

JUDGING THE ALGORITHM

\\

*A case study on the risk
assessment tool for gender-based
violence implemented in the Basque
country*

Nov 17, 2022

Ana Valdivia (*Oxford Internet Institute*)
Cari Hyde-Vaamonde (*King's College London*)
Julián García Marcos (*Magistrate in Spanish Court*)



HOW IT STARTED...

- Event on AI and judicial system in February 2021.
- Hype about how AI could reduce the workload of judges.
- Lack of concern about the risks, limits and pitfalls.

Granada, referencia en Derecho y la Inteligencia Artificial

Publicado el lunes, 15 febrero 2021



ROADMAP

1. INTRODUCTION

2. THE EPV-R

**3. TECHNICAL, LEGAL AND
USER PERSPECTIVE**

4. CONCLUSIONS

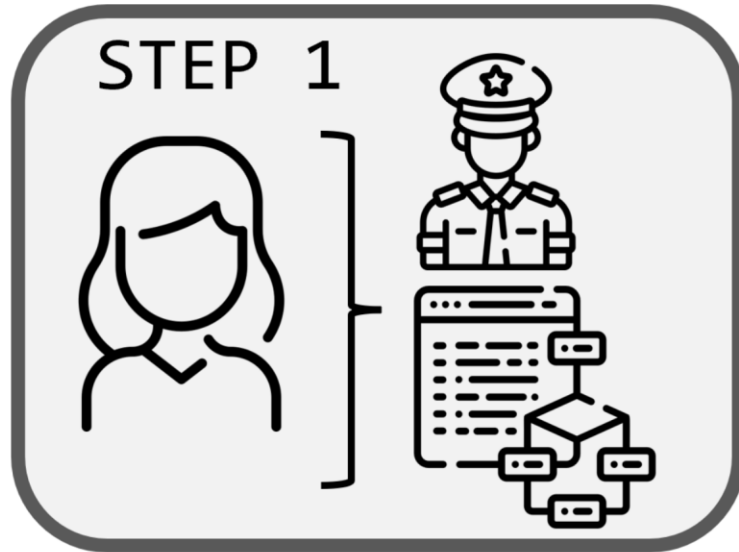
INTRODUCTION

The case of M.



THE EPV-R

A RISK ASSESSMENT TOOL FOR GENDER-BASED VIOLENCE



1. Women report gender violence to the police. The police analyses the case and assesses the risk of violence using the EPV's algorithm which consists of 20 psychometric items.



2. The report with the output given by the EPV (severe or non-severe violence) is sent to the courtroom.



3. The judge reads the report and together with the EPV's scores takes a decision based on the assessed risk of violence.

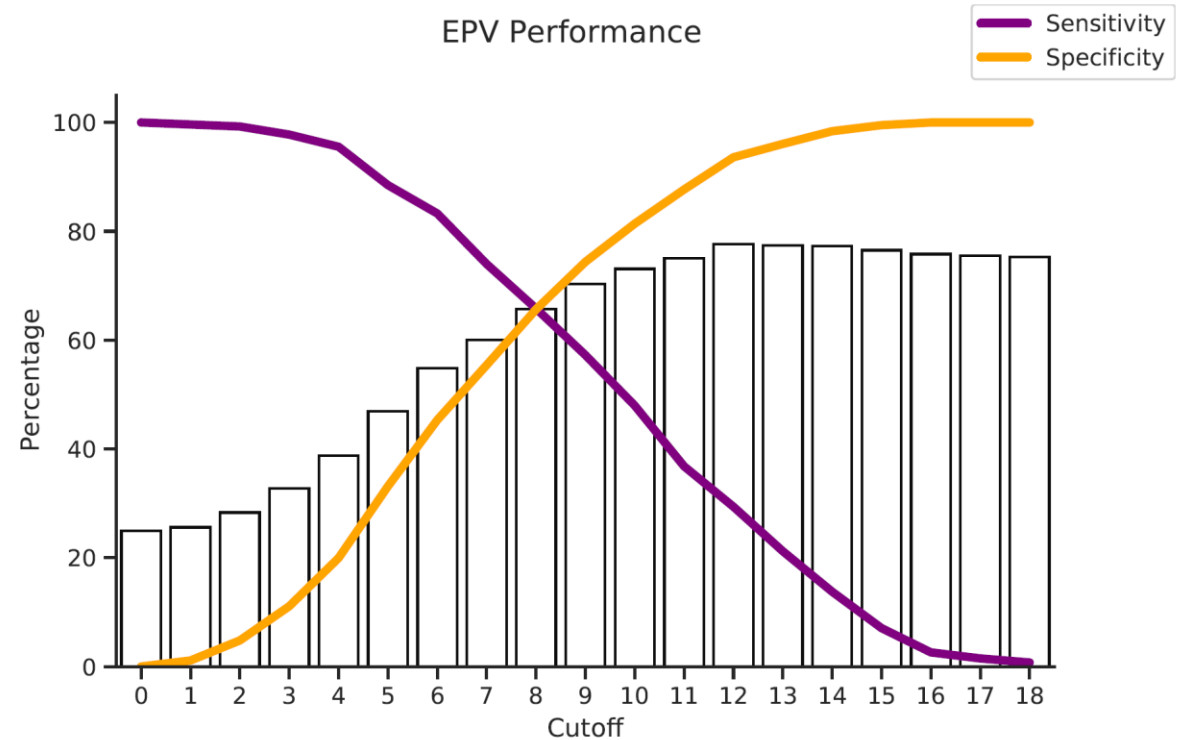
THE EPV-R

- The questionnaire contains some questions such as:
 - Is the male batterer or victim an immigrant?
 - Severe threats or threatening to kill in the past month?
 - Clear intention of causing severe or very severe injuries?
 - Very intense jealousy
 - Victim's perception of danger.

Name:	File:	
Date:	Assessor:	
I. Personal data		Assessment (0 or 1)
1. Male batterer or victim is an immigrant		
II. Couple relationship status		Assessment (0 or 1)
2. Recently separated or in the process of separation		
3. Recent harassment of victim or breaking the restraining orders		
III. Type of violence		Assessment (0 or 1)
4. Existence of physical violence that can cause injuries		
5. Physical violence in the presence of the children or other relatives		
6. Increase in the frequency and severity of the violent incidents in the past month		
7. Severe threats or threatening to kill in the past month		
8. Threatening with dangerous objects or with weapons of any kind		
9. Clear intention of causing severe or very severe injuries		
10. Sexual aggressions in the couple relationship		
IV. Male batterer's profile		Assessment (0 or 1)
11. Very intense jealousy or controlling behaviors toward partner		
12. History of violent behaviors with previous partner		
13. History of violent behaviors with other people (friends, work mates, etc.)		
14. Abuse of alcohol and/or drugs		
15. History of mental illness and dropping out of psychiatric or psychological treatments		
16. Cruel, disparaging behaviors directed at the victim and lack of remorse		
17. Justification of violent behavior due to aggressor's own state (alcohol, drugs, stress) or to victim's provocation		
V. Victim's vulnerability		Assessment (0 or 1)
18. Victim's perception of danger of death in the past month		
19. Attempts to drop charges or going back on the decision to leave or report the aggressor to the police		
20. Victim's vulnerability because of illness, solitude, or dependence		
Severe violence risk assessment		
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Low (0-4)	Moderate (5-9)	High (10-20)

TECHNICAL, LEGAL AND USER PERSPECTIVE

- From a technical perspective:
 - The efficacy of the risk assessment tool was assessed by analysing the **trade-off** between **sensitivity** (true positive rate: *TP rate*) and **specificity** (true negative rate: *TN rate*):
 - “Thus, for example, a total score of 10, considered high risk, includes 48% of the severe aggressors, which means that one half obtain lower scores, and only 18% of the less severe aggressors obtain this score (false positive (Echeburúa et al., 2009, p. 932).”
- Yet in the context of gender-based violence, **FP and FN do not play the same role** which implies that ‘reasonable equilibrium’ between both error rates might not be desirable.



Cut-off between sensitivity or TPR (purple) and specificity or TNR (orange) for each score of the EPV tool. Accuracy (bars) is higher when specificity is also higher given that the dataset is unbalanced (269 severe violence cases (positives) vs. 812 non-severe cases (negatives)). Source: Echeburúa et al. (2009).

TECHNICAL, LEGAL AND USER PERSPECTIVE

From a technical perspective

	Predicted	
	Predicted severe	Predicted non-severe
	Severe	non-severe
Actual	129 (TP)	140 (FN)
	151 (FP)	661 (TN)

Table 1: Confusion matrix of the EPV (cutoff score = 10). The number of FN (140) is higher than the number of TP (129), which implies that the tool is more likely to classify severe cases as non-severe when the obtained punctuation is 10. Source: Echeburúa *et al.* (2009).

TECHNICAL, LEGAL AND USER PERSPECTIVE

From a technical perspective

		Predicted	
		Predicted severe	Predicted non-severe
Actual	Severe	129 (TP)	140 (FN)
	non-severe	151 (FP)	661 (TN)

Table 1: Confusion matrix of the EPV (cutoff score = 10). The number of FN (140) is higher than the number of TP (129), which implies that the tool is more likely to classify severe cases as non-severe when the obtained punctuation is 10. Source: Echeburúa *et al.* (2009).

TECHNICAL, LEGAL AND USER PERSPECTIVE

- From a technical perspective:
 - Opaque implementation
 - Paradox of efficiency
 - Feedback loop



TECHNICAL, LEGAL AND USER PERSPECTIVE

From a user perspective

Police report

Folio N.º: 26

COMUNICACION DE VALORACION DE RIESGOS SOBRE VICTIMA DE VIOLENCIA DE GENERO (VG)

REFERENCIA: [REDACTED]

En [REDACTED] COMISARIA, a las 01:08 horas del día 18 de OCTUBRE de 2020, las o los etzainas con n.º profesional [REDACTED] y [REDACTED], que actúan como Instructor/a y Secretario/a respectivamente de la presente, para comunicar la VALORACIÓN DE RIESGOS de [REDACTED], informan de lo siguiente:

La Ertzaintza, a fin de realizar la valoración de riesgo de violencia grave en las relaciones de pareja cuando el agresor es un varón y la víctima una mujer, se ha dotado de la herramienta denominada "Escala de Predicción del Riesgo de violencia grave contra la pareja (EPV-R)", confeccionada por el Catedrático de Psicología Clínica de la Facultad de psicología de la UPV-EHU, Enrique Echiburua, con la colaboración de la Consejería de Seguridad del Gobierno Vasco (Ertzaintza).


Esta escala recoge un total de 20 ítems o indicadores de riesgo que hay que identificar realizando una investigación sobre los hechos concurrentes en cada caso. A fin de disponer de la mayor fiabilidad, la valoración de riesgo se realiza siempre que sea posible por más de una o un agente de la Ertzaintza.

El objeto de realizar una valoración de la situación de riesgo en la que se encuentre la víctima es articular los medios de protección personal adecuados a cada caso y momento, a fin de prevenir nuevas agresiones y protegerla adecuadamente.

Aplicación de la Escala EPV-R:

Los ítems que la componen se han seleccionado a partir de su capacidad discriminante de una violencia grave y de su coherencia psicológica con el conjunto. Se agrupan en cinco apartados: datos personales, situación de la relación de

Fdo.: Instructora/Instructor Fdo.: Secretaria/Secretario


EUSKO JAURLARITZA
GOBIERNO VASCO
SEGURTASUNA
Agencia
Departamento de Seguridad
Asesores
Comisión de Género

Folio N.º: 27

COMUNICACION DE VALORACION DE RIESGOS SOBRE VICTIMA DE VIOLENCIA DE GENERO (VG)

pareja, tipo de violencia, perfil del agresor y vulnerabilidad de la víctima.

La valoración de riesgo es un proceso continuo que se actualiza a medida que se tiene conocimiento de información que permita identificar nuevos indicadores de riesgo o descartar la concurrencia de algunos ítems anteriormente tenidos en cuenta. Por este motivo, la determinación del nivel de riesgo es la fotografía de la situación de la violencia de pareja en un momento dado.

Fruto de la aplicación de la escala, resultan cuatro posibles niveles de riesgo:

- Nivel de riesgo Básico: baja probabilidad de que puedan darse nuevos incidentes violentos.
- Nivel de riesgo Moderado: alguna probabilidad de que puedan darse nuevos incidentes violentos.
- Nivel de riesgo Alto: alta probabilidad de que puedan darse nuevos incidentes violentos.
- Nivel de riesgo Especial: muy alta probabilidad de que puedan darse nuevos incidentes violentos.

En función del nivel de riesgo resultante, la Ertzaintza aplica diferentes medidas policiales de protección que, abarca desde la oferta de formación en medidas de autoprotección, entrevistas y visitas aleatorias, comprobaciones aleatorias mediante teléfono, oferta de traslado al juzgado para la primera comparecencia, etc., hasta la oferta de protección permanente durante las 24 horas del día o de tramitación ante el órgano judicial de la solicitud de su seguimiento por medios telemáticos de control, según el nivel de riesgo que concurra.

Folio N.º: 28

COMUNICACION DE VALORACION DE RIESGOS SOBRE VICTIMA DE VIOLENCIA DE GENERO (VG)

Una vez aplicada la escala EPV-R sobre el caso de la víctima de violencia de género [REDACTED], la Valoración de Riesgos es la siguiente:

- Expediente n.º: [REDACTED]

La valoración de riesgo realizada a fecha 18 de OCTUBRE del año 2020, arroja un resultado de BASICO.

Observaciones:

Todo lo cual se pone en su conocimiento a los efectos oportunos.

CONCLUSION

- This paper has been prepared by **authors of diverse disciplines (law and computer science)** to highlight a practice that has widely gone unreported.
- It is perfectly possible that **the judge in question will assume strong probative value** / weight on to the EPV-R assessment, such in the M's case.
- We propose an additional perspective that could potentially avoid unintended consequences on the use of these tools and overcome risks, harms and limitations that have been identified through this paper: **data feminism, design justice**.



THANKS !



Cari Hyde-Vaamonde (*King's College London*)

Julián García Marcos (*Magistrate in Spanish Court*)



APPENDIX



TECHNICAL, LEGAL AND USER PERSPECTIVE

From a user perspective:

- Suddenly, **I realised that something didn't work with the tool**
- In some cases, from my subjective point of view, there was an **obvious risk for the victim** but the tool did not correctly reflect the risk assessment.
- In one of the cases **the victim and her daughter were spied on**, chased home by the aggressor, disturbed late at night, watched from the window...
- The victim had previously been assaulted and her daughter suffered from psychological sequelae derived from the events
- I thought It was a case where the victims needed protection...
- But...
- The **risk assessment was BASIC (which means, basically, NO RISK)**
- And that was just one of the cases I was completely puzzled about...