Project Description Outline

Names of group members

Thijmen Breeschoten, Chiara Capresi, Naomi Penfold, Alistair Trucco

Roles & responsibilities of each member

Thijmen Breeschoten worked on:

- Initial data wrangling/cleaning of the A&E hospital activity datasets.
- Plot visualisations for A&E attendance KPI measurements.
- Data analyses, including hypothesis testing for summer/winter comparison per health board.
- Selecting insights and drawing conclusions for impact of seasonal effects on A&E attendances across Scotland.

Chiara Capresi worked on:

- Analysing and cleaning data about the impact of Covid in hospital admissions.
- Producing visualisations about hospital admissions' trend over time, from the beginning of covid emergency up to now.

Naomi Penfold worked on:

- Exploring data about inpatient stays, treatment wait time and A&E attendances
- Analysing and cleaning data, and producing visualisations, about hospital occupancy
- Initial R Shiny design and scripts: global, ui and server; drawing wireframes throughout project
- Coding consistent themes (for ggplot outputs and R Shiny dashboard)
- Writing documentation (github README)

Alistair Trucco worked on:

- translating nhs location data to be useable in leaflet
- Analysis and visualisations for delayed discharge data
- t-test workflow

Everyone worked on:

- General data exploration
- Presentation deck
- Dashboard R Shiny framework
- Collaborative discussions and planning
- Git repository management
- Project description outline
- Drawing final conclusions

Brief description of dashboard topic

The main purpose of our dashboard was to gain an understanding of how acute care provision changes by season (summer versus winter) and with the impact of the COVID-19 pandemic. We have focussed on comparing health boards to the national average, and understanding demographic groups that drive any patterns or are most affected by an effect.

The dashboard outlines our topic in terms of showing:

- Summer v winter A&E attendances
- Impact of COVID on the three metrics above, by showing before, during and after COVID on time series plots

Our dashboard consists of:

- One selector, which filters selected plots for individual health board data
- Four tabs:
    - First tab explores the seasonal impact on A&E attendances for each health board in Scotland. Four plots are presented (description from left-right and top-bottom):
        - i) Map, reactive to the health board selector, zooming in and showing the location of each A&E department within the health board of choice. By clicking on each location, the specialisation of the A&E department is given (such as 'Minor injury unit').
        - ii) Barplot, reactive to the health board selector, showing the total number of attendances at the A&E health board of choice split by season (summer and winter) per year. Only pre-covid years are included to exclude Covid-specific effects. The line gives the average number of attendances per season (including all four seasons of the year) per year.
        - iii) Barplot, reactive to the health board selector, showing the total number of A&E attendances split by deprivation level for most recent pre-covid years.
        - iv) Barplot, reactive to the health board selector, showing the total number of A&E attendances split by age group for most recent pre-covid years.
    - Three tabs looking at metrics over time from before COVID until now (2022/23) in order to visualise impact of COVID:
        - Hospital admissions: this tab shows:
            - (i) a not reactive heatmap of Scotland, showing the average of hospital admissions during the period mentioned above for each single hospital in the country.
            - (ii) a reactive line plot that shows, for any selected health board (and for the entire Scotland as well), the variation over time in the average of hospital admissions.
            - (iii) a third line plot, still reactive per health board, that highlights the differences in this trending three age's classes: under 5, 5 to 64, 65 and over.
        - Hospital occupancy: this tab shows:
            - (i) a heatmap of average inpatient occupancy for each hospital across Scotland, it is not reactive to the health board selector
            - (ii) a heatmap as above, that is reactive to health board selection, zooming in to only show hospitals in that health board
            - (iii) a time series plot of average occupancy from 2017 Q1 to 2022 Q4 for Scotland (a fixed blue line) and the selected health board (a purple line, which changes with the selector input)
        - Delayed discharges: this tab shows:
            - (i) a heatmap of mean change in bed-days by health board before and after covid, it is not reactive to the health board selector.
            - (ii) a time series plot of average delayed bed-days in a month from July 2016 to May 2023 for Scotland split by age (18-74, 75+, all ages) on the selected health board.

Stages of the project

- Data exploration to understand the available data, its quality, and what would be possible to analyse, including identifying meaningful KPIs for the client
- Communication with the client to understand key terms ("episode", "acute")
- Planning & dashboard wireframe – iteratively developed throughout

- Choosing datasets
- Git branching & version control
- Statistical analyses
- Visualisations to create plots: including time series, bar charts and maps (using Leaflet)
- Communicating key insights through presentation and on the dashboard
- Documentation of code and producing the README

Which tools were used in the project

- Zoom (daily stand-ups and occasional mob programming)
- Slack/ WhatsApp (communication)
- Excalidraw (for planning and ‘brainstorm’ sessions)
- Git/GitHub (collaboration & version control)
- Rstudio (data analysis and visualisation)
    - Packages including: tidyverse, janitor, lubridate, bslib, plotly, leaflet, sf and shiny
- RShiny (reactive dashboard)

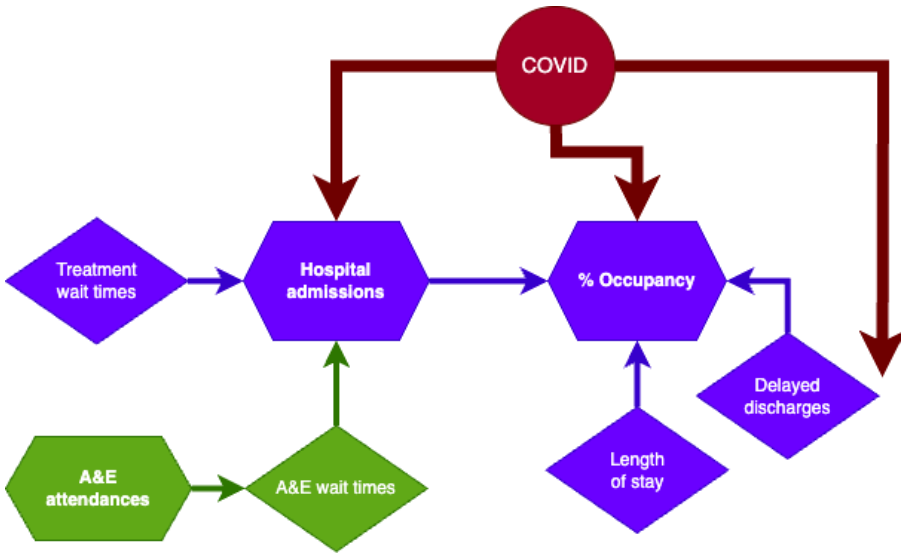
How did you gather and synthesise requirements for the project?

We synthesised the information given in the brief by our clients, representatives of Public Health Scotland. After internal discussions and agreeing on goals we further discussed requirements with our clients during a second discussion in order to confirm that goals were meeting client expectations and to clarify some domain-specific terminology.

We prioritised four key indicators to focus on, given the time limit and for the reasons described below (see next section).

Motivations for using the data you have chosen

From the data available, we first identified and selected KPIs of importance to the client, Public Health Scotland, as performance measurements. We were able to explore several key performance indicators (KPIs) of acute care provision, including client intake (A&E attendances and hospital admissions), measures of service workload within the service (wait times in A&E and for hospital treatment, length of stay, bed occupancy), and an outflow metric that affects service capacity and resources (delayed discharge).



Key indicators of acute care provision included in the available datasets.

KPI data were further analysed for impact of: A) seasonal fluctuations and influences, and B) covid effects. Those KPI's showing an influence were selected and further analysed.

(A) To investigate seasonality, we focussed on A&E attendances. First, because an impact of seasonality is expected at the A&E department due to seasonality-impacted injuries (e.g. outdoor activities, viral activity and transmission peaks). Second, because we found a seasonal pattern here (also in A&E wait times) whereas we did not find any seasonality in other indicators.

(B) To understand the impact of COVID-19 pandemic, we focused on three key indicators that cover the flow of clients into, within and out of hospital care: hospital admissions, bed occupancy, and delayed discharge. We could have looked at other indicators (in other datasets also mentioned in this documents) instead or as well, but chose to focus on these three in the interests of time and simplicity, while also covering different factors affecting service provision.

  (a) For the analysis of COVID-19 impact on hospital admissions, we used a dataset that contains information about the number of any kind of hospital admissions (not only admissions of people affected by Covid) divided per health board. This was because we were interested in highlighting the differences in the measure of this metric caused by the impact of Covid. This dataset contains also some demographic information about patients, like age groups and gender, which allowed us to make demographic considerations about age.

  (b) For the analysis of COVID-19 impact on occupancy (a measure of used hospital capacity, in terms of staffed-bed-days), we used a dataset that contains quarterly data on occupancy for each hospital location in each health board from Q1 2017 to Q4 2022. There was no demographic information available for these measures.

  (c) For the analysis of COVID-19 impact on delayed discharges, a dataset was used containing monthly records of delayed discharge days divided by health board and two age groups. The data ranges from July 2016 to May 2023. This was used for examining the difference before and after COVID and how it affected the rest of the healthcare pipeline.

Data quality and potential bias, including a brief summary of data cleaning and transformations

According to the About tab on the dataset page/dedicated page online, the data quality is not perfect, however we did not find any major issues that prevented analyses. In each dataset,

there were “qf” columns that highlighted any missing data or additional information. No information here led us to exclude any data points, we used all the data available.

The dataset may not be biased because the data has been collected by NHS services rather than being self-reported, therefore there is no self-selection bias. However, there may be bias introduced by the NHS processes and workflow, which we do not know about and have no control over.

To clean the datasets, we:

- Used janitor package to clean variable names (we prefer tidy format, snake_case)
- Checked for any NA values that may need excluding or recoding

We also wrangled datasets to produce the summary data need for visualisations, including:

- Joining KPI datasets with location information and coordinates, to use for heatmaps
- Summarising or averaging KPI values across specialties, age groups, or otherwise, in order to present a health board and national average. Where necessary, we excluded aggregated data within datasets, such as “All of Scotland” values or “All age groups” so that measures did not double-count individual episodes. Age classes were created depending on specific visualisation and ensuring we were providing informative insights for that indicator
- Recoding date information to POSIX format to be usable in time series plots
- Encoding summer and winter months, and before and after COVID time periods, in order to compare between these groups
- Recoded/prepared data to be included in statistical significance testing (seasonal effect on A&E attendances and covid effect on delayed discharges)

How is the data stored and structured

This project uses data from Public Health Scotland and NHS Scotland, which contains public sector information licensed under the Open Government Licence v3.0.

The Shiny dashboard presents data from the following specific datasets:

- Monthly A&E activity (including attendances and demographics data) and waiting times: https://www.opendata.nhs.scot/dataset/monthly-accident-and-emergency-activity-and-waiting-times
- COVID-19 Wider Impacts - Hospital Admissions: https://www.opendata.nhs.scot/dataset/covid-19-wider-impacts-hospital-admissions
- [for occupancy] Beds Information in Scotland: https://www.opendata.nhs.scot/dataset/hospital-beds-information
- Delayed Discharges in NHSScotland: https://www.opendata.nhs.scot/dataset/delayed-discharges-in-nhsscotland
- NHS Scotland Hospital Locations: https://www.opendata.nhs.scot/dataset/hospital-codes/resource/c698f450-eeed-41a0-88f7-c1e40a568acc

Exploration notebooks also include work using these additional datasets:

- Treatment wait times: https://www.opendata.nhs.scot/dataset/stage-of-treatment-waiting-times
- Inpatient and day cases activity (including length of stay):
    - Activity by Board of Treatment and Specialty: https://www.opendata.nhs.scot/dataset/inpatient-and-daycase-activity/resource/c3b4be64-5fb4-4a2f-af41-b0012f0a276a

    - Activity by Board of Treatment, Age and Sex: https://www.opendata.nhs.scot/dataset/inpatient-and-daycase-activity/resource/00c00ecc-b533-426e-a433-42d79bdea5d4
    - Activity by Board of Treatment and Deprivation: https://www.opendata.nhs.scot/dataset/inpatient-and-daycase-activity/resource/4fc640aa-bdd4-4fbe-805b-1da1c8ed6383

Following cleaning and wrangling, the data is in the form of cleaned csv documents within the github repository. This means that the shiny app source code files can run for anyone who has cloned the repository, because our shiny app scripts read in these cleaned csv files.

The raw data were linked datasets, in that they have matching information about health boards and also the variables within (e.g. sex). We retained the consistent naming of health boards and locations within our data, and used consistent column names, to enable us to join and compare across datasets.

Ethical and legal considerations of the data

We have followed these principles of ethical data science:

- **Responsible use of the data**, i.e. not publish misleading results from the data where possible, and clearly and reproducibly explain how results were obtained. We have published our project openly on Github, and through this anyone could reproduce our Shiny dashboard. However, we have provided context for the project (in the Github README), including stating that this project was not requested by and had no involvement from Public Health Scotland or NHS Scotland, and any results or insights are not intended to be used in real life.
- **No identifiable individuals** – the data we have used (and shared in the github repository) does not contain identifiable information about individuals: it is aggregated metrics about the service and we are not able to speak to any individual's diagnosis, treatment or service experience. We have shared data that has been derived from the data made openly available by NHS Scotland and Public Health Scotland (who already have their own data governance and control measures to prevent release of identifiable information).

The datasets are covered by the Open Government Licence v3.0, which means anyone is free to use and adapt the data as they please, so long as the data attribution statement is either included in the product or linked to. We have included the data source and licence links in the git repository README as well as in the Shiny dashboard.