	Policy and Guidance	Page 1 of 4
	ISMS-07-02	Effective date: 29 Nov 2018
	TRE Data Validation	Version: 1.1

Guidance for validation of datasets received from data providers

Description of problem

Researchers are often provided with de-identified datasets of routinely-collected data, such as electronic health records (EHR), for research purposes. In many cases a data provider will have to prepare an extract of the dataset which will then be transferred to the researchers. Due the complexity of the datasets, which are often poorly described, the process of preparing a data extract can introduce errors. Such errors could include, for example:

- missing / additional fields
- missing / additional rows
- unexpected format of fields
- unexpected format of files
- unexpected coding of fields (e.g. missing values, dates)
- potentially disclosive information being released (e.g. in free text fields)

Such errors may result in:


- delays in data processing (prepared code is not compatible with the dataset)
- delays resulting from having to request a corrected extract
- knock-on impact on results of analyses, and potentially problems replicating results or getting outputs disclosure checked
- accidental re-identification of patients

These recommendations aim to help minimise the risk of such errors occurring, to encourage validation of datasets to detect potential issues as soon as possible, and to encourage researchers to have a process in place if errors are detected at a later date.

Recommendations

1) Develop a clear data specification in advance of receiving data

Having a clear specification will have three important results: firstly, the researchers should gain an understanding of the datasets that will help them plan their analyses and have realistic expectations about the quality and utility of the data; secondly, it should be useful for the data providers when preparing the extraction; and thirdly, it can be used to validate the data extraction.

	Policy and Guidance	Page 2 of 4
	ISMS-07-02	Effective date: 29 Nov 2018
	TRE Data Validation	Version: 1.1

Researchers should think about what information will be needed to answer the research question and find out if this is available, in what format, for how many patients and for what date range.

The data specification should be designed by the researchers in close collaboration with the data providers. In the first instance, researchers should request a data specification or data dictionary from the data providers. If this is unavailable, then a new specification will need to be developed. It can also be useful to seek guidance from a clinician/practitioner familiar with the data collection process on the meaning/purpose of the different fields. Questions to ask about potential fields for the specification could be whether the field is entered by users or machine-generated, levels of accuracy, and the meanings of blanks and missing values.

Minimising the risk of re-identification of patients

It is important to consider the potential risk of re-identification of patients within the dataset. Although the datasets should have direct identifiers removed (including names, addresses, exact dates of birth), the richness of the datasets creates the risk of re-identification. Factors that increase the risk of re-identification include: the number of fields requested, the frequency of coded events within fields, and the provision of free text fields. We recommend that researchers:

- Request the minimum number of fields needed to answer the research question
- Apply some kind of minimum frequency rule, e.g. do not request codes used less than 10 times in the dataset
- Do not request free text information unless absolutely necessary. Discuss requirements for free text fields with the data provider as these are particularly risky. There are algorithms for detecting and possibly removing identifiers from free text, and also for converting free text into medical codes. Some are available on the open web, and others are available from CHC researchers such as Goran Nenadic. These could be gathered and used as needed by the research team.


Examples of details to specify in advance

The following list covers some of the key details to agree in advance of a data extract.

- Exactly what fields are needed, and their names
- The format of each of the fields requested
- How missing data is coded within each field
- How fields such as dates are to be coded
- Which patients to include in the extract (e.g. provide a list of inclusion/exclusion codes, ages, locations)
- What date range is required
- What format the files will be, and the character chosen to delimit the fields

Common sources of errors

These errors are known to have led to problems for data extractions:

	Policy and Guidance	Page 3 of 4
	ISMS-07-02	Effective date: 29 Nov 2018
	TRE Data Validation	Version: 1.1

The separator used to delimit fields when data is provided in flat text files. Commas are frequently used as separators but can easily appear *within* a data field. This will result in incorrect definition of fields when the file is read as a table. (This would be apparent if there were more fields than expected in the files). Recommendations:

- Do not request comma-separated files. Instead, request a less frequently used character (e.g. tabs or pipes). Also, advise the data provider to do a search-and-replace within the dataset *before* performing the extraction to remove any instances of your chosen separator within fields.


The presence of carriage returns / line breaks within fields when data is provided in flat text files. This is a particular problem when free text fields are requested: any character may be present within blocks of text. Presence of line breaks within fields will result in additional rows being generated. Recommendations:

- Request the data provider to do a search-and-replace for carriage returns/line breaks within fields before performing the extraction

2) Perform validation checks upon receipt of data

Validation checks should be performed as soon as possible after receipt of the data – it will be easier for the data extractor to resolve any issues in a timely manner, and will ensure errors are detected before the analyses are begun.

- Ask the data provider for the following details:
 - Number of files
 - Number and labels for fields within each file
 - Number of rows within each file
- Compare the files received to the details above. Open each of the files as data tables in your software of choice. Check the number of fields and rows – the incorrect number of fields can imply problems with the delimitation of fields, the incorrect number of rows can indicate erroneous line breaks.
- Check the format of fields against the data specification
- Check the data – is it coded as expected (e.g. missing fields, dates)? Simple descriptive statistics will help visualise the data and may highlight unexpected values
- Carefully review any free text fields for potentially disclosive information. Perhaps take a random sample of the data and review the text.

	Policy and Guidance	Page 4 of 4
	ISMS-07-02	Effective date: 29 Nov 2018
	TRE Data Validation	Version: 1.1

3) Have a process in place if problems are discovered later

You need a plan for what to do if problems are encountered in any of the validation steps above. Who needs to know within the research group, elsewhere in the institution, and externally e.g. at the organisation who provided the data? Familiarise the group with information security incident procedure, for example if you find identifying information in a supposedly de-identified dataset.

UNCONTROLLED IF STORED LOCALLY OR PRINTED