Exploratory Analysis of Sexual Convictions

Data Analytics Lab

January 2020

This study is an examination of the sexual crime data available to West Midlands Police with a view to considering the factors that influence outcomes. In particular, the factors that may reduce the probability of making a charge and the probability of victims withdrawing their complaint.

1 Table of Contents

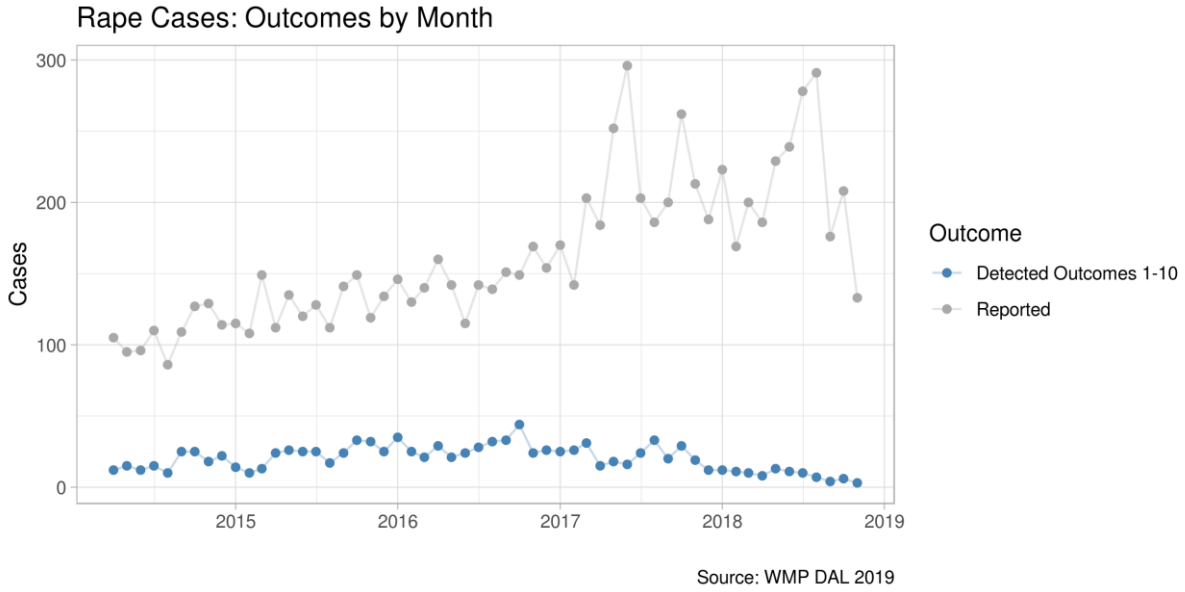
2 Introduction

ONS figures show that the incidence of recorded rape has been increasing dramatically. It is suggested that rising rape figures are partly due to an actual increase in the prevalence of violent sexual crime, and partly the result of victims being more willing to disclose. In particular, there has been an increase in reporting of historical cases of rape committed in earlier years.

The ONS advise (ONS 2018) that the majority of cases do not come to the attention of the police. One factor that is thought to affect reporting decisions by victims of rape is the high level of attrition in bringing rape cases to court and securing a conviction. Based on statistics from the crime survey for England and Wales ~26% more sexual crime offences are committed than are reported nationally.

Of the offences that do come to the attention of the police, many do not progress through the criminal justice system. Over 50% of sexual offences recorded by WMP do not proceed further through the criminal justice system due either to evidential difficulties, no suspect being identified, or investigations indicating that no crime took place or a false allegation has been made. This high percentage is a reflection of the challenges involved in investigating sexual offences.

Of the cases that do progress further, there is a clear year on year decreasing trend in the proportion of cases resulting in a charge. This decline may be, in part, due to resource pressures on the police following a substantial increase in recorded sexual offences. See the figure below. It is clear that as more cases are reported, less have resulted in a charge (outcome 1-10 in the chart below).



Rape is a statutory offence in England and Wales. The law requires the following points to be proved (beyond a reasonable doubt):

- No consent /and
- Penetration of mouth/anus by penis /or
- Penetration of vagina by an object held or manipulated by the hand

Rape under section 4 is a gender-neutral offence

For rape, consent is fundamental. The onus is to prove that the victim did not consent to the activity *AND* that the suspect did not believe that the victim consented (based on CPS definition www.cps.gov.uk/sexual-offences). If children under 16 years of age are involved, then it is not necessary to prove consent. Marriage does not provide grounds for consent. Consent can be conditional on specific actions and can be withdrawn at *any* point. Also, failure to resist does not constitute consent.

ACPO and the CPS have a protocol on the interactions with the Police Service (2015, reviewed 2017). Section 9 details the interaction between the police and the CPS. The investigating officer should arrange a consultation with a rape specialist as soon as possible. This should be within seven days at most and 24 hours if a suspect is detained in custody.

This rape specialist prosecutor will hold a pre-trial interview with any witnesses, though having witnesses is known to be rare. The legal prosecution will also discuss the case as soon as is practical, and the investigating officers will contact the Independent Sexual Violence Advisor (ISVA) to allow the legal team to understand the victim's situation.

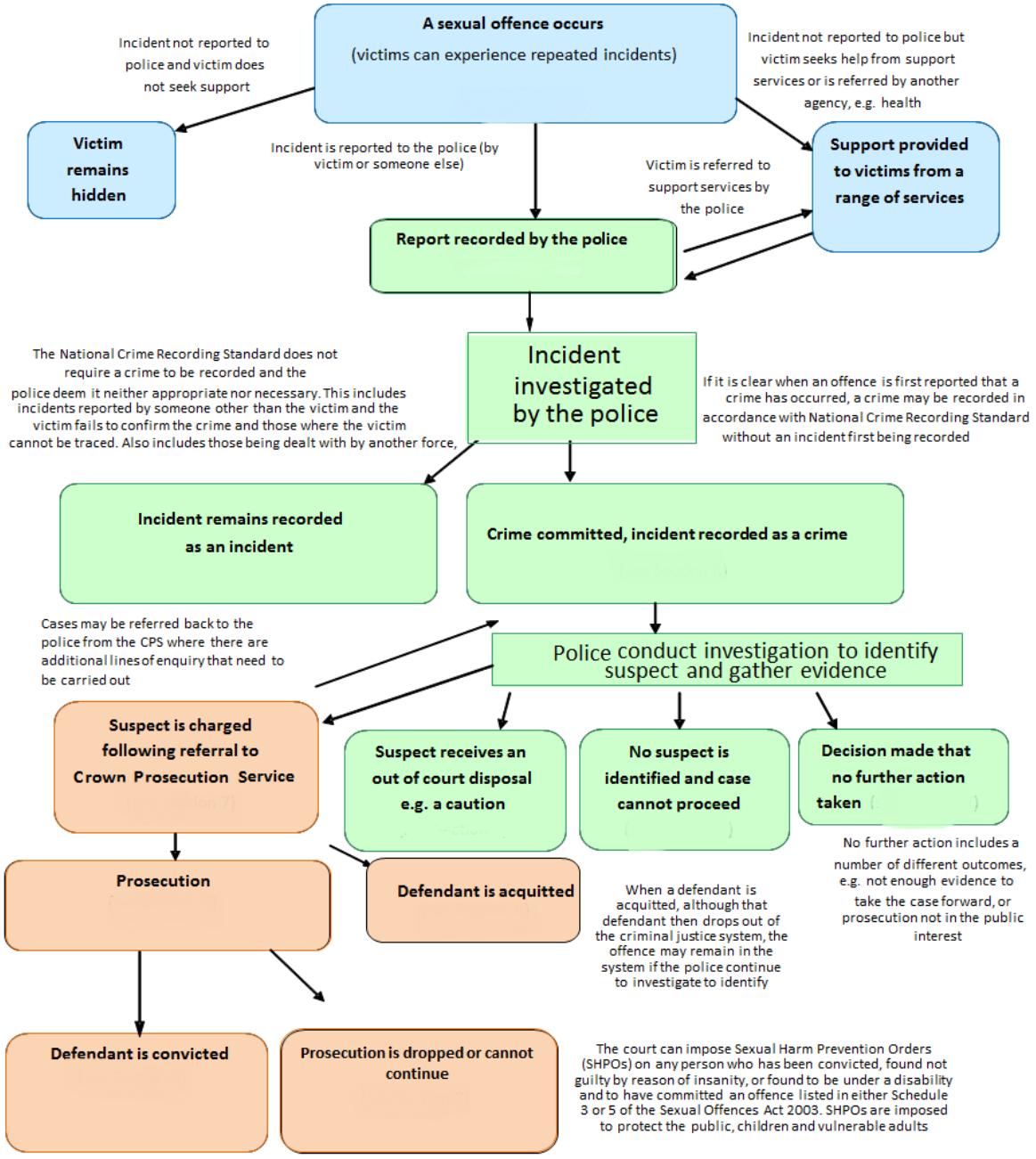
It is not uncommon (~31% cases) that victims withdraw their complaint. In this case, a statement is taken, preferably with an ISVA or similar present. This covers the rationale for the dropping of the complaint and if there was any pressure on the victim or whether the claim was untrue. It is possible to carry on a prosecution without the victim, though this is undesirable. If the police or the CPS decide to drop the case or reduce the charge, the victim will be informed with the offer of support of the most applicable kind. Even though these procedures are in place, a considerable number of victims do not see the prosecution through to charging or trial.

This study is an examination of the sexual crime data available to West Midlands Police with a view to considering the factors that influence outcomes. In particular, the factors that may reduce the number of victims withdrawing their complaint.

The key attrition points as an incident progresses into and through the criminal justice system are

- Whether an incident is reported / recorded
- Whether an incident is "no crimed"
- Whether the Police investigation gathers a sufficient body of evidence to recommend a case to the CPS
- Whether the CPS recommends to proceed to trial.
- Whether the IP withdraws support for an investigation.

This investigation examines only the data available from the Police recording the incident to the decision to **charge** or *close* a case with another outcome.



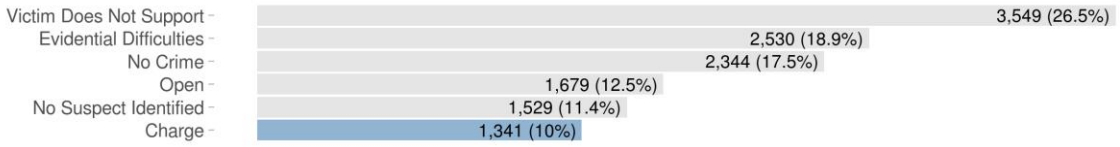
*How sexual offences progress through the criminal justice system (**Source: ONS**). This investigation examines only the 'green' section: from the Police recording the incident to the decision to **charge** or close the case with another outcome*

3 Exploratory Data Analysis

3.1 Data

Data were extracted for the period January 2014 - October 2018 from the Crimes, Socrates, and Oasis databases for all penetrative crimes.

Clear up codes are aggregated into the groupings:



Source: WMP DAL 2019

Also reported are the "*Outcomes 1-10*" which are used in WMP internal reporting. In addition to the cuc codes used in the **Charge** grouping, this includes the cuc codes:

Cuc Code	Description	Incident Count
56	The offender has died (all offences)	69
60	Sufficient evidence to charge, but cps decided not in the public interest to prosecute	12 (which is ~0.1% of all cases, and equivalent to 1% of charged cases).
61	Sufficient evidence to charge, but police decided not in the public interest to prosecute	12

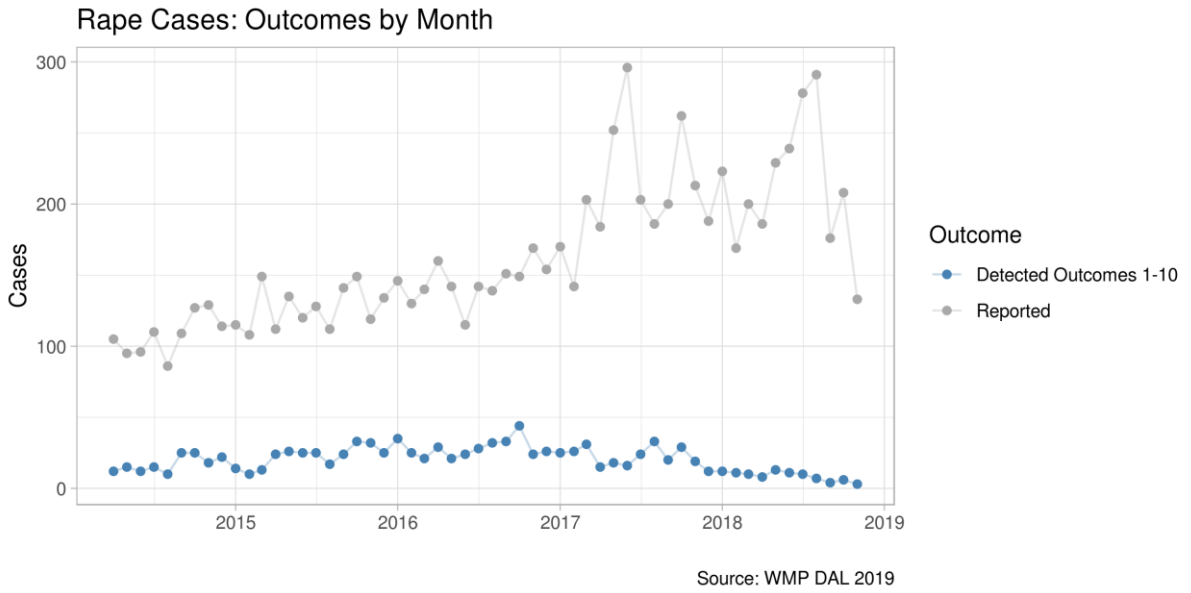
Table: Other clearup codes.

There is little difference between **Charge** and *Outcomes 1-10*.

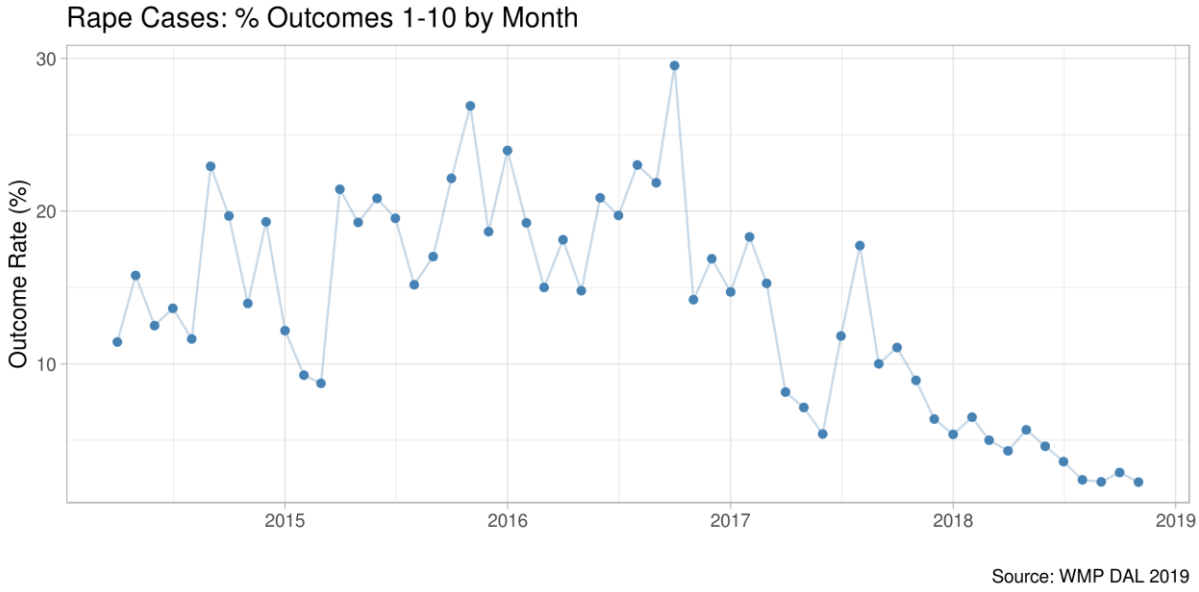
3.2 Cases by Month

1,627 crimes are ongoing at the end of the extracted period. These are predominantly from the most recent periods, though there are a number of cases that have remained open over many years. (We will see later that 20% of cases that are charged are open for longer than a year.)

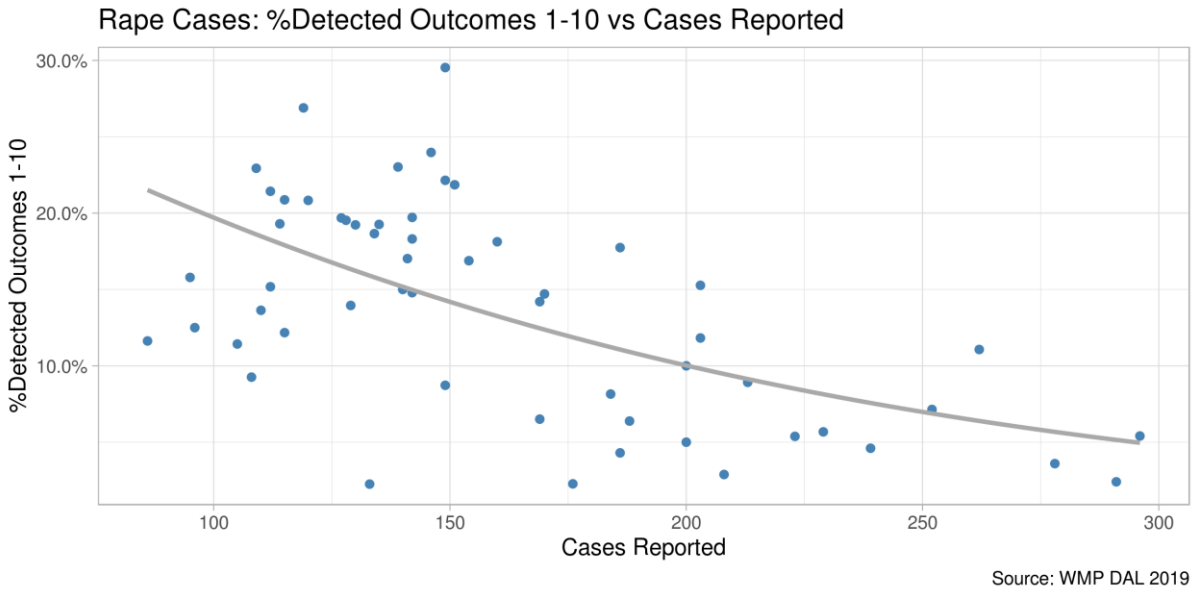
There is a clear increasing trend in the number of cases reported since 2014. The number of cases cleared up mirror this increase. *Outcomes 1-10*, which typically result in a **charge** have decreased.



This drop off in **charges** is clearer in the plot of *%Outcomes 1-10* in isolation.



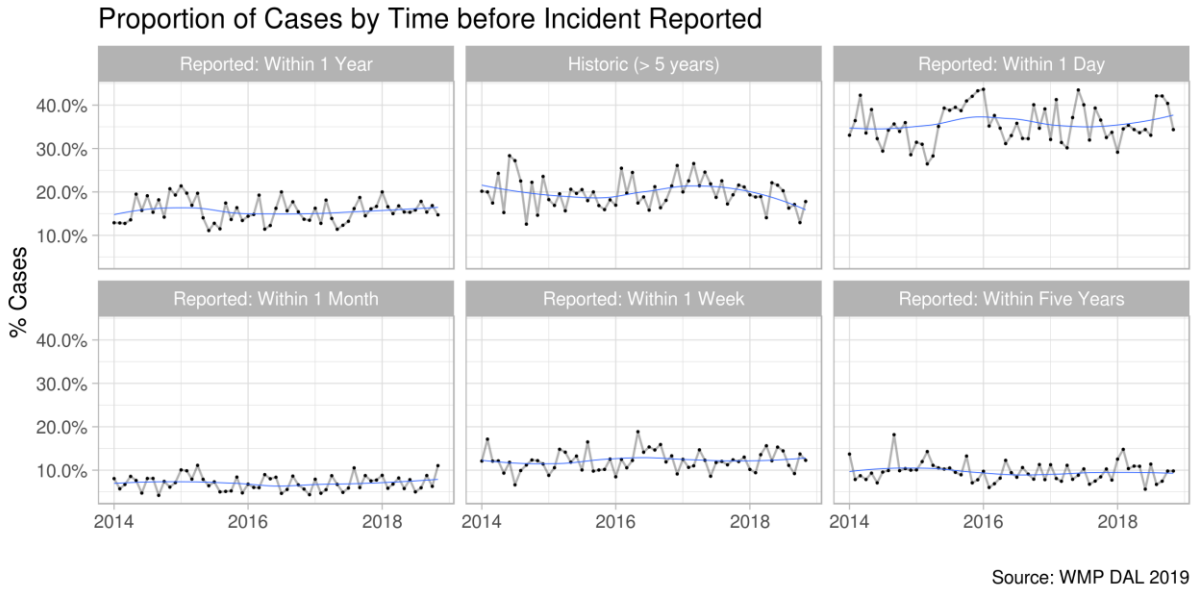
Plotting the percentage of *outcomes 1-10* against the number of cases reported for each month in the period January 2014 - October 2018 shows a negative relationship between the volume of incidents reported and the number of positive outcomes. Plausibly, higher caseloads generate more administrative work leading to less focus on individual cases with the result that fewer crimes are **charged**.



The composition of cases has remained relatively unchanged throughout the period 2014-2018. For example, there has not been an increase in historical reporting in this period.

- ~35% of incidents are reported on the same day.
- ~20% are reported over 5 years after the event.

Each of the groupings is exclusive and does not include the other groups. For example, "reported within 1 week" does **not** include incidents "reported with 1 day".

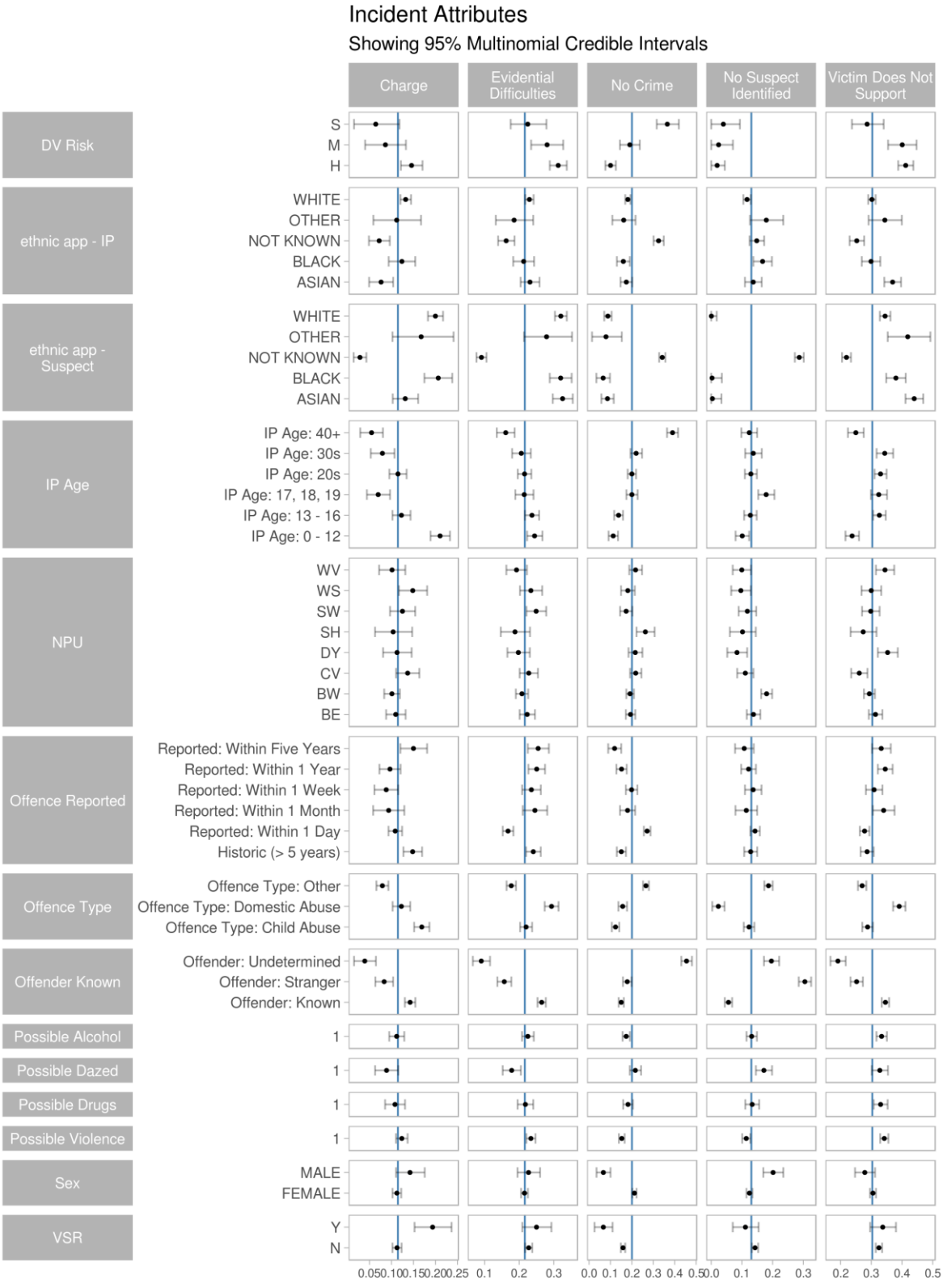


3.3 Explanatory Factors

An outcome is subject to a wide variety of external factors in addition to calendar effects. The graphic below shows the proportion of outcomes by each factor. The blue line represents the overall proportion of cases in each clear up category. Reading down looking for irregularities:

- A case is more likely to be **charged** if the IP is below 13, and less likely in the age range 17-19.
- A case is less likely to be **charged** if the IP is 40+ and more likely to result in **No Crime**.
- The IP is less likely to support when there is a high domestic violence risk, and also when the suspect is Asian.

Factors based on textual analysis of the investigation notes for *alcohol*, *drugs*, *violence* and the IP being *dazed* or losing consciousness show little impact on the likelihood to charge. Though, the victim is less likely to support where violence or alcohol is involved.

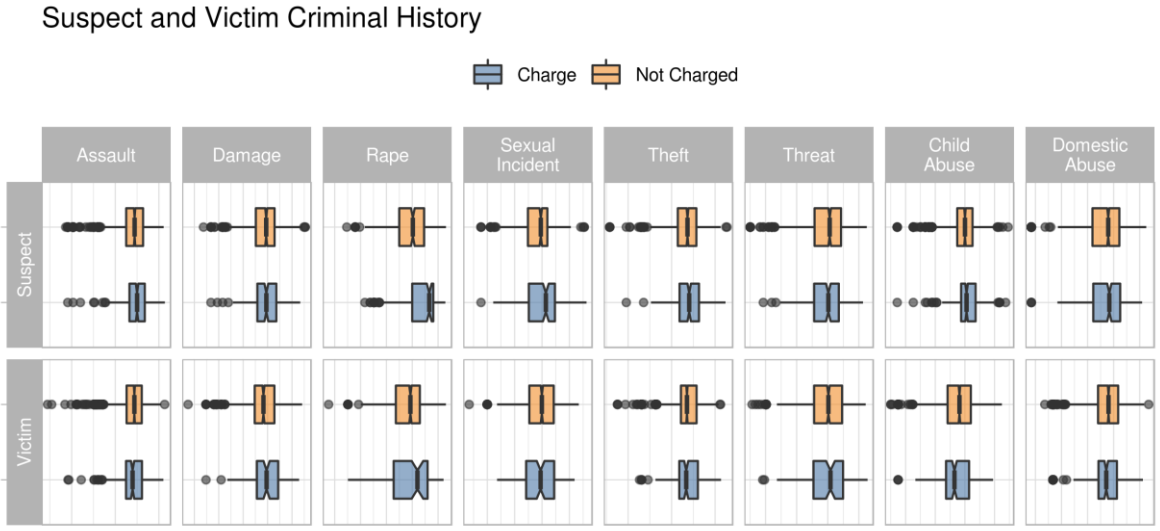


Source: WMP DAL 2019

The boxplots below compare outcomes based on the suspects' and victims' criminal history.

Criminal history is based on an exponentially weighted moving 10 year average. More recent crimes have more weight. Crimes older than 10 years at the time the incident are not included.

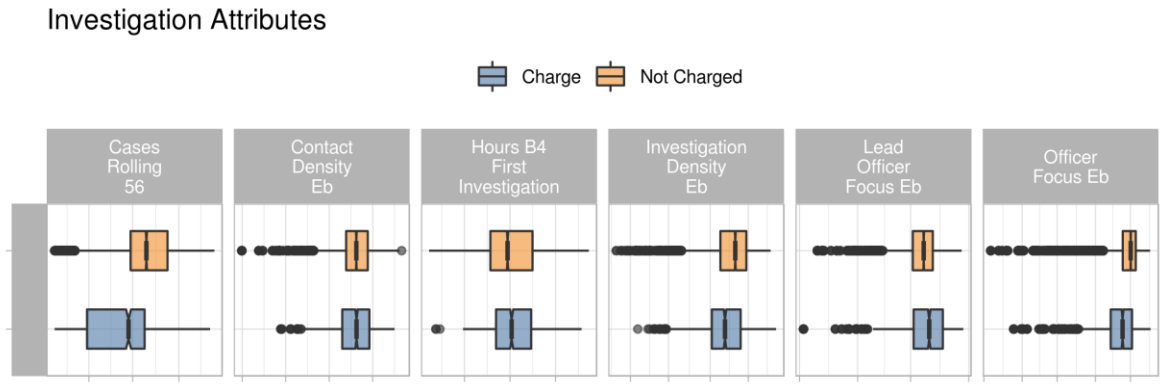
- A case is more likely to be charged if the suspect has a recent history of rape or other sexual incidents. Also, where the victim has previously reported an incident of rape.
- There is a slight reduction in the likelihood of a charge where the IP has previously been involved in an incident of child abuse.



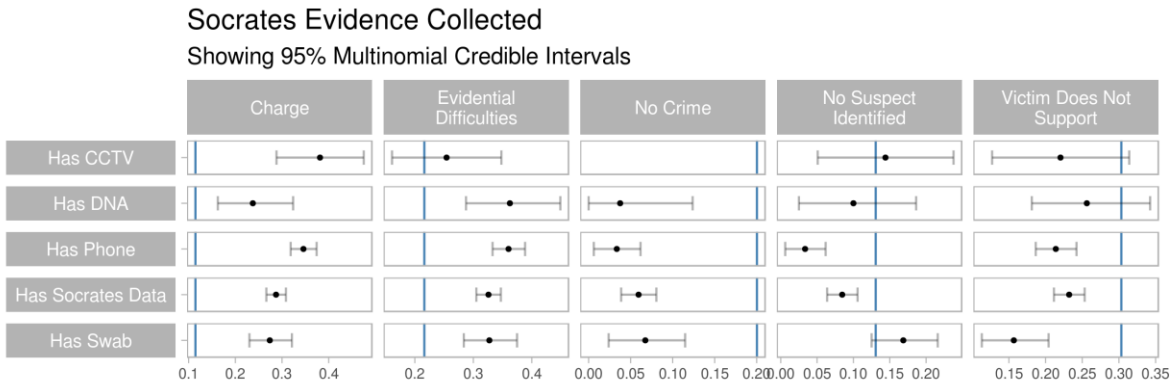
Source: WMP DAL 2019

The graphic below compares outcomes based on attributes of the investigation.

- Higher caseloads are associated with fewer crimes being **charged**.
- The number of officers working a case is captured by the variable *officer focus*, which is defined here as the number of unique officers / number of investigation notes. The fewer officers working a case, the more likely that a crime is charged. (This might also be interpreted as a measure of the complexity of a case)
- A case is far more likely to be **charged** if forensic data is available (20% of cases). Collecting phone evidence being the most advantageous.



Source: WMP DAL 2019

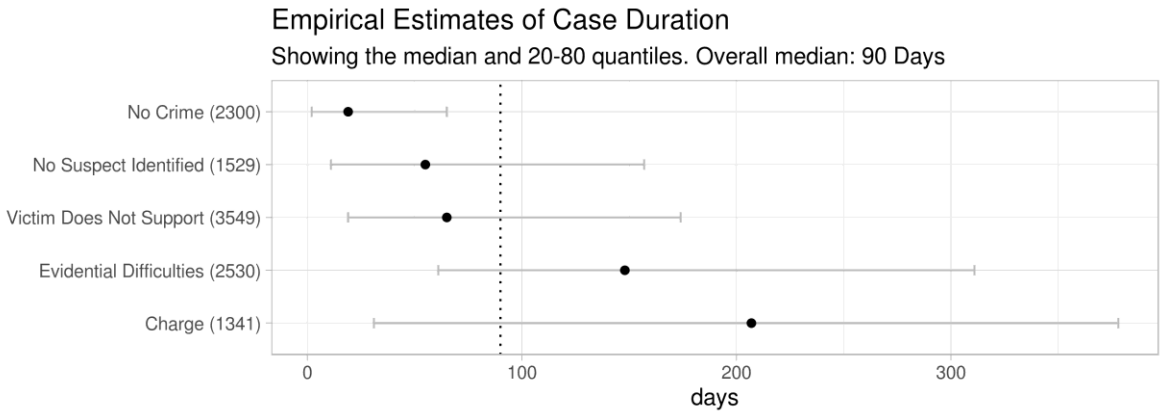


Source: WMP DAL 2019

3.4 Case Duration

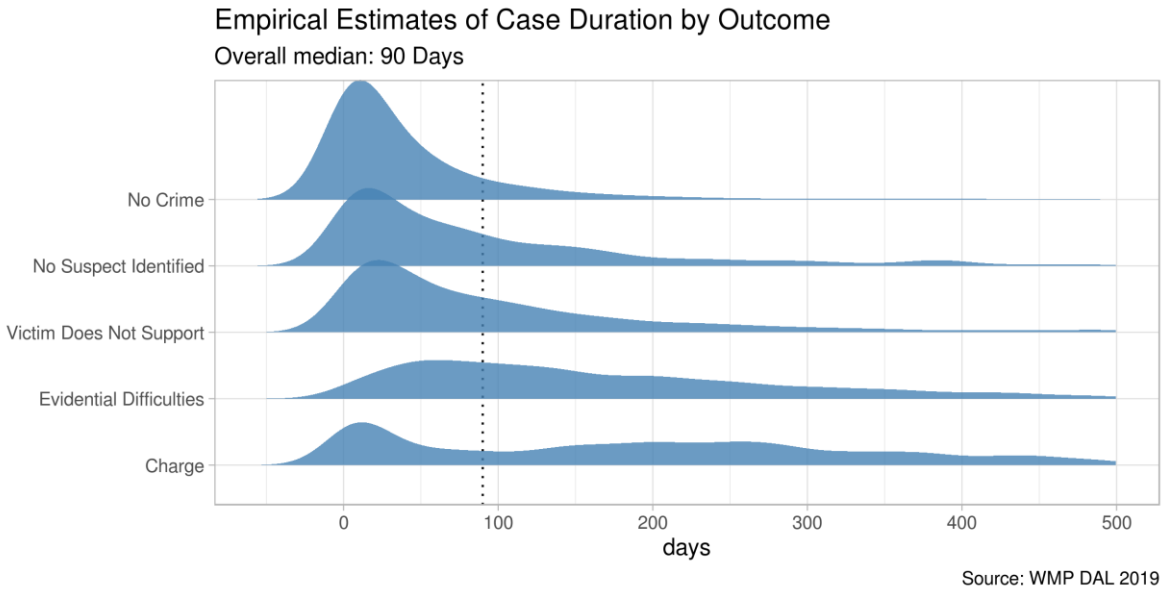
The median time that a case is open is 90 Days. (A Kaplan-Meier correction accounting for in-progress cases does not change this).

The median duration varies greatly between outcomes. **No Crime** cases are identified relatively early, with 80% closed within two months. Incidents leading to a **charge** have a median duration of 207 days (~7 months), with around 20% taking over a year.

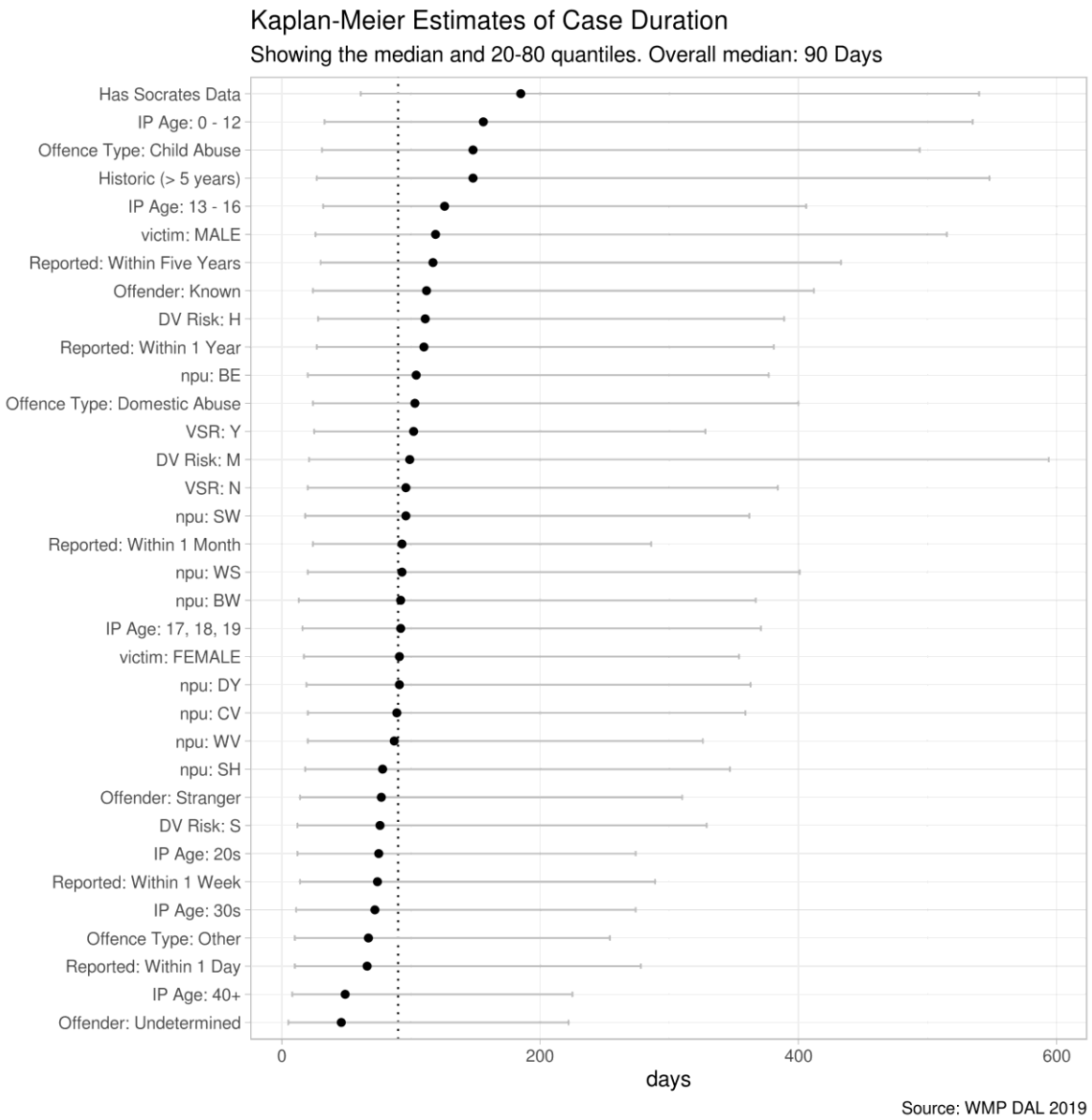


Source: WMP DAL 2019

The shape of the "Charge" density (below) suggests that while some cases are clearcut and processed quickly, there are a large number of more complex investigations.

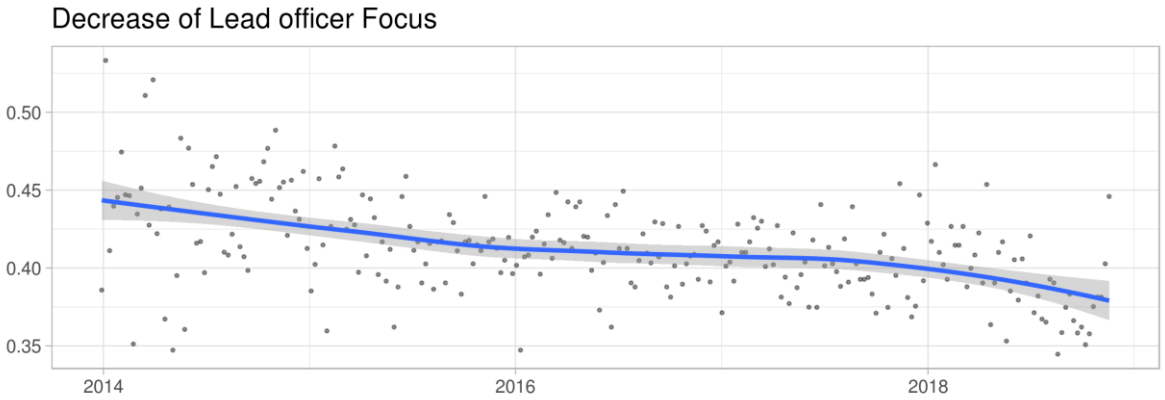


Reviewing the impact of features of investigations on the overall case duration, we note that the features that are associated with an increase in case duration (eg. having forensic evidence) are plausibly associated with an increased likelihood to **charge**, and features associated with a decreased case duration are associated with the endpoints **No Crime** and **No Suspect Identified**.


Kaplan-Meier Estimates of Case Duration
Showing the median and 20-80 quantiles. Overall median: 90 Days
Has Socrates Data
IP Age: 0 - 12
Offence Type: Child Abuse
Historic (> 5 years)
IP Age: 13 - 16
victim: MALE
Reported: Within Five Years
Offender: Known
DV Risk: H
Reported: Within 1 Year
npu: BE
Offence Type: Domestic Abuse
VSR: Y
DV Risk: M
VSR: N
npu: SW
Reported: Within 1 Month
npu: WS
npu: BW
IP Age: 17, 18, 19
victim: FEMALE
npu: DY
npu: CV
npu: WV
npu: SH
Offender: Stranger
DV Risk: S
IP Age: 20s
Reported: Within 1 Week
IP Age: 30s
Offence Type: Other
Reported: Within 1 Day
IP Age: 40+
Offender: Undetermined
0
200
400
600
days
Source: WMP DAL 2019

4 Findings

The incidence of recorded rape and penetrative sexual crimes has been increasing dramatically in the study period, and this has met with a year on year decreasing trend in the proportion of cases resulting in a charge. This decline may be, in part, due to resource pressures. One of the most important explanatory variables is the proportion of an investigation dealt with by a single officer. The data indicate that this metric has been falling over the analysis period.



Source: WMP DAL 2019

The highest impacts on the outcome of an investigation are depicted in the nomograms below. A measure to the right is consistent with support of an outcome, and measures to the left are compatible with a decrease in the odds of an outcome.

The nomograms included below are for rape only. Nomograms including all penetrative sexual crimes are in the appendix to this document.

The presence or absence of scene of crime data has a high impact on the outcome. A phone taken into evidence is consistent with an increase in the odds of a charge and a decrease in the odds of a victim not supporting.

There is a complex relationship between IP age and support of a case. The likelihood of not supporting is low for children, peaks in the late teens and slowly decreases from the mid-thirties.

There is overall a positive effect of maintaining contact with the IP. However, this is not linear and a high density of contact is related to a victim not supporting. This is likely where WMP are *chasing* as a result of the IP disengaging from the process.

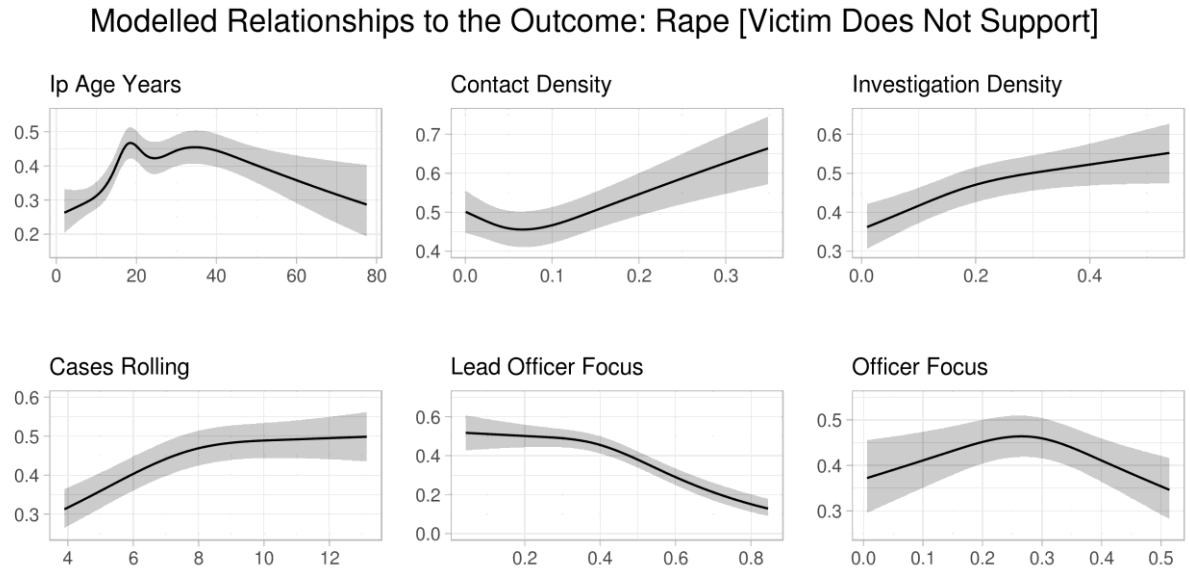
The number of open cases has a negative impact on victim support and the likelihood of charging. Conceivably this is related to the officer focus and case continuity. The model suggests that maintaining officer continuity (lead officer focus) has a positive effect on the number of cases resulting in a charge and also reducing victims' not supporting a case.

The time before reporting has little impact and is most likely *controlled for* by the availability of forensic data.

The number of officers working a case is captured by the variables *officer focus*, which is defined here as "the unique officers / number of investigation notes". The more officers working a case, the more likely the victim will not support. (This might also be interpreted as a measure of the complexity of a case as well as the degree to which officers are moved between cases and departments).

The model suggests that a high activity level in combination with multiple officers (as captured by *lead officer focus*) working a case may discomfort an IP and result in disengagement from the process (or officers may be chasing IPs that have already come to an internal decision not to continue with the case).

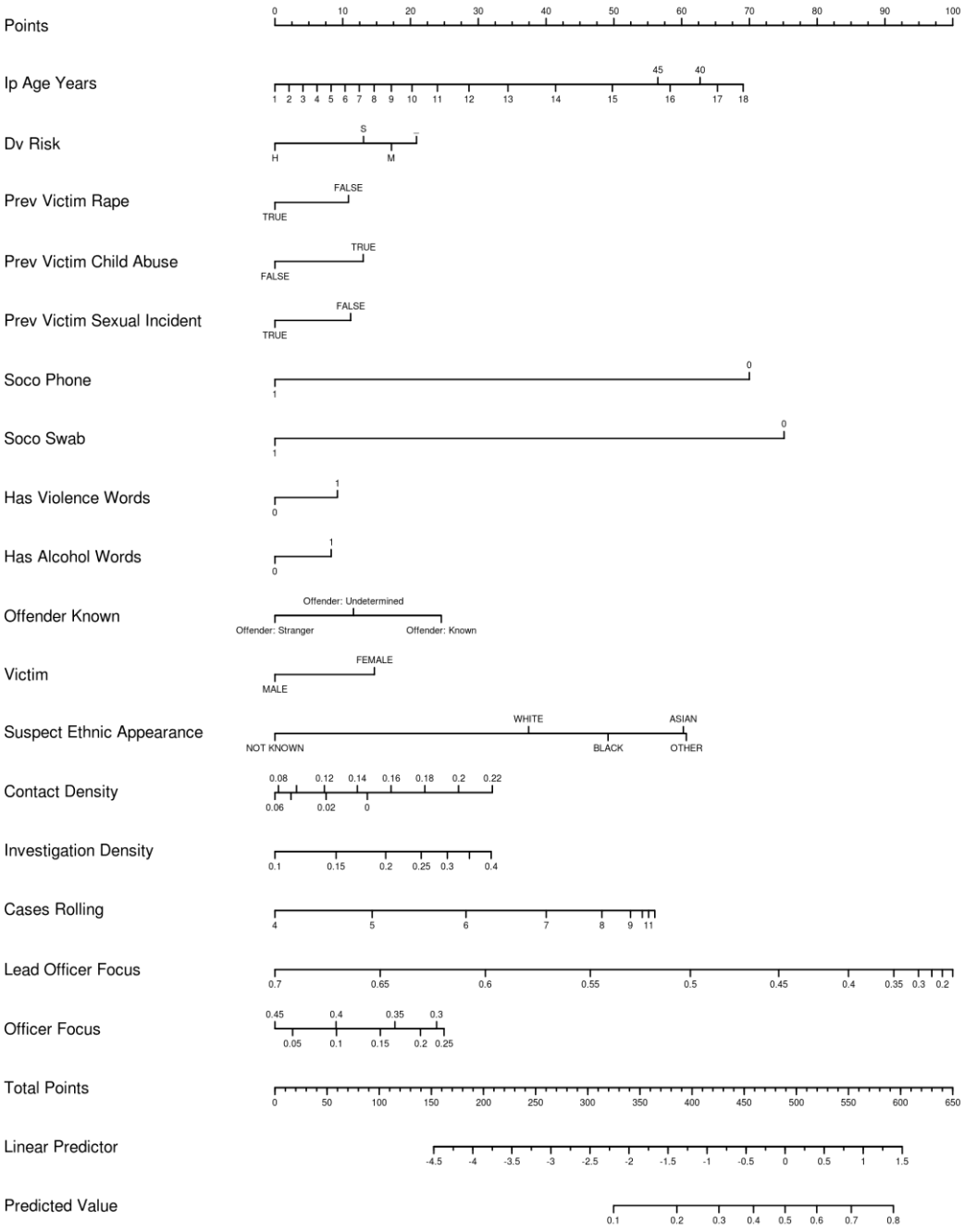
4.1 Victim does not Support



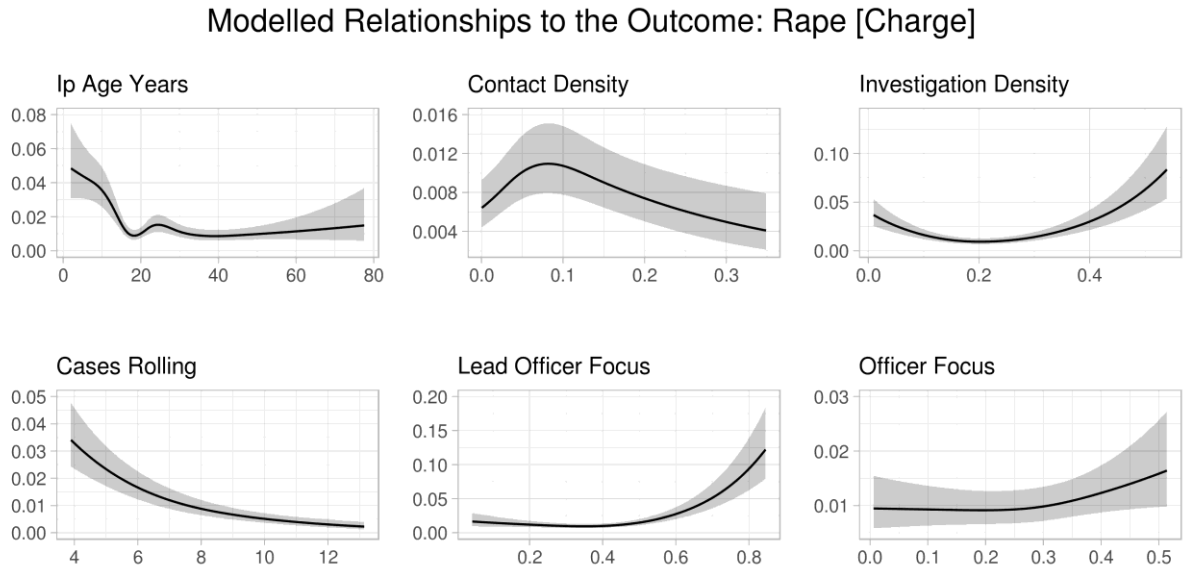
It is also noteable on the nomogram that

- An IP with a previous history of rape or a sexual incident is more likely to support.
- An IP with previous history of child abuse is less likely to support.
- A case involving violence or alcohol is less likely to be supported.
- A victim is less likely to support where the suspect is known.
- A victim is less likely to support where the suspect is Asian.

Nomogram: Rape [Victim Does Not Support]



4.2 Charge



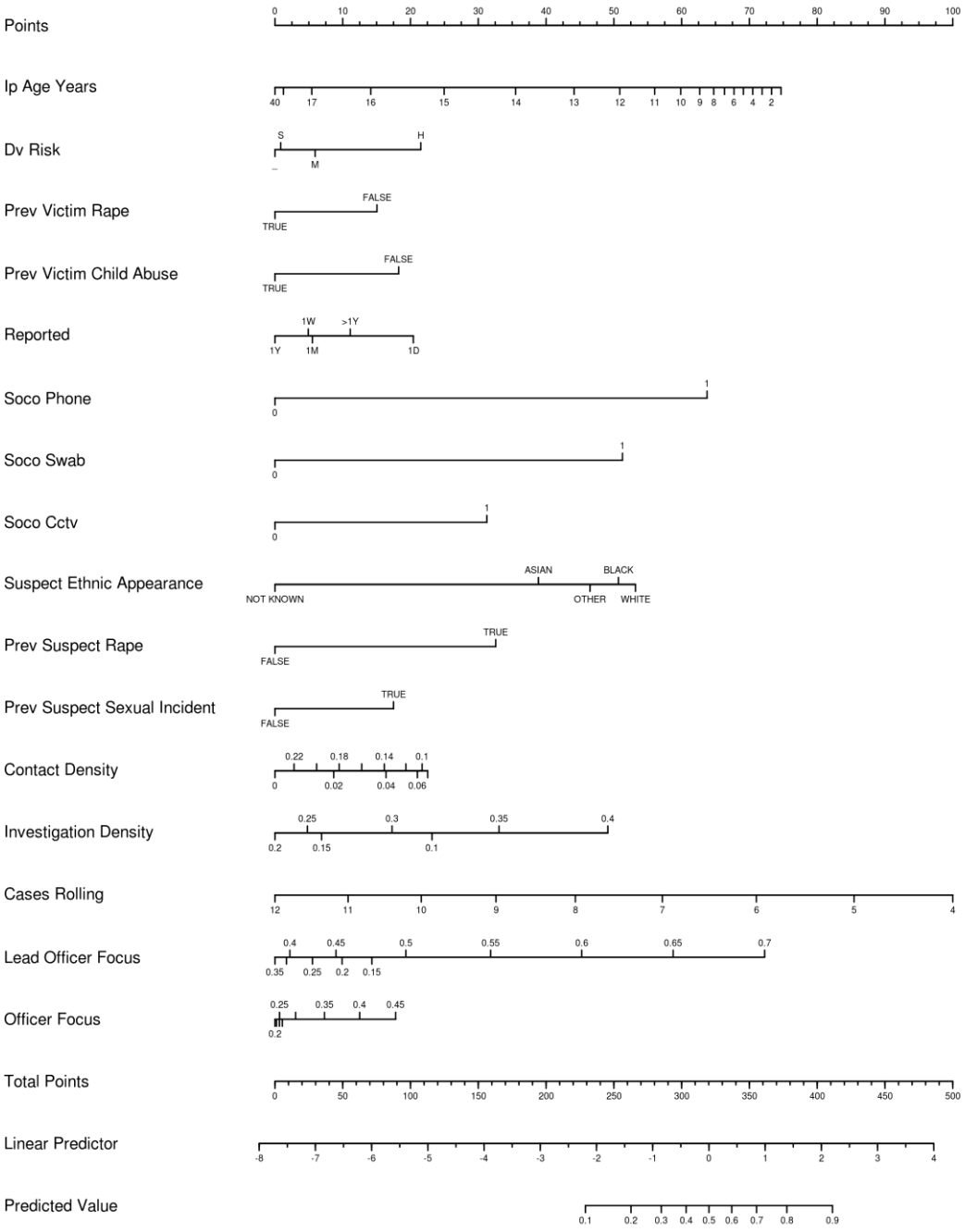
There is a threshold effect of lead officer focus. If the lead officer is responsible for less than around 40% of the activity, there is little impact on the likelihood to charge.

As noted previously, there is overall a positive effect of maintaining contact with the IP. However, we again see a non-linear effect. This is likely where WMP are *chasing* as a result of the IP disengaging from the process.

It is also apparent that:

- A case involving an IP with previous history of rape or child abuse is less likely to be charged.
- A case involving an IP with a high DV risk is more likely to be charged.
- A case involving a suspect with a previous history of rape is more likely to be charged.
- A case reported on the same day as the incident is more likely to be charged.

Nomogram: Rape [Charge]



5 Effects on Resourcing

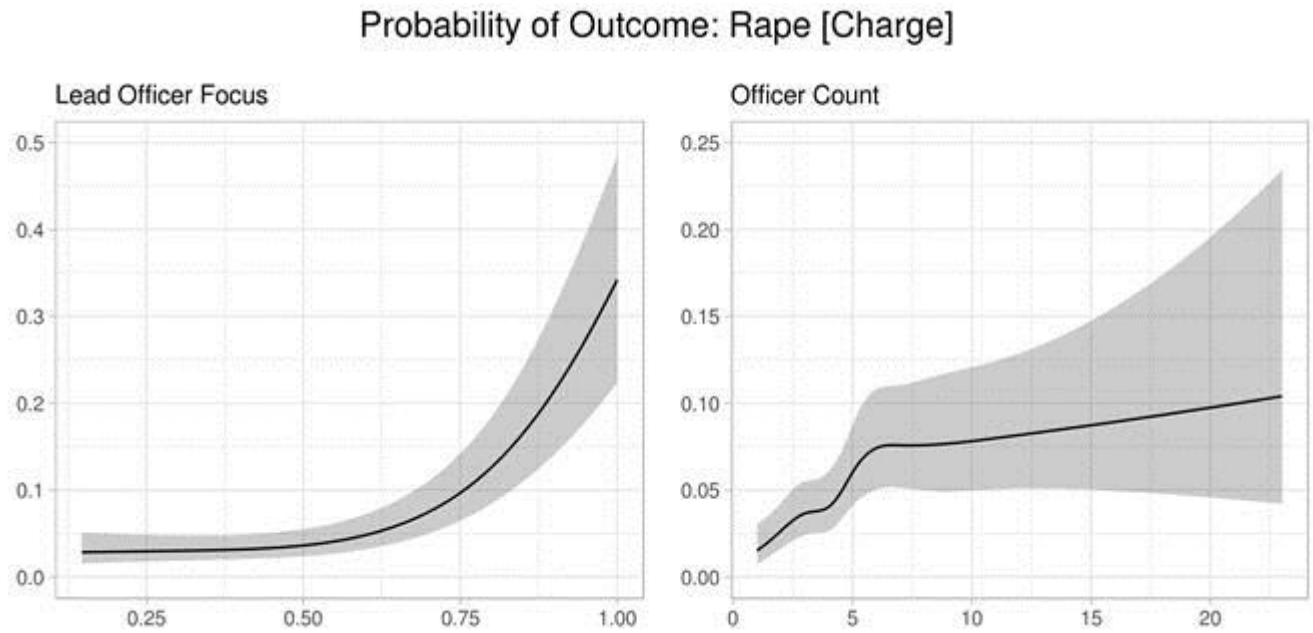
The above analyses show the importance of resources, both in terms of their level and how they are used. This section examines resources in further detail.

Consider the process of adding more officers to a case in order to improve the outcome of an investigation. At some point, adding more officers will cause issues such as officers getting in each other's way or the duplication of work. When investigations move between officers one after the other this introduces problems such as "getting up to speed", and for the IP, an inconsistent point of contact.

Here we refit the logistic model to the raw number of officers, and we include only the outcome "Charge" and "Victim Does not Support".

The partial plots below agree with the previous model: the more work done by a single officer, the more likely we are to see a charge.

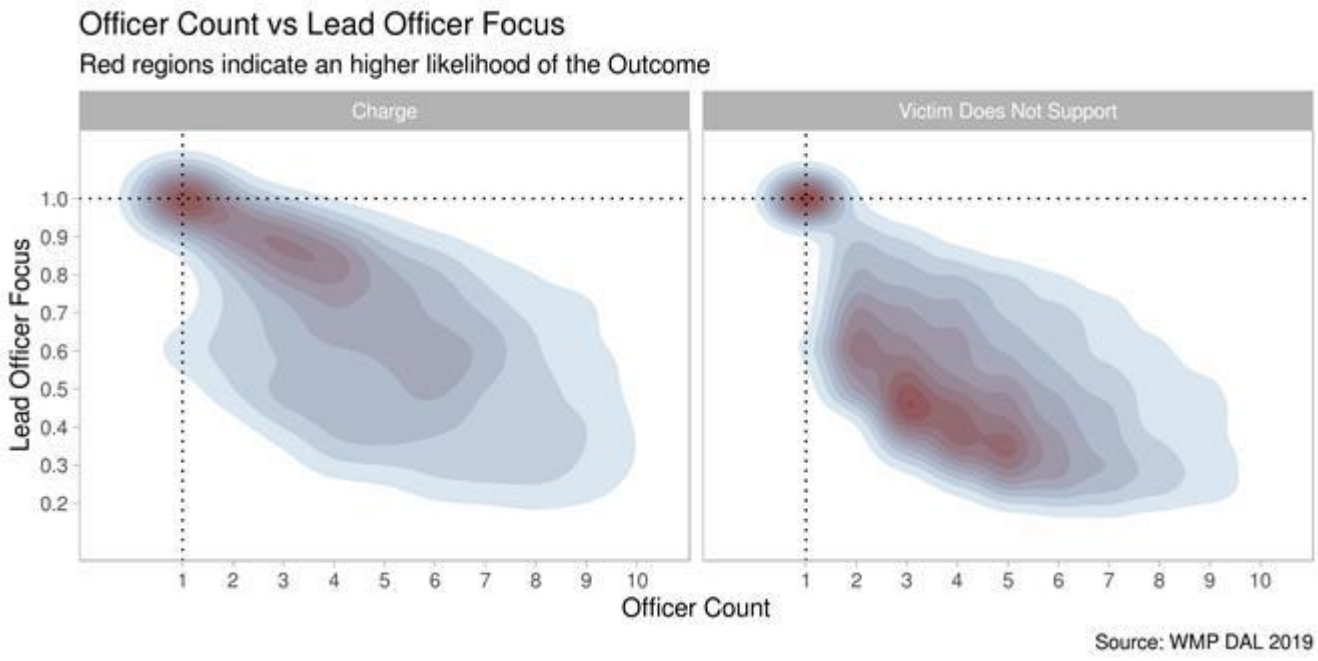
However, adding officers also increases the likelihood of a charge. This is misleading because the two variables are related - the more officers working on a case the less focus the lead officer has.



Plotting the two variables together.

- The highest probability of a *charge* is when a single officer is responsible for the majority of the work.
- Up to 4 officers can assist as long as the lead officer performs 80% or more of the work.
- The highest probability of a *victim not supporting* also occurs when a single officer is responsible for the majority of the work. This likely represents IPs that are insensitive to any WMP action and will not support.

- When the lead officer performs less than 80% of the work, there is a high likelihood of the IP not supporting.

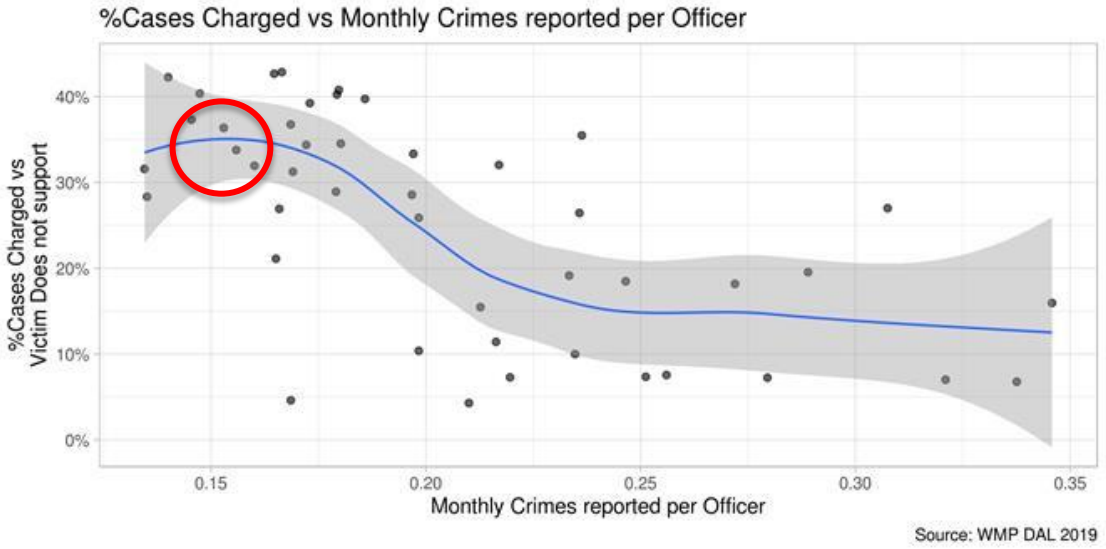


In 2014 around 100 new cases of rape were reported every month. The number of monthly reported cases has increased year on year to approximately 220 in 2018. In the intervening period, the number of officers working on rape cases has remained relatively unchanged.

(For an officer to be working on a case we include here any officer adding an investigation note, not just officers working within the PPU).

Cases can be open for many months, leading to an on-going resource requirement.

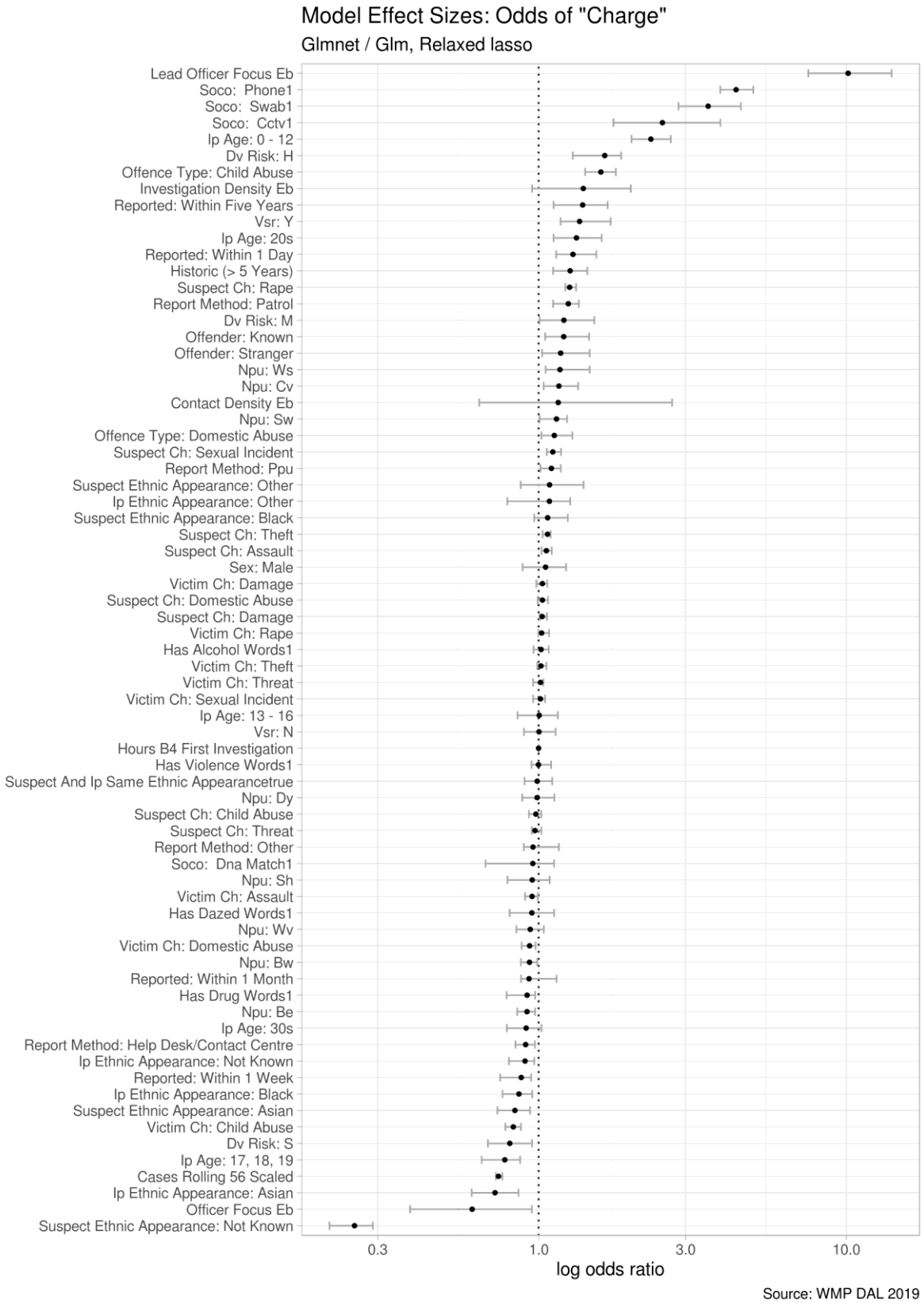
From the graph below, to maximise the proportion of charged cases, we require one officer for every 0.15 crimes reported in a month. This proportion equates to 1,460 officers. (~ 70% increase).

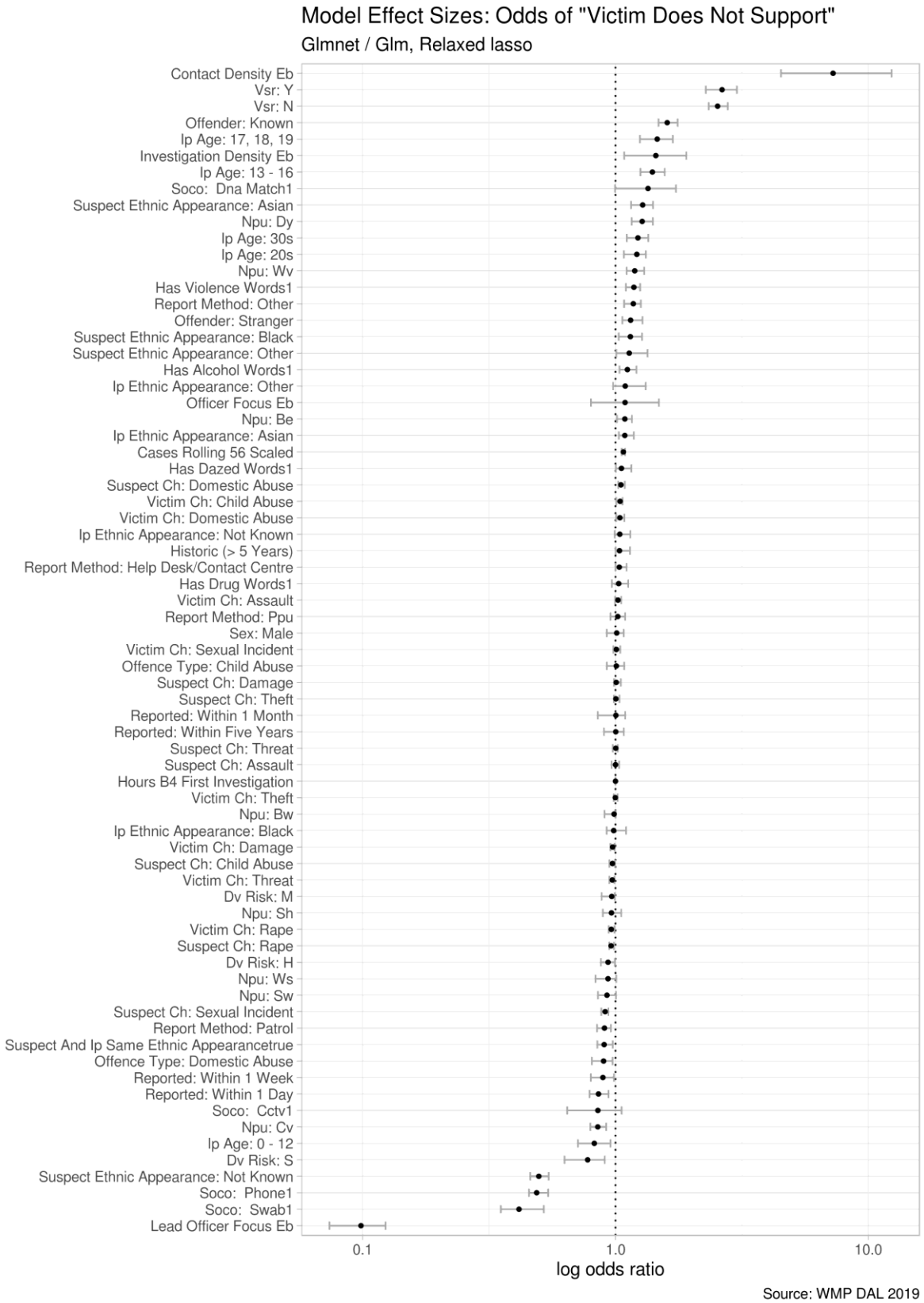


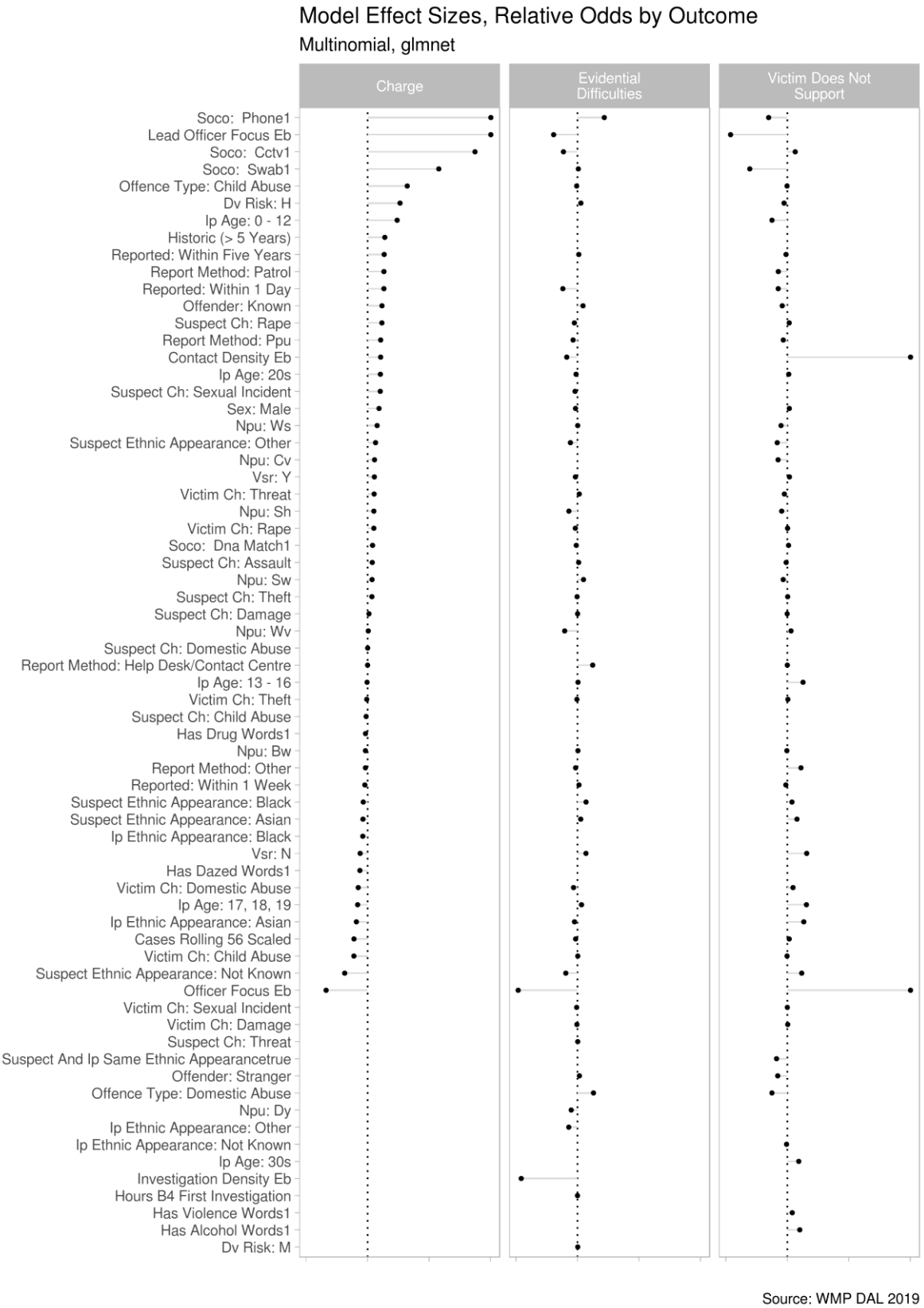
6 Appendix

6.1 Relative Odds by Outcome.

A line to the right is consistent with support of an outcome, and line to the left is compatible with a decrease in the odds of an outcome. For example, a phone taken into evidence is consistent with an increase in the odds of a charge.

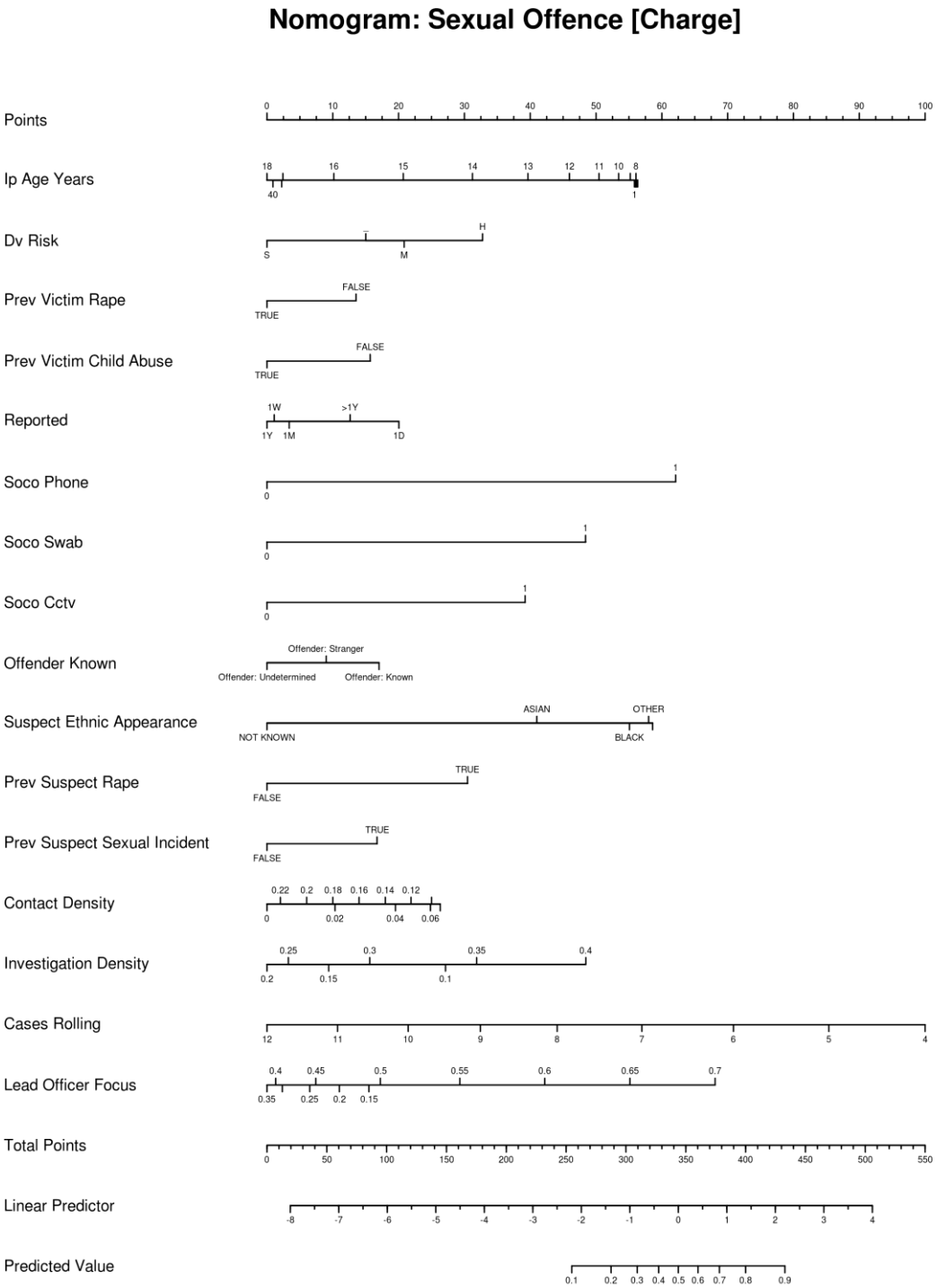

Model Effect Sizes: Odds of "Charge"
Glmnet / Glm, Relaxed lasso
Lead Officer Focus Eb
Soco: Phone1
Soco: Swab1
Soco: Cctv1
Ip Age: 0 - 12
Dv Risk: H
Offence Type: Child Abuse
Investigation Density Eb
Reported: Within Five Years
Vsr: Y
Ip Age: 20s
Reported: Within 1 Day
Historic (> 5 Years)
Suspect Ch: Rape
Report Method: Patrol
Dv Risk: M
Offender: Known
Offender: Stranger
Npu: Ws
Npu: Cv
Contact Density Eb
Npu: Sw
Offence Type: Domestic Abuse
Suspect Ch: Sexual Incident
Report Method: Ppu
Suspect Ethnic Appearance: Other
Ip Ethnic Appearance: Other
Suspect Ethnic Appearance: Black
Suspect Ch: Theft
Suspect Ch: Assault
Sex: Male
Victim Ch: Damage
Suspect Ch: Domestic Abuse
Suspect Ch: Damage
Victim Ch: Rape
Has Alcohol Words1
Victim Ch: Theft
Victim Ch: Threat
Victim Ch: Sexual Incident
Ip Age: 13 - 16
Vsr: N
Hours B4 First Investigation
Has Violence Words1
Suspect And Ip Same Ethnic Appearancetrue
Npu: Dy
Suspect Ch: Child Abuse
Suspect Ch: Threat
Report Method: Other
Soco: Dna Match1
Npu: Sh
Victim Ch: Assault
Has Dazed Words1
Npu: Wv
Victim Ch: Domestic Abuse
Npu: Bw
Reported: Within 1 Month
Has Drug Words1
Npu: Be
Ip Age: 30s
Report Method: Help Desk/Contact Centre
Ip Ethnic Appearance: Not Known
Reported: Within 1 Week
Ip Ethnic Appearance: Black
Suspect Ethnic Appearance: Asian
Victim Ch: Child Abuse
Dv Risk: S
Ip Age: 17, 18, 19
Cases Rolling 56 Scaled
Ip Ethnic Appearance: Asian
Officer Focus Eb
Suspect Ethnic Appearance: Not Known
0.3
1.0
3.0
10.0
log odds ratio
Source: WMP DAL 2019


Model Effect Sizes: Odds of "Victim Does Not Support"
Glmnet / Glm, Relaxed lasso
Contact Density Eb
Vsr: Y
Vsr: N
Offender: Known
Ip Age: 17, 18, 19
Investigation Density Eb
Ip Age: 13 - 16
Soco: Dna Match1
Suspect Ethnic Appearance: Asian
Npu: Dy
Ip Age: 30s
Ip Age: 20s
Npu: Wv
Has Violence Words1
Report Method: Other
Offender: Stranger
Suspect Ethnic Appearance: Black
Suspect Ethnic Appearance: Other
Has Alcohol Words1
Ip Ethnic Appearance: Other
Officer Focus Eb
Npu: Be
Ip Ethnic Appearance: Asian
Cases Rolling 56 Scaled
Has Dazed Words1
Suspect Ch: Domestic Abuse
Victim Ch: Child Abuse
Victim Ch: Domestic Abuse
Ip Ethnic Appearance: Not Known
Historic (> 5 Years)
Report Method: Help Desk/Contact Centre
Has Drug Words1
Victim Ch: Assault
Report Method: Ppu
Sex: Male
Victim Ch: Sexual Incident
Offence Type: Child Abuse
Suspect Ch: Damage
Suspect Ch: Theft
Reported: Within 1 Month
Reported: Within Five Years
Suspect Ch: Threat
Suspect Ch: Assault
Hours B4 First Investigation
Victim Ch: Theft
Npu: Bw
Ip Ethnic Appearance: Black
Victim Ch: Damage
Suspect Ch: Child Abuse
Victim Ch: Threat
Dv Risk: M
Npu: Sh
Victim Ch: Rape
Suspect Ch: Rape
Dv Risk: H
Npu: Ws
Npu: Sw
Suspect Ch: Sexual Incident
Report Method: Patrol
Suspect And Ip Same Ethnic Appearancetrue
Offence Type: Domestic Abuse
Reported: Within 1 Week
Reported: Within 1 Day
Soco: Cctv1
Npu: Cv
Ip Age: 0 - 12
Dv Risk: S
Suspect Ethnic Appearance: Not Known
Soco: Phone1
Soco: Swab1
Lead Officer Focus Eb
0.1
1.0
10.0
log odds ratio
Source: WMP DAL 2019

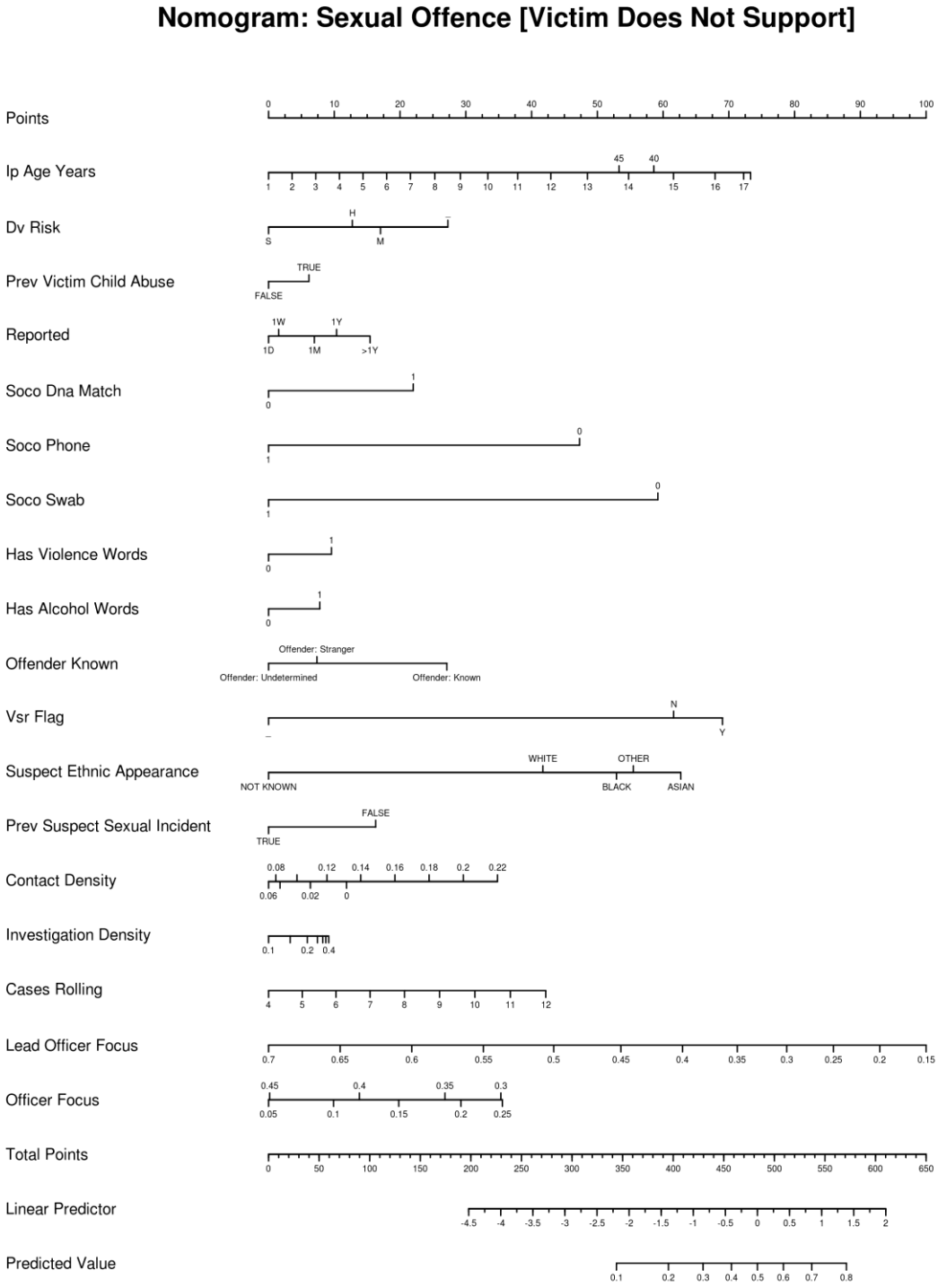


Source: WMP DAL 2019

6.2 Nomogram, all penetrative crimes: Charge



6.3 Nomogram, all penetrative crimes: Victim does not Support



7 References

ONS. 2018. “Sexual Offending: Victimisation and the Path Through the Criminal Justice System - Office for National Statistics.” https://www.ons.gov.uk/peoplepopulationandcommunity/crimeandjustice/articles/sexualoffendingvictimisationandthepaththroughthecriminaljusticesystem/2018-12-13.

8 ANNEX – Methodology

8.1 Introduction

This analysis is based on observational data for crime records extracted from the WMP crime database for the period from January 2014 to October 2018. The unit of analysis is a combination of recorded crime reference, victim, and suspect.

Data were extracted from the Crimes, Socrates, and Oasis databases for the period from January 2014 - October 2018. Based on searches for *RAPE* and *PENETRATION* but without the terms *CONSENT* or *IMAGE* (though holding images of certain natures are crimes in and of themselves this was beyond the remit of this investigation).

This results in a total of 12,945 crimes for analysis of which 11,318 are not ongoing investigations. 317 of the crimes have multiple suspects, giving 11,702 units of analysis.

Making causal inferences based on observational data without an *a priori* model based on theory is problematic. This approach has the potential for confounding and encourages the use of convenience model specifications that do not include fundamental explanatory variables. These issues can lead to differences in magnitude, or even direction, of estimated effect sizes between different modelling techniques.

With only observational data available, variables of interest were selected firstly in consultation with a subject matter expert in the PPU (Public protection referral unit). The SME also assisted in the specification of an initial causal diagram. This approach avoids many of the issues related to including explanatory and control variables only on the justification of correlation with the dependent variable. These variables were then used to build an initial logistic regression model.

Informed by this initial regression model, several further data mining techniques were applied. Confounding, bias and model misspecification, are extant issues when data mining. To the extent that there is an agreement with the initial full model, the results are confirmatory.

8.2 Variable Selection

Variables of interest were selected in consultation with a subject matter expert from the PPU unit. In addition to features of the incident itself, these variables relate to:

- Victim Credibility
- Evidence Available
- Investigation Features
- Suspects and Victims prior criminal history.

Crimes and Oasis

Variable	Type	Comments	
cuc category grouping	factor	Final Clearup Category of the Incident	Charge: 1341, Evidential Difficulties: 2530, No Crime: 2344, No Suspect Identified: 1529, Open: 1679, Other: 409, Victim Does Not Support: 3549
npu	factor	Neighbourhood Policing Unit.	BE: 2418, BW: 3713, CV: 1620, DY: 1121, SH: 639, SW: 1410, WS: 1139, WV: 1321
vsr flag	factor	Victim Support Requested	NA: 1478, N: 11262, Y: 641
dv risk	factor	Risk of Domestic Violence. High, Medium, Standard	NA: 10159, H: 2129, M: 631, S: 462
report method desc	factor		FRONT OFFICE: 518, HELP DESK/CONTACT CENTRE: 5813, PATROL: 1805, PPU: 2360, OTHER: 2885
offence type desc	factor		Other: 6269, Child Abuse: 3888, Domestic Abuse: 3224
victim sex	factor		FEMALE: 12258, MALE: 1123
has witness	logical		0.2%
offender known	factor		Undetermined: 1878, Known: 8675, Stranger: 2828
reported	factor	Same day, week, month, historic	Within 1 Day: 4811, 1 Week: 1629, 1 Month: 930, 1 Year: 2082, 5 Years: 1276, Historic (> 5 years): 2653
ip age years	numeric	IP Age at the time of the offence	Mean 22.4, SD: 12.6, Median: 19.2
suspect age years	numeric	Suspect Age at the time of the offence	Mean 29.1, SD: 12.9, Median: 26.7
days b4 reporting	numeric	Days before crime was reported	Mean 1727, SD: 3914, Median: 10.1
days b4 soco	numeric	Days before Scene of Crime data was collect	Mean 39.3, SD: 101.3, Median: 4.8
days b4 finished	numeric	Days an Incident is Open	
days b4 finished censored	numeric	Days an Incident is Open (+ Crimes that are still open)	
hours b4 first investigation	numeric	Hours between reporting and first investigation note	
suspect ethnic appearance	factor		WHITE: 4380, ASIAN: 1510, BLACK: 1256, NOT KNOWN: 5990, OTHER: 245

Variable	Type	Comments	
ip ethnic appearance	factor		WHITE: 7926, ASIAN: 1640, BLACK: 1235, NOT KNOWN: 2172, OTHER: 408
ip age group	factor	Grouping of the ip age in years	IP Age: 0 - 12: 2396, 13 - 16: 2721, 17, 18, 19: 1717, 20s: 3080, 30s: 1651, 40+: 1816

Scene of Crime

Variable	Type	Comments	
has soco	logical	Is there scene of crime data associated with this incident?	22.7%
soco dna match	logical	Is there a dna match to a suspect?	1.6%
soco swab	logical	Were swabs taken?	4.8%
soco phone	logical	Is the phone of the IP or Suspect available?	13.8%
soco cctv	logical	Is CCTV available?	1.1%

Investigation Notes

A word2vec classifier (Mikolov, Le, and Sutskever 2013) was trained on a text corpus consisting of crime descriptions from investigation notes. The resulting word vectors were then used to identify synonyms for words related to violence, alcohol, drugs and a loss of consciousness.

The resulting synonyms, including some misspellings, were then used to classify each incident.

- violence words: *threat, afraid, afriad, fear, intimida, terrified, duress, violence, frightened, scared, aggressive, abusive, suffocat, torture, strang, choke, hit, punch, knife, slap, stab, beat, slash, fist, kick, butt, strike, overpower, tying, gunpoint, ligature, tie, whack, pin, smack, windpipe, wind, headbutt.*
- alcohol words: *drunk, drink, tipsy, lager, drank, brandy, vodka, cider, cocktails, wine, alcohol, rum, wkd, prosecco, beer, pint, pints, whiskey, champagne, carling, strongbow, desperado, malibu, cans, stella, archers, gin, smirnoff, frosty, jaegerbombs, lambrini, budweiser.*
- drug words: *drug, cocaine, amphetamines, mkat, heroin, herion, heroine, crack, mdma, canabis, ketamin, drugged, poppers, spiked, pills, spliffs, sniffed, snort, powder, inhal.*

- dazed words: *groggy, conciousness, disorientated, intoxic, daze, unconcious, blacked, blacking, drowsy, spaced, dizzy, haz, numb, woozy.*

Variable	Type	Comments	
has violence words	logical	Based on text mining of the Investigation Notes	50.0%
has alcohol words	logical		30.2%
has drug words	logical		17.2%
has dazed words	logical		12.5%

Derived

Variable	Type	Comments	
cases rolling 56	numeric	Competeing caseload. Based on a rolling daily mean over the prior 8 weeks.	
officer focus eb	numeric	No of Officers / No of Investigation Notes	
lead officer focus eb	numeric	No of Notes by the lead officer / No of Investigation Notes	Here lead officer is the officer responsible for the most notes
investigation density eb	numeric	No of Days with an Investigation Note / Elapsed Days	
contact density eb	numeric	No of Contact Attempts / No of Investigation Notes	

The four proportions (officer focus, lead officer focus, investigation density, and contact density) are transformed using empirical Bayes.

All of the data is used to form a prior for the underlying distribution of the proportion – for example, the proportion of investigation relating to contact attempts. The data for a specific observation then evaluates a posterior belief. This transformation shrinks investigations with relatively few investigation notes to the mean of the empirical distribution. The posterior estimate is then interpretable as the evidence that an observation differs from the overall distribution mean.

In this case, the prior is expressed as a beta distribution $X \sim \text{Beta}(\alpha_0, \beta_0)$. The hyperparameters of this beta distribution are found by fitting a beta-binomial distribution to the data using maximum likelihood. This gives more consideration to crimes with a higher number of notes and is less sensitive to noise than fitting a Beta distribution directly.

When we evaluate any individual to crime, we start with the overall prior, and update based on the attributes of the incident. For example, for contact density, this is evaluated as $\frac{\text{No of Contact Attempts}+\alpha_0}{\text{No of Investigation Notes}+\alpha_0+\beta_0}$

Criminal History

Variable	Type	Comments
s domestic abuse,	numeric	Suspects' previous history
s child abuse,	numeric	
s suspect assault,	numeric	
s suspect damage,	numeric	
s suspect rape,	numeric	
s suspect sexual incident,	numeric	
s suspect theft,	numeric	
s suspect threat,	numeric	
v domestic abuse,	numeric	Victims' previous history
v child abuse,	numeric	
v victim assault,	numeric	
v victim damage,	numeric	
v victim rape,	numeric	
v victim sexual incident,	numeric	
v victim theft,	numeric	
v victim threat	numeric	

The number of crime records for each nominal of each time is aggregated at the quarter level. From this a 10 year exponential moving average is calculated such that more recent crimes have a higher weighting, with exponential decay to zero weighting after 40 quarters (using the ratio = 2/(40 + 1)).

8.3 Models

Logistic regression

We start with a main effects model with no interactions based on the variables identified by the SME. This is a *one vs all* model where the outcome is compared with all other outcomes. Open crimes are not included.

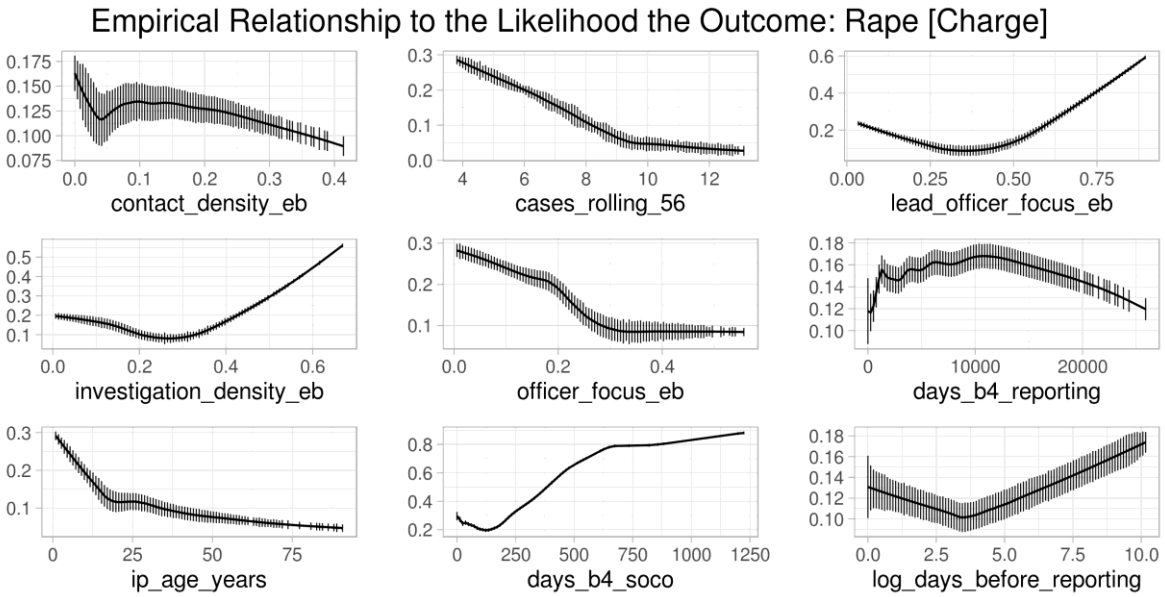
Four models are fitted:

- *"Charge"* outcome for all sexual incidents.
- *"Charge"* outcome for crimes categorised as rape.
- *"Victim Does not support"* outcome for all sexual incidents.
- *"Victim Does not support"* outcome for crimes categorised as rape.

Preparation

The logistic model assumes that continuous explanatory variables are linear with the logit of the dependent variable, that effects are additive, and that observations are independent.

Non-parametric regression using loess to estimate the relationship between the probability of a charge and the continuous explanatory variables reveals non-linearity and non-monoticity. This will be an issue if uncorrected. Here we transform the variables using restricted cubic splines.



We performed a redundancy analysis over the explanatory variables. This applies parametric additive models (using regression splines) to determine how well each variable can be predicted from the remaining variables. Here the variable has_soco is predictable (R^2 > 0.7) from the other variables so we drop it from the model.

There is no remaining issue of high correlation between explanatory variables. (The highest correlation between the remaining variables is 0.48 between offender_known: Stranger and offender_known: Known).

The Model fitting is based on likelihood with a regularisation penalty (based on 5 fold cross validation) to avoid overfitting.

Fitted Model

Based on the fit, if the intent was parsimony and predictive ability only, there are some variables that could be dropped without impacting the model's ability to seperate the

classes. Here, we are interested in the variable effect sizes and their value in explaining the outcomes of incidents.

```
Logistic Regression Model for Charged Rape.


               Wald Statistics

 Factor                        Chi-Square d.f. P
 ip_age_years                    89.24     5   <.0001
  Nonlinear                      65.12     4   <.0001
 dv_risk                         12.53     3   0.0058
 prev_victim_rape                 6.01     1   0.0142
 prev_victim_child_abuse         15.86     1   0.0001
 prev_victim_domestic_abuse       0.81     1   0.3683
 prev_victim_sexual_incident      0.05     1   0.8272
 reported                        19.53     4   0.0006
 soco_dna_match                   2.74     1   0.0977
 soco_phone                     265.54     1   <.0001
 soco_swab                       72.96     1   <.0001
 soco_cctv                        9.55     1   0.0020
 has_violence_words               0.21     1   0.6444
 has_alcohol_words                0.11     1   0.7411
 has_drug_words                   0.05     1   0.8257
 has_dazed_words                  0.56     1   0.4557
 offender_known                   2.92     2   0.2319
 vsr_flag                         4.14     2   0.1260
 victim                           0.00     1   0.9886
 mo_repeat_victim                 0.00     1   0.9974
 suspect_ethnic_appearance      119.77     4   <.0001
 prev_suspect_rape               37.08     1   <.0001
 prev_suspect_child_abuse         0.03     1   0.8698
 prev_suspect_sexual_incident    12.00     1   0.0005
 contact_density_eb              17.41     2   0.0002
  Nonlinear                      17.35     1   <.0001
 investigation_density_eb       119.12     2   <.0001
  Nonlinear                     114.78     1   <.0001
 cases_rolling_56               198.37     2   <.0001
  Nonlinear                       0.95     1   0.3293
 lead_officer_focus_eb          165.23     2   <.0001
  Nonlinear                      47.66     1   <.0001
 officer_focus_eb                 5.51     2   0.0637
  Nonlinear                       2.97     1   0.0850
 TOTAL NONLINEAR                268.94     9   <.0001
 TOTAL                         1151.90    47   <.0001


                     Model Likelihood     Discrimination    Rank Discrim.
                        Ratio Test           Indexes           Indexes
 Obs          8277    LR chi2    1791.13    R2       0.362    C       0.863
  0           7207    d.f.            47    g        1.840    Dxy     0.726
  1           1070    Pr(> chi2) <0.0001    gr       6.296    gamma   0.726
 max |deriv| 7e-11                         gp       0.160    tau-a   0.163
                                           Brier    0.085

                                    Coef     S.E.   Wald Z Pr(>|Z|)
 Intercept                           2.3371 0.5050   4.63 <0.0001
 ip_age_years                       -0.0335 0.0268  -1.25 0.2110
 ip_age_years'                      -1.6432 0.4273  -3.85 0.0001
 ip_age_years''                     16.1736 3.1684   5.10 <0.0001
 ip_age_years'''                   -26.3269 4.8566  -5.42 <0.0001
 ip_age_years''''                   14.0504 2.6539   5.29 <0.0001
 dv_risk=H                           0.4329 0.1339   3.23 0.0012
 dv_risk=M                           0.0438 0.2307   0.19 0.8495
 dv_risk=S                          -0.0654 0.2596  -0.25 0.8011
```

```
prev_victim_rape                       -0.3521 0.1436  -2.45 0.0142
prev_victim_child_abuse                -0.4257 0.1069  -3.98 <0.0001
prev_victim_domestic_abuse             -0.0893 0.0993  -0.90 0.3683
prev_victim_sexual_incident             0.0280 0.1283   0.22 0.8272
reported=1W                            -0.3774 0.1377  -2.74 0.0061
reported=1M                            -0.3904 0.1730  -2.26 0.0241
reported=1Y                            -0.5211 0.1318  -3.95 <0.0001
reported=>1Y                           -0.2250 0.1214  -1.85 0.0637
soco_dna_match                         -0.4347 0.2625  -1.66 0.0977
soco_phone                              1.5328 0.0941  16.30 <0.0001
soco_swab                               1.3406 0.1569   8.54 <0.0001
soco_cctv                               0.7964 0.2577   3.09 0.0020
has_violence_words                     -0.0384 0.0831  -0.46 0.6444
has_alcohol_words                      -0.0317 0.0960  -0.33 0.7411
has_drug_words                         -0.0243 0.1103  -0.22 0.8257
has_dazed_words                        -0.0995 0.1334  -0.75 0.4557
offender_known=Offender: Known          0.1373 0.1701   0.81 0.4194
offender_known=Offender: Stranger      -0.0607 0.1839  -0.33 0.7414
vsr_flag=N                              0.3598 0.2365   1.52 0.1282
vsr_flag=Y                              0.6783 0.3348   2.03 0.0428
victim=MALE                            -0.0019 0.1312  -0.01 0.9886
mo_repeat_victim                       -0.0005 0.1420   0.00 0.9974
suspect_ethnic_appearance=ASIAN        -0.3388 0.1166  -2.91 0.0037
suspect_ethnic_appearance=BLACK        -0.0595 0.1116  -0.53 0.5940
suspect_ethnic_appearance=NOT KNOWN    -1.2581 0.1175 -10.71 <0.0001
suspect_ethnic_appearance=OTHER        -0.1718 0.2479  -0.69 0.4884
prev_suspect_rape                       0.7774 0.1277   6.09 <0.0001
prev_suspect_child_abuse                0.0178 0.1087   0.16 0.8698
prev_suspect_sexual_incident            0.4230 0.1221   3.46 0.0005
contact_density_eb                     10.5675 2.6040   4.06 <0.0001
contact_density_eb'                   -21.6238 5.1917  -4.17 <0.0001
investigation_density_eb               -9.4542 1.0290  -9.19 <0.0001
investigation_density_eb'              14.7400 1.3758  10.71 <0.0001
cases_rolling_56                       -0.3499 0.0434  -8.07 <0.0001
cases_rolling_56'                       0.0562 0.0576   0.98 0.3293
lead_officer_focus_eb                  -2.1090 0.8116  -2.60 0.0094
lead_officer_focus_eb'                  6.6861 0.9684   6.90 <0.0001
officer_focus_eb                       -0.5438 0.8945  -0.61 0.5432
officer_focus_eb'                       1.8392 1.0678   1.72 0.0850
```

Calibration and Diagnostics

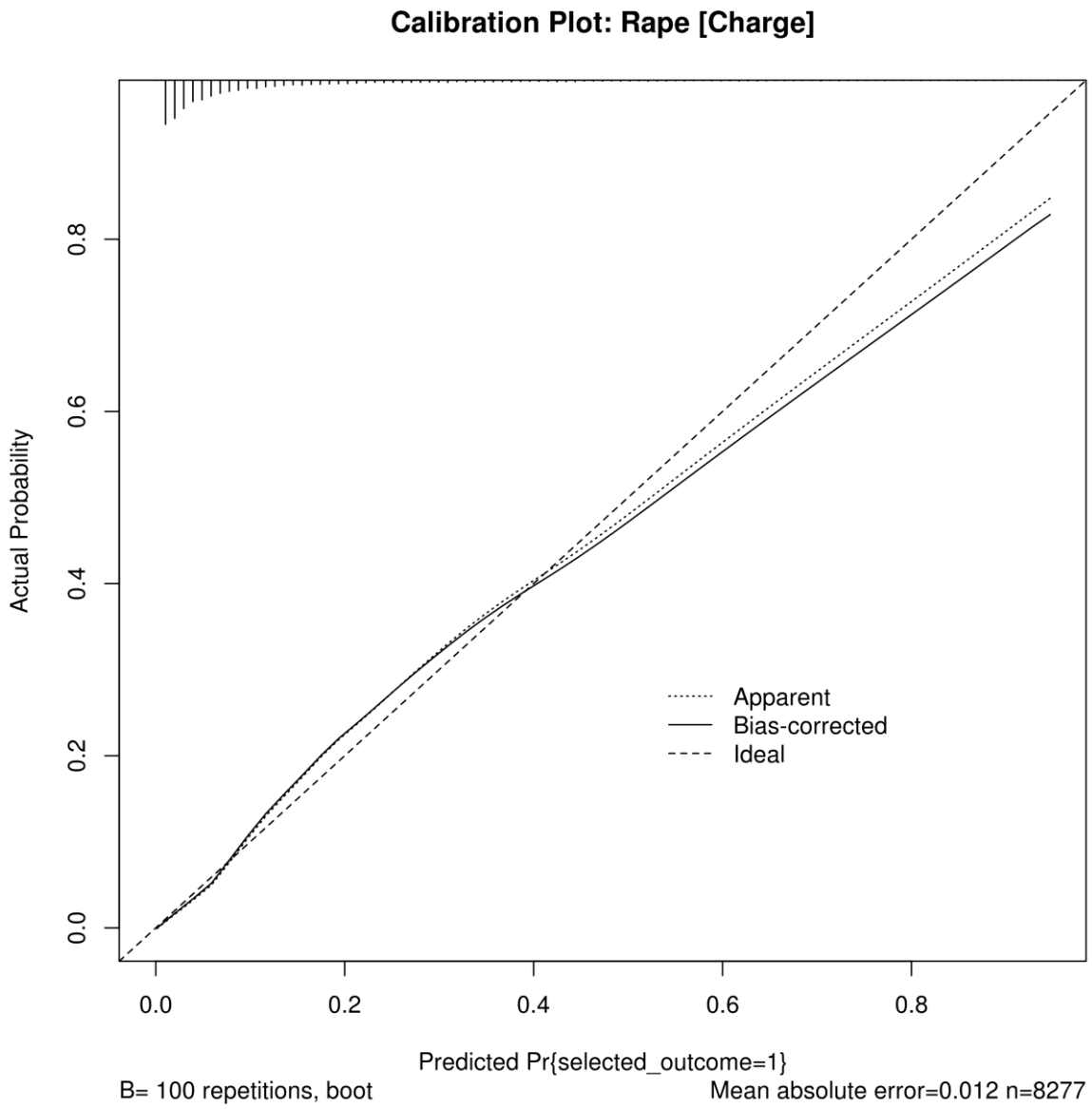
Colinearity as measured by VIF show no issues. The highest VIF is 3.3 (This does not include the continuous variates using restricted cubic splines where high colinearity is to be expected.)

There is some evidence of high influence observations related to the variable soco_cctv. This is as a result of the low incidence of crimes with CCTV available.

Validating the model using 100 bootstrap resamples shows some issues with the fit of the model. The output probability is not well calibrated as shown on the calibration plot and this is reflected in a hosmer-lemeshow-goodness-of-fit test ($p < 1E6$). This has little impact on the ability of the model to separate between different outcomes, but does indicate we should be judicious quoting results on a probabilistic scale.

```
          index.orig training    test optimism index.corrected   n
Dxy           0.7230   0.7272  0.7175   0.0097          0.7133 100
R2            0.3600   0.3652  0.3538   0.0114          0.3486 100
Intercept     0.0000   0.0000 -0.0323   0.0323         -0.0323 100
Slope         1.0000   1.0000  0.9724   0.0276          0.9724 100
Emax          0.0000   0.0000  0.0119   0.0119          0.0119 100
D             0.2147   0.2177  0.2106   0.0071          0.2076 100
```

```
U            -0.0002  -0.0002  0.0001  -0.0004          0.0001 100
Q             0.2149   0.2179  0.2104   0.0075          0.2075 100
B             0.0846   0.0838  0.0853  -0.0015          0.0862 100
g             1.8259   1.8507  1.7998   0.0509          1.7750 100
gp            0.1593   0.1597  0.1580   0.0017          0.1575 100

n=8277   Mean absolute error=0.012   Mean squared error=0.00036
0.9 Quantile of absolute error=0.025
```



Model effect Sizes

The out of sample AUC for "charge", and "Victim does not support" is 0.857 and 0.701 respectively for rape crimes. (Compared to 0.871 and 0.707 models including all sexual crimes).

These are models have a great ability to separate the outcome classes and predict well between "Charged", "Victim does not support", and other outcomes.

Other linear models

Relaxed Lasso

To improve out-of-sample predictions, lasso regression (Tibshirani 1996) penalises model complexity by forcing the the value of model coefficients towards zero. This improves the variance of the model at the expense of obtaining biased (in the statistical sense of the term) coefficients. Bootstrapping then leads to over optimistic confidence intervals due to shrunken standard errors.

An approach to reducing this bias is to run feature selection and model fitting in 2 steps:

- a lasso to perform feature selection
- an unconstrained (or further lasso) model fit to the reduced set of features.

This is known as the relaxed lasso (Meinshausen 2007). The magnitude of the relaxed Lasso coefficients is typically larger than that of the Lasso coefficients.

However, the lasso is intended to be a one-stop solution. Applying it as a feature selection technique to feed variables into another model does not penalise for the model selection. Again, possibly leading to some bias. (This can be observed when bootstrapping. The lasso technique does not always yield the same set of explanatory variables). The purpose here is parsimony over predictive accuracy. However, estimates are trustworthy only in magnitude and direction.

The output of this model is labelled relaxed lasso in the appendix.

Bayesian Penalisation

Similar to the lasso approach. We apply the rstanarm package to perform penalised Bayesian estimation. This allows us to apply lasso shrinkage to individual variables.

The output of this model is labelled rstanarm in the appendix.

The parameter estimates are consistent with the relaxed lasso fit.

Mining for high-level interactions

Higher-level interaction were investigated by adding all second-level and third-level interaction terms. A lasso glm was then used to perform feature selection.

This approach necessarily introduces correlated predictors, lasso tends to choose one and push the others to 0, therefore omitting a significant proportion of informative variables.

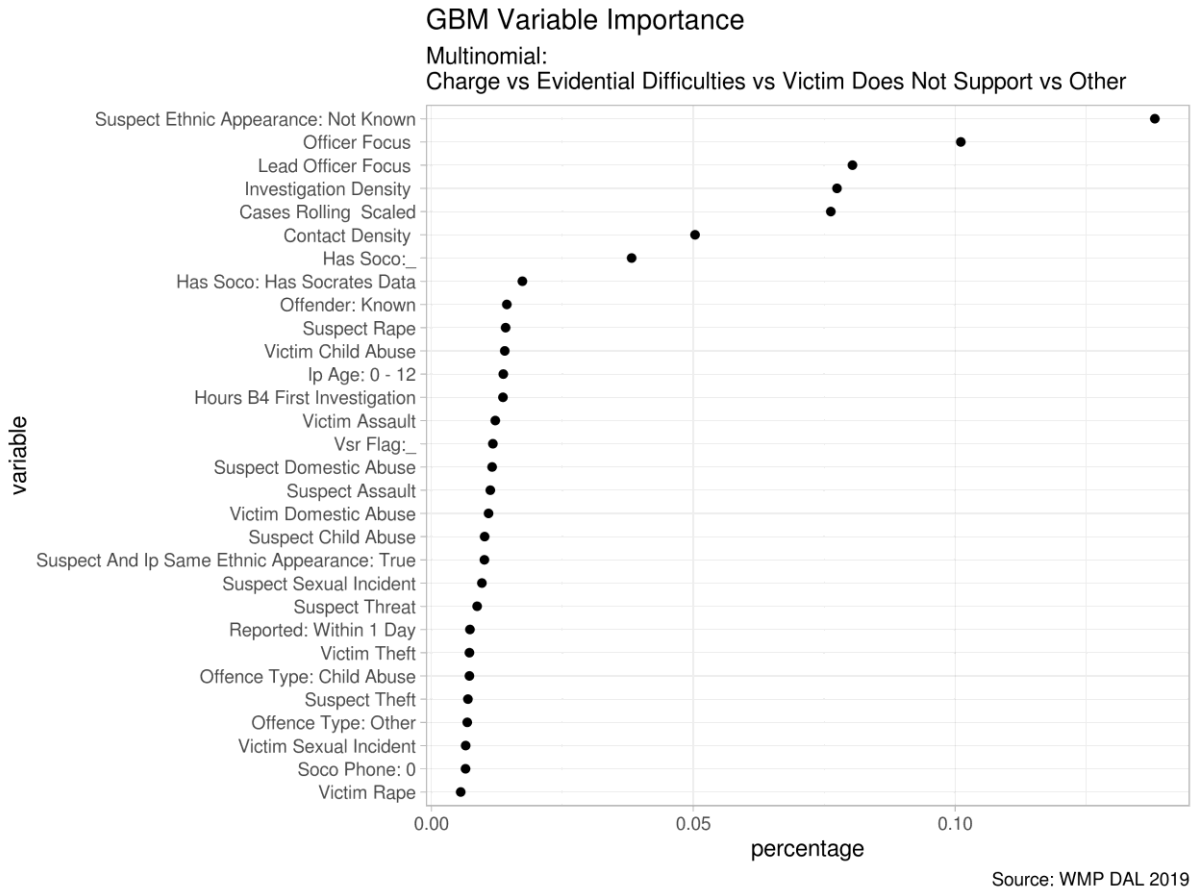
No interactions were found in addition to main effects with a business interpretation that are stable between bootstraps. This is indicative that there are no highly predictive interaction terms.

Tree Based Models

As an alternative assessment of feature importance, we fit a variety of one-vs-all, and multinomial models to the data using a gradient boosting model (similar to random

forests) and decision trees. Random forests do not have scale or linearity assumptions and are more able to include complex interaction terms between the variables.

The variable importance agrees well with the logistic model, and also to an analysis of weight of evidence of each of the individual explanatory variables.

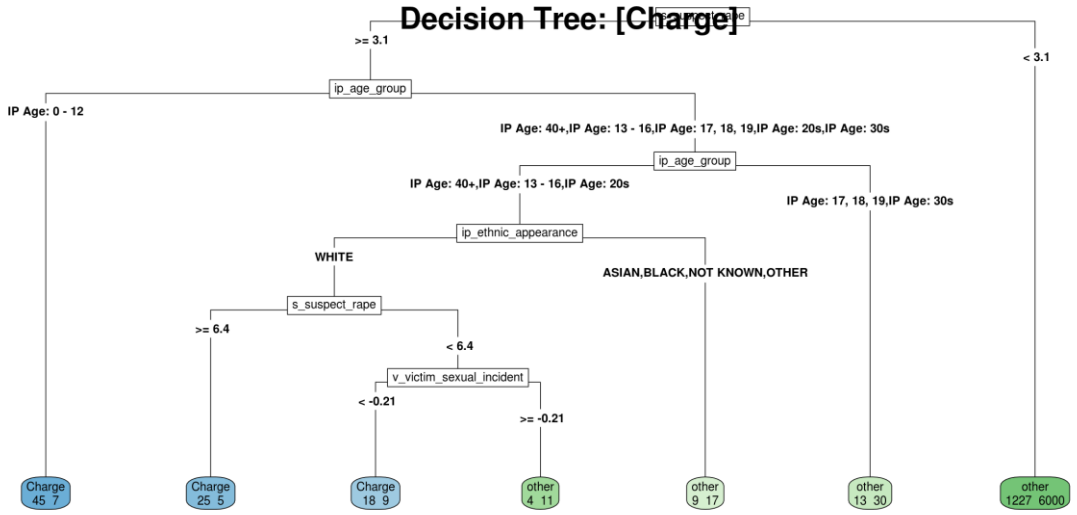


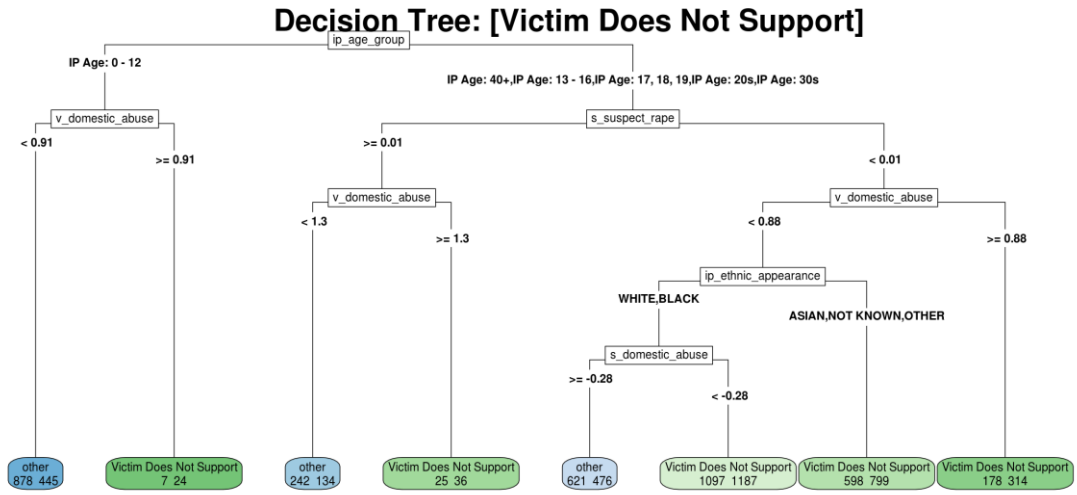

```
Exploratory data analysis and variable screening for binary classification models using inf
ormation theory (WOE and IV).

                                Variable           IV      PENALTY         AdjIV
19             suspect_ethnic_appearance 7.997949e-01 9.466215e-02  0.7051327465
8                              has_soco 6.612953e-01 4.429103e-02  0.6170042291
11                           soco_phone 5.416652e-01 3.523867e-02  0.5064265616
22               cases_rolling_56_scaled 6.014685e-01 1.056034e-01  0.4958651811
24                      officer_focus_eb 3.898257e-01 5.383678e-02  0.3359889128
25                  lead_officer_focus_eb 3.537420e-01 6.078280e-02  0.2929591674
26               investigation_density_eb 2.830179e-01 6.096413e-02  0.2220538081
21 suspect_and_ip_same_ethnic_appearance 2.650093e-01 4.731203e-02  0.2176972337
7                             ip_age_group 2.614923e-01 5.119354e-02  0.2102987149
3                         offence_type_desc 1.504596e-01 1.462649e-02  0.1358330706
32                           s_suspect_rape 1.273008e-01 4.194524e-03  0.1231062631
34                         s_suspect_theft 1.297273e-01 2.181667e-02  0.1079106170
13                          offender_known 1.489855e-01 4.169127e-02  0.1072941932
29                           s_child_abuse 1.088186e-01 2.863881e-03  0.1059546735
27                       contact_density_eb 1.685690e-01 7.147131e-02  0.0970977084
35                         s_suspect_threat 9.620380e-02 6.350482e-03  0.0898533182
30                        s_suspect_assault 8.851240e-02 1.990326e-03  0.0865220783
33               s_suspect_sexual_incident 9.497373e-02 1.113295e-02  0.0838407800
31                         s_suspect_damage 8.157190e-02 4.125302e-03  0.0774465931
4                        report_method_desc 1.014738e-01 2.878137e-02  0.0726924455
10                             soco_swab 6.840946e-02 1.439614e-03  0.0669698414
23            hours_b4_first_investigation 6.543093e-02 9.362334e-03  0.0560686002
```

28	s_domestic_abuse	5.640775e-02	3.903009e-03	0.0525047394
20	ip_ethnic_appearance	7.503119e-02	2.260996e-02	0.0524212351
14	reported	6.967269e-02	2.708818e-02	0.0425845043
40	v_victim_rape	4.919296e-02	1.825489e-02	0.0309380752

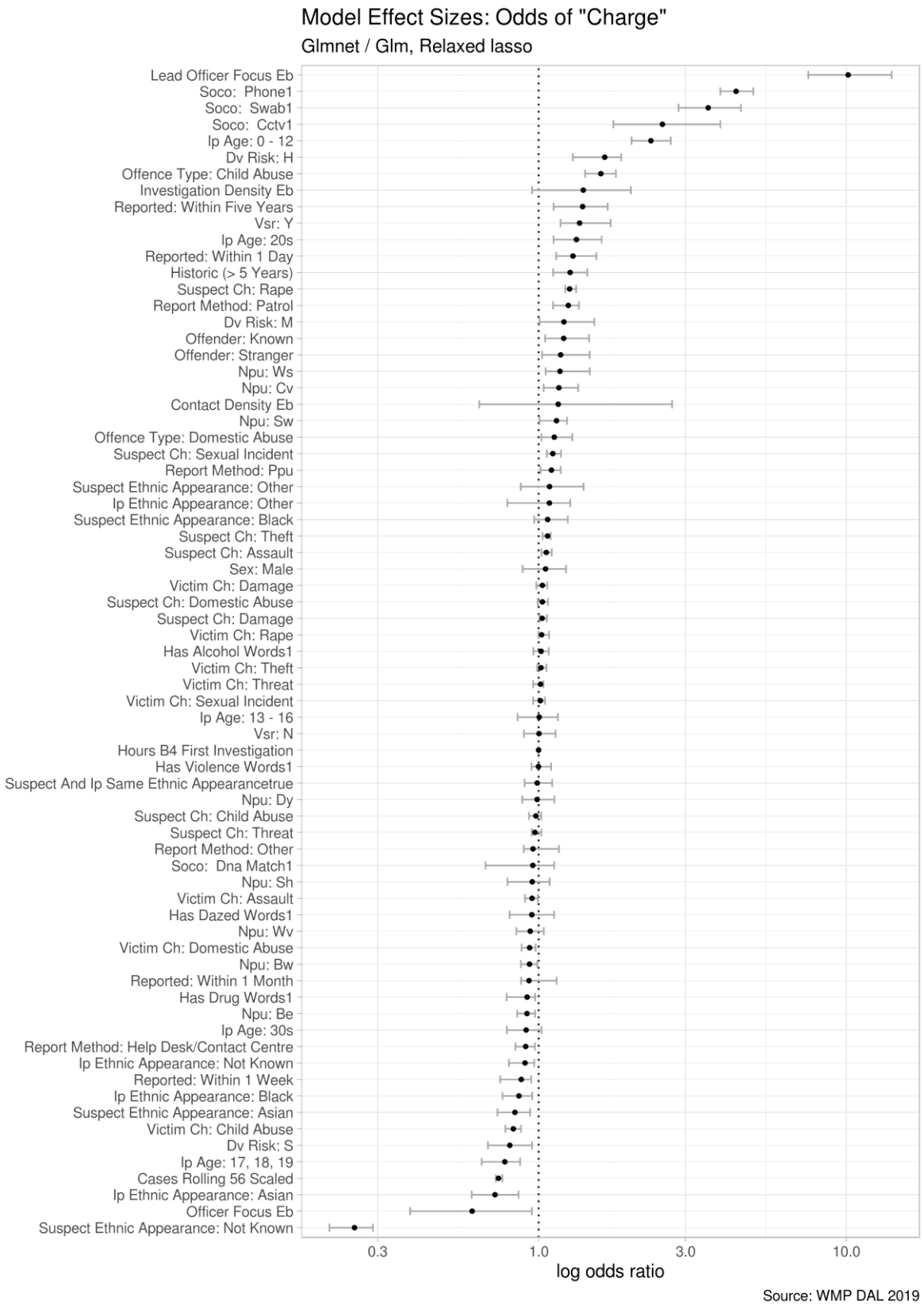
The AUC of one-vs-all models is no better than the logistic models. This is indicative that there are no highly predictive interaction terms not included in the linear models.

Decision trees based on CART (Breiman, Friedman, Olshen, and Stone 1984) though not highly discriminative demonstrate the high leverage of the suspect's previous criminal history on the decision to charge.



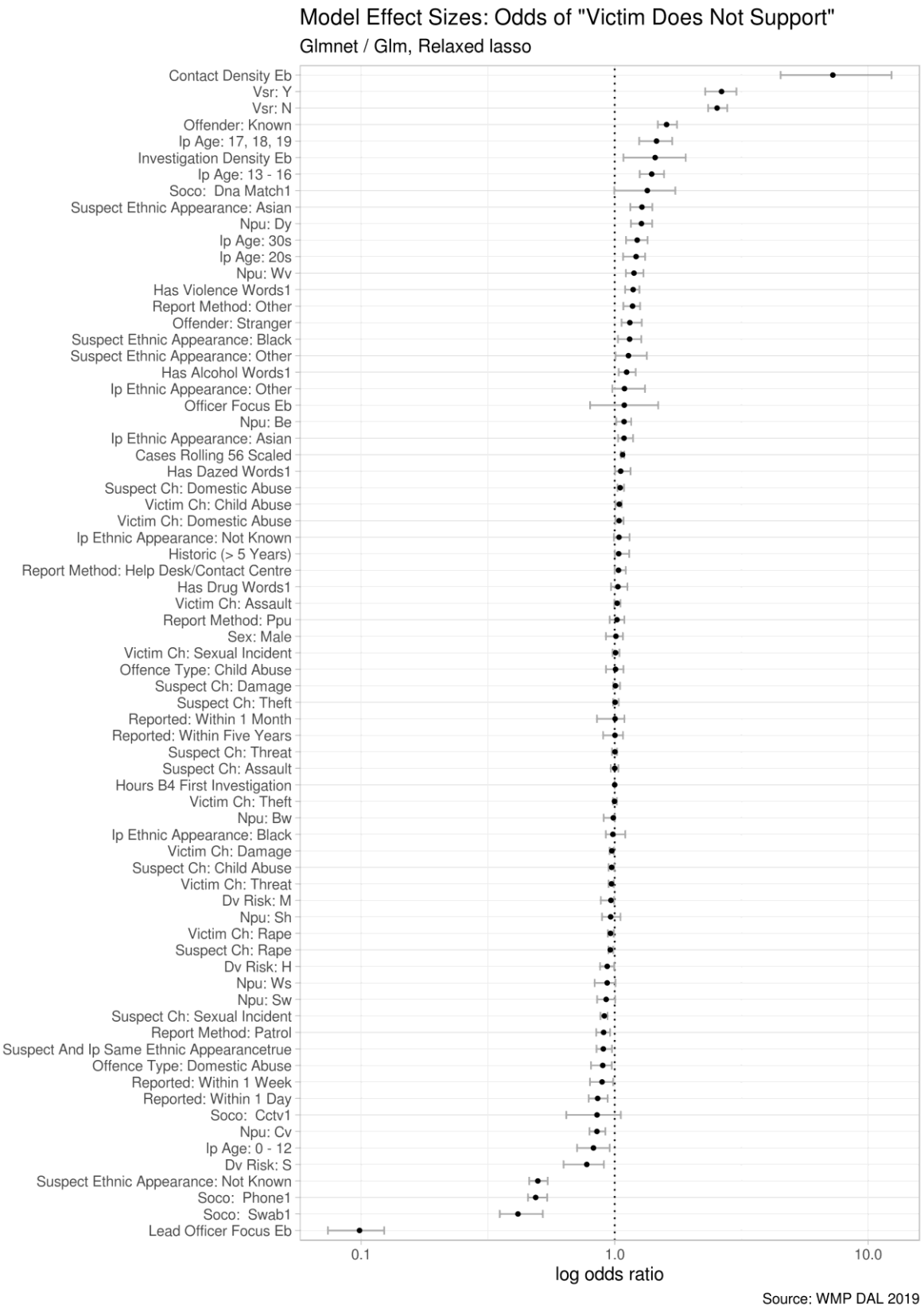


Relaxed Lasso, Relative Odds of Charge.



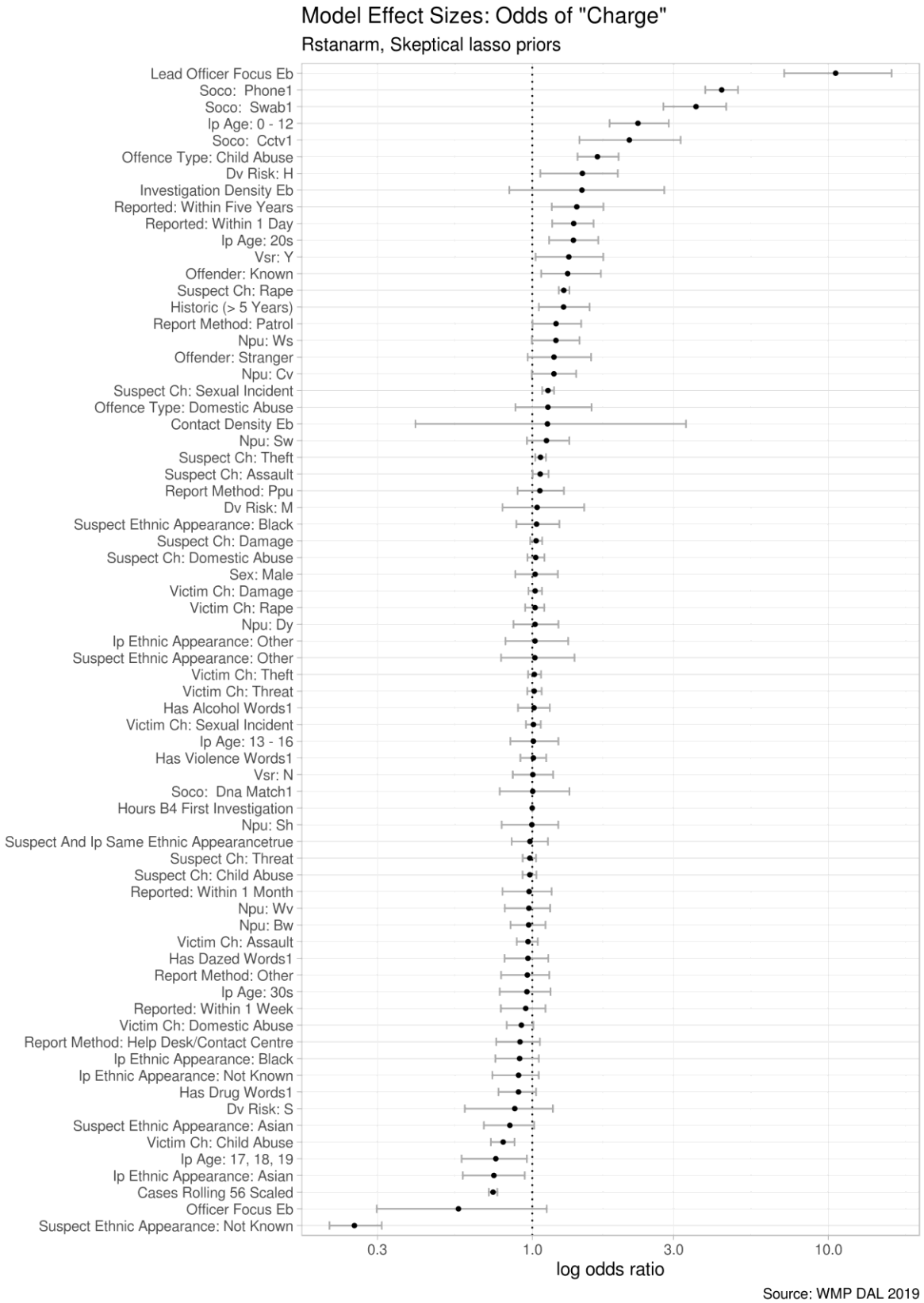
Source: WMP DAL 2019

Relaxed Lasso, Relative Odds of Victim Does Not Support.



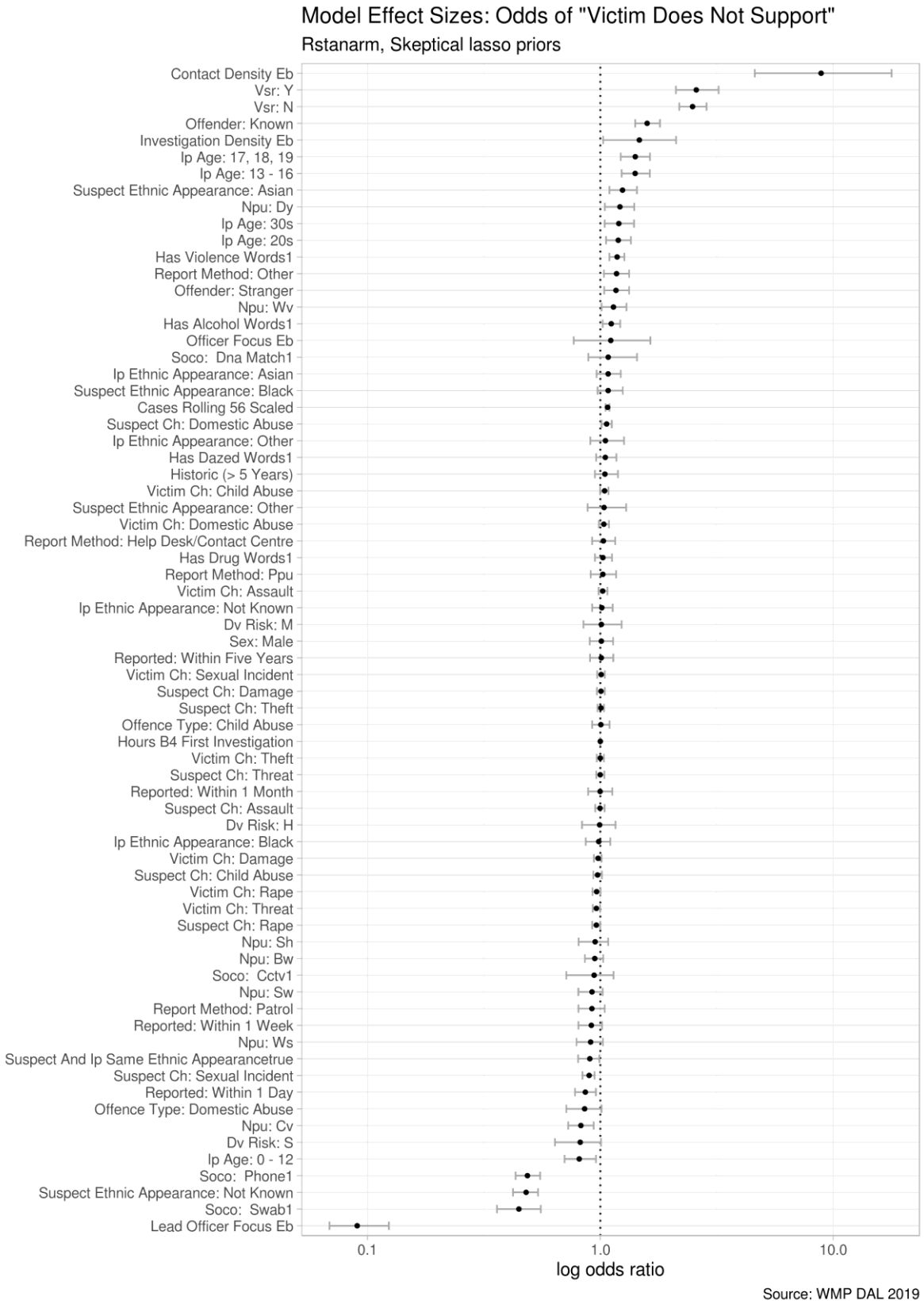
Source: WMP DAL 2019

Bayesian Regression, Relative Odds of Charge.



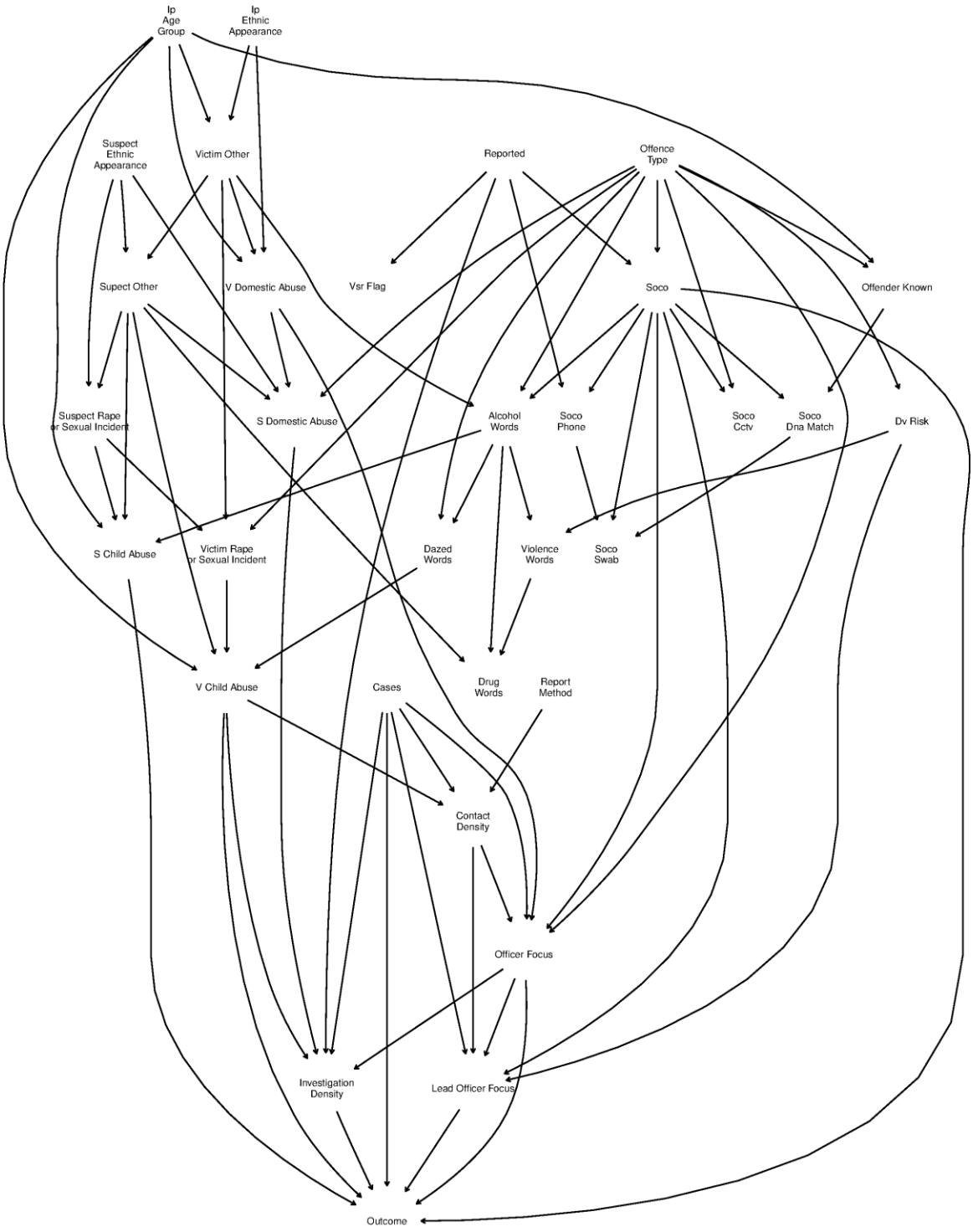
Source: WMP DAL 2019

Bayesian Regression, Relative Odds of Victim Does Not Support.



Empirical Directed Acyclic Graph (DAG)

The figure below show an empirical bayesian network structure learned from the crime data for the "Victim Does Not Support" outcome. This largely agrees with the logistic models based on SME input. For example, the outcome is directly related to the number of open cases, the lead officer focus, the investigation density, the availability of Scene of Crime evidence and previous history of abuse.



• References

Meinshausen, Nicolai. 2007. “Relaxed Lasso.” *Computational Statistics & Data Analysis* 52 (1): 374–93.

Mikolov, Tomas, Quoc V. Le, and Ilya Sutskever. 2013. “Exploiting Similarities Among Languages for Machine Translation.” *CoRR* abs/1309.4168. http://arxiv.org/abs/1309.4168.

Tibshirani, Robert. 1996. “Regression Shrinkage and Selection via the Lasso.” *Journal of the Royal Statistical Society: Series B (Methodological)* 58 (1): 267–88.