CRIMINOLOGY
RESEARCH GRANT

Automatic text mining of family and domestic violence records: An approach for future research

George Karystianis
Armita Adily
Rina Carines Cabral
Tony Butler

Report to the Criminology
Research Advisory Council
Grant: CRG 34/15–16

July 2021

This project was supported by a Criminology Research Grant.
The views expressed are the responsibility of the author and are not necessarily those of the Council.

© Australian Institute of Criminology 2021

Apart from any fair dealing for the purpose of private study, research, criticism or review, as permitted under the *Copyright Act 1968* (Cth), no part of this publication may in any form or by any means (electronic, mechanical, microcopying, photocopying, recording or otherwise) be reproduced, stored in a retrieval system or transmitted without prior written permission. Inquiries should be addressed to the publisher.

Published by the Australian Institute of Criminology
GPO Box 1936 Canberra ACT 2601
Tel: (02) 6268 7166
Email: front.desk@aic.gov.au
Website: www.aic.gov.au/crg

ISBN 978 1 922478 12 2 (Online)

This research was supported by a Criminology Research Grant. The views expressed are those of the author and do not necessarily reflect the position of the Criminology Research Advisory Council or the Australian Government.

This report was reviewed through a double-blind peer review process.

Edited and typeset by the Australian Institute of Criminology.

Contents

v Acknowledgements

vi Acronyms

vii Abstract

viii Executive summary

viii Family and domestic violence police records

viii Method

ix Results

ix Conclusion

1 Introduction

1 Family and domestic violence and mental illness

2 Web Computerised Operational Policing System

2 Police event narratives

3 Text mining

3 Study scope

4 Research aims

4 Research significance

5 Methods

5 Definitions

6 Use of text mining to extract information of interest

9 Data linkage between WebCOPS and NSW Health administrative data

10 Comparing the extracted mental disorder mentions against the WebCOPS structured fields

10 Predictive analytics methods

11 Results

11 Mental disorder mentions in police narratives

14 Mentions of abuse types and victim injuries

18 Data linkage between WebCOPS and NSW Health administrative data

18 Comparing the extracted mental disorder mentions against the WebCOPS structured fields

19 Text mining FDV event narratives to report trends in domestic violence

21 Predictive analytics results

22 Limitations

23 Discussion

25 References

28 Appendix

Figures

20 Figure 1: Trends in the five most common mental disorder mentions for male POIs in FDV police event narratives, 2005–2016

20 Figure 2: Trends of the five most common mental disorder mentions for female POIs in FDV police event narratives, 2005–2016

21 Figure 3: Trends of the most common types of victim injuries in police narratives of FDV events

Tables

11 Table 1: FDV event narratives with mentions of mental disorder (including traumatic brain injury and dementia), Jan 2005 – Dec 2016

12 Table 2: Self and third-party mentions of types of mental disorder in FDV event narratives at the first level of ICD-10 for POIs and victims, Jan 2005 – Dec 2016

13 Table 3: Self and third-party mentions of the top 10 mental disorders at the second level of ICD-10 in the FDV event narratives for POIs and victims, Jan 2005–Dec 2016

14 Table 4: Self and third-party mentions of the top 10 mental disorders at the third level of ICD-10 in the FDV event narratives for POIs and victims, Jan 2005 – Dec 2016

15 Table 5: FDV event narratives with mentions of abuse types

16 Table 6: FDV event narratives according to the number of mentions of abuse types

17 Table 7: FDV event narratives with mentions of victim-sustained injury types

17 Table 8: FDV event narratives according to the number of injury types

18 Table 9: FDV events with TM extracts for mentions of mental disorder in WebCOPS event narratives against the number of FDV events with a diagnosis of mental disorder for victims or POIs, 2005–2016

19 Table 10: Comparison of FDV events with mentions of mental disorder in a WebCOPS fixed field versus TM extracted information

21 Table 11: Performance of naïve Bayes, random forest and logistic regression methods for predicting breach of AVOs by POIs involved in several FDV events

28 Table A1: ICD-10 classifications of mental and behavioural disorders used to map the extracted and standardised mental health disorder mentions at three levels, plus additional categories

Acknowledgements

The authors would like to thank the New South Wales Police Force for their assistance with this project, particularly Dr Chris Devery, Dr Christie Wallace, John Blanchette, Erin Sharland, and Nicole Grant. This research was supported by a Criminology Research Grant (34/15–16) from the Australian Institute of Criminology.

Acronyms

ABS	Australian Bureau of Statistics
AIHW	Australian Institute of Health and Welfare
APDC	Admitted Patient Data Collection
AVO	Apprehended violence order
CHeReL	Centre for Health Record Linkage
EDDC	Emergency Department Data Collection
FDV	Family and domestic violence
ICD-10	International Classification of Diseases, 10th revision
IPV	Inter-partner violence
NSWPF	New South Wales Police Force
POI	Person of interest
TM	Text mining
WebCOPS	Web Computerised Operational Policing System
WHO	World Health Organization

Abstract

This body of work focuses on the design and implementation of an automated text-mining method that extracted information from a large-scale set of family and domestic violence (FDV) police records (*n*=492,393) and on whether this information can be used for further research. We report an evaluation of the method and the results from its application to the large dataset. These include trends over time for mentions of mental health problems obtained from police records for individuals involved in FDV events; specificity and sensitivity when comparing the extracted mental disorder mentions with NSW Health data; and the development of a predictive analytics approach to breaches of apprehended violence orders based on the extracted information. Findings indicate not only that text mining the free-text FDV police records can yield substantially useful and previously unknown information but also that text mining can fuel predictive analytics that can indicate high-risk offenders in the FDV area, impacting early prevention and intervention policies in FDV cases.

Executive summary

Family and domestic violence police records

The NSW Police Force (NSWPF) attends a number of family and domestic violence (FDV) events each year, recording details of these events as both structured (ie coded) data and unstructured, free-text narratives. The longer, unstructured descriptive text contains information typically not available in the structured data, such as the types of abuse (including physical, psychological, emotional, and financial) by persons of interest (POIs), the injuries sustained by victims, and mental disorder descriptors for POIs and victims.

Method

We designed and implemented an automatic text-mining method to determine whether we could systematically identify this information from the free-text narratives within a large-scale dataset of FDV records supplied by the NSWPF. We demonstrated some FDV trends for male and female POIs. We also linked the extracted mental health mentions with the official diagnoses obtained from NSW Health data—namely, the Admitted Patient Data Collection (APDC) and the Emergency Department Data Collection (EDDC)—for the same individuals. Finally, we created a predictive model that suggests whether a repeated offender is likely to breach their apprehended violence order (AVO) based on past information. We used a training set of 200 recorded FDV events to design our approach, which is based on syntactical patterns in the text. We tested our method on an evaluation set of 100 FDV events with a returned average precision of 92.0 percent for POI and victim mental health mentions and a precision of 90.2 percent and 85.0 percent for abuse type and victim injuries, respectively.

Results

We applied our methodology to a dataset of almost half a million of FDV events (492,393) and reported the results. In this dataset, 77,995 (17%) events had mentions of mental disorders (including traumatic brain injury and dementia) for the POIs and/or the victims. 'Mood [affective] disorders' (eg bipolar disorder, depression) had the highest rates of mentions of mental illness among POIs (15,330, 18.7%) and victims (4,946, 22.7%). We found 71.32 percent (351,178) of events with mentions of abuse types, and more than one-third of events (177,117, 35.97%) mentioned victim injuries. 'Emotional/verbal abuse' (117,488, 33.46%) was the most common abuse type, followed by 'punching' (86,322 events, 24.58%) and 'property damage' (78,203 events, 22.27%). 'Bruising' was the most common form of injury sustained (51,455 events, 29.03%), with 'cut/abrasion' (51,284 events, 28.93%) and 'red marks/signs' (42,038 events, 23.71%) ranking second and third, respectively. After linking the extracted mental disorder mentions to official diagnoses obtained by NSW Health, our sensitivity was 15.6 percent—that is, of all FDV events, 15.6 percent had a mention of mental disorder that corresponded to a diagnosis from NSW Health (Figures 1 and 2). We identified some trends over time based on the extracted information. In police narratives for male POIs mentions of all common types of mental disorder increased over time, and for both male and female POIs mentions of 'major depressive disorder, single episode' increased over time. Finally, our predictive model indicated that, based on past information from the free-text narratives, there is a likelihood of 70 percent that a POI will breach their AVO.

Conclusion

The results suggest text mining can be used to automatically extract information from police-recorded FDV events that can support further public health research in the area, such as examining the relationship between abuse types and victim injuries; the relationship between gender and abuse types; and the risk of escalation for victims of domestic violence. Potential also exists for this extracted information to be linked to other information sources on diagnosis of mental health problems, and for these data to be used as inputs into models that can predict future offending by repeat FDV perpetrators.

Introduction

Family and domestic violence (FDV) is a major public health problem with substantial social, health and economic impacts (Ayre J et al. 2016; KPMG 2016; AIHW 2018). It generally occurs between family members but is most frequent in inter-partner (domestic) relationships (AIHW 2018). Acts of violence among those in family and domestic relationships may consist of physical, sexual, emotional or psychological abuse (Australian Government Department of Social Services 2010; Coumarelos 2019). The common FDV events reported to the police are threats, assaults, malicious damage, maliciously damaging property with the intent to injure a person, stalking, intimidation, communication offences such as harassment via repeated phone calls or text messages, cruelty to animals, driving offences, sexual offences, indecent assault and breaches of an apprehended violence order (AVO) (Grech & Burgess 2011). While there is no accurate picture of the prevalence of FDV, as many events are not reported to the police (Birdsey & Snowball 2013), it is estimated that in Australia one in six women and one in 16 men aged 15 years or over have experienced physical or sexual violence by a current or previous partner (ABS 2017). Domestic violence was found to be responsible for a range of avoidable diseases and injuries among adult women including depressive disorders, anxiety disorders, early pregnancy loss, suicide and self-inflicted injuries, alcohol use disorders, and preterm birth and low birthweight complications (AIHW 2018). In 2015–16, the estimated cost of violence against women and their children in Australia was $22b (KPMG 2016). FDV has an intergenerational effect, with children who have witnessed abuse experiencing multiple health problems (Bedi & Goddard 2007).

Family and domestic violence and mental illness

Previous studies suggest a relationship between mental illness and FDV (Oram et al. 2013; Trevillion et al. 2012). However, there is a research gap regarding the extent of this relationship among the perpetrators and victims (Solomon et al. 2005; Trevillion et al. 2012; Van Dorn et al. 2012). Specifically, research about the rates of mental illnesses among some groups who are at an increased risk of FDV is neglected (Oram et al. 2013; Solomon et al. 2005).

A systematic review of 41 studies of women and men who were 16 years or older demonstrated an increased likelihood of FDV victimisation across all categories of mental illnesses (Trevillion et al. 2012). With limited high-quality research, only four studies from the systematic review were included in a meta-analysis that showed a higher risk of inter-partner violence (IPV) victimisation among women with depressive disorders (odds ratio 2.77; 95% CI 1.96–3.92), anxiety disorders (odds ratio 4.08; 95% CI 2.39–6.97) and post-traumatic stress disorder (odds ratio 7.34; 95% CI 4.50–11.98) compared to women with no mental illnesses (Trevillion et al. 2012).

There is also limited high-quality evidence about the extent and patterns of FDV carried out by people with mental disorders (Solomon et al. 2005). Few empirical studies suggest a higher incidence of IPV among people with mental disorders including post-traumatic stress disorder, borderline personality disorder, psychopathic traits, and adult attachment challenges (McGinn et al. 2015; Solomon et al. 2005). A meta-analysis of six studies with a combined sample of 222 IPV perpetrators showed that 53 percent had a history of traumatic brain injury, compared with a prevalence of traumatic brain injury in the general population ranging from 10 percent to 38.5 percent (Farrer et al. 2012).

Web Computerised Operational Policing System

In New South Wales, data on events attended by the police are recorded in the Web Computerised Operational Policing System (WebCOPS). WebCOPS is an integrated law enforcement system designed to focus on operational policing functions to enable the police to capture, access and analyse crime information and intelligence on an organisation-wide basis. It is estimated that 1.5 million events are entered into WebCOPS each year (NSW Police Force 2011). Information about the events recorded in the WebCOPS contains both structured data (in fixed fields) and free text (as event narratives). The quantitative data in the fixed fields contain demographic and criminological information that is routinely used in surveillance and research. In the WebCOPS, FDV events are identified based on the relationship between a victim and the person of interest (POI).

Police event narratives

An event narrative is one of the features of the WebCOPS records. It is the free-text description of the event attended by the police, with its length varying depending on the severity and duration of the event. Event narratives are used by the police to guide the development of charge sheets in those cases where matters proceed to court; otherwise they have little practical value and are rarely used for research purposes. They contain information about the individuals involved and a description of relevant events leading up to, during and following the event, along with any action taken by police. Police narratives are generally detailed with reasonably complete accounts of the events from the police perspective.

Until our research, the use of these police event narratives to inform FDV debate was non-existent, but they have been used in a limited fashion to enhance larger quantitative analyses, by selecting a small sample of event narratives and manually analysing them to extract information of a qualitative nature. For example, offences involving small numbers of events (eg attempted abductions; Fitzgerald & People 2006) or offences that are confined to a particular geographic area or year have used the narratives. However, this manual approach is painstaking and time-consuming and uses only a fraction of the available narratives. One study undertook a manual analysis of narratives relating to fraud to identify the dollar value of the fraud, the parties involved, and, in the case of credit cards, how the cards were obtained. This study used 1,000 narratives from between 2012 and 2013, representing only two percent of a possible 49,457 events, and 500 narratives from between 2008 and 2009, representing 1.4 percent out of a possible 35,664 narratives (Macdonald & Fitzgerald 2014).

This study suggested that ‘there is no systematic way to extract information from these [police] narratives other than by manual review’ (Macdonald & Fitzgerald 2014). However, the application of text-mining methodologies may mean a change to this situation, with the potential to extract meaningful information from large bodies of data such as these.

Text mining

Text mining (TM) has been applied for over 30 years in many biomedical domains for the automatic recognition of important clinical concepts and events, with notable results (Abbe et al. 2015; Savova et al. 2010; Spasić et al. 2014). TM refers to the process of identifying targeted information from unstructured textual sources (Feldman & Sanger 2007; Kao & Poteet 2007). TM aims to automatically process unstructured data by extracting meaningful information from text (eg mentions of mental disorders for offenders and victims from police records) and by making the information contained in the text accessible for subsequent analysis to support further epidemiological or clinical research. Its output can be visualised using statistical and graphical techniques (eg cluster analysis) to represent the data, while extracted elements can be used as variables in predictive data modelling.

Study scope

Despite the rich descriptive content of the WebCOPS event narratives regarding the context and individuals involved in FDV events, these narratives are underused as a source of data to generate research insights. In a series of studies, we examined the feasibility and applicability of automated TM as a research approach to identify correlates of FDV from a large corpus of free-text event narratives that may have not been quantified in the fixed fields of the WebCOPS. Specifically, in our series of studies we addressed the following:

- the feasibility of TM in identifying the correlates of FDV from a large-scale free-text dataset (Karystianis et al. 2019, 2018);
- the use of the extracted information to fuel predictive modelling; and
- whether TM of free text can be applied as a public health and criminological research approach.

Research aims

We formulated four research aims within the scope of the study:

- to develop and implement an automated text-mining technique for the WebCOPS police narratives to extract features and identify potential patterns of FDV with a statistically acceptable level of accuracy;
- to investigate the frequency and patterns of mental disorders (ie mental illness, traumatic brain injury, and drug and alcohol misuse) of those involved in FDV events and of contextual features of FDV (ie abuse types, victim injuries) as identified in the police event narratives;
- to determine whether recurrent episodes of FDV can be predicted based on the data acquired; and
- to determine whether text mining of event narratives enhances the prediction of FDV beyond the structured data.

Research significance

This study is consistent with the NSW Premier's priority of reducing domestic violence reoffending by 25 percent by 2023 (NSW Government nd). We demonstrated the feasibility and applicability of text mining as an approach for future FDV research to better understand and identify pathways leading to family and domestic violence events. We built a predictive model that uses the extracted information to identify high-risk cases in which a repeated offender of FDV could breach their AVO, thus contributing to improved prevention activities.

Methods

We used event narratives from 492,393 FDV events provided by the NSW Police Force (NSWPF), occurring between January 2005 and December 2016 (Karystianis et al. 2019, 2018). These FDV events were identified because they were flagged in WebCOPS as 'domestic violence related', or because the violence description was coded as 'domestic', or because the relationship between the victim and the POI included any of the following: 'spouse/partner' (including ex-spouse/ex-partner), 'boyfriend/girlfriend' (including ex-boyfriend/ex-girlfriend), 'parent/guardian' (including step/foster), 'child' (including step/foster), 'sibling,' 'other member of family' (including kin), or 'carer'. The structured data for these events covered the following incident (ie offence) categories: various types of assaults; breaches of AVOs; homicides; malicious damage to property; and offences against another person such as intimidation, kidnapping, abduction, and harassment. Other kinds of FDV events (eg sexual assault) were not included in the structured data provided to the researchers.

In our study of this FDV cohort, we aimed to identify:

- mentions of mental disorders for POIs and victims;
- types of abuse engaged in by POIs (ranging from physical to emotional and social forms of abuse); and
- injuries inflicted on victims.

Definitions

Mental disorders

All ICD-10 (WHO 2017) classifications of mental and behavioural disorders and additional classifications including for dementia and brain injury (Table A1 in *Appendix*).

Abuse types

All forms of physical, emotional, psychological, social, economic abuse and neglect in family and domestic relationships (Table A2 in *Appendix*).

Victim injuries

Any type of victim-sustained physical injury.

Use of text mining to extract information of interest

The text-mining approach for this study included the design and application of rules that were based on syntactical patterns observed in text combined with dictionary terms for the recognition of mental disorders, abuse types and victim injuries in the event narratives. It consisted of the following steps:

- the creation of three dictionaries of terms to identify mentions of mental disorder, abuse types and victim injuries;
- the design and implementation of rules to capture mental disorders, abuse types and victim injuries mentions in context;
- the standardisation and mapping of the extracted mental disorder mentions to ICD-10 and other classifications; and
- the aggregation of multiple mentions in each narrative to achieve event-level annotation.

A total of 300 randomly selected event narratives were used for training, development (used to enhance the performance of the rules) and evaluation purposes (100 for each).

Dictionaries

In an attempt to capture all possible mental disorder related terms in the free-text police narratives, we used the classification of 'Mental and behavioral disorders' according to the World Health Organization's International Classification of Diseases, Tenth Revision (ICD-10) (WHO 2017). We also manually inspected a random sample of 200 events (training and development sets) to compile a dictionary of informal terms describing the ICD-10 mental and behavioural disorders (and others including traumatic brain injury and dementia), after consultation with experts in the field of mental health. This included frequently used synonyms, misspellings (eg 'schitzophrenia') and other indicative descriptive sentences (eg 'abuses alcohol', 'anger issues') that suggested a mental illness or disorder. We also included mentions of unspecified mental disorders reported in the narratives (eg 'victim is suffering from a severe mental disorder'), mentions of psychotropic medications by name or drug class (eg 'the victim takes Valium', 'accused takes a number of antidepressants'), and mentions of prescription drug abuse, substance abuse, and drug-induced disorders.

For abuse types and victim injuries, we similarly inspected the same random sample of 200 events and manually crafted dictionaries by reviewing several sources to categorise abuse types (KPMG 2016; Makkai 2017; White Ribbon Australia 2018). We also included systemic variations such as plural forms, past and present tenses, and common misspellings (eg 'harrassment', 'assalting') that were frequent in the event narratives. Non-specific forms of violence (eg 'bashing,' 'smack,' 'assaulted,' and 'clipping') were categorised as 'assault (unspecified)'. These term dictionaries contained nouns, verbs, phrases and sentences, and a total of 44 abuse types and 17 common injury types were included.

Rules

Our approach for the automatic recognition of mental disorders, abuse types and victim injuries involved the engineering of rules (Karystianis et al. 2018; Karystianis et al. 2019). The rules were based on common syntactical patterns observed in text (in our random sample of 200 events) that make use of preposition, verbs, and noun phrases. Combined with our term dictionaries, the rules are triggered when the related pattern is encountered in text, opting to identify the mental disorder, abuse type or victim injury of interest (eg 'The victim suffers from schizophrenia' [indicating mental illness of the victim], 'The defendant punched [indicating abuse type from the POI] the victim'). We also used concept enumeration, since it frequently appeared in the training and development sets (eg 'Injuries: Swollen hand, soreness and scratch under left eye [mentions of victim's injuries]'). To generate our rules, we used the General Architecture for Text Engineering or GATE software (Cunningham 2013).

Mapping mentions of mental disorder to ICD-10

Since the extracted mental disorder mentions were highly variable (eg synonyms, misspellings), any further analysis required them to be mapped into standard mental health concepts such as the ICD-10 classifications of 'Mental and behavioral disorders'. This was done automatically through a heuristic algorithm that relies on groups of representative terms for various ICD-10 categories. If a given mention matched one term from a specific ICD-10 category, then it was mapped to that category. The mapping was done at three levels with the first level being the most generic (26 categories), representing the overall type of mental health disorder.

In addition to the original 18 ICD-10 categories, with the guidance of a psychiatrist, we created eight customised categories to capture those mentions where no direct mapping to the ICD-10 was obvious. Four of these eight categories involved mentions of psychotropic medications ('medications—antidepressants', 'medications—antianxiety', 'medications—antipsychotics', 'medications—neuroleptics'). For example, in event narratives where a medication class (eg antidepressant) or a brand name (eg Zoloft) was specified, we mapped these mentions to the category 'medications—antidepressants.' The other four additional categories were 'drug prescription abuse', 'substance abuse (unspecified)', 'traumatic brain injury', and 'unspecified drug-induced disorder.' For cases where we recognised that either the victim or the perpetrator had an unknown mental disorder or an unknown drug-induced mental disorder, these were assigned to the categories of 'unspecified mental disorder' or 'unspecified drug-induced disorder', respectively. Cases in which mental disorder mentions were more specific were mapped to lower-level ICD-10 categories, with the second and third levels having 62 and 98 categories, respectively. A fourth level of ICD-10 classification (27 categories) was recorded in some narratives. However, for the purpose of reporting the results, we merged this level with the third level of classification. For example, instead of reporting 'other impulse control disorders' (third level), we included 'intermittent explosive disorder' (fourth level) in the third classification level for the representation of results only. Thus, although there were no explicit mentions of 'other impulse disorders' (for example), this mapping did not result in any information loss. All the ICD-10 levels used can be seen in the *Appendix* in Table A1.

Elimination of multiple mentions

More than one syntactical pattern may be matched in an event narrative and may refer to one or more mentions of abuse types, victim injuries or mental disorders related to a perpetrator or a victim including duplicates. This led to the extraction of highly variable mentions of the same abuse types (eg 'punch,' 'punched,' and 'punching') and victim injuries (eg 'bruised,' 'bruises,' and 'purple marks') in an event narrative which could have been duplicates. Each mention is mapped to its 'canonical' representative, with only one mention for each abuse type or injury retained for each event. For example, if, in a domestic violence event narrative, we extracted three mentions of the abuse type 'punching' and two mentions of the abuse type 'kicking,' we only annotate two abuse types—'punching' and 'kicking'—at the FDV event level.

We also eliminated duplicate mentions in the FDV event narratives of the extracted mental health mentions mapped to the ICD-10 categories. This led to narrative-level unification, since unique mentions of mental disorders for either victims or POIs were present in each event.

Evaluation

The method was evaluated on a set of 100 unseen, randomly chosen police narratives with mentions of mental disorders. The set was manually inspected and annotated by two domain experts in domestic violence and psychiatry, who identified mentions of mental disorders for POIs and victims, and by the first and second authors, who identified mentions of abuse types and victim injuries. The agreement for the mental disorder mentions between the first set of annotators was 90 percent and for the abuse types and the victim injuries from the second set of annotators was 91 percent; both were calculated as the absolute agreement rate (Kim 2006) and suggest consistent and reliable annotations by the experts.

The text-mining methodology was evaluated at the narrative level (after the mapping and elimination of any duplicate mentions) using the following standard metrics for the mentions of mental disorders, abuse types and victim injuries:

- precision—the number of true positives against the number of true positives and false positives;
- recall—the number of true positives against the number of true positives and false negatives; and
- F1-score—the harmonic mean between precision and recall (Ananiadou 2006).

We defined true positive as the detection of a correct mention in an event; false positive as the extraction of any unrelated mention that had not been annotated manually; false negative as a correct mention that was not detected by our method; and true negative as the case where our method did not identify any mentions when none were annotated. The average precision for the extraction of mental disorder mentions for either POIs or victims was 92.0 percent, whereas the precision for abuse types and victim injuries was 90 percent and 85 percent respectively. This suggests that text mining *can* reliably extract targeted information from police-recorded FDV events.

Data linkage between WebCOPS and NSW Health administrative data

We also explored the agreement between the mentions of mental disorders in the police event narratives and the NSW Ministry of Health diagnosis information obtained for each victim and/or POI. We linked the data extracted from the WebCOPS with two administrative data collections from NSW Ministry of Health that contain diagnostic information, the Admitted Patient Data Collection (APDC) and the Emergency Department Data Collection (EDDC). To our knowledge, this is the first attempt to apply a 'big data' approach to link a large corpus of text-mined data with administrative data collections in the area of justice health.

The linkage between the WebCOPS, APDC and EDDC was performed by the Centre for Health Record Linkage (CHeReL) using probabilistic matching techniques. Data custodians of each data collection provided CHeReL with demographic details for each individual, such as name, address, date of birth and gender, with an encrypted source record number to apply probabilistic methods to identify the records from each data collection that are likely to belong to the same person. CHeReL then assigned a specific person number for each individual to merge de-identified data from each data collection for data analysis.

The text mining methodology was unable to associate the extracted mental disorder mention with a specific POI or victim, if more than two individual POIs or victims were present. Thus, we focused only on those DV events that included a single POI and a single victim—a total of 416,441 DV events out of 492,393. The victims and POIs in these events were identified using the WebCOPS identifier for each individual, namely the Criminal Name Index. Those with any diagnosis of a mental disorder in either the APDC or EDDC were flagged as having a mental disorder, regardless of the diagnosis type or the number of admissions or presentations to an emergency department. We calculated the sensitivity and specificity of the mentions of the mental disorder extracted from the event narratives in identifying the correct diagnosis of mental disorder for victims and POIs using APDC and EDDC data. These measures were calculated as follows:

- sensitivity—number of FDV events with diagnosis of mental disorder for victims or POIs in either the APDC or EDDC and mentions of mental disorder in the police narratives/number of FDV events with diagnosis of mental disorder for victims or POIs in APDC or EDDC; and
- specificity—number of FDV events with no diagnosis of mental disorder for victims or POIs in either the APDC or EDDC and no mention of mental disorder in the event narratives/number of FDV events with no diagnosis of mental disorder for victims or POIs in APDC or EDDC.

Comparing the extracted mental disorder mentions against the WebCOPS structured fields

The extracted mental disorder mentions from the event narratives were compared against the WebCOPS structured data (fixed fields) to examine the utility of text mining for enhancing FDV information from the police reports. As mentioned above, we used mental disorder data only for those events that had one victim and one POI. Therefore, only events with a single victim and a single POI were included in the analysis comparing the mentions of mental disorder in the WebCOPS fixed field with mentions in the event narratives (416,441 events).

Predictive analytics methods

We used the data extracted by text mining as potential predictors for breaches of AVO. We aimed to predict whether a POI would breach their AVO based on information from their past FDV events. To our knowledge, this is the first attempt to examine the feasibility of using TM derived data in a predictive analysis study.

We first filtered the FDV events to those involving a single victim and a single POI (416,441 events). From this dataset, we selected POIs that were involved in two or more FDV events, resulting in the use of 234,562 events. This dataset was transformed into a sequential one consisting of 71,170 POIs who had been involved in multiple FDV events. We restricted the analysis to those POIs who had been involved in a maximum of three previous events to calculate the probability of an AVO breach. The sequential data classified the various fields of our dataset (both text-mined and fixed fields) into two groups: POI details (eg gender, country of origin, date of birth) and event details (eg types of abuse, victim injuries, premise types) in order to semi-profile the POI and the event. We selected three algorithms for this prediction exercise— naïve Bayes, logistic regression, and random forest—since these methods perform well with similar types of large-scale datasets and also offer interpretability regarding which features affect the returned performance.

We split this new dataset (*n*=71,170) into 80 percent (56,936) as the training set and 20 percent (14,234) as the evaluation set, a common practice for measuring the performance of machine-learning algorithms. From these, there were 32,223 (training) and 8,108 (evaluation) POIs with a breach of AVO recorded as an offence.

Results

Mental disorder mentions in police narratives

Of 492,393 FDV events, 77,995 (16%) had mentions of mental disorder (including traumatic brain injury and dementia) for POIs and/or victims. Of these, 77 percent (60,032 events) had mentions for perpetrators only, 16 percent (12,852 events) had mentions for victims only, and six percent (5,111 events) had mentions of mental disorder for both perpetrators and victims (Table 1).

Table 1: FDV event narratives with mentions of mental disorder (including traumatic brain injury and dementia), Jan 2005 – Dec 2016

	Number of events	Unique mentions of mental disorder
POIs only	60,032	81,942
Victims only	12,852	21,290
POIs and victims	5,111	–
Total	**77,995**	**103,232**

Table 2 presents the generic classifications for types of mental disorders that were mentioned in the narratives for victims and POIs (level 1). Almost one-third of the mentions of mental disorder (26,598) for POIs and one-fourth (4,851) for victims were unspecified but still had been noted down by the attending police officer(s). 'Mood [affective] disorders' (eg bipolar disorder, depression) had the highest number of mentions of mental disorder among POIs (15,330, 18.7%) and victims (4,946, 22.7%). Nearly 12.0 percent of perpetrators (9,848) and 10.4 percent of victims (2,224) had mentions of 'behavioural and emotional disorders with onset usually occurring in childhood and adolescence' (eg 'anger issues', 'loses control'). 'Intentional self-harm' was more frequently mentioned for POIs (3,271) than victims (949). However, the rate of mentions was higher for victims than POIs (4.4% vs 3.9%). 'Anxiety, dissociative, stress-related, somatoform and other non-psychotic mental disorders' scored the third-highest number of mentions of mental disorder among victims of FDV events (2,261, 10.6%). However, the numbers of these disorders were higher among POIs with mentions of mental disorder (3,755, 4.5%). Overall, mentions of 'intellectual disability' among POIs (1,517) were higher than the victims (939). Yet, among those with mentions of mental disorder, their rates were higher among victims (4.4%) than POIs (1.8%). Mentions of 'traumatic brain injury' (eg 'the victim has suffered a brain injury due to a car accident') were reported for nearly one percent of POIs and victims (688 and 250 mentions, respectively).

Table 2: Self and third-party mentions of types of mental disorder in FDV event narratives at the first level of ICD-10[a] for POIs (unique mentions *n*=81,942) and victims (unique mentions *n*=21,290), Jan 2005 – Dec 2016

Mental disorder group	Number of mentions (POIs)	Number of mentions (victims)
Unspecified mental disorder	26,598	4,851
Mood [affective] disorders	15,330	4,946
Behavioural and emotional disorders with onset usually occurring in childhood and adolescence	9,848	2,224
Mental and behavioral illnesses due to psychoactive substance use	6,790	1,259
Schizophrenia, schizotypal, delusional, and other non-mood psychotic disorders	5,771	1,032
Anxiety, dissociative, stress-related, somatoform and other non-psychotic mental disorders	3,755	2,261
Intentional self-harm	3,271	949
Substance abuse	2,852	370
Pervasive and specific developmental illnesses	1,775	501
Intellectual disability	1,517	939
Illnesses of adult personality and behaviour	1,340	420
Injury of unspecified body region	800	265
Traumatic brain injury	688	250
Mental illnesses due to known physiological conditions	559	649
Medications—antidepressants	400	130
Symptoms and signs involving cognition, perception, emotional state and behaviour	189	79
Medications—antipsychotics	142	16
Medications—anti-anxiety	91	24
Other degenerative diseases of the nervous system	62	52
Unspecified drug-induced illnesses	57	1
Chromosomal abnormalities, not elsewhere classified	53	39
Behavioural syndromes associated with physiological disturbances and physical factors	31	23
Systematic atrophies primarily affecting the central nervous system	11	6
Diseases of the nervous system	6	3
Drug prescription abuse	5	1
Medications—neuroleptics	1	0

a: The most generic categories representing the overall type of mental illness

At the second level of the classification of mental disorders, 'major depressive disorder, single episode' ranked first for both POIs and victims, with 8,944 (18.7%) and 3,269 (22.2%) mentions, respectively, although the rates were higher for victims. First- or third-person reports of 'alcohol abuse' were the second-highest reported mentions of mental disorders among POIs (5,829, 12.1%) and fifth-highest reported among victims (1,180, 8.0%). 'Dementia, unspecified' was mentioned 644 times for victims (4.3%) and 546 times for POIs (1.1%) (Table 3).

Table 3: Self and third-party mentions of the top 10 mental disorders at the second level of ICD-10 in the FDV event narratives for POIs and victims, Jan 2005 – Dec 2016

Mental disorder	Number of mentions (POIs)	Number of mentions (victims)
Major depressive disorder, single episode	8,944	3,269
Alcohol abuse	5,829	1,180
Bipolar disorder	5,449	1,553
Other behavioural and emotional disorders with onset usually occurring in childhood and adolescence	4,888	776
Schizophrenia	4,880	849
Attention deficit hyperactivity disorder	3,980	1,312
Other anxiety disorders	2,446	1,714
Pervasive developmental disorder	1,721	477
Specific personality disorders[a]	1,310	372
Intellectual disability, unspecified	1,225	779
Dementia, unspecified[b]	546	644

a: Not in the top 10 mental disorders for victims
b: Not in the top 10 mental disorders for POIs

At the third level of classification of mental disorders, 'bipolar disorder, unspecified' ranked first in mentions for both POIs (5,445, 21.59%) and victims (1,553, 21.36%) with similar rates (Table 4). However, it was observed that in POIs 'unspecified behavioral and emotional disorders with onset usually occurring in childhood and adolescence' were second in mentions (4,888, 19.38%) unlike with victims that had 'anxiety disorder, unspecified' (1,459, 20.07%).

Table 4: Self and third-party mentions of the top 10 mental disorders at the third level of ICD-10 in the FDV event narratives for POIs and victims, Jan 2005 – Dec 2016

Mental disorder	Number of mentions (POIs)	Number of mentions (victims)
Bipolar disorder, unspecified	5,445	1,553
Unspecified behavioral and emotional disorders with onset usually occurring in childhood and adolescence	4,888	776
Schizophrenia, unspecified	4,630	821
Anxiety disorder, unspecified	2,336	1,459
Autism	956	329
Oppositional defiant disorder	811	114
Suicide attempt	800	265
Cyclothymic disorder	780	90
Post-traumatic stress disorder	767	379
Asperger syndrome	758	146
Paranoid personality disorder	638	157

Mentions of abuse types and victim injuries

Given the relatively accurate results of the method in identifying abuse types and victim injuries for the evaluation phase, we applied our text-mining approach to the complete corpus of 492,393 FDV events. At least one mention of an abuse type was identified in over 71.3 percent (351,178) FDV event narratives. One-third of these events (177,117, 35.9%) had mentions of victim injuries in the narratives.

Of the 44 identified abuse types, ‘emotional/verbal abuse’ (117,488, 33.5%) was the most common, followed by ‘punching’ (86,322, 24.6%) and ‘property damage’ (78,203, 22.3%) (Table 5). More than half of the identified abuse types in the narratives were non-physical (55.4%), compared with 11.9 percent of events in the structured fields. In particular, ‘intimidation’ was reported in the fixed fields for 11.9 percent of all events and identified in free-text narratives for 21.6 percent of events. A total of 35.5 percent (124,498 events) of FDV event narratives contained only one identified abuse type, whereas 33.8 percent (118,819 events) of FDV events included three to five different abuse types (Table 6). Table A2 in the *Appendix* contains all the recorded abuse types and a definition for each.

Table 5: FDV event narratives with mentions of abuse types (*n*=351,178)

Abuse type	Events, *n* (%)[a]
Assault (unspecified)	171,323 (48.8)
Emotional/verbal abuse	117,488 (33.5)
Punching	86,322 (24.6)
Property damage	78,203 (22.3)
Intimidation	75,662 (21.6)
Grabbing	66,728 (19.0)
Pushing	62,794 (17.9)
Scratching	20,493 (5.8)
Physical restraining	20,014 (5.7)
Kicking	19,435 (5.5)
Slapping	17,474 (5.0)
ADVO breach	16,903 (4.8)
Attempting to hit with an object or weapon	13,592 (3.9)
Hair pulling/dragging by hair	13,048 (3.7)
Choking	11,325 (3.2)
Spitting	9,341 (2.7)
Hitting with an object or weapon	8,387 (2.4)
Other	7,135 (2.0)
Pulling	6,373 (1.8)
Victim being thrown around	5,255 (1.5)
Lunging	4,685 (1.3)
Possession of personal effects	3,265 (0.9)
Blocking	3,163 (0.9)
Harassment	3,100 (0.9)
Stalking	2,940 (0.8)
Self-harming	2,597 (0.7)
Biting	2,285 (0.7)
Dragging	2,216 (0.6)
Shaking	2,098 (0.6)
Stabbing	1,903 (0.5)
Forced entry	1,779 (0.5)
Headlocking	1,482 (0.4)
Chasing	1,324 (0.4)
Kneeing	1,321 (0.4)
Gagging	1,161 (0.3)
Elbowing	225 (0.1)
Limb twisting	173 (0.1)

Table 5: FDV event narratives with mentions of abuse types (*n*=351,178) (cont.)	
Abuse type	**Events, *n* (%)[a]**
Headbutting	148 (0.04)
Sexual assault	125 (0.04)
Preventing child access	91 (0.03)
Social restriction	40 (0.01)
Financial control	29 (0.01)
Attempting to set fire to premises	28 (0.01)
Ordered dog attack	1 (<0.01)

a: Percentages rounded to one decimal point. Two decimal points were reported for negligible percentages

Table 6: FDV event narratives according to the number of mentions of abuse types (*n*=351,178)	
Number of abuse types	**Events, *n* (%)[a]**
1	124,498 (35.5)
2	89,342 (25.4)
3–5	118,819 (33.8)
6–9	17,951 (5.1)
>10	568 (0.2)
Total	**351,178 (100.0)**

a: Percentages rounded to one decimal point

Types of victim injuries and number of sustained injuries in event narratives are given in Table 7 and Table 8. ‘Bruising’ was the most frequently mentioned as a victim injury (51,455, 29.0%), followed by ‘cut/abrasion’ (51,284, 28.9%) and ‘red marks/signs’ (42,038, 23.7%). Nearly 10 percent of victim-sustained injuries were ‘fractures’ (Table 7). Only one form of injury was reported for victims in 105,812 FDV events (59.6%), and 43,499 FDV events (24.4%) had victims with two forms of injury (Table 7).

Table 7: FDV event narratives with mentions of victim-sustained injury types (*n*=177,117)	
Injury type	**Events, *n* (%)[a]**
Bruising	51,455 (29.0)
Cut/abrasion	51,284 (28.9)
Red mark(s)	42,038 (23.7)
Swelling	32,581 (18.4)
Soreness	26,729 (15.1)
Other	19,778 (11.2)
Bleeding	19,154 (10.8)
Fracture(s)	17,531 (9.9)
Lump	9,482 (5.4)
Grazing	7,305 (4.1)
Black eye(s)	2,994 (1.7)
Scratching	2,399 (1.4)
Bite mark(s)	2,350 (1.3)
Stab wound(s)	2,346 (1.3)
Burn mark(s)	1,382 (0.8)
Broken tooth	620 (0.4)
Tear off nail(s)	7 (<0.1)

a: Percentages rounded to one decimal point

Table 8: FDV event narratives according to the number of injury types (*n*=177,117)	
Number of injury types	**Events, *n* (%)[a]**
1	105,812 (59.6)
2	43,499 (24.4)
3-4	25,717 (14.5)
5-6	2,490 (1.4)
≥7	89 (0.1)
Total	**177,117 (100.0)**

a: Percentages rounded to one decimal point

Data linkage between WebCOPS and NSW Health administrative data

Table 9 compares the number of events with mental disorder mentions against those with a diagnosis of mental disorder for victims and POIs in the EDDC or APDC.

Table 9: FDV events with TM extracts for mentions of mental disorder in WebCOPS event narratives against the number of FDV events with a diagnosis of mental disorder for victims or POIs, 2005–2016 (*n*=416,441)

		Diagnosis of mental disorder from APDC/EDDC in FDV event		
		Yes	No	Total
Mention of mental disorder in FDV event narrative	Yes	19,799	44,788	64,587
	No	106,693	245,161	351,855
Total		**126,492**	**289,950**	**416,441**

Note: APDC=Admitted Patient Data Collection; EDDC=Emergency Department Data Collection

The 15.6 percent sensitivity indicates that mental disorders were mentioned in event narratives for 15.6 percent of FDV events for a victim or POI where the person also had a diagnosis of mental disorder in either the EDDC or APDC since 2005. Specificity (84.6%) shows that 84.6 percent of events with no diagnosis of mental disorder in either the APDC or EDDC had no mention of mental disorder in the police narratives. The concordance of mentions of mental disorder in police narratives was the ratio of the correct assessment of police narratives against all FDV events: (19,799+245,161)/416,441=63.6 percent.

Comparing the extracted mental disorder mentions against the WebCOPS structured fields

Mental disorders

We compared the events with extracted mentions of mental disorders against the events that were tagged with ‘mental illness related’ for victims and POIs in the respective WebCOPS fixed field. While only one percent of the FDV event narratives (4,296) were tagged as ‘mental illness related’ in the fixed field, mentions of mental disorder for victims and POIs were identified in 15.8 percent (64,587) of the event narratives (Table 10).

Table 10: Comparison of FDV events with mentions of mental disorder in a WebCOPS fixed field versus TM extracted information

	Number of FDV events with TM extracted mentions of mental disorder				
	No information extracted	Mention of mental disorder for POI	Mention of mental disorder for victim	Extracted information for POI and victim	Total
Number of FDV events tagged as 'mental illness related'	1,457	2,344	217	278	4,296
Number of FDV events not tagged as 'mental illness related'	NA	46,810	10,950	3,988	61,748
Total	**1,457**	**49,154**	**11,167**	**4,266**	**66,044**

Note: NA=not available; TM=text mining

Intimidation

More than half of the identified abuse types in the narratives were non-physical (55.4%), compared to 11.9 percent of events in the structured data; specifically, 'intimidation' was reported in 11.9 percent of all events in the quantitative data and was identified in the free-text narratives of 21.6 percent of events.

Text mining FDV event narratives to report trends in domestic violence

We explored the use of extracted information from FDV police event narratives to report the trends of common types of mental disorder and victim injuries between 2005 and 2016.

Mental disorders

Trends of the most commonly reported types of mental disorder for men and women POIs are given in Figures 1 and 2. Mentions of all common types of mental disorders increased in police narratives for male offenders over time (Figure 1). 'Alcohol abuse' started with the greatest number of mentions in 2005 (175), remaining so until 2010, when 'major depressive disorder, single episode' sharply increased (especially in 2013) with the highest number of mentions in 2016. Comparatively, the five commonly reported mental illnesses had a more stable trend among female POIs. 'Major depressive disorder, single episode' had the greatest number of mentions across every year, in constrast with the male POIs where 'alcohol abuse' prevailed. There was a general increasing trend for every mental disorder over the given time period. Additionally, for both male and female POIs, mentions of major depressive disorder increased over time.

Figure 1: Trends in the five most common mental disorder mentions for male POIs in FDV police event narratives, 2005–2016 (*n*)

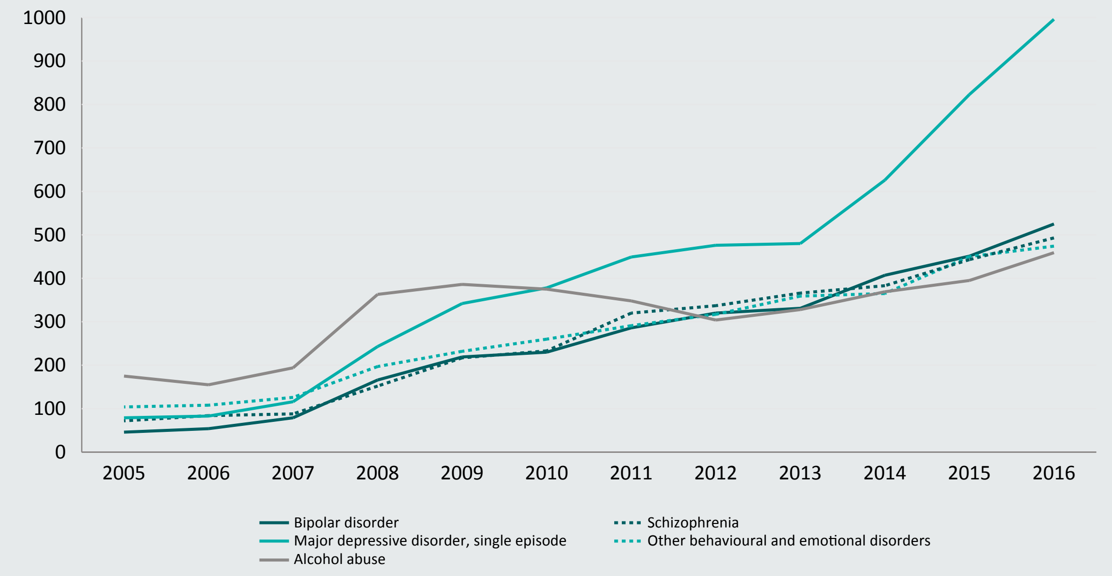
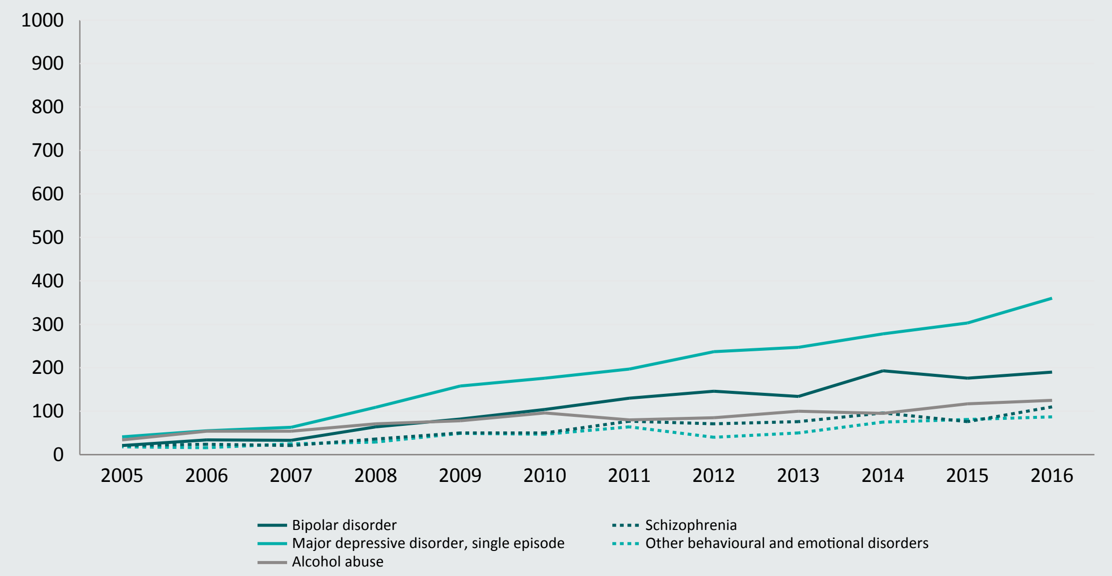


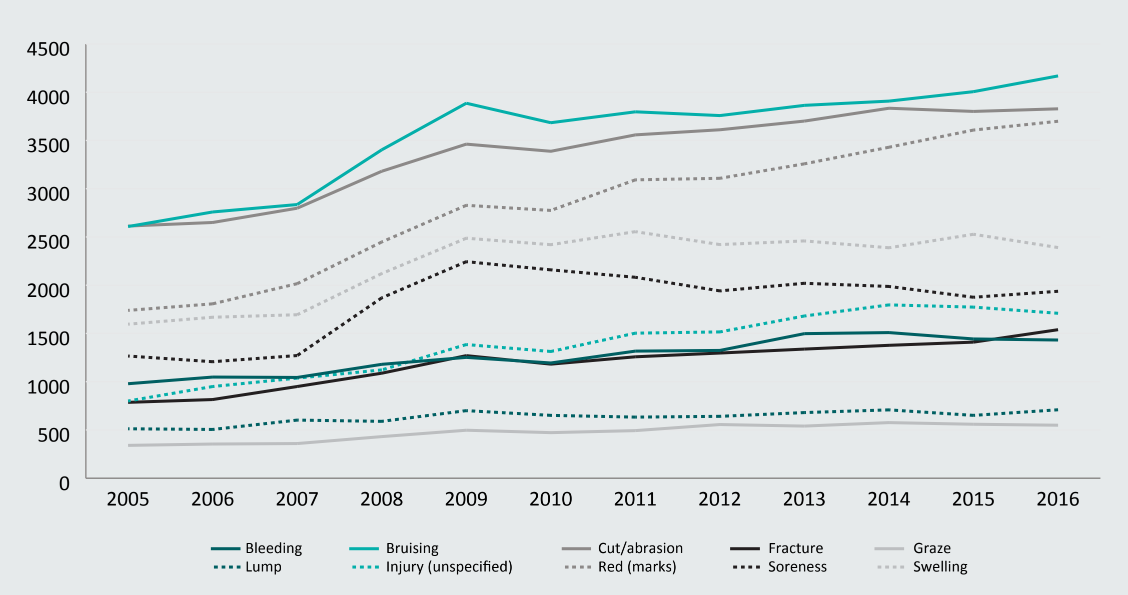
Figure 2: Trends of the five most common mental disorder mentions for female POIs in FDV police event narratives, 2005–2016 (*n*)



Victim injuries

Overall, victim injuries increased over time in the police event narratives (Figure 3). Reports of victim-sustained 'bruising', 'cut/abrasions', and 'red marks' increased more than other injury types.

Figure 3: Trends of the most common types of victim injuries in police narratives of FDV events, 2005–2016 (*n*)



Predictive analytics results

Breach of AVOs

We used three methods on an FDV dataset that includes a group of POIs with repeated events to predict whether a POI will breach their AVO, based on information from previous events. In an evaluation set of 14,234 POIs with repeated FDV events, logistic regression returned the best performance with a 70.0 percent F1-score followed by random forest and naïve Bayes with respective F-scores of 62.0 percent and 57.0 percent (Table 11).

Table 11: Performance of naïve Bayes, random forest and logistic regression methods for predicting breach of AVOs by POIs involved in several FDV events (*n*=71,170)

	Precision (%)	Recall (%)	F1-score (%)	AUROC
Naïve Bayes	63	52	57	57
Random forest	64	60	62	51
Logistic regression	65	76	70	67

Note: AUROC=area under the receiver operating characteristic curve

These results are promising as a first attempt, suggesting that information recorded for previous FDV events can be used to predict a 70 percent probability of a POI committing an AVO breach. From the NSW Police Force perspective, if a future FDV event occurs involving a particular POI, they can examine any previous FDV events that involved this POI and apply the predictive model, which will suggest how likely a breach of AVO by this particular POI is. Results returned by the logistic regression indicate that there is room to improve the model's predictability, demonstrating that text-mined information from FDV event narratives can potentially be used in predictive research and inform existing intervention policies for certain offence types.

Limitations

Our text-mining approach could have missed cases due to more specialised or explicit mentions of abuse types occurring in domestic violence events. Despite incorporating all types of abuse, there are still likely to be some cases where we did not identify explicit types of abuse. Similarly, although we included the basic and most common forms of injuries to victims, there may be less frequent and lesser known abuse types (eg use of acid, malnutrition) that may lead to specific injuries that were not included by our approach. Additionally, the implementation of spellchecking algorithms could assist in the future identification of any misspelled abuse types or injuries and potentially improve the performance of our method.

Since it was difficult to be sure which party a mental disorder mention relates to when more than one POI or victim is included in the same event, we decided to focus on FDV events that contained only a single POI and a single victim. This reduced the dataset that we used to fuel the data linkage and the predictive analytics from 492,393 events to 416,441 events.

Discussion

We conducted a series of studies to explore the feasibility and applicability of text mining of WebCOPS police event narratives as an approach to enhance information about family and domestic violence. These text-mined data were also combined with data from NSW Health administrative collections.

We successfully demonstrated the feasibility of text mining for identifying mentions of mental disorders, types of abuse and victim injuries in a large, population-based corpus of police event narratives (*n*=492,393). To our knowledge, this novel approach that combines the methods of text mining, data linkage and predictive analytics with population-level data has never been undertaken in the justice area. We then explored the utility of the information extracted by text mining the event narratives by comparing it with the WebCOPS fixed-fields dataset, which is the standard source of FDV information. We also demonstrated that there is valuable information in the event narratives that is not routinely recorded in the fixed fields. For example, if a victim sustained minor injuries and experienced verbal abuse—which can be warning signs for repeat and serious FDV events—these details are likely to be reported only in the police narratives. We successfully examined the application of text-mining results in a predictive analysis to explore whether this past information can be used to predict a future breach of an AVO. For future use of our predictive model, a different way of pre-processing the data (eg adding different or more features to characterise POIs and events) or the application of different machine learning algorithms is likely to improve the performance of the analysis.

We visualised the trends over time using the text-mined information from the police event narratives (2005–2016) suggesting the utility of text mining for public health surveillance purposes. Changes in these trends (eg mentions of mental disorders or injuries) can inform awareness about FDV-related factors and may inform prevention and early intervention strategies. Having established a text mining pipeline, it is now possible to undertake this approach on a regular basis to identify new and emerging patterns of abuse, victim injuries, weapons, victim types and so on that can inform health, welfare and police responses as well as policy formulation.

We examined the agreement between mentions of mental disorders in the police narratives and actual diagnoses of mental disorders for victims and POIs in these events through the linkage with the NSW Health data. The diagnosis data, however, was limited to those involving admission to hospital or presentation to an emergency department with a mental disorder. The agreement between mentions of mental disorder from text mining and APDC/EDDC presentations was satisfactory (63.6%). This is most likely due to mental disorders often being treated in the primary health care setting by GPs, meaning it would not appear in the APDC or EDDC. Future research could include linkage to Medicare data.

To our knowledge, this study is the first attempt to capture the various forms of abuse and types of victim injuries contained in a large corpus of police event narratives. While our data were mostly selected from FDV events involving physical abuse that were recorded in the WebCOPS structured format, the most common forms of abuse were non-physical. We also captured a range of mentions of injuries for which victims may have not sought medical attention and which would not be available for public health surveillance, which relies on medical attendances. Therefore, event narratives represent the only source of data for some types of abuse and injury in FDV events. We acknowledge that many forms of FDV and victim injuries are frequently not reported to the police and that the capture of non-physical forms of abuse and less serious injuries such as bruising from the police narratives was the by-product of investigating physical abuse events. Nonetheless, our findings, which show high rates of non-physical forms of abuse and common types of victim injuries, are supported by previous studies (Muelleman et al. 1996; Outlaw 2009). Due to frequent violence against women in domestic situations, more attention has been given to investigating physical forms of abuse in intimate partner relationships (Outlaw 2009). Still other forms of violence (eg psychological abuse and financial control) have important health and social consequences and can even be signs of physical abuse, which could result in serious bodily harm or even murder (Outlaw 2009). In future, text mining could be used to examine the relationship between abuse types and victim-sustained injuries and specific characteristics of victims and POIs. Further, the use of text mining as a predictive approach to investigate the escalation of abuse types and sustained injuries shows potential and may contribute significantly to the shaping of future policies that will benefit victims of domestic abuse.

Finally, our endeavours demonstrate that text mining as a research tool can also be applied to other areas of public health and criminology research to provide unique insights that would not otherwise be available.

References

URLs current as at December 2020

Abbe A, Grouin C, Zweigenbaum P & Falissard B 2015. Text mining applications in psychiatry: A systematic literature review. *International Journal of Methods in Psychiatric Research* 25(2): 86–100

Ananiadou S, Kell DB & Tsujii JI 2006. Text mining and its potential applications in systems biology. *Trends in Biotechnology* 24(12): 571–579

Australian Bureau of Statistics (ABS) 2017. *Personal safety survey, 2016*. Canberra: Australian Bureau of Statistics. https://www.abs.gov.au/statistics/people/crime-and-justice/personal-safety-australia

Australian Government Department of Social Services 2010. *National plan to reduce violence against women and their children 2010–2022*. Canberra: Department of Social Services

Australian Institute of Health and Welfare (AIHW) 2018. *Family, domestic and sexual violence in Australia, 2018*. Canberra: AIHW. https://www.aihw.gov.au/reports/domestic-violence/family-domestic-sexual-violence-in-australia-2018/summary

Ayre J et al. 2016. *Examination of the burden of disease of intimate partner violence against women in 2011: Final report*. Sydney: Australia's National Research Organisation for Women's Safety

Bedi G & Goddard C 2007. Intimate partner violence: What are the impacts on children? *Australian Psychologist* 42: 66–77

Birdsey E & Snowball L 2013. Reporting violence to police: A survey of victims attending domestic violence services. *Crime and Justice Statistics* October 2013(91)

Bundschus M, Dejori M et al. 2008. Extraction of semantic biomedical relations from text using conditional random fields. *BMC Bioinformatics* 9: 207

Coumarelos C 2019. *Quantifying the legal and broader file impacts of domestic and family violence*. Justice issues paper no. 32. Sydney: Law and Justice Foundation of New South Wales

Cunningham HTV, Robets A & Vontcheva K 2013. Getting more out of biomedical documents with GATE's full lifecycle open source text analysis. *PLoS Computational Biology* 9(2)

Farrer TJ, Frost RB & Hedges DW 2012. Prevalence of traumatic brain injury in intimate partner violence offenders compared to the general population: A meta-analysis. *Trauma, Violence, & Abuse* 13: 77–82

Feldman R & Sanger J 2007. *The text mining handbook: Advanced approaches in analyzing unstructured data*. New York, Cambridge University Press

Fitzgerald J & People J 2006. *Victims of abduction: Patterns and case studies*. Crime and Justice Bulletin no. 94. Sydney: NSW Bureau of Crime Statistics and Research

Grech K & Burgess M 2011. *Trends and patterns in domestic violence assault: 2001 to 2010*. Bureau Brief no. 61. Sydney: NSW Bureau of Crime Statistics and Research

Hristovski D, Friedman C et al. 2006. Exploiting semantic relations for literature-based discovery. *AMIA Annual Symposium Proceedings*: 349–353

Jimeno A, Jimenez-Ruiz E et al. 2008. Assessment of disease named entity recognition on a corpus of annotated sentences. *BMC Bioinformatics* 9: S3

Kao A & Poteet SR 2007. *Natural language processing and text mining*. London: Springer

Karystianis G et al. 2019. Automated analysis of domestic violence police reports to explore abuse types and victim injuries: Text mining study. *Journal of Medical Internet Research* 21(3): e13067

Karystianis G et al. 2018. Automatic extraction of mental health disorders from domestic violence police narratives: Text mining study. *Journal of Medical Internet Research* 20(9): e11548

Kim JD, Tsujii JI 2006. Corpora and their annotation. In S Ananiadou & J McNaught (eds), *Text mining for biology and biomedicine*. Boston: Artech House: 179–211

KPMG 2016. *The cost of violence against women and their children in Australia: Final report*. Prepared for Australian Government Department of Social Service

Macdonald W & Fitzgerald J 2014. *Understanding fraud: The nature of fraud offences recorded by NSW Police*. Crime and Justice Bulletin no. 180. Sydney: NSW Bureau of Crime Statistics and Research

McGinn T, Taylor B, McColgan M & Lagdon S 2015. Survivor perspectives on IPV perpetrator interventions: A systematic narrative review. *Trauma, Violence, & Abuse* 17(3): 239–255

Mouzos J & Makkai T 2004. *Women's experiences of male violence: Findings from the Australian component of the International Violence Against Women Survey (IVAWS)*. Research and public policy series no. 56. Canberra: Australian Institute of Criminology. https://www.aic.gov.au/publications/rpp/rpp56

Muelleman RL, Lenaghan PA and Pakieser RA 1996. Battered women: Injury locations and types. *Annals of Emergency Medicine* 28(5): 486–492

NSW Government nd. Premier's Priorities. https://www.nsw.gov.au/premiers-priorities

NSW Police Force 2011. *New computer operating system brings NSW Police Force into the 21st century*. Media release, 15 September

Oram S, Trevillion K, Feder G & Howard LM 2013. Prevalence of experiences of domestic violence among psychiatric patients: Systematic review. *British Journal of Psychiatry* 202: 94–99

Outlaw M 2009. No one type of intimate partner abuse: Exploring physical and non-physical abuse among intimate partners. *Journal of Family Violence* 24(4): 263–272

Savova G et al. 2010. Mayo clinical text analysis and knowledge extraction system (cTAKES): Architecture, component evaluation and applications. *Journal of the American Medical Informatics Association* 17(5): 507–513

Solomon PL, Cavanaugh MM & Gelles RJ 2005. Family violence among adults with severe mental illness: A neglected area of research. *Trauma, Violence & Abuse* 6(1): 40–54

Spasić I, Livsey J, Keane JA & Nenadić G 2014. Text mining of cancer-related information: Review of current status and future directions. *International Journal of Medical Informatics* 83(9): 605–623

Trevillion K, Oram S, Feder G & Howard LM 2012. Experiences of domestic violence and mental disorders: A systematic review and meta-analysis. *PLOS One* 7(12): e51740

Van Dorn R, Volavka J & Johnson N 2012. Mental disorder and violence: Is there a relationship beyond substance use? *Social Psychiatry and Psychiatric Epidemiology* 47(3): 487–503

White Ribbon Australia 2018. *Physical abuse: Signs of physical abuse*. https://www.whiteribbon.org.au/understand-domestic-violence/types-of-abuse/physical-abuse/

World Health Organization 2017. *The International Classification of Disease 10: Mental and Behavioural Disorders*

Appendix

Table A1: ICD-10 classifications of mental and behavioural disorders used to map the extracted and standardised mental health disorder mentions at three levels, plus additional categories			
First level	**Second level**	**Third level**	**Fourth level**
Mental disorders due to known physiological conditions	Vascular dementia	–	–
	Unspecified dementia	–	–
	Delirium	–	–
	Unspecified mental disorder due to known physiological condition	–	–
Mental and behavioral disorders due to psychoactive substance use	Alcohol related disorders	–	–
	Opioid related disorders	–	–
	Cannabis related disorders	–	–
	Cocaine related disorders	–	–
	Other stimulant related disorders	–	–
	Nicotine dependence	–	–
	Other psychoactive substance related disorders	–	–

Table A1: ICD-10 classifications of mental and behavioural disorders used to map the extracted and standardised mental health disorder mentions at three levels, plus additional categories (cont.)

First level	Second level	Third level	Fourth level
Schizophrenia, schizotypal, delusional, and other non-mood psychotic disorders	Schizophrenia	Paranoid schizophrenia	–
		Disorganised schizophrenia	–
		Catatonic schizophrenia	–
		Undifferentiated schizophrenia	–
		Residual schizophrenia	–
		Other schizophrenia	–
		Unspecified schizophrenia	–
	Schizotypal disorder	–	–
	Delusional disorders	–	–
	Brief psychotic disorder	–	–
	Shared psychotic disorder	–	–
	Schizoaffective	–	–
	Unspecified psychosis not due to a substance or known physiological condition	–	–
Mood [affective] disorders	Manic episode	–	–
	Bipolar disorder	Bipolar disorder, unspecified	–
		Other bipolar disorders	Bipolar II disorders
	Major depressive disorder, single episode	Postpartum depression	–
	Major depressive disorder, recurrent	Other recurrent depressive disorders	
	Persistent mood disorders	Cyclothymic disorder	–
		Dysthymic disorder	–
		Other persistent mood disorders	Disruptive mood dysregulation disorder
	Unspecified mood disorder	–	–

Table A1: ICD-10 classifications of mental and behavioural disorders used to map the extracted and standardised mental health disorder mentions at three levels, plus additional categories (cont.)

First level	Second level	Third level	Fourth level
Anxiety, dissociative, stress-related, somatoform and other nonpsychotic mental disorders	Phobic anxiety disorder	Agoraphobia	–
		Social phobias	Social phobia, generalised
		Specific isolated phobias	Arachnophobia
			Claustrophobia
			Acrophobia
			Androphobia
			Gynaecophobia
		Other phobic anxiety disorders	–
		Phobic anxiety disorder, unspecified	–
	Other anxiety disorders	Panic disorder	–
		Generalised anxiety disorder	–
		Anxiety disorder, unspecified	–
	Obsessive compulsive disorders	Hoarding disorder	–
		Excoriation disorder	–
		Obsessive compulsive disorder, unspecified	–
	Reaction to severe stress and adjustment disorders	Acute stress reaction	–
		Post-traumatic stress disorder	–
		Adjustment disorders	–
	Dissociative and conversion disorders	Dissociative amnesia	–
		Dissociative fugue	–
		Dissociative stupor	–
		Other dissociative and conversion disorders	Dissociative identity disorder
		Dissociative and conversion disorder, unspecified	–
	Somatoform disorders	Somatization disorder	–
		Undifferentiated somatoform disorder	–
		Hypochondrial disorders	Body dysmorphic disorder
			Hypochondriasis

Table A1: ICD-10 classifications of mental and behavioural disorders used to map the extracted and standardised mental health disorder mentions at three levels, plus additional categories (cont.)

First level	Second level	Third level	Fourth level
	Other nonpsychotic mental disorders	Depersonalization-derealization syndrome	–
		Pseudobulbar affect	–
		Nonpsychotic mental disorder, unspecified	–
Behavioral syndromes associated with physiological disturbances and physical factors	Eating disorders	Anorexia nervosa	–
		Bulimia nervosa	–
		Other eating disorders	Binge eating disorder
			Avoidant food intake disorder
	Sleep disorders not due to a substance or known physiological condition	Insomnia not due to a substance or known physiological condition	Primary insomnia
			Adjustment insomnia
			Paradoxical insomnia
			Psychophysiologic insomnia
	Sexual dysfunction not due to a substance or known physiological condition	Hypoactive sexual desire disorder	–
	Abuse of non-psychoactive substances	Abuse of steroids or hormones	–
Disorders of adult personality and behavior	Specific personality disorders	Paranoid personality disorder	–
		Schizoid personality disorder	–
		Antisocial personality disorder	–
		Borderline personality disorder	–
		Histrionic personality disorder	–
		Obsessive compulsive personality disorder	–
		Avoidant personality disorder	–
		Dependent personality disorder	–
		Other specific personality disorders	Narcissistic personality disorder
		Personality disorder, unspecified	–

Table A1: ICD-10 classifications of mental and behavioural disorders used to map the extracted and standardised mental health disorder mentions at three levels, plus additional categories (cont.)

First level	Second level	Third level	Fourth level
	Impulse disorders	Pathological gambling	–
		Pyromania	–
		Kleptomania	–
		Trichotillomania	–
		Other impulse disorders	Intermittent explosive disorder
		Impulse disorder, unspecified	–
	Gender identity disorders	Transsexualism	–
		Dual role transsexualism	–
		Gender identity disorder	–
	Paraphilias	Fetishism	–
		Transvestic fetishism	–
		Exhibitionism	–
		Voyeurism	–
		Paedophilia	–
		Sadomasochism	–
		Other paraphilias	Frotteurism
	Other personalities of adult and personality behaviour	Factitious disorder	–
	Unspecified disorder of adult personality and behavior	–	–
Intellectual disabilities	Mild intellectual disabilities	–	–
	Moderate intellectual disabilities	–	–
	Severe intellectual disabilities	–	–
	Profound intellectual disabilities	–	–
	Unspecified intellectual disabilities	–	–

Table A1: ICD-10 classifications of mental and behavioural disorders used to map the extracted and standardised mental health disorder mentions at three levels, plus additional categories (cont.)

First level	Second level	Third level	Fourth level
Pervasive and specific developmental disorders	Specific developmental disorders of speech and language	Phonological disorder	–
		Expressive language disorder	–
		Mixed receptive-expressive language disorder	–
		Other developmental disorders of speech and language	Childhood onset fluency disorder
			Social pragmatic communication disorder
	Specific developmental disorders of scholastic skills	Specific reading disorder	–
		Mathematics disorder	–
		Other developmental disorders of scholastic skills	Disorder of written expression
	Pervasive developmental disorders	Autism	–
		Rett's syndrome	–
		Asperger's syndrome	–
		Pervasive developmental disorder, unspecified	–
	Unspecified disorder of psychological development	–	–
Behavioral and emotional disorders with onset usually occurring in childhood and adolescence	Attention-deficit hyperactivity disorders	–	–
	Conduct disorders	Conduct disorder, unspecified	–
		Oppositional defiant disorder	–
	Emotional disorders with onset specific to childhood	Separation anxiety disorder of childhood	–
	Disorders of social functioning with onset specific to childhood and adolescence	Selective mutism	–
		Reactive attachment disorder of childhood	–
		Disinhibited attachment disorder of childhood	–

Table A1: ICD-10 classifications of mental and behavioural disorders used to map the extracted and standardised mental health disorder mentions at three levels, plus additional categories (cont.)

First level	Second level	Third level	Fourth level
	Tic disorder	Transient tic disorder	–
		Chronic motor or vocal tic disorder	–
		Tourette's disorder	–
	Other behavioural and emotional disorders	Unspecified behavioural and emotional disorders	–
Unspecified mental disorder			–
Other degenerative diseases of the nervous system	Alzheimer's disease	Alzheimer's disease, unspecified	–
	Other degenerative diseases of the nervous system, not elsewhere classified	Frontotemporal dementia	–
Systemic atrophies primarily affecting the central nervous system	Huntington's disease	–	–
Injury of unspecified body region	Injury of unspecified body region	Unspecified injury	Suicide attempt
Symptoms and signs involving cognition, perception, emotional state and behavior	Symptoms and signs involving emotional state	Other symptoms and signs involving emotional state	Homicidal and suicidal ideations
Chromosomal abnormalities, not elsewhere classified	Down syndrome	Down syndrome, unspecified	–
Intentional self-harm	–	–	–
Unspecified diseases of the nervous system	–	–	–
unspecified drug-induced disorders	–	–	–
Medications—neuroleptics	–	–	–
Medications—antipsychotics	–	–	–
Medications—anti-anxiety	–	–	–
Medications—antidepressants	–	–	–
Traumatic brain injury	–	–	–
Substance abuse	–	–	–
Drug prescription abuse	–	–	–

Note: This schema contains the additional eight categories that were used to capture those mentions where no direct mapping to the ICD-10 was obvious

CRG reports

CRG 34/15–16

Dr George Karystianis is a Research Fellow at the Justice Health Research Program, School of Population Health, UNSW Sydney.

Dr Armita Adily is a Research Fellow at the Justice Health Research Program, School of Population Health, UNSW Sydney.

Rina Carines Cabral is a data scientist who worked on this project and has graduated with a Master of Data Science degree from the University of Sydney's School of Computer Science.

Professor Tony Butler is the Program Head, Justice Health Research Program, School of Population Health, UNSW Sydney.

www.aic.gov.au/crg