Information statement
Evaluation of automated de-identification of general practice free text health records

This information is only for practices contributing to or willing to contribute to Data for Decisions (D4D). It details information about a new research and technical study that the Data for Decisions team are undertaking.

What is the aim of this project?

As part of participation in D4D, you have been providing us (the Department of General Practice Data for Decisions team) with regular sets of de-identified data from your electronic medical record (EMR). We wish to eventually expand the D4D research initiative to include additional **de-identified** free text data, specifically progress notes and pathology test results. This will allow us to better research important areas such as adolescent health and family violence.

To do this, we first need to understand how well the de-identification methods we currently use work on these specified free text fields and determine whether alternate methods might work better.

What does participation involve?

If you consent to participate, no work will be required from you other than your completion and return of the consent form. **One instance** of progress notes and pathology test result data from your practice will then be securely de-identified and extracted using the GRHANITE® data extraction tool that is already at your practice (if you are a D4D partner; if you are not, please speak to Dr Rachel Canaway, details below). We will apply the same processes we use when extracting the usual data for D4D.

Why is this project needed?

There are a number of new methods of de-identifying free text data being used overseas; however, we need to determine whether they work for Australian general practice data and how they compare to the de-identification methods currently used by the D4D team when extracting data from other parts of the EMR. In this project, we are trialling and evaluating these existing methods and may develop new methods.

Automatic anonymising or pseudonymising (i.e. replacing names with pseudonyms from extensive and culturally diverse name dictionaries) can present limitations; for instance, progress notes can include names of diseases that include person names. If de-identification methods change, for example, a diagnosis of ‘Crohn’s disease’ to ‘Smith’s disease’ or ‘Anonymised disease’, that is not helpful for researchers who have no way of knowing what the disease being referred to was prior to it being ‘de-identified’.

Why are free text data important?

These data are very important for research and audit and feedback activities as they include information about social circumstances, clinical presentations and actions undertaken that are not captured elsewhere in the EMR. Currently, this information can only be extracted manually by researchers visiting a health care facility and examining records on-site, which results in researchers' exposure to patient identifying information (i.e. poorer privacy protection). Enabling more cost-effective analysis of things such as social circumstance will make a huge difference in allowing researchers to safely and effectively study areas of great need.

This project aims to identify safer and more efficient methods for extracting free text data while preserving patient privacy. Ultimately, the aim is that, through research, greater understanding of a broader range of conditions and clinical practices can be gained that could contribute to better care and outcomes for patients.

Does my practice have to take part?

Your contribution to this study is voluntary. If you decide not to take part, or consent and then later withdraw, this will not impact your participation in Data for Decisions or your relationship with the University of Melbourne in any way.

How will this data be used?

The research team will be trialling and evaluating new and enhanced de-identification methods. As part of this research, the D4D technical team are seeking to move from anonymising data (e.g. changing 'Dr Smith' to 'Dr [Anonymous]') to pseudonymising data (e.g. changing 'Dr Smith' to 'Dr Jones'). This means that, in the rare occurrence that a name or other identifying information is missed during de-identification, someone looking at the data will not know whether they are viewing a 'real' name or not. Pseudonymising is considered one of the safest ways to protect privacy when extracting data from EMR.

After this research is complete, the research team will undertake a further, separate study with the de-identified free-text data using natural language processing, which is a technique used to facilitate computer-assisted analysis to understand how people use language and to analyse the privacy protected free text data so it might lead to more meaningful use. You can opt now to be part of this next study (if you decide to participate in the current de-identification study), or you can wait and be contacted again later to seek your willingness to allow the extracted EMR data to be subject to the additional analyses.

How will this data be protected?

Data from progress notes and pathology test result fields will be de-identified by GRHANITE® before extraction, using the methods currently used by D4D. This means that just like the data fields currently extracted, the progress notes and pathology test results will also have person identifying information removed before it leaves your practice.

All data will be stored on secure servers at the University of Melbourne as per the usual D4D processes and can only be accessed by approved members of the research team. The data will not leave the secure confines of the University of Melbourne data environment nor will it be used for other projects.

The data used for this project will be destroyed 5-years after publication of findings from the study. There is no way that your practice or patients could be identified in any publications arising from this study.

If I consent now will our free text fields continue to be extracted regularly by GRHANITE®?

No, we are seeking your consent to extract the specified free text fields once, at one time point. In the consent form you have the option for your data to be used for the first phase or both phases of the study: (1) the study to optimise de-identification methods, and (2) the second phase which will employ natural language processing techniques to increase the meaningful and cost-efficient use of privacy-protected free text EMR data. We will have more detailed information about the second phase of the study available later.

In the future, depending on the findings from this study, you might later be approached for consent to allow regular extraction of these specified free-text fields for their incorporation into the Patron primary care data repository. Once in Patron the data would be available for other researchers to use, pending their ethics and Data Governance Committee approvals. You are NOT being asked now to provide consent for ongoing extraction of the specified free text fields.

Who is involved?

The Data for Decisions team in the Department of General Practice is collaborating with a group of data and technical specialists from a number of departments within the University of Melbourne, including researchers associated with the Melbourne Data Analytics Platform. This project is part of a larger evaluation of de-identification methods for free text data from both general practice and hospital settings.

The full group of investigators is listed below and on the consent form. All investigators will be required to follow the strict D4D protocols and sign agreements regarding data access and use.

Who can I contact for additional information and to take part?

If you have any questions, either before or after taking part in this research, please contact us via the details below.

Dr Rachel Canaway, D4D / VicREN Manager
Department of General Practice, University of Melbourne VIC 3010 Australia
T: (03) 8344 3392
E: rcanaway@unimelb.edu.au

Research team

A/Prof Douglas Boyle, Prof Karin Verspoor, Dr Noel Faux, Ms Priyanka Pillai, Ms Kim Doyle, Dr Simon Mutch, Dr Daniel Capurro, Prof Jane Hocking, Prof Lena Sanci, Prof Wendy Chapman, Ms Carol El-Hayek, Mr Warwick Strangward, Mr Roger Ward, Ms Alaina Vaisey

Ethics approval

This project was approved by the University of Melbourne Human Research Ethics Committee on 27 July 2020. (HREC ID: 2057196.1)

If you have concerns about the conduct of this research, please contact:

Manager, Human Research Ethics and Integrity, the University of Melbourne
(03) 8344 2073
humanethicscompaints@unimelb.edu.au
Ethics ID: 2057196.1

GENERAL PRACTICE CONSENT FORM

Project: Evaluation of automated de-identification of general practice free text health records

Investigators:

A/Prof Douglas Boyle, Prof Karin Verspoor, Dr Noel Faux, Ms Priyanka Pillai, Ms Kim Doyle, Dr Simon Mutch, Dr Daniel Capurro, Prof Jane Hocking, Prof Lena Sanci, Prof Wendy Chapman, Ms Carol El-Hayek, Mr Warwick Strangward, Mr Roger Ward, Ms Alaina Vaisey

1. I/we consent to this general practice participating in this research study: *Evaluation of automated deidentification of general practice free text health records*, a project that is part of the *Data for Decisions* program of research. I/we are satisfied that the details of the project have been explained.
2. I/we have read the Information Statement (Plain Language Statement v1.0 HREC ID: 2057196.1 Date: July 2020) and have been provided with a copy to keep.
3. I/we have had the opportunity to ask questions and discuss the project and have received adequate information to inform the decision for the general practice to participate.
4. I/we agree that the researchers will extract de-identified progress notes and pathology test results data from this practice as described in the Information Statement.
5. I/we understand that this data will be stored on secure servers at the University of Melbourne as per *Data for Decisions* protocols and can only be accessed by approved members of the research team.
6. I/we understand that this data will be used by the research team to trial and evaluate new de-identification methods.
7. I/we understand that after signing and returning this consent form via email it will be retained by the University of Melbourne. A copy of this signed consent form will be retained by the general practice.
8. I/we understand that this project is funded by the University of Melbourne.
9. I/we acknowledge that:

    a. This general practice is free to withdraw from this project at any time without explanation or prejudice and this will not impact on the practice's participation in *Data for Decisions* or the relationship with the University of Melbourne;

    b. I/we have been informed that the confidentiality of the data provided by this practice will be safeguarded subject to any legal requirements;

    c. I/we have been informed that this consent form and information will be stored securely at the University of Melbourne and destroyed 5 years after publication of the study results; and

    d. In accordance with the law of Victoria, I/we understand that it is possible for data to be subject to subpoena, or freedom of information request.

Additional Consent Option (Please circle if appropriate)

a. I/we do consent for the de-identified progress notes and pathology tests results used in this study to be **used for additional analysis in a future study** investigating natural language processing techniques, if the outcomes of this study, after peer review, indicate that it is safe to do so.

General practice name: *General practice stamp*	
Name of authorised person:	
Signature: *Position in the practice:*	*Date:*
Name of authorised person:	
Signature: *Position in the practice:*	*Date:*