Substance use in psychiatric crisis: relationship to violence

Supplementary Material

Nicola J. Kalk, John E. Robins, Kezia R. Ross,
Megan Pritchard, Michael T. Lynskey, Vivienne Curtis, Katherine I. Morley

Contents

1 Phenotype Algorithms

1.1 Background

The CRIS database is complex as the patient journey system (PJS) is used by many different clinical services with different medical records needs. Consequently, information relevant to the definition of a single variable may be recorded in multiple database tables and free-text documents. We developed phenotype algorithms to combine data from these sources in the most efficient way, using a staged approach to variable definition:

1. Identify sources of structured information: These are data sources that contain information recorded in a pre-specified way, such as describing patient gender or ethnicity using a pre-specified list of options. Data sources were identified via discussions with clinical and informatics colleagues, and examination of database table structure.

2. Examine content of each structured information source: We examined how frequently the relevant database fields were populated with non-null values.

3. Consider unstructured data sources and natural language processing (NLP): These are data sources where information is not entered based on a pre-specified list of values; this may include things such as long-form clinical notes. If relevant NLP algorithms for use with CRIS data already existed, we used these to extract information from clinical notes within the time period.

4. Consider manual note scoring: If information was only available in clinical notes and could not easily be captured via NLP algorithms (e.g. suicide risk), we considered whether it could be converted into structured data via manual note scoring.

5. Integrate structured and unstructured data: For those variables where data were extracted from more than one source, EHR phenotype algorithms were developed to define the strategy used to integrate and reconcile these data.

1.2 Data Sources

We included data from a number of different sources in PJS. Although the cPoS has its own structured proforma for collecting information, this was not implemented until approximately 6 months after the cPoS opened. Thus we had to combine information from this proforma with other data sources in order to extract the variables required for all detentions in the data set. A brief description of the different data sources, including whether we used structured or unstructured information from each source, is provided in Table S1.

Table S1: Included CRIS data sources. Description of CRIS data sources from which data were extracted. Whether the data source contained structured (S) and/or unstructured (U) information used in generating phenotype algorithms is indicated.

Source	**Description**	**Data Types**	
		S	U
Electronic Patient Record (EPR)	Records core patient sociodemographic information.	✓	
Place of Safety Proforma (PSP)	Custom form designed to record large range of information relating to detention characteristics including mode of transportation, use of restrictive interventions by police, alcohol and drug intoxication, detention outcome and destination, clinical notes from all health and social care professionals relating to detention.	✓	✓
Diagnosis (D)	Can be used to record up to 6 ICD-10 diagnoses and allows Multi-Axial ICD-10 recording. Diagnoses are only 'registered' within NHS Trust reporting and performance systems via this form. A diagnosis can be assigned at any point, but is formally required: (i) following a mental state assessment and formulation; (ii) at inpatient discharge; (iii) at transfer/discharge from a community-based team.	✓	
Current Drug and Alcohol (CDA)	Record of patient's current drug and alcohol use. Includes specific fields for substance use on day of assessment, and on the day prior to assessment.	✓	
Urine Drug Screen (UDS)	Records urine screen results, which includes tests for a selection of prescribed medications and illicit drugs (cannabis, .	✓	
Alcohol Use Disorders Identification Test (AUDIT)	Records the results of the Alcohol Use Disorders Identification Test	✓	
Risk Assessment (RA)	Risk assessment should be recorded for all new service-users and updated whenever there is a significant change to the risk profile or at key intervals during the patient journey, such as at points of transfer between teams. Covers: risk to others (e.g. violence, sexual offending, stalking); risk to self (e.g. deliberate or accidental self-harm, self-neglect); risk from others (abuse, neglect, or exploitation).	✓	✓
Events (E)	Notes for all clinical events		✓
Mental Health Act (MHA)	Records all information about use of the Mental Health Act in clinical care	✓	

1.3 Algorithms

We devised strategies for integrating and reconciling multiple sources of information, where available, for key variables. These algorithms were implemented in R during data processing, but where the sources of information were complex, diagrams were developed to aid explanation and discussion with clinical colleagues (Morley 2014; Denaxas & Morley 2015). Table S2 shows which data sources were used in the derivation of each variable in the data set.

Table S2: Data sources used included in variable definition. EPR: centralised electronic patient records; PSP: Place of Safety Proforma; D: diagnosis; CDAA: Current Drug And Alcohol form; UDS: urine drug screen; AUDIT: Alcohol Use Disorders Identification Test form; RA: risk assessment form; E: event notes; MHA: Mental Health Act records; NLP: natural language processing algorithm output.

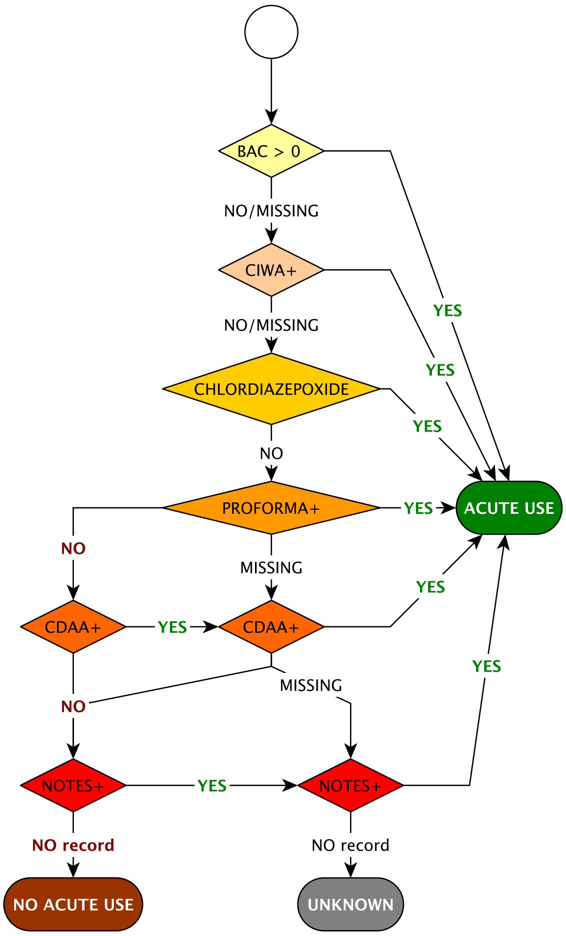
Variable	EPR	PSP	D	CDAA	UDS	AUDIT	RA	E	MHA	NLP
Age	✓									
Gender	✓									
Ethnicity	✓	✓						✓		
Housing status	✓	✓					✓	✓		
Alcohol use		✓		✓		✓		✓		✓
Cannabis use		✓		✓	✓			✓		
Stimulant use		✓		✓	✓			✓		
Opiate use		✓		✓	✓			✓		
Sedative use		✓		✓	✓			✓		
Synthetic cannabinoid use		✓		✓				✓		
Party drug use		✓		✓				✓		
Psychiatric diagnosis			✓							✓
Detention outcome		✓						✓	✓	

For most phenotypes, integration of information from multiple sources was straightforward, with preference given to information originating from the PoS proforma. However, for some phenotypes, such as those relating to substance use, there was a range of information available which had to be reconciled and integrated. For these phenotypes, more complex algorithms were developed as detailed below.

1.3.1 Acute Alcohol Use

For this phenotype, the aim was to identify recent alcohol use (i.e. in the previous 24-48 hours leading up to detention), as opposed to long-term dependence. Consequently, we gave priority to information that provided immediate, biological indications of the effects of alcohol: blood alcohol concentration and/or alcohol withdrawal assessment, as recorded in clinical notes, and administration of chlordiazepoxide (medication used in the treatment of alcohol withdrawal), which was drawn from NLP output (see Figure S1). We then included information from two structured forms - the PoS proforma record of alcohol intoxication, and the Current Drug and Alcohol form which records substances used on the day of detention and the day prior. Finally, we incorporated information from clinical notes on self-reported alcohol use, and the assessments of police, paramedics, and clinical staff. If there was no information regarding alcohol use, but there was a record of using other substances, then we classified this as no evidence of alcohol use. However, if there was no information about any substance use for a particular detention, acute alcohol use was classified as missing.

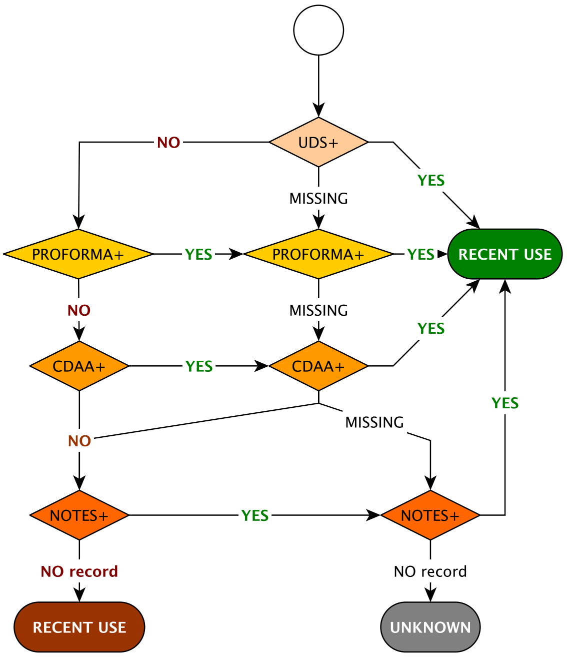
Figure S1: Diagram of algorithm for integrating sources of information on acute alcohol use. BAC: blood alcohol concentration; CIWA: Clinical Institute Withdrawal Assessment of Alcohol score; PROFORMA+: record of alcohol intoxication in PoS proforma; CDAA+: record of acute alcohol use in Current Drug And Alcohol form; NOTES+: record of acute alcohol use from clinical note scoring.



1.3.2 Recent Illicit and Licit Drug Use

For all substances of interest, apart from alcohol, the same sources of information, or a subset, were available and thus a single algorithm could be used. As for alcohol use, we prioritised biological measures of recent use, so first examined data from urine drug screens where available. Following that, we prioritised the same sources of structured information and data extracted from clinical notes as the alcohol use algorithm (see Figure S2). The algorithm was applied to create variables indicating recent use of cannabis, stimulants (amphetamines, cocaine), opiates (heroin, methadone, buprenorphine), sedatives, synthetic cannabinoids, party drugs (MDMA, G, GHB/GBH).

Figure S2: Diagram of algorithm for integrating sources of information on recent substance use. UDS: urine drug screen; PROFORMA+: record of intoxication due to substance in PoS proforma; CDAA+: record of recent use of drug in Current Drug And Alcohol form; NOTES+: record of recent use of drug from clinical note scoring.



1.3.3 Detention Outcome and Destination

Detention outcome was ultimately classified as: *Admission - Informal*, *Admission - Section 2/3*, *Discharge - Mental Health follow-up*, or *Discharge - GP follow-up*. The destinations included: *Inpatient care*, *GP*, *Community Mental Health Team*, or *Home Treatment Team*. Consequently, information about one variable could be used to infer the value of the other variable if it was missing. We initially used structured information from the PoS proforma to derive these measures. Where this information was missing, we incorporated information from the Events table, and finally from manual scoring of clinical notes. We cross-checked this information against records in the Mental Health Act table to determine whether admissions were informal or under Section 2 or Section 3 of the MHA. If we could not determine the conditions under which someone was admitted, this was treated as missing.

1.3.4 Past Year Psychiatric Diagnoses

Information on psychiatric diagnoses recorded during the past year (up to and including the current detention episode) were drawn from the structured diagnosis table and from the output of a previously developed natural language processing algorithm (Perera 2016). Table S3 displays

ICD-10 codes included and the categorisation used for the phenotype algorithm. Where multiple diagnoses existed that belonged to different categories, diagnoses relating to psychotic illnesses were retained in preference to those relating to non-psychotic illnesses. Individuals with a diagnosis in the learning disorder and/or neurodevelopmental, or organic disorders categories, were excluded from the data set (see main text).

Table S3: ICD-10 codes included in past-year psychiatric diagnoses phenotype. LD indicates learning disability.

Code	Description	Categorisation
F02.*	Dementia in other diseases classified elsewhere	Organic disorders
F03	Unspecified dementia	Organic disorders
F05.*	Delirium, not induced by alcohol and other psychoactive substances	Organic disorders
F06.*	Other mental disorders due to brain damage and dysfunction and to physical disease	Organic disorders
F07.*	Personality and behavioural disorders due to brain disease, damage and dysfunction	Organic disorders
F09.*	Unspecified organic or symptomatic mental disorder	Organic disorders
F20.*	Schizophrenia	Psychotic illness
F21.*	Schizotypal disorder	Psychotic illness
F22.*	Persistent delusional disorders	Psychotic illness
F23.*	Acute and transient psychotic disorders	Psychotic illness
F25.*	Schizoaffective disorders	Psychotic illness
F28	Other nonorganic psychotic disorders	Psychotic illness
F29	Unspecified nonorganic psychosis	Psychotic illness
F30.*	Manic episode	Psychotic illness
F31.*	Bipolar affective disorder	Psychotic illness
F32.*	Depressive episode	Non-psychotic illness
F33.*	Recurrent depressive disorder	Non-psychotic illness
F34.*	Persistent mood [affective] disorders	Non-psychotic illness
F38.*	Other mood [affective] disorders	Non-psychotic illness
F39	Unspecified mood [affective] disorder	Non-psychotic illness
F40.*	Phobic anxiety disorders	Non-psychotic illness
F41.*	Other anxiety disorders	Non-psychotic illness
F42.*	Obsessive-compulsive disorder	Non-psychotic illness
F43.*	Reaction to severe stress, and adjustment disorders	Non-psychotic illness
F44.*	Dissociative [conversion] disorders	Non-psychotic illness
F45.*	Somatoform disorders	Non-psychotic illness
F50.*	Eating disorders	Non-psychotic illness
F52.*	Sexual dysfunction, not caused by organic disorder or disease	Non-psychotic illness
F54.*	Psychological and behavioural factors associated with disorders or diseases classified elsewhere	Non-psychotic illness
F60.*	Specific personality disorders	Non-psychotic illness
F61	Mixed and other personality disorders	Non-psychotic illness
F62.*	Enduring personality changes, not attributable to brain damage and disease	Non-psychotic illness
F63.*	Habit and impulse disorders	Non-psychotic illness
F68.*	Other disorders of adult personality and behaviour	Non-psychotic illness
F69	Unspecified disorder of adult personality and behaviour	Non-psychotic illness
F70	Mild mental retardation	LD/neurodevelopmental
F71	Moderate mental retardation	LD/neurodevelopmental
F72	Severe mental retardation	LD/neurodevelopmental
F78	Other mental retardation	LD/neurodevelopmental
F79	Unspecified mental retardation	LD/neurodevelopmental
F80.*	Specific developmental disorders of speech and language	LD/neurodevelopmental
F81.*	Specific developmental disorders of scholastic skills	LD/neurodevelopmental
F82	Specific developmental disorder of motor function	LD/neurodevelopmental
F84.*	Pervasive developmental disorders	LD/neurodevelopmental
F88	Other disorders of psychological development	LD/neurodevelopmental
F89	Unspecified disorder of psychological development	LD/neurodevelopmental
F90.*	Hyperkinetic disorders	LD/neurodevelopmental
F91.*	Conduct disorders	LD/neurodevelopmental
F93.*	Emotional disorders with onset specific to childhood	LD/neurodevelopmental
F94.*	Disorders of social functioning with onset specific to childhood and adolescence	LD/neurodevelopmental
F95.*	Tic disorders	LD/neurodevelopmental
F98.*	Other behavioural and emotional disorders with onset usually occurring in childhood and adolescence	LD/neurodevelopmental
F99	Mental disorder, not otherwise specified	Non-psychotic illness
G10	Huntington's disease	Organic disorders
Q90	Down's syndrome	LD/neurodevelopmental

2 Multiple Imputation

2.1 Approach

Imputation was conducted using Multiple Imputation by Chained Equations (MICE) as implemented in the MICE package for the R software (van Buuren 2011) following previously published guidance for implementation and reporting (White 2011, Sterne 2009, Hayati Rezvan 2015). We initially investigated variable-wise and participant-wise missing data, and the distribution of variables in complete cases versus those with missing data. As recommended, the multiple imputation model included the primary and secondary outcomes, all predictors, accounting for pre-planned interactions, and relevant auxiliary variables (Kontopantelis 2017, Harel 2018, Perkins 2018, Tilling 2016); details provided in Table S4. A total of 50 imputed data sets were generated. Variable distributions from observed and imputed data sets were compared, and results from analyses of the individual datasets combined using Rubin's rules (Rubin 1987).

Table S4: Variables included in multiple imputation model

Type	Description
Outcomes	Occurrence of violent incident during detention
	Use of restrictive interventions
Predictors	Age and gender
	Acute alcohol use
	Recent use of cannabinoids and/or stimulants
	Psychotic symptoms
	Interaction between recent use of cannabinoids/stimulants and psychotic symptoms
Auxiliary	Presence of PoS proforma
	Past-year psychiatric diagnosis
	Detention outcome (discharge or admission and under what conditions)
	Detention destination
	Housing status
	Ethnicity
	Indicator of first or subsequent detention

We carried out a diagnostic check of the imputation model by comparing variable distributions from observed and imputed data sets, and compared results of the primary analysis using multiply imputed data to those from a complete record analysis. We also conducted sensitivity analyses to assess the impact of departures from the Missing At Random assumption. As Hayati Rezvan *et al.* (2015) note, there are multiple approaches to testing this assumption. van Buuren (2018) discusses using a δ-adjustment to explore the impact of variation in the missing data, had they been observed; following this we modelled a range of scenarios for the 'true' values of the missing data for our two main predictors - psychotic symptoms and cannabinoid/stimulant use. We generated imputations as described above (same model and number of data sets) under six different scenarios, one for each possible value that the missing data for each variable could take. van Buuren (2018) notes these types of scenarios are extreme and sensitivity analyses should preferably be based on realistic scenarios. However, our data could potentially be missing for very different reasons. It could be missing because individuals were so behaviourally disturbed that assessment was not possible. Conversely, if individuals were calm, recording absence of psychotic symptoms or substance use may not have had high clinical priority. This sensitivity analysis allowed us to explore how these different scenarios may affect the robustness of our results.

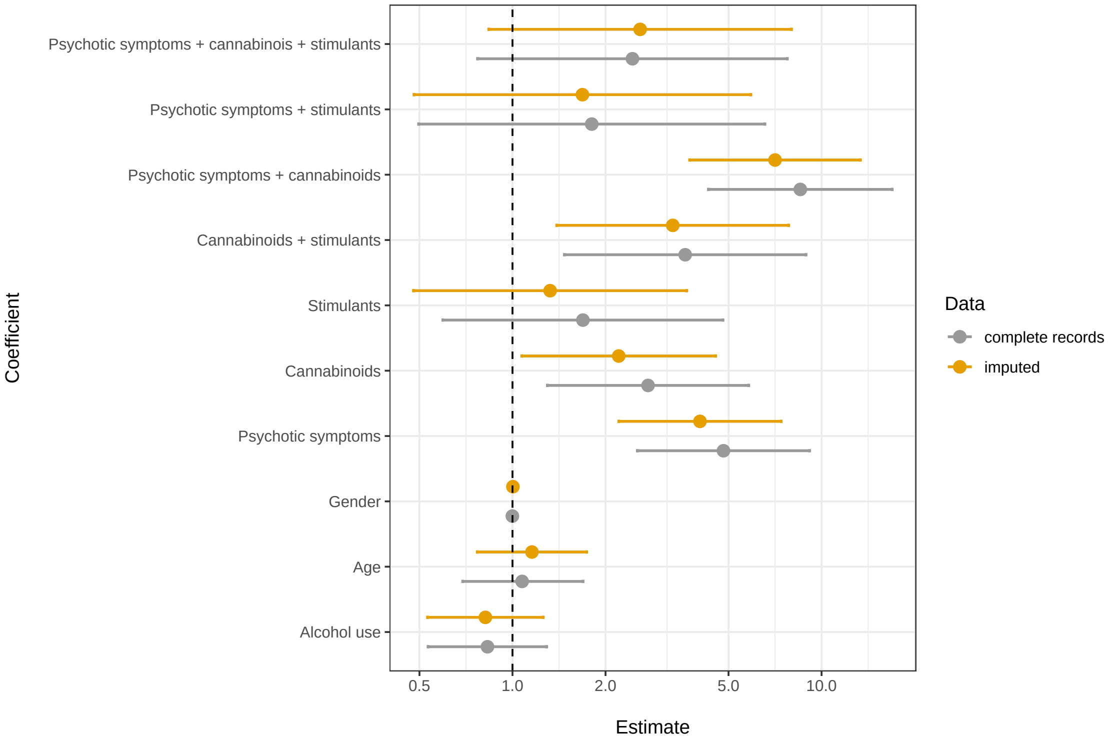
2.2 Results

We found little variation in the distribution of variables between the observed and imputed data sets. There was only slight variation in the point estimates for the regression coefficients between the imputed data and the observed data, which as expected given that data were only missing for a few variables, and the overall level of missing data was relatively low (see Table S5 for results from complete record analysis and Figure S3 for graphical comparison).

Table S5: Results from the complete record regression analysis

Variable	β	LCI	UCI	P
Psychotic symptoms	1.57	0.94	2.23	$<<$ 0.0001
Recent cannabinoid use	1.01	0.25	1.76	0.01
Recent stimulant use	0.52	-0.62	1.50	0.32
Recent cannabinoid and stimulant use	1.29	0.34	2.16	0.01
Acute alcohol use	-0.19	-0.63	0.25	0.41
Psychotic symptoms × cannabinoids	-0.44	-1.38	0.51	0.36
Psychotic symptoms × stimulants	-1.51	-3.25	0.09	0.07
Psychotic symptoms × cannabinoids and stimulants	-1.96	-3.48	-0.57	0.01
Gender	0.07	-0.37	0.53	0.75
Age	-0.0001	-0.02	0.02	0.91

Figure S3: Comparison of estimates (odds ratios with 95% confidence intervals) from the imputed and complete record analyses examining associations with primary outcome of violence during detention.



The estimates from the regression analyses of the data sets imputed for the six different sensitivity analysis scenarios were relatively consistent (see Figure S4 and Table S6). The confidence intervals for the estimates across the scenarios overlapped and the point estimates were very similar. There were a few exceptions, such as for the estimates of the odds ratios relating to cannabinoid use under the scenario in which all missing substance use data were labelled as cannabinoid use, but this is expected, particularly given that these groups had smaller numbers of detentions. Given that these scenarios represent extremes (i.e. we would not expect that for all detentions with missing substance use information the individuals had recently used cannabinoids), the results suggest that overall the analyses are robust.

Figure S4: Comparison of estimates (odds ratios with 95% confidence intervals) from the main multiple imputation analysis and sensitivity analyses under the different scenarios for missing data values.

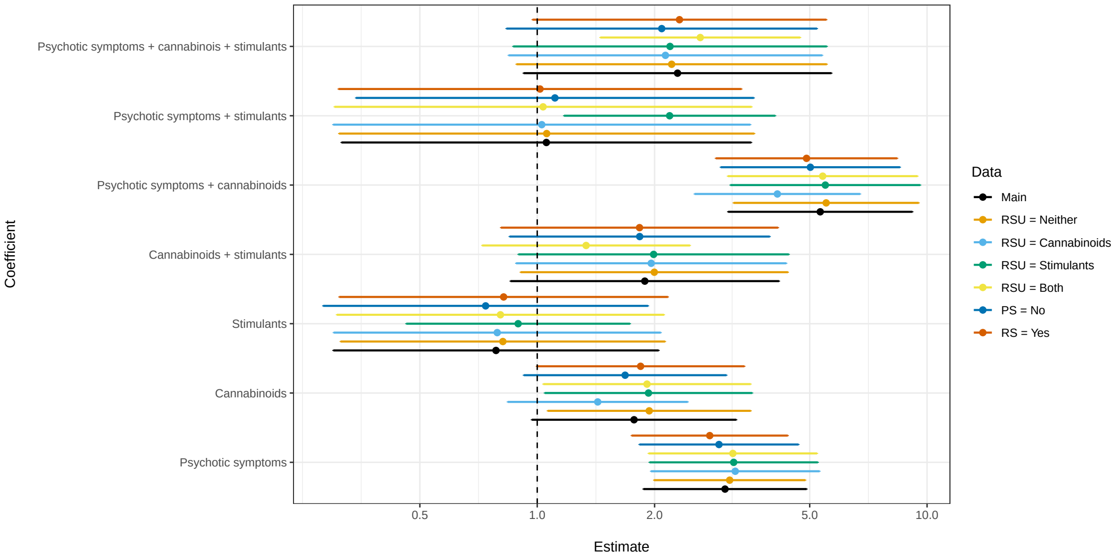


Table S6: Comparison of regression estimates from multiple imputation sensitivity analyses. Estimates shown as regression coefficients (95% confidence interval). RSU_m: indicates value of missing data for recent cannabinoid/stimulant use; PS_m: indicates value of missing data for psychotic symptoms.

Coefficient	Main	RSU_m = ‘Neither’	RSU_m = ‘Cannabinoids’	RSU_m = ‘Stimulants’	RSU_m = ‘Both’	PS_m = ‘No’	PS_m = ‘Yes’
Psychotic symptoms	1.1 (0.6 - 1.6)	1.1 (0.7 - 1.6)	1.2 (0.7 - 1.7)	1.2 (0.7 - 1.7)	1.2 (0.7 - 1.7)	1.1 (0.6 - 1.5)	1 (0.6 - 1.5)
Recent cannabinoid use	0.6 (0 - 1.2)	0.7 (0.1 - 1.3)	0.4 (-0.2 - 0.9)	0.7 (0 - 1.3)	0.6 (0 - 1.3)	0.5 (-0.1 - 1.1)	0.6 (0 - 1.2)
Recent stimulant use	-0.2 (-1.2 - 0.7)	-0.2 (-1.2 - 0.8)	-0.2 (-1.2 - 0.7)	-0.1 (-0.8 - 0.5)	-0.2 (-1.2 - 0.7)	-0.3 (-1.3 - 0.7)	-0.2 (-1.2 - 0.8)
Recent cannabinoid and stimulant use	0.6 (-0.2 - 1.4)	0.7 (-0.1 - 1.5)	0.7 (-0.1 - 1.5)	0.7 (-0.1 - 1.5)	0.3 (-0.3 - 0.9)	0.6 (-0.2 - 1.4)	0.6 (-0.2 - 1.4)
Psychotic symptoms × cannabinoids	0 (-0.8 - 0.8)	-0.1 (-0.9 - 0.7)	-0.1 (-0.8 - 0.6)	-0.1 (-0.9 - 0.7)	-0.1 (-0.9 - 0.7)	0 (-0.8 - 0.8)	0 (-0.8 - 0.8)
Psychotic symptoms × stimulants	-0.8 (-2.4 - 0.7)	-0.9 (-2.4 - 0.7)	-0.9 (-2.5 - 0.7)	-0.3 (-1.2 - 0.7)	-0.9 (-2.5 - 0.7)	-0.7 (-2.2 - 0.9)	-0.8 (-2.3 - 0.7)
Psychotic symptoms × cannabinoids & stimulants	-0.9 (-2.2 - 0.3)	-1 (-2.3 - 0.2)	-1.1 (-2.3 - 0.2)	-1.1 (-2.3 - 0.2)	-0.5 (-1.3 - 0.4)	-0.9 (-2.2 - 0.3)	-0.8 (-2 - 0.4)
Acute alcohol use	-0.3 (-0.7 - 0.04)	-0.3 (-0.7 - 0.05)	-0.3 (-0.7 - 0.04)	-0.3 (-0.7 - 0.04)	-0.3 (-0.7 - 0.03)	-0.4 (-0.7 - 0.02)	-0.4 (-0.7 - -0.01)
Gender	-0.1 (-0.4 - 0.3)	-0.1 (-0.4 - 0.3)	0 (-0.3 - 0.3)	-0.1 (-0.5 - 0.2)	-0.1 (-0.4 - 0.3)	-0.1 (-0.4 - 0.3)	0 (-0.4 - 0.3)
Age	-0.01 (-0.02 - 0.01)	-0.01 (-0.02 - 0.01)	-0.01 (-0.02 - 0.01)	-0.01 (-0.02 - 0.01)	-0.01 (-0.02 - 0.01)	-0.01 (-0.02 - 0.01)	-0.01 (-0.02 - 0.01)

3 Sensitivity Analyses

3.1 Approach

We conducted the main analyses at the *individual* level. Some individuals appeared in our full data set more than once as they were detained multiple times during the study period, but our main analyses were restricted to the first detention of each person during the study period. To assess the impact of this on results, we re-analysed our data including data from all detentions. As this was a sensitivity analysis and the data were complex due to only a minority of individual being detained multiple times, we used a simple approach and did not account for the non-independence of multiple detentions from the same person. We conducted this sensitivity analysis using both the multiply imputed data and complete records.

3.2 Results

As shown in Figure S5 and Tables S7 and S8, the results from the sensitivity analysis were attenuated compared to those based on the analysis of only the first detention episode, in that the point estimates for the odds ratios (ORs) indicated weaker associations between the two main predictors (recent substance use and psychotic symptoms) and the outcome of physical violence. This is expected given that where individuals were known to have behaved violently in previous detentions, clinical management measures to reduce the risk of this would be applied earlier, thus reducing the possibility of observing our outcome.

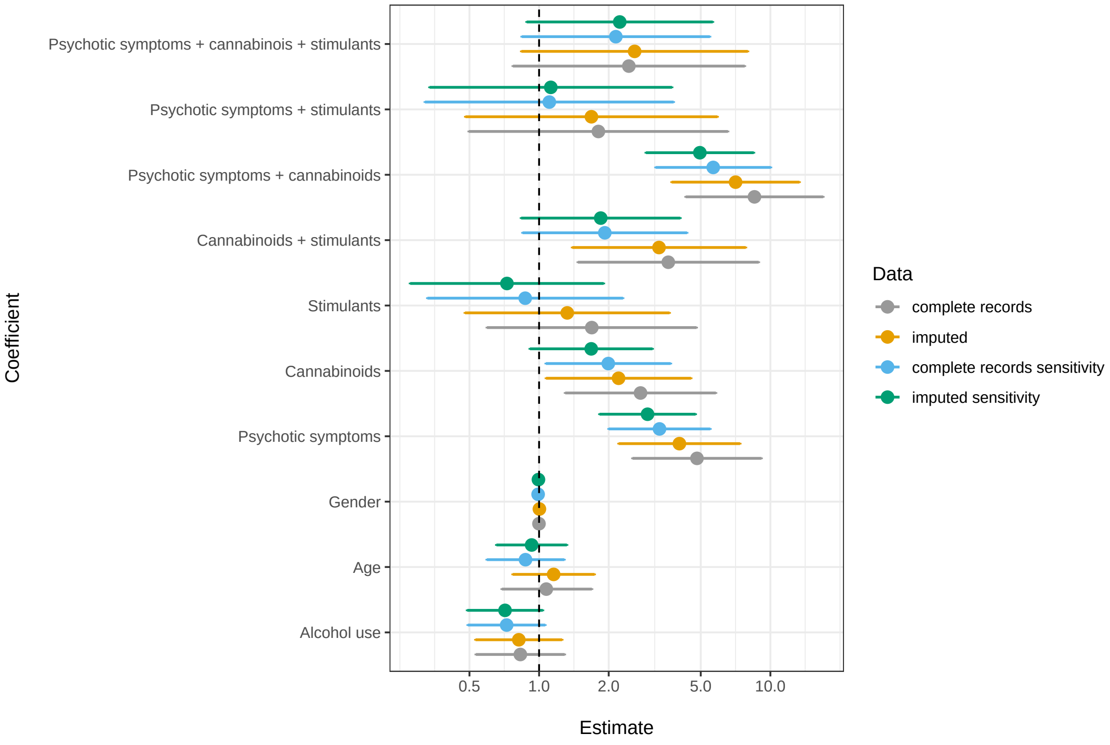
Table S7: Comparison of interaction results from the sensitivity analysis using multiply imputed data or complete records. Estimates are displayed as odds ratios (95% confidence interval); p-value.

Psychotic symptoms	**Recent cannabinoid/stimulant use**			
	Neither	**Cannabinoids**	**Stimulants**	**Both**
Complete				
No	ref.	2.0 (1.1-3.7); 0.03	0.8 (0.3-2.3); 0.8	1.9 (0.9-4.4); 0.1
Yes	3.3 (2.0-5.6); << 0.001	5.7 (3.2-10.0); << 0.001	1.1 (0.3-3.8); 0.9	2.2 (0.8-5.5); 0.1
Imputed				
No	ref.	1.7 (0.9-3.1); 0.1	0.7 (0.3-1.9); 0.5	1.9 (0.8-4.1); 0.13
Yes	2.9 (1.8-4.8); << 0.001	5.0 (2.9-8.5); << 0.001	1.1 (0.3-3.8); 0.9	2.2 (0.9-5.6); 0.09

Table S8: Comparison of regression coefficient estimates from the sensitivity analysis using multiply imputed data or complete records.

	Imputed		**Complete records**	
Variable	**β (95% CI)**	**P**	**β (95% CI)**	**P**
Psychotic symptoms	1.08 (0.6 - 1.56)	<< 0.0001	1.20 (0.69 - 1.71)	<< 0.0001
Recent cannabinoid use	0.52 (-0.09 - 1.13)	0.10	0.69 (0.05 - 1.30)	0.03
Recent stimulant use	-0.32 (-1.28 - 0.64)	0.52	-0.14 (-1.24 - 0.75)	0.78
Recent cannabinoid and stimulant use	0.61 (-0.18 - 1.41)	0.13	0.65 (-0.23 - 1.43)	0.12
Psychotic symptoms × cannabinoids	0.002 (-0.80 - 0.81)	1.00	-0.15 (-0.97 - 0.68)	0.71
Psychotic symptoms × stimulants	-0.64 (-2.20 - 0.92)	0.42	-0.96 (-2.65 - 0.58)	0.23
Psychotic symptoms × cannabinoids and stimulants	-0.89 (-2.13 - 0.35)	0.16	-1.09 (-2.37 - 0.16)	0.09
Acute alcohol use	-0.34 (-0.72 - 0.04)	0.08	-0.32 (-0.71 - 0.06)	0.10
Gender	-0.07 (-0.42 - 0.28)	0.68	-0.14 (-0.52 - 0.25)	0.49
Age	-0.01 (-0.02 - 0.01)	0.40	-0.01 (-0.03 - 0.01)	0.26

Figure S5: Comparison of estimates (odds ratios with 95% confidence intervals) from the four analyses examining associations with primary outcome: full imputed data set; full complete records data set; sensitivity analysis of imputed data; sensitivity analysis of complete records data set.



References

Carpenter *et al.* Sensitivity analysis after multiple imputation under missing at random: A weighting approach. *Stat Methods Med Res.* 2007;16(3):259–75.

Denaxas & Morley. Big biomedical data and cardiovascular disease research: Opportunities and challenges. *European Heart Journal - Quality of Care and Clinical Outcomes.* 2015;1(1):9-16

Harel *et al.* Multiple Imputation for Incomplete Data in Epidemiologic Studies. *Am J Epidemiol.* 2018;187(3):576–84.

Hayati Rezvan *et al.* The rise of multiple imputation: a review of the reporting and implementation of the method in medical research. *BMC Med Res Methodol.* 2015;15(1):30.

Héraud-Bousquet *et al.* Practical considerations for sensitivity analysis after multiple imputation applied to epidemiological studies with incomplete data. *BMC Med Res Methodol.* 2012;12.

Kontopantelis *et al.* Outcome-sensitive multiple imputation: A simulation study. *BMC Med Res Methodol.* 2017;17(1):1–13.

Morley *et al.* Defining Disease Phenotypes Using National Linked Electronic Health Records: A Case Study of Atrial Fibrillation. *PLoS One.* 2014;9(11):e110900.

Perera *et al.* Cohort profile of the South London and Maudsley NHS Foundation Trust Biomedical Research Centre (SLaM BRC) Case Register: current status and recent enhancement of an Electronic Mental Health Record-derived data resource. *BMJ Open.* 2016;6:e008721.

Perkins *et al.* Principled Approaches to Missing Data in Epidemiologic Studies. *Am J Epidemiol.* 2018;187(3):568–75.

Rubin. *Multiple Imputation for Nonresponse in Surveys.* New York: John Wiley & Sons Inc.; 1987.

Sterne *et al.* Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ.* 2009;339:157–60.

Tilling *et al.* Appropriate inclusion of interactions was needed to avoid bias in multiple imputation. *J Clin Epidemiol.* 2016;80:107–15.

van Buuren. *Flexible Imputation of Missing Data.* 2nd ed. FL: Boca Raton: CRC/Chapman & Hall; 2018.

van Buuren *et al.* Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software.* 2011;45(3):1–67.

White *et al.* Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in Medicine.* 2011;30(4):377–99.