

PEER REVIEW HISTORY

BMJ Open publishes all reviews undertaken for accepted manuscripts. Reviewers are asked to complete a checklist review form (<http://bmjopen.bmj.com/site/about/resources/checklist.pdf>) and are provided with free text boxes to elaborate on their assessment. These free text comments are reproduced below.

ARTICLE DETAILS

TITLE (PROVISIONAL)	Protocol for the Data-linkage Alcohol Cohort Study (DACs): Investigating mortality, morbidity, and offending among people with an alcohol-related problem using linked administrative data
AUTHORS	Peacock, Amy; Chiu, Vivian; Leung, Janni; Dobbins, Timothy; Larney, Sarah; Gisev, Natasa; Pearson, Sallie-Anne; Degenhardt, Louisa

VERSION 1 – REVIEW

REVIEWER	Arun Sondhi Therapeutic Solutions (Addictions) Ltd., London, UK
REVIEW RETURNED	05-Apr-2019

GENERAL COMMENTS	<p>The paper is a protocol for a data linkage study examining service utilisation of acute and problematic alcohol misusers in contact with ED and I/P departments in Australia. This is to be achieved through multiple data linkages across a number of sources.</p> <p>The protocol is clear and the design/analytical strategy is well-stated. The four aims are also clearly defined and appropriate to the design. The focus on specific cohorts e.g. 15-24 year olds (line 27, page 7), those who are readmitted and multiple representations (page 8) are necessary and relevant. The design benefits from being cautious about what it aims to deliver although there may be a few additions that may be worth considering.</p> <p>I have a few minor considerations for the authors given the scope of the study. The study will match patients from an 'index' treatment event e.g. first episode at an ED and I/P to other services. I wondered whether a limitation to be stated is the absence of linking to specialist alcohol services. Could the authors state why this wasn't included in the study design as it is an obvious gap in utilisation and maybe a confounding factor in subsequent engagement in acute services and the wider criminal justice system?</p> <p>The methodological consideration section (page 9) discusses the inclusion of new participating EDs - there is a likely issue with data quality. Can the authors mention how they plan to examine (and possibly adjust for) variable data quality within acute services especially and the other services? In other words, how confident are the authors that the data is of sufficient quality to complete the aims as suggested? Describing a data cleaning process would be helpful as part of the methods section.</p> <p>The probabilistic matching process looks standard pp6-7, but can the authors had more detail on this? Will the probabilistic match include adjusting for name changes/transposing of names etc. Do</p>
-------------------------	---

	the authors plan to comment on the success of matching across datasets and the possible extent of false-negative reporting? Links to CJS datasets (especially) will be fraught with aliases and whether the study design has built in any consideration of adjusting the matching criteria to adjust for any possible missing matches?
REVIEWER	Michaël Schwarzinger Translational Health Economics Network (THEN), Paris, France
REVIEW RETURNED	05-May-2019
GENERAL COMMENTS	The proposed study is of interest. I have only minor comments: 1) Page 8, lines 7-15: what is meant by a "formal or statistical" admission/discharge? This section is not clear. 2) Page 8, lines 37-42: cause-specific mortality rates of interest could be detailed.

VERSION 1 – AUTHOR RESPONSE

Reviewer(s)' Comments to Author:

Reviewer: 1. Reviewer Name: Arun Sondhi. Institution and Country: Therapeutic Solutions (Addictions) Ltd., London, UK. Please state any competing interests or state 'None declared': None declared

1. The study will match patients from an 'index' treatment event e.g. first episode at an ED and I/P to other services. I wondered whether a limitation to be stated is the absence of linking to specialist alcohol services. Could the authors state why this wasn't included in the study design as it is an obvious gap in utilisation and maybe a confounding factor in subsequent engagement in acute services and the wider criminal justice system?

Response: We agree with the reviewer that the lack of a linked dataset systematically capturing all engagement with specialist alcohol treatment services is a limitation of this study, which is why we had flagged this in the 'Study Limitations' section and 'Methodological Considerations' section. Ideally, we would have linked the cohort to the Alcohol and Other Drug Treatment Services National Minimum Dataset, however linkage to this dataset was not available for this study. In saying this, the NSW Mental Health Ambulatory Data Collection captures ambulatory mental health care dedicated to the assessment, treatment, rehabilitation or care of non-admitted patients. This may include mental health day programs, psychiatric outpatients and outreach services, care by hospital-based consultation liaison services, care in community residential settings, and mental health promotion and prevention services. This dataset has not been used in research to capture alcohol-related prevention and intervention service delivery; until we have the capacity to study the data, we cannot know the quality of information provided. We also know that it will not capture all prevention and intervention activities related to alcohol. We have consequently now adapted the 'Methodological Considerations' section to read:

"Thirdly, data on alcohol consumption, as well as intervention and treatment for problematic alcohol use, cannot be systematically ascertained from the included data sources. The MH-AMB captures ambulatory mental health service provided to non-admitted individuals. To the authors' knowledge, this data set has not been used to quantify engagement in prevention and intervention for alcohol-related problems in previous linkage studies. Until we have the capacity to study the data, we cannot know the quality of information provided however, there may be the capacity to explore use of this data set to capture such engagement and take this into consideration in analyses." (page 9)

2. The probabilistic matching process looks standard pp6-7, but can the authors had more detail on this? Will the probabilistic match include adjusting for name changes/transposing of names etc. Do

the authors plan to comment on the success of matching across datasets and the possible extent of false-negative reporting?

Response: We thank the reviewer for this suggestion. We have added the following details to the 'Datasets and Linkage' section:

"On identifying the base cohort, the CHeReL will extract linked data for these individuals from a range of routinely-collected administrative data sets using the probabilistic record linkage software ChoiceMaker^{14,15}. Identifying information (i.e., name, address, date of birth and gender) for each dataset is included in the Master Linkage Key (MLK) constructed by the CHeReL. The ChoiceMaker software uses an exact 'blocking' algorithm to search for valid matches in the MLK to identify all matching records. A combination of two techniques is used to determine whether each potential match denotes (or possibly denotes) the same person, comprising: i) a probabilistic decision, which is computed using a machine learning technique, and ii) absolute rules, which include upper and lower probability cut-offs which initially start at 0.75 and 0.25 for a linkage and are adjusted for each individual linkage to ensure false links are minimised. The parameters for the extract from the MLK are set such that no true matches are missed if full identifiers are available. Extensive quality assurance measures ensure the false positive rate for linkage is less than 0.5% and the false negative rate for linkage is less than 0.1%¹⁵. All datasets except the NSW Re-offending Database (ROD) are routinely contained within the MLK. The internal matching process of the NSW Re-offending Database (ROD) has been validated (specificity of 99.9 percent and a sensitivity of 93.8 percent¹⁶), and linkage of records within the MLK to those within ROD will follow the same process as above." (page 6)

The CHeReL have implemented rigorous quality assurance measures, including manual clerical review of records where, for example:

- Date of birth and first 4 letters of first name and first 4 letters of surname differ
- Males where date of birth and surname differ
- Compare given names and surname (first 3 characters of given name and first 4 characters of surname after stripping spaces and non-alpha characters)

The most recent manual clerical review yielded a false positive rate of 3/1000 linkages. For this reason, the CHeReL state that the overall percentage of records and persons affected is very small. If researchers find inconsistent information in linked records, it is recommended that these records be excluded from the analysis and the number and proportion of records excluded be reported in the methods section of the study report. We have consequently added the following statement to the 'Methodological Considerations' section:

"Finally, the false positive rate for linkage is less than 0.5% and the false negative rate for linkage is less than 0.1%¹⁵. We will compare time-independent information (e.g., date of birth, date of death) across datasets to identify inconsistencies that may be indicative of false positive linkages. These participants will be excluded from the cohort and identified in all reporting on final cohort composition." (page 9)

3. The methodological consideration section (page 9) discusses the inclusion of new participating EDs - there is a likely issue with data quality. Can the authors mention how they plan to examine (and possibly adjust for) variable data quality within acute services especially and the other services? In other words, how confident are the authors that the data is of sufficient quality to complete the aims as suggested? Describing a data cleaning process would be helpful as part of the methods section.

Response: We appreciate the reviewer's concern about the quality of linkage from acute services. We refer the reviewer to the response above, which overviews the process implemented by the CHeReL to ensure reliable linkage of records. We also overview the key data cleaning steps in the 'Methodological Considerations' section with respect to additional checks by the research team to confirm linkage quality (see response above). A full description of the data cleaning process is beyond the scope of this paper (and challenging to fully specify a priori without study of the data). However,

we are confident that the EDDC data quality is sufficient to complete the aims of this study. The EDDC data have also been used extensively in previous linkage projects. Examples include:

- Perera, J. et al. (2018). Presentations to NSW emergency departments with self-harm, suicidal ideation, or intentional poisoning, 2010–2014. *Medical Journal of Australia* 208;8; 348-353.
- Carr, V. J., et al. (2016). New South Wales Child Development Study (NSW-CDS): an Australian multiagency, multigenerational, longitudinal record linkage study. *BMJ Open* 6;2:e009023.
- Mitchell, R. J., et al. (2013). Can SNOMED CT as implemented in New South Wales, Australia be used for road trauma injury surveillance in emergency departments?. *Health Information Management Journal* 42;2; 4-8.
- O'Connell, Dianne L., et al. (2014). Acute hospital-based services utilisation during the last year of life in New South Wales, Australia: methods for a population-based study. *BMJ Open* 4;3:e004455.

We have added text in the manuscript to indicate that, where necessary, we will run analyses with and without the EDs that were added later in the series to establish if this has changed the study outcomes to account for variable data quality from inclusion of new participating emergency departments in the EDDC.

4. Links to CJS datasets (especially) will be fraught with aliases and whether the study design has built in any consideration of adjusting the matching criteria to adjust for any possible missing matches?

Response: All datasets, excepting the NSW Re-offending Database (ROD), are routinely contained within the MLK. The internal matching process of the NSW Re-offending Database (ROD) has been validated (specificity of 99.9 percent and a sensitivity of 93.8 percent). Linkage of records within the MLK to those within ROD will follow the same process as above, including rate of false positive and false negative linkages.

“All datasets, excepting the NSW Re-offending Database (ROD), are routinely contained within the MLK. The internal matching process of the NSW Re-offending Database (ROD) has been validated (specificity of 99.9 percent and a sensitivity of 93.8 percent¹⁶), and linkage of records within the MLK to those within ROD will follow the same process as above.” (page 6)

Reviewer: 2

Reviewer Name: Michaël Schwarzingler. Institution and Country: Translational Health Economics Network (THEN), Paris, France. Please state any competing interests or state 'None declared': None declared. Please leave your comments for the authors below. The proposed study is of interest. I have only minor comments:

1. Page 8, lines 7-15: what is meant by a "formal or statistical" admission/discharge? This section is not clear.

Response: We have revised the definition of “period of stay” and “episode of care” (which formerly incorporated these terms) as:

“We will estimate the number of hospital separations in two ways, as a count of episodes of care and of periods of stay¹⁹. A period of stay will be defined as the complete period of care from admission to hospital until separation. A period of stay may consist of multiple episodes of care, the latter defined as a period of a specific care type (e.g., receipt of acute care and rehabilitation may be coded as two episodes-of-care within one period-of-stay).” (page 7)

2. Page 8, lines 37-42: cause-specific mortality rates of interest could be detailed.

Response: We have provided the following information about cause-specific mortality rates:

“We will classify deaths in accordance with guidance for clustering major causes of mortality²¹, with a focus on causes of death wholly or partly attributable to alcohol consumption²².” (page 8)

VERSION 2 – REVIEW

REVIEWER	Arun Sondhi Therapeutic Solutions (Addictions) Ltd.
REVIEW RETURNED	18-Jun-2019
GENERAL COMMENTS	Thank your for your clear and concise explanatory notes to my comments. I look forward to reading the findings from the study in due course.