Feedback loops: Using data insight to stimulate data sharing across NHS Accident and Emergency departments.

B209460

03/07/2022

Word count - 1954

# Introduction

The sharing and analysis of routine data plays a vital role in performance management of NHS A&E departments, allowing staff and management to understand trends and employ strategies that mitigate risk.  As Dame Fiona Caldicott has noted, “The duty to share information can be as important as the duty to protect patient confidentiality.”[[1]](#endnote-1)

# Data

I have utilised data downloaded from the open-source A&E attendances dataset[[2]](#endnote-2). The dataset Is part of the NHSRdatasets package which has been created to support skills development in the NHS-R community. The dataset consists of 12,765 rows of data and six columns with different classes as follows:

1. **Period:** Date variable, the date of each submission to the a&e attendances dataset. Each submission represent one month.
2. **Org\_code:** Factor variable, the NHS org\_code reference for each a&e department
3. **Type:** Factor variable, the department type for each org\_code (either 1, 2 or other)
4. **Attendances:** Integer variable, the number of attendances to a&e within period
5. **Breaches:** Integer variable, thee number of attendances that breaches the four hour waiting time
6. **Admission**: Integer variable, the number of attendances that were admitted to hospital.

The data is complete, no further action was required to handle missing data. The data has been collected for the years 2016/17 through 2018/19 (Apr-Mar). This dataset was saved in csv format within the Rawdata folder, with the title “ ae\_attendances.csv”.

For my data capture tool, I have taken a subset of this data, entitled “[ae\_attendances\_VariablesForAnalysis.csv](https://noteable.edina.ac.uk/user/3n6bew7i3pe75jngbuwezh/edit/B209460/B209460-Assessment/RawData/ae_attendances_VariablesForAnalysis.csv)” which includes the variables period, org\_code, attendances and breaches. A further variable was added to this dataset, Index (integer) in order to maintain reference to the raw datasets when subset into test and training datasets for the development of my data capture tool. Finally, the data was divided into test and training data and 12 random records were selected into the “ae\_attendances\_test.csv” to test my data capture tool.

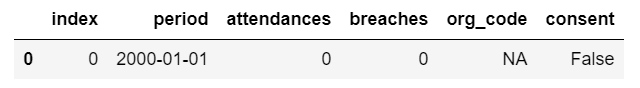
# Data Capture Tool

The data capture tool was developed using a python script. The following packages were downloaded:

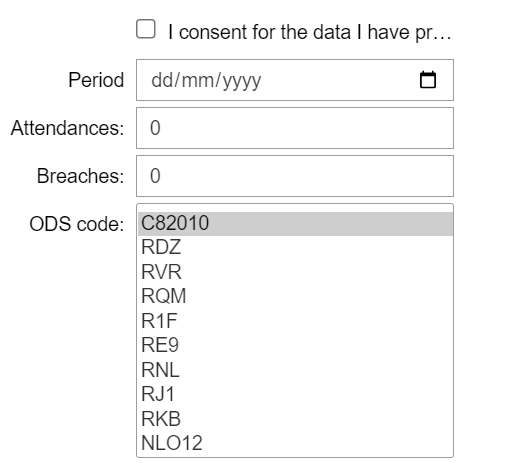
**Pandas**: supports the creations, manipulation and use of dataframes within python and is an important package for data scientists working with data in table or array form.

**Ipywidgets:** allows for the creation of widgets of multiple types and is an easy-to-use way to capture user-entered data.

In order to create the data capture tool, I first created an empty data frame using the pandas package. This dataframe contained all of the variables described above with one additional variable – “Consent”, which is a Boolean variable (true or false) and will be used to capture consent to process the captured data in accordance with the University of Edinburgh’s Research Data Management policy[[3]](#endnote-3).



Widgets were then created for each variable that the data capture tool will capture and then organised into a single vertical form to ease data input. A message to users was then included that explained the purpose of the collected data and thanked users for their input.



# Data Management

The following data management plan is designed to ensure that the data collected and used within this project (including scripts and other project resources) adheres to the FAIR principles (data is Findable, Accessible, Interoperable and Reusable)[[4]](#endnote-4).

**Data collection**

The above describes the data collection in detail. Collected data will be stored in CSV format to maximise interoperability with different software and coding languages.

**Data quality and standards**

Quality of the NHS datasets used to test the data capture tool will be checked for missing values and consistency of variable types. The data capture tool will ensure the capture of quality data through widgets (which will permit only correct variable types).  Timeliness of data capture is an important metric of data quality, which will be monitored through monthly spot checks by the research lead and direct contact with submitting org units.  These measures of quality closely follow those proposed by Laura Sebastian-Coleman. [[5]](#endnote-5)

**Metadata standards and data documentation**

The Research Data Alliance standards will be followed for the collection of metadata. And a data dictionary will be created and maintained using the Datameta package (R) to outline each of the collected variables.  Documentation for the creation of this dictionary will be included within the R scripts and the dictionary will be widely available through the project resources (shared via Github).  A written report will be provided which fully outlines all steps in data creation, documentation and storage.  Also available via Github.

**Ethics & Legal Compliance**

No person identifiable data will be stored within this project.  The project proposal will be subjected to an ethical review board before the commencement of data capture.  Sharing and use of the collected data will be limited to researchers who are able to demonstrate training in data and research ethics and upon review of a proposal for use of the collected data.

**Backup**

A backup of the GitHub repository will be taken on a three-monthly basis by the project lead and stored on a USB drive.

**Sharing**

All collected data will be stored long-term for potential re-use for further studies.  Such studies will be reviewed before provision of access, however, the collected data may provide a useful resource for future studies, especially in combination with other datasets and so it is anticipated to be shared with research communities, particularly those working within the NHS across the United Kingdom and other health institutions worldwide.  Data will be made available through the university of Edinburgh discussion boards and will be used to prompt discussion among students and further analysis.  The final report, including access to the raw and analysed datasets and all scripts and materials will be published via online journals and clear instruction on how to request access to the dataset will be provided. Finally, a user license will be provided within the root folder of the Github repository to further allow the use of the data and tool for anyone with repository access

**Coding**

Two scripts were used to analyse the data and to create the data capture tool. Both scripts were developed in logical and well annotated ways and care was taken to ensure that each section of code was well annotated describing the purpose of each coding chunk and some description of any R or python packages that were employed to obtain results. Care was also taken to use coding that is as simple as possible in order that the code will be easily understood by any parties wishing to re-use it of re-produce the results. The coding languages python and R were chosen as they are the most widely used languages for data science and therefore their use will further aid the ease at which results can be re-created.

Finally, the file structure is in un-spaced format to aid interoperability and mirrors normal file structure formats for research / coding practices (segregating Raw from processed data and using separate folders for scripts, outputs, tables and figures). Naming of data sets is also consistent and clear to improve the accessibility of this data. This structure is synced and shared using a Github repository and clear version control is used to provide additional information and the ability to see changes to coding practices and the development of the project. Access to the repository will be controlled using an access key.  Access to the repository will be monitored by the project lead to ensure no access beyond the stated use within a project proposal.

**Data storytelling**

Data captured using the data capture tool will be analysed to provide insight for NHS staff and Management that can be used to drive A&E efficiency. A performance metric will be calculated by taking the percentage of attendances that resulted in a breach of the four-hour wait time each month. This will then be plotted against time to outline seasonal fluctuations in breaches to the wait time. This information will be valuable to NHS staff and management to allow them to plan for anticipated breaches in the wait time.

**Future work**

Future work using the collected dataset will include the creation of a linear regression model to further predict breaches to the four hour wait time. In order to complete this analysis, an additional variable will be added to the data capture tool – department type (factor). A linear regression model will then be created that utilises period (month), department type and attendance figures to predict the point at which any org unit is likely to breach its four hour waiting time limit. This will provide much more detailed information into potential future breaches and is likely to uncover unforeseen insights and lead to new research that can improve A&E management practices.

**Reflective practice**

**Data management process**

In working through my data management plan on DMPonline and through reading Michener’s rules for data management plans[[6]](#endnote-6), I realised that there are some areas of this work that I need to further develop my understanding. In particular, the balance between data ethics and providing data that is widely available and accessible, yet also having due regard for unintended consequences of data misuse. I think that this area is perhaps still unclear within my plan and reflects a need for further future study but much would come down to the ethics committee reviewing a data plan and it’s use.

**Coding practice**

During the first half of this course, there were a lot of areas that I didn’t quite understand, for example I didn’t know what a data capture tool was, or it’s intended purpose and I didn’t understand the link between downloaded data, insight, and the data capture tool. When developing the code for the data capture tool, I think this poor understand caused some poorly annotated and poorly structured code. However, I followed Joel Lee’s tips for writing clearer code which I think helped. In the future, I will ask questions to fully understand the task to improve code legibility and therefore re-usability and access.

**Response to feedback**

The feedback I received from my discussion post in week two informed the development of my data management plan. In particular, I was encouraged to think a lot more widely in terms of data storage and also to explain my ideas more fully. The course discussion boards and webinars also provided a lot of space to improve coding practice and solve common issues. In particular, though group discussion, I was able to fix issues with my data dictionary.

**Skills developed**

Data management is an important part of my work with a research team. The team and I produce a large amount of analysis and utilise different datasets. However, often previous analysis is hard to find or understand, leading to work that has to be repeated. The FAIR guidance that I have learnt on this course, as well as the use of Github and how to structure folders and analysis will prove invaluable for my work going forward. I plan to use these developments to discuss with my colleagues how better to share work and make it more findable, accessible, interoperable, and reusable

1. O’Hanlon, D.S. (2017). Why data sharing matters for excellent care | NHS Confederation. [online] www.nhsconfed.org. Available at: <https://www.nhsconfed.org/articles/why-data-sharing-matters-excellent-care> [Accessed 3 Jul. 2022]. [↑](#endnote-ref-1)
2. Hutson, G., Jemmett, T., Mainey, C., Turner, Z. and community, NHS-R. (2021). NHSRdatasets: NHS and Healthcare-Related Data for Education and Training. [online] R-Packages. Available at: https://cran.r-project.org/web/packages/NHSRdatasets/index.html [Accessed 3 Jul. 2022]. [↑](#endnote-ref-2)
3. The University of Edinburgh. (2022). Research Data Management Policy. [online] Available at: https://www.ed.ac.uk/information-services/about/policies-and-regulations/research-data-policy [Accessed 3 Jul. 2022]. [↑](#endnote-ref-3)
4. GO FAIR. (2017). FAIR Principles - GO FAIR. [online] Available at: https://www.go-fair.org/fair-principles/ [Accessed 3 Jul. 2022]. [↑](#endnote-ref-4)
5. Sebastian-Coleman, L. (2013). Data Quality Metric - an overview | ScienceDirect Topics. [online] www.sciencedirect.com. Available at: https://www.sciencedirect.com/topics/computer-science/data-quality-metric [Accessed 3 Jul. 2022]. [↑](#endnote-ref-5)
6. Michener, W.K. (2015). Ten Simple Rules for Creating a Good Data Management Plan. *PLOS Computational Biology*, 11(10), p.e1004525. doi:10.1371/journal.pcbi.1004525. [↑](#endnote-ref-6)