Replicating 'Bias Behind Bars' under Bayesian setting*

Logistic regression on sssessment scores of Canadian inmates under Bayesian setting

Yi Su

22 December 2020

Abstract

Racial discrimination against the First Nations has been a problem in Canada since the first colonist step on the sides of the St. Lawrence river. While this problem has been controversial among the general public, is it still a possible case in the scope of prison systems? The Globe and Mail investigation indicates that Black and Indigenous inmates experience worse assessment scores solely based on their race. In this report, we focus on replicating the results of the Globe's investigation under Bayesian settings, instead of the original Frequentist settings, using the same dataset from the Correctional Service of Canada. Although the magnitudes are slightly different, the replicated results agree with the Globe's findings that Black and Indigenous inmates do get worse assessment scores base on their race.
keywords: Logistic Regression, Bias, Racial Discrimination, Correctional Service of Canada, The Globe and Mail, Bayesian

1 Introduction

For the majority of the general public, the part of the world behind bars is likely to be one of the least favorable places to be. Not only because that prisons have a less favorable physical environment, but also because it is filled by heinous criminals. This is, to some extent, a known fact. However, are the prisoners really all that heinous? The answer would be no. Although they all did commit a severe crime to be put in prison, it does not necessarily mean they are all heinous since the severity of their crimes is different. In the prison system, there are various scores, designed by the Correctional Service of Canada (CSC), assessing a criminal's potential threat to inmates of the prison, and also the potential ability to reintegrate to the society after their sentence. These scores are crucial to the inmates since these determine the overall correctional plan and will affect the inmates' life after sentence.

In general, these scores are risk assessments, assessing an inmate's future risk to the public using all past and present information for the design of a correctional plan. The problem arises from these scores, while these measures are crucial to a prisoner's in-prison life and out-prison life, these are not all determined under a standardized and objective measure, thus bias naturally exists when subjectivity enter. Some of these are measured purely by standardized tests and objective data, but some are purely based on the officer's judgment. In case of bias due to subjectivity, it is natural to consider some of the long-lasted biases among the general public like race and gender bias. Racial discrimination against the First Nations has been a problem in North America even before Canada existed as a country, and it would not be a surprise if some degree of discrimination exists among prison systems. The Globe and Mail investigation (Cardoso 2020a) identified that there is a bias against Black and First Nations (indigenous) inmates among the assessments. Through a story-like article, Tom Cardoso explains the potential issue in CSC's assessment systems and the impact of such an issue on an inmate's life both during and after the sentence.

The focus of this report is to replicate the Globe's findings using a Bayesian setting instead of the original Frequentist setting and examine if the findings still hold statistically. We replicate exactly the original procedure using the same dataset originally obtained from CSC (Canada 2018). However, instead of

*Code available at: https://github.com/YiSu2000/Replicate-Bias-Behind-Bars

replicating all three logistic models, we only replicate the first two models analyzing the impact of race on security and reintegration scores. Reproducibility of an article is an important factor for the article to be persuasive to its audience, and it can be assessed through replicating its results.

Although the original codes are not for public shares, Tom did post a methodology instruction article titled "How we did it" (Cardoso 2020b). This article introducing the original methodology of "Bias Behind Bars" is not a detailed instruction paper and there are many hidden details to the exact methodology. When encountering such an issue, we proceed using our interpretation of the context of the procedure. The final result is slightly different from the article due to these issues, as well as the use of Bayesian settings. However, the result should agree with the article that there is some bias against Black and Indigenous inmates on CSC assessment scores. We will discuss more detail on the hidden detail interpretation issues in the sections that encounter it. This report validates the reproducibility of the first two statistical models discussed in "Bias Behind Bars". In this report, "Bias Behind Bars" is referred to as the article, and "How we did it" is referred to as the methodology instruction.

In order to replicate the result of the Globe (Cardoso 2020b), we focus on two of the assessment scores, the offender security level, and reintegration level, for having the largest impact on an inmate's life behind the bars. Both of these scores are assessed by the CSC officers using a series of specialized tests and interviews, which are all assigned when the inmates first arrive at the federal prison. The offender security level is a set of measures from minimum, medium, to maximum, and maximum bring the 'worst' score. Each inmate with the following measure will be assigned to a facility or area suiting their security level, for example, an inmate with a maximum score will be assigned to an area with a maximum security level. The reintegration score estimates an inmate's potential to re-enter society without committing a new offend, and this score is also crucial for the parole hearings. The reintegration is designed to have low, medium, and high levels, assigned using a set of actuarial and non-actuarial assessments, and low bring the 'worst' score. A more detailed discussion of the scores is included in *Section 2.1 & 2.2*.

For the remaining part of the report, *Section 2* includes a discussion on the CSC dataset and some exploratory data analysis for a deeper understanding of the dataset. The specific procedure of the investigation by the Globe and the logistic models used are included in *Section 3*. The model results and comparisons to the original results are included in *Section 4*. Lastly, we make discuss some potential future improvements to the procedure and draw conclusions base on the comparison results in *Section 5*.

2 The CSC 2012-2018 Data

The dataset we use is a Correctional Service of Canada (CSC) dataset from their inmate's database, and this dataset was requested through several levels of bureaucracy (including consent from the head of CSC) by Tom Cardoso, the author of "Bias Behind Bars" (Cardoso 2020a). The dataset records Canadian inmate entries within the CSC database from 2012 to 2018, and for that reason, it can be considered as a panel dataset. It is free of all personal information, containing 744958 entries from 50116 inmates and 25 variables in total (technically it contains 26, but the "judge" variable was set all blank by CSC). Specifically, each year's data given by the CSC is a snapshot of their full database on March 31 of the corresponding year. March 31 is the end of CSC's fiscal year, and for that reason, the unit of year in this dataset is a fiscal year like YE1112 means the fiscal year 2011-2012, instead of the "common" year unit like 2015.

Table 1 displays an example of observations in the CSC dataset, restricting to variables we discuss and use in the context of "Bias Behind Bars" (Cardoso 2020a). All Tables (excluding model results) in this report was made in `R` (R Core Team 2020), using `kable` package (Dowle and Srinivasan 2019) and the extension of it `kableExtra` (Zhu 2020).

The variables recorded include the offender security score and reintegration potential score, which are the two response variables we are interested in. As well as the characteristic that should be controlled to achieve less biased results, including age, gender, year recorded, sentence type, and static score. The sentence types are separated into determinate and indeterminate (a life sentence), indicating the severity of the crime to some degree. The static score corresponds to the risk level of the offender, it is assigned using Static Factors Assessment, a CSC tool measuring the inmates' past involvement with the criminal justice system (some

are crime records). Thus, a higher static score means the inmate has had a decent record with the criminal justice system in the past, and thus indicates an inmate's criminal history to some extent. Fiscal years and offense ID appear in CSC coding format, which is not unique to each inmate, but to that specific year or offense type. Meanwhile, offender ID and sentence ID are individual-specific for every sentence the offender attempted between the 2012-2018 fiscal years.

Table 1: Sample Observations

Fiscal Year	Sentence ID	Offender Number	RACE	Gender	Age	Jurisdiction	Sentence Type	Offender Security Level	Static	REintegration Potential	Offence ID
YE1112	U40A00014615	82.5071	White	MALE	23	FEDERAL	DETERMINATE	MINIMUM	LOW	HIGH	U40A00073751
YE1112	V30A00004376	642.9544	White	MALE	72	FEDERAL	INDETERMINATE	MINIMUM	MEDIUM	HIGH	V30A00035957
YE1112	V50A00010599	54589.0040	Indigenous	MALE	64	FEDERAL	DETERMINATE	MINIMUM	MEDIUM	MEDIUM	V50A00049146
YE1213	O48101001656	78917.0009	White	MALE	51	FEDERAL	INDETERMINATE	MINIMUM	HIGH	MEDIUM	I42101008580

The article also generated numerical offense severity, by hand-matching over 700 offense types recorded in this dataset with the Uniform Crime Reporting Survey's offense category codes (product of Statistics Canada). In that way, the new offense codes can be matched with Statistics Canada's weighting system for Crime Severity Index (CSI), which represents a numerical measure of severity of the offense. However, this procedure of hand-matching over 700 offense types is not very realistic given the time constraints of this report especially when it was available later in the timeline. Nevertheless, we still thank Rohan Alexander for gathering the complete 2018 CSI weight document from Tom Cardoso (it is a public document, but is not open for download on Statistics Canada). This action will cause our replicate model to have one less control variable to include, which is the most severe offense of an inmate. The resultant issues will be discussed in *Section 5*.

For the remaining of the report, we perform data manipulation, result visualization, and statistical model training in `R` (R Core Team 2020). In the data manipulation steps, the packages `tidyr` (Wickham 2020), `tidyverse` (Wickham et al. 2019), `skimr` (Waring et al. 2020), and `naniar` (Tierney et al. 2020) were used. And all figures presented in this section were produced using the `ggplot2` (Wickham 2016) package in `R`. The replication of data cleaning begins with removing all entries for inmates in provincial jurisdiction. This step was not explicitly step written in the methodology instruction, it was included in a bracket, and we assume it is a necessary step to take. However, we already come to a difference with the article results in the first step, the author mentioned he had 741,738 entries after removing the provincial jurisdictions, but our result showed 741829 entries without missing value in jurisdictions. Fortunately, this is not a major concern, but it rings the bell that differences should be expected in most further steps.

Next, we implement our usual steps when dealing with raw data. We removed the missing values in the variables we are interested in (three sores, age, gender, year, sentence type). This reduces the dataset size to 686540 entries. For the 34 race categories in the dataset, we replicate the article methodology of classifying Indigenous races to "Indigenous" race category, and classify all races other than White, Black, and Indigenous, into "other" race category. This classification allows the model to focus on differences between White, Black, and Indigenous groups. The Indigenous classification reference we used were directly from the race grouping categories in the CSC dataset, while the article used Ontario's Data Standards for the Identification and Monitoring of Systemic Racism, which is a guide for identifying racial disparities in data. Assuming the CSC follows a similar, and ethical classification system, we should not encounter major differences with the article, if any, in the classification of races.

Although not explicitly mentioned in the article, we recognized a significant amount of entries that share the same fiscal year ID, sentence ID, and offender ID, and the only difference between these entries are the offense IDs. The assumption to explain this phenomenon is that when there are multiple offenses in one 'crime action', the CSC system had to put multiple entries in to accommodate that, as offense ID and type only records one offense at a time. This assumption should be a plausible explanation since it is common to see crimes that involve breaking into houses, robbery, sexual assault, and the use of violent force in one crime. After removing all replicated observations, the dataset size is reduced to 142700 entries. In *Section 2.1* and *Section 2.2*, the high degree of similarity in size of the model-specific sub-dataset verifies that this action is appropriate. Even though there might be some differences in the details of judging replicated observations, our method should be similar to the actual procedure of the article to some degree.

Figure 1 shows the distribution of race groups among the 20358 entries each representing a unique inmate in early 2016 ('early' since it's collected at end of 15-16 fiscal year). Indigenous inmates account for 23.67% of inmates in our data, this is slightly less than the article's composition of 25.5% Indigenous inmates in early 2016. The Black inmate proportion of 7.94% is slightly less than the article result of 8.7%, and the percentage of inmates other than Black, White, and Indigenous is 8.66%, which is lower than the article result of 10.1%. These deviations suggest that either our approach classifies too many inmates as White, or something went different during the data cleaning steps. It is more likely to be the latter case since the White inmates remain untouched throughout the race classifications. These differences are expected since the deviation from article results will cumulate throughout the procedure by assuming the specific steps taken to get the article results. By over-classifying more inmates as White, and less in all other race groups, we expect our model estimates to be different, but the significance of the effects should be similar overall.

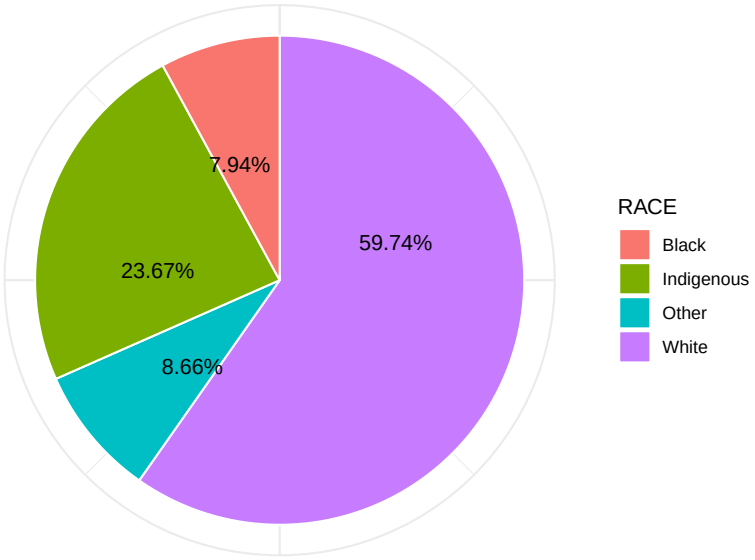

Figure 1: Distribution of race in 2016

Before introducing the model-specific sub-datasets, it is important to check for any empirical results that might suggest any difference between race groups. Figure 2 shows the distribution of offender security scores/levels among the four race groups. In the case of offender security scores, we are interested in the worst level, which is maximum. It can be seen that Indigenous and Black inmates do have a higher proportion of inmates with maximum security scores, 5% points more on average than White inmates. Meanwhile, Indigenous and Black inmates have a significantly less amount of minimum security scores comparing to White and Other inmates, indicating it is also the case that Indigenous and Black inmates have more medium security scores.

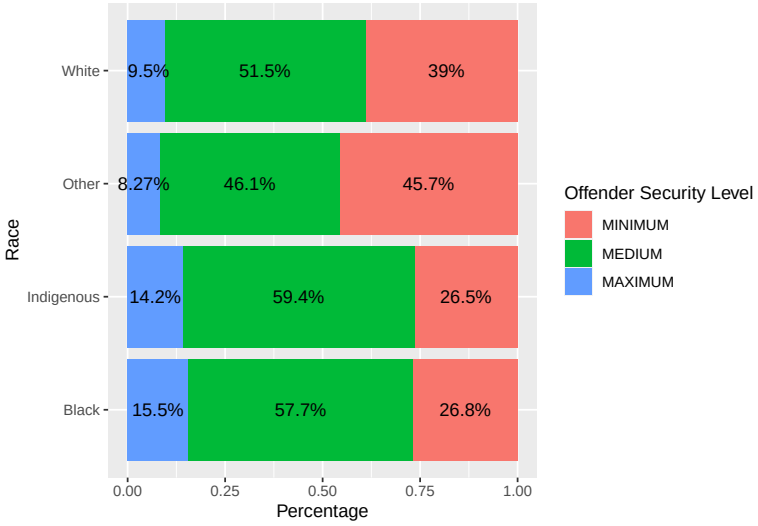

Figure 2: Security score distribution among race

Figure 3 shows the distribution of reintegration potential scores among race groups. The worst level of reintegration score is low, for having the worst potential of reintegrating into society. Figure 3 suggests that Indigenous inmates have a 15% points higher proportion of inmates with low reintegration scores than White inmates, and 15% points less with high reintegration scores so that the distribution of medium score is similar to White inmates. In general, the distributions of reintegration scores are similar between Black inmates and White inmates, suggesting we may not find any evidence that Black inmates are treated differently than White inmates when assessing reintegration potential scores.

Note that these are just naive implication of potential racial discrimination when assessing offender security level, and does not represent anything causal before we regress the scores with race, and many other control variables to get closer to a causal link between the two.

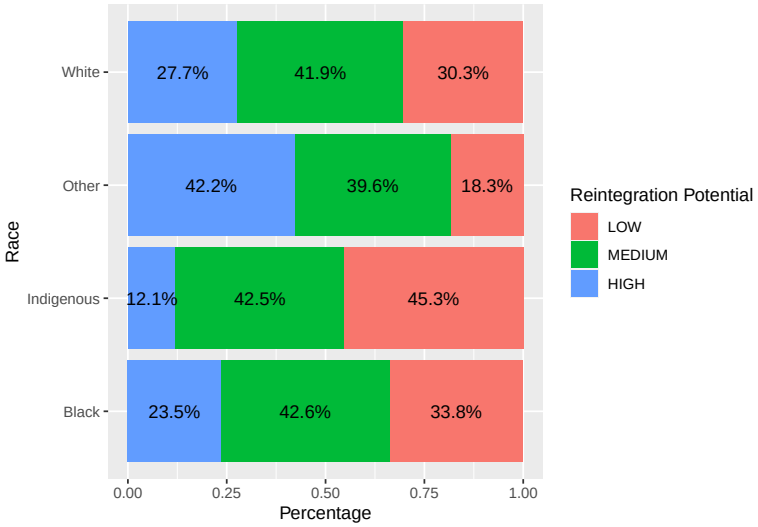

Figure 3: Reintegration score distribution among race

The article 'Bias Behind Bars'(Cardoso 2020a) tested three models in total, two models assessing the impact of race on the probability of getting the worst offender security score and reintegration potential score, and one model assessing the impact of reintegration score on probability of re-offending. Due to the time constraint

of this report and ambiguity in the methodological instructions of the third model, we only replicated the first two models assessing the impact of race on the two assessment scores. For each model, we subset the CSC dataset into sub-datasets since the offender security score assessments are different from reintegration potential scores.

The main issue with the instruction for the third model is how to subset the CSC data. The model should be using the re-offend indicator variable as the response and using reintegration potential as the predictor along with other controls, and the model itself has no ambiguity. Our understanding of the subset for this model was to generate the re-offend indicator dummy variable for all entries based on if the offender appeared in the dataset multiple times with unique sentence IDs. However, the article mentioned that it should have 28110 total entries for this subset. This makes the whole process ambiguous since there are a lot more than 28110 entries if we are to compare all inmates that re-offended and never re-offended. Since the results using the dataset generated with our understanding are far from the article's result (Cardoso 2020a), we excluded the third model and its results.

2.1 Offender Security level Model Subset

As mentioned in the introduction, the offender security level is one of the most important scores affecting an inmate's life behind bars. It has a set of measures from minimum, medium, to maximum, and maximum bring the 'worst' score. The CSC officer uses an actuarial setting questionnaire called the Custody Rating Scales to measure the security level of an inmate when they first arrive at the designated institution. The questionnaire is focused on the severity of the main offense of the current sentence, and any history of drug use or alcohol abuse. Although the security score can be overridden later, the inmate will be transferred to the institution suiting their security level. And the maximum level is considered to be the worst score since an inmate with maximum security level will share a facility in the same institution with other maximum level inmates, which are likely to be murdered or committed other severe crimes and have lower degrees of personal freedom. Meanwhile, the low-security level inmates will be surrounded by inmates committing fraud or other non-violent crimes, and have more freedom relatively.

Since the offender security level was set when they first arrive and is not likely to be over-ride in most cases, the dataset for the security score model will only contain the first entry of each inmate with their unique offender ID. We do so by locating the entry of each inmate with the lowest age and also filtering for inmates that were in custody at the time to eliminate the community service sentences. This is slightly different from the description in the methodology instructions, which was "the dataset looked only at inmates who'd just begun their sentence since that's when they are assessed with the Custody Rating Scale" (Cardoso 2020b). It is unsure if the article meant to filter for the entries when they just enter, or just filter for all new inmates at the beginning of 2018.

The response variable of the model is a dummy variable, generated as 1 if the inmate got a maximum security level, and 0 otherwise. This generation is required by the logistic regression we introduce in *Section 3*. Lastly, we follow the article to have two sub-models for each score, one for females and one for males, and the sub-dataset will be divided by gender for it.

The sub-dataset using our interpretation leaves 36285 entries, each representing a unique inmate, with 34252 male and 2033 female. While the article suggests a sub-dataset with 22922 entries, with 21439 males and 1483 females. We tried to filter for only new inmates with entries recorded in the fiscal years 2017-2018, but that only leaves 3155 unique entries. Thus, we proceed with the sub-dataset with 34352 entries for the offender security level models, specifically the male security level model and the female security level model.

Instead of checking score distribution for males and females using a bar plot similar to Figure 2, we check it using a frequency table. Since the former form does not show count and percentage at the same time, and the count does play an important role for a smaller dataset like the female dataset.

Table 2 shows the distribution of security scores for males. The overall distribution is similar to the complete dataset, but it appears that there is a slightly more percentage of maximum security level inmates combined with a lot more percentage of medium security level inmates for both Black and Indigenous inmates. Resulting

in approximately 10% points less Black and Indigenous inmates to have minimum security level. However, their gap between the White inmates in terms of maximum security level percentage still holds in general.

Table 2: Distribution of Security Score for Male

Race	Offender Security Level	Count	Percentage
White	MINIMUM	5561	27.6%
White	MEDIUM	12405	61.5%
White	MAXIMUM	2213	11%
Indigenous	MINIMUM	1394	17.1%
Indigenous	MEDIUM	5552	68.3%
Indigenous	MAXIMUM	1183	14.6%
Other	MAXIMUM	320	10.7%
Other	MINIMUM	977	32.8%
Other	MEDIUM	1681	56.4%
Black	MAXIMUM	528	17.8%
Black	MEDIUM	1910	64.4%
Black	MINIMUM	528	17.8%

Table 3 shows the distribution of security scores for females. There is two major difference between the security score of female distribution to the complete dataset distribution. Only 3.85% (6 individuals) of inmates in Other race groups were assigned with a maximum security level, this is significantly lower than the former 8.27%. And only 8% (12 individuals) of Black inmates were assigned with a maximum security level, again, significantly lower than the former 15.5%. These are the drawbacks of dividing the dataset by gender since the sample size of females in the CSC dataset is naturally less than males. In this case, there are only a total of 150 Black female inmates for the security score model, and its distribution is not much similar even to the male sub-dataset. Thus, the female security score model can be expected to have a different result than the male model, and it may not be statistically significant due to the relatively small sample size under Bayesian settings.

Table 3: Distribution of Security Score for Female

Race	Offender Security Level	Count	Percentage
White	MINIMUM	452	44.1%
White	MEDIUM	491	47.9%
White	MAXIMUM	83	8.09%
Indigenous	MEDIUM	413	58.9%
Indigenous	MINIMUM	190	27.1%
Indigenous	MAXIMUM	98	14%
Other	MEDIUM	62	39.7%
Other	MINIMUM	88	56.4%
Other	MAXIMUM	6	3.85%
Black	MEDIUM	42	28%
Black	MINIMUM	96	64%
Black	MAXIMUM	12	8%

2.2 Reintegration Potential Model Subset

The reintegration potential score is a measure of the inmate's potential ability to re-enter society after sentence and the likelihood of not re-offending. Leveling from the low, medium, to high, and low is the worst reintegration score indicating a low likelihood of re-entering the society without re-offending. The importance of reintegration potential lies in the parole hearings, a higher reintegration potential helps to gain trust from the parole boards. This score is regularly assessed partially base on Static Factors Assessment, and the other parts depend on the inmate's gender and/or race group. If the inmate is Indigenous or female, the CSC would use the Dynamic Factors Identification and Analysis test, which is a measure of the inmate's life experiences. Otherwise, the CSC uses the Statistical Information on Recidivism scale, which is another measure predicting the likelihood of re-offend. The latter measure was never used on Indigenous or female inmates with the intention to prevent cultural bias. Notice that both the Static Factors Assessment and the

Dynamic Factors Identification assess the severity of past criminal history and life experiences, which are based on the administering officer's subjective judgment and is where the potential issue lies.

Since the reintegration potential scores are assessed each year throughout the sentence, the sub-dataset for the reintegration score model will naturally contain more entries since we are allowing multiple entries for each inmate. Since we already eliminated replicated observations, the sub-dataset was defined by simply filtering for all inmates in custody to eliminate all community service status. The resultant dataset contains 91787 entries, 87909 males and 3878 females. Comparing to the article's size of 90524 entries, 86762 males and 3762 females, our dataset is closer to the desired size than previous attempts. Similar to the previous section, the response dummy variable was generated to be 1 for having low reintegration potential, and 0 otherwise.

Table 4 shows the distribution of reintegration scores among males in the model-specific dataset. It appears that inmates with high reintegration scores dropped by approximately 10% points on average for all race groups, increasing the medium and low scores' share. This could be the result of keeping only inmates in custody who may have lower reintegration potential on average, and those excluded inmates in community status may have higher integration potential scores in general.

Table 4: Distribution of Reintegration Potential for Male

Race	Reintegration Potential	Count	Percentage
White	HIGH	6673	13.1%
White	LOW	22443	44.1%
White	MEDIUM	21815	42.8%
Indigenous	HIGH	1107	4.98%
Indigenous	LOW	12687	57%
Indigenous	MEDIUM	8445	38%
Other	MEDIUM	3212	46.7%
Other	HIGH	1710	24.8%
Other	LOW	1960	28.5%
Black	LOW	3501	44.6%
Black	MEDIUM	3284	41.8%
Black	HIGH	1072	13.6%

Table 5 shows the distribution of reintegration scores among females in the same dataset. The same pattern of decreasing amount of high reintegration score inmates is observed for White, Indigenous, and Other race groups, but with varying magnitude. White inmates have a reduction of approximately 10% points, 4% points for Indigenous inmates, and 12% points for Other race inmates. However, for Black inmates, a 14% points increase in the proportion of high reintegration potential inmates was observed. This could be due to the small sample size of Black female inmate entries, 261 entries among the 3878 entries. Smaller size led to higher variation within the Black female inmates, and thus unusual pattern from the other race groups, and the Black male inmates.

Table 5: Distribution of Reintegration Potential for Female

Race	Reintegration Potential	Count	Percentage
White	MEDIUM	1110	58.9%
White	HIGH	348	18.5%
White	LOW	426	22.6%
Indigenous	MEDIUM	871	59.7%
Indigenous	LOW	461	31.6%
Indigenous	HIGH	128	8.77%
Other	LOW	43	15.8%
Other	HIGH	81	29.7%
Other	MEDIUM	149	54.6%
Black	MEDIUM	126	48.3%
Black	HIGH	99	37.9%
Black	LOW	36	13.8%

3 Model

The models for both scores share the general form of a binary logistic regression model fitted under Bayesian settings. Binary logistic regression assesses the logit probability of an event to occur, in other words, the log-likelihood of the dummy response variable to equal to 1. Due to the mathematical property of logistic functions which range from 0 to 1 in the response variables, logistic regression is the natural choice to consider when dealing with binary (dummy) response variables for preventing extrapolation in the predicted response values. As mentioned in the previous sections, we created the dummy response variables to assess the log-likelihood of getting the worst offender security score and reintegration potential score using race groups as the main explanatory variable, and including the individual characteristics and other measures from the CSC as control variables.

The regression models are fitted in `R` (R Core Team 2020) using the `brms` package (Bürkner 2017) and `tidybayes` package (Kay 2020). The `brms` package is designed for Bayesian regression models, essentially brings the merit of `Stan` using simple `R` syntax and through similar algorithms like Markov chain Monte Carlo (MCMC). MCMC chains allow the model to develop the parameter gradually, and `brms` uses its variant form which is the No-U-Turn Sampler (NUTS) to adapt the samples for posterior distributions, the default 4 chains and 1000 iteration (not including warm-up) were used to fit the model with the help of `brm` function in `brms`.

The main difference between our Bayesian setting and the article's (Cardoso 2020a) Frequentist setting is on the interpretation of parameters, in other words, how we treat the results assessing the effects of being indifferent race groups. The Frequentist setting interprets the parameters as an unknown constant, which is waiting to be estimated by the logistic regression resultant coefficients. The Bayesian setting interprets the parameters as random variables, and we estimate the distribution of that random variable using Bayesian logistic regressions. In the Bayesian setting, we can set prior belief about the distribution of those random variables which directly affects the estimated distribution (posterior). In this report, we use the default priors of `brms`, which are t distributions gathered automatically by `brms`, because there is not have much prior belief about the distribution of those effects. The equation for the two models is essentially the same except for the response variables since they include the same explanatory variable and control variables.

$$\begin{aligned} Pr(MaxSecurity_{i,t} = 1) = logit^{-1}(\beta_0 + \beta_1 Black_{i,t} + \beta_2 Indigenous_{i,t} + \beta_3 Other_{i,t} + \\ \delta_1 Age_{i,t} + \sum \alpha_j Static_{i,t} + \sum \epsilon_k Year_{i,t} + \delta_2 SentenceType_{i,t}) \end{aligned}$$

$$\begin{aligned} Pr(LowReint_{i,t} = 1) = logit^{-1}(\beta_0 + \beta_1 Black_{i,t} + \beta_2 Indigenous_{i,t} + \beta_3 Other_{i,t} + \\ \delta_1 Age_{i,t} + \sum \alpha_j Static_{i,t} + \sum \epsilon_k Year_{i,t} + \delta_2 SentenceType_{i,t}) \end{aligned}$$

The equations for the two binary logistic regression models are displayed as above. The response variables are dummy variables discussed in *Section 2*, which are equal to 1 if the inmate got the worst level of scores, maximum for security level, and low for reintegration potential. Remember that each model will be fitted twice, one time for male inmates only and one time for female inmates only, each based on the corresponding sub-dataset discussed in *Section 2*. The race is the main explanatory variable we are interested in since we are trying to replicate the results of 'Bias Behind Bars' (Cardoso 2020a). Three dummy variables were set up so that they equal to 1 if the inmate is Indigenous, Black, or Other race respectively, and if one of these dummies equals to 1, the other two equals 0 automatically. White is manually set as the reference group, meaning it's effect is implicitly integrated into the intercept (β_0) to avoid multicollinearity with the other three dummies. Due to this dummy variable coding format, the interpretation of the $\beta_{1,2,3}$ is the average percent difference in the likelihood of getting the worst scores for being Indigenous, Black, or Other race alone, holding all other control variables constant.

Besides the race groups being the main interest of the article and of this report, these choices of control variables are based on the methodology instructions (Cardoso 2020b), but with the exclusion of the variable

representing the weight of the most severe offense of an inmate. As mentioned in the previous sections, the control variables include age, static score, year, and sentence type. Age is a discrete numeric variable, and interpretation of δ_1 is the average increase in the likelihood of getting the worst scores for each additional increase in age, holding all other explanatory variables constant. We did not modify ages into age groups because it is not explicitly mentioned in the methodology instructions, and should not cause much difference in the estimates for the effect of race dummies. The static score is a categorical variable, as discussed in *Section 2*, it levels from the low, medium, to high. The `brms` automatically create dummy variables for categorical variables in a similar format as we did for race groups. However, we did not manually set a specific reference group so `brms` will pick the reference group automatically, which may be different across models. So we wrote all other categorical variables in summation formats along with their corresponding effect parameters. The α_js are the two coefficients for the two non-reference dummy variables, with interpretation similar to the βs.

The year is also a categorical variable representing the fiscal year of the entry, by automatically creating the dummy variables, adding year is essentially adding year fixed effects into the model. Adding time fixed effects into the model allows to control for any variation that varies across years only, but constant across individual/provinces. Again, remember it should have six dummy variables in the summation for the total of seven fiscal years in the CSC dataset. However, the article did not add individual fixed effects and the possible individual specific time trends which are some useful controls to look into. Individual/province fixed effects have a similar interpretation as time fixed effects, it controls for variation across individuals/provinces that's constant across years.

The reason why no individual/province terms were not discussed in the article and the methodology instructions, thus, we did not add the individual fixed effects for replication purposes. It could be the case that no evidence or arguments are supporting that individuals/provinces are different from each other from 2012 to 2018. The ϵ_ks are the six coefficients for the six non-reference dummy variables, with interpretation similar to the βs. For future improvements of the model, adding individual fixed effects and individual-specific year trends should help stabilize the estimated effects in general. The addition of individual fixed effects or even province fixed effects, since province data is recorded in the CSC dataset, should be assessed carefully for any evidence of over-fitting. There are more than 35000 unique inmates across seven years, and the amount of individual dummies to add is significant, which could potentially fracture the posterior distributions of the estimates. In that case, controlling for states is a better choice if there is reason to believe CSC administrative officers are different across provinces, and across years at the same time if one considers adding interaction terms.

Sentence type is a categorical variable with only two levels, determinate or indeterminate, so `brms` will only create one dummy for it. Thus, δ_2 is interpreted as the average difference in the likelihood of getting the worst score between the reference and non-reference sentence types, holding all other explanatory variables constant.

Figure 8, figure 9, figure 10, and figure 11 in the appendix displays the trace plots of the parameters during the MCMC chains of all four fitted models, these plots were created in `R` (**???**) using functions in `brms` (Bürkner 2017) package. An ideal trace plot would be in a consistent and rapid up-and-down shape with no long term trend. All of the trace plots appear to satisfy that requirement, and there are no divergent transitions during the MCMC chains which would be indicated in red by `brms` if any. These findings show evidence that our fitted model is not fractured, and the reliability of these fitted models is assessed in the next section, *Section 4*.

4 Results

The displayed regression results focus on the three main race effects, which is the focus of this report. White inmates are set as the reference group, so its effect is integrated into the intercepts. For each mean estimate of the race effects, the estimated standard error and 95% credible intervals are shown below them. Standard errors measures how 'spread' the distribution of the estimates is around the mean estimate. A lower standard error indicates more concise the distribution is around the mean. The interpretation of the 95% credible intervals is different from the well-known confidence intervals, which can be interpreted as if we keep taking

samples and fitting the model then the true value of the parameter lies in the confidence interval 95% of the time. Credible intervals are Bayesian measures, and it can be interpreted as there's a 95% probability that the magnitude of the race effects lie within the interval.

The main difference between the confidence interval and credible interval is that the bounds of confidence intervals are considered to be random, while the bounds of credible intervals are considered as fixed constants. Notice that this is the difference in the reverse of interpretation of parameters between Frequentist and Bayesian settings. If the 95% credible interval contains zero, then the magnitude of the estimated effect is not statistically significant from zero. Otherwise, the estimated effect would be statistically significant from zero at the 5% significance level, meaning the chance of getting the estimated value is less than 5% if the true distribution of the effect does have zero mean.

The corresponding Receiver Operating Characteristic (ROC) curves for each model are also displayed to replicate all results mentioned in the methodology instruction (Cardoso 2020b). The ROC curves were produced using the `pROC` package (**???**) in `R` (**???**). The ROC curve plots the False Positive Percentage (FPP) and True Positive Percentage (TPP) when using the model to predict back on the sub-dataset for the binary response of getting the worst assessment scores. When making predictions, each point on the ROC represents a threshold value of treating the prediction as 1 or 0. For example, if the threshold is 0.5 when the fitted model uses an entry to predict the response variable and got a prediction of 0.6, it will be treated as 1. The True Positive Rate/Percentage (TPP) is the rate of successfully predicting the response variable, and the False Positive Rate/Percentage (FPP) is the rate of falsely predicting the response variable. The ROC plots these two rates for every possible threshold of identifying the binary response, and an ideal ROC should be as close to the upper left corner as possible for the highest TPP and the lowest FPP.

The Area Under the Curve (AUC) is a numeric measure of the ROC curves, and it can be used as a measure to compare two logistic regression models to find the model with better prediction power. As the name of AUC says, it measures the area under the ROC curve in percentage, 50% means the model is randomly guessing between the binary outcomes, indicating the model has no predictive power. In general, the higher the AUC, the better the model is at predicting the binary response variable. An AUC larger than 70% indicates the model has a solid prediction power, and AUC larger than 90% indicates it is a very good model, but an AUC close to 99.99% will indicate over-fitting since it is unrealistic in most cases.

4.1 Security Socre Model

Table 6 : *Security Level Model Results*

	Male	Female
Black	0.21	0.23
	(0.06)	(0.37)
	(0.09, 0.33)	(-0.50, 0.93)
Indigenous	-0.05	0.15
	(0.04)	(0.18)
	(-0.14, 0.04)	(-0.21, 0.51)
Other	-0.07	-0.74
	(0.07)	(0.49)
	(-0.21, 0.07)	(-1.76, 0.15)

Table 6 displays the results of the logistic regression model on the likelihood of being assigned with the maximum offender security level. For the male model, the estimate suggests that Black male inmates are 21% more likely to get maximum security level than White male inmates holding all other controls constant. This result is statistically significant from zero since the corresponding 95% credible interval does not contain zero. Comparing to the article's (Cardoso 2020a) estimate of 23.8% which is also statistically significant, our result is slightly lower. Same as the article results, we find no statistically significant effect for male Indigenous inmates to be different from White inmates, nor is the magnitude of this effect itself significant. Notice that

the model suggests that male inmates with Other race groups are also indifferent to White inmates when assessing security levels. However, as mentioned in the article, it does not mean there is no effect just because our model not detecting an effect on Indigenous inmates. The only conclusion that can be drawn on the non-significant results is that our model does not provide evidence to support that there is an effect.

Notice that the female security level model disagrees with the article results, all estimates in the female model are not statistically significant for containing zero in the 95% credible intervals. The article results suggest there is a significant effect for Indigenous female inmates, but not for Black female inmates. This disagreement could be a result of having a large standard error in the female model, which makes it harder for the identification of the effect since the posterior distribution is more spread.

Figure 4 and figure 5 display the ROC curves for the male and female security level models, both showing an AUC larger than 70% meaning both are models with good prediction powers. However, unlike the article suggesting the male model have a higher AUC than the female model, the current female model is showing a small amount of extra prediction power compare to the male model. Given that none of the race effects in the female model are statistically significant, this implies that the control variables play a more important role in security level assessments for females than for males. Both ROC curves indicate that an optimal choice of threshold would result in a TPP around 80% with FPP around 25%, which is generally considered as a good amount of trade-off between TPP and FPP.

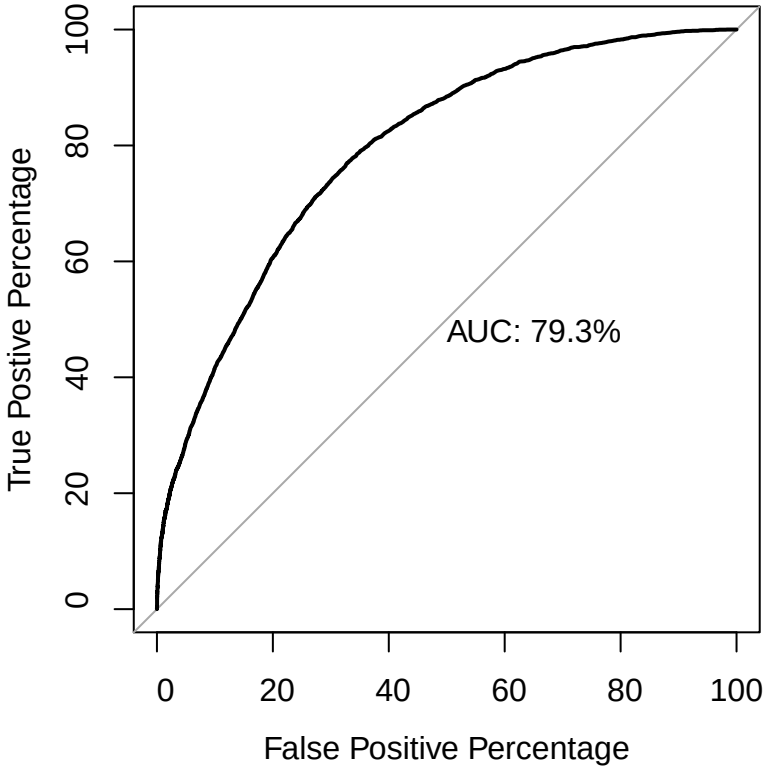

Figure 4: ROC Male Security Level Model

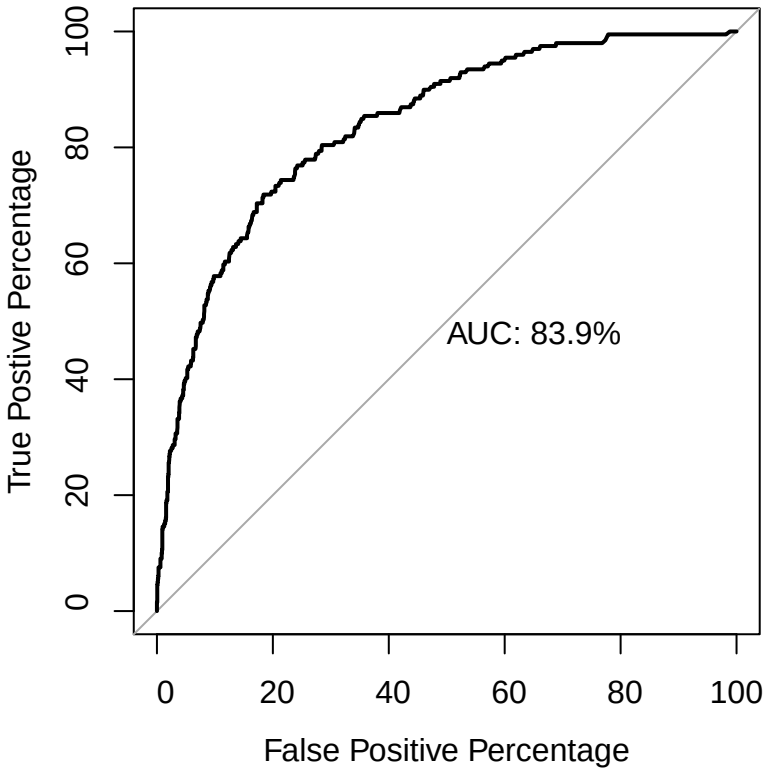

Figure 5: ROC Female Security Level Model

4.2 Reintegration Score Model

Table 7 : *Reintegration Potential Model Results*

	Male	Female
Black	-0.10	-0.34
	(0.03)	(0.22)
	(-0.15, -0.04)	(-0.79, 0.09)
Indigenous	0.45	-0.02
	(0.02)	(0.10)
	(0.41, 0.49)	(-0.21, 0.17)
Other	-0.59	0.09
	(0.03)	(0.21)
	(-0.66, -0.53)	(-0.33, 0.50)

Table 7 shows the results of the logistic regression model on the likelihood of being assigned with the low reintegration potential score. For the male model, the estimate suggests that Indigenous male inmates are 45% more likely to get low reintegration potential score than White male inmates holding all other controls constant. This result is highly statistically significant from zero since the estimated standard error is as low as 0.02, resulting in a 95% credible interval from 0.41 to 0.49. Our estimate is significantly higher than the article suggested 29.5% for Indigenous male inmates. The Black male inmates are actually 10% less likely to get a low reintegration potential score than White male inmates holding all other controls constant, this is also statistically significant. This is larger in magnitude than the suggested 6.1% in the article. Notice that inmates in other races are 59% less likely to get low reintegration potential score than White male inmates holding all other controls constant, and is also statistically significant from zero. This suggests that the CSC

officers might be in favor of those minority groups when assessing their reintegration potential.

Similar to the female model for offender security level, the resultant race effects from the female reintegration potential model are all not statistically significant from zero. Since the article did not discuss the results from the female reintegration potential model, it is assumed that the article did not get any statistically significant results for the female race effects.

Figure 6 and figure 7 display the ROC curves for the male and female reintegration potential models, both showing an AUC around 80% meaning both models have very good prediction power. Unlike the article suggesting the male model have a higher AUC than the female model, the two models for reintegration potential have similar AUC. The male model ROC curve indicates an optimal choice of threshold would result in a TPP around 90% with FPP around 30%. Meanwhile, the female model ROC curve indicates an optimal choice of threshold would result in a TPP around 80% with FPP around 20%. Given that they have similar AUC, this demonstrates the trade-off between TPP and FPP: higher TPP would result in higher FPP. As a metric measure of ROC curves, the AUC is similar for the two ROC curves since their optimal trade-off ratio is similar.

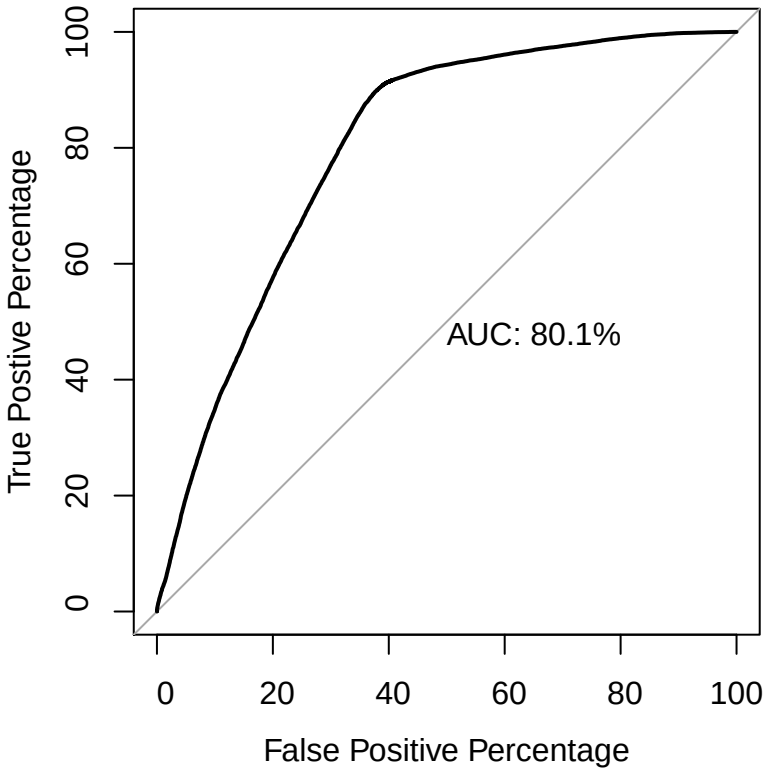

Figure 6: ROC Male Reintegration Model

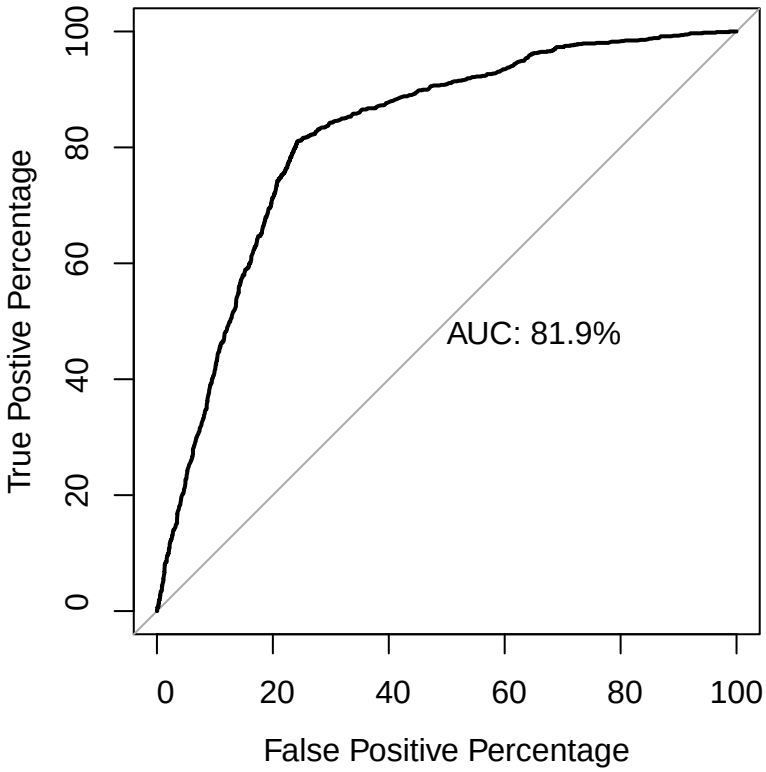

Figure 7: ROC Female Reintegration Model

5 Discussion and Conclusion

Although the magnitudes are slightly different from the suggested results in "Bias Behind Bars" and "How we did it" (the methodology instruction), our Bayesian logistic regression models on male inmates indicate the same overall results in terms of both statistical significance and direction of the effects. This proves the robustness of those effects by achieving highly analogical results even when we used potentially different assumptions in the data manipulation steps and under different statistical model assumptions.

For offender security score assessments, the model suggests that Black inmates were more likely to be assigned with the worst security score and get sent to maximum security facilities or institutions after controlling for all kinds of background variables. The reason behind this may not be necessarily caused by the subjectivity of CSC officers against Black inmates, but a result of social stigma against the Black population in North America in the past century. The CSC officers may decide to exaggerate the potential threat of Black offenders based on their impressions and past experiences.

A research that was done a decade ago, (Bertrand and Mullainathan 2004), showed that employers are more likely to give interview and hire chances to White sounding names rather than Black sounding names in North America, which reflects that the underlying social stigma or race discrimination does exist. Nonetheless, CSC officers were trained against all kinds of racial discrimination, and the suggested results of the male security score model could be seen as an indication that CSC officers are relying too much on past experiences when assessing security levels. This is not saying that they have seen significantly more Black felons, inmates with very severe offenses, than White felons, but to say that CSC officers tend to remember more cases of Black felons because of the underlying social bias against Black inmates.

On the other hand, the male reintegration potential model indicates that Indigenous inmates were treated unfairly during reintegration potential assessments comparing to White inmates, and are more likely to get assigned with low reintegration potential. This could be a result of the special assessment scheme when the

inmate is Indigenous or female. The CSC would use the Dynamic Factors Identification and Analysis test when assessing the reintegration potential of Indigenous or female inmates. The overall non-significant results from the female model with high prediction power implies this possibility from the other side, there might be no racial difference in reintegration potential assessments if all inmates used the same set of measurements. However, this possible explanation cannot be proven without more details on the difference between the Dynamic Factors Identification and Analysis test and the usual procedure for other inmates, which was not discussed in detail in the article and the methodology instructions.

These results follow from the results in "Bias Behind Bars", suggesting that CSC do need to adjust their assessment systems to compensate for the human factor, and potentially adjust the administrative officer training. One viable adjustment could be employing officers with the same race group as the inmate to be assessed, which theoretically eliminates any social bias against certain racial groups. Any improvement in CSC assessments needs time to be verified against racial discrimination and even take another 7 years to gather enough data. With the "Bias Behind Bars" making a huge impact in the world of the criminal justice system, attempts to improve the assessment system by the CSC can be foreseen in the near future.

5.1 Potential Problems for Exluding the Most Severe Offence Variable

In *Section 2*, the article's control variable representing the most severe offense of an inmate was dropped from this report for its heavy workload given our time constraints when the necessary information was finally obtained. Although the model results indicate it does not lead to a severe consequence against the article results and the static scores essentially control for inmate's past offense history, there exist potential problems for not controlling the most severe offense of an inmate.

The most severe offense can be seen as another way of controlling for individual fixed effects by treating the inmates with the same degree of severity together. If this was the intention of Tom, the author of "Bias Behind Bars", then our model should control for the offender IDs or province for fixed effects. The most severe offense should also further stabilize the model results, potentially driving the magnitude of model estimates closer to article estimates even under Bayesian settings.

However, even if the process of hand matching over 700 offense types was done, it is expected to be different from what Tom's hand matched since the detailed offense types require some subjective judgment with hand matching. Tom mentioned in the reply to the email requesting for the CSI weights that, using Neuro-linguistic programming (NLP) could potentially sort the CSC offense types to the CSI weights in a less manual way. But that would be out of our ability and awaits future attempts to practice.

5.2 Next Steps

Besides the potential problems with excluding the most severe crime variable, the main constraint of the current models is that there are no statistically significant results for any race effects in both female score models. The general procedure of increasing an estimated effects' statistical significance is to reduce its estimated standard error for a more concise posterior distribution, thus reducing the chance that the 95% credible interval contains 0. The current estimated standard errors are equal or larger than the mean estimates in magnitude, for example, the mean estimate for the effect of being Black is 0.15, but the estimated standard error is 0.18.

A common measure of reducing estimated standard error for effects is to increase the size of the dataset to be fitted, as demonstrated by the male models with significantly larger dataset size and significantly smaller standard errors for the effects. This is not valid unless the CSC give consent to a larger subset of their database because the current observational dataset size is fixed for both male and females. Note that, the dataset size needs to increase by 4 times the current size to reduce the standard error by 2 times. In general, asking for the CSC to release a couple of decades of entries from their database is not realistic because of the bureaucratic procedures.

Other options include increasing the amount of variation in the explanatory variables and reducing measurement errors. Unfortunately, both options are not applicable, there is no control on the variable of race groups among the female inmates since the dataset is given by the CSC and is fixed for change without a solid

and reasonable argument, and the measurement errors are random in general since the race of an inmate is self-reported. However, if the inmates believe in report themselves as the minority race groups would be in favor of security level and reintegration potential assessments, fixing race groups to what was reported in the first entry for every inmate could effectively reduce measurement error. On the other hand, this action may result in a decrease in variation of race groups within the female sub-dataset so the trade-off between the two results in an unknown effect on estimated standard errors.

It is also unfortunate that the third logistic regression model assessing the likelihood for an inmate to re-offend base on reintegration potential was dropped because of the ambiguous instructions for the corresponding dataset. We encourage future attempts to get in touch with Tom Cardoso, the author of "Bias Behind Bars", for more details on that specific sub-dataset. Including the third model would help demonstrate the social bias against the Black and/or Indigenous population is not true, the article results suggest that given the assignment of low reintegration potential, White inmates are more likely to re-offend than Black and Indigenous inmates. A proof of the robustness of this result could help to remove the social bias furthermore.

Appendix

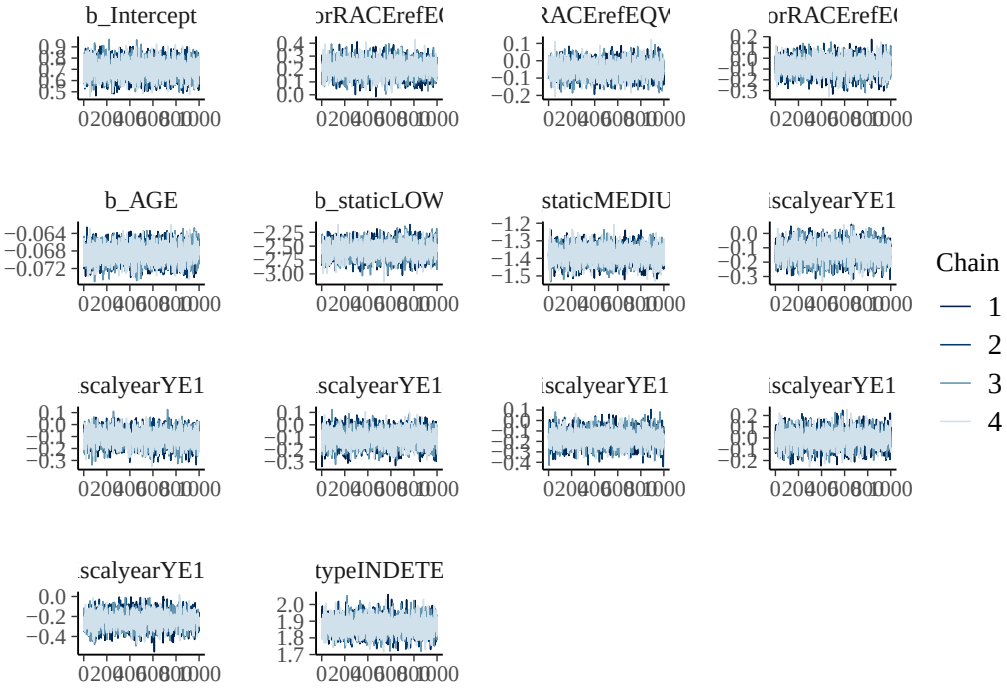

Figure 8: Trace plots for male security level model

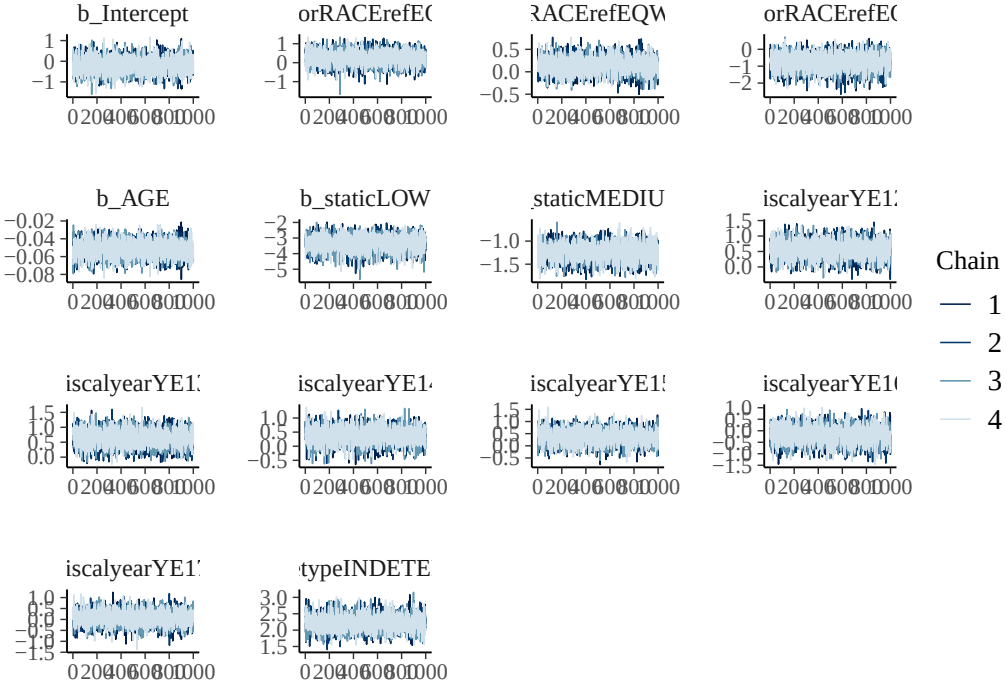

Figure 9: Trace plots for female security level model

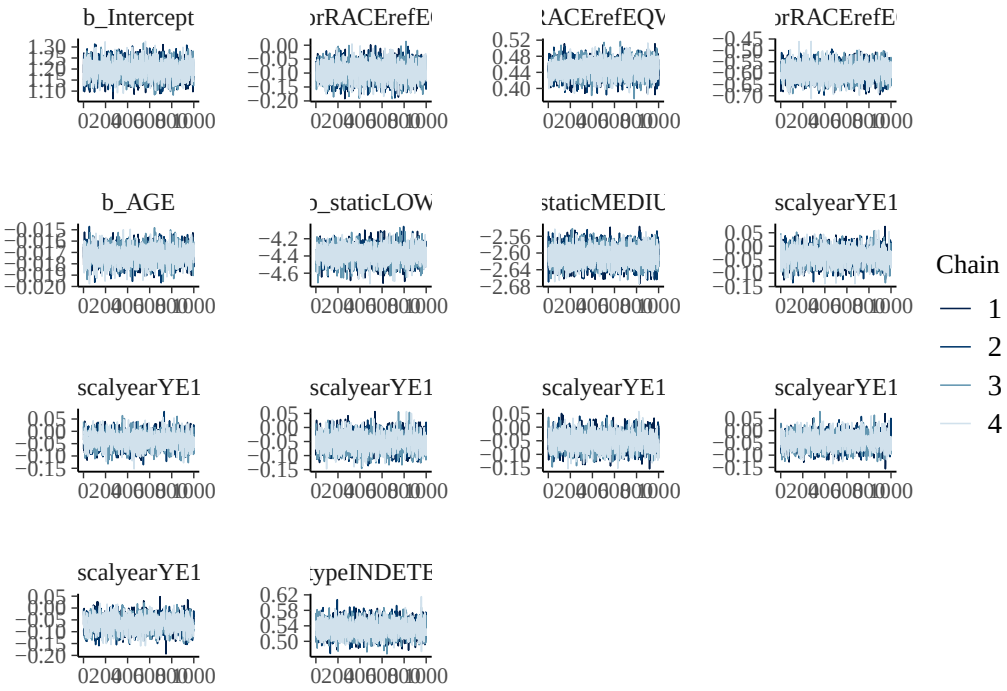

Figure 10: Trace plots for male reintegration potential model

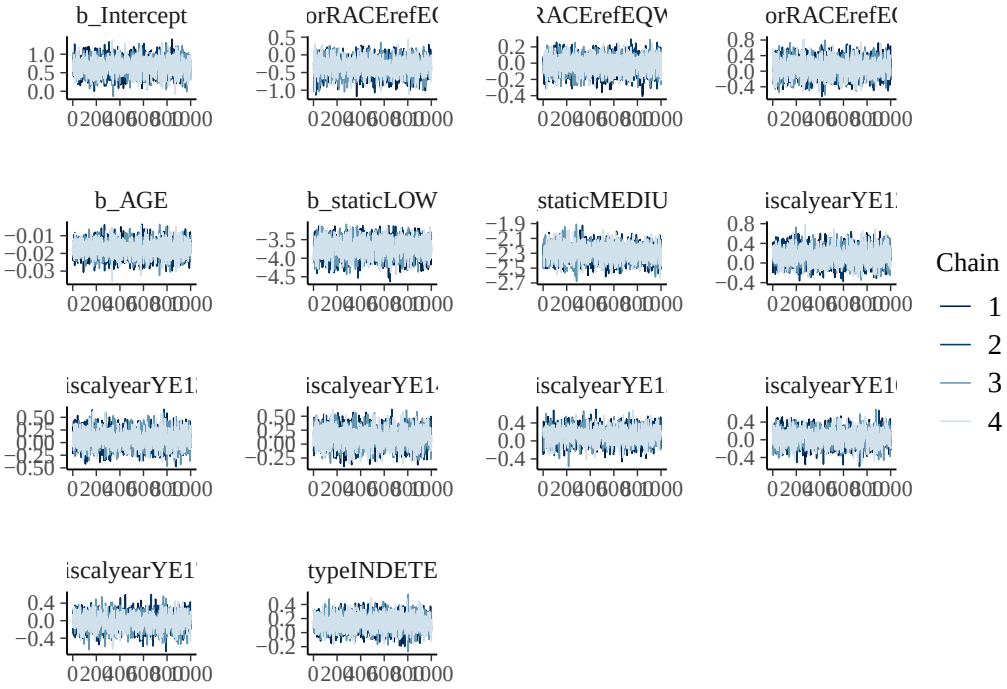

Figure 11: Trace plots for female reintegration potential model

References

Bertrand, Marianne, and Sendhil Mullainathan. 2004. "Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination." *American Economic Review*.

Bürkner, Paul-Christian. 2017. "brms: An R Package for Bayesian Multilevel Models Using Stan." *Journal of Statistical Software* 80 (1): 1–28. https://doi.org/10.18637/jss.v080.i01.

Canada, Correctional Service of. 2018. *Commissioner's Directives*. www.csc-scc.gc.ca/acts-and-regulations/005006-0001-en.shtml.

Cardoso, Tom. 2020a. "Bias Behind Bars: A Globe Investigation Finds a Prison System Stacked Against Black and Indigenous Inmates." *The Globe and Mail*. www.theglobeandmail.com/canada/article-investigation-racial-bias-in-canadian-prison-risk-assessments/.

———. 2020b. "How We Did It: How the Globe Uncovered Systemic Bias in Prisoners' Risk Assessments." *The Globe and Mail*. https://www.theglobeandmail.com/canada/article-investigation-racial-bias-in-canadian-prisons-methodology/.

Dowle, Matt, and Arun Srinivasan. 2019. *Data.table: Extension of 'Data.frame'*. https://CRAN.R-project.org/package=data.table.

Kay, Matthew. 2020. *tidybayes: Tidy Data and Geoms for Bayesian Models*. https://doi.org/10.5281/zenodo.1308151.

R Core Team. 2020. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Tierney, Nicholas, Di Cook, Miles McBain, and Colin Fay. 2020. *Naniar: Data Structures, Summaries, and Visualisations for Missing Data*. https://CRAN.R-project.org/package=naniar.

Waring, Elin, Michael Quinn, Amelia McNamara, Eduardo Arino de la Rubia, Hao Zhu, and Shannon Ellis. 2020. *Skimr: Compact and Flexible Summaries of Data*. https://CRAN.R-project.org/package=skimr.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

———. 2020. *Tidyr: Tidy Messy Data*. https://CRAN.R-project.org/package=tidyr.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Zhu, Hao. 2020. *KableExtra: Construct Complex Table with 'Kable' and Pipe Syntax*. https://CRAN.R-project.org/package=kableExtra.