 – 0.84)

Fuente: FAZEL et al., 2012 (valores extraídos de la Tabla contenida en la p. 10)

Fácilmente puede apreciarse cómo los valores de los diferentes indicadores de capacidad predictiva varían notablemente, no sólo dentro de cada grupo de herramientas (por ejemplo, los instrumentos que miden el riesgo de reincidencia violenta tienen una sensibilidad muy elevada, pero una especificidad muy baja), sino también entre unos grupos y otros de instrumentos de valoración (la sensibilidad es muy elevada en los que miden el riesgo de reincidencia violenta o sexual, pero mucho más baja en los que estiman el riesgo de reincidencia general; sin embargo, respecto de la especificidad ocurre a la inversa).

Volvamos ahora al ejemplo imaginario con el que iniciábamos este epígrafe: el hipotético juez que tiene que decidir sobre la imposición de una medida de seguridad a un sujeto, al que una herramienta de valoración del riesgo ha clasificado en la categoría de riesgo alto de reincidencia sexual. Y supongamos que el juez interroga al perito por la calidad y precisión de dicha estimación de riesgo. Si el perito le contesta que dicha estimación de riesgo de reincidencia se ha hecho con una herramienta que en diversos estudios de validación ha arrojado unos valores medios de área bajo la curva de 0,74, y una sensibilidad del 88%, probablemente el juez asumirá que es una herramienta "buena". Por el contrario, si el perito le informa de que según esos

mismos estudios la herramienta de valoración del riesgo tiene una especificidad de 0,34 y un valor predictivo positivo del 23%, el juez tenderá a considerarla "peor". Porque el mero hecho de saber que algo que puede oscilar entre el 0 y el 100 se ha quedado en el 34% o en el 23% genera en las personas que carecen del conocimiento especializado necesario para valorar la trascendencia exacta de esa información la impresión de que es "peor" que si ha alcanzado el 74% o el 88%.

Interesa subrayar que tanto en un caso como en el otro se le estaría dando al juez información *correcta* sobre la capacidad predictiva del instrumento; la diferencia estriba en cuál de los diversos parámetros con los que ésta puede medirse es el que se ha elegido proporcionarle. Pero como el juez[8] normalmente carecerá de los conocimientos específicos de criminología y de estadística necesarios para saber cuál de los indicadores de capacidad predictiva es más relevante para la decisión que tiene que tomar, carecerá de criterio para decidir cuál de las dos respuestas a su pregunta sobre la precisión de la estimación de reincidencia es más pertinente.

De todo ello deriva que la cuestión decisiva a la hora de juzgar la capacidad predictiva de los instrumentos de valoración del riesgo de reincidencia es saber cuál o cuáles son los indicadores relevantes. Y cuál o cuáles son los indicadores relevantes en cada caso depende del contexto. Una primera diferenciación a este respecto resulta obvia: hay indicadores que pueden ser muy útiles y pertinentes en la investigación criminológica, pero que quizá no sean los más decisivos cuando se trata de adoptar decisiones en el contexto de la justicia penal, pues evidentemente los objetivos que se persiguen en uno y otro ámbito son muy distintos, y también lo son los criterios de legitimidad a que deben sujetarse las decisiones. Puede haber cuestiones esenciales en una decisión judicial (por ejemplo, el respeto a principios como el de proporcionalidad o el de presunción de inocencia) que sean irrelevantes en la investigación empírica, y a su vez puede haber cuestiones que sean fundamentales en este último ámbito, pero que no tengan trascendencia en el judicial o penitenciario. Por otro lado, también dentro del campo de la justicia penal seguramente habrá que diferenciar entre distintos supuestos, pues cada uno de los indicadores de capacidad predictiva podrá ser más o menos relevante en función de la clase de decisión que se tenga que tomar en cada caso.

3. Indicadores de capacidad predictiva relevantes en el ámbito jurídico-penal

¿Cuáles son, pues, los indicadores de capacidad predictiva más relevantes para las decisiones judiciales en las que la valoración del riesgo de reiteración delictiva es uno de los criterios a considerar? Para responder esta pregunta vamos a ponerla en relación con la otra cuestión que habíamos dejado planteada al inicio del epígrafe anterior: qué significa exactamente la expresión

[8] Hablamos de "juez" para simplificar, y porque es la autoridad a quien corresponde adoptar las decisiones sobre la imposición o no de medidas de seguridad, sobre la suspensión o no de la condena, sobre la libertad condicional, y sobre la ratificación o no de las decisiones de las Juntas de Tratamiento en materia penitenciaria. Pero lo mismo cabría decir, probablemente, sobre muchos otros operadores jurídicos, en especial los Fiscales, que han de informar sobre la adopción de todas estas decisiones. Por otro lado, y como se expondrá *infra* con más detalle, también es importante la impresión que se genera en otro tipo de público sobre la capacidad predictiva de las herramientas estructuradas de valoración del riesgo, como por ejemplo los parlamentaros que en un momento dado tienen que decidir sobre la aprobación de una reforma legal que introduce consecuencias jurídicas limitativas de derechos que dependen del grado de peligrosidad del sujeto.

"riesgo alto" de reincidencia, porque ambas cosas están muy relacionadas.

3.1. Diferencia entre riesgo relativo y riesgo absoluto

Hay diferentes formas de explicar lo que significa el "riesgo alto" como posible resultado de una valoración de riesgo hecha con un instrumento estructurado. Una sería decir: "los sujetos que el instrumento X clasifica en un nivel de riesgo alto tienen 10 veces más probabilidad de reincidir que los sujetos que dicha herramienta clasifica como de riesgo bajo". Esta afirmación nos informa sobre el mayor riesgo de reincidir de los clasificados como de riesgo alto *en relación con* los clasificados como de riesgo bajo. Es por tanto una información sobre el mayor *riesgo relativo* que presenta un grupo respecto de otro[9]. Pero es dudoso que esta forma de cuantificar el riesgo alto proporcione al juez la información que éste necesita para muchas de las decisiones que ha de tomar. Porque no es posible hacerse una idea mínimamente aproximada de lo que realmente significa que el riesgo sea "10 veces más alto" si no conocemos cuál es la probabilidad de reincidencia del grupo de comparación. Si el grupo de bajo riesgo tiene una probabilidad de reincidencia del 2%, 10 veces más es sólo 20%. Pero si el grupo de bajo riesgo tiene una probabilidad de reincidencia del 8%, 10 veces más es el 80%. Y la diferencia puede ser muy significativa.

La segunda forma de responder la pregunta de qué significa "riesgo alto" es por tanto proporcionar directamente esas probabilidades de reincidencia que corresponden a cada grupo. Con ello se da información sobre el *riesgo absoluto* de reincidencia de cada grupo, es decir, sobre la probabilidad de que el suceso (la reincidencia) ocurra realmente en ese grupo en un periodo de tiempo dado. Una forma de informar sobre el riesgo absoluto sería por ejemplo decir que el 20% de los sujetos clasificados como de riesgo alto con el instrumento X vuelven a cometer un delito violento en los siguientes 5 años, mientras que sólo lo hacen el 2% de los clasificados con ese mismo instrumento como de riesgo bajo[10].

Puede haber decisiones para las que sea relevante la información sobre el riesgo relativo. Estimo que esto puede ocurrir sobre todo en el ámbito penitenciario, donde hay que gestionar a grupos de personas. Por ejemplo, si los recursos para proporcionar determinado tratamiento o programa en un centro penitenciario son insuficientes para atender a todos los internos que lo solicitan, la Administración puede considerar adecuado reservarlo para el grupo de los que presenten mayor riesgo de reincidencia[11]. En este caso lo importante será saber qué internos tienen mayor riesgo

[9] Utilizamos en este trabajo la expresión 'riesgo relativo' en el sentido amplio de "mayor o menor riesgo de un grupo o individuo en relación con otro grupo o individuo", tal y como lo hemos descrito en el texto. No nos referimos, por tanto, al otro sentido más técnico en que puede utilizarse el término, pues "riesgo relativo" designa también un indicador específico de capacidad predictiva que indica cuántas más veces tiende a ocurrir el evento en el grupo de sujetos expuestos al factor de riesgo, en relación con el grupo no expuesto, y que se expresa en el cociente entre el riesgo en el grupo que presenta la variable problema y el riesgo en el grupo de referencia (LOINAZ, 2017:94).

[10] En el campo de la valoración del riesgo se asume que son aplicables a un individuo concreto las probabilidades de reincidencia del grupo en el que dicho individuo resulta encuadrable por la puntuación que ha obtenido en la herramienta de valoración del riesgo. En los últimos años sin embargo algunos autores han cuestionado la validez de esta premisa, pero se trata de una cuestión compleja en la no es posible profundizar aquí porque excede del objeto de estudio propio de este trabajo. Cfr. al respecto con posturas enfrentadas, COOKE Y MICHIE, 2010; COOKE Y MICHIE 2011; HART Y COOKE, 2013; HARRIS, LOWENKAMP Y HILTON, 2015, o MOSSMAN, 2015.

[11] En sentido parecido, HARRIS / LOWENKAMP / HILTON (2013:123), afirmando que la información sobre el riesgo relativo es especialmente relevante cuando se tata de gestionar los recursos disponibles en una determinada institución clínica o penitenciaria.

que los otros, y puede ser secundario el dato de si dicho riesgo más alto es del 40, del 60 o del 80%.

Pero cuando se trata de decisiones judiciales, en las que no se gestionan grupos de personas sino que se imponen consecuencias jurídicas a individuos en atención a sus características particulares, estimo que la información pertinente no es si el sujeto A tiene un riesgo 5 o 10 veces mayor o menor que el sujeto B, sino cómo de alto es el riesgo del sujeto A, para después sobre esa base decidir si dicho riesgo justifica o no la adopción de la decisión de que se trate en el caso concreto. Se necesita, por tanto, conocer el riesgo absoluto de reincidencia asociado a dicho individuo.

Hay que analizar entonces cuál o cuáles de los diferentes indicadores que existen para medir la validez predictiva de los instrumentos estructurados de valoración del riesgo ofrecen información sobre este tipo de riesgo, y cuáles resultan más relevantes. Pero en este punto nos encontramos con el problema de que en las investigaciones sobre valoración del riesgo de reincidencia se han estudiado y desarrollado con mucha más profundidad los indicadores de riesgo relativo que los de riesgo absoluto, o – en la denominación que suele darse a estas dos dimensiones diferentes de capacidad predictiva en la literatura criminológica especializada – los indicadores de discriminación (o riesgo relativo) que los de calibración (o riesgo absoluto) [12]. Así lo destacan diversos autores (SINGH, 2013; HELMUS et al, 2012, HANSON, 2017), y sólo recientemente algunos han comenzado a tomar cierta conciencia de este extremo, y a advertir sobre la necesidad de desarrollar y perfeccionar también indicadores que sirvan para medir cómo de fiables son las estimaciones de riesgo absoluto de reincidencia (cfr. por ej. HANSON, 2017).

De momento, y a salvo de la evolución que en este punto pueda tener lugar en un futuro próximo, seguramente el indicador de riesgo absoluto más conocido y tradicional (y más fácil de entender para el profano) sea el valor predictivo. Por ello, y aun siendo conscientes de que en lo que sigue efectuaremos una explicación excesivamente simplificada de un problema que es muy complejo[13], en aras de la claridad expositiva nos parece oportuno seleccionar el valor predictivo como indicador básico, y compararlo con la sensibilidad y la especificidad.

3.2. Sensibilidad, especificidad y valores predictivos. La medicina como ejemplo

La sensibilidad y la especificidad son dos de los indicadores de capacidad predictiva más conocidos y utilizados. Se calculan sobre la base de una tabla de contingencia de 2x2 en la que se organiza la información sobre la valoración del riesgo y la reincidencia realmente observada en 4 categorías: verdaderos positivos, falsos positivos, verdaderos negativos, y falsos negativos.

La Tabla 2 ofrece los resultados de un estudio que se hizo hace algunos años sobre una muestra de 163 delincuentes sexuales excarcelados en Cataluña, a los que se administró una herramienta de valoración del riesgo (el SVR-20), y tras una media de tiempo de 4 años se comparó la

[12] "Calibration refers to how well a risk assessment tool's predictions of risk agree with actual observed risk, whereas discrimination refers to how well an instrument is able to separate those who went on to be violent from those who did not" (SINGH, 2013:8). Cfr. Asimismo MUÑOZ LORENTE / LÓPEZ-OSSORIO 2016:136.

[13] Aunque sea relativamente sencillo entender qué es el valor predictivo y qué tipo de información proporciona, a la hora de utilizarlo en investigación y comparar por ejemplo el porcentaje de predicciones correctas obtenidas en unos estudios y otros puede ser necesario complementar el análisis con otros parámetros estadísticos, o calcular márgenes de error; cfr. un ejemplo en ROSSEGGER et al, 2014:3 y ss.

estimación de riesgo con la reincidencia real (aquí por reincidencia se entiende reincidencia sexual, y además penitenciaria, es decir, comisión de un nuevo delito y encarcelamiento).

Tabla 2. Riesgo estimado y reincidencia observada en un grupo de agresores sexuales en Cataluña

		Reincidencia sexual **observada**		Total	Porcentaje de predicciones correctas (valor predictivo)
		SÍ	NO		
Reincidencia sexual **pronosticada**	SÍ	17 Verdaderos positivos	28 Falsos positivos	45	37,7% (valor predictivo positivo)
	NO	7 Falsos negativos	111 Verdaderos negativos	118	94,07% (valor predictivo negativo)
Total		24	139	163	
Porcentaje de detecciones correctas		70,8% sensibilidad	79,9% especificidad		

Fuente: reproducción de la tabla contendida en PÉREZ RAMÍREZ y otros (2008:209), a la que he añadido la última columna, pues el cálculo del valor predictivo no estaba incluido en la tabla original.

Los autores de este estudio utilizaron la sensibilidad y la especificidad para dar cuenta del rendimiento del SVR-20, en los siguientes términos: "el SVR-20 predice correctamente el 79,9% de los no reincidentes (especificidad o verdaderos negativos) y el 70,8% de los reincidentes (sensibilidad o verdaderos positivos), con un total de clasificaciones correctas del 78,5%." (PÉREZ RAMÍREZ y otros, 2008:209). Y en efecto, así es. La sensibilidad es la capacidad de la herramienta para detectar correctamente los casos que sí cumplen el criterio – en nuestro caso, haber reincidido – y se calcula dividiendo el total de verdaderos positivos (en el ejemplo, 17) entre el total de sujetos que efectivamente reincidieron (en el ejemplo, 24). La sensibilidad también recibe el nombre de fracción de verdaderos positivos (LOINAZ, 2017:89). La especificidad por el contrario es la capacidad de la herramienta para detectar aquellos casos que no cumplen el criterio – en nuestro caso, los no reincidentes – y se calcula dividiendo el total de verdaderos negativos (111) entre el total de sujetos que no reincidieron (139). La especificidad también se denomina fracción de verdaderos negativos (LOINAZ, 2017:89). Conociendo ambos valores puede calcularse la eficacia diagnóstica de la herramienta, que consiste en dividir el total de clasificaciones correctas (verdaderos positivos más verdaderos negativos) entre el total de la muestra, y que en este caso da ciertamente como resultado 78,5%.

Adviértase que la sensibilidad y la especificidad miden cuántos sujetos de los que reinciden o no

lo hacen ha sido capaz de identificar correctamente el instrumento. La sensibilidad dice: sabiendo cuántos han delinquido, vamos a ver cuántos de esos habíamos sido capaces de identificar con nuestro test. Y la especificidad dice: sabiendo cuántos no delinquieron, veamos cuántos de esos detectó correctamente la herramienta. Sin embargo, es dudoso que ésta sea la información más relevante en el contexto forense, cuando un juez tiene que tomar una decisión sobre qué consecuencias jurídicas aplicar, o cómo ejecutarlas, sobre un individuo concreto. Para entender por qué, puede resultar ilustrativa una comparación con la medicina[14]: el texto que se reproduce a continuación valora la utilidad de la sensibilidad y de la especificidad cuando se realiza una prueba diagnóstica para detectar en un paciente la presencia de una enfermedad:

> "Sensibilidad. Es la probabilidad de clasificar correctamente a un individuo enfermo, es decir, la probabilidad de que para un sujeto enfermo se obtenga en la prueba un resultado positivo. La sensibilidad es, por lo tanto, la capacidad del test para detectar la enfermedad. […]
>
> Los conceptos de sensibilidad y especificidad permiten, por lo tanto, valorar la validez de una prueba diagnóstica. Sin embargo, *carecen de utilidad en la práctica clínica*. Tanto la sensibilidad como la especificidad proporcionan información acerca de la probabilidad de obtener un resultado concreto (positivo o negativo) en función de la verdadera condición del enfermo con respecto a la enfermedad. Sin embargo, *cuando a un paciente se le realiza alguna prueba, el médico carece de información a priori acerca de su verdadero diagnóstico*, y más bien *la pregunta se plantea en sentido contrario*: ante un resultado positivo (negativo) en la prueba, ¿cuál es la probabilidad de que el paciente esté realmente enfermo (sano)?. Así pues, resulta obvio que hasta el momento sólo hemos abordado el problema en una dirección. Por medio de los *valores predictivos* completaremos esta información" (PITA FERNÁNDEZ / PÉRTEGAS DÍAZ, 2003, cursivas añadidas)

Si en este texto sustituimos la terminología médica por la jurídica, obtendríamos algo así:

> Sensibilidad. Es la probabilidad de clasificar correctamente a un individuo reincidente, es decir, la probabilidad de que para un sujeto reincidente se obtenga en la herramienta de valoración del riesgo un resultado de riesgo alto. La sensibilidad es, por lo tanto, la capacidad del test para detectar la reincidencia. […]
>
> Los conceptos de sensibilidad y especificidad permiten, por lo tanto, valorar la validez de una herramienta de valoración del riesgo. Sin embargo, *carecen de utilidad en la práctica judicial*. Tanto la sensibilidad como la especificidad proporcionan información acerca de la probabilidad de obtener un resultado concreto (positivo o negativo) en función de la verdadera condición del sujeto con respecto a la reincidencia. Sin embargo, *cuando a un sujeto se le realiza alguna valoración de riesgo, el juez carece de información a priori acerca de su verdadera condición de reincidente*, y más bien *la pregunta se plantea en sentido contrario*: ante un resultado de riesgo alto (o bajo) en la prueba, ¿cuál es la probabilidad de que el paciente realmente reincida (o no)?. Así pues, resulta obvio que hasta el momento sólo hemos abordado el problema en una dirección. Por medio de los valores predictivos completaremos esta información

Parece, por tanto, que si las decisiones judiciales son en algún sentido asimilables a las decisiones clínicas, los valores predictivos pueden ser más relevantes que la sensibilidad y la especificidad en el contexto forense.

El valor predictivo positivo es la proporción de individuos que el instrumento clasificó como de alto riesgo y que efectivamente reincidieron después. Se obtiene dividiendo el total de

[14] Tanto la sensibilidad como la especificidad – y lo mismo puede decirse del resto de indicadores de capacidad predictiva – no se usan sólo en el contexto de las valoraciones de riesgo de reincidencia o de violencia, sino que son aplicables a muchas otras ramas del conocimiento donde se utilicen pruebas o tests para estimar la probabilidad de ocurrencia de determinados fenómenos, como puede ser el caso de una prueba médica, la precisión de detección de un radar, etc.

verdaderos positivos (en el ejemplo de la Tabla 2, 17) entre el total de sujetos a los que el instrumento clasificó como de riesgo alto (en el mismo ejemplo, 45). Y el valor predictivo negativo es la proporción de individuos que la herramienta clasificó como de bajo riesgo y que después efectivamente no reincidieron. Se obtiene dividiendo el total de verdaderos negativos (en el ejemplo de la Tabla 2: 111) entre el total de sujetos a los que el instrumento clasificó como de riesgo bajo (en el mismo ejemplo: 118). Los valores respectivos para el ejemplo que hemos utilizado en la Tabla 2 están indicados en la última columna de la tabla: el valor predictivo positivo en aquel estudio fue de del 37,7%, y el valor predictivo negativo del 94%.

Tanto la sensibilidad y la especificidad como los valores predictivos se obtienen a partir de la misma información que ofrece una tabla de contingencia de 2x2: calculando los porcentajes de falsos y de verdaderos positivos. Pero la diferencia más relevante – a los efectos que aquí interesan – entre ambos pares de indicadores es que tanto la sensibilidad como la especificidad tienen una orientación retrospectiva, mientras que los valores predictivos tienen una orientación prospectiva: las primeras miran hacia el pasado, mientras que los segundos miran hacia el futuro (SINGH, 2013:11, 12). Antes hemos afirmado que lo que dice la sensibilidad es: sabiendo cuántos han delinquido, vamos a ver cuántos de esos habíamos sido capaces de identificar con nuestro test. Pues bien, el valor predictivo, por el contrario, dice: sabiendo cuántos hemos identificado como peligrosos, vamos a ver cuántos de ellos realmente han delinquido después. Es el porcentaje de casos en los cuales la realidad ha confirmado la predicción.

Como ponía de manifiesto el fragmento del artículo sobre pruebas diagnósticas en medicina que hemos reproducido líneas arriba, la situación en la que normalmente se encuentra el juez que tiene que adoptar una decisión teniendo en cuenta – entre otros criterios – una valoración de riesgo de reincidencia se asimila a la situación en la que se encuentra el médico que debe decidir si administra o no un tratamiento, o si realiza una prueba de confirmación más invasiva, a la vista del resultado de un test diagnóstico. La pregunta relevante es: a la vista de los resultados de esta herramienta de valoración del riesgo (o prueba diagnóstica), que me indica que el resultado es de riesgo alto (o positivo), ¿cuál es la probabilidad de que el sujeto realmente reincida después (o realmente tenga la enfermedad)? Y la respuesta a esa pregunta la da el valor predictivo positivo, no la sensibilidad.

> En relación con un sujeto al que el SVR-20 ha dado una puntuación correspondiente a riesgo alto (en el ejemplo de la Tabla 2), una sensibilidad del 70,8% (como la que se obtuvo en aquel estudio) indica que, si este individuo es realmente reincidente, la probabilidad de que la herramienta lo haya clasificado como de riesgo alto es del 70,8%. Un valor predictivo positivo del 37,7% (como el que se obtuvo en aquel estudio) indica que, habiendo recibido una clasificación en la categoría de riesgo alto, la probabilidad de que ese individuo realmente reincida es del 37,7%.

4. Indicadores de capacidad predictiva más comúnmente utilizados en Criminología: el área bajo la curva ROC (AUC)

Hemos visto que los valores predictivos son un indicador relevante para valorar la información que nos proporciona una herramienta estructurada de valoración del riesgo sobre la probabilidad de reincidencia de un sujeto cuando sobre la base de dicha información tienen que tomarse decisiones en el ámbito forense. Sin embargo, los valores predictivos no son los estadísticos

usados con más frecuencia en la investigación criminológica sobre estas herramientas. Por el contrario, es mucho más habitual que en esta clase de estudios se proporcione información sobre la sensibilidad y la especificidad, y también sobre otro indicador del que no hemos hablado hasta ahora: el área bajo la curva ROC (o AUC, por sus siglas en inglés: *area under the curve*), que se ha convertido *de facto* en la medida estándar por excelencia para informar sobre la capacidad predictiva de las herramientas estructuradas de valoración del riesgo: como afirma SINGH, (2013:16), "it has become the *de facto* standard in the field"[15].

> Y un repaso a los estudios publicados en España en los últimos 10 o 15 años sobre las herramientas de valoración del riesgo corrobora que éste es el caso también en nuestro país. El estudio de PÉREZ RAMÍREZ et al (2008) que hemos utilizado como ejemplo en el epígrafe anterior proporcionaba como medidas de la validez predictiva del SVR-20 la sensibilidad, la especificidad y el área bajo la curva ROC. No incluía referencia alguna a los valores predictivos. El estudio de CAPDEVILA CAPDEVILA et al (2015) sobre la tasa de reincidencia penitenciaria en Cataluña en el año 2014, que incluye información sobre una herramienta de valoración del riesgo utilizada en el sistema de gestión penitenciaria en Cataluña – el RisCanvi – proporciona como indicadores de validez predictiva la sensibilidad, la especificidad y la *odds ratio* (CAPDEVILA CAPDEVILA et al, 2015:151 y 152). Tampoco menciona los valores predictivos. Otro estudio más reciente sobre el RisCanvi ofrece datos de área bajo la curva y análisis de supervivencia, pero tampoco valores predictivos, aunque en este caso se indica la proporción de sujetos clasificados como de riesgo alto, medio o bajo que reincidieron con nuevos delitos violentos en el tiempo de seguimiento (respectivamente, 32, 16 y 9%; cfr. ANDRÉS-PUEYO / ARBACH-LUCIONI / REDONDO, 2018:262 y s.). En el año 2006 se publicó un estudio sobre la aplicabilidad de dos herramientas actuariales (VRAG y SAQ) a la población penitenciaria española. La capacidad predictiva se midió a través del área bajo la curva ROC, sin hacer referencia a los valores predictivos (BALLESTEROS REYES / GRAÑA GÓMEZ / ANDREU RODRÍGUEZ, 2006). Y algo similar puede decirse de otros estudios como los de NGUYEN / ARBACH LUCIONI / ANDRÉS PUEYO, 2011:286-288 (sensibilidad, especificidad, área bajo la curva ROC y otros indicadores, pero sin referencia a valores predictivos), o NGUYEN VO / ANDRÉS PUEYO, 2016 (modelos de regresión logística y área bajo la curva). El trabajo de MUÑOZ LORENTE y LÓPEZ-OSSORIO sobre utilización en el contexto forense de las valoraciones psicológicas del riesgo de violencia ofrece una panorámica de los instrumentos más utilizados señalando como único parámetro de rendimiento el área bajo la curva (MUÑOZ LORENTE / LÓPEZ-OSSORIO, 2016: Tabla2).

Como ya hemos advertido antes, las finalidades de una investigación criminológica no coinciden con los objetivos que persigue un juez cuando impone o deja de imponer determinada consecuencia jurídica a un delincuente concreto, por lo que el hecho de que los valores predictivos, relevantes en este último ámbito, no suelan aparecer mencionados en los estudios criminológicos sobre valoración del riesgo no necesariamente tendría que ser una omisión reprochable o un defecto metodológico[16]. Sin embargo, las herramientas de valoración del riesgo no son un mero objeto de investigación teórica o fundamental desligado de cualquier aplicación práctica, sino que, muy al contrario, se diseñan precisamente con el objetivo de que sean

[15] Cfr. asimismo SINGH y PETRILA, 2013:3 ("ROC curve analysis continues to be the dominant statistical technique used to test instruments") o MOSSMAN, 2013:28 ("By the middle of the 21st century's first decade, ROC indices had become investigators' standard tools for describing instruments that assess the risk of future violence and the recidivism potential of sex offenders").

[16] De hecho, los valores predictivos como indicadores de validez predictiva presentan un problema que a veces los hace poco idóneos para ser utilizados como criterio de comparación entre estudios con muestras o instrumentos diferentes, y es que son muy sensibles a las tasas de prevalencia del fenómeno. Esta fue de hecho una de las razones por las cuales a partir de los años 90 se fue imponiendo el uso de otros indicadores, como el área bajo la curva ROC y otros indicadores de riesgo relativo (cfr. sobre todo ello SINGH, 2013; MOSSMAN, 1994; MOSSMAN, 2013).

utilizadas en el ámbito penitenciario y judicial (además de en otros, como la hospitalización psiquiátrica), como ayuda para la toma de decisiones. Siendo esto así, sí sería deseable que incluyeran entonces información sobre los valores predictivos, no necesariamente en vez de, pero sí al menos junto con otros indicadores que también puedan ser relevantes desde otros puntos de vista. Porque no hacerlo puede generar malos entendidos sobre la calidad de las herramientas de valoración del riesgo y sobre el papel que pueden y/o deben jugar en la toma de decisiones en el ámbito penal y penitenciario. Para comprender por qué vamos a ver brevemente qué significa y qué información aporta el área bajo la curva ROC, el estadístico más utilizado tanto a nivel nacional como internacional para evaluar la capacidad predictiva de las herramientas estructuradas de valoración del riesgo.

4.1. Qué es la curva ROC: un ejemplo imaginario

Las herramientas actuariales de valoración del riesgo no proporcionan clasificaciones dicotómicas de los sujetos (en peligrosos – no peligrosos), sino que generalmente contienen múltiples niveles de riesgo posibles: muy bajo, bajo, moderado, alto, extremo, etc., en los que se ubica al sujeto en función de la puntuación que haya obtenido tras la valoración de los distintos factores de riesgo. En consecuencia, quien utiliza la herramienta como ayuda para la toma de decisiones puede optar por situar la frontera entre riesgo aún asumible y riesgo ya intolerable en diferentes puntos de la escala, y en función de dónde le parezca oportuno situar el punto de corte obtendrá diferentes predicciones, cada una de las cuales producirá un número distinto de aciertos y errores. El área bajo la curva ROC es un indicador de capacidad predictiva que se adapta bien a esta característica de la moderna estimación del riesgo. Puede describirse diciendo que se trata de "un índice global de discriminación, igual a la probabilidad de que un individuo violento seleccionado al azar reciba una clasificación de riesgo superior (mayor puntuación total, categoría actuarial de riesgo, o evaluación de riesgo profesional estructurada) que un individuo no violento seleccionado al azar" (Singh 2013, p. 15). Como es un indicador bastante complejo, la mejor manera para comprender cómo funciona es servirse de un ejemplo. Para ello utilizaremos a continuación el estudio imaginario con el que MOSSMAN introdujo en los años 90 este estadístico en el ámbito de la valoración de riesgo (MOSSMAN, 1994)[17].

> Los párrafos que siguen contienen una descripción bastante técnica y matemática del funcionamiento de la curva ROC; el lector que no quiera perderse en detalles y cálculos numéricos puede pasar directamente al epígrafe 4.2., en el que se valora el tipo de información que aporta la AUC; no obstante, consideramos conveniente explicar su funcionamiento con cierto detalle, especialmente para que se comprenda mejor lo que después se expondrá en las conclusiones, sobre la complejidad de las valoraciones de riesgo, y sobre los peligros de fundamentar decisiones judiciales o instituciones jurídico-penales en el conocimiento que proporciona la metodología estadística.

Imaginemos que tenemos una herramienta estructurada que clasifica a los sujetos en 5 niveles de riesgo. Para comprobar cómo de bien funciona, la aplicamos a una muestra de 1000 individuos, y al cabo de un tiempo estipulado observamos cuál ha sido la reincidencia real, y la comparamos con el nivel de riesgo que previamente nuestra herramienta había asignado a cada individuo. Supongamos que de los 1000 sujetos, 100 reinciden (un 10%), y que los datos son los que siguen:

[17] Puede verse un ejemplo parecido en MOSSMAN, 2013.

Tabla 3: clasificación de 1000 sujetos en cinco niveles de riesgo y comparación con la delincuencia evidenciada

	Nivel de riesgo asignado por el instrumento (1= muy bajo; 5= muy alto)					Total
	1	**2**	**3**	**4**	**5**	
Reinciden	7	9	15	19	50	100
No reinciden	450	172	135	83	60	900
Total	457	181	150	102	110	1000
Ratio de VP o sensibilidad		0.93	0.84	0.69	0.50		
Ratio de FP o 1-especificidad		0.50	0.31	0.16	0.07		

Fuente: adaptación de la Tabla 2 incluida en MOSSMAN, 1994: 785.

Como puede verse, 457 sujetos habían sido clasificados en el nivel de riesgo más bajo (nivel 1). Si los comparamos con su comportamiento real observamos que la gran mayoría (450) no reincidió; sin embargo, 7 sí lo hicieron, a pesar de que habían sido considerados como de riesgo muy bajo. Por otro lado, de las 181 personas que fueron clasificadas en el segundo nivel de riesgo (nivel 2) 172 no reincidieron, pero 9 sí lo hicieron. Y así sucesivamente podemos observar en la Tabla 3 el número total de personas que fueron clasificadas en los niveles 3, 4 y 5 de riesgo, y dentro de cada nivel cuántos de ellos reincidieron y cuántos no lo hicieron.

Obsérvese que hay errores en todos los niveles de riesgo: algunas de las personas consideradas inicialmente como de riesgo bajo delinquen a pesar de ello (son los falsos negativos), y también algunas de las personas inicialmente valoradas como de riesgo alto sin embargo no lo hacen (son los falsos positivos). Ninguna herramienta de valoración del riesgo es perfecta, siempre habrá errores. La cuestión es qué tipo de errores prefiere cometer o evitar la persona que toma decisiones sobre la base de la información que la herramienta le aporta.

Ahora supongamos que, disponiendo de esta información sobre el funcionamiento de la herramienta, vamos a utilizarla como base para adoptar la decisión sobre conceder o denegar la libertad condicional a un nuevo grupo de internos que hemos evaluado con ella. Supongamos también que el criterio decisivo para adoptar tal decisión sea el riesgo de comisión de nuevos delitos, y que la autoridad a la que corresponda adoptar la decisión (digamos, para abreviar, que sea un juez) tiene libertad para decidir cuál es nivel a partir del cual considera que el riesgo es inasumible. E imaginemos, para terminar, que tenemos dos jueces diferentes, que deben tomar la decisión sobre la base de la misma información.

El primer juez (Juez 1) es una persona muy preocupada por la inseguridad ciudadana, y es de la opinión de que facilitar a los penados un regreso escalonado a la sociedad a través de un régimen de libertad controlada no reduce sino que incrementa las tasas de criminalidad. En consecuencia, decide conceder la libertad condicional sólo a los internos que han sido clasificados en el nivel más bajo de riesgo (nivel 1), pues sólo éste le parece un riesgo asumible, y denegársela a todos los demás (los clasificados en los niveles 2, 3, 4 y 5). ¿Cómo de buena sería esta decisión? Para valorarlo, podemos fijarnos en los números de aciertos y errores que produciría:

Tabla 4: Sensibilidad, especificidad, valores predictivos y errores en la decisión del Juez 1

	SÍ reinciden	**NO reinciden**	**Total**	
Riesgo alto: niveles 2-5 (se deniega la libertad condicional)	93	450	543	17,1% valor predictivo positivo
Riesgo bajo: nivel 1 (se concede la libertad condicional)	7	450	457	98,5% valor predictivo negativo
Total	100	900	1000	
	93% sensibilidad	50% especificidad		

El Juez 1 sólo ha concedido la libertad condicional a los 457 sujetos clasificados en el nivel 1 de riesgo (el más bajo), y se la ha denegado a todos los demás, es decir, a un total de 543 (la suma de los 181, 150, 102 y 110 clasificados en los niveles 2, 3, 4 and 5 respectivamente). Si el instrumento de valoración del riesgo funciona en este caso tal y como funcionó en la primera ocasión en que lo aplicamos y las muestras son similares, podemos asumir que la reincidencia real será aproximadamente la misma que entonces se evidenció (un 10%), y su distribución por niveles de riesgo también. En consecuencia, podemos asumir que de los 457 sujetos a los que el Juez 1 ha concedido la libertad condicional 7 van a reincidir. Y que de los 543 a los que el juez ha denegado la libertad condicional, 93 habrían reincidido (la suma de los 9, 15, 19 y 50 reincidentes en los niveles 2, 3, 4 y 5 de la Tabla 3), y 450 no (la suma de los 172, 135, 83 y 60 no reincidentes en los niveles 2, 3, 4 y 5 de la Tabla 3).

Usando el instrumento de la manera en que lo hecho el Juez 1 (es decir, situando el umbral de discriminación entre los niveles de riesgo 1 y 2) ha sido capaz de detectar a 93 sujetos como de alto riesgo de entre los 100 que han reincidido, esto es, ha tenido una sensibilidad del 93%. Pero no ha funcionado tan bien al clasificar como de bajo riesgo a los sujetos que finalmente no reincidieron: de los 900 que no lo hicieron, la herramienta consideró equivocadamente como de alto riesgo a la mitad (450), lo que significa que la especificidad ha sido sólo del 50%.

Por otro lado, al aplicar el instrumento de esta forma 543 personas fueron clasificadas como de

alto riesgo, de las que sin embargo sólo hubieran reincidido 93. El valor predictivo positivo es sólo del 17,1% (pues 93 es el 17,1% de 543). Por el contrario, el valor predictivo negativo es mucho más alto, del 98,5%, porque de las 457 personas que quedaron clasificadas como de riesgo bajo, el 98,5% (esto es, 450) efectivamente no reincidieron.

Vemos por tanto que la decisión del Juez 1 evidencia una muy buena capacidad predictiva por lo que hace a la predicción de los sujetos con riesgo bajo: el 98,5% de los que han sido puestos en libertad no van a delinquir de nuevo, habrá sólo 7 falsos negativos. Sin embargo, no puede decirse lo mismo por lo que respecta al grupo de sujetos que se ha considerado de riesgo demasiado alto como para dejarlos en libertad: más del 80% de las personas a las que el Juez 1 no está concediendo la libertad condicional por miedo a que reincidan no lo habrían hecho de haber tenido la oportunidad de demostrarlo. Su decisión ha generado un número muy elevado de falsos positivos (450), que sin embargo resultan "invisibles", puesto que la decisión de mantenerlos en prisión impide comprobar la existencia de estos errores. Por último, la Tabla 4 evidencia que el número total de errores que ha producido la decisión del Juez 1 es de 457: 7 falsos negativos y 450 falsos positivos.

Pero también es posible tomar decisiones diferentes utilizando la misma información proporcionada por esta herramienta imaginaria de estimación de riesgo. Supongamos que tenemos un segundo juez (Juez 2), de mentalidad más progresista, al que preocupan sobre todo los efectos negativos que la estancia en prisión produce en los condenados, y que está convencido de que una utilización generosa del tercer grado y de la libertad condicional favorece en gran medida la reinserción (y con ello, la evitación de futuros delitos). Por ello, decide denegar la libertad condicional sólo al grupo de sujetos clasificados en el nivel más elevado de riesgo (nivel 5) porque estima que sólo estos representan un peligro inasumible, y concedérsela a todos los demás.

Si construimos una tabla de contingencia de 2x2 para esta decisión, igual que lo hicimos en el caso del Juez 1, veremos que los números son muy diferentes ahora:

Tabla 5: Sensibilidad, especificidad, valores predictivos y errores en la decisión del Juez 2

	SÍ reinciden	**NO reinciden**	**Total**	
Riesgo alto: nivel 5 (se deniega la libertad condicional)	50	60	110	45,5% valor predictivo positivo
Riesgo bajo: niveles 1-4 (se concede la libertad condicional)	50	840	890	94,4% valor predictivo negativo
Total	100	900	1000	
	50% sensibilidad	93,3% especificidad		

Como consecuencia de la decisión del Juez 2, 890 personas van a salir en libertad condicional (la suma de los 457, 181, 150 y 102 sujetos clasificados en los niveles de riesgo 1, 2, 3 y 4), de las cuales reincidirán 50, y 840 no lo harán. De los 110 que permanecerán en prisión, 50 habrían reincidido en caso de haber sido liberados, y 60 no. La sensibilidad es mucho más baja ahora, sólo del 50%, pero a cambio la especificidad ha subido al 93,3%. También se observan diferencias por lo que hace a los valores predictivos, especialmente en el valor predictivo positivo: ha subido al 45,5%, mientras que el valor predictivo negativo ha bajado un poco, pero se mantiene por encima del 90% (94,4%). Si atendemos a los errores que derivan de esta decisión, veremos que también son muy distintos de los que producía la decisión del Juez 1: el número total de errores ha disminuido drásticamente (de 457 ha bajado a 110), y también su distribución ha cambiado, pues está mucho más equilibrada entre los falsos negativos (50) y los falsos positivos (60).

Ahora cabría preguntarse, ¿cuál de las dos decisiones es mejor? A primera vista podría parecer que la segunda, ya que produce un número total de errores mucho menor. Pero alguien a quien preocupe mucho la seguridad pública podría objetar que aunque el número de errores sea menor en la decisión del Juez 2, el número de falsos negativos ha subido desde 7 hasta 50, y si estuviéramos hablando de delincuentes violentos que tras salir de la prisión vuelven a cometer delitos graves como violaciones, homicidios o lesiones graves, quizá sería preferible asumir – como en la decisión del Juez 1 – un número mayor de falsos positivos, a cambio de reducir los falsos negativos lo más posible y evitar con ello la comisión de algunos nuevos delitos muy graves. Por el otro lado, sin embargo, podría argüirse en sentido contrario que la libertad es un derecho fundamental del que nadie puede ser privado arbitrariamente, y que mantener en prisión a 543 personas sólo para evitar la comisión de 93 nuevos delitos – esto es, mantener en prisión a una proporción de 8 sujetos inofensivos por cada dos reincidentes – es una medida desproporcionada. Es decir, resulta evidente que la pregunta no puede ser respondida sólo exponiendo la información estadística que aporta la herramienta de valoración del riesgo. Estas herramientas no pueden decirnos cuál es la mejor decisión, sólo pueden informarnos del número y de la clase de errores en que probablemente incurriremos en función de cómo decidamos utilizarlas. Y siempre habrá errores, porque ningún instrumento de valoración del riesgo es perfecto. La cuestión decisiva es qué número y qué clase de errores nos parecen asumibles o insoportables y por qué razones, desde la perspectiva no sólo de combatir la delincuencia de manera eficaz, sino también de hacerlo de manera justa.

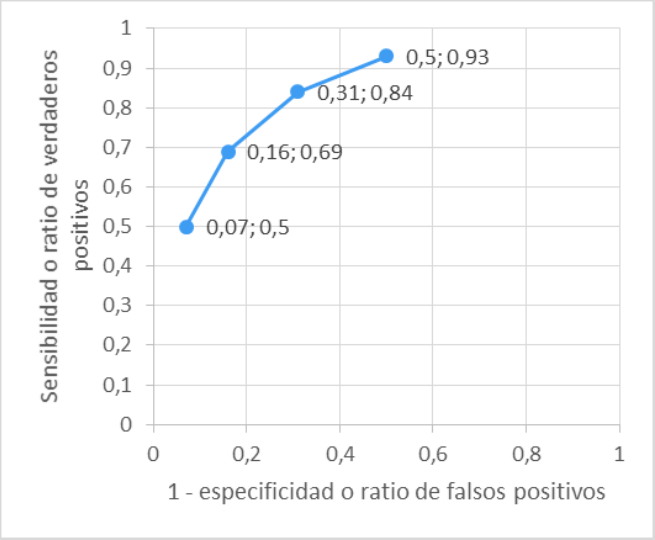
Al final de este trabajo volveremos brevemente sobre este problema (que es el crucial en toda esta materia), pero antes retomemos la cuestión que pretendíamos resolver en este epígrafe: la de qué es la curva ROC y el área que queda bajo esa curva, y qué información aporta.

Si observamos de nuevo las Tablas 4 y 5 (con los resultados de las decisiones del Juez 1 y del Juez 2) veremos que cada una de ellas produce una combinación de valores de sensibilidad y de especificidad diferentes: 0.93 / 0.50 en el primer caso, y 0.50 / 0.93 en el segundo. Pues bien, igual que hemos construido estas tablas para dos jueces imaginarios que quisieran situar la frontera entre el riesgo asumible y el no asumible en los niveles 1-2, y 4-5, respectivamente, podríamos construir las tablas que resultarían si otros imaginarios jueces situaran dicho límite en los niveles 2-3 y 3-4, y calcular los valores de sensibilidad y de especificidad, así como los números y la clase

de errores, que resultarían en cada caso. En las dos últimas filas de la Tabla 3 se ofrecen los valores de sensibilidad y de 1-especificidad para estos nuevos posibles puntos de corte: son, respectivamente, 0.84 y 0.31, y 0.69 y 0.16[18].

Conociendo estos pares de cifras, que corresponden a cada punto de corte que decidamos establecer, podríamos representarlos gráficamente en un sistema de coordenadas, donde la sensibilidad fuera el eje de ordenadas y 1-especificidad el eje de abscisas. Y obtendríamos algo como lo siguiente:

Figura 1: Aproximación a la curva ROC para los datos de la Tabla 3



Como puede observarse, los cuatro pares de valores de sensibilidad y 1-especificidad para cada punto en el que decidamos situar el corte entre riesgo alto y bajo en nuestra herramienta de valoración del riesgo pueden ser unidos con una curva. Y esa curva es la curva ROC. ROC es el acrónimo de *Receiver Operating Characteristics*, o característica operativa del receptor. La curva ROC representa todas las diferentes combinaciones posibles entre valores de sensibilidad y de 1-especificidad para cada posible punto de corte que podamos establecer en una herramienta de valoración del riesgo. Puede describirse también diciendo que es la representación de la razón de verdaderos positivos y la razón de falsos positivos para cada punto de corte en la escala (LOINAZ, 2017:90 y ss.; MOSSMAN, 1994; SINGH, 2013). En la Figura 1 sólo se ve el fragmento de la curva que corresponde a los 4 puntos de corte calculados según los distintos niveles de riesgo que contiene el instrumento imaginario de que nos estamos sirviendo como ejemplo, pero para cada herramienta de valoración del riesgo puede calcularse la curva completa, desde la esquina inferior izquierda hasta la superior derecha del sistema de coordenadas (cfr. ulteriores referencias al respecto en SINGH, 2013:16)

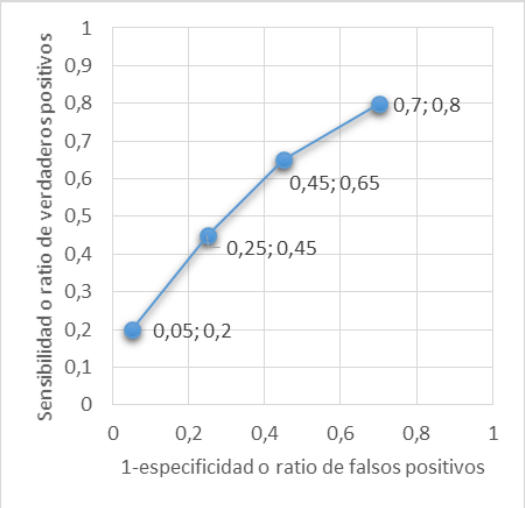
La curva ROC que corresponde a cada herramienta de valoración del riesgo tiene una pendiente

[18] La sensibilidad es de 0.69 si se sitúa el umbral de discriminación entre los niveles de riesgo 2 y 3, y de 0.31 si se coloca entre los niveles 3 y 4. En cuanto a la especificidad, las cifras que aparecen en la Tabla 3 no son directamente la especificidad sino su complementario, 1-especificidad. Porque como enseguida se explicará en el texto, es con este último valor con el que se construye la curva ROC.

diferente: puede haber curvas con pendientes muy pronunciadas que ascienden rápidamente y tienden a acercarse a la esquina superior izquierda del gráfico, como la de la Figura 2, y otras curvas con aspecto más plano que se acercan a la diagonal entre las esquinas inferior izquierda y superior derecha del gráfico, como la de la Figura 3:

Figura 2:

Figura 3:



El área completa (la superficie) de estos sistemas de coordenadas es siempre 1, puesto que los valores máximos de la sensibilidad y de 1-especificidad son, en ambos ejes, 1. Y cada curva ROC que dibujemos deja por debajo de sí un área, que recibe el nombre de área bajo la curva. Ese área que queda por debajo de la curva tiende al valor 1 en las curvas pronunciadas como la de la Figura 2 (se acercará a 1 tanto más cuanto más pegada esté a la esquina superior izquierda), mientras el área bajo las curvas planas como la de la Figura 3 tiende a 0.5 (puesto que la diagonal o bisectriz del primer cuadrante divide la superficie total del sistema de coordenadas, que es 1, por la mitad). En el ámbito de la valoración del riesgo se considera que un área bajo la curva de 1 sería una clasificación perfecta, mientras que una curva de 0.5 no sería mejor que el azar, porque no sería mejor que diferenciar a los sujetos reincidentes de los que no lo son lanzando una moneda al aire (MOSSMAN, 2013:31)

4.2. Información que aporta la curva ROC y limitaciones

El valor del área bajo la curva que corresponde al ejemplo imaginario de MOSSMAN del cual nos hemos estado sirviendo en este epígrafe (la curva ROC dibujada en la Figura 1) es de 0.856 (MOSSMAN, 1994:785). Si el valor máximo que puede alcanzar el área bajo la curva es de 1, parece que 0.86 es un valor elevado, y que por tanto nos encontramos ante un instrumento de valoración del riesgo con una buena capacidad predictiva.

> Con todo, no hay acuerdo sobre cuáles son los valores del área bajo la curva que deben considerarse satisfactorios[19]. SINGH / DESMARAIS / VAN DORN (2013) revisaron 47 estudios que utilizaban el área bajo la

[19] "More than any other performance indicator reported in the violence risk assessment literature, the AUC is interpreted according to benchmarks as to what constitutes a small, moderate, or large magnitude effect size

> curva como indicador de capacidad predictiva, y mientras algunos consideran moderada la capacidad predictiva a partir de un valor de AUC de 0.60, otros sólo lo hacen a partir de 0.65, y algunos incluso sólo a partir de 0.70; en cuanto a la capacidad predictiva alta, para algunos lo es a partir de un valor de AUC de 0.70, mientras que para otros lo sería a partir de 0.75 o 0.80. Otros autores proponen aplicar criterios aún más estrictos, y considerar que la precisión de las clasificaciones de riesgo es moderada con valores de AUC entre 0.80 y 0.90, y elevada sólo a partir de 0.90 (SJÖSTEDT Y GRANN, 2002). En otro estudio reciente, sin embargo, se considera 'pobre' la AUC si está por debajo de 0.55, 'regular' o 'aceptable' (*fair*) si está entre 0.55 y 0.63, 'buena' entre 0.64 y 0.71, y 'excelente' a partir de 0.71 (DESMARAIS / JOHNSON / SINGH, 2018:29). Como puede observarse, la variabilidad de criterios es bastante notable.

Y aquí es donde aparece la cuestión decisiva: ¿qué significa ese valor de 0.856? Significa que si escogemos al azar un sujeto que efectivamente ha reincidido hay un 86% de probabilidades de que ese sujeto haya obtenido en ese instrumento de estimación del riesgo una puntuación de mayor riesgo que un sujeto no reincidente también escogido al azar. Es decir, el valor del área bajo la curva ROC es una medida de riesgo relativo o un índice de discriminación: indica que un sujeto que haya recibido una puntuación más alta en el instrumento tiene mayor probabilidad de reincidir que otro sujeto que con ese mismo instrumento haya obtenido una puntuación menor (MOSSMAN, 2006). Lo que refleja es la capacidad del instrumento actuarial para discriminar cuáles son los sujetos dentro de la muestra que tienen más riesgo de reincidir que otros. Pero no dice nada sobre cuál es la probabilidad de riesgo (absoluto) asociada a los sujetos clasificados en cada nivel de riesgo. En particular, ese valor de 0.856 no significa ni que los sujetos clasificados como de riesgo alto tengan un 86% de probabilidades de reincidir, ni que el porcentaje total de detecciones ni de predicciones correctas hechas con ese instrumento sea del 86%. Recordemos los valores predictivos y el número de errores que resultaban de las decisiones tomadas por el Juez 1 y el Juez 2 reflejados en las Tablas 4 y 5: ese instrumento de predicción imaginario, que tiene un área bajo la curva de 0.86, proporciona estimaciones con números de errores muy dispares según cómo se use, es decir, según dónde se coloque el umbral de discriminación. Y ni siquiera en el mejor de los casos (el que producía un menor número global de errores, Juez 2) el valor predictivo positivo alcanzaba el 50%. Queremos insistir en este punto: puede haber instrumentos con valores altos de área bajo la curva ROC, superiores al 80%, pero cuyo valor predictivo no llegue ni siquiera en el mejor de los casos al 50%.

Por tanto, valores de área bajo la curva ROC de 1 no representan una predicción perfecta, sino una discriminación perfecta: lo que significan es que ese instrumento siempre daría a un sujeto violento elegido al azar una puntuación más alta que a un sujeto no violento elegido igualmente al azar (MOSSMAN, 2013:31; SINGH, 2013:17). Pero ni mucho menos significan que los individuos clasificados como de riesgo alto tengan una probabilidad de reincidir del 100%, ni que los clasificados como de riesgo bajo la tengan del 0%. Hay que tener en cuenta que la curva ROC se construye sobre los valores de sensibilidad y de especificidad, de modo que tiene – igual que ellos – una orientación fundamentalmente retrospectiva, y no prospectiva:

> "El análisis con curvas ROC se desarrolló como una metodología para el diagnóstico, en lugar de como una metodología para el pronóstico (Cook, 2008). Es decir, el análisis con curvas ROC y la AUC responden a la pregunta, '¿pudo haberse predicho un suceso negativo que ya ha tenido lugar?'. Esta no es la situación

(Singh et al., 2013). However, these benchmarks were never intended to be used to evaluate the performance of predictive models (Mossman, 2013), and there is considerable variation in rules-of-thumb, suggesting that caution is warranted when using them" (SINGH 2013:18). Cfr. asimismo MOSSMAN, 2013:35 y s.

> a la que se enfrentan día a día los profesionales que trabajan en instituciones psiquiátricas o correccionales, que más bien necesitan respuesta para esta otra pregunta: '¿se demostrará como correcta en el futuro la predicción que he hecho?'. Se trata de dos preguntas esencialmente diferentes." (SINGH, 2013:17)

El área bajo la curva tiene además otras limitaciones, de las que a los efectos de este trabajo interesa destacar especialmente una[20]: como probablemente el lector ha podido comprobar por sí mismo en las líneas precedentes, es un estadístico nada fácil de entender, sobre todo para el profano pero a veces ni siquiera para los propios especialistas en la materia. En un trabajo publicado hace algunos años se revisaron 47 estudios que evaluaban precisamente a través de la AUC el funcionamiento de diversas herramientas de valoración del riesgo de violencia; pues bien, sólo 16 de los 47 ofrecían una interpretación de lo que significa este parámetro, pero lo que es más llamativo, en 14 de esos 16 estudios la interpretación que se hacía era equivocada, pues o bien afirmaban que el área bajo la curva es el porcentaje de sujetos cuyo comportamiento fue correctamente predicho por el instrumento, o bien afirmaban que es la proporción de sujetos considerados de alto riesgo que efectivamente reincidió, es decir, la confundían con el valor predictivo positivo (SINGH / DESMARAIS / VAN DORN, 2013). Si esto ocurre entre académicos expertos en el ámbito de la valoración del riesgo, no cabe sino esperar que las confusiones sean aún mayores entre personas especializadas en otros ámbitos tan diferentes como el de la justicia penal.

4.3. Consecuencias para el Derecho penal.

Como hemos visto el área bajo la curva ROC (u otros índices de discriminación o medidas de riesgo relativo) es la forma más extendida en la literatura criminológica para valorar la capacidad predictiva de las herramientas estructuradas de valoración del riesgo. Además, este indicador suele alcanzar en los diferentes instrumentos de valoración del riesgo de violencia valores cercanos o superiores al 0.70, de los que se informa cumplidamente en las publicaciones científicas, y que en la bibliografía criminológica especializada suelen calificarse como aceptables, o satisfactorios[21]. Sin embargo, el dato de que el área bajo la curva ROC sea elevada puede ser una información completamente irrelevante para muchas de las decisiones que tienen que tomarse en el ámbito de la justicia penal y para las cuales la peligrosidad del sujeto es uno de los criterios a los que el operador jurídico obligatoriamente tiene que atender.

Al mismo tiempo, los instrumentos de valoración del riesgo de violencia evidencian valores predictivos positivos mucho más bajos, casi nunca superiores al 50%[22]. Sin embargo, esta información raramente se ofrece en los estudios al respecto, que tampoco suelen incluir otro tipo de datos sobre el riesgo absoluto de reincidencia, como sería por ejemplo simplemente decir qué proporción de los sujetos clasificados en cada nivel de riesgo por el instrumento en cuestión reincidió en el tiempo de seguimiento estipulado. Y sin embargo es la información sobre el riesgo absoluto la que muchas veces puede ser más relevante en las decisiones judiciales.

---

[20] Para más información al respecto, cfr. MOSSMAN 2013; SJÖSTEDT y GRANN 2002; SINGH, 2013.

[21] Cfr. los valores de AUC reflejados supra en la Tabla 1. También MOSSMAN (2013:34) afirma que los modernos instrumentos estructurados de valoración del riesgo de violencia suelen arrojar valores de AUC de entre 0.65–0.80. No obstante, recuérdese lo que se ha advertido *supra* en el texto sobre la ausencia de un consenso fundamentado acerca de qué valores de la AUC deben considerarse elevados y cuáles no.

[22] Cfr. asimismo los valores reflejados en la tabla 1. Cfr también MARTÍNEZ GARAY, 2014, con información sobre los valores predictivos observados en estudios recientes en Alemania y en España.

Es por este conjunto de circunstancias por lo que al inicio de este trabajo nos referíamos al riesgo de que pueda extenderse un optimismo excesivo sobre el rendimiento de estos instrumentos, optimismo que puede resultar peligroso en el ámbito de la justicia penal. La información que circula sobre la capacidad predictiva de las modernas herramientas estructuradas de valoración del riesgo de reincidencia o de violencia tiende a destacar sus fortalezas (su aceptable poder de discriminación entre individuos con mayor o menor riesgo de reincidir), que pueden ser irrelevantes en un proceso penal, y sin embargo a no insistir mucho en algunas de sus debilidades (como su escaso poder de identificar a los individuos con mayor riesgo de reincidencia), que por el contrario pueden ser decisivas desde el punto de vista de la justicia penal.

Por ejemplo, un juez al que se le informa de que un sujeto ha obtenido una puntuación indicativa de riesgo alto de reincidencia en delitos sexuales con alguna de estas herramientas, y además se le explica que la capacidad predictiva de dicha herramienta es alta porque tiene un valor de área bajo la curva de 0.80, podría malinterpretar que 0.80 es el porcentaje de casos en que las predicciones hechas con dicho instrumento han sido confirmadas por la realidad, o que es la probabilidad de reincidencia que corresponde a los sujetos clasificados como de riesgo alto. En ambos casos estaría equivocado. Pero, lo que es más grave, esa incorrecta comprensión de la información recibida podría conducir a que el juez aplicase a esa persona una medida de seguridad restrictiva de libertad porque entienda que una probabilidad del 80% de volver a cometer delitos sexuales justifica dicha limitación de derechos, cuando en realidad el valor predictivo positivo (esto es, el porcentaje de sujetos calificados como de riesgo alto con dicho instrumento y que han evidenciado nuevos delitos sexuales graves en un periodo de seguimiento de 5 años) quizá no llega al 15%. Y si el juez hubiera sabido esto, quizá su decisión habría sido distinta.

> Por supuesto, el mismo tipo de confusión puede producirse respecto de la información referente a los sujetos que hayan recibido la calificación de riesgo bajo. Sin embargo, aquí las consecuencias de una mala interpretación por parte del operador jurídico son menos graves, porque los valores predictivos negativos suelen ser, como se desprende de los datos reflejados en la Tabla 1, mucho más elevados que los positivos, e incluso más elevados que los valores de la AUC. Aplicado al ejemplo del párrafo anterior: el juez al que se le informa de que un sujeto ha obtenido una puntuación indicativa de riesgo *bajo*, y además se le explica que la capacidad predictiva de la herramienta con la que ha sido evaluado el sujeto es alta porque tiene un valor de área bajo la curva de 0.80, podría perfectamente malinterpretar que 0.80 es el porcentaje de casos en que las predicciones hechas con dicho instrumento han sido confirmadas por la realidad, o el porcentaje de casos en que los sujetos clasificados como de riesgo bajo efectivamente no vuelven a reincidir. En ambos casos estará equivocado pero dicha confusión no es demasiado grave porque ese porcentaje (el valor predictivo negativo) probablemente será también de 0.80 o incluso superior – cfr. los valores de la Tabla 1 – de manera que el malentendido seguramente no hará al juez tomar una decisión distinta de la que habría adoptado en caso de haber comprendido correctamente lo que significa la AUC y haber conocido el valor predictivo negativo.

Las confusiones y los malos entendidos no tienen por qué limitarse, por otro lado, únicamente al significado del área bajo la curva, sino que puede ocurrir también en relación con otros indicadores: en otro lugar hemos ejemplificado cómo puede fácilmente malinterpretarse la información que se proporciona sobre el grado de acierto en la predicción si no diferencia suficientemente entre la sensibilidad – que también puede alcanzar en estos estudios valores

elevados – y el valor predictivo (MARTÍNEZ GARAY, 2016). Y sin embargo, es posible que al menos por ahora no sea el contexto forense aquel en el que resulte más peligrosa la falta de una información transparente sobre la fiabilidad de las herramientas de valoración del riesgo de reincidencia. En efecto, algunas investigaciones apuntan a que los jueces y tribunales no son – de momento – especialmente receptivos a las valoraciones de peligrosidad realizadas con la ayuda de herramientas estructuradas, y siguen confiando en el tradicional juicio clínico no estructurado tanto o más que en aquéllas[23]. Con ello no estamos afirmando que este estado de cosas sea bueno porque el juicio clínico sea mejor o más preciso que las valoraciones de riesgo estructuradas; nos limitamos a destacar el hecho de que la irrupción del moderno paradigma de la valoración del riesgo no parece haber aumentado en exceso la confianza de los jueces en las valoraciones de peligrosidad, sino que estos parecen mantener al respecto una cautela similar (sea ésta mucha o poca) a la que tradicionalmente han exhibido en relación con el juicio clínico tradicional. Con todo, está por ver durante cuánto tiempo se mantendrá este estado de cosas.

A nuestro juicio es en el ámbito de las políticas públicas de lucha contra la delincuencia, e incluso en el ámbito de la doctrina penal, donde los efectos de un optimismo exagerado sobre la capacidad predictiva de los modernos instrumentos estructurados de valoración del riesgo pueden tener efectos más nocivos. Si se extiende el convencimiento de que es posible estimar con un grado alto o al menos moderado de acierto el riesgo de reincidencia, sobre todo el riesgo de nueva comisión de delitos graves, es probable que los legisladores se muestren más dispuestos a introducir (y los juristas a aceptar) consecuencias jurídicas limitativas de derechos que descansan sobre la base de la peligrosidad, como los *civil commitments* en el ámbito estadounidense, la custodia de seguridad en Alemania y otros países, o penas como la prisión permanente, cuyo régimen de revisión en España depende de que se constate la ausencia de peligrosidad del sujeto[24]. Porque si creemos que es factible comprobar la existencia de su presupuesto de aplicación, hay más argumentos para defender que son consecuencias jurídicas no arbitrarias o incluso necesarias para una protección eficaz de la sociedad. Pero si se es consciente de que es muy difícil detectar con precisión a los delincuentes más peligrosos, y de que los pronósticos de alta peligrosidad tienen niveles de acierto muy bajos (en términos de valor predictivo positivo), la legitimidad de estas consecuencias jurídicas es mucho más cuestionable, porque significa reconocer que en la mayor parte de los casos se aplicarán a sujetos poco peligrosos a los que se limitarán drásticamente sus derechos fundamentales sin que de ello derive ningún aumento apreciable del nivel de protección para la sociedad.

La sentencia del Tribunal Constitucional alemán de 5 de febrero de 2004 es a nuestro juicio un ejemplo de este tipo de riesgos. En dicha resolución el tribunal confirmó la constitucionalidad de una reforma legal que había flexibilizado los presupuestos de aplicación de la custodia de

[23] Cfr. la bibliografía indicada al respecto en la nota 4.

[24] Art. 92.1 CP: "1. El tribunal acordará la suspensión de la ejecución de la pena de prisión permanente revisable cuando se cumplan los siguientes requisitos: […]
c) Que el tribunal, a la vista de la personalidad del penado, sus antecedentes, las circunstancias del delito cometido, la relevancia de los bienes jurídicos que podrían verse afectados por una reiteración en el delito, su conducta durante el cumplimiento de la pena, sus circunstancias familiares y sociales, y los efectos que quepa esperar de la propia suspensión de la ejecución y del cumplimiento de las medidas que fueren impuestas, pueda fundar, previa valoración de los informes de evolución remitidos por el centro penitenciario y por aquellos especialistas que el propio tribunal determine, la existencia de un pronóstico favorable de reinserción social. […]".

seguridad (medida de seguridad privativa de libertad que se puede imponer en ese país, para ejecutarse tras el cumplimiento de la pena de prisión, a los delincuentes que se considera que tienen un riesgo alto de reincidencia con delitos muy graves), y eliminado el plazo máximo de 10 años que hasta entonces tenía como límite cuando se le imponía a un delincuente por primera vez. El tribunal afirmó que el internamiento prolongado que supone la custodia de seguridad no vulnera la dignidad de la persona siempre que resulte imprescindible debido al peligro continuado que supone el sujeto. La comprobación de dicho peligro – que se convierte, asumido dicho punto de partida, en esencial para la legitimidad de la medida – no le pareció al tribunal especialmente problemática, porque en su opinión los estudios publicados al respecto indicarían que el conocimiento acerca de los factores de riesgo habría mejorado notablemente en los últimos años, lo que permitiría efectuar pronósticos relativamente buenos y fiables sobre una parte de los delincuentes[25]. Y el Tribunal concluyó que el pronóstico constituye una base apropiada para la decisión precisamente en los casos, poco frecuentes, de peligro muy elevado, en los que la reforma cuya constitucionalidad se discutía permitía prolongar indefinidamente la duración de la custodia de seguridad. En nuestra opinión, y tras todo lo que se ha explicado en los párrafos anteriores, dicha confianza en que el riesgo de reincidencia, especialmente el de reincidencia violenta grave, pueda ser estimado con suficiente seguridad resulta como mínimo cuestionable, y es probable que las expectativas de los operadores jurídicos y de las autoridades responsables de las políticas penales y penitenciarias sobre lo que es posible predecir con un mínimo de rigor estén bastante alejadas de lo que la evidencia empírica disponible permite sostener (cfr. SCURICH, 2016:178; DOUGLAS / PUGH / SINGH / SAVULESCU / FAZEL, 2017:135; FAZEL / SINGH / DOLL / GRANN, 2012:5).

5. La necesidad de información mejor, y más transparente

Que evaluar el rendimiento de las herramientas de valoración del riesgo de violencia únicamente a través de la AUC (o con alguna otra medida aislada de riesgo relativo) proporciona una información muy incompleta sobre su funcionamiento es algo que está empezando a ser reconocido en el ámbito de la criminología (SZMUCKLER / EVERITT / LEESE, 2011; SHEPHERD / SULLIVAN, 2017). Diversos autores recomiendan complementar el análisis de la capacidad predictiva incluyendo también otra información: por ejemplo, SINGH, 2013, propone que se ofrezcan no uno sino varios indicadores de capacidad predictiva y que se informe de sus respetivas limitaciones; DOUGLAS et al, 2017, recomiendan que se informe con claridad sobre los números de falsos positivos y falsos negativos; o ROSSEGGER et al 2014, que aconsejan que se ofrezcan indicadores tanto de discriminación como de calibración. Otros autores subrayan que en todo caso cualquier decisión judicial no puede basarse únicamente en la probabilidad de reincidencia o de violencia proporcionada por cualquier instrumento de valoración del riesgo,

[25] Sentencia del Tribunal Constitucional alemán, 2 BvR 2029/01, de 5.2.2004, número marginal 102. La sentencia cita expresamente un trabajo de NEDOPIL (2002), en el que efectivamente se afirma, como sostiene el Tribunal, que el conocimiento sobre los factores de riesgo de reincidencia ha mejorado notablemente en los últimos años, pero donde el autor también subraya que son cada vez más claras las fronteras de lo que es posible predecir, e insiste en las limitaciones que plantean las tasas de prevalencia bajas o desconocidas, en la imposibilidad de realizar predicciones a largo plazo, en la insuficiencia de los conocimientos empíricos existentes, y en las dificultades inherentes a la predicción de fenómenos tan complejos como el comportamiento humano.

porque en ella hay implicadas muchas otras cosas (MOSSMAN, 2013).

En relación especialmente con el valor informativo de la AUC, COOKE y MICHIE han denunciado que se trata de un indicador particularmente opaco, al cual resultan aplicables las críticas que en otros ámbitos se han formulado a los indicadores de riesgo relativo (COOKE / MICHIE, 2013) [26]. En efecto, algunos autores han señalado que en general la información dada en términos de riesgo absoluto es mucho más transparente y fácil de entender para el público – tanto especializado como profano – que la información proporcionada en términos de riesgo relativo. GIGERENZER et al (2008) sostienen que en el ámbito de la medicina informar sobre riesgos relativos sin especificar cuáles son las tasas de ocurrencia de los fenómenos puede considerarse una mala práctica porque conduce al receptor de la información a sobreestimar la magnitud del beneficio. Si por ejemplo se afirma que un medicamento reduce el riesgo de padecer determinada enfermedad en un 50%, ello puede significar que el riesgo desciende desde un 20% a un 10%, pero también que baja desde un 0.0002% a un 0.0001%. En ambos supuestos hay una reducción del riesgo relativo en un 50%, pero la relevancia clínica del efecto en un caso y otro es muy distinta (GIGERENZER et al, 2008:66). Como también lo son – añadiríamos – las decisiones de política sanitaria que deban adoptarse a partir de esta información (por ejemplo, financiar o no el tratamiento con fondos públicos, o incluirlo en el calendario de vacunación obligatorio). En Derecho penal cabría efectuar un razonamiento similar: afirmar que en determinada herramienta estructurada de valoración del riesgo la clasificación de un sujeto en una categoría de riesgo superior (por ejemplo, pasar de "riesgo moderado" a "riesgo alto") correlaciona con un aumento del riesgo relativo de reincidencia del 50% puede significar tanto que el riesgo moderado es un 40% de probabilidades de reincidencia y el alto un 80%, como que el riesgo moderado es 10% y el alto 20%. En ambos casos el riesgo relativo aumenta en un 50%, pero la relevancia de la información para la toma de una decisión judicial es muy distinta.

Habida cuenta que la información sobre la capacidad predictiva de las herramientas de valoración del riesgo de reincidencia puede plantear (serias) dificultades de comprensión para el público al que se dirige, es necesario un esfuerzo para facilitar la información de la manera más clara posible. Porque para evitar las confusiones y los malos entendidos no es imprescindible ser un experto en estadística (aunque sean necesarios unos conocimientos mínimos al respecto), sino que muchas veces basta con presentar la información de manera clara y transparente para que un público con unas habilidades matemáticas básicas sea capaz de comprender correctamente su significado (GIGERENZER et al, 2008).

6. Otras cuestiones problemáticas: estimaciones de riesgo en diferentes contextos

Con todo, de lo dicho hasta aquí no debe concluirse que en nuestra opinión los problemas que plantea la utilización de estimaciones estructuradas de riesgo de violencia en el Derecho penal se solucionen simplemente ofreciendo información sobre los valores predictivos, o incluyendo algún otro tipo de datos sobre el riesgo absoluto de reincidencia asociado a cada categoría de

[26] Según COOKE y MICHIE, el área bajo la curva es un indicador especialmente opaco porque se construye a partir de dos indicadores que también lo son: la sensibilidad y la especificidad, que son poco transparentes porque son probabilidades condicionadas (COOKE / MICHIE, 2013, así como GIGERENZER et al, 2008:77)

riesgo. De hecho, y a pesar de que pueda resultar paradójico después de todo lo dicho hasta ahora, a veces el valor predictivo puede no ser el indicador más relevante, o esconder otros problemas.

Para entender por qué a veces el valor predictivo puede tener sólo una relevancia relativa en la toma de determinadas decisiones, u ofrecer una información poco fiable, puede resultar ilustrativa la investigación hecha por LÓPEZ-OSSORIO et al sobre la eficacia predictiva de la valoración del riesgo en procesos por violencia de género (LÓPEZ-OSSORIO et al, 2016). Estos autores estudiaron una muestra de 407 casos de mujeres que habían denunciado ser víctima de este tipo de violencia, y a las que se hizo una valoración policial de riesgo para determinar el nivel de protección de que debían gozar mientras se investigaban los hechos. La valoración se hizo aplicando el protocolo de valoración policial del riesgo (en adelante, VPR), una herramienta actuarial que clasifica los casos en cinco niveles: riesgo no apreciado, riesgo bajo, medio, alto y extremo. De las 407 mujeres que fueron evaluadas, 211 (51.8%) presentaron riesgo «no apreciado», 157 eran casos de riesgo «bajo» (38.6%), 38 casos de riesgo «medio» (9.3%) y un caso de riesgo «alto» (0.2%). Una vez hecha la evaluación del riesgo se hizo un seguimiento a tres y a seis meses de las nuevas denuncias presentadas contra el mismo agresor. La tasa de reincidencia (medida como nuevas denuncias) fue del 4.91% (20 casos) a tres meses, y del 12.04% (49 casos) a seis meses. Los autores del estudio informan que el 81.6% de los nuevos delitos denunciados fueron de violencia leve, y el 18.4% de violencia grave (entendiendo por violencia tanto la física como la psíquica, así como la sexual y el acoso) (LÓPEZ-OSSORIO et al, 2016:5).

Para el análisis de la eficacia predictiva del VPR los autores combinaron los cinco niveles de riesgo en dos: no-presencia y presencia (que incluyó los niveles bajo, medio y alto), lo que permitió disponer de una variable dicotomizada para calcular sensibilidad, especificidad y valores predictivos a partir de una matriz de 2 × 2. Los datos que se obtuvieron fueron los siguientes:

Tabla 6. Estimación policial de riesgo de nueva victimización por violencia de género y reincidencia registrada en un intervalo de 3 meses posteriores a las evaluaciones

Riesgo estimado	Reincidencia SÍ	Reincidencia NO	Total	Valor predictivo
Sí	17	179	196	Positivo: 8.6%
NO	3	208	211	Negativo: 98.5%
Total	20	387	407	
	Sensibilidad: 85%	Especificidad: 53.7%		

Fuente: elaborada a partir de los datos de la tabla 2 en LÓPEZ-OSSORIO et al, 2016:5

Los autores informan en el estudio sobre la sensibilidad, la especificidad y los valores predictivos, en el seguimiento a tres meses, como puede verse en la Tabla 6, y también aportan el valor del

área bajo la curva (0.714) así como un análisis por medio del modelo de regresión logística, y la *odds ratio* (6.585). A los seis meses los valores de sensibilidad, *odds ratio* y área bajo la curva fueron más bajos: 61.2%, 0.58 y 1.826, respectivamente (en el estudio no se ofrecen los valores predictivos en la estimación de riesgo a seis meses).

Los autores valoran la validez predictiva del formulario VPR como satisfactoria, afirmando que "las estimaciones del protocolo VPR son adecuadas tal y como está configurado en la actualidad", especialmente por lo que hace a la predicción a tres meses. Añaden que "las predicciones en este intervalo temporal demuestran la existencia de una relación significativa entre el pronóstico de riesgo de las valoraciones y los episodios de reincidencia." Por lo que respecta a la capacidad de la herramienta para estimar el riesgo a seis meses, los autores afirman que la utilidad del VPR en este caso es baja, ya que aumenta la probabilidad de falsos negativos pues la sensibilidad disminuye al 61.2%; aunque se mantiene igual en la identificación de casos negativos sin incrementar el riesgo de falsos positivos, puesto que la especificidad continua en el 53.6% (LÓPEZ-OSSORIO et al, 2016:6).

¿Resultan significativos a la hora de valorar el rendimiento del VPR los valores predictivos que se evidenciaron en este estudio? El valor predictivo negativo es de más del 98%. Esto significa que de cada 100 denuncias por violencia de género que fueron considerados como casos de ausencia de riesgo por la policía, en 98 efectivamente la víctima no ha vuelto a sufrir nuevas agresiones, y por tanto la decisión de no aplicar medidas cautelares de protección se evidenció a posteriori como acertada. Es un porcentaje de acierto en nuestra opinión bastante elevado, que supone que el VPR es un buen indicador para seleccionar los casos de bajo riesgo de reincidencia. El valor predictivo positivo, por el contrario, es bajísimo: sólo el 8,6%. De cada 100 casos de violencia de género en los que se aplicó algún tipo de medida de protección a la víctima, en apenas nueve tuvo lugar una nueva agresión. ¿Es éste un mal dato? No necesariamente. Porque como acertadamente indican los autores del estudio, "el protocolo de valoración policial de riesgo implementado en el sistema VioGen requiere que, inmediatamente después de la valoración policial del riesgo con el formulario VPR se pongan en práctica una serie de medidas de protección que pretenden garantizar la seguridad de las víctimas, tratando de evitar precisamente la reincidencia. De ahí que sea esperable que esta intervención policial influya de algún modo en los parámetros sobre sensibilidad y especificidad del formulario (aciertos de ocurrencia y no ocurrencia)". Nuestra única discrepancia con esta valoración sería que la aplicación de medidas policiales de protección no distorsiona solo la sensibilidad sino también otro parámetro, precisamente el valor predictivo positivo. Porque cabe pensar que si no se hubiera aplicado ninguna medida de protección de la víctima en ninguno de los 196 casos en los que el VPR consideró que había algún grado de riesgo, quizá en lugar de 17 nuevos episodios de agresión habría habido más. Es decir, que la aplicación de la consecuencia jurídica prevista para la existencia de riesgo ha podido contribuir a disminuir precisamente el mismo riesgo que se había pronosticado. En otras palabras, un valor predictivo positivo muy bajo en este estudio podría ser interpretado como una confirmación de la eficacia preventiva de las medidas policiales de protección de las víctimas de violencia de género.

El problema es que no podemos estar seguros. Para saber si el altísimo porcentaje de "falsos positivos" en este estudio es en realidad una prueba de que el VPR sobreestima el riesgo de

reincidencia en materia de violencia de género, o es por el contrario un indicio de eficacia de la prevención policial, sería necesario tener una muestra equivalente de mujeres a las que, a pesar de haber obtenido una evaluación positiva de riesgo, no se les hubiera aplicado medida alguna de protección, y entonces comparar la reincidencia en ambas muestras a tres y a seis meses. Pero, como resulta obvio, no parece admisible llevar a cabo esta clase de "diseño experimental" en la práctica (cfr. BUSHWAY / SMITH, 2007, subrayando este tipo de problemas metodológicos y las dificultades que existen para superarlos).

¿Qué consecuencias cabe extraer entonces de estos datos? En nuestra opinión, el alto valor predictivo negativo demostrado por el VPR en el estudio de LÓPEZ-OSSORIO et al avala la utilización de esta herramienta como criterio de exclusión, es decir, cabe confiar en que un resultado de ausencia de riesgo obtenido con dicho instrumento muy probablemente irá acompañado de una ausencia real de nuevas agresiones y puede ser considerado un buen argumento para no adoptar sobre el denunciado medidas limitativas de derechos, al menos en el corto plazo[27]. Por el contrario, cuando el VPR indica la presencia de riesgo, el bajísimo valor predictivo evidenciado en el estudio no es por sí solo criterio suficiente para contraindicar la adopción de medidas de protección. Y ello porque muchas de las medidas policiales de protección de la víctima no limitan derechos fundamentales del denunciado (como por ejemplo proporcionar un teléfono de emergencia a la víctima, mantener con ella contactos telefónicos o personales, realizar vigilancias de su domicilio y de trabajo, etc.[28]), por lo que su implementación puede contribuir a reducir riesgos de nuevas agresiones sin perjudicar a una persona cuya culpabilidad aún no se ha demostrado. Se gana en seguridad sin para ello limitar derechos de terceros[29]. La situación cambia, por supuesto, cuando como consecuencia de la valoración de riesgo se apliquen medidas de protección para la víctima que sí impliquen limitaciones de derechos del denunciado, por ejemplo una orden judicial de alejamiento y prohibición de comunicación. En estos casos la valoración ha de ser más diferenciada, pero incluso aquí hay elementos que podrían aconsejar implementar las medidas de protección aun a pesar del escaso valor predictivo positivo del VPR: por ejemplo, el hecho de que dichas medidas son siempre de duración limitada mientras se instruye el proceso y están en continua revisión, o que en caso de dictarse finalmente una sentencia condenatoria se pueden abonar a la condena impuesta, por lo que acaban formando parte del castigo merecido por el culpable.

> La cuestión tiene, desde luego, muchas aristas: en sentido contrario cabe argüir que el abono de la medida cautelar a la pena no es posible cuando el proceso termina con sentencia absolutoria. También el hecho de que las nuevas agresiones sean en su gran mayoría clasificables como leves (el 81.6%, según LÓPEZ-OSSORIO et al, 2016:5) puede ser un argumento en contra de la legitimidad de aplicar de manera generalizada medidas de protección que supongan fuertes restricciones de derechos para el denunciado. En todo caso, falta en el estudio citado información más detallada para pronunciarse con mayor fundamento sobre estos extremos, pues el estudio no diferencia el número y clase de nuevas agresiones en función del nivel de riesgo en que se había colocado al agresor, es decir, no sabemos si las 17 nuevas

[27] El hecho de que el estudio no informe de los valores predictivos en el seguimiento a seis meses impide extender esta conclusión a periodos de tiempo superiores a tres meses.

[28] Cfr. las diversas medidas de protección posibles, ordenadas según el nivel de riesgo que se haya asignado al caso, en la Instrucción 7/2016, de la Secretaría de Estado de Seguridad, por la que se establece un nuevo protocolo para la valoración policial del nivel de riesgo de violencia de género.

[29] El problema será, en todo caso, económico: si existen recursos personales y materiales para proporcionar protección a todas las mujeres con evaluación positiva de riesgo aunque éste se considere bajo.

> agresiones (verdaderos positivos) que se produjeron en los tres primeros meses tras la evaluación corresponden a supuestos clasificados como de riesgo bajo, medio, o alto.

En definitiva, la peligrosidad del sujeto es un criterio que según la legislación vigente debe ser tenido en cuenta en muy diversos tipos de decisiones judiciales o penitenciarias, cada una de las cuales puede estar condicionada de manera distinta por el necesario respeto a criterios jurídicos como los principios de proporcionalidad y de presunción de inocencia. Por ello, no pueden hacerse afirmaciones generales sobre el nivel de capacidad predictiva mínimo que debe mostrar una herramienta de valoración del riesgo para que su uso pueda ser considerado aceptable en cualquier contexto, ni es siempre el mismo indicador de capacidad predictiva el relevante. Si acaso, la única afirmación que puede hacerse con pretensiones de generalidad es la de que siempre es necesario extremar la cautela, asegurarse en todos los casos de que se comprende bien el significado de la información proporcionada y las limitaciones de la misma, y no olvidar que en ningún caso un determinado nivel de riesgo estimado justifica por sí solo la adopción de ninguna decisión.

7. Discusión

De todo lo expuesto hasta aquí se desprenden a nuestro juicio varias consecuencias.

1. La primera, que la valoración de la capacidad predictiva de los instrumentos estructurados de valoración del riesgo de reincidencia es un asunto muy complejo, porque se expresa con muchos tipos de indicadores diferentes, porque estos son en ocasiones parámetros estadísticos y matemáticos muy técnicos y de difícil comprensión para el profano, y también porque la forma de presentar la información no es todo lo clara que sería deseable y tiende a subrayar las fortalezas de estos instrumentos y a dejar en la sombra sus debilidades.

> Y todo ello sin entrar en otro tipo de problemas, de mayor alcance, que afectan a este tipo de investigación científica y que han sido denunciados por la literatura especializada, no sólo para el ámbito de la psicología o de la criminología, pero en los que no podemos detenernos aquí. Por apuntar sólo unas ideas, SIONTIS et al (2015) han advertido que en medicina proliferan nuevos métodos de estimación de riesgos que en la mayor parte de las ocasiones no están suficientemente validados por estudios ulteriores hechos por autores independientes, y, lo que es también muy relevante, cuando estos sí se llevan a cabo suelen dar como resultado que el modelo de estimación de riesgos analizado tiene una validez predictiva inferior a la que se anunció en el estudio original. Y el descenso en la capacidad predictiva es aún mayor cuando el estudio de validación ha sido realizado por autores distintos a los del estudio original. La escasez de estudios de validación también es un hecho en psicología, lo que ha llevado a algún autor a afirmar que más del 50% de los nuevos descubrimientos que se pretenden haber realizado en esta disciplina podrían ser simplemente falacias aún no descubiertas ('*unchallenged fallacies*', IOANNIDIS, 2012). Además, cuando se hacen estudios de validación sobre herramientas de valoración de riesgo de reincidencia no siempre se respetan en el diseño los criterios establecidos en el estudio original, lo que dificulta interpretar los resultados como confirmación o refutación de la capacidad predictiva de la herramienta de que se trate (ROSSEGGER et al, 2013). Y también se ha señalado la existencia de sesgos de autor en los estudios sobre eficacia predictiva de las herramientas estructuradas de valoración del riesgo de violencia (SINGH / GRANN, /FAZEL, 2013).
>
> Por otro lado, no cabe desconocer que los conflictos de intereses puestos de relieve desde hace tiempo en relación con la industria farmacéutica y la medicina son aplicables *mutatis mutandis* al campo de la valoración del riesgo de reincidencia. Sin ir más lejos, las herramientas de valoración del riesgo de

violencia se comercializan, en ocasiones por empresas con ánimo de lucro; el software necesario para implementarlas y los manuales de instrucciones se venden, se cobra por sesiones de entrenamiento a los funcionarios que deberán aplicarlas, etc (SINGH / GRANN / FAZEL, 2013). Para todo ello, es necesario convencer de que su rendimiento es óptimo, y a ello contribuye la publicación de estudios científicos que avalan su capacidad predictiva.

Por ello en nuestra opinión cualquier explicación de la capacidad predictiva de las herramientas estructuradas de valoración del riesgo de violencia que trate de simplificar la cuestión en exceso resulta sospechosa, especialmente si va acompañada de la recomendación de utilizar estas herramientas en las resoluciones judiciales o administrativas sobre el tratamiento de los acusados o condenados[30].

Y es que la tecnificación y sofisticación estadística de las modernas herramientas de valoración del riesgo de reincidencia o de violencia no puede hacernos olvidar que siempre que se aplican como criterios para la toma de alguna decisión en el ámbito penal se está haciendo a la vez un juicio (explícito o implícito) sobre cuáles son los riesgos que *deben* asumirse o evitarse, o las restricciones de derechos que resultan *aceptables* o no, todo lo cual presupone como es obvio decisiones valorativas. Cualquier decisión adoptada sobre la base de la información que estos instrumentos proporcionan presupone una opción político-criminal. Puede haber ámbitos en los que se considere prioritario minimizar el riesgo de falsos negativos (no detectar como peligrosos a sujetos que sí continuarán delinquiendo) aun a costa de aceptar la existencia de un alto porcentaje de falsos positivos (véase el ejemplo de la investigación de LÓPEZ-OSSORIO, 2016, citado *supra*). Aquí podría ser decisiva una herramienta con una sensibilidad elevada. Pero si por el contrario en otro tipo de decisiones lo que debe evitarse es imponer privaciones de libertad a sujetos que no van a seguir delinquiendo, entonces interesará un instrumento con una especificidad y un valor predictivo positivo altos. Sin embargo, lo fundamental en todos estos casos es justificar por qué resultan aceptables, o no, en unos contextos sí y en otros no, determinados riesgos. Entender la aplicación de las valoraciones estructuradas de riesgo como una mera cuestión técnica que puede ser resuelta con un programa informático y un algoritmo oculta toda esta problemática, que es en realidad la decisiva desde el punto de vista jurídico y de derechos fundamentales.

2. Por ello, nos parecen cuestionables posiciones como la manifestada por la APA reiteradamente en sus *amici briefs* sobre casos de imposición de pena capital, a los que nos hemos referido al inicio de este trabajo. En primer lugar porque la contraposición que hacen entre juicios clínicos sobre peligrosidad emitidos por expertos, de un lado, y el conocimiento profano o *lay knowledge* que supuestamente es el propio de las valoraciones estructuradas de riesgo de reincidencia, por otro, puede dar la imagen de que estas últimas son una cuestión más sencilla, hecha con procedimientos más simples, cuya comprensión está al alcance de cualquiera aunque no sea

[30] En nuestra opinión incurre en este defecto el conocido video de Anne MILGRAM en una TEDTalk, que bajo el título "Why smart statistics are the key to fighting crime" reclama aplicar en el contexto forense (y especialmente en las decisiones relativas a la prisión provisional) herramientas actuariales basadas en *big data* por analogía con el método estadístico que Billy Beane, gerente del equipo de béisbol Oakland Athletics, implantó con gran éxito para fichar jugadores según la conocida película *Moneyball*. MILGRAM aboga por lo que ella denomina "*moneyballing criminal justice*" afirmando que el método funcionó en el béisbol y funciona en la justicia penal. (Puede verse la conferencia online en: https://www.ted.com/talks/anne_milgram_why_smart_statistics_are_the_key_to_fighting_crime/transcript?quote=1850).

experto en estadística, ni en criminología, ni en psicología. Pero en este punto hay que diferenciar dos cosas. La *aplicación* de una herramienta actuarial de valoración del riesgo puede ser, en efecto, un procedimiento bastante simple. En ocasiones no lleva más de una hora y consiste en rellenar un cuestionario de unos pocos ítems a partir de información fácilmente accesible en el expediente judicial o penitenciario (como por ejemplo edad en el momento de la excarcelación, número de antecedentes penales, existencia o no de delitos sexuales anteriores, sexo de la víctima, etc.[31]), sin necesidad siquiera de tener delante al sujeto evaluado. Sin embargo, y como he tratado de ilustrar en este trabajo con el ejemplo del área bajo la curva, no hay nada de sencillo ni en la *construcción* de las herramientas de valoración del riesgo, ni en su *interpretación*. La complejidad de los cálculos, la selección de los puntos de corte y las decisiones valorativas que ello implica, el significado exacto de cada estadístico, la relevancia de cada uno, los márgenes de error a que están sometidos, etc. son materias complejísimas y oscuras, y el lego en estadística es completamente manipulable por el experto. Ya hemos visto que sólo la forma en que se presentan unos determinados resultados puede ser determinante para crear una imagen más o menos favorable sobre el funcionamiento de una de estas herramientas. Pero a ello hay que sumar otros factores. De un lado, en ocasiones algunos de los ítems que deben ser comprobados requieren una valoración por parte del evaluador (por ejemplo, decidir si el sujeto tiene o no "actitud hostil o valor procriminal", "desajuste infantil", si presenta "temeridad" o "irresponsabilidad", incluidos como ítems en el RisCanvi), de manera que la 'objetividad' del resultado incluye en realidad en buena medida apreciaciones personales del evaluador[32]. Por otro lado, existe evidencia de que no todos los evaluadores siguen al pie de la letra el manual de instrucciones, y cumplimentan los formularios con mayor o menor discrecionalidad según confíen más o menos en el instrumento que aplican o en su propia experiencia personal[33]. A todo ello hay que añadir que algunos ítems dependen de información personal que ha de facilitar el sujeto y sobre la que puede mentir, o no recordar, etc. En definitiva, tras la *aparente* simplicidad de los resultados de las evaluaciones estructuradas del riesgo de violencia hay una maraña de números, fórmulas y cálculos bastante complicados, y la *impresión* de mayor transparencia y objetividad que sin duda generan puede ocultar una buena proporción de apreciaciones subjetivas o sesgadas, que sin embargo quedan invisibilizadas por los números y las fórmulas[34].

De otro lado, y por lo que hace a la contraposición que efectúa la APA entre opinión 'experta'

---

[31] Cfr. la descripción de los 10 ítems, similares a estos, que constituyen el cuestionario del Static-99R (una de las herramientas actuariales de valoración del riesgo de violencia más utilizadas en los EEUU para determinar el internamiento post-condena de los *sexual violent predators*) en NGUYEN VO / ANDRÉS PUEYO, 2016.

[32] Cfr. los ítems del RisCanvi en CAPDEVILA CAPDEVILA et al, 2015:125 y s. Muchas otras herramientas de valoración del riesgo también incorporan ítems de este estilo, cfr. HANNAH-MOFFAT / MAURUTTO / TURNBULL, 2009:402.

[33] Cfr. sobre este conjunto de circunstancias HANNAH-MOFFAT, 2015; SKEEM, 2013:302y s. y HANNAH-MOFFAT / MAURUTTO / TURNBULL, 2009, quienes añaden que cuando los profesionales que cumplimentan las herramientas de valoración del riesgo no están de acuerdo con el resultado final porque no coincide con el que según su experiencia y conocimiento del caso les parece más adecuado, generalmente optan por inflar o reducir artificialmente las puntuaciones hasta obtener una cifra final que les parezca más ajustada, en lugar de corregir expresamente el resultado inicial haciendo constar en el expediente esta modificación y las razones que la motivan (2009:405)

[34] Como afirma HANNAH-MOFFAT, 2015: "Many practitioners have commented that they strategically exercise their discretion when filling out risk assessments. However, this exercise of discretion is rarely visible because risk tools are effectively "black boxes" that sanitize the subjective input of practitioners to produce simple "objective" scores".

cuando es un psiquiatra o psicólogo el que emite un juicio clínico sobre el riesgo de reincidencia, y conocimiento 'lego' o no especializado (*lay knowledge*) que sería el que transmitiría quien informa sobre los resultados de la aplicación de una valoración estructurada de riesgo de reincidencia, cabe señalar que también quienes diseñan las herramientas de valoración del riesgo y pueden acudir a los tribunales a declarar sobre sus resultados son "expertos" – quizá no en psiquiatría, pero probablemente en criminología, estadística, sociología o psicología – y muy probablemente ostenten un título universitario, que en la mayor parte de los casos será el de doctor en alguna de estas disciplinas, cuando no serán incluso catedráticos en alguna prestigiosa institución de enseñanza superior. Nos estamos seguros de que haya razones para pensar que un jurado o un juez será menos impresionable por esos últimos 'expertos' y el conocimiento 'profano' que exhiban, que por los tradicionales psicólogos o psiquiatras[35].

También nos parece cuestionable en los informes de la APA su insistencia en desacreditar los juicios clínicos de peligrosidad en los casos de pena capital, pero a la vez subrayar la mucha mejor calidad y fiabilidad de las estimaciones estructuradas, aconsejando que se utilicen éstas como apoyo para demostrar la peligrosidad, que en muchas de las jurisdicciones de EEUU que conservan este tipo de pena es uno de los criterios fundamentales para imponerla. Porque presentar la cuestión en estos términos puede llevar a pensar que dicha peligrosidad *puede demostrarse científicamente* (con valoraciones actuariales). Cuando hay muchas dudas de que éste sea el caso: recuérdense los bajos porcentajes de valor predictivo positivo de la Tabla 1, especialmente cuando la delincuencia que se intenta predecir es la más violenta (y por ello la menos frecuente). Que las herramientas actuariales funcionen bien para algunas cosas no significa que lo hagan siempre, y el contexto de la detección de los individuos más peligrosos para aplicarles las consecuencias jurídicas más graves es precisamente aquél en el que los números de errores (falsos positivos) son los más altos. Lo que ha cambiado en los últimos 40 o 50 años es el modo en que se estima la peligrosidad y los métodos que se usan para presentar la información, pero no sustancialmente el porcentaje de acierto en las predicciones, que sigue siendo similar al que evidenciaron los estudios sobre los casos Baxstrom y Dixon en los años 70 y 80 del siglo pasado (en este sentido, expresamente, MÜLLER / STOLPMANN, 2015).

3. Y esto engarza con otra consideración importante. La *creencia* en que algo puede ser medido con precisión favorece la tendencia a considerarlo un fundamento admisible para una consecuencia jurídica. Si creyéramos que la estimación del riesgo de reincidencia es un constructo evanescente plagado de prejuicios y sesgos, imposible de medir con fiabilidad, seríamos probablemente mucho más reacios a establecerlo como presupuesto de consecuencias jurídicas que pueden llegar a limitar muy drásticamente los derechos fundamentales de los afectados, porque entenderíamos que dichas consecuencias, aplicadas sobre una base tan endeble, serían arbitrarias y por ende ilegítimas. Pero si por el contrario *creemos* que el riesgo de reincidencia es un concepto sólido, fundamentado en múltiples investigaciones científicas publicadas en revistas

[35] Diversas investigaciones apuntan que los individuos confían más en la evidencia percibida como 'empírica' que en la que no se presenta con esta etiqueta, y que resulta difícil resistir el poder de los números y de los resultados de cálculos algorítmicos especialmente cuando no se es experto en estadística, aparte de que proporcionar una determinada estimación actuarial de riesgo puede favorecer el sesgo de anclaje, y condicionar el resto de información que se reciba sobre ese caso, que será interpretada a la luz de la impresión que generó la primea cifra o valoración (cfr. sobre ello CHRISTIN / ROSENBLAT / BOYD, 2015 y RECENT CASES, 2017:1536, con ulteriores referencias).

internacionales de prestigio, y que puede ser medido de forma objetiva y fiable, estaremos más dispuestos a aceptar consecuencias jurídicas basadas en este presupuesto con mucha mayor facilidad (incluso, la pena de muerte). La presentación de la información sobre el rendimiento de la valoración estructurada del riesgo de reincidencia y de violencia de forma excesivamente optimista favorece nuestra predisposición a optar por un modelo de Derecho penal basado en ese criterio.

Porque el Derecho penal puede articularse sobre el peligro de comisión de nuevos delitos como criterio central a la hora de imponer determinadas consecuencias jurídicas, o no hacerlo. Un sistema jurídico puede prever medidas de seguridad post-condena para delincuentes peligrosos, o no tenerlas. Puede configurar la libertad condicional como de concesión automática tras el cumplimiento de determinada parte de la condena, o como de concesión discrecional en función del peligro de reincidencia que represente el sujeto. Puede tener condenas con duración máxima predeterminada de acuerdo a la gravedad del hecho, o flexible en función del riesgo de reincidencia que se aprecie en el sujeto en sucesivas revisiones. La bibliografía especializada que subraya la solidez científica y la utilidad para el ámbito forense de las herramientas estructuradas de valoración del riesgo de violencia, y especialmente la asunción acrítica de sus afirmaciones desde el campo del Derecho, puede terminar influyendo en nuestras concepciones sobre el castigo legítimo. Como afirma HARCOURT, si en vez de haber perfeccionado tanto la medición del riesgo de reincidencia dispusiéramos de algo así como un termómetro para medir la culpabilidad, por ejemplo si científicos biomédicos hubieran desarrollado algún tipo de escáner cerebral para medir con precisión el grado de intencionalidad de la conducta, quizá nos mostráramos mucho más dispuestos a establecer la magnitud del injusto culpable cometido como criterio decisivo en muchas decisiones judiciales y penitenciarias, antes que el riesgo de reincidencia (HARCOURT, 2007:189 y *passim*).

4. La penetración de dinámicas estructuradas y actuariales en la valoración del riesgo de reincidencia es aún limitada en nuestro país, pero creemos que puede afirmarse que está en expansión[36]. Aunque no es común que se utilice esta metodología por ejemplo cuando se pide un informe forense sobre la peligrosidad como base para adoptar una medida de seguridad, la valoración estructurada del riesgo de reincidencia sí es una realidad en el ámbito penitenciario, tanto en el territorio dependiente de la Secretaría General de Instituciones Penitenciarias como en la Comunidad Autónoma de Cataluña que tiene transferida la gestión penitenciaria: en el primero, con la utilización de la tabla de variables de riesgo y la de concurrencia de circunstancias peculiares, especialmente en la concesión de permisos de salida y propuestas de concesión de libertad condicional (DAUNIS RODRÍGUEZ, 2016), y, por lo que respecta a Cataluña, por la introducción del sistema RisCanvi en todo el ámbito de la gestión penitenciaria (ANDRÉS-PUEYO, 2013; ANDRÉS-PUEYO, 2017; ANDRÉS-PUEYO / ARBACH-LUCIONI / REDONDO, 2018). Y como hemos visto ha empezado a ser utilizado en otros contextos como el de la valoración policial del riesgo en la violencia de género.

Las críticas que en este trabajo se han hecho a la utilización irreflexiva de la valoración del riesgo

[36] BRANDARIZ GARCÍA indica diversos factores que pueden haber contribuido al significativo retraso de España en incorporarse a lo que este autor denomina 'modelo gerencial-actuarial' de justicia penal, (2016:27 y ss.), pero también señala que es probable que su influencia aumente y perdure en el tiempo (2016:257).

de reincidencia en el contexto penal no significan, en modo alguno, que debamos despreciar los resultados de las investigaciones hechas sobre esta materia, ni mucho menos cerrar los ojos ante la información sobre la realidad que nos aportan. La información que proporciona la investigación empírica sobre las herramientas estructuradas de valoración del riesgo puede ser muy valiosa, pero ha de ser analizada con cautela, y valorada desde una perspectiva crítica.

8. Conclusiones

1. La valoración del riesgo de violencia o de reincidencia con herramientas estructuradas (ya sean puramente actuariales o de juicio clínico estructurado) es un procedimiento complejo, y la comprensión de sus resultados no es absoluto sencilla y requiere un mínimo conocimiento de algunos conceptos estadísticos básicos. A una correcta interpretación de su capacidad para predecir con acierto el riesgo de violencia no contribuye la forma en que usualmente se presentan los resultados de las investigaciones en este ámbito en la literatura criminológica especializada, que tiende a destacar aquellos aspectos en que estas herramientas funcionan mejor y a subrayar mucho menos aquéllos otros en que sus resultados son más pobres, a lo cual hay que sumar que la información suele proporcionarse con medidas de riesgo relativo, muy complejas de entender para el profano, y haciendo hincapié en indicadores que muchas veces tienen escasa relevancia para la práctica judicial.

2. Todo ello puede contribuir a que se extienda un optimismo injustificado sobre el rendimiento de estas herramientas, que en el ámbito forense puede favorecer la toma de decisiones limitativas de derechos fundamentales de los acusados o condenados carentes de suficiente fundamento, y en el ámbito de la política criminal puede alentar la proliferación de instituciones jurídicas cuyo presupuesto sea el riesgo de reincidencia en la creencia de que dicho riesgo puede comprobarse empíricamente con facilidad, cuando ello no es así.

3. Es necesario mejorar la calidad y la transparencia de la información que se proporciona sobre las herramientas estructuradas de valoración del riesgo de reincidencia o de violencia, y ser conscientes de que cualquier simplificación excesiva en este punto es engañosa. Las valoraciones del riesgo efectuadas con estos instrumentos, por mucho que se presenten como objetivas, científicas, y fundamentadas en datos empíricos, nunca lo pueden ser en la misma medida en que lo es por ejemplo una predicción meteorológica. Las valoraciones estructuradas del riesgo de violencia incluyen valoraciones y pre-comprensiones sobre lo que se mide y cómo se mide – no siempre explícitas –, y la forma en que se presenta la información sobre sus resultados puede sugestionar indebidamente a un jurado o a un juez de forma similar a como puede hacerlo un informe clínico sobre peligrosidad emitido por un psiquiatra.

4. Desde el punto de vista jurídico, fundamentar consecuencias penales limitativas de derechos en la peligrosidad del sujeto más que en la gravedad del injusto culpable cometido no es menos problemático ahora de lo que lo era en los años 70 u 80 del siglo pasado, ni ha mejorado notablemente la capacidad para identificar con precisión a los individuos que probablemente reincidirán con la comisión de delitos graves. Quizá lo que la investigación sugiere en este punto es precisamente lo contrario de lo que nos gustaría escuchar: que si bien podemos discriminar con un relativo grado de acierto entre individuos con mayor o menor riesgo de reincidencia

dentro de un grupo, e identificar con un grado de acierto elevado a los menos peligrosos, seguimos sin poder estimar con un grado de precisión superior al azar quiénes serán los que continuarán cometiendo los delitos más graves. El conocimiento de esta realidad, si bien desacredita – en nuestra opinión – las pretensiones de dotar de fundamento empírico a los regímenes agravados de responsabilidad para los sujetos supuestamente más peligrosos, por el contrario proporciona una muy buena base argumentativa para ampliar las medidas alternativas a la prisión para los casos de riesgo bajo.

5. Dar entrada en el modelo de responsabilidad penal a las valoraciones de riesgo debe estar rodeado siempre de cautelas. Porque incluso aunque se mantenga la respuesta penal dentro de los márgenes del injusto culpable en los casos de riesgo alto, sin permitir agravaciones de la responsabilidad penal ni del régimen penitenciario en estos supuestos, establecer flexibilizaciones o mitigaciones para los de riesgo bajo comporta establecer una diferencia de trato entre personas basada en pronósticos e independiente de la gravedad del hecho cometido. Que esto resulte aceptable o no, no solo depende de que el riesgo (bajo o alto) pueda ser medido con precisión, sino de si se ajusta a nuestras convicciones sobre la justicia y sobre los derechos fundamentales, y sobre cuáles deben ser los parámetros cuyo respeto convierte en legítima la coacción estatal a través del *ius puniendi*.

9. Limitaciones

El objeto de este trabajo se ha centrado en el análisis de la información que circula acerca de la capacidad predictiva de las herramientas estructuradas de valoración del riesgo de reincidencia, y de qué modo puede dar lugar a malentendidos que favorezcan un optimismo excesivo al respecto por parte de los juristas y del público en general. Pero no aborda otros problemas metodológicos importantes de estas herramientas, como por ejemplo la misma ambigüedad inherente a aquello que se predice[37], las limitaciones que suponen las cifras negras cuando la información se obtiene a partir de archivos oficiales, la información que se pierde por la falta de respuesta o por las mentiras deliberadas de los sujetos a los que se aplican estas herramientas, y en general las limitaciones inherentes al enfoque cuantitativo en el que se basa toda la investigación sobre valoración estructurada del riesgo[38]. Por otro lado, este trabajo se ha centrado en la *predicción del riesgo* de reincidencia y en dicho riesgo como presupuesto para la adopción de consecuencias jurídicas limitativas de derechos en el ámbito del Derecho penal. Pero no analiza otro ámbito esencial como es el de la *gestión del riesgo*[39], en el que varias de las afirmaciones que aquí se han efectuado deberían ser matizadas, pues es un ámbito que se rige por principios en

[37] ¿"Delincuencia" son nuevos arrestos, que quizá terminan en absolución, sólo nuevas condenas, también lo que el sujeto informa aunque no haya sido objeto de investigación oficial? ¿Es "delincuencia" también el quebrantamiento de condiciones impuestas durante los periodos de supervisión en libertad que determinan la aplicación de una consecuencia jurídica más grave? ¿Qué entendemos por "violencia"?: ¿agresión física, intimidación, fuerza sobre las cosas, insultos e injurias? ¿Es "violencia sexual" también el exhibicionismo ante adultos, que al menos en España es considerado atípico?

[38] Sobre dichas limitaciones, que en nuestra opinión no se deben despreciar, cfr. las interesantes consideraciones de YOUNG, 2015.

[39] "La gestión del riesgo hace referencia a la aplicación de los conocimientos disponibles generados en los estudios de valoración del riesgo para minimizar la frecuencia actual de las conductas violentas y delictivas así como sus efectos" (ANDRÉS-PUEYO / REDONDO ILLESCAS, 2007:165)

parte distintos a los propios del Derecho penal. En el epígrafe 6 ya se ha indicado por ejemplo que la gestión eficaz del riesgo predicho puede ser precisamente uno de los factores que distorsionen el acierto de la predicción, lo que hace que en determinados contextos el valor predictivo positivo pierda relevancia como indicador del rendimiento de una herramienta de valoración del riesgo. También en el epígrafe 3.1. se ha apuntado que los indicadores de riesgo relativo pueden ser de mucha utilidad en el ámbito penitenciario cuando se tiene que decidir asignar recursos escasos de tratamiento a los individuos que, dentro de dicha población, se estime que más los necesitan. Pero la cuestión es mucho más compleja y no era objeto del presente trabajo analizarla[40].

10. Bibliografía

AMERICAN PSYCHIATRIC ASSOCIATION (1982), *Barefoot v. Estelle: Brief Amicus Curiae* (https://www.psychiatry.org/psychiatrists/search-directories-databases/library-and-archive/amicus-briefs, consultado por última vez el 23-10-2017)

AMERICAN PSYCHIATRIC ASSOCIATION (2007), *US v. Fields: Brief Amicus Curiae* (http://www.apa.org/about/offices/ogc/amicus/fields.pdf, consultado por última vez el 23-10-2017)

AMERICAN PSYCHIATRIC ASSOCIATION (2011), *Coble v. Texas: Brief Amicus Curiae* (http://www.apa.org/about/offices/ogc/amicus/coble.pdf, consultado por última vez el 23-10-2017)

ANDRÉS PUEYO, Antonio (2013), "Peligrosidad criminal: análisis crítico de un concepto polisémico", en DEMETRIO CRESPO (dir.), *Neurociencias y Derecho Penal. Nuevas perspectivas en el ámbito de la culpabilidad y tratamiento jurídico-penal de la peligrosidad*, Edisofer – BdF, pp. 483-503.

ANDRÉS PUEYO, Antonio (2013), "Valoració del risc i gestió de la reincidència: la utilitat del RisCanvi en la reinserció", en CID et al (coords.), *De l'execució de penes a la reinserció*, Barcelona: UAB, pp. 67-70.

ANDRÉS-PUEYO, Antonio (2017): "Predicción de la reincidencia penitenciaria en Cataluña por medio del RisCanvi", en ORTS BERENGUER / ALONSO RIMO / ROIG TORRES (eds.), *Peligrosidad criminal y Estado de Derecho*, Tirant lo Blanch, Valencia, pp. 371-388.

ANDRÉS PUEYO, Antonio / ECHEBURÚA, Enrique (2010), "Valoración del riesgo de violencia: instrumentos disponibles e indicadores de aplicación", *Psicothema*, vol. 22(3), pp. 403-409.

ANDRÉS PUEYO, Antonio / REDONDO ILLESCAS, Santiago (2007), "Predicción de la violencia: entre la peligrosidad y la valoración del riesgo de violencia", *Papeles del Psicólogo*, vol. 28(3), pp. 157-173.

ANDRÉS-PUEYO, Antonio / ARBACH-LUCIONI, Karin / REDONDO, Santiago (2018), "The RisCanvi: A New Tool for Assessing Risk for Violence in Prison and Recidivism", en SINGH et al (eds.), *Handbook of Recidivism Risk/Needs Assessment Tools*, John Wiley & Sons, pp. 255-268.

[40] Sobre algunas dificultades que el modelo de la valoración y gestión del riesgo puede suponer para el desistimiento y la reinserción social de los delincuentes cfr. MCNEILL, 2017.

ARBACH-LUCIONI, K., DESMARAIS, S., HURDUCAS, C., CONDEMARIN, C., KIMBERLIE, D., DOYLE, M. SINGH, J. (2015), "La práctica de la evaluación del riesgo de violencia en España", *Revista de la Facultad de Medicina*, 63, pp. 357–366. http://dx.doi.org/10.15446/revfacmed.v63n3.48225

BALLESTEROS REYES, Alicia / GRAÑA GÓMEZ, José Luis / ANDREU RODRÍGUEZ, José Manuel (2006), "Valoración actuarial del riesgo de violencia en centros penitenciarios", *Psicopatología clínica, Legal y Forense* (6), pp. 103-117.

BRANDARIZ GARCÍA, José Ángel (2016), *El modelo gerencial-actuarial de penalidad: Eficiencia, riesgo y sistema penal*, Dykinson.

BUSHWAY, Shawn / SMITH, Jeffrey (2007), "Sentencing Using Statistical Treatment Rules: What We Don't Know Can Hurt Us", *Journal of Quantitative Criminology* 23, pp. 377–387.

CAPDEVILA CAPDEVILA, M. (Coord,), BLANCH SERENTILL, M., FERRER PUIG, M., ANDRÉS PUEYO, A., FRAMIS FERRER, B., COMAS LÓPEZ, N., … MORA ENCINAS, J. (2015), *Tasa de reincidencia penitenciaria 2014*. Centro de Estudios Jurídicos y Formación Especializada, Generalitat de Catalunya.

CERVELLÓ DONDERIS, Vicenta (2014), "Peligrosidad criminal y pronóstico de comportamiento futuro en la suspensión de la ejecución de la pena", *La ley penal* nº 106, enero-febrero 2014.

CHRISTIN, Angèle / ROSENBLAT, Alex / BOYD, Danah (2015), "Courts and Predictive Algorithms (10.27.2015)" (https://datasociety.net/output/data-civil-rights-courts-and-predictive-algorithms/, consultada por última vez el 10.1.2018).

COOKE, Davis J. / MICHIE, Christine (2010), "Limitations of diagnostic precision and predictive utility in the individual case: a challenge for forensic practice", *Law and Human Behaviour*, 34, 259-274. doi: 10.1007/s10979-009-9176-x.

COOKE, David J. / MICHIE, Christine (2011), "Violence risk assessment. Challenging the illusion of certainty" en B. MCSHERRY & P. KEIZER (Eds.), *Dangerous people. Policy, prediction and practice*, New York-London: Routledge, pp. 147-161.

COOKE, David J. / MICHIE, Christine (2013), "Violence risk assessment: From prediction to understanding, from what? to why?", en C. LOGAN & E. JOHNSTONE (eds), *Managing clinical risk: A guide to effective practice*. Routledge, pp. 3-26.

DAUNIS RODRÍGUEZ, Alberto (2016), "Criterios para la valoración de la peligrosidad y el riesgo en el ámbito penitenciario,", *Cuadernos de política criminal*, Nº 120, pp. 239-279.

DESMARAIS, Sarah L. / JOHNSON, Kiersten / SINGH, Jay P. (2018), "Performance of recidivism Risk Assessment Instruments in US Correctional Settings", en SINGH et al (editores), *Handbook of Recidivism Risk/Needs Assessment Tools*, Wiley Blackwell, pp. 3-29.

DIAMOND, Bernard L. (1974), "The psychiatric prediction of dangerousness", *University of Pennsylvania Law Review*, Vol. 123, pp. 439-452.

DOUGLAS, T. / PUGH, J. / SINGH, I. / SAVULESCU, J. / FAZEL, S. (2017), "Risk assessment tools in criminal justice and forensic psychiatry: The need for better data", *European Psychiatry* 42, pp. 134–137.

FAZEL / SINGH / DOLL / GRANN (2012), "Use of risk assessment instruments to predict violence

and antisocial behaviour in 73 samples involving 24827 people: systematic review and meta-analysis", *British Medical Journal*, 2012; 345:e4692

GIGERENZER, Gerd / GAISSMAIER, Wolfgang / KURZ-MILCKE, Elke / SCHWARTZ, Lisa M. / WOLOSHIN, Steven (2008), "Helping Doctors and Patients Make Sense of Health Statistics", *Psychological Science in the Public Interest*, 8 (2), pp. 53-96.

HANNAH-MOFATT, Kelly (2015), "The Uncertainties of Risk Assessment: Partiality, Transparency, and Just Decisions", *Federal Sentencing Reporter* 27 (4), pp. 244-247. DOI: 10.1525/fsr.2015.27.4.244.

HANNAH-MOFFAt, Kelly / MAURUTTO, Paula / TURNBULL, Sarah (2009), "Negotiated Risk: Actuarial Illusions and Discretion in Probation", *Canadian Journal of Law and Society* 24 (3), pp. 391-409.

HANSON, R. Karl (2017), "Assessing the Calibration of Actuarial Risk Scales. A Primer on the E/O Index", *Criminal Justice and Behavior*, 44 (1), 26–39. DOI: 10.1177/0093854816683956

HARCOURT, Bernard E. (2007), *Against prediction : profiling, policing, and punishing in an actuarial age*, Chicago [etc.] : University of Chicago Press.

HARRIS, G. T., LOWENKAMP, C. T. & HILTON, N. Z. (2015), "Evidence for Risk Estimate Precision: Implications for Individual Risk Communication", *Behavioral Sciences and the Law*, 33, pp. 111-127. doi: 10.1002/bsl.2158.

HART, Stephen D. / COOKE, David J. (2013), "Another look at the (im-)precision of individual risk estimates made using actuarial risk assessment instruments", *Behavioral Sciences and the Law*, 31, pp. 81-102. doi: 10.1002/bsl.2049.

HEILBRUN, Kirk (2009), *Evaluation for risk of violence in adults*, New York : Oxford University Press.

HELMUS, Leslie / HANSON, R. Karl / THORNTON, David / BABCHISHIN, Kelly M. / HARRIS, Andrew J. R. (2012), "Absolute recidivism rates predicted by Static-99r and Static-2002r sex offender risk assessment tools vary across samples. A meta-analysis", *Criminal Justice and Behavior* 39 (9), pp. 1148-1171, doi: 10.1177/0093854812443648

IOANNIDIS, John P.A. (2012), "Why Science Is Not Necessarily Self-Correcting", *Perspectives on Psychological Science* 7(6), pp. 645–654.

KRAUSS, Daniel A. / SCURICH, Nicholas (2013), "Risk Assessment in the Law: Legal Admissibility, Scientific Validity, and Some Disparities between Research and Practice", *Behavioral Sciences and the Law*, 31, pp. 215–229. DOI: 10.1002/bsl.2065

LOINAZ, Ismael (2017), *Manual de evaluación del riesgo de violencia. Metodología y ámbitos de aplicación*, Pirámide, Madrid.

LÓPEZ-OSSORIO, Juan José / GONZÁLEZ-ÁLVAREZ, José Luis / ANDRÉS-PUEYO, Antonio (2016), "Eficacia predictiva de la valoración policial del riesgo de la violencia de género", *Psychosocial Intervention* 25, pp. 1-7.

MARTÍNEZ GARAY, Lucía (2014), "La incertidumbre de los pronósticos de peligrosidad: consecuencias para la dogmática de las medidas de seguridad" (epígrafe 5 redactado en coautoría con Francisco MONTES SUAY), *InDret. Revista para el análisis del Derecho*, nº 2/2014, pp. 1 – 77.

MARTÍNEZ GARAY, Lucía (2014), "Errores conceptuales en la estimación de riesgo de reincidencia. La importancia de diferenciar sensibilidad y valor predictivo, y estimaciones de riesgo absolutas y relativas", *Revista Española de Investigación Criminológica*, 14, pp. 1-31.

MCNEILL, Fergus (2017), "Las Consecuencias Colaterales del Riesgo", *InDret. Revista para el análisis Del Derecho* 1/2017, pp. 1-19.

MONAHAN, John (1981), *Predicting violent behaviour. An assessment of clinical techniques*, Sage Publications, London.

MOSSMAN, Douglas (1994), "Assessing predictions of violence: Being accurate about accuracy", *Journal of Consulting and Clinical Psychology* 62 (4), pp. 783-792.

MOSSMAN, Douglas (2006), "Another look at interpreting risk categories", *Sexual Abuse: A Journal of Research and Treatment*, 18(1), pp. 41-63. doi: 10.1177/107906320601800104.

MOSSMAN, Douglas (2013), "Evaluating Risk Assessments Using Receiver Operating Characteristic Analysis: Rationale, Advantages, Insights, and Limitations", *Behavioral Sciences and the Law* (31), pp. 23-39 DOI: 10.1002/bsl.2050

MOSSMAN, Douglas (2015), "From Group Data to Useful Probabilities: The Relevance of Actuarial Risk Assessment in Individual Instances", *The Journal of the American Academy of Psychiatry and the Law*, 43(1), pp. 93-102.

MÜLLER, Jürgen L. / STOLPMANN, Georg (2015), „Legalbewährung nach rechstkräftiger Ablehnung einer nachträglichen Anordnung der Unterbringung in der Sicherungsverwahrung", *Monatsschrift für Kriminologie und Strafrechtsreform*, 98(1), pp. 35-47.

MUÑOZ VICENTE, José Manuel / LÓPEZ-OSSORIO, Juan José (2016), "Valoración psicológica del riesgo de violencia: alcance y limitaciones para su uso en el contexto forense", *Anuario de Psicología Jurídica* (26), pp. 130–140.

NGUYEN VO, Thuy / ANDRÉS PUEYO, Antonio (2016), *Validez predictiva del SVR-20 y la Static-99 en una muestra de agresores sexuales en Cataluña*, Centre d'Estudis Jurídics i Formació especialitzada, Generalitat de Catalunya.

NGUYEN, Thuy / ARBACH-LUCIONI, Karin / ANDRÉS-PUEYO, Antonio (2011), "Factores de riesgo de la reincidencia violenta en población penitenciaria", *Revista de Derecho Penal y Criminología*, 3ª Época, nº 6, pp. 273-294.

NEDOPIL, Norbert (2002), „Prognosebegutachtungen bei zeitlich begrentzten Freiheitsstrafen – Eine sinnvolle Lösung für problematische Fragestellungen?", *Neue Zeitschrift für Strafrecht*, 2002 (Heft 7), pp. 344-349.

PÉREZ RAMÍREZ, Meritxell / REDONDO ILLESCAS, Santiago / MARTÍNEZ GARCÍA, Marian / GARCÍA FORERO, Carlos / ANDRÉS PUEYO, Antonio (2008), "Predicción de riesgo de reincidencia en agresores sexuales", *Psicothema* (20-2), pp. 205-210.

PITA FERNÁNDEZ, S. / PÉRTEGAS DÍAZ, S. (2003), "Pruebas diagnósticas: Sensibilidad y especificidad" Unidad de Epidemiología Clínica y Bioestadística. Complexo Hospitalario Universitario de A Coruña (España). *Cadernos de Atención Primaria* 10, pp. 120-124 .

RECENT CASES (2017), "Criminal Law – Sentencing Guidelines – State v. Loomis. Wisconsin Supreme Court Requires Warning Before Use of Algorithmic Risk Assessments in Sentencing", *Harvard Law Review* 130, pp. 1530-1537.

ROSSEGGER, Astrid / GERTH, Juliane / SEEWALD, Katharina / URBANIOK, Frank / SINGH, Jay P. / ENDRASS, Jérôme (2013), «Current obstacles in replicating risk assessment findings: a systematic review of commonly used actuarial instruments», *Behavioural Sciences and the Law*, vol. 31, pp. 154 – 164.

ROSSEGGER, Astrid / ENDRASS, Jérôme / GERTH, Juliane / SINGH, Jay P. (2014), "Replicating the Violence Risk Appraisal Guide: A Total Forensic Cohort Study", *PLoS ONE* 9 (3): e91845. doi:10.1371/journal.pone.0091845

SCURICH, Nicholas (2016), "Structured Risk Assessment and Legal Decision-Making", en M.K. MILLER, B.H. BORNSTEIN (eds.), *Advances in Psychology and Law*, Springer International Publishing Switzerland, pp. 159-183.

SHEPHERD, Stephan M. / SULLIVAN, Danny (2017), "Covert and Implicit Influences on the Interpretation of Violence Risk Instruments", *Psychiatry, Psychology and Law*, 24:2, pp. 292-301.

SINGH, Jay P. (2013), "Predictive validity performance indicators in violence risk assessment: a methodological primer", *Behavioural Sciences and the Law* (31), pp. 8-22.

SINGH, Jay P. / GRANN, Martin / FAZEL, Seena (2013), "Authorship Bias in Violence Risk Assessment? A Systematic Review and Meta-Analysis", *PLoS ONE* 8 (9): e72484. doi:10.1371/journal.pone.0072484

SINGH, Jay P. / DESMARAIS, Sarah L. / VAN DORN, Richard A. (2013), "Measurement of Predictive Validity in Violence Risk Assessment Studies: A Second-Order Systematic Review", *Behavioral Sciences and the Law* (31), pp. 55–73.

SINGH, Jay P. / PETRILA, John (2013), "Measuring and Interpreting the Predictive Validity of Violence Risk Assessments: An Overview of the Special Issue", *Behavioral Sciences and the Law* (31), pp. 1–7.

SIONTIS, George C.M. / TZOULAKI, Ioanna / CASTALDI, Peter J. / IOANNIDIS, John P.A. (2015), "External validation of new risk prediction models is infrequent and reveals worse prognostic discrimination", *Journal of Clinical Epidemiology* 68, pp. 25-34.

SJÖSTEDT, Gabrielle / GRANN, Martin (2002): " Risk Assessment: What is Being Predicted by Actuarial Prediction Instruments?, *International Journal of Forensic Mental Health*, 1:2, pp. 179-183, DOI: 10.1080/14999013.2002.10471172

SKEEM, Jennifer (2013), "Risk Technology in Sentencing: Testing the Promises and Perils (Commentary on Hannah-Moffat, 2011)", *Justice Quarterly*, 30:2, pp. 297-303.

STEADMAN, Henry J. (2000), "From Dangerousness to Risk Assessment of Community Violence: taking stock at the turn of the century", *Journal of the American Academy of Psychiatry and the Law* (28), pp. 265-71.

STEADMAN, Henry J. / COCOZZA, Joseph (1978), "Psychiatry, dangerousness and the repetitively

violent offender", *The Journal of Criminal Law & Criminology*, vol. 69(2), pp. 226-231.

SZMUKLER, G. / EVERITT, B. / LEESE, M. (2012), "Risk assessment and receiver operating characteristic curves", *Psychological Medicine*, 42, pp. 895–898.

VIVES ANTÓN, Tomás S. (1974), "Métodos de determinación de la peligrosidad", en VVAA, *Peligrosidad social y medidas de seguridad (La Ley de peligrosidad y rehabilitación social de 4 de agosto de 1970)*, Universidad de Valencia, pp. 389-417.

YOUNG, Jock (2015), *La imaginación criminológica*, Marcial Pons, Madrid.