Journal of Experimental Criminology
https://doi.org/10.1007/s11292-023-09596-4

Effect of conditional release on violent and general recidivism: A causal inference study

Marzieh Karimi-Haghighi[1] · Carlos Castillo[2] · Songül Tolan[3] · Kristian Lum[4]

Accepted: 6 November 2023

Abstract

Objectives To study the effect of Conditional Release (C.R.) on recidivism. To compare this effect along different recidivism risk levels, to evaluate whether risk-assessment-based policies that prioritize people in lower risk categories for release maximally reduce recidivism.

Methods We use a dataset of 22,726 incarcerated persons released from 87 prison centers in Spain. We apply multiple causal inference methods including Propensity Score Matching (PSM), Inverse Propensity score Weighting (IPW), and Augmented Inverse Propensity Weighting (AIPW) to determine Average Treatment Effect (ATE) of C.R. on recidivism.

Results Granting C.R. significantly reduces violent and general recidivism risks.

Conclusions The results suggest that C.R. can promote a safe and supervised return to the community while protecting public safety. ATEs obtained through causal inference methods suggest that granting C.R. exclusively to low-risk inmates does not lead to the maximum reduction of recidivism, and hence we propose alternatives to be studied further.

Keywords Causal inference · Conditional release · Criminal justice · Machine learning · Recidivism

Songül Tolan has left the JRC and joined the German federal ministry of employment and social affairs in October 2022. This paper represents the views of the author and is not meant to represent the position of any of her affiliated institutions.

✉ Marzieh Karimi-Haghighi
m.karimihaghighi@gmail.com

[1] Ph.D. graduate from Department of Technologies of Information and Communication, Universitat Pompeu Fabra, Barcelona, Catalonia, Spain

[2] ICREA, Barcelona, Catalonia, Spain

[3] JRC, European Commission, Seville, Spain

[4] University of Pennsylvania, Philadelphia, PA, USA

Published online: 23 December 2023

Introduction

Studies on US prison population show that almost 1% of the US adult population is incarcerated which is 5 to 10 times higher than the rates in Western European and other liberal democracies (Travis et al., 2014; Loeffler & Nagin, 2022). Looking at the latest European prison population rate reported in January 2021, about 0.1% of the European population is incarcerated (Aebi et al., 2022). Although incarceration rates in European countries are not nearly as high as in the United States, in all countries the rise of "mass incarceration" over the last half century has caused an increasing attention to assessing the effects on crime rates as well as the social and economic costs (Raphael & Stoll, 2009; Durlauf & Nagin, 2011; Spelman, 2020; Loeffler & Nagin, 2022).

Measures to address mass incarceration have to consider the tradeoff between the social and economic costs of potentially unnecessary incarceration and a potentially increased risk to public safety as more people with a probably higher likelihood of re-offense are released. However, from a mid-to-long term perspective and according to the literature of the past decades, this trade-off is not as strong as it initially seems, especially if we go beyond the simplistic view of its underlying dichotomous decision (release or no release) and consider a variety of alternative rehabilitation focused interventions. In fact, for some offender types and depending on the rehabilitation programme, an extended period of incarceration as opposed to rehabilitation-focused early release programmes may yield a net-positive effect on overall reoffense rates (Kuriakose, 2019; Petrich, 2021). Understanding the effects that such programmes may have on overall recidivism rates can help navigate these considerations.

In this paper, we study the effect of Conditional Release (C.R.) on recidivism. C.R. can, by definition, help reduce the number of people who are incarcerated. However, it is unclear how increased availability of C.R. will impact recidivism and public safety.

Some studies suggest that incarceration or the length of incarceration has a deterrent effect on recidivism (Drago et al., 2009), whereby people refrain from committing crimes for fear of the resulting sanctions (Loeffler & Nagin, 2022). Generally, crime prevention avoids both the costs of crime and the costs of punishment (Marchese di Beccaria, 1819; Becker, 1968; Cotter, 2020). However, incarceration often fails to achieve deterrence from recidivism, and triggers punishment and costs (Loeffler & Nagin, 2022). Incarceration punishment may reduce crime during the incapacitation, when the person is physically separated from free society, however, beyond that it has a chastening impact on the incarcerated person. Punishment may affect future criminality of a person through different mechanisms, some of which such as rehabilitation may reduce future criminal involvement (Bhuller et al., 2020; Gilman & Walker, 2020), whereas others such as social stigma may increase criminal involvement (Link & Phelan, 2001).

Prison conditions and prison experience are very important in the determination of the direction and magnitude of the incarceration effect. The effects are heterogeneous as we can see when comparing the findings of studies on Scandinavia-based prisons, which are more orientated towards rehabilitation (Benko, 2018; Bhuller et al., 2020; Lappi-Seppälä, 2012; Hjalmarsson and Lindquist, 2020), with studies on US-based prisons, which are more orientated towards punishment (Cullen et al., 2000; Beckett & Sasson, 2003; Weaver, 2007; Cullen & Gilbert, 2012). Other studies have shown that prisons do not reduce recidivism, but could have a criminogenic effect that increases the risk of recidivism (Chen & Shapiro, 2007; Mears & Bales, 2009; Jolliffe & Hedderman, 2015). In addition, the effectiveness of intermediate sanctions (penalties that exist between prison and probation) is mediated only through the provision of appropriate cognitive-behavioral treatments (Gendreaue et al., 2000; Cullen et al., 2011). Prison incarceration with a focus on rehabilitation can be largely crime preventive. Rehabilitation programs such as employment training services during the sentence, which is common in Scandinavian countries, correctional substance abuse treatments, and generally high quality prisons can decrease future criminal involvement, including recidivism (Bhuller et al., 2020; Sondhi et al., 2020; Tobón, 2020; Andrews & Bonta, 2010). In addition to rehabilitation during the sentence, noncustodial sanctions which are partially or totally alternative to prison such as community sentences, electronic monitoring and semi-liberty [1] (Cid, 2009; Henneguelle et al., 2016; Yukhnenko, 2019; Statistics, Northern Ireland and Research Agency, 2019; Monnery et al., 2020; Williams & Weatherburn, 2022; Andersen & Telle, 2022), as well post-prison interventions such as employment, housing, and social reintegration support can also help reduce recidivism risk (Western, 2018; Kirk, 2020; Harding & Harris, 2020).

In hopes of reducing incarceration rates without substantially increasing crime, decision makers commonly use violence risk assessment tools when making noncustodial decisions such as conditional release. The main purpose of such tools is to prevent criminal violence and its consequences, but they also help prison management identify people with a greater risk of recidivism and allocate rehabilitation efforts accordingly. Ideally, accurate risk assessment may help place low-risk defendants into alternative programs to prison (Andrés-Pueyo et al., 2018). Accordingly, cases with an assigned low risk level have higher chances of receiving conditional release compared to the cases who are assessed as high risk. In this regard, these tools have to diagnose correctly and target the proper person to be released early, which in turn may reduce the recidivism rate (Austin, 2006).

The rest of this paper is organized as follows. After outlining previous studies in "Related work" section, we summarize "*Our Contribution*". Risk assessment tools and conditional release in the legislation of the country under study are described in the "Risk assessment and conditional release" section. In the "Dataset" section, the data used in this study is described and analysed with respect to recidivism and conditional release variables. The procedure and methods are presented in the "Methodology" section. Results are given in the "Results" section, and they are discussed, and the paper concluded in the "Discussion and conclusion" section. Finally, limitations of the study are outlined in the "Limitations" section and prospective research directions are introduced in the "Futurework" section.

Related work

Criminologists have long studied the effects of punishment or treatment on recidivism, where they have drawn from various different methods. In this section, we focus on

[1] Semi-liberty is an interesting mix of custody and liberty. Offenders under semi-liberty are free during weekdays, but have to stay at night and during weekends in dedicated correctional facilities, often called "halfway houses" or "semi-liberty centers" (Monnery et al., 2020).

causal inference methods such as Instrumental Variables (IV), Regression Discontinuity (RD), and other statistical methods (VanderWeele, 2015).

Instrumental Variables (IV) methods are used to approximate the methodological conditions of randomized control trials, by conditioning on a variable that is exogenous to the treatment status and filtering out selection bias that normally contaminates the estimated impact of treatment on the outcome of interest (Angrist et al., 1996). The most commonly used instrumental variable in the studies dealing with the effects of custodial/noncustodial sanctions is the random assignment of cases to judges (Loeffler & Nagin, 2022). Some research suggests that extraneous factors such as hunger or a bad mood can influence judge decisions (Danziger et al., 2011; Cho et al., 2016; Eren & Mocan, 2018; Heyes & Saberian, 2019), despite a societal agreement that such factors should not influence judicial decisions. In particular, it has been observed that the proportion of favorable rulings decreases with serial order within a session but goes back to the initial level after a session break that includes eating a meal (Danziger et al., 2011). Another study revisited this finding and claimed that their analyses do not provide conclusive evidence for the hypothesis that mood influences legal rulings (Glöckner, 2016). In fact, the observed downward trend could be explained by selective dropout of favorable cases due to rational time management, censoring of data and autocorrelation.

Regression Discontinuity (RD) is another strategy that addresses selection bias in estimates of treatment effects on the outcome. RD was first developed in education research (Thistlethwaite & Campbell, 1960). In RD research designs, the assignment of units to treatments is performed based on a score-based system. In these scoring systems, when the assignment is discontinuous and deterministic at some threshold value along the score, any sudden changes in the outcome of interest can be causally attributed to the effects of treatment (Loeffler & Nagin, 2022).

Other statistical methods include methodologies encompassing regression models and inverse probability weighting that have been developed for the estimation of a treatment effect on an outcome. These include methods such as propensity score matching (PSM) (Rosenbaum & Rubin, 1983), regression adjustment (RA) (Vansteelandt & Daniel, 2014), inverse probability weighting (IPW) (Bray et al., 2019), and augmented inverse probability weighting (AIPW) (Glynn & Quinn, 2010).

On Table 1, we summarize several causal inference studies on the effects of custodial and noncustodial sanctions on recidivism. We explain them in the following sections in two categories of studies on incarceration effects on recidivism and alternatives to prison effects on recidivism.

Effects of full incarceration and prison time on recidivism

There are many studies on the effect of incarceration on recidivism (Loeffler & Nagin, 2022). The most used method in these studies is an Instrumental Variables (IV) approach, which is used to estimate the causal impact of incarceration on recidivism by controlling for an exogenous variation in the assignment of cases (Green & Winik, 2010; Loeffler, 2013; Mueller-Smith, 2015; Gupta et al., 2016; Harding et al., 2017; Bhuller et al., 2020). In a study of the District of Columbia's Superior Court,

Table 1 Causal inference studies on the effects of custodial and noncustodial sanctions on recidivism

Study	Num. of observations & location	Follow-up after release	Method	Sanction	Cause	Effect on recidivism
(Chen & Shapiro, 2007)	1,205 USA	1-3 years	RD	Custodial	Prison conditions (security levels)	Criminogenic
(Mears & Bales, 2009)	58,752 Florida, USA	3 years	PSM	Custodial	Supermax incarceration	Criminogenic
(Marie, 2009)	260,000 ENG & WLS, UK	1-2 years	RD	Noncustodial	Early release (EM)	Preventative
(Green & Winik, 2010)	1,003 Washington D.C., USA	4 years	IV	Custodial	Prison time	No effect
				Noncustodial	Probation time	No effect
(Jolliffe & Hedderman, 2015)	5,500 ENG & WLS, UK	1 year	PSM	Custodial	Incarceration	Criminogenic
(Henneguelle et al., 2016)	2,827 France	5 years	IV	Noncustodial	Electronic monitoring	Preventative
(Mitchell et al., 2017)	96,254, Florida, USA	3 years	RD	Custodial	Incarceration	No effect
(Rhodes et al., 2018)	304,000 USA	3 years	IV RD	Noncustodial	Prison length reduction	No effect
(Zapryanova, 2020)	700,000 Georgia, USA	3 years	IV	Custodial	Prison time	No effect
				Noncustodial	Probation time	No effect
(Bhuller et al., 2020)	23,373 Norway	5 years	IV	Custodial	Prison time (rehabilitation)	Preventative
(Meier et al., 2020)	804, Israel	1-6 years	IV	Noncustodial	Parole	Preventative
(Monnery et al., 2020)	1,445, France	5 years	IV	Noncustodial	Semi-liberty	Preventative
(Sondhi et al., 2020)	59,150 England, UK	1 year	RA, IPW,	Custodial	RNR Treatment	Preventative Criminogenic
			AIPW & IPWRA		Pharm. Treatment	
(Gilman & Walker, 2020)	1,478 Washington, USA	1 year	IPW	Custodial	Treatment ("Step Up")	Preventative
(Andersen & Telle, 2022)	48,636 Norway	1-3 years	IV	Noncustodial	Electronic monitoring	Preventative

drug-related persons are assigned randomly to different judicial calendars on which judges gave out sentences that varied substantially in terms of prison and probation time (Green & Winik, 2010). Their results show that variations in prison and probation time have no noticeable effect on recidivism rates. Also, in another study on the cases from the state of Georgia, the causal effect of prison and parole time on recidivism is estimated by relying on two instrumental variables (Zapryanova, 2020). The results are consistent, and show that time in parole has no significant effect on recidivism and time in prison has a negative effect of 1.04 percentage points only if a person recidivates while on parole, which seems to have no effect on overall recidivism. However, the results are different in a different IV study on data from Norway, which shows that time spent in prison with a focus on rehabilitation can be preventive and reduces further criminal behavior (Bhuller et al., 2020). One reason for this could be that the Norwegian prison system is successful in increasing participation in rehabilitation programs such as job training and encouraging employment. In an investigation on data from Texas, using IV estimates, it is found that incarceration generates modest incapacitation effects and sizable social costs to society (Mueller-Smith, 2015).

In addition to judge IV studies on imprisonment effects, Regression Discontinuity (RD) is applied in estimating the effects of incarceration on recidivism. An example of RD research includes estimating the causal effects of prison conditions (custodial security classification levels) on recidivism which suggests that harsher prison conditions lead to more post-release crime (Chen & Shapiro, 2007). Another RD study shows that processing juveniles in the adult system may not uniformly increase offending and may reduce offending in some circumstances (Loeffler & Grunwald, 2015). Also, using the RD approach, it is shown that prison has no effect on the reconviction rates of persons offended due to drug crimes in Florida (Mitchell et al., 2017).

There are also various studies which examine the effect of incarceration on recidivism using statistical methods. In a study on data from the Florida Department of Corrections, the effects of "supermax housing" (a highly restrictive type of incarceration) on 3-year recidivism outcomes is investigated using propensity score matching analysis. They show that supermax incarceration may increase violent recidivism (Mears & Bales, 2009). Another study using propensity score matching on cases from several regions of the UK shows that incarceration slightly increases the chances of reoffending (Jolliffe & Hedderman, 2015).

Effects of alternatives to prison on recidivism

Multiple studies have sought to determine whether programs providing an alternative to prison reduce recidivism, and to measure the extent of this reduction (Vass, 1990; Dynia & Sung, 2000; Cid, 2009).

Most previous research uses IV methods. One study in France, by using IV estimates, shows that converting entire sentences into electronic monitoring (sentence at home under electronic monitoring instead of incarceration) has long-lasting beneficial effects on recidivism rates (Henneguelle et al., 2016). The estimates suggest that this conversion can reduce the probability of reconviction by 6-7 percentage points after five years. Similarly, in another paper, the effect of an electronic monitoring program in Norway is evaluated on the recidivism rate using IV design (Andersen & Telle, 2022). Their results show a reduction of about 15 percent in two-year recidivism rates and approximately 0.3 offences on average in the one-year recidivism frequency. In a study in Israel, it is shown that the parole requests of cases appearing further from the judges' last break are more often denied by the judges (Meier et al., 2020). Exploiting this behavioral pattern in an instrumental variable, the authors estimate that early release decisions driven by exogenous factors reduce the propensity to recidivate.

In another study, semi-liberty is also introduced as a suitable alternative to prison which has a beneficial effect on recidivism (Monnery et al., 2020). In this study, it is found that under treatment exogeneity and conditional independence, semi-liberty results in a reduction of 22% to 31% in cases' recidivism in the five years after release.

There are few Regression Discontinuity (RD) studies on the effect of alternatives to prison on recidivism. A RD study based on data from England and Wales shows that early release on electronic monitoring (EM) can reduce the probability of rearrest by 5 to 7 percent (Marie, 2009). In other research using both IV and RD methods, it is found that average length of prison stay can be reduced by 7.5 months with a small impact on recidivism (Rhodes et al., 2018).

Regarding statistical methods, we find comparatively less research on the effect of alternatives to prison on recidivism. The impact of multiple component treatments on reoffending of incarcerated people with an alcohol use disorder in England is investigated using multiple treatment effect estimators (RA, IPW, and AIPW, and IPWRA) (Sondhi et al., 2020). The results show that a Risk Need Responsivity (RNR) program is the most effective intervention compared to other treatments and represents a lower recidivism rate for treated cases compared to the control group. By contrast, pharmacological treatment results in a statistically significant higher level of reoffending in treated cases relative to the control group. In another study, a treatment program (named "*Step Up*") for youth and families experiencing Child to Family Violence (CFV) is evaluated and its effects on the three outcomes of general recidivism, assault-related recidivism and domestic violence-related recidivism are estimated using an IPW estimator (Gilman & Walker, 2020). The results show that, even when including youth who did not fully complete the program, there is a significantly lower risk of general recidivism for treated cases compared to the control group; and for program completers, the effects are even more pronounced.

The evidence on the effect of incarceration and alternative programs on recidivism is mixed and seems to depend on location, differences in the objectives of the incarceration system, and type of offense. Comparing these effects in Table 1, we can see that generally the literature suggests that custodial sanctions have, at best, no effect or even a criminogenic effect on recidivism, except for rehabilitation-focused incarceration. However, non-custodial alternative programs to prison mostly show preventative effects and to a small extent show no effect on recidivism.

Our Contribution

In this work, we use machine learning (ML) supported causal inference methods to study the effect of Conditional Release (C.R.) on general and violent recidivism risk within 2 to 5 years of a person's release from prison. A comparison of the means of several demographic and penitentiary features between C.R. cases and cases released after completing their entire sentences (something called Definitive Release (D.R.)) and between men and women reveal relevant differences between the former and striking differences between the latter groups. Therefore, we conduct separate analyses for men and women, by also creating two independent models. A diagram for the methodology steps used in this paper is shown on Fig. 1.

We use causal inference methods, which involve several computational steps, i.e., the creation of a predictive model of C.R. propensity and of predictive models of general and violent recidivism risks. The models are obtained using different ML methods. General input features to the models include demographics, penitentiary variables, and risk items and computed risk scores and levels of a risk assessment tool named *RisCanvi*. The best predictions in terms of AUC-ROC (Area Under ROC Curve) are applied to the causal inference methods to calculate the Average Treatment Effect (ATE) of C.R. on general and violent recidivism. The causal methods we use are popular impact evaluation methods in economics and social science in case of non-random assignment of individuals to alternative policies in observational studies (Athey, 2015; Athey & Wager, 2019). The methods include Propensity Score Matching (PSM) (Rosenbaum & Rubin, 1983), Inverse-Propensity score Weighting (IPW) (Bray et al., 2019), and Augmented Inverse-Propensity Weighting (AIPW) (Glynn & Quinn, 2010).

We determine heterogenous effects of C.R. on recidivism by estimating the Conditional Average Treatment Effect (CATE) using Causal Forest and Double Machine Learning (Athey et al., 2019; Nie & Wager, 2021; Chernozhukov et al., 2016). To observe the treatment effect consistency with the risk estimated by the *RisCanvi* risk assessment tool, we compare ATE values separately for cases that have been estimated by the *RisCanvi* tool to have low, medium, or high risk.

In this paper, by recidivism we mean "reincarceration," which means that the released inmates are sentenced and return to prison after the commission of a new crime. There are several limitations to considering reincarceration as recidivism which are mentioned in the "Limitations" section. Although considering recidivism as the

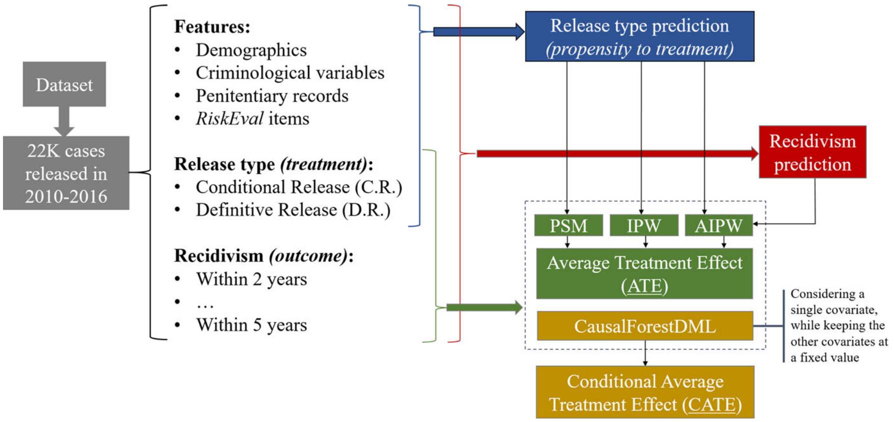
Dataset
22K cases released in 2010-2016
Features:
• Demographics
• Criminological variables
• Penitentiary records
• *RiskEval* items
Release type *(treatment)*:
• Conditional Release (C.R.)
• Definitive Release (D.R.)
Recidivism *(outcome)*:
• Within 2 years
• …
• Within 5 years
Release type prediction *(propensity to treatment)*
Recidivism prediction
PSM
IPW
AIPW
Average Treatment Effect (ATE)
CausalForestDML
Considering a single covariate, while keeping the other covariates at a fixed value
Conditional Average Treatment Effect (CATE)

Fig. 1 Methodology diagram

commission of any new crime incorporates a broad range of outcomes, reincarceration is a recidivism measurement used in Spain for decades with the aim of reducing mass incarceration and is applied due to the following reasons:

- Reincarceration is a clear measurement that provides a straightforward way to track individuals' interactions and involvement with the criminal justice system and makes it easier to collect and analyze data. Reincarceration data such as our data that was captured by the local *Rehabilitation Department* of the Department of Justice of Catalonia in Spain, is typically collected and reported consistently by correctional agencies, making it easier to compare recidivism rates across different jurisdictions and time periods. This standardized data can also help policymakers and researchers make informed decisions.
- Reincarceration data can be used to hold the criminal justice system accountable for decisions related to parole, probation, and reentry programs.
- Tracking reincarceration rates can help policymakers assess the effectiveness of various criminal justice policies and interventions aimed at reducing reincarceration. It allows for evidence-based decision-making regarding changes in sentencing, rehabilitation programs, and reentry services. When comparing recidivism rates across different jurisdictions or populations, considering reincarceration as a measure allows for more equitable comparisons, as it accounts for differences in law enforcement practices, criminal justice policies, and resources.
- Although the optimal result is desistance of any future criminal activity, the prevention of more severe crimes represents a positive impact on public safety that has to be measured (King & Elderbroom, 2014). Hence, in studies (like our research) that are focused on more serious crimes such as violent recidivism, reincarceration is an appropriate metric that can be measured.

Risk assessment and conditional release

Risk assessment instrument

With public safety as one of the fundamental goals of intervention with incarcerated persons, the need for accurate risk assessments has intensified in recent decades. The adoption of structured risk assessment tools has made major progress in the past 40 years. Although these tools are far from perfect, they are more accurate compared to unguided professional judgement used to assess the risk for violence in the 1980s (Hanson, 2005). These tools are used in many socially relevant contexts such as public health, information security, project management, auditing, and criminal justice (Raz, 2001; Alberts & Dorofee, 2003; Allen et al., 2006; Anenberg et al., 2016). In the field of criminal justice, they are applied in different areas such as pre-trial risk assessment, sentencing, probation, and parole (Kehl & Kessler, 2017; Lowenkamp, 2009; Monahan & Skeem, 2016; Miron et al., 2021; Wright et al., 1984; Funk, 1999; Meredith et al., 2007). The risk estimated by these tools can be linked to an intervention consistent with the computed risk. The expectation is that persons assessed with low risk should have lower rates of being sentenced to prison, shorter sentences, higher rates of being paroled and receive lower levels of supervision compared to high-risk cases (Austin, 2006). However, the goals of community protection do not require an exclusive focus on low-risk cases, but can also be effectively promoted when more resources and services are directed towards higher risk cases (Hanson, 2005). Indeed, in the present study, we show, using causal inference methods, that conditional release as an intervention can reduce recidivism, and that this reduction is more pronounced in people deemed higher risk.

RisCanvi was introduced as a multi-level risk assessment protocol for violence prevention in the prison system of Catalonia in Spain in 2009 (Andrés-Pueyo et al., 2018). This protocol is applied multiple times during a person's period in prison; the official recommendation is to do this every six months, or at the discretion of the case manager. *RisCanvi* is not a questionnaire. Instead, each person is interviewed by trained professional case workers. Two versions of the *RisCanvi* protocol were created, an abbreviated one of 10 items for screening (*RisCanvi*-S), and a complete one of 43 items (*RisCanvi*-C). The risk items are listed in Table 14 in Appendix A. Risk items can be categorized into five different categories: Criminal/Penitentiary, Biographical, Family/Social, Clinical, Attitudes/Personality. These items can also be divided into static factors (which cannot be altered, such as "age of starting violent activity") and dynamic factors (which can change, such as "pro-criminal or antisocial attitudes"). In the original *RisCanvi* protocol, risk is determined for each incarcerated person relative to four possible outcomes: self-directed violence, violence in the prison facilities, committing further violent offenses, and breaking prison permits. A fifth risk score was introduced later for general recidivism (Singh, 2018). The outcome of *RisCanvi*-S can be "high-risk" or "low-risk". If the outcome is low-risk for all five criteria, the same *RisCanvi*-S protocol is repeated after six months. Otherwise, in the case of high-risk levels or significant change in a person's situation, the complete version *RisCanvi*-C is applied. The outcome of *RisCanvi*-C can be "high-risk", "medium-risk", or "low-risk". When the risk levels measured by *RisCanvi*-C are medium or high, the next evaluation is again a *RisCanvi*-C; otherwise, *RisCanvi*-S is used.

The *RisCanvi* tool has a performance that is similar to those used internationally for the same purpose. Much of the tool's design is based on the Canadian protocol "Level of Service Inventory-Revised" (Andrews & Bonta, 2000), adapted to the local context, including a thorough review of the input features (Andrés-Pueyo et al., 2018). We remark it is fairly different to COMPAS (Brennan et al., 2009), because it is not a proprietary system, and it has been created by and for a public administration. Its predictive capacity in terms of AUC is about 0.72 which is neither much higher nor much lower than protocols used elsewhere, such as the ones used in US (COMPAS with AUC 0.67 (Brennan et al., 2009), ORAS (Latessa, 2010) with AUC 0.66) and Canada (PCL-R (Hare, 2003) with AUC 0.72). Considering what is reported in the literature and what is deployed in various countries, we do not believe that with current technologies and processes, predictive systems for this task can be substantially more accurate than this.

Conditional release

“Conditional release” (abbreviated C.R. throughout this paper) is similar to “parole" in the USA and is described in detail in the legislation of the country under study. C.R. in Spain is mainly regulated by article 90 of the Criminal Code, last amended in 2015. It occurs when an incarcerated person who meets some requirements is released before completing the full period of the sentence. Cases that are not granted C.R. and are only released at the end of the sentence are described as “definitive release” (abbreviated D.R. throughout this paper). Regarding the cases in our study, each penitentiary center is associated to one court (“Court of Penitentiary Oversight”); each of these courts often has a single judge. In each center, C.R. requests are proposed by a treatment committee to the judge for the persons who are in “third degree” classification[2] and have completed, in most cases, 75% or, in some few cases[3], 67% of the sentencing time. Accordingly, when this committee believes that the person presents a low risk to society if released early, they prepare a request for C.R. and present it to the judge. This request is based on the risk factors of *RisCanvi*, but does not explicitly include the computed *RisCanvi* level, however, in the majority of cases, a high *RisCanvi* risk level makes it unlikely that a request for C.R. will be presented.

Finally, judicial discretion is exercised with respect to C.R. by a Penitentiary Oversight Judge. In deciding whether to grant or deny C.R., the judge must consider elements such as the prisoner’s personality, his/her background, the circumstances of the crime committed, the relevance of the legal rights that could be affected by a repetition of the crime, his/her behavior during the serving of the sentence, and family and social circumstances. The judge can impose conditions including the prohibition to contact victims, the prohibition to contact people who can help in the commission of new crimes (e.g., previous criminal associates), and the obligation to maintain a stable place of residence and to report to a probation officer with a specified regularity. The judge can also decree a mandatory treatment program, which in the case of drug addiction treatment includes submitting to mandatory drug testing.

C.R. can reduce recidivism through two mechanisms: by submitting to supervision and by providing support. Supervision, which in Spain can be considered as similar to what in the US is Intensive Supervision Probation (ISP), includes oversight of the conditions imposed by the judge for C.R., under penalty of having the C.R. revoked. Such conditions are specifically designed to reduce the probability of a new crime being committed. For instance, if alcohol played a role in the crime(s) leading to prison sentencing, being forced to undergo a treatment program for alcoholism and being subject to random testing for alcohol is expected to reduce the likelihood of recidivism. Support is provided directly through probation officers and social workers who are employed by the Department of Justice, and indirectly through intermediary Non-Governmental Organizations (NGOs) that receive subsidies and develop a number of

[2] “Third degree” classification in Spain is a semi-freedom regime in which the person may spend up to 16 hours per day outside of prison; this regime is granted by correctional officers after the person has had good behavior, has received permits to spend brief periods outside of prison, and has returned from those permits.

[3] Persons who participate actively (the law says “continuously”) in treatment programs, or programs for the reparation of the victims.

training and treatment programs. The primary goal of the support received is to increase the human capital (knowledge and skills) and social capital (supportive connections) of the excarcerated person. Ideally, this support builds the capacity to successfully overcome important challenges, such as finding a job, and thus reduces the likelihood of recidivism due to, e.g., being jobless for an extended period of time.

Dataset

The anonymized dataset used in this study has been provided by 87 prison centers of Catalonia in Spain and consists of 26,305 prison releases between 2010 and 2016 which are definitive (72%) or conditional (28%) releases among 22,726 individuals.[4] Our research plan was reviewed by the correctional department in charge of these centers, who approved our anonymized data access request. The anonymized data processing was also approved by our university's data protection officer.

The cases in the dataset have been evaluated with *RisCanvi* every 6 months. Persons who have only *RisCanvi*-S evaluation are low-risk cases and cases with both *RisCanvi*-S and *RisCanvi*-C evaluations are the ones who have been evaluated as high-risk in *RisCanvi*-S and then assigned to *RisCanvi*-C. Looking at previous work on the dynamics of the *RisCanvi* evaluations (Karimi-Haghighi & Castillo, 2021), the risk in general tends to be reduced with imprisonment. In other words, there is no criminogenic effect of imprisonment in this context. For each case, we sought the latest *RisCanvi* (*RisCanvi*-S or *RisCanvi*-C) evaluation, considering it valid for the purposes of predicting recidivism if it was done at most 9 months before the release date. There are many cases without a complete recent evaluation in the 9 months before their release. As these cases are considered "close" to their date of release, due to resource constraints and available staff, their evaluation process is not prioritized. This results in a large number of missing items due to some interviews not being performed. The largest number of removed samples (about 43% of the total) were due to incomplete *RisCanvi* evaluations, although we tried to impute missing items using their values from the previous valid evaluation. After imputing the missing items in *RisCanvi* evaluations (the imputation process is explained in the Appendix A.1) and removing cases with incomplete evaluations, we remained with 15,029 evaluations which are presented per release year in Table 2. As can be seen, the number of evaluations has increased with each passing year, as *RisCanvi* is adopted more consistently and thoroughly through the entire prison system. This variation might be a potential limitation of our study, however, in most of the release years (from 2012 to 2016) which consists of the majority of our dataset (72%), valid evaluations are between 50% and 70%.

We consider 220 features (feature list is found on Table 15 in Appendix B) including 23 demographic features, 146 penitentiary features, 43 *RisCanvi* items and 8 *RisCanvi* risk levels and scores (4 risk levels and 4 risk scores). There were also cases with missing values (about 10% of the total data) in some demographics and penitentiary

[4] Note that there are more prison releases than studied individuals as the data can record more than one prison release per person. Additionally, a small number of cases, not included in our dataset, are released due to other reasons, including being pardoned or successfully asking for a retrial finding them not guilty.

Table 2 *RisCanvi* evaluations per release year

Release year	2010	2011	2012	2013	2014	2015	2016
No. of released cases	3,494	3,766	4,152	4,010	3,999	3,596	3,288
No. of released cases with a valid evaluation (valid evaluation is at most 9 months prior to release)	634 (18%)	1,776 (47%)	2,320 (56%)	2,501 (62%)	2,702 (68%)	2,582 (72%)	2,514 (76%)
No. of Screening (*RisCanvi*-S) evaluations	172	670	1,021	1,036	916	836	807
No. of Complete (*RisCanvi*-C) evaluations	462	1,106	1,299	1,465	1,786	1,746	1,707

Table 3 Average recidivism rates two to five years after release for people released in 2010-2016 (total no. of releases = 26,305)

Recidivism type	within 2 years	within 3 years	within 4 years	within 5 years
Any (General)	16.4%	19.6%	21.9%	23.6%
Violent	4.2%	5.0%	5.5%	5.8%

features for which there were no values to be imputed. After dropping these cases, as well as very few special cases (twelve) that underwent *RisCanvi* evaluation but were not sentenced, 12,250 cases remain in the final dataset used in our analysis.

Recidivism

We obtained general and violent recidivism rates of the total cases in four follow-up periods within release date. The rates are shown in Table 3. Logically, the probability of committing a new crime after being released from prison is non-decreasing over the span of time spent out of prison.

In Fig. 2, we can also observe general and violent recidivism rates in the four follow-up periods for each release year. In all the follow-up periods, the highest rates of general and violent recidivism are observed in the persons released in earlier years of 2010 and 2011. Comparing recidivism rates of 2016 to recidivism rates of 2010, we find that there has been a minimum 6 and 2 percentage points decrease in general and violent recidivism rates, respectively. This decreasing rate is part of a global trend on the reduction of recidivism and crime rate (Velázquez, 2018; Tonry, 2014).

General recidivism rates have decreased in cases released between 2010 and 2014 in all follow-up periods, but the rates have been unchanged or increased for the cases released after 2014. There has been a decreasing rate of violent recidivism for the cases released between 2010 and 2015 within two years of their release. However, when looking at the recidivism rates for more than 2 years follow-up, this decrease happens for persons released between 2010 and 2013 and in 2015, while we observe an increased rate of violent recidivism for release years 2014 (except for the follow-up period of two years) and 2016 (except for the follow-up period of five years).

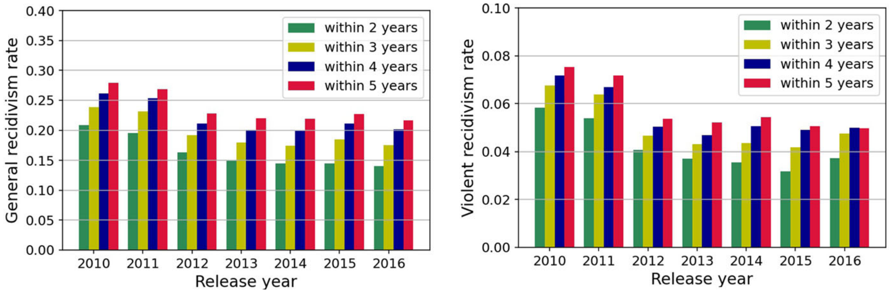

Fig. 2 Recidivism rates in four follow-up periods within each release year (total no. of releases in 2010-2016 = 26,305)

The rate we obtained for general recidivism of the cases released in 2010 (3,494 cases) in the follow-up period of five years (28%) is almost inline with what is estimated by the Center of Legal Studies and Specialized Training of the community (Capdevila et al., 2015). They estimate a general recidivism rate of 30% within 5.5 years of the cases released in 2010 (3,414 cases that are almost the same cases of our study who were released in 2010). In their report, there is a rate drop of 10 percentage points in general recidivism of the persons released in 2010 compared to the cases released in 2002.

For different groups of people released between 2010 and 2016, we can observe general and violent recidivism rates within 5 years of their release in Table 4.

Most groups are self-explanatory:

- *Single/other* refers to their civil status.
- *With deportation* are cases in which, normally due to violations of immigration law, the person is expected to be removed from the country immediately after being released.
- *Base crime types* include violent or non-violent crimes against property, crimes against people, gender-based violence, crimes against sexual freedom, drug-related crimes, traffic-related crimes, and others.
- *Permission* is an ordinary short leave from prison during the base sentence due to some reasons such as death or serious illness of a direct family member of the person.
- *Prison degree* can be 1st: cases that are isolated from the general prison population, 2nd: cases who are in general prison population, and 3rd: cases who must spend 8 hours in prison every day but can be outside 16 hours per day.
- *With electronic surveillance* are cases who have an electronic surveillance mechanism, usually an ankle bracelet. They are in the 3rd degree, but instead of coming to prison at night, they can sleep at their own home, as the ankle bracelet can send an alert if they are not at home at night.
- *Dependent units* are special, managed housing units outside of prison that can be assigned to those in the 3rd degree.

Comparing recidivism rates of these groups against the base rate (overall prevalence), we can see that cases younger than 30 at the time of release, who are national, with single civil status, pending deportation, who are in prison because of a violent crime or crime against property, convicted to more than 5 years sentence, with more than one previous prison entry, with rejected permission or no permission request, with mostly degree regression during their sentence (i.e., mostly being moved to a more restricted environment instead of a freer one), having (very) severe rules violations within prison, with (no) lower points in the prison evaluations, who have been relocated to another module within prison multiple times, who at least once went to special supervision, psychiatry, or nursing modules, and persons who were classified in the 1st and 2nd degree before their release have relatively higher violent and general recidivism rates compared to the related base rates. There are lower general and violent recidivism rates for cases with electronic surveillance and persons who spent time in dependent units.

Table 4 Recidivism rates within five years of release for different groups (no. of releases in 2010-2016 in the final cleaned data = 12,250)

Group	Size	General recidivism within 5 years Base rate: 17.1%	Violent recidivism within 5 years Base rate: 4.7%
Male	11,335 (93%)	17.3%	4.9%
Female	915 (7%)	13.7%	2.0%
Age at release time ≤ 30	2,712 (22%)	23.0%	6.9%
Age at release time > 30	9,538 (78%)	15.4%	4.1%
National	7,544 (62%)	19.0%	5.4%
Foreigner	4,706 (38%)	13.9%	3.5%
Single	6,929 (57%)	20.2%	6.0%
Other	5,321 (43%)	13.0%	3.0%
With deportation	1,418 (12%)	22.5%	6.8%
Without deportation	10,832 (88%)	16.4%	4.4%
Violent base crime	4,317 (35%)	19.4%	7.6%
Non-violent base crime	7,933 (65%)	15.8%	3.2%
(Non) violent base crime against property	3,742 (31%)	27.0%	7.8%
Other types of base crime	8,508 (69%)	12.6%	3.3%
Base crime sentence < 5 years	9,427 (77%)	16.6%	4.0%
Base crime sentence ≥ 5 years	2,823 (23%)	18.6%	7.0%
Previous prison entries > 1	3,231 (26%)	30.2%	8.6%
Previous prison entries ≤ 1	9,019 (74%)	12.4%	3.3%
Permission rejection or no permission request	3,571 (29%)	26.0%	7.0%
Permission acceptance	8,679 (71%)	13.4%	3.8%
Mostly degree regression	1,797 (15%)	28.6%	9.2%
Others	10,453 (85%)	15.1%	3.9%
(Very) severe rules violations within prison	4,206 (34%)	26.3%	8.2%
Light/no rules violations within prison	8,044 (66%)	12.3%	2.9%
(No) lower evaluation points (level C and D)	4,532 (37%)	24.0%	7.0%
Higher evaluation points (level A and B)	7,718 (63%)	13.0%	3.4%
Module changes > 7 (median)	5,825 (48%)	20.7%	6.5%
Module changes ≤ 7	6,425 (52%)	13.8%	3.1%
Special supervision module ≥ 1	2,300 (19%)	30.1%	10.1%
No special supervision module	9,950 (81%)	14.1%	3.5%

Table 4 continued

Group	Size	General recidivism within 5 years Base rate: 17.1%	Violent recidivism within 5 years Base rate: 4.7%
Nursing module ≥ 1	2,262 (18%)	21.0%	7.0%
No nursing module	9,988 (82%)	16.2%	4.2%
Psychiatry module ≥ 1	524 (4%)	30.0%	11.6%
No psychiatry module	11,726 (96%)	16.5%	4.4%
Last prison degree before release: 1st & 2nd	5,535 (45%)	26.4%	7.7%
Last prison degree before release: 3rd	6,715 (55%)	9.4%	2.2%
With electronic surveillance	1,572 (13%)	5.2%	1.0%
Others	10,678 (87%)	18.8%	5.3%
In dependent units	414 (3%)	9.9%	2.9%
Others	11,836 (97%)	17.3%	4.8%

Conditional release (C.R.) vs. Definitive release (D.R.)

In line with the main objective of this study to estimate the effect of Conditional Release (C.R.) on recidivism, we first look at C.R. rates in comparison to recidivism rates over the years and at descriptive statistics of relevant features for both C.R. cases and Definitive Release (D.R.) cases separately.

Table 5 shows that the C.R. rate has increased yearly from 22.6% in 2010 to 34.5% in 2016. While there are year-by-year variations in the amount of this increase in C.R., there does not seem to be any discontinuity or sudden change during the observation period. This increase is part of a strong policy push applied by the Dept. of Justice to request C.R. for more cases, which has led treatment committees to request for C.R. more often. It does not reflect any change in C.R. legislation, or any change that we are aware of in the criteria applied by judges.

The increase in C.R. rates can also be a reason for the recidivism rate decrease during these years as shown previously in Fig. 2. We can see that in the release years with the lowest C.R. rate (2010,2011) the recidivism rates within 2 to 5 years follow-up periods are the highest. Also, in release year 2014, which has a very small increase in C.R. rate, there has been no or a small recidivism rate decrease within 2 to 5 follow-up years. Even violent recidivism rates increased within follow-up periods higher than 3 years.

Table 5 Conditional release (C.R.) rate per year (total no. of releases in 2010-2016 = 26,305)

Release year	2010	2011	2012	2013	2014	2015	2016
No. of released cases	3,494	3,766	4,152	4,010	3,999	3,596	3,288
Conditional Release (C.R.) rate	22.6%	24.2%	26.5%	29.9%	30.4%	31.8%	34.5%

Considering conditional release as an intervention, or treatment that may reduce recidivism, we show some descriptive statistics of C.R. cases as treatment group in comparison to the ones of D.R. persons as control group. The comparison is presented in terms of some demographics, penitentiary features and recidivism within several years of release in Table 6.

As can be seen, the observed mean difference of D.R. and C.R. cases is statistically significant in variables such as gender, age at release time, civil status, violent base crime, deportation, permission request, acceptance, and rejection, number of nursing, psychiatry, and supervision modules, degree regression and progress, (very) severe rules violations within prison, number of previous prison entry, base crimes against people or property, gender-based violence crime, base crimes related to drugs or traffic, having electronic surveillance, being in dependent units, *RisCanvi* risks of self-directed violence, violence in the prison facilities, violent recidivism, and breaking prison permits, general recidivism within 2 to 5 years, and violent recidivism within 2 to 5 years.

Comparing the mean value of the variables in C.R. and D.R. cases shows the following. Cases that are less likely to receive C.R. (and hence, more likely to be released under D.R.) tend to be: cases with single civil status, who are in prison because of violent crime or non-violent crimes against property, pending deportation, with no permission request or rejected permissions, spending time in nursing, psychiatry, or special supervision modules, with mostly degree regression, who remain in the 2nd prison degree, having (very) severe rules violations within prison, with previous prison entry, and having higher risk scores of self-directed violence, violence in the prison facilities, violent recidivism, and breaking prison permits. On the other hand, cases that are more likely to receive C.R. tend to be persons with married civil status, who were more often granted and enjoyed permissions, having always or mostly degree progressions, who are in prison because of a drug-related crime, with electronic surveillance, or lived in dependent units.

Both general and violent recidivism are more likely in cases released with D.R. compared to C.R. General recidivism within 2 to 5 years of prison release happens 12 to 17 percentage points more in cases with D.R. compared to ones with C.R. Violent recidivism within the same periods happens 3 to 5 percentage points more in cases with D.R. compared to C.R. We can also compare general and violent recidivism rates of the C.R. and D.R. cases within 2 to 5 years of their release per each release year in Fig. 3. We can see at least 14 percentage points and 3 percentage points higher general and violent recidivism rates respectively for D.R. cases compared to the C.R. persons in all follow-up periods within each release year.

From this observation alone, however, one cannot conclude that C.R. causes a reduction in recidivism risk. C.R. is granted almost exclusively to cases who are deemed to have lower risk, as the relevant legislation indicates that people who receive C.R. must have a “favorable individualized prognosis of social reintegration”. Therefore, we need to control for this selection through causal inference methods, if we want to study the causal effect of C.R. on recidivism.

Table 6 Descriptive statistics: control (D.R.) vs. treatment (C.R.) (no. of releases in 2010-2016 in the final cleaned data = 12,250)

Variable	Mean value in control (D.R.) (8,162 cases)	Mean value in treatment (C.R.) (4,088 cases)	P-value (D.R. vs. C.R.)
Male[1]	0.94	0.89	< 0.0001
Age at release	38.78	40.67	< 0.0001
Single[1]	0.60	0.50	< 0.0001
Married[1]	0.19	0.28	< 0.0001
Violent base crime[1]	0.39	0.28	< 0.0001
Deportation[1]	0.15	0.04	< 0.0001
Rejected permission[1]	0.24	0.03	< 0.0001
Accepted and enjoyed permission[1]	0.56	0.96	< 0.0001
No permission request[1]	0.18	0.01	< 0.0001
Number of nursing modules	0.49	0.39	0.001
Number of psychiatry modules	0.13	0.04	< 0.0001
Number of special supervision modules	0.82	0.18	< 0.0001
Mostly degree regression[1]	0.20	0.05	< 0.0001
Mostly degree progress[1]	0.24	0.52	< 0.0001
Remaining in the 2nd prison degree[1]	0.48	0.00	< 0.0001
Always degree progress[1]	0.20	0.51	< 0.0001
Severe prison rules violations[1]	0.25	0.15	< 0.0001
Very severe prison rules violations[1]	0.16	0.05	< 0.0001
Number of previous prison entry	1.55	1.11	< 0.0001
Base crime: Against people	0.15	0.12	< 0.001
Base crime: Gender-based violence	0.13	0.06	< 0.0001
Base crime: Against property[2]	0.24	0.12	< 0.0001
Base crime: Drugs	0.10	0.38	< 0.0001
Base crime: Traffic	0.12	0.09	< 0.0001
With electronic surveillance[1]	0.05	0.29	< 0.0001
In dependent unit[1]	0.02	0.06	< 0.0001
Self-directed violence risk score	10.83	2.57	< 0.0001
Score of violence in the prison facilities	9.70	1.22	< 0.0001
Violent recidivism risk score	10.26	1.12	< 0.0001

Table 6 continued

Variable	Mean value in control (D.R.) (8,162 cases)	Mean value in treatment (C.R.) (4,088 cases)	P-value (D.R. vs. C.R.)
Breaking prison permits risk score	0.05	-1.46	< 0.0001
General recidivism[1] within 2 years	0.14	0.02	< 0.0001
General recidivism[1] within 3 years	0.17	0.04	< 0.0001
General recidivism[1] within 4 years	0.20	0.05	< 0.0001
General recidivism[1] within 5 years	0.23	0.06	< 0.0001
Violent recidivism[1] within 2 years	0.04	0.01	< 0.0001
Violent recidivism[1] within 3 years	0.05	0.01	< 0.0001
Violent recidivism[1] within 4 years	0.06	0.01	< 0.0001
Violent recidivism[1] within 5 years	0.06	0.01	< 0.0001

[1] 0=no, 1=yes
[2] non-violent

Methodology

The distribution of the features (feature list is found in Table 15 in Appendix B) by gender reveals significant differences between men and women which are explained in "Gender differences" section. Considering conditional release as a treatment, we focus on the causal effect of this treatment on general and violent recidivism. The average treatment effect is obtained in "Average treatment effect (ATE)" section using statistical methods such as Propensity Score Matching (PSM), Inverse Propensity

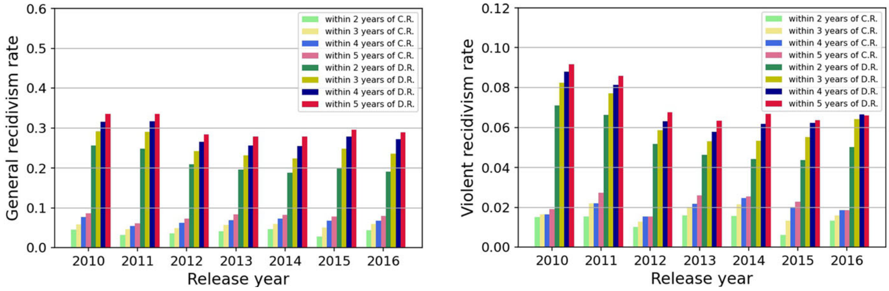

Fig. 3 General and violent recidivism rates in C.R. (light-color bars) and D.R. (dark-color bars) cases (total no. of releases in 2010-2016 = 26,305)

score Weighting (IPW), and Augmented Inverse Propensity Weighting (AIPW). All of these methods rely on the propensity to treatment which is estimated in "Propensity to conditional release (C.R.)" section using Machine Learning (ML) models. In the AIPW method, we also need to obtain a model for the outcome (general/violent recidivism). We present the models for both general and violent recidivism prediction in "General and violent recidivism prediction" section. Finally, in "Conditional average treatment effect (CATE)" section, we determine the treatment effect heterogeneity by estimating Conditional Average Treatment Effect (CATE) using Generalized Random Forest and Double Machine Learning methods.

Gender differences

We observe significant differences between men and women within our dataset, so we prefer to treat them differently and create separate ML models for these two groups (Skeem et al., 2016; Collins, 2010; Huebner et al., 2010).

Table 7 shows descriptive statistics of relevant features of this study separately for men and women. The mean difference is statistically significant for almost all listed variables. In general, these include differences in the types of crimes committed by men and women, in their behavior and activities within prison and progression to C.R., and in their recidivism rates.

Also, testing a global model of propensity to C.R. (which is trained using all population consisting both men and women) on women and men results in different predictive performances which can be interpreted as an algorithmic bias. So we prefer not to use a global model to prevent this algorithmic discrimination.

Propensity to conditional release (C.R.)

We consider Conditional Release (C.R.) as a treatment, and hence people released with C.R. are the treatment group, and people released with Definitive Release (D.R.) are the control group. Among releases between 2010 and 2016, there are 11,335 releases for men; 32% of them are C.R. (treatment) and 68% are D.R. (control). Similarly, there are 915 releases for women, 48% of them are C.R. and 52% are D.R.

For both men and women, we estimate the propensity to treatment (C.R.) using different Machine Learning (ML) models such as Logistic Regression (LR), Multi-Layer Perceptron (MLP) and Random Forest (RF). Input features (feature list is found in Table 15 in Appendix B) to the models consist of 21 demographics, some penitentiary features (142 for men and 89 for women, the difference is due to the fact that some penitentiary centers include only men and some include only women), 43 *RisCanvi* items, 8 *RisCanvi* risk levels and scores (4 risk levels and 4 risk scores). In order to account for the fact that risk assessment tools are trained on historical data to predict the risk of recidivism in the future, we split into training and test set accordingly. In more detail, we use the cases with releases between 2010 and 2014 for training each model (7,482 cases in men model and 592 for women model) and test the models using cases released in 2015 and 2016 (3,549 cases in men model and 310 in women model). In each model, the test set does not include any cases of the training set, which

Table 7 Descriptive statistics: men vs. women (no. of releases in 2010-2016 in the final cleaned data = 12,250)

Variable	Mean value in men (11,335 cases)	Mean value in women (915 cases)	P-value (men vs. women)
National[1]	0.61	0.67	< 0.001
Single[1]	0.57	0.51	< 0.001
Violent base crime[1]	0.36	0.20	< 0.0001
Deportation[1]	0.12	0.04	< 0.0001
No. of rejected permissions	0.84	0.64	0.1
No. of accepted and enjoyed permissions	9.68	10.66	< 0.1
No. of activities	16.80	19.90	< 0.0001
No. of nursing modules	0.41	1.03	< 0.0001
No. of special supervision modules	0.56	1.23	< 0.0001
No. of light prison rules violations	0.04	0.28	< 0.0001
No. of severe prison rules violations	0.56	0.80	< 0.0001
No. of very severe prison rules violations	0.46	0.62	0.001
Mostly degree progress[1]	0.33	0.39	< 0.0001
Remaining in the 2nd prison degree[1]	0.33	0.20	< 0.0001
Always degree progress[1]	0.30	0.37	< 0.0001
With electronic surveillance[1]	0.11	0.30	< 0.0001
In dependent unit[1]	0.03	0.11	< 0.0001
Previous prison entry[1]	0.83	0.71	< 0.0001
Base crime: Against people	0.14	0.09	< 0.0001
Base crime: Gender-based violence	0.11	0.02	< 0.0001
Base crime: Against property (violent)	0.11	0.09	0.1
Base crime: Against property[2]	0.20	0.28	< 0.0001
Base crime: Drugs	0.18	0.34	< 0.0001
Base crime: Traffic	0.11	0.05	< 0.0001
Self-directed violence risk score	8.49	2.93	< 0.0001
Score of violence in the prison facilities	7.25	2.14	< 0.0001
Violent recidivism risk score	7.84	-0.67	< 0.0001
Breaking prison permits risk score	-0.44	-0.68	< 0.1
C.R.[1]	0.32	0.48	< 0.0001

Table 7 continued

Variable	Mean value in men (11,335 cases)	Mean value in women (915 cases)	P-value (men vs. women)
General recidivism[1] within 2 years	0.10	0.08	< 0.1
General recidivism[1] within 3 years	0.13	0.10	0.01
General recidivism[1] within 4 years	0.15	0.12	< 0.01
General recidivism[1] within 5 years	0.17	0.14	< 0.01
Violent recidivism[1]within 2 years	0.03	0.01	< 0.001
Violent recidivism[1] within 3 years	0.04	0.01	< 0.0001
Violent recidivism[1] within 4 years	0.05	0.02	< 0.001
Violent recidivism[1] within 5 years	0.05	0.02	< 0.0001

[1] 0=no, 1=yes
[2] non-violent

is why the total size of the training and test sets are smaller than their related total population.

Finally, we ensure that the distribution of treatment cases in training and test set of each model is almost balanced. The percentage of C.R. cases in the training and test set of the model for men is 31% and 36%, respectively and C.R. cases in the training and test sets of the women is 47% and 50%, respectively.

General and violent recidivism prediction

To compute the causal effect of conditional release (C.R.) on general and violent recidivism using the Augmented Inverse Propensity Weighted (AIPW) method, we construct models for both the outcomes (general and violent recidivism within 2 to 5 years of release) and the propensity to be assigned to C.R. Using different ML algorithms (such as LR, MLP, and RF) and the same training and test sets used in C.R. propensity models of men and women, we obtain eight prediction models of general and violent recidivism outcomes within 2 to 5 years of release for each group. The input features of these models are the same features used in the models for the propensity to C.R. plus the actual treatment variable (C.R.).

Average treatment effect (ATE)

We compute the Average Treatment Effect (ATE) of C.R. on general and violent recidivism using various causal inference methods. In order to obtain consistent estimates of the causal effect, the following conditions need to hold:

- Stable Unit Treatment Value Assumption (SUTVA) (Angrist et al., 1996): We assume that the Stable Unit Treatment Value Assumption (SUTVA) holds such that the recidivism risk of a person is unaffected by the particular assignment of C.R. to other cases.
- Common Support (Overlap) (Caliendo & Kopeinig, 2008): Common support means that there is complete 'overlap' in the distribution of propensity scores across treatment and control cases to find adequate matches. This condition is also satisfied in our study, which is shown in the propensity score distribution plots (Figs. 4 and 5) in "Results" section.
- Conditional Independence (Dawid, 1979): Conditional independence or unconfoundedness requires, that conditional on all confounders used in the model, the assignment of C.R. is random. This condition cannot be tested but the different estimates of the propensity to treatment yield notably high AUC values which can be attributed to the amount and the criminological relevance of the confounders that we use. This can be seen in "Results" section. This suggests that we include most of the relevant confounders to predict treatment assignment and the two recidivism outcomes.

When these conditions have been fulfilled, then there is 'strong ignorability' of how an individual came to be treated relative to the outcome (Rosenbaum & Rubin, 1983). Strong ignorability implies that no systematic, unobserved, pretreatment differences exist between treated and control subjects that are related to the response under study (Joffe & Rosenbaum, 1999).

In the following, we explain the causal inference methods of Propensity Score Matching (PSM), Inverse Propensity score Weighting (IPW), and Augmented Inverse Propensity Weighting (AIPW) that we use to obtain the Average Treatment Effect of conditional release on general and violent recidivism.

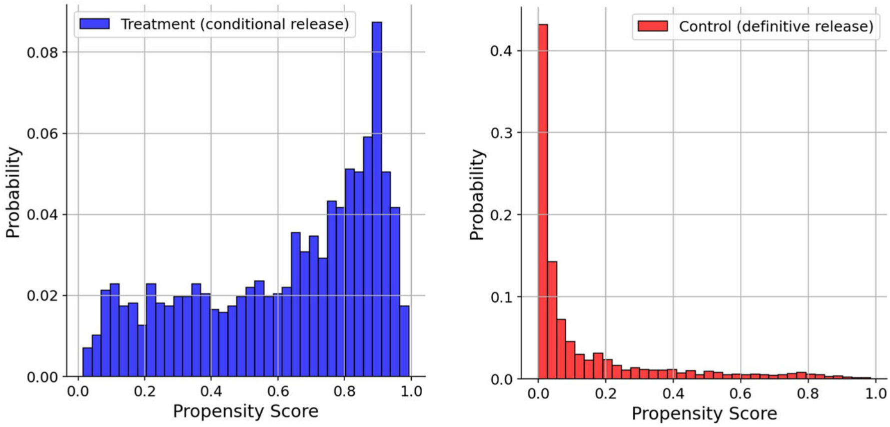

Fig. 4 Distribution of the propensity to treatment (C.R.) for **men** in our sample

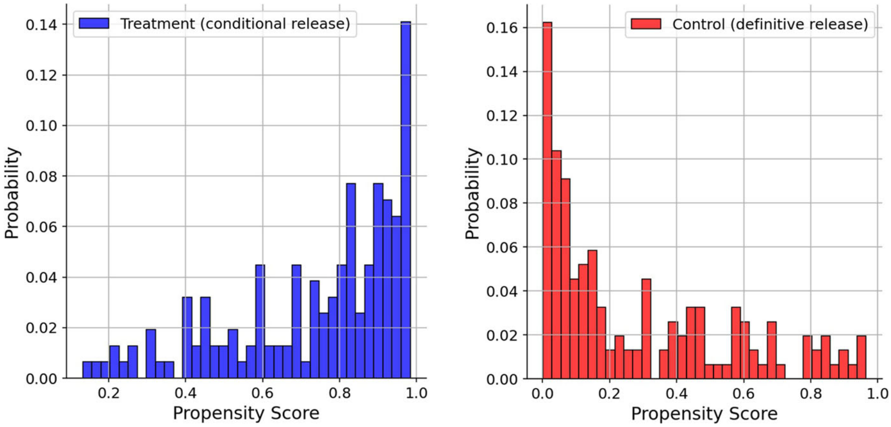

Fig. 5 Distribution of the propensity to treatment (C.R.) for **women** in our sample

Propensity score matching (PSM)

In the method of propensity score matching (Rosenbaum & Rubin, 1983), we do the matching by ranges of propensity scores. Thus, we sort the data by propensity scores and then stratify it into buckets (four in our case) of almost the same size (for men 3 buckets of the exact size of 887 and one bucket with 888 test cases and for women 3 buckets of the exact size of 77 and one bucket with 79 test cases). In our work, we obtain the ATE by subtracting the mean recidivism of non-treated (control) cases from treated ones in each bucket.

Inverse-propensity score weighting (IPW)

The basic idea of this method is weighting the outcome measures by the inverse of the individual's treatment propensity so that similar baseline characteristics are obtained (Bray et al., 2019). In this method, the treatment effect for individual i is obtained using Eq. 1:

$$TE_i = \frac{W_i Y_i}{p_i} - \frac{(1 - W_i) Y_i}{1 - p_i} \tag{1}$$

W_i indicates treatment (1 for treated and 0 for control cases), p_i represents probability of receiving treatment (propensity score of treatment), and Y_i indicates recidivism (1 if recidivate and 0 if not recidivate) for individual i.

The IPW method places more weights on observations from the control group with a high treatment propensity and vice versa for observations in the treatment group, improving on the covariate balance. In other words, the untreated units with higher estimated probability of being treated and the treated units with lower estimated probability of being treated receive higher weights. At last, the model is estimated using data of those that are more similar (thus more comparable) to each other. "Extracting" data on similar observation units mimics natural experiments.

If the propensity scores were known (which is the case here), then this estimator will be unbiased for the ATE (Tsiatis, 2006). Furthermore, when the propensity scores are estimated consistently, then this estimator is consistent for the ATE. In our study, looking at the propensity score distributions of the treatment and control groups in "Results" section (Figs. 4 and 5), we can see the consistency of these estimates especially for men group. The IPW estimator is also widely believed to have poor small sample properties when the propensity score gets close to zero or one for some observations. Specifically, treatment cases with very low propensity scores and control cases with very high propensity scores will provide extreme contributions to the estimate (Glynn & Quinn, 2010). However, in our study, according to the propensity score distribution of the treatment and control group which is shown in "Results" section (Figs. 4 and 5), the percentage of treatment cases with very low propensity scores and control cases with very high propensity scores are very low (less than 1%).

Augmented inverse-propensity weighted (AIPW)

This method combines both the properties of the regression-based estimator and the IPW estimator. It has an augmentation part $(W_i - p_i)\widehat{Y_i}$ to the IPW method, in which $\widehat{Y_i}$ is the estimated probability of recidivism using all features applied to the propensity score model plus the treatment variable. In other words, in this method, two models are used; a binary regression model for the propensity score, and a regression model for the outcome variable. So, this estimator yields doubly robust estimations which requires only either the propensity or outcome model to be correctly specified but not both. Comparing this estimator to IPW and PSM estimators, it is shown that the AIPW estimator has comparable or lower mean square error than the other two estimators. When the propensity score and outcome models are both properly specified and, when one of the models is misspecified, the AIPW estimator is superior (Glynn & Quinn, 2010). This double-robustness property gives the AIPW estimator a tremendous advantage over most other estimators in that with the AIPW estimator the researcher has more hope of getting a reasonable answer in complicated real-world situations where there is uncertainty about both the treatment assignment process and the outcome model. We can compute the AIPW treatment effect on individual i as in Eq. 2:

$$TE_i = \frac{W_i Y_i - (W_i - p_i)\widehat{Y_i}}{p_i} - \frac{(1 - W_i)Y_i - (W_i - p_i)\widehat{Y_i}}{1 - p_i} \tag{2}$$

In IPW and AIPW, we obtain the individual treatment effect TE_i, which is the difference between the outcomes if the person is treated (treatment) and not treated (control). In other words, this effect is the difference of recidivism probability when the person is granted C.R. and not granted C.R. A negative value shows a reduced recidivism risk and a positive value indicates an increased recidivism risk. The resulting ATE is the average over all individual treatment effects.

Conditional average treatment effect (CATE)

To determine heterogenous effects of C.R. on recidivism, we estimate the Conditional Average Treatment Effect (CATE) using Generalized Random Forest and

Double Machine Learning (Athey et al., 2019; Nie & Wager, 2021; Chernozhukov et al., 2016). Generalized Random Forests are flexible methods for estimating treatment effect heterogeneity with Random Forests. The specific application of this algorithm to estimate CATE is what researchers call Causal Forests. These estimators are used as final models for CATE estimation within the EconML (**econml**) package. CATE is the ATE conditioned on membership in a subgroup. Using *econml.dml.CausalForestDML* in the EconML package, we obtain the CATE by considering a single covariate, while keeping all the other covariates at a fixed value (median). The *econml.dml.CausalForestDML* combines a Causal Forest with Double Machine Learning to residualize the treatment and outcome, which again yields doubly robust estimates.

Results

We observe the predictive performance of the Machine Learning (ML) models introduced in "Methodology" section for the propensity to receive conditional release (C.R.) and general and violent recidivism prediction of men and women groups in "Predictive Performance of ML models" section. Due to double-robustness property of the Augmented Inverse Propensity Weighted (AIPW) estimator, we present the computed Average Treatment Effect (ATE) by this estimator in "Average treatment effect (ATE)" section for both gender groups. The ATE results obtained by the two other statistical methods of Propensity Score Matching (PSM) and Inverse Propensity score Weighting (IPW) are shown in Appendix C. In "ATE by risk level" section, we compare the obtained ATE values in cases with three different risk levels (high, medium, and low) of the *RisCanvi* risk assessment tool. Finally, the results of Conditional ATE (CATE) on membership in different subgroups are given in "Conditional average treatment effect (CATE)" section.

Predictive Performance of ML models

We find that Logistic Regression yields the most accurate prediction of the propensity to receive conditional release (C.R.) for both men and women. The results are shown in Table 8 in terms of AUC-ROC which stands for "Area under the ROC (Receiver Operating Characteristics) Curve". This metric is used to measure the performance of the classification models at various threshold settings. ROC is a probability curve with TPR (True Positive Rate) against FPR (False Positive Rate) and AUC measures the entire two-dimensional area underneath the ROC curve. The Higher the AUC, the better the classification model is at distinguishing between positive and negative classes. Obtained estimates from our ML models are well calibrated for both groups (calibration curves omitted for brevity). According to the AUC values in Table 8, ML models show accurate results for both groups especially for men. The models will be used in the computation of the ATE. We can also observe the distribution of the C.R. propensity scores in treatment and control groups of men and women in Figs. 4 and 5. As can be seen, for both men and women there is an overlap in the distribution of

Table 8 AUC-ROC of propensity to conditional release (C.R.) prediction

Group	Size (test set)	AUC-ROC	Model
Men	3,549	0.92	Logistic Regression
Women	310	0.89	Logistic Regression

the propensity scores of treatment and control cases to find adequate matches. This is a necessary condition to be able to apply our causal inference methods. Also, for both groups the distributions are well spread between 0 and 1.

We can also observe the predictive performance of the ML-based models of general and violent recidivism within 2 to 5 years of release in terms of AUC-ROC in Table 9. The algorithm used for all models is Random Forest with 500 estimators and a maximum depth of 2. As can be seen, all models show high AUC for both risk outcomes and in both groups especially for women. In general recidivism prediction, the AUC value is 0.74-0.77 and 0.84-0.87 for men and women groups respectively. The AUC results for violent recidivism prediction show values of 0.77-0.80 and 0.78-0.83 for men and women respectively. These outcome models will be used in the computation of ATE in the AIPW method.

Average treatment effect (ATE)

Our goal is to determine whether conditional release (C.R.) has a causal effect on general and violent recidivism within 2 to 5 years of a person's release. The Average Treatment Effect (ATE) computed using Augmented Inverse-Propensity Weighting (AIPW) method is shown on Tables 10 and 11 for men and women respectively. We can observe negative confidence intervals in all ATE values for general recidivism risk of men and women, which suggests a causal effect of C.R. in the reduction of general recidivism risk for both groups within 2 to 5 years of their release. In the ATE values obtained from AIPW method for violent recidivism, negative confidence intervals are found for men within 3 years and women within 5 years of their release. However, in the AIPW results obtained for other follow-up periods of the two groups, which are shown in italics, the confidence intervals contain the value zero, from which we cannot establish whether there is a change in the violent recidivism risk due to C.R.

The ATE obtained using Propensity Score Matching (PSM) is shown in Appendix C on Tables 16 and 17 for men and women respectively. For men with lower propensity

Table 9 AUC-ROC of general and violent recidivism prediction using Random Forests

	Recidivism within	2 years	3 years	4 years	5 years
Men (3,549 cases)	General	0.77	0.75	0.75	0.74
	Violent	0.80	0.80	0.78	0.77
Women (310 cases)	General	0.87	0.86	0.85	0.84
	Violent	0.83	0.81	0.78	0.79

Table 10 AIPW results estimating ATE and its 95% confidence interval [lower-ci, upper-ci]. Almost all of the confidence intervals, with the exception of those in italics, lie entirely in the negative region

Men (3,549 cases)	AIPW on general recidivism			AIPW on violent recidivism		
	lo-ci	ATE	up-ci	lo-ci	ATE	up-ci
within 2 years	-0.06	-0.04	-0.02	*-0.02*	*-0.01*	*0.002*
within 3 years	-0.08	-0.05	-0.03	-0.03	-0.02	-0.003
within 4 years	-0.08	-0.06	-0.03	*-0.03*	*-0.01*	*0.01*
within 5 years	-0.09	-0.06	-0.03	*-0.03*	*-0.01*	*0.01*

to receive C.R., we can observe a more negative ATE of C.R. on both general and violent recidivism within all follow-up periods. This means that if men with low probability of getting C.R. (high risk cases) have chances to receive C.R., their general and violent recidivism probability within 5 years of their release would be respectively 34 percentage points and 11 percentage points lower than if they would not receive C.R. For women, the effects are similar for the group with the lowest propensity to receive C.R. but slightly different from the effects for men for the higher propensity groups. For violent recidivism of women, we find a stronger ATE in buckets with lower propensity to receive C.R. but in buckets with medium-high and high C.R. probability there is no significant effect of C.R. on violent recidivism. However, considering all cases together, the ATE of C.R. on general and violent recidivism (within all follow-up periods) is negative for both men and women.

The ATE results obtained from Inverse-Propensity score Weighting (IPW) are shown on Tables 18 and 19 in Appendix C for men and women respectively. According to these results, all confidence intervals in ATE values lie entirely in the negative region for general and violent recidivism of both men and women. This is a strong indication that C.R. reduces the risk of violent and general recidivism for men and women within 2 to 5 years of their release.

ATE by risk level

In this section, we explore the heterogeneity of the computed ATEs by three different risk levels (high, medium, and low) of violent recidivism risk (REVI risk) as obtained from the *RisCanvi* risk assessment tool. In Table 12 violent recidivism base rates

Table 11 AIPW results estimating ATE and its 95% confidence interval [lower-ci, upper-ci]. Almost all of the confidence intervals, with the exception of those in italics, lie entirely in the negative region

WOMEN	AIPW on general recidivism			AIPW on violent recidivism		
(310 cases)	lo-ci	ATE	up-ci	lo-ci	ATE	up-ci
within 2 years	-0.10	-0.06	-0.02	*-0.02*	*-0.01*	*0.004*
within 3 years	-0.14	-0.08	-0.02	*-0.03*	*-0.01*	*0.002*
within 4 years	-0.16	-0.10	-0.04	*-0.04*	*-0.02*	*0.001*
within 5 years	-0.18	-0.12	-0.05	-0.04	-0.02	-0.001

Table 12 Violent recidivism base rates per REVI level. Violent recidivism probability is higher for men having higher REVI risk assessments. Result can not be established for women due to small sample size

			Violent recidivism within:							
	Size		2 years		3 years		4 years		5 years	
REVI level	men 3,549	women 310	men	women	men	women	men	women	men	women
Low	2,594	286	1%	1%	2%	1%	2%	1%	3%	2%
Medium	609	17	6%	6%	8%	6%	10%	6%	10%	6%
High	346	7	9%	0%	11%	14%	12%	14%	13%	14%

are shown for different REVI risk levels. As can bee seen, for men there is a clear correlation between the base rates and the *RisCanvi* risk levels in all follow-up periods which means that the estimated REVI risk level by *RisCanvi* is consistent with the violent recidivism rates within 2 to 5 years of release. This correlation is not as clear for women due to small sample size. Risk level for women is almost always “low risk” in our sample so that 286 women have low risk, but only 17 have medium risk, and only 7 have high risk, which makes statistics relating REVI risk and recidivism unreliable. Hence, we only compare ATE values by different REVI risk levels for men.

On Table 13, the ATE of C.R. on violent recidivism of men (within 2 to 5 years of release) obtained from AIPW is shown for three different REVI risk levels. Similar results obtained from IPW estimator are also shown on Table 20 in Appendix C.1. ATE values with confidence intervals consisting value zero are not reliable and shown in italics. Comparing other ATE values (with confidence intervals not including value zeros), we can see the most negative ATE of C.R. on violent recidivism in cases with medium REVI risk level in both AIPW and IPW results. These results show that granting C.R. to men with higher REVI risk (medium) yields a stronger reduction in violent recidivism risk compared to granting C.R. only to the cases with a low REVI risk level. According to these results, we note that the risk estimated by a risk assessment tool should not be linked to treatment (C.R.) of a case. By dedicating more resources toward higher risk detected cases than lower risk ones the community protection can be effectively promoted (Hanson, 2005).

Conditional average treatment effect (CATE)

We measured Conditional Average Treatment Effect (CATE) in different groupings according to base crime and criminological features. This is important because it would be relevant to know if granting C.R. to specific groups could yield a stronger reduction in recidivism compared to other groups. However, we found no differences worth reporting (mostly one percentage point or less), and in all groups we studied the effect of C.R. is a reduction of recidivism, and the reduction is similar in magnitude.

Discussion and conclusion

In this paper we studied the effect of conditional release (C.R.) on violent and general recidivism of persons who were released from several prison centers in Catalonia in

Table 13 ATE-AIPW of C.R. on violent recidivism and its 95% confidence interval [lower-ci, upper-ci] in different REVI risk levels of **men**. Confidence intervals shown in italics contain zero value and are not reliable. Granting C.R. to men with medium REVI risk yields a stronger reduction in violent recidivism risk compared to low REVI risk cases

	AIPW on violent recidivism within 2 years			AIPW on violent recidivism within 3 years		
REVI level	lo-ci	ATE	up-ci	lo-ci	ATE	up-ci
Low (2,594)	-0.02	-0.01	-0.003	-0.03	-0.02	-0.01
Medium (609)	-0.05	-0.03	-0.01	-0.07	-0.04	-0.01
High (346)	*-0.05*	*0.04*	*0.13*	*-0.06*	*0.03*	*0.12*
	AIPW on violent recidivism within 4 years			AIPW on violent recidivism within 5 years		
REVI level	lo-ci	ATE	up-ci	lo-ci	ATE	up-ci
Low (2,594)	*-0.03*	*-0.005*	*0.02*	*-0.02*	*-0.001*	*0.02*
Medium (609)	-0.09	-0.05	-0.02	-0.08	-0.05	-0.02
High (346)	*-0.06*	*0.03*	*0.12*	*-0.06*	*0.03*	*0.12*

Spain between 2010 and 2016. Due to noticeable differences in men and women in our dataset with respect to some penitentiary features and the performance of the predictive models, we treated them differently by creating separate ML models for these two groups (Skeem et al., 2016; Collins, 2010; Huebner et al., 2010). Considering separate analysis and modeling for groups with significant differences, as the gender group in our study, can lead to more reliable and effective results compared to considering a global behavioral model for all populations.

We computed the Average Treatment Effect (ATE) of C.R. on both general and violent recidivism of men and women using several statistical causal inference methods such as Propensity Score Matching (PSM) (Rosenbaum & Rubin, 1983), Inverse-Propensity score Weighting (IPW) (Bray et al., 2019), and Augmented Inverse-Propensity Weighted (AIPW) (Glynn & Quinn, 2010) methods. In contrast to previous work, we applied different methods for ATE computation which leads to more robust results. These methods require a precise prediction of the propensity to receive treatment (C.R.) and the probability of the studied outcome (recidivism). For both the treatment propensity and the outcome probability we obtain high predictive performance in terms of AUC. This suggests that our data explains most of the variations in treatment and outcome which supports our identification strategy. The obtained ATE values from all of the methods mostly show that C.R. reduces the risks of violent and general recidivism of men and women within 2 to 5 years of their release. This effect of C.R. on recidivism probability reduction can be attributable to the greater post-release supervision, support mechanism, training, and rehabilitative programs provided to those granted C.R. in comparison to D.R. cases.

We studied the ATE of C.R. on the risk reduction of groups with different risk categories. To this end, we compared ATE values of men with 3 different risk levels of violent recidivism (REVI risk) estimated in *RisCanvi* risk assessment tool. Comparison could not be established for women due to the small sample size. The comparison showed that granting C.R. to men with medium REVI risk can be more effective in reducing their violent recidivism probability compared to granting C.R. to the cases with low REVI risk level.

We believe that the findings of our research can be beneficial for practitioners and policymakers. These results can be used in the design of processes for C.R. by sometimes considering higher risk cases, while at the same time trying not to divert the resources away from the vulnerable groups, as currently there is a tremendous pressure to focus resources on defendants who are assessed as low-risk (Bonta & Andrews, 2007; Andrés-Pueyo et al., 2018). As we showed, in some higher risk cases there would be more reduction in their violent recidivism probability if they are granted C.R. This can be a response to the question of how we can promote the reduction of risk and not merely its assessment (Monahan & Skeem, 2016).

We would also emphasize that any jurisdiction that uses structured evaluations and C.R., should conduct periodic studies, similar to ours, to monitor the effectiveness of C.R. in recidivism reduction and the efficiency of the risk assessment tool in the C.R. assignment process.

Additionally, we believe our results may also contribute to supporting the idea of doing structured evaluations in countries where they are not currently done, as efficient use of such structured tools in conjunction with C.R. may lead to reductions in recidivism.

Policy Implications

A recommendation that these results suggest is that the risk estimated by a risk assessment tool should not be the only basis for granting treatment (C.R.). In fact, our results show that granting C.R. to higher risk detected cases can yield improvements in community safety by reducing overall recidivism rates. However, risk assessment, as currently used, mainly serves as a motivation to grant C.R. to low-risk incarcerated persons. This usage has two main problems: First, it assumes risk is static, but according to the "third generation" of risk assessment tools, we should address dynamic factors that can be changed to reduce risk. Instead of determining risk, we should move towards needs assessment and intervention, based on the risk-need-responsivity (RNR) principle, and look at what needs an individual has that can be met to reduce their risk (Bonta, 1996; Barabas et al., 2018). Causal inference methods that we used in this study are applicable in such frameworks which do not assume a fixed future for a person. These methods can help understand how risk can be changed, and hence design effective interventions. According to the RNR principle, to achieve effective rehabilitation, risk instruments have to be evidence-based and level of rehabilitation service should go with the level of risk, type of criminogenic need, and learning style and motivations (responsivity) of the individual being treated (Bonta & Andrews, 2007).

Second, even if the risk assessment tool is used to determine risk, it is unclear that the best for society is to grant C.R. only to low-risk cases, as a robust conclusion from our analysis is that C.R. greatly reduces the chances of recidivism for higher risk cases. This may seem to contradict the literature related to risk estimates for flight risk (Kleinberg, 2018), which uses such estimates to grant bail to low-risk defendants. However, pre-trial and C.R. applications of risk assessment instruments, which tend to be considered as two analogous settings by computer scientists, should not be treated in the same way. Hence, we believe that the connection between risk assessment and C.R. requires a deep examination in light of these results.

This causal inference study for C.R. application provides a path towards effectively supporting incarcerated persons, less incarceration, and prison systems with capacity of C.R. programs in which cost of C.R. is lower than the cost of incarceration. Also, causal inference methods such as the ones we used allow to perform observational studies, as criminal justice is a domain in which some types of direct experimentation might be unethical or harmful. We also used a large dataset and our results hold across substantially diverse prison centers. We stress that the methodology we described is broadly applicable. Our findings are likely to be specific to this particular dataset, but show the general effectiveness of the methodology in this setting.

Limitations

Considering reincarceration as recidivism ("penal recidivism") has some limitations, including narrow scope due to underreporting crimes (Beck & Shipley, 1989), differential legal outcomes with respect to some demographics (Pettit & Gutierrez, 2018; Ropes Berry, 2020), and ignoring the public safety impact of some non-imprisonable low-risk offenses (Roberts, 2004). However, studies of reincarceration have been used in Spain for several years to understand how to reduce mass incarceration, and this has the potential to enhance the monitoring of individuals' interactions and their engagement with the criminal justice system, while also ensuring accountability within the criminal justice system itself.

We are studying a dynamic setting in which the usage of *RisCanvi* increases over the years. While in the majority of our dataset we observe between 50% and 70% of valid evaluations (valid evaluation is at most 9 months prior to release), this variation might be a potential limitation of our study.

In causal inference studies, one of the conditions that needs to be satisfied to obtain consistent estimates of the causal effect is conditional independence or unconfoundedness, which cannot be tested. This condition requires that, conditional on all confounders used in the model, the assignment of treatment is random. The high AUC values we obtained in predicting treatment assignment and risk outcomes suggests that this assumption is supported.

Future work

Some potential areas for future research, drawn from this study, can be outlined as follows.

In causal inference studies, more scenarios can be defined to design an intervention using other features or different combinations of features. Additionally, changes in the situation of an inmate other than C.R. can also be considered interventions, such as being classified in the 3rd prison degree – people who must spend 8 hours in prison every day but can be outside up to 16 hours per day.

Given that risk evaluations are done periodically, there exists the possibility of studying the effect of C.R. on recidivism for cases with different risk evolution curves prior to their release. This can in principle help in treatment assignment decisions, if current reductions in risk are found to be correlated with future risk reductions.

In addition to the statistical causal inference methods we used in this study, other causal inference methods can also be used to measure the effect of a treatment on an outcome, such as instrumental variables (Angrist et al., 1996) and regression discontinuity (Thistlethwaite & Campbell, 1960).

Other risks that are relevant for the management of prisons can also be investigated using the methodologies we used in this study. These include self-directed violence, violence to other inmates or prison staff, and risk of breaking prison permits.

Appendix

A *RisCanvi* risk items

Risk items of the two *RisCanvi* versions (*RisCanvi*-C and *RisCanvi*-S) are shown on Table 14 (Andrés-Pueyo et al., 2018). There are 6 items which are common in both versions.

A.1 *RisCanvi* items imputation

The number of items in *RisCanvi*-S (10 items) is less than *RisCanvi*-C (43 items). However, 6 items of *RisCanvi*-S match 6 items in *RisCanvi*-C and the remaining 4 items are combinations of other *RisCanvi*-C items. To have 43 numbers of items, which is also more informative, in cases with only *RisCanvi*-S evaluation (these are low-risk cases), we imputed the 33 remaining items using low risk values of *RisCanvi*-C dynamic items and values of penitentiary or demographic features in case of static items. For the cases with *RisCanvi*-C as the latest valid evaluation (which is at most 9 months before the release date), if there is a valid *RisCanvi*-S evaluation before that, we use its non-empty items to impute the missing items in the *RisCanvi*-C.

Table 14 *RisCanvi* risk items

RisCanvi Complete items (S = shared with Screening)
(1) Violent base offense
(2) Age at the time of the base offense
(3) Intoxication during performing the base offense
(4) Victims with injuries
(5) Length of criminal convictions
(6) Time served in prison
(7) History of violence (S)
(8) Start of the criminal or violent activity (S)

Table 14 continued

RisCanvi Complete items (S = shared with Screening)
(9) Increase in frequency, severity and diversity of crimes
(10) Conflict with other incarcerated persons
(11) Failure to accomplishment of penal measures
(12) Disciplinary reports
(13) Escape or absconding
(14) Grade regression
(15) Breaching prison permit
(16) Poor childhood adjustment
(17) Distance from residence to prison
(18) Educational level
(19) Problems related with employment
(20) Lack of financial resources (S)
(21) Lack of viable plans for the future
(22) Criminal history of family or parents
(23) Difficulties in the socialization or development in the origins family
(24) Lack of family or social support (S)
(25) Criminal or antisocial friends
(26) Member of social vulnerable groups
(27) Relevant criminal role
(28) Gender violence victims (only women)
(29) Responsibility for the care of family
(30) Drug abuse or dependence
(31) Alcohol abuse or dependence
(32) Severe mental disorder
(33) Sexual promiscuity and/or paraphilia
(34) Limited response to psychological and/or psychiatric treatments (S)
(35) Personality disorder related to anger or violent behaviour
(36) Poor stress coping
(37) Self-injury attempts or behaviour (S)
(38) Pro criminal or antisocial attitudes
(39) Low mental ability
(40) Recklessness
(41) Impulsiveness and emotional instability
(42) Hostility
(43) Irresponsibility
Other *RisCanvi* Screening items
(1) Institutional/prison misconduct
(2) Escapes or breaches of permits
(3) Problems with drugs or alcohol use
(4) Hostile or pro criminal attitudes

B Features

List of features used in this study is shown in three categories of demographics, penitentiary features, and *RisCanvi* features on Table 15.

Table 15 List of features. "Y/N" are boolean features, and "Num" are numerical features

Demographics	Penitentiary features	*RisCanvi* features
Age at release (Num)	Prison center {1,...,87} (Y/N)	Items {1,...,43}[3] (Num)
Male (Y/N)	No permission request (Y/N)	Risk level (Num) in:
Female (Y/N)	Permission rejection (Y/N)	-Self-directed violence
Country birth (Y/N)	Enjoyed permission (Y/N)	-Violence in prison facilities
Country residence (Y/N)	Not enjoyed accepted permission (Y/N)	-Violent recidivism
Country nationality (Y/N)	No. of enjoyed permissions (Num)	-Breaking prison permits
Single (Y/N)	No. of rejected permissions (Num)	Risk score (Num) in:
Married (Y/N)	Activity (Y/N)	-Self-directed violence
Divorced (Y/N)	No. of activities (Num)	-Violence in prison facilities
Separated (Y/N)	No. of module changes (Num)	-Violent recidivism
Widow (Y/N)	No. of nursing modules (Num)	-Breaking prison permits
Deported (Y/N)	No. of psychiatry modules (Num)	
Country language 1 (Y/N)	No. of special supervision module (Num)	
Country language 2 (Y/N)	No. of regressions to 1st degree (Num)	
Education level (Num)	No. of regressions to 2nd degree (Num)	
Nationality {1,...,8} (Y/N)	No. of progresses to 2nd degree (Num)	
	No. of progresses to 3rd degree (Num)	
	No. of degree change (Num)	
	Mostly degree regression (Y/N)	
	Mostly degree progress (Y/N)	
	No. of degree regression (Num)	
	No. of degree progress (Num)	
	First prison degree (Num)	
	Last prison degree (Num)	
	Degree evolution {0,...,4} (Y/N)	
	Light rules violations (Y/N)	
	Severe rules violations (Y/N)	
	Very severe rules violations (Y/N)	
	No. of light rules violations (Num)	
	No. of severe rules violations (Num)	
	No. of very severe rules violations (Num)	
	Violent base crime (Y/N)	
	With CL[1] (Y/N)	

Table 15 continued

Demographics	Penitentiary features	*RisCanvi* features
	Without CL (Y/N)	
	Fully-paid CL (Y/N)	
	Not fully-paid CL (Y/N)	
	Unable to pay CL (Y/N)	
	Evaluation points (Num)	
	Previous prison entries (Y/N)	
	No. of previous prison entries (Num)	
	Sentence day (Num)	
	Sentence duration class (Num)	
	Penalty admission type {1,2,3} (Y/N)	
	Base crime types {1,...,8}[2] (Y/N)	
	With electronic surveillance (Y/N)	
	In dependent units (Y/N)	

[1] “Civil Liability” (CL) is a monetary compensation imposed in addition to time in jail. There are two basic cases: civil liability (with CL) and no civil liability (without CL). The former is further divided into sub-classes including whether civil liability was paid in full (fully-paid CL), not paid in full (not fully-paid CL), or the person declared him/herself unable to pay

[2] Against people, Gender-based violence, Against sexual freedom, Against property (violent), Against property (non-violent), Drugs, Traffic, and others

[3] The items are listed on Table 14

C ATE results from IPW and PSM methods

The Average Treatment Effect (ATE) obtained from Propensity Score Matching (PSM) are shown on Table 16 and Table 17 for men and women respectively. Also, the ATE values computed using Inverse-Propensity score Weighting (IPW) are shown on Table 18 and Table 19 for both gender groups.

Table 16 ATE obtained for **men** (3,549 cases) using Propensity Score Matching with four buckets and for all. Negative numbers indicate that the probability of recidivism of those who treated (i.e., with C.R.) is lower

Propensity	P(T|X)=low		P(T|X)=med-low		P(T|X)=med-high		P(T|X)=high		All	
Treatment size	0.2%		7.6%		32.8%		59.3%		35.8%	
ATE of C.R. on Recidivism type:	general	violent	general	violent	general	violent	general	violent	general	violent
within 2 years	-0.21	-0.07	-0.08	-0.02	-0.04	-0.01	-0.02	-0.01	-0.11	-0.04
within 3 years	-0.26	-0.09	-0.11	-0.03	-0.05	-0.02	-0.02	-0.01	-0.14	-0.05
within 4 years	-0.31	-0.10	-0.12	-0.03	-0.05	-0.01	-0.01	-0.01	-0.16	-0.05
within 5 years	-0.34	-0.11	-0.14	-0.03	-0.05	-0.01	-0.03	-0.01	-0.17	-0.05

Table 17 ATE obtained for **women** (310 cases) using Propensity Score Matching with four buckets and for all. Negative numbers indicate that the probability of recidivism of those who treated (i.e., with C.R.) is lower

Propensity	P(T|X)=low		P(T|X)=med-low		P(T|X)=med-high		P(T|X)=high		All	
Treatment size	0.6%		19.9%		34.6%		44.9%		50.3%	
ATE of C.R. on Recidivism type:	general	violent	general	violent	general	violent	general	violent	general	violent
within 2 years	-0.24	-0.05	-0.02	0.00	-0.09	0.00	0.00	0.00	-0.13	-0.03
within 3 years	-0.29	-0.07	0.01	0.00	-0.14	0.00	0.00	0.00	-0.16	-0.03
within 4 years	-0.33	-0.07	-0.03	-0.02	-0.14	0.00	0.00	0.00	-0.19	-0.04
within 5 years	-0.37	-0.08	-0.03	-0.02	-0.18	0.00	0.01	0.00	-0.21	-0.05

Table 18 IPW results estimating ATE and its 95% confidence interval [lower-ci, upper-ci]. All of the confidence intervals lie entirely in the negative region

MEN	IPW on general recidivism			IPW on violent recidivism		
(3,549 cases)	lo-ci	ATE	up-ci	lo-ci	ATE	up-ci
within 2 years	-0.09	-0.08	-0.06	-0.03	-0.02	-0.01
within 3 years	-0.11	-0.09	-0.07	-0.04	-0.03	-0.02
within 4 years	-0.13	-0.10	-0.07	-0.04	-0.03	-0.01
within 5 years	-0.14	-0.11	-0.08	-0.04	-0.02	-0.002

Table 19 IPW results estimating ATE and its 95% confidence interval [lower-ci, upper-ci]. All of the confidence intervals lie entirely in the negative region

WOMEN	IPW on general recidivism			IPW on violent recidivism		
(310 cases)	lo-ci	ATE	up-ci	lo-ci	ATE	up-ci
within 2 years	-0.12	-0.08	-0.04	-0.03	-0.01	-0.0003
within 3 years	-0.16	-0.10	-0.05	-0.03	-0.02	-0.002
within 4 years	-0.18	-0.12	-0.07	-0.04	-0.02	-0.004
within 5 years	-0.20	-0.14	-0.08	-0.05	-0.03	-0.01

C.1 ATE-IPW by risk level

ATE of Conditional Release (C.R.) on violent recidivism of men (within 2 to 5 years of release) obtained from IPW is shown for three different REVI risk levels on Table 20.

Table 20 ATE-IPW of C.R. on violent recidivism and its 95% confidence interval [lower-ci, upper-ci] in different REVI risk levels of **men**. Confidence intervals shown in italics contain zero value and are not reliable. Granting C.R. to men with medium REVI risk yields a stronger reduction in violent recidivism risk compared to low REVI risk cases

	IPW on violent recidivism within 2 years			IPW on violent recidivism within 3 years		
REVI level	lo-ci	ATE	up-ci	lo-ci	ATE	up-ci
Low (2,594)	-0.02	-0.01	-0.002	-0.02	-0.01	-0.01
Medium (609)	-0.08	-0.06	-0.04	-0.10	-0.08	-0.05
High (346)	*-0.14*	*-0.04*	*0.06*	*-0.16*	*-0.06*	*0.04*
	IPW on violent recidivism within 4 years			IPW on violent recidivism within 5 years		
REVI level	lo-ci	ATE	up-ci	lo-ci	ATE	up-ci
Low (2,594)	*-0.03*	*-0.003*	*0.02*	*-0.02*	*0.001*	*0.02*
Medium (609)	-0.12	-0.09	-0.06	-0.12	-0.09	-0.06
High (346)	*-0.17*	*-0.07*	*0.03*	*-0.18*	*-0.08*	*0.03*

Funding Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature.

References

Statistics, Northern Ireland and Research Agency. (2019). *The enhanced combination order October 2015 to November 2018*. Northern Ireland: Tech. rep.

Aebi, M.F., Cocco, E., Molnar, L., Tiago, M.M. (2022). Prisons and prisoners in Europe 2021: Key findings of the SPACE I report. https://wp.unil.ch/space/files/2022/05/Aebi-Cocco-Molnar-Tiago2022Prisons-and-Prisoners-in-Europe-2021Key-Findings-SPACE-I-220404.pdf.

Alberts, C.J., Dorofee, A.J. (2003). Managing information security risks: The OCTAVE approach. Addison-Wesley Professional.

Allen, R. D., et al. (2006). Auditor risk assessment: Insights from the academic literature. In: *Accounting horizons,20*(2), pp. 157–177.

Andersen, S. N., & Telle, K. (2022). Better out than in? The effect on recidivism of replacing incarceration with electronic monitoring in Norway. In: *European journal of criminology,19*(1), pp. 55–76.

Andrés-Pueyo, A., Arbach-Lucioni, K., Redondo, S. (2018). The RisCanvi: A new tool for assessing risk for violence in prison and recidivism. In: *Recidivism risk assessment: A handbook for practitioners*, pp. 255-268.

Andrews, D. A., & Bonta, J. (2010). Rehabilitating criminal justice policy and practice. In: *Psychology, Public Policy, and Law,16*(1), pp. 39.

Andrews, D.A., Bonta, J. (2000). The level of service inventory-revised. Multi-Health Systems Toronto.

Anenberg, S. C., et al. (2016). Survey of ambient air pollution health risk assessment tools. In: *Risk analysis,36*(9), pp. 1718–1736.

Angrist, J. D., Imbens, G. W., & Rubin, D. B. (1996). Identification of causal effects using instrumental variables. In: *Journal of the American statistical Association,91*(434), pp. 444–455.

Athey, S. (2015). Machine learning and causal inference for policy evaluation. In: *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 5-6.

Athey, S., & Wager, S. (2019). Estimating treatment effects with causal forests: An application. In: *Observational studies,5*(2), pp. 37–51.

Athey, S., Tibshirani, J., & Wager, S. (2019). Generalized random forests. In: *The annals of statistics,47*(2), pp. 1148–1178.

Austin, J. (2006). How much risk can we take-The misuse of risk assessment in corrections. In: *Fed probation,70*, pp. 58.

Barabas, C., et al. (2018). Interventions over predictions: Reframing the ethical debate for actuarial risk assessment. In: *Conference on fairness, accountability and transparency*. PMLR, pp. 62-76.

Becker, G.S. (1968). Crime and punishment: An economic approach. In: *The economic dimensions of crime*. Springer, pp. 13-68.

Beckett, K., & Sasson, T. (2003). *The politics of injustice: Crime and punishment in America*. Sage Publications.

Beck, A. J., & Shipley, B. E. (1989). *Recidivism of prisoners released in 1983*. Office of Justice Programs, Bureau of Justice: US Department of Justice.

Benko, J. (2018) The radical humaneness of Norway Halden prison: The goal of the Norwegian Penal system is to get inmates out of it. In: *NY times mag*. Retrieved 10.

Bhuller, M., et al. (2020). Incarceration, recidivism, and employment. In: *Journal of political economy,128*(4), PP. 1269–1324.

Bonta, J. (1996). *Risk-needs assessment and treatment*. In

Bonta, J., & Andrews, D. A. (2007). Risk-need-responsivity model for offender assessment and rehabilitation. In: *Rehabilitation,6*(1), PP. 1–22.

Bray, B. C., et al. (2019). Inverse propensity score weighting with a latent class exposure: Estimating the causal effect of reported reasons for alcohol use on problem alcohol use 16 years later. In: *Prevention science,20*(3), pp. 394–406.

Brennan, T., Dieterich, W., & Ehret, B. (2009). Evaluating the predictive validity of the COMPAS risk and needs assessment system. In: *Criminal justice and behavior,36*(1), pp. 21–40.

Caliendo, M., & Kopeinig, S. (2008). Some practical guidance for the implementation of propensity score matching. In: *Journal of economic surveys,22*(1), pp. 31–72.

Capdevila, M., et al. (2015). Taxa de reincidéncia penitenciária 2014. In.

Chen, M. K., & Shapiro, J. M. (2007). Do harsher prison conditions reduce recidivism? A discontinuity-based approach. In: *American law and economics review,9*(1), pp. 1–29.

Chernozhukov, V., et al. (2016). Double/debiased machine learning for treatment and causal parameters. In: arXiv:1608.00060.

Cho, K., Barnes, C. M., & Guanara, C. L. (2016). Sleepy punishers are harsh punishers: Daylight saving time and legal sentences. In: *Psychological science*,

Cid, J. (2009). Is imprisonment criminogenic? A comparative study of recidivism rates between prison and suspended prison sanctions. In: *European journal of criminology,6*(6), pp. 459–480.

Collins, R. E. (2010). The effect of gender on violent and nonviolent recidivism: A meta-analysis. In: *Journal of criminal justice,38*(4), pp. 675–684.

Cotter, R. (2020). *Length of incarceration and recidivism*. DC, USA: Tech. rep. Washington.

Cullen, F. T., Jonson, C. L., & Nagin, D. S. (2011). Prisons do not reduce recidivism: The high cost of ignoring science. In: *The Prison Journal,91*(3 suppl), pp. 48S-65S.

Cullen, F. T., Fisher, B. S., & Applegate, B. K. (2000). Public opinion about punishment and corrections. *Crime and Justice, 27*, 1–79.

Cullen, F., & Gilbert, K. (2012). *Reaffirming rehabilitation*. Routledge.

Danziger, S., Levav, J., & Avnaim-Pesso, L. (2011). Extraneous factors in judicial decisions. In: *Proceedings of the national academy of sciences,108*(17), pp. 6889–6892.

Dawid, A. P. (1979). Conditional independence in statistical theory. In: *Journal of the royal statistical society: Series B (Methodological),41*(1), pp. 1–15.

Drago, F., Galbiati, R., & Vertova, P. (2009). The deterrent effects of prison: Evidence from a natural experiment. In: *Journal of political economy*, *117*(2), pp. 257–280.

Durlauf, S. N., & Nagin, D. S. (2011). Imprisonment and crime: Can both be reduced?. In: *criminology & public policy*, *10*(1), pp. 13–54.

Dynia, P., & Sung, H.-E. (2000). The safety and effectiveness of diverting felony drug offenders to residential treatment as measured by recidivism. *In: Criminal justice policy review*, *11*(4), pp. 299–311.

Eren, O., & Mocan, N. (2018). Emotional judges and unlucky juveniles. In: *American economic journal: Applied economics,10*(3), pp. 171–205.

Funk, S. J. (1999). Risk assessment for juveniles on probation: A focus on gender. In: *Criminal justice and behavior*, *26*(1), pp. 44–68.

Gendreau, P., et al. (2000). The effects of community sanctions and incarceration on recidivism. In: *Forum on corrections research. Vol. 12. 2*. Correctional Service of Canada, pp. 10-13.

Gilman, A. B., & Walker, S. C. (2020). Evaluating the effects of an adolescent family violence intervention program on recidivism among court-involved youth. In: *Journal of family violence*, *35*(2), pp. 95–106.

Glöckner, A. (2016). The irrational hungry judge effect revisited: Simulations reveal that the magnitude of the effect is overestimated. In: *Judgment and decision making*, *11*(6), pp. 601.

Glynn, A. N., & Quinn, K. M. (2010). An introduction to the augmented inverse propensity weighted estimator. In: *Political analysis*, *18*(1), pp. 36–56.

Green, D. P., & Winik, D. (2010). Using random judge assignments to estimate the effects of incarceration and probation on recidivism among drug offenders. In: *Criminology,48*(2), pp. 357–387.

Gupta, A., Hansman, C., & Frenchman, E. (2016). The heavy costs of high bail: Evidence from judge randomization. In: *The journal of legal studies*, *45*(2), pp. 471–505.

Hanson, R. K. (2005). Twenty years of progress in violence risk assessment. In: *Journal of interpersonal violence*, *20*(2), pp. 212–217.

Harding, D. J., & Harris, H. M. (2020). After prison: Navigating adulthood in the shadow of the justice system. *Russell Sage Foundation*.

Harding, D. J., Morenoff, J. D., et al. (2017). Short-and long-term effects of imprisonment on future felony convictions and prison admissions. In: *Proceedings of the national academy of sciences*, *114*(42), pp. 11103–11108.

Hare, R.D. (2003). Psychopathy checklist-revised. In: *Psychological assessment*.

Henneguelle, A., Monnery, B., & Kensey, A. (2016). Better at home than in prison? The effects of electronic monitoring on recidivism in France. In: *The journal of law and economics,59*(3), pp. 629–667.

Heyes, A., & Saberian, S. (2019). Temperature and decisions: evidence from 207,000 court cases. In: *American economic journal: Applied economics,11*(2), pp. 238–65.

Hjalmarsson, R, Lindquist, M.J. (2020). The health effects of prison. In.

Huebner, B. M., DeJong, C., & Cobbina, J. (2010). Women coming home: Long-term patterns of recidivism. In: *Justice quarterly,27*(2), pp. 225–254.

Joffe, M. M., & Rosenbaum, P. R. (1999). Invited commentary: Propensity scores. In: *American journal of epidemiology,150*(4), pp. 327–333.

Jolliffe, D., & Hedderman, C. (2015). Investigating the impact of custody on reoffending using propensity score matching. In: *Crime & delinquency,61*(8), pp. 1051–1077.

Karimi-Haghighi, M., Castillo, C. (2021). Efficiency and fairness in recurring data-driven risk assessments of violent recidivism. In: *Proceedings of the 36th annual acm symposium on applied computing*, pp. 994-1002.

Kehl, D.L., Kessler, S.A. (2017). Algorithms in the criminal justice system: Assessing the use of risk assessments in sentencing. In.

King, R. S., & Elderbroom, B. (2014). *Improving recidivism as a performance measure*. DC: Urban Institute Washington.

Kirk, D. S. (2020). *Home free: Prisoner reentry and residential change after hurricane Katrina*. USA: Oxford University Press.

Kleinberg, J., et al. (2018). Human decisions and machine predictions. In: *The quarterly journal of economics,133*(1), pp. 237–293.

Kuriakose, M. (2019). Recidivism: Effect of incarnation length on offenders: A review of literature. In: *International journal of indian psychlogy*, *7*(2), Lappi-Seppälä, T. (2012). Penal policies in the Nordic countries 1960–2010. In: *Journal of scandinavian studies in criminology and crime prevention*, *13*(sup1), pp. 85–111.

Latessa, E. J., et al. (2010). The creation and validation of the Ohio Risk Assessment System (ORAS). In: *Fed. probation*, *74*, pp. 16.

Link, B. G., & Phelan, J. C. (2001). Conceptualizing stigma. In: *Annual review of sociology*, *27*(1), pp. 363–385.

Loeffler, C. E. (2013). Does imprisonment alter the life course? Evidence on crime and employment from a natural experiment. In: *Criminology,51*(1), pp. 137–166.

Loeffler, C. E., & Grunwald, B. (2015). Processed as an adult: A regression discontinuity estimate of the crime effects of charging nontransfer juveniles as adults. In: *Journal of research in crime and delinquency,52*(6), 890–922.

Loeffler, C. E., & Nagin, D. S. (2022). The impact of incarceration on recidivism. *Annual review of criminology, 5*, 133–152.

Lowenkamp, C. T. (2009). The development of an actuarial risk assessment instrument for US pretrial services. In: *Fed. probation*, *73*, 33.

Marchese di Beccaria, C. (1819). An essay on crimes and punishments. 47183. Philip H. Nicklin.

Marie, O. (2009). The best ones come out first! Early release from prison and recidivism a regression discontinuity approach. Tech. rep.

Mears, D. P., & Bales, W. D. (2009). Supermax incarceration and recidivism. *Criminology, 47*(4), 1131–1166.

Meier, A., Levav, J., Meier, S. (2009). Early release and recidivism. In.

Meredith, T., Speir, J. C., & Johnson, S. (2007). Developing and implementing automated risk assessments in parole. In: *Justice research and policy*, *9*(1), pp. 1–24.

Miron, M., et al. (2021). Evaluating causes of algorithmic bias in juvenile criminal recidivism. In: *Artificial intelligence and law,29*(2), pp. 111–147.

Mitchell, O., et al. (2017). The effectiveness of prison for reducing drug offender recidivism: A regression discontinuity analysis. In: *Journal of experimental criminology*, *13*(1), pp. 1–27.

Monahan, J., & Skeem, J. L. (2016). Risk assessment in criminal sentencing. *Annual review of clinical psychology, 12*, 489–513.

Monnery, B., Wolff, F.-C., & Henneguelle, A. (2020). Prison, semiliberty and recidivism: Bounding causal effects in a survival model. In: *International review of law and economics*, *61*, 105884.

Mueller-Smith, M. (2015). The criminal and labor market impacts of incarceration. In: Unpublished Working Paper 18.

Nie, X., & Wager, S. (2021). Quasi-oracle estimation of heterogeneous treatment effects. In: *Biometrika*, *108*(2), pp. 299–319.

Petrich, D. M., et al. (2021). Custodial sanctions and reoffending: A meta-analyticreview. In: *Crime and justice*, *50*(1), pp. 353–424.

Pettit, B., & Gutierrez, C. (2018). Mass incarceration and racial inequality. In: *American journal of economics and sociology*, *77*(3–4), pp. 1153–1182.

Raphael, S., & Stoll, M. A. (2009). *Do prisons make us safer?: The benefits and costs of the prison boom*. Russell Sage Foundation.

Raz, T., & Michael, E. (2001). Use and benefits of tools for project risk management. In: *International journal of project management*, *19*(1), pp. 9–17.

Rhodes, W., et al. (2018). Relationship between prison length of stay and recidivism: A study using regression discontinuity and instrumental variables with multiple break points. In: *Criminology & public policy*, *17*(3), pp. 731–769.

Roberts, J.V. (2004). The virtual prison: Community custody and the evolution of imprisonment. Cambridge University Press.

Ropes Berry, K., et al. (2020). The intersectional effects of race and gender on time to reincarceration. In: *Justice quarterly37*(1), pp. 132-160.

Rosenbaum, P.R., Rubin, D.B. (1983). The central role of the propensity score in observational studies for causal effects. In: *Biometrika 70*(1), pp. 41-55.

Singh, J. P., et al. (2018). *Handbook of recidivism risk/needs assessment tools*. John Wiley & Sons.

Skeem, J., Monahan, J., Lowenkamp, C. (2016) Gender, risk assessment, and sanctioning: The cost of treating women like men. In: *Law and human behavior40*(5), p. 580.

Sondhi, A., Leidi, A., Best, D. (2020). Estimating a treatment effect on recidivism for correctional multiple component treatment for people in prison with an alcohol use disorder in England. In: *Substance abuse treatment, prevention, and policy 15*(1), pp. 1-12.

Spelman, W. (2020). The limited importance of prison expansion. In: *Crime, inequality and the State*. Routledge, pp. 150-164.

Thistlethwaite, D.L., Campbell, D.T. (1960). Regression-discontinuity analysis: An alternative to the ex post facto experiment. In: *Journal of Educational psychology51*(6).

Tobón, S. (2020). Do better prisons reduce recidivism? Evidence from a prison construction program. In: *The review of economics and statistics*, pp. 1-47.

Tonry, M. (2014). Why crime rates are falling throughout the Western world. In: *Crime and justice 43*(1), pp. 1-63.

Travis, J., Western, B., Redburn, F.S. (2014). *The growth of incarceration in the United States: Exploring causes and consequences*. In.

Tsiatis, A.A. (2006). *Semiparametric theory and missing data*. In.

VanderWeele, T. (2015). *Explanation in causal inference: Methods for mediation and interaction*. Oxford University Press.

Vansteelandt, S., Daniel, R.M. (2014). On regression adjustment for the propensity score. In: *Statistics in medicine33*(23), pp. 4053-4072.

Vass, A.A. (1990). *Alternatives to prison: Punishment, custody and the community*. Sage London.

Velázquez, A. G. T. (2018). *The changing state of recidivism: Fewer people going back to prison*. US: Tech. rep.

Weaver, V.M. (2007). Frontlash: Race and the development of punitive crime policy. In: *Studies in American political development21*(2), pp. 230-265.

Western, B. (2018). *Homeward: Life in the year after prison*. Russell Sage Foundation.

Williams, J., Weatherburn, D. (2022). Can electronic monitoring reduce reoffending? In: *Review of economics and statistics104*(2), pp. 232-245.

Wright, K.N., Clear, T.R., Dickson, P. (1984). Universal applicability of probation risk-assessment instruments: A critique. In: *Criminology22*(1), pp. 113-134.

Yukhnenko, D., et al. (2019). Recidivism rates in individuals receiving community sentences: A systematic review. In: *PloS one14*(9), e0222495.

Zapryanova, M. (2020). The effects of time in prison and time on parole on recidivism. In: *The journal of law and economics63*(4), pp. 699-727.

Marzieh Karimi-Haghighi PhD, is a computer science graduate from the Department of Technologies of Information and Communication at Universitat Pompeu Fabra in Barcelona. Her research interests include improving ML-based risk assessment in complex data settings in terms of algorithmic fairness and ML-supported causal inference effect of an intervention on a risk outcome.

Carlos Castillo PhD, is an ICREA Research Professor at Universitat Pompeu Fabra in Barcelona, where they lead the Web Science and Social Computing research group. Their background is information retrieval, and currently they focus on algorithmic fairness and crisis informatics.

Songül Tolan PhD, is an economist at the German Federal Ministry of Employment and Social Affairs, dealing with labor and social policy related questions of the green and digital transition. Her research interests include the impact of AI on the labor market, regulation, and social policies.

Kristian Lum PhD, is a research associate professor at the Data Science Institute of the University of Chicago. Her research includes the development of new statistical methods and open source software packages for fairness, record linkage, and population estimation. Her main focus is on fairness, accountability and transparency.