MAC-MIGS CDT

PhD in Mathematics

Understanding the Public Health Waiting Times Landscape in Scotland:

Finding Key Drivers and Forecasting Demand

Authors:
Meritxell Brunet Guasch
Christopher A Oldnall
Sofie Verhees

Supervisors:
Dr. Ayse Arik
Dr. Timothy Cannings
Prof. George Strefaris

Industry Partners:
Ms Amy McKeon
Mr Kennneth Nicolson
Dr Filip Zmuda

March 21, 2022

Abstract

This report presents a data-driven analysis of cancer waiting times in Scotland. We analysed open data provided by Public Health Scotland with records of the numbers of patients that were referred and/or treated within 31 and 62 days in different regions. Our goal when analysing this data was twofold. First, we aimed to explore the effect of the COVID-19 pandemic on the number of cancer referrals and diagnoses. Secondly, we aimed to identify the important factors underlying variation in cancer referrals and diagnoses, such as region or cancer type. Most of our work relied on the application of Generalised Linear Models. Overall, this report shows that the pandemic resulted in a reduction of rates of referral and diagnosis of cancer in Scotland. Moreover, our analysis suggests that these are significantly determined by region, cancer type and age. However, the open source data provided by Public Health Scotland is not sufficient to draw conclusions on the cancer waiting times in Scotland.

For model files and assets please see:
`www.github.com/SofieBV/PHS-project`

Contents

Acronyms

BIC Bayesian information criterion. 12

GLM Generalised linear model. 9, 13

HBs Health boards. 4, 6

PHS Public Health Scotland. 4–6

Acknowledgements

With thanks to the supervisory team, **Dr Ayse Arik**, **Dr Timothy Canninngs** and **Professor George Streftaris** for their mathematical insight and guidance. Furthermore thank you to Public Health Scotland, in particular **Ms Amy McKeon**, **Mr Kennneth Nicolson** and **Dr Filip Zmuda**, for their input and vital expert knowledge of the public health landscape.

Author Contributions

Meritxell Brunet Guasch
xell.brunetguasch@ed.ac.uk

Throughout the project Xell's role was to initially look into change point detection and how it may be utilised. As the project progressed her role was to look more at the implementation of the general linearised models within the team and in particular take a lead on how these could be applied to the diagnosis data set.

Christopher A Oldnall
chris.oldnall@ed.ac.uk

Chris' role within the project at the beginning was to consider how the data looked, and brainstorm potential techniques that could be employed. Past this, alongside assisting with the initial model implementation, he focused on the structure of the report and took the lead a lead on the project copywriting.

Sofie Verhees
s.b.verhees@sms.ed.ac.uk

Beginning with creating analogous figures to those of PHS, Sofie's role was to analyse the data and ensure that any anomalies were explored. Following through from this Sofie took a lead on the programming side of the project, constructing the generalised linear models pertinent to application on the waiting times data.

Part I
Introduction

> "Medicine is a science of uncertainty and an art of probability."
>
> William Osler

Health, illness and disease have concerned humans since their origins, with earliest medical traditions dating back to Ancient Greece. Approaching the end of the 19th-century and the rapid growth of cities, the need for systematic sanitary measures led to the beginning of public health measures, as well as the documentation of developed medicines, methodology and treatments [10]. The ability to document medical practice and medical records progressed through the 20th-century, with larger studies commencing. This was the start of the "big-data" age of medicine. By 1952 accurately recorded super-studies, such as Doll et. al's study on the smoking habits of male doctors [4], were able to be carried out. In Doll's case, the work would later go on to prove statistically significant evidence that smoking was a cause of lung cancer (as early as 1956). Similar studies pointed the potential of recording and analysing medical data to improve treatment strategies and public health measures.

Performing systematic analysis of public health data is important to understand the state of the system, inform policy making and forecast future demand. Public Health Scotland (PHS), the organization leading health data management in Scotland, has recognized the potential stored in the data they collect. Consequently, they aim to collaborate with academics and students from both The University of Edinburgh and Heriot Watt University to analyse their data sets. This project is the first result of this collaboration, and aims to provide a precedent for future work.

Waiting Times, Cancer and COVID-19

An important measure of public health institutions are waiting times, i.e. length of time between diagnosis, referral and start and end of treatment. Data on waiting times is useful to health authorities and governments in terms of setting standards and targets, as well as keeping record of the saturation of health services [19].

Waiting times are particularly relevant in cancer, a disease in which prognosis and probability of survival is strongly dependent on an early diagnosis and subsequent treatment. Research in England has shown that cancer waiting times delayed by two months can cause up to a loss of 0.7 life-years for a referred patient [20]. The importance of reducing waiting times for cancer patients led to the introduction of a 31-day waiting standard between referral and start of treatment in the UK [8]. In this report, we aim to provide insight on the cancer waiting time landscape in Scotland by analysing PHS data on the meeting of waiting times standards.

In 2020, public healthcare systems were strongly compromised by the COVID-19 pandemic, a disease caused by the virus SARS-CoV-2 that has resulted in over 450 million infections and 6 million deaths worlwide [12]. In the UK, the pandemic has had profound effects on the National Health Service, which had to accommodate unprecedented numbers of patients due to COVID-19. This resulted in a decrease in patients treated for other conditions, as well as an increase on the waiting time for non urgent treatment [24]. Studies showed that, during COVID-19, less patients with acute coronary syndrome were admitted to the hospital in England [6]. Similarly, COVID-19 caused a decrease in number of patients referred, diagnosed and treated for colorectal cancer in England [7]. Conversely, in the context of cancer waiting times in Scotland, PHS reported that the standards were met in the same proportion of cases throughout the pandemic (see Figure 1). However, no studies were performed addressing the absolute number of patients referred for such standards.

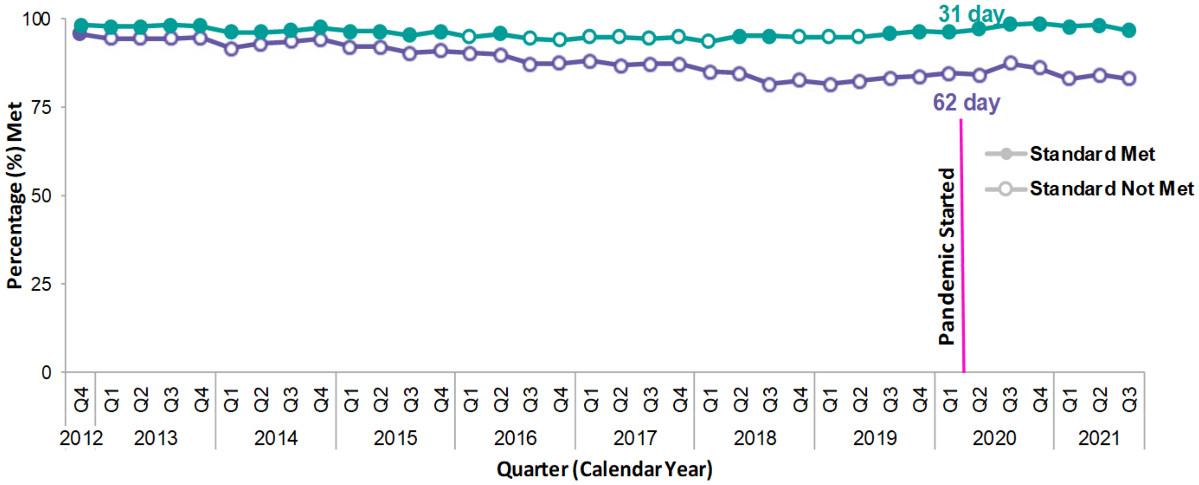

Figure 1: NHS Scotland performance against the 31 and 62 day standards [16].

Research Questions

This report presents a statistical analysis of waiting times data provided by PHS. Our first goal is to investigate how COVID-19 affected the number of patients who were being referred for cancer treatment. Form here comes our first research question.

Research Question 1:
How has COVID-19 affected the number of eligible referrals of cancer patients under the 31 and 62 day standards in Scotland?

Moreover, we aim to use additional information stored in the data set (e.g. cancer type and region of referral) to better understand the factors underlying variation in the number of referrals for the cancer waiting times standards. Hence, we construct our second research question as follows.

Research Question 2:
What factors contribute to the variation of referral numbers for cancer treatment under the standards in Scotland?

Finally, we apply the same methodology developed to answer the previous questions to a different data set containing numbers of cancer diagnosis. This aims to answer the following question.

Research Question 3:
What is the impact of COVID-19 and other drivers on the number of cancer diagnoses in comparison to the number of individuals being referred for treatment under the 31 and 62 day standard in Scotland?

For ease of exploration of this report, we note its structure:

- Part II explores the background of PHS, the data and methodologies to be employed,
- Part III presents the data analysis results and how these relate to our research questions,
- Part IV discusses the results obtained as well as directions for future work and practices.

Part II
Background and Methodology

> "Mathematics consists in proving the most obvious thing in the least obvious way."
>
> George Polya

1 Data

PHS is the organisation associated with the Scottish government and NHS Scotland, who "have access to and collaborate on an enormous range of data both on Scotland's health and wellbeing, and on health and social care services. This includes a wealth of data and intelligence vital to helping people access quality services, like the cancer services data" [14]. Some of the data sets which are of interest to the public and researches are placed onto an open data platform whereby anybody can access them. Throughout this report we will only look at open source data, which contain a small subset of the full data that is overseen by PHS. A summary of the data sets which we have identified as being useful particularly for the purposes of Part III can be found in Appendix B.

In Scotland, health care services are divided into 14 regions, which are referred to as Health boards (HBs). Consequently, most of the data sets generated by PHS classify the data by health boards in which the patients are referred and treated. Figure 2 shows how these health boards are distributed on the Scottish map. These HBs can again be divided for cancer data into three super-regions: WOSCAN, NCA, and SCAN. WOSCAN embraces the four NHS boards in the West of Scotland: Ayrshire & Arran, Forth Valley, Greater Glasgow & Clyde and Lanarkshire. SCAN is the South East Scotland Cancer Network bringing together the NHS boards: Borders, Dumfries & Galloway, Fife, and Lothian. Finally NCA in the north of Scotland encompasses the boards: Grampian, Highland, Orkney, Shetland, Tayside, and Western Isles. HBs differ between them in terms of demographic characteristics. When comparing data of different HBs, it will become important to normalise by their population size, for which we use estimates provided by PHS (See Appendix B).

1.1 Cancer Waiting Times: The 31 and 62 day Standards

From the range of waiting times data collected by PHS, which ranges from drug and alcohol treatment to operations we focus on cancer waiting times in this report, for which two open data sets are available. Each of these data sets contains the numbers of patients referred and treated within the 31 and 62 days standards in Scotland, with the following additional information

- Year and quarter of referral (from 2012 to 2021)
- Cancer Type (see Table 1)
- Health Board in which the patient was referred (see Figure 2)

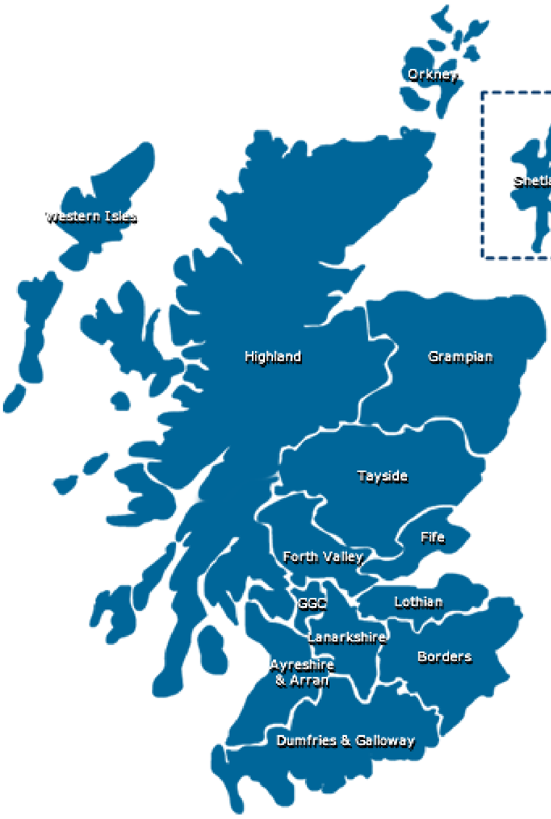

Figure 2: Scotland Health Board Areas as of 2022. Figure obtained from [9].

In Scotland, this data has been reported quarterly since 2012. These two aforementioned standards are defined below.

Definition 1: 31 and 62 Day Standard

We state that the two standards, as defined by the Scottish Government [18], are

- 31-day target from decision to treat until first treatment, regardless of the route of referral.
- 62-day target from urgent referral with suspicion of cancer, including referrals from national cancer screening programmes, until first treatment.

In our first analysis of the data set, we realised that there was a mismatch between patients eligible for treatment and treated patients in different regions. This led to us beginning to analyse the data by eye in addition to sense checks. Following our observations, an update to the data was issued by PHS to ensure it was aligned with that being used to create their reports. From this we were then able to recreate some figures, similar to those by PHS, surrounding the numbers of refereed and treated patients under the standards such as Figure 3. Even though the ratio of referred and treated patients stays mostly the same over the years, as figure 1 also suggests, you can see a drop in both the number of referrals and treated at the start of 2020.

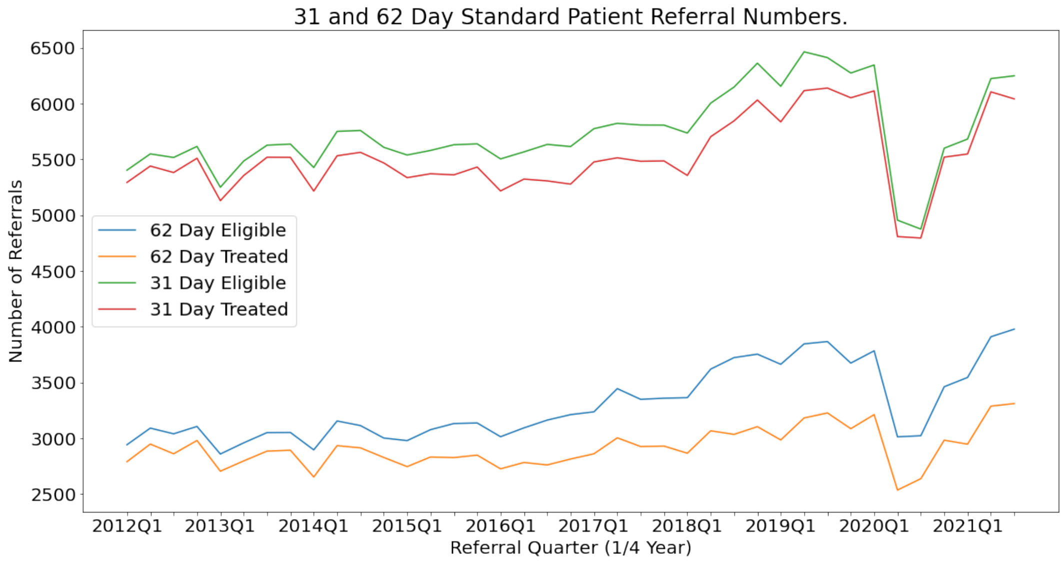

Figure 3: A plot which shows the time evolution of patients eligible and treated within the 31 and 62 day standards from 2012 onwards.

1.2 Cancer Diagnosis

In order to compare and contrast our results, we decided to perform a parallel analysis using an alternative data set. We analyse a data set containing weekly data on cancer diagnosis across different HBs from 2019 to 2021. This is also open source data provided by PHS [17]. This data is richer than the waiting times data in that it contains enhanced demographic information of patients diagnosed. In particular, the data set includes counts of the number of cancer diagnosis in Scotland, with the following additional information:

- Year, month and week of diagnosis (from 2019 to 2021)
- Cancer Type (see Table 1)
- Health Board in which the patient was diagnosed (see Figure 2)
- Sex of the patient (male or female)
- Age Group (0-49, 50-69, 70+)

As we can see in Table 1, the cancer diagnosis data considers more cancer types and a finer classification than the referrals and treated data-set.

Moreover, Figure 4 shows the number of patients diagnosed with cancer is higher than the number of patients referred for treatment for both the 31 and 62 day standard, even if we only consider the cancer types that appear on both data sets. This is expected, because a set eligibility criteria for the standards excludes some cancers for a plethora of reasons, such as if the patient had a clinically complex pathway to their treatment [15]. Consequently, there is a considerable fraction of cancer patients that do not meet the criteria to be eligible to meet the waiting times standards, which compromises our ability to draw conclusions about cancer waiting times form the 31 and 62 day standard data.

Referrals/Treated Cancer Types	Total Diagnosis Cancer Types
	All Malignant Neoplasms
	Bladder
	Bonde and connective tissue
	Brain
Breast	Breast
Cervical	Cervical
Colorectal	Colorectal
Head and neck	Head and neck
Lymphoma	Hodgkin Lymphoma
	Non-Hodgkin lymphoma
	Kidney
	Leukaemias
	Non-Melanoma Skin Cancer
Melanoma	Malignant Melanoma of the skin
	Mesothelioma
	Multiple Myeloma and malignant plasma cell Neoplasms
Upper GI	Liver and intrahepatic Bile Ducts
	Oesophagus
	Pancreas
	Stomach
Ovary	Ovary
Urological	Penis
	Prostate
Lung	Trachea, Bronchus, and Lung
	Testis
	Thyroid
	Uterus
	Vagina
	Vulva

Table 1: Comparison of cancer categories for referral and diagnosis data. Cancer types that appear in both data sets are highlighted. See Appendix B for category sources.

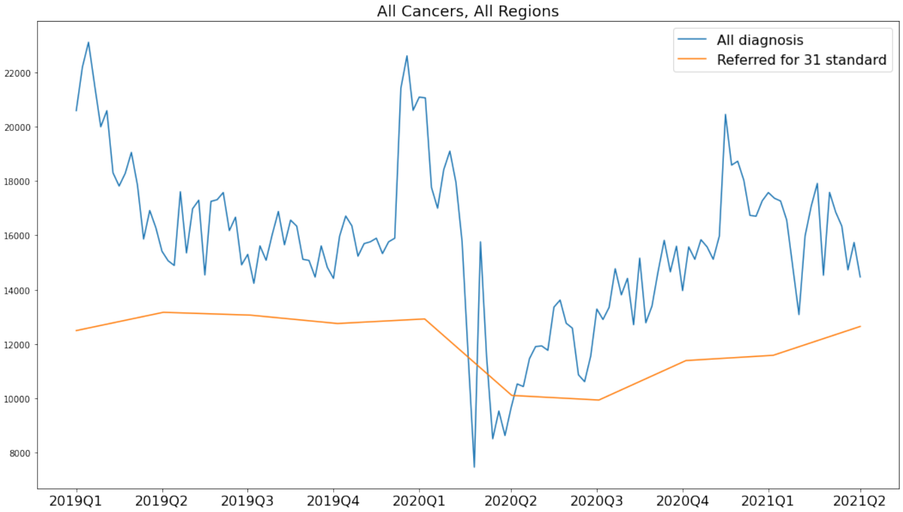

Figure 4: Number of patients diagnosed with cancer (blue), referred for the 31 day standard (orange) and treated within the 31 day standard (green) from 2019 to 2021. Data aggregated for all HBs in Scotland and all cancer types that are both in the diagnosis and the referral data set (see Table 1). See Appendix B for category sources.

2 Generalised Linear Models

The three data sets analysed in this report (31 day referrals, 62 day referrals, diagnoses), contain different variables, such as region and cancer type, that might affect the number of patients referred/treated/diagnosed. The GLM provides a useful framework for comparing how several explanatory, or *independent variables*, affect the outcome of *dependent variables*. In GLMs, each outcome of the dependent variables is assumed to be generated from a particular distribution of the exponential family. Some distributions belonging to the exponential family are: normal, binomial, poisson and gamma [2].

2.1 Basic Form

Let us firstly explore a formal definition of a GLM before we start to proceed with how they can be amended for other terms to be included.

Definition 2: Generalised Linear Model

A generalised linear model (GLM) consists of the following three components:

1. A (common) distribution from the exponential family for the independent response variables $Y_1, ..., Y_n$.
2. A linear predictor
$$\eta_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}.$$
3. A differentiable, strictly monotone link function such that,
$$g(\mu_i) = \eta_i,$$
where $\mu_i = \mathbb{E}[Y_i]$

As we started to explore the cancer referral and diagnosis datasets, we decided that it would be sensible to assume that the distribution was Poisson as had been assumed in similar research [1]. This was because of the following reasons:

- We are dealing with 'rate' data - that is to say we need to have a positive parameter for the model, since we are dealing with numbers of individuals entering a waiting times system.
- The data is highly varied, especially when analysing specific cancer types separately, as shown in section 4.3. Therefore having a model with a mean equal to its variance allows for this large variation in the data [11].

Let us have a look at the form of a Poisson GLM for the number of referred cancer patients.

Definition 3: Poisson GLM

Let Y_t be a random variable representing the number of referred patients for cancer treatment in a certain quarter, t, with mean θ_t. Then if we assume

$$Y_t \sim \text{Poisson}(\theta_t),$$

we have the following model,

$$\log(\theta_t) = \beta_0 + \beta_1 t.$$

Note that we can also include more β coefficients to model the effect of other covariates.

2.2 Normalisation by Population Size

We now consider the case in which the expected number of counts θ depends on a categorical variable representing the region of referral:

$$\theta_{t,r} = \beta_0 + \beta_1 + \beta_{2,r},$$

where the coefficient $\beta_{2,r}$ takes a different value for every possible region r. When applying GLMs to population data from different categories or regions r, it is important to normalise by population size in order to obtain comparable results for each region coefficient. To do so, instead of considering number of occurrences, we consider rates of occurrences. In other words, we divide the parameter θ by population size:

$$\frac{\theta_{t,r}}{E_{t,r}} = \beta_0 + \beta_1 + \beta_{2,r},$$

where $E_{t,r}$ denotes the population size of each region r at a given time t. Thus, we obtain the following model

$$\log\left(\frac{\theta_{t,r}}{E_{t,r}}\right) = \beta_0 + \beta_1 + \beta_{2,r}.$$

Using the properties of logarithms, the left-hand side can be rearranged as follows

$$\log\left(\frac{\theta_{t,r}}{E_{t,r}}\right) = \log(\theta_{t,r}) - \log(E_{t,r}).$$

The term $\log(E_{t,r})$ is known as the ‘offset’ term. Combining the results above, we obtain the following form of the GLM with a normalisation by population.

Definition 4: Poisson GLM Normalised by Population Size

Let $Y_{t,r}$ be the number of cancer referrals, now also depending on ‘r’, the categorical variable region. Then given that we have $Y_{t,r} \sim \text{Poisson}(\theta_{t,r})$, we have that the GLM normalised by population is given by,

$$\log(\theta_{t,r}) = \log(E_{t,r}) + \beta_0 + \beta_1 t + \beta_{2,r},$$

where $E_{t,r}$ represents the population of region r at time t.

2.3 Time Alterations

Another consideration that we will look to implement in the next part, will be that of considering the functional form of a time dependence within the GLM. As shown in figure 3, a linear trend might not suffice as the data seems to have a slight curve. Therefore we could start to include some other form for the time variable. An example where we choose a second order polynomial function, which incorporates t^2, is given below.

Example 2.1: t^2 Poisson GLM

Given a quadratic relationship in time for a Poisson GLM with $Y_t \sim \text{Poisson}(\theta_t)$ the number of referred cancer patients, then we give the relationship as,

$$\log(\theta_t) = \beta_0 + \beta_1 t + \beta_2 t^2.$$

2.4 Inclusion of Change Point Indicator

Consider next that we might end up trying to model a set of time series data which has a change point, such as the time when the pandemic started. Whilst in section 3 we will explore how one might start to identify a change point within the data, we need to consider how we could do this using a GLM and what terms we would incorporate for this change. Let this coefficient be γ_0. As we want this coefficient to come into effect after a certain time ϵ, we multiply it by the indicator function $\mathscr{I}(t-\epsilon)$ defined as

$$\mathscr{I}(t-\epsilon) = \begin{cases} 0 & \text{if} \quad t \leq \epsilon, \\ 1 & \text{if} \quad t > \epsilon. \end{cases}$$

To explore the effect of the pandemic on the period of time after the pandemic started, we need to know the effect of the coefficient on the time trend and therefore the coefficient is also multiplied by $(t-\epsilon)$. This gives the following model:

Definition 5: Change Point Impacted Poisson GLM

Given a change point in a time series Y_t, where $Y_t \sim \text{Poisson}(\theta_t)$ is the number of cancer referrals, then we give the following generalised linear model,

$$\log(\theta_t) = \beta_0 + \beta_1 t + \gamma_0 (t-\epsilon)\mathscr{I}(t-\epsilon).$$

2.5 Inclusion of Recovery Period

Finally, we can include a more complex term for the impact of the pandemic. From figure 3, it seems reasonable to assume that, following the impact of COVID-19, there was an immediate recovery period afterwards. To model the initial drop in Y_t, we use the term $\gamma_0 \mathscr{I}(t-\epsilon)$, where the indicator function guarantees that this occurs at the start of the pandemic at time ϵ. To model the recovery, we need a coefficient γ_1 taking affect after $t = \epsilon$, which depends on the time trend. This gives the expression,

$$\gamma_1 (t-\epsilon)\mathscr{I}(t-\epsilon).$$

Hence we can defined the model as follows:

Definition 6: Recovery Included Poisson GLM

Given a change point followed by an immediate recovery in a time series Y_t, where $Y_t \sim \text{Poisson}(\theta_t)$ is the number of cancer referrals, then we give the following generalised linear model,

$$\log(\theta_t) = \beta_0 + \beta_1 t + \gamma_0 \mathscr{I}(t-\epsilon) + \gamma_1 (t-\epsilon)\mathscr{I}(t-\epsilon).$$

2.6 Model Selection Criteria

Different models might perform differently in terms of being able to describe past behaviour of data or being able to predict what may happen in the future. The way in which we have identified as being an efficient way of choosing which model is performing best is that of the Bayesian information criterion, as defined below.

Definition 7: BIC

Given a fitted model for some response data, then we may calculate,

$$\text{BIC} = -2 \times \text{maximised log-likelihood} + p \times \log(n),$$

where p are the number of independently fitted parameters in the model and n are the number of observations we are fitting to [5].

BIC is not a metric to evaluate a model itself, but allows us to compare models against another. In particular, the BIC provides a good balance between goodness of fit and model complexity. In the absence of other reasons for choosing a model—for example subject matter information about the relevance of some covariates—we can choose the model with the smallest BIC value.

When comparing GLMs with different combinations of explanatory variables, we can use BIC to choose between them by employing iterative variable selection methods. We use a forward stepwise selection procedure [3], which works as follows:

- Fit the model with just one covariate, for each covariate available.
- Add to the model the covariate that gives the best goodness of fit score, provided the score is better than without the covariate or a level α of significance is attained.
- For all the remaining covariates, repeat the process.
- Variables selected at a step may be removed in a later step should the model be improved without it by adding in an additional covariate.

Part III
Data Analysis

> "Data are just summaries of thousands of stories—tell a few of those stories to help make the data meaningful."
>
> Dan Heath

In this section, we implement the GLMs presented in Part II, in addition to discussing other techniques which can be used to answer our research questions. This will be done by utilising the data sets described in Section 1.

3 The impact of COVID-19

Recall our first research question concerning the impact of COVID-19 on cancer waiting times.

Research Question 1:
How has COVID-19 affected the number of eligible referrals of cancer patients under the 31 and 62 day standards in Scotland?

All data sets considered are time series data, that is, sequences of observations taken at successive points in time. Therefore, it is reasonable to analyse whether or not and when the pandemic has impacted cancer waiting times by determining the existence and location of a *change point* in the time series data.

Definition 8: Change Point in Time Series Data

Given a finite time series as an ordered sequence of observations $(x_1, x_2, \ldots, x_T)$, consider a subset of the time series $x_{a:b} = (x_a, x_{a+1}, \ldots, x_b)$ and denote its joint distribution as $p(x_{a:b})$.

If for a chosen time τ,

$$p(x_{1:\tau}) \neq p(x_{\tau+1:T}),$$

then we say that a change point has occurred at the time τ.

In the context of Research Question 1, we are interested in associating the position of a change point in the data under study with the COVID-19 pandemic.

3.1 Change Point Detection Algorithms

A common approach to identify change points in time-series data are change point detection algorithms [23]. These find the time point(s) in which change point(s) occur. An advantage of these methods is that they do not require *a priori* knowledge of the location of the change point. Our goal was to use change point detection algorithms to identify the location of change points and assess whether or not they matched with the beginning of the pandemic.

There exists a large library of algorithms that perform change point detection on time series data [23]. In this project, we employed the package *ruptures*, which implements a variety of change point detection methods in Python [22]. We choose four methods: pelt, binary segmentation, window based search and dynamic programming. These are all offline methods, meaning they receive and process the entire data set at the same time, considering changes in the whole time series. All methods detect change points through minimising a cost function over possible numbers and locations of change-points. However, the definition of the cost function and the computational approach to minimise the cost function differs between methods, resulting in different outputs. Thus, it is common practice to employ several methods.

Figure 5 shows the results obtained using the previously mentioned algorithms to the 31 days referrals data. First, we notice that there is indeed an observable drop in the number of treated patients coinciding with the start of the pandemic (indicated by the red section). However, there is discrepancy between the change points detected by different algorithms, with most not finding a change point at the beginning of the pandemic. A potential reason for this is that change point detection aims to place changes on points that better explain the overall time series, rather than explaining punctual events such as a sudden drop. Moreover, the data is noisy, with oscillations caused by several factors that dilute the strength of the signal of the drop in the first quarter of 2020. Different methods finding different change points suggests that this is not the best approach to identify the impact of COVID-19 on our time-series data.

3.2 GLMs for Pandemic Impact

Our initial analysis using change point detection algorithms was unable to provide clear insight on the impact of the pandemic on cancer waiting times data. Consequently, we consider an alternative strategy to identify change points relying on GLMs. We fit GLMs to our data with and without including terms representing the pandemic. Comparison of the different models allows us to determine if the effect of COVID-19 was significant. Instead of blindly looking for change points as in change point detection, we choose ϵ to be the first quarter of 2020 ('2020Q1' in the data). This change point location was based on the fact that the first COVID-19 cases in Scotland were recorded in this quarter. Further exploration of different values for ϵ, as can be found in section 3.3, also indicate '2020Q1' to be the best choice.

In order to compare different GLMs, we use forward stepwise model selection as explained in section 2.6. The first step is to fit a basic GLM and then to add different more complex terms. From the different possibilities of terms to include in the model, each explained in 2.2, we chose to always include the normalisation by population size. Then there were three main models of interest: the basic GLM, the inclusion of the pandemic, and the inclusion of the recovery period. Trying all three versions with a linear time trend and a quadratic time trend, creates the large variation of models given below.

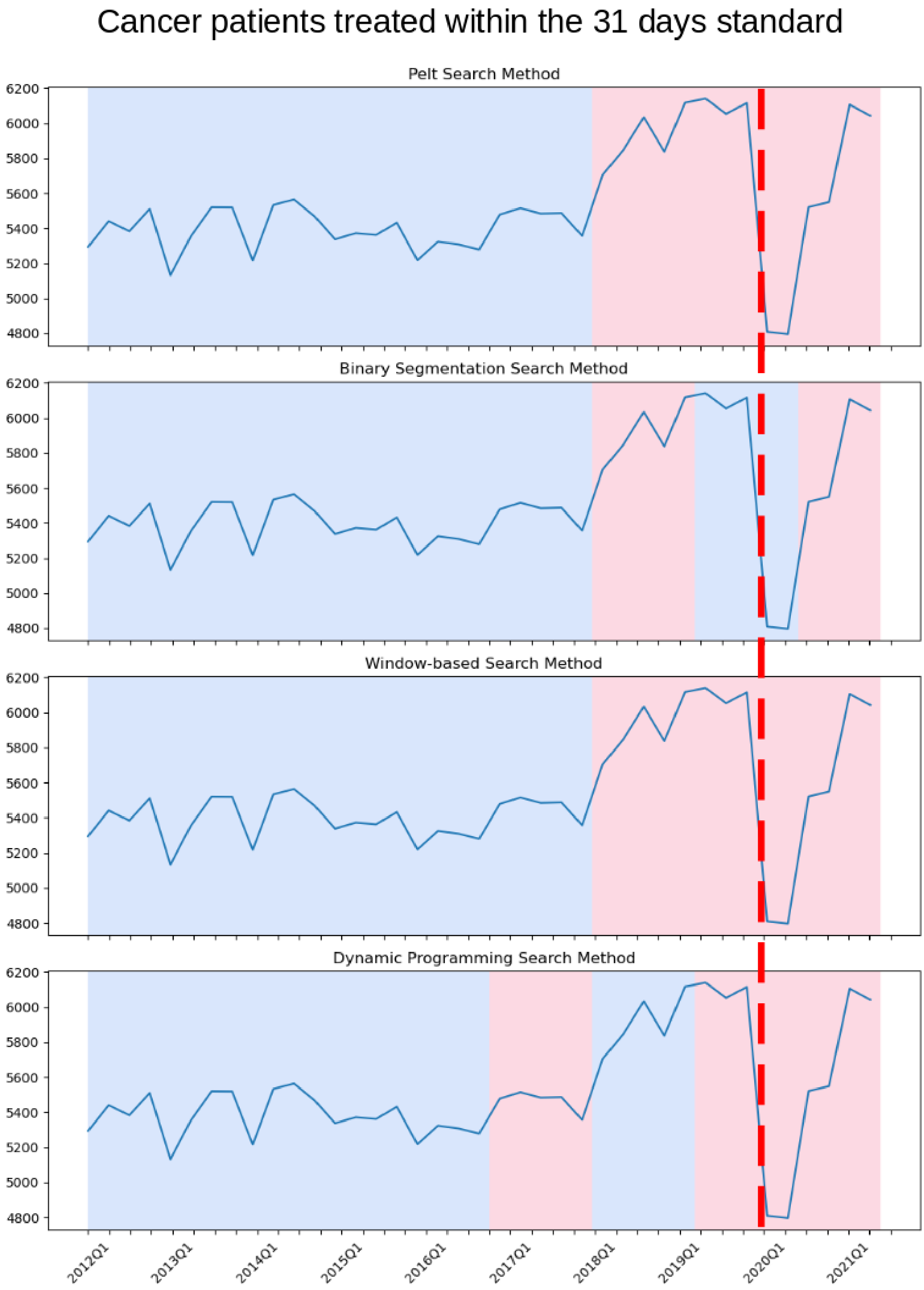

Figure 5: Results obtained by implementing different change point detection algorithms to the 31-day data. These are implemented in Python using the package *ruptures* [22].

model 3.2.1: $\log(\theta_t) = \log(E_t) + \beta_0 + \beta_1 t,$

model 3.2.2: $\log(\theta_t) = \log(E_t) + \beta_0 + \beta_1 t + \beta_2 t^2,$

model 3.2.3: $\log(\theta_t) = \log(E_t) + \beta_0 + \beta_1 t + \gamma_0 (t-\epsilon)\mathscr{I}(t-\epsilon),$

model 3.2.4: $\log(\theta_t) = \log(E_t) + \beta_0 + \beta_1 t + \beta_2 t^2 + \gamma_0 (t-\epsilon)\mathscr{I}(t-\epsilon),$

model 3.2.5: $\log(\theta_t) = \log(E_t) + \beta_0 + \beta_1 t + \gamma_0 \mathscr{I}(t-\epsilon) + \gamma_1 (t-\epsilon)\mathscr{I}(t-\epsilon),$

model 3.2.6: $\log(\theta_t) = \log(E_t) + \beta_0 + \beta_1 t + \beta_2 t^2 + \gamma_0 \mathscr{I}(t-\epsilon) + \gamma_0 (t-\epsilon)\mathscr{I}(t-\epsilon).$

These six models were fitted on both the 31 day referral data and the 62 day referral data aggregated for all cancer types and regions. The coefficients with confidence intervals and BIC for all six models and two datasets can be found in appendix A. The actual data plotted in comparison to the fitted models with significant coefficients for the 31 day standard are given by figure 6. For all models except for model 3.2.4, all coefficients were significant. Model 3.2.4 has an insignificant coefficient β_2, which is why this model has been discarded. Next to that, the BIC decreases if the complexity of the model increasing, making model 6 the model best fitting the data.

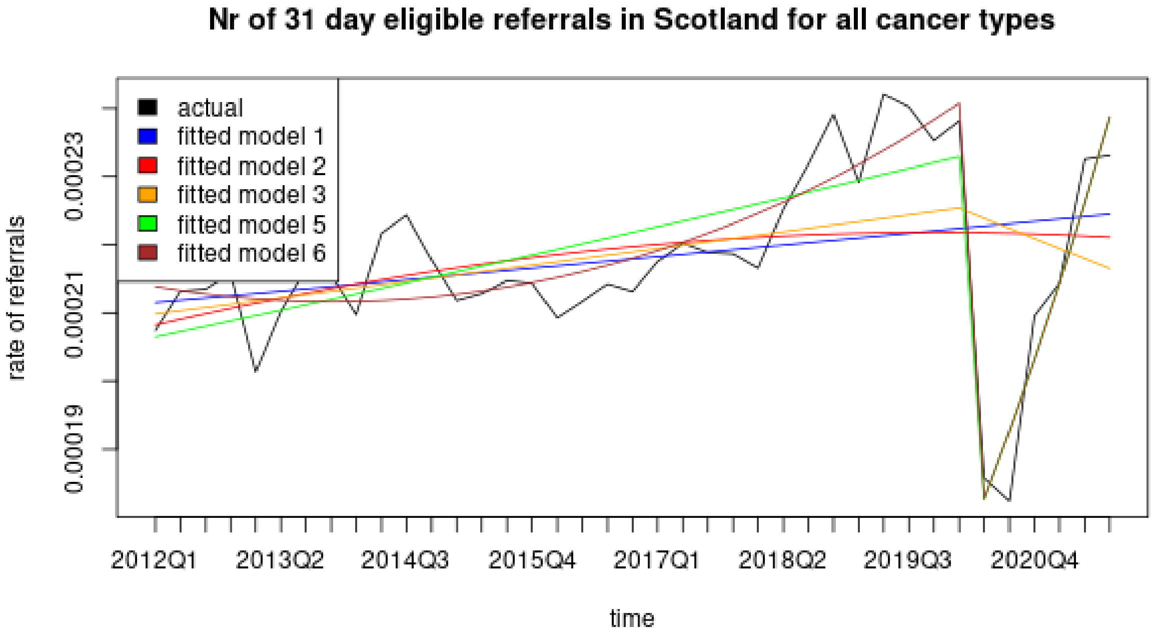

Figure 6: Number of referrals eligible for 31 day standard and corresponding fitted models, where model 3.2.1 as defined in 2.2, model 3.2.2 as defined in 2.3, model 3.2.3 in 2.4, and model 3.2.4 in 2.5.

As shown in the figure, the first model, given by the blue line, indeed seems to fit the mean of the rate of referrals for the 31 day standard. There seems to be a slightly upward trend, suggesting the number of referrals increase over time. Second, we added a nonlinear time-term for model 2, given by the red line. There is a slight curve in this model, but it looks mostly similar to the linear model in time. The result of model 3 is given by the orange line in figure 6. The significant coefficient $\gamma_0 = -0.008914$, suggests that the pandemic had a decreasing effect on the time trend on the rate of referrals for 31 day standard. The figure shows this decrease clearly. However, γ_0 is quite small, which could be because it bundles both the pandemic and recovery time together. This would mean it is not able to model the drop as clearly as model 3.2.5 or 3.2.6.

Models 3.2.5 and 3.2.6 give drastically different fitted lines because of the clear drop at time ϵ. This drop models the data more accurately, which can also be seen in the BIC as this has more than halved (a much bigger decrease than between other models). On the other hand, it can be questioned if this model would be useful for predictions. A discussion around this point can be found in Part IV.

The actual data plotted in comparison to the fitted models with significant coefficients for the 62 day standard are given by figure 7. Model 3.2.1 and 3.2.2 both have significant coefficients, with model 3.2.2 having a smaller BIC. This is similar to the models for the 31 day referral data. Model 3.2.3 is again insignificant, suggesting that the relationship with time is not quadratic. Surprisingly, model 3.2.4 has coefficient γ_0 that is insignificant, suggesting that the pandemic did not have a significant influence on the 62 day referral data. Continuing to model 3.2.5 and 3.2.6, the coefficients equivalent to the impact of the pandemic are significant. Thus it is necessary to split the recovery period from the drop in rate of referrals for the model to find a significant change due to the pandemic.

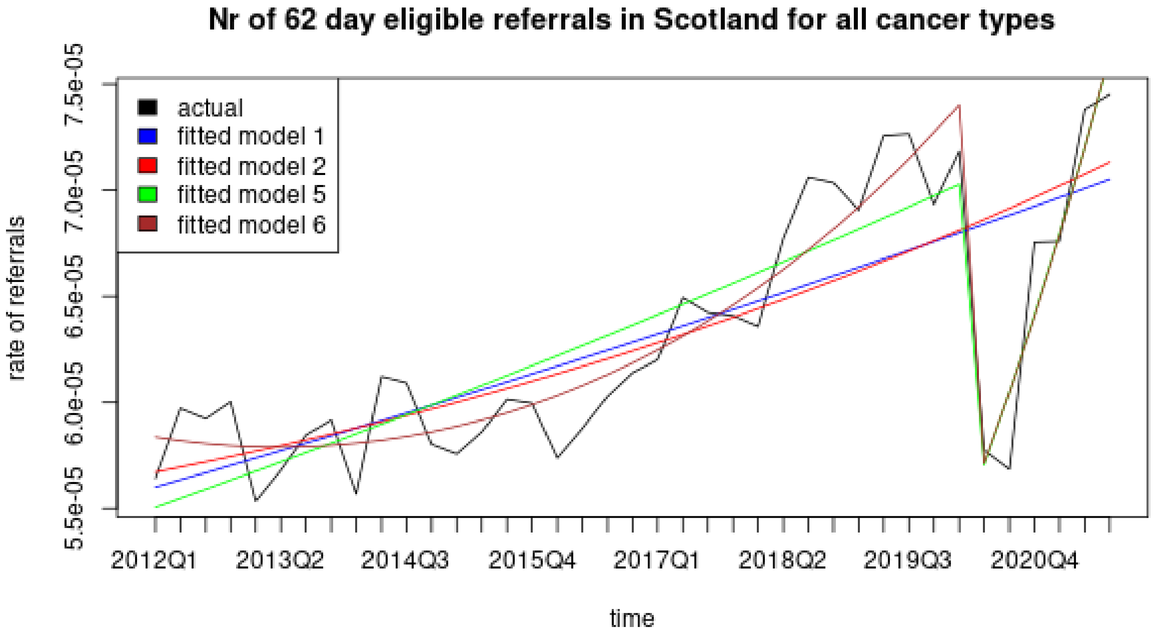

Figure 7: Number of referrals eligible for 62 day standard and corresponding fitted models, where model 1 as defined in 2.2, model 2 as defined in 2.3, model 3 in 2.4, and model 4 in 2.5.

All coefficients that model the effect of COVID-19 are shown in table 2. This shows that there is a significant drop of rate of referrals at the point of the pandemic, and a significant positive influence on the time trend after the pandemic once recovery has set in.

	coefficient	31 day referrals	62 day referrals
model 3.2.3	γ_0	-0.008914	-0.001360
model 3.2.5	γ_0	-0.2965	-0.2667
	γ_1	0.04967	0.05060
model 3.2.6	γ_0	-0.3275	-0.3154
	γ_1	0.04165	0.03766

Table 2: Pandemic coefficients of fitted models for 31 day and 62 day referrals, where the highlighted coefficient has insignificant p-value 0.551, while all other coefficients are significant with p-values smaller than $2e-12$. γ_0 gives the effect of the pandemic on the time trend in model 3.2.3 and the size of the drop in model 3.2.4 and 3.2.5, while γ_1 represents the effect of the recovery on the time trend.

3.3 Optimisation of ϵ

Even though the first recorded cases of COVID-19 were in the first quarter of 2020, it could be that the change point in the number of cancer referrals happened months later. In the above section we decided via intuition to set ϵ to 2020 Quarter 1. In order to determine this was a sensible choice, we explored how different values for ϵ change the results and determine which hypothetical change point location better explains the data. Picking from above both model 3.2.3 and model 3.2.5 — that is the models where we have only linear time periods and an inclusion of a pandemic term — we then ran the model with a variety of choices of ϵ ranging from the last two quarters of 2019 through till the end of 2020. Firstly the result of this simulation for model 3.2.3 can be seen in figure 8.

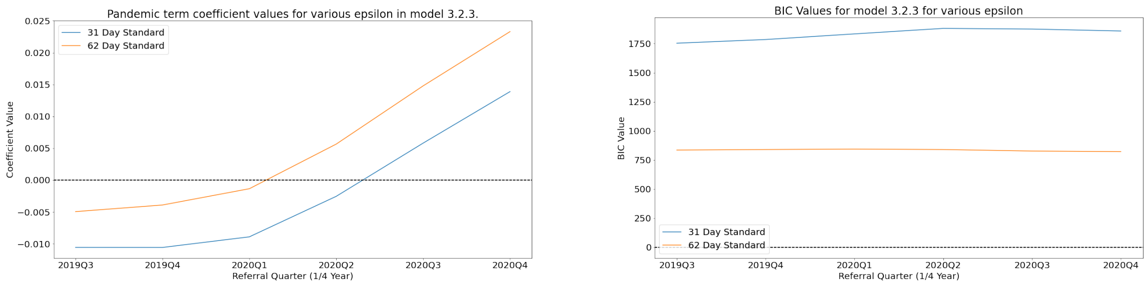

Figure 8: Values of the pandemic inclusion term coefficient for a range of selected epsilon for model 3.2.3.

Figure 8 would suggest that we have made a sensible choice in picking the first quarter of 2020 in both models. Whilst the figure shows one should pick earlier an earlier quarter, we know it would not make sense to choose a date in 2019 to represent the pandemic. A potential reason as to why we might be seeing a lower coefficient value in a pre-pandemic time within model 3.2.3 is due to our pandemic term being an additional coefficient to the time-trend coefficient which already exists in the model. Should one choose to normalise by the time-trend coefficient, potentially a more firm indication of 2020Q1 being the change point could be obtained for model 3.2.3.

Furthermore we see that the value of the BIC is increased slightly as the coefficient value also increases in the 31 day standard case, however for the 62 day standard the BIC stays relatively the same regardless of the choice of ϵ. This reaffirms that model 3.2.3 is not optimal when trying to model the impact of the pandemic.

Next the results of the simulation for model 3.2.5 can be found in figure 9 below.

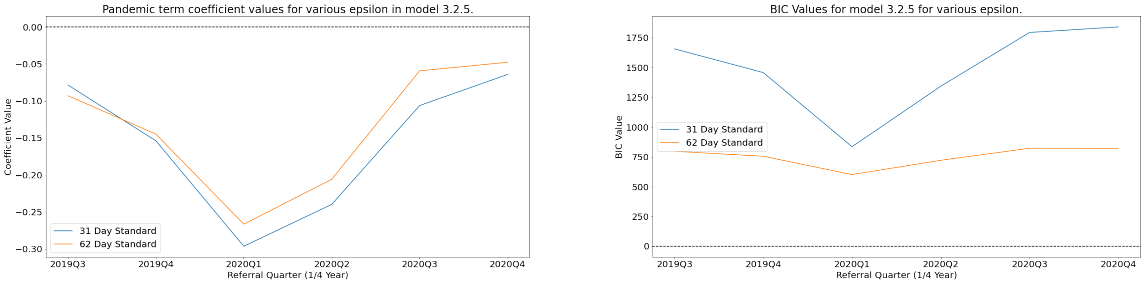

Figure 9: Values of the pandemic inclusion term coefficient for a range of selected epsilon for model 3.2.5.

Figure 9 indicates that the choice of ϵ to be at the first quarter of 2020 is optimal. This is shown by the most negative value being given by the coefficient and is reinforced by the smallest BIC score for both the 31 and 62 day standards.

For both model 3.2.3 and 3.2.5 we see that regardless of the choice of epsilon, the pandemic inclusion term is always a statistically significant coefficient.

4 Driving Factors

In the previous section, we showed that there has been a reduction in the number of referrals for the 31 and 62 day standard coinciding with COVID-19. However, prior to COVID-19, there was additional variation in the number of referrals, including a meeting of the 31-day standard but not the 62-day standard. This suggests the existence of alternative drivers for variation on the amount of patients being referred and treated within the waiting times standards. Assessing this is the goal of our second research question.

> **Research Question 2:**
> What factors contribute to the variation of cancer patients referred for waiting times standards in Scotland?

4.1 Region and Cancer Type As Covariates

Constrained to the demographic data provided with the eligible referral data for 31 or 62 day standards, we can create a GLM that includes categorical variables for region and cancer type (14 regions and 10 cancer types as explained in section 1). This GLM is defined as follows:

Let $Y_{t,c,r}$ be the number of patients referred for cancer treatment with,

- t: quarter of referral,
- c: cancer type,
- r: health board region of referral.

Assuming that $Y_{t,c,r} \sim \text{Poisson}(\theta_{t,c,r})$ then we can model the key drivers for the number of referrals as,

$$\textbf{model 4.1.1:} \quad \log(\theta_{t,c,r}) = \log(E_{t,r}) + \beta_0 + \beta_1 t + \beta_{3,c} + \beta_{4,r},$$

where the population size, $E_{t,r}$ is now dependent on both time and region and $\beta_{3,c}$ and $\beta_{4,r}$ are categorical variables. Again, we also fit a model that includes the impact of COVID-19:

$$\textbf{model 4.1.3:} \quad \log(\theta_{t,c,r}) = \log(E_{t,r}) + \beta_0 + \beta_1 t + \beta_{3,c} + \beta_{4,r} + \gamma_0 (t-\epsilon)\mathscr{I}(t-\epsilon).$$

We tried the same variations of terms as in the previous section, giving equivalent models for 3.2.1-3.2.6 with additional coefficients $\beta_{3,c}$ and $\beta_{4,r}$. The coefficients β_1, β_2 and γ_0 of these models were insignificant for their equivalent models in section 4.3. Similarly to this section, the BIC decreased as the model became more complex, as shown in appendix A. However, any fitted model had almost the exact same coefficients $\beta_{3,c}$ and $\beta_{4,r}$ and corresponding confidence intervals. Therefore, we only plot the coefficients for the most complex model with the lowest BIC,

$$\textbf{model 4.1.6:} \quad \log(\theta_{t,c,r}) = \log(E_{t,r}) + \beta_0 + \beta_1 t + \beta_2 t^2 + \beta_{3,c} + \beta_{4,r} + \gamma_0 \mathscr{I}(t-\epsilon) + \gamma_1 (t-\epsilon)\mathscr{I}(t-\epsilon).$$

These coefficients are shown in Figures 10 and 11 for the 31 day referral and 62 day referrals data, respectively. Overall, the four cancer types that have the most impact on the number of referrals are: urological, breast, lung and colorectal cancer. Meanwhile ovarian and cervical cancer have the least impact on our model. The coefficients for the regions are overall closer to 0, suggesting that cancer type has a bigger influence on the number of referrals than region. We also see a difference between the distribution of coefficients between the 31 and 62 day standard. Even though both show Western Isles to have a high impact and Lanarkshire to have a low impact on the number of referrals, the coefficients for the number of referrals eligible for the 31 day standard additionally show a higher impact for Greater Glasgow & Clyde, Dumfries & Galloway, and Tayside, and a lower impact for Orkney.

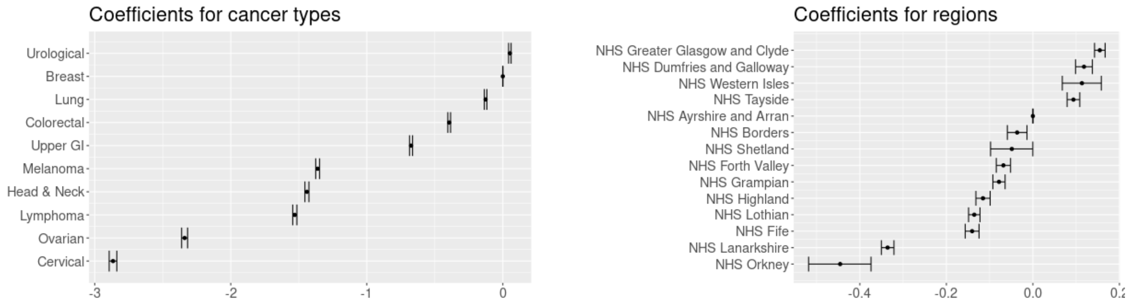

Figure 10: Coefficients $\beta_{3,c}$ and $\beta_{4,r}$ and their respective confidence intervals for cancer types and regions for 31 day referrals as resulted from model 4.1.6.

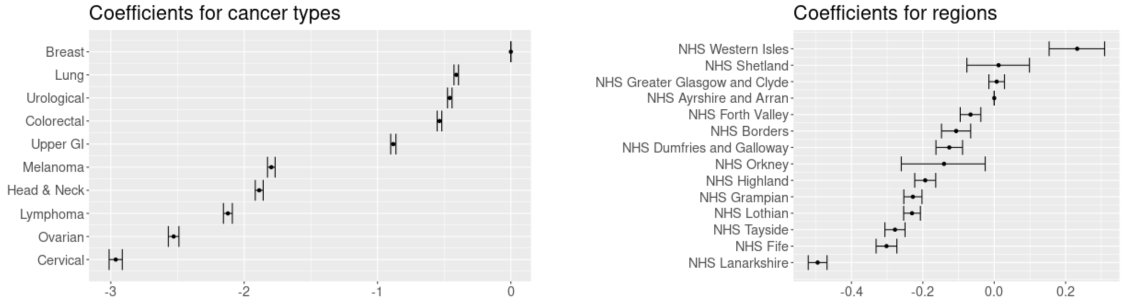

Figure 11: Coefficients $\beta_{3,c}$ and $\beta_{4,r}$ and their respective confidence intervals for cancer types and regions for 62 day referrals as resulted from model 4.1.6.

4.2 Interactions of Important Factors and the Pandemic

The previous section suggests that some regions and cancer types have a higher or lower impact on the number of referrals than others. Therefore, we might expect the impact of COVID-19 to also vary for certain regions or cancer types. In this section, we will look at the interactions between the regions and pandemic impact as well as interactions between the cancer types and pandemic impact. This could tell us how COVID-19 has differently affected the number of referred cancer patients for different regions and cancer types. To do this, we use the GLM,

$$\textbf{model 4.2.3:}\quad \begin{aligned}\log(\theta_{t,c,r}) =& \log(E_{t,r}) + \beta_0 + \beta_1 t + \beta_{3,c} + \beta_{4,r} + \gamma_0(t-\epsilon)\mathscr{I}(t-\epsilon)\\ &+ \gamma_{2,c}(t-\epsilon)\mathscr{I}(t-\epsilon) + \gamma_{3,r}(t-\epsilon)\mathscr{I}(t-\epsilon),\end{aligned}$$

where $\gamma_{2,c}$ is the categorical interaction term between γ_0 and different cancer types and $\gamma_{3,r}$ is the categorical interaction term between γ_0 and different regions. The GLM with nonlinear time trend coefficient, β_2, was discarded because of insignificant coefficients. The extra pandemic recovery term γ_1 could be added, but it would cause many more interaction terms and therefore this simpler model was fitted.

Figures 12 and 13 show the interaction coefficients for the regions and cancer types. The significant pandemic coefficients γ_0 are -0.0435313 for the 31 day referrals, and -0.0376035 for the 62 day referrals. Both coefficients have p-values smaller than $9e-7$. The first notable thing about the coefficients is their large confidence intervals. This could be because the COVID-19 coefficient γ_0 only takes affect from '2020Q1' to '2021Q3', so there is fewer data points on which these coefficients are based, in comparison to coefficients $\beta_{3,c}$ and $\beta_{4,r}$. Note that the cancer types with the lowest coefficients in the previous section 4.1, ovarian and cervical cancer, are the cancer types with the largest confidence intervals. This again suggests that a lack of data could be the cause for the large confidence intervals. The large confidence intervals in turn contribute to the fact that there are no clear outliers the same for both the 31 and 62 day referrals.

For 31 day standard, figure 12 does suggest that the number of referred patients with upper gastrointestinal cancer has been less affected by the pandemic than the number of referred patients with ovarian or lymphoma cancer. It also suggests that the number of referrals of Orkney and the Western Isles have been less impacted by the pandemic than Borders.

For the 62 day referrals, figure 13 suggests that the number referrals for head & neck, urological, and upper gastrointestinal cancer have been less impacted by covid-19 than lung and ovarian cancer. The region coefficients are so close together with relatively large confidence intervals that it is hard to speak of any clear differences between regions.

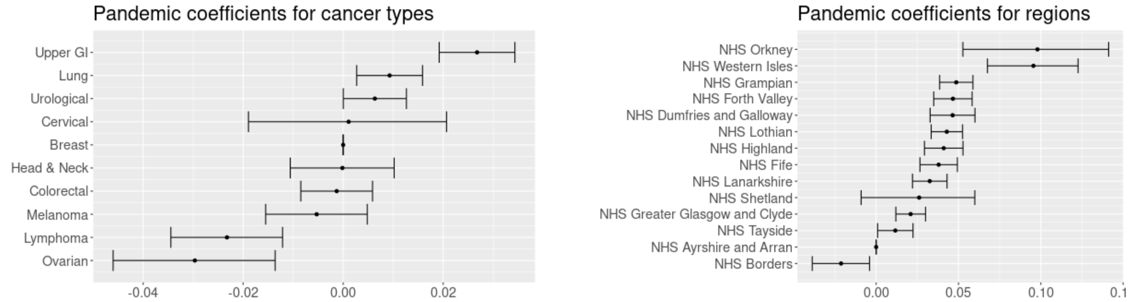

Figure 12: Interaction coefficients $\gamma_{2,c}$ and $\gamma_{3,r}$ for different cancer types and regions, respectively, for the 31 day standard resulting from fitting model 4.2.3.

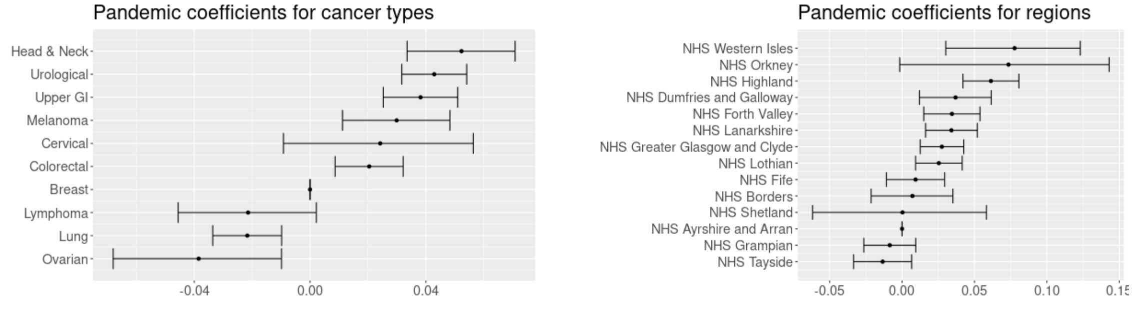

Figure 13: Interaction coefficients $\gamma_{2,c}$ and $\gamma_{3,r}$ for different cancer types and regions, respectively, for the 62 day standard resulting from fitting model 4.2.3.

4.3 A GLM for Specific Regions and Cancer Types

To explore in further detail the differences between cancer types and regions, we fit a model on data for one specific region and one specific cancer type. As this would result in many different models, we picked the most common cancer types, breast and lung cancer, to analyse. To encompass all referrals, we chose the three super-regions WOSCAN, NCA, and SCAN, to model separately. Fitting the same six models as in section 3.2 gives figures 14 for the 31 day referrals and figure 15 for the 62 day referrals. Models with insignificant coefficients β_1, β_2, γ_0 or γ_1 are discarded. All significant coefficients can be found in table 3.

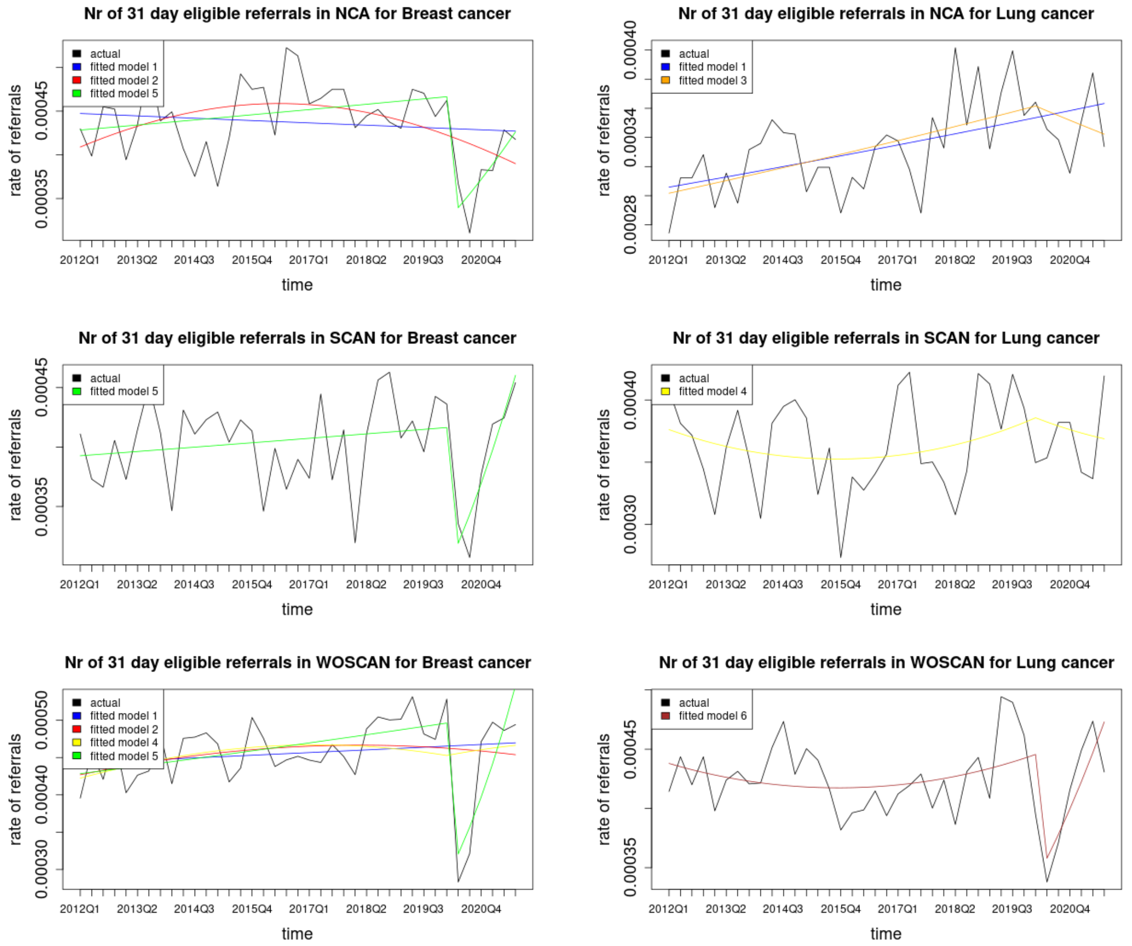

Figure 14: For the 31 day standard, rate of eligible referrals and their fitted models with significant coefficients for different regions (NCA, SCAN, WOSCAN) and different cancer types (breast, lung).

Firstly, it is notable that different models give a better fit with significant coefficients for different regions and different cancer types, suggesting again that region and cancer type are important factors impacting the number of referrals. Next to that, none of the fitted models have significant coefficients for the number of referrals with lung cancer in NCA, suggesting that a general model for all cancer types and all regions is not able to describe the data for specific regions and cancer types. This also makes it hard to compare different regions and cancer types as coefficients are not directly comparable.

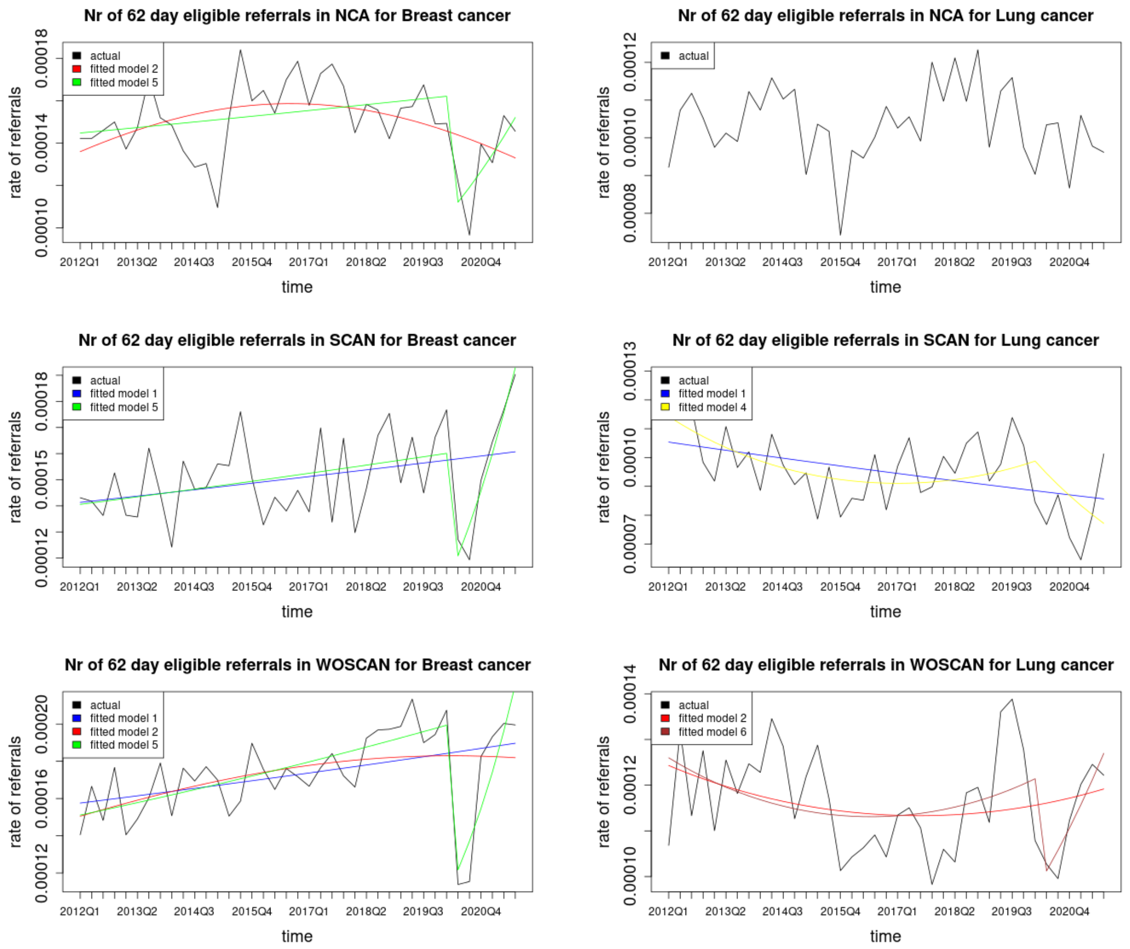

Figure 15: For the 62 day standard, rate of eligible referrals and their fitted models with significant coefficients for different regions (NCA, SCAN, WOSCAN) and different cancer types (breast, lung).

When observing the rate of referrals with breast cancer in any region, fitting model 4.3.5 gives significant coefficients, while this model never results in significant coefficients for lung cancer referrals. Looking at the data in figures 14 and 15, we see that this seems to align with the fact that the rate of referrals for breast cancer is better modelled with a drop, while the rate of referrals for lung cancer does not follow a steep drop when the pandemic started.

Comparing the coefficients γ_0 for breast cancer in different regions in table 3, shows that for both the 31 and 62 day standard, the impact of COVID-19 on breast cancer referrals is the highest in WOSCAN and lowest in SCAN. Similarly, the recovery coefficient γ_1 for breast cancer is highest in WOSCAN and lowest in SCAN. This suggests that even though WOSCAN was most impacted by the pandemic, the rate of referrals also increased the most on the time trend in the recovery period. Another interesting observation is the fact that for the both standards in regions SCAN and WOSCAN, the coefficient β_1 is negative for lung cancer, while positive for breast cancer. This suggests that the rate of referrals in these regions decreases for lung cancer while increases for breast cancer. This is unexpected, because incidence rate of all cancer types usually increase over time.

standard	region	cancer type	model	β_1	β_2	γ_0	γ_1	BIC
31 day	NCA	breast	4.3.1	−0.00121				558
			4.3.2	0.01322	0.00038			520
			4.3.5	0.00266		−0.36303	0.04280	468
		lung	4.3.1	0.00453				413
			4.3.3	0.00565		−0.01480		411
	SCAN	breast	4.3.5	0.00184		−0.33987	0.07152	479
		lung	4.3.4	−0.00883	0.00030	−0.01966		533
	WOSCAN	breast	4.3.1	0.00157				878
			4.3.2	0.00728	−0.00015			869
			4.3.4	0.00978	−0.00024	0.01181		868
			4.3.5	0.00459		−0.54304	0.10196	579
		lung	4.3.6	−0.00652	0.00022	−0.2733	0.04659	512
62 day	NCA	breast	4.3.2	0.01678	−0.00046			386
			4.3.5	0.00354		−0.42997	0.05746	368
		lung	-					-
	SCAN	breast	4.3.1	0.00338				384
			4.3.5	0.00406		−0.37473	0.08942	355
		lung	4.3.1	−0.00545				358
			4.3.4	−0.02268	0.00057	−0.05818		351
	WOSCAN	breast	4.3.1	0.00488				564
			4.3.2	0.01205	−0.00019			561
			4.3.5	0.00866		−0.61317	0.11074	418
		lung	4.3.2	−0.00845	0.00019			391
			4.3.6	−0.01222	0.00035	−0.22507	0.03301	385

Table 3: Model coefficients for fitted models per region and cancer type. Only models with all coefficients having p-values < 0.05 are included, where coefficients in dark gray boxes have p-value < 0.05, in light gray boxes have p-value < 0.01, and all other coefficients have p-value < 0.001.

5 Diagnosis Data

In this section, we analyse the diagnosis data set to study the effect of the pandemic and demographic factors on cancer diagnosis, as a proxy to better understand how the previous shape the waiting times landscape.

Research Question 3:
What is the impact of COVID-19 and other factors on the number of cancer diagnoses in Scotland? How does it compare to the number of individuals being referred for treatment under the 31 and 62 day standard in Scotland?

5.1 The Impact of COVID-19

We now follow the same methods as in section 3 to analyse the impact of the pandemic on cancer diagnosis using different GLMs by comparing models. We start from the simplest model that normalises the number of diagnosis by the population size of the region and considers a linear trend in time (model 5.1.1), and then additional linear trends representing the pandemic (model 5.1.2) and the recovery period (model 5.1.3). The cancer diagnosis data is reported weekly instead of quarterly, so we choose the beginning of the pandemic ϵ to be the third week of March 2020. We do not consider quadratic terms since the analysis in section 3 showed that they were not always statistically significant.

model 5.1.1: $\log(\xi_t) = \log(E_t) + \beta_0 + \beta_1 t,$

model 5.1.2: $\log(\xi_t) = \log(E_t) + \beta_0 + \beta_1 t + \gamma_0 (t-\epsilon)\mathscr{I}(t-\epsilon),$

model 5.1.3: $\log(\xi_t) = \log(E_t) + \beta_0 + \beta_1 t + \gamma_0 \mathscr{I}(t-\epsilon) + \gamma_1 (t-\epsilon)\mathscr{I}(t-\epsilon).$

We fit these models to the cancer diagnosis data, aggregating over all the other covariates (region, cancer type, age and sex). Table 4 shows the values of the coefficients, their confidence interval, and the BIC score of each model. In all cases, all coefficients were found statistically significant with p-values smaller than $2e-12$. The linear time trend is weak in all models (order of 10^{-3}), and negative for models 5.1.1 and 5.1.3, and positive for 5.1.2. In all cases, the pandemic coefficient is negative and of order 10^{-1}. Finally, in model 5.1.3, the recovery term is positive, but 3 orders of magnitude smaller than the pandemic term. We notice that the latter model, which includes both the pandemic and the recovery term, has lowest BIC, meaning it is the model that best describes the data. This result suggests that there was a significant decrease in the number of cancer diagnosis starting with the beginning of the pandemic followed by a weaker but positive recovery.

Comparing this to Table 2, we see that the pandemic has had a similar affect in the rates of cancer diagnosis and referrals for the 31 and 62-day standards. In all cases, the time series data is better explained by models that include both the pandemic and the recovery terms, with the coefficient of the first taking values around -0.3. Thus, our analysis of cancer diagnosis data supports the results obtained in the previous study using 31 and 62-day referral data, even though this has lower time resolution.

	coefficient	value	confidence interval $(2.5\%, 97.5\%)$	BIC
model 5.1.1	β_1	-0.00119	$(-0.00131, -0.00108)$	89793.8
model 5.1.2	β_1	0.00239	$(0.00216, 0.00262)$	89679.9
	γ_0	-0.3097	$(-0.3268, -0.2927)$	
model 5.1.3	β_1	-0.00237	$(-0.002681, -0.00207)$	89450.8
	γ_0	-0.3618	$(-0.3798, -0.3438)$	
	γ_1	0.01126	$(0.01076, 0.01169)$	

Table 4: Coefficients, confidence intervals and BIC scores of models 5.1.1-5.1.3 fitted to cancer diagnosis data. All coefficients were significant with p-values smaller than $2e-12$.

5.2 Important factors

In order to study the contribution of different factors to the variation in cancer diagnosis data, we follow the same methods as in Section 4 and fit the data to GLMs with different combinations of explanatory variables. Given the additional demographic data provided in the cancer diagnosis data, we create a GLM that includes categorical variables for region and cancer type, as well as sex and age group.

Let $X_{t,c,r,s,a}$ be the number of cancer diagnosis,

- t: week of diagnosis,
- c: cancer type,
- r: health board region of referral,
- s: sex,
- a: age group [1].

Assuming that $X_{t,c,r,s,a} \sim$ Poisson($\xi_{t,c,r,s,a}$) we can model the important factors for the number of diagnosis as

$$\textbf{model 5.2.1:} \quad \log(\xi_{t,c,r,s,a}) = \log(E_{t,r}) + \beta_0 + \beta_1 t + \beta_{3,c} + \beta_{4,r} + \beta_{5,s} + \beta_{6,a}.$$

Following our results in the previous section, we include two terms to model the impact of COVID-19:

$$\textbf{model 5.2.2:} \quad \log(\xi_{t,c,r,s,a}) = \log(E_{t,r}) + \beta_0 + \beta_1 t + \beta_{3,c} + \beta_{4,r} + \beta_{5,s} + \beta_{6,a} + \gamma_0(t-\epsilon)\mathscr{I}(t-\epsilon) + \gamma_1(t-\epsilon)\mathscr{I}(t-\epsilon).$$

We consider different GLMs, starting from model 5.1.1 as the simplest model and up to model 5.2.2 as the most complicated model. Table 5 shows the BIC score for different models. Remarkably, we notice that adding a categorical variable for cancer causes the BIC to be reduced by half. Addition of all other variables also results in lower BIC, i.e. in models that fit the data better, except for the addition of sex. Thus, the GLM that better fits the data is:

$$\textbf{model 5.2.3:} \quad \log(\xi_{t,c,r,s,a}) = \log(E_{t,r}) + \beta_0 + \beta_1 t + \beta_{3,c} + \beta_{4,r} + \beta_{6,a} + \gamma_0(t-\epsilon)\mathscr{I}(t-\epsilon) + \gamma_1(t-\epsilon)\mathscr{I}(t-\epsilon).$$

[1] Population studies usually treat age as a numerical variable. We treat is as categorical because the cancer diagnosis data set only specifies the age as belonging to one of 0-49, 50-69 or 70+.

	$\beta_{3,c}$	$\beta_{4,r}$	$\beta_{5,s}$	$\beta_{6,a}$	γ_0	γ_1	BIC
model 5.1.1							89793.8
	■						43351.5
	■	■					41258.5
	■	■	■				41348.1
	■	■		■			39782.5
	■	■	■	■			39562.7
	■	■	■		■	■	38127
model 5.2.3	■	■		■	■	■	36254.8
model 5.2.2	■	■	■	■	■	■	36968.3

Table 5: BIC scores of GLMs fitted to cancer diagnosis data. A blue shade indicates that the GLM contains the corresponding variable. All GLMs contain the population size offset, the intercept β_0 and the linear time trend $\beta_1 t$.

Figures 16 and 17 display the values of the coefficients for different regions, cancer types and age of model 5.2.3. On the one hand, all coefficients for cancer type are significant with p-values smaller than $2e-12$, showing the importance of the type of cancer in the rates of diagnosis. As expected, the cancer types with highest diagnosis rates are breast, colorectal, lung and prostate cancer. On the other hand, some coefficients for regions are not statistically significant, possibly because some HBs have small population sizes. We do notice that the HB corresponding to Greater Glasgow & Clyde has the largest rate of cancer diagnosis, whilst Lothian has one of the lowest. Finally, our analysis aggregating over all cancer types suggests that age is a strong driver for cancer diagnosis, with significant and large differences between age groups. As expected, the rate of cancer diagnosis is significantly larger for people over 50, and largest for people over 70. However, sex is not a significant significant factor, as we can conclude from both the fact that the coefficients are not significant and that the model including sex as a categorical variable has one the larger BIC scores.

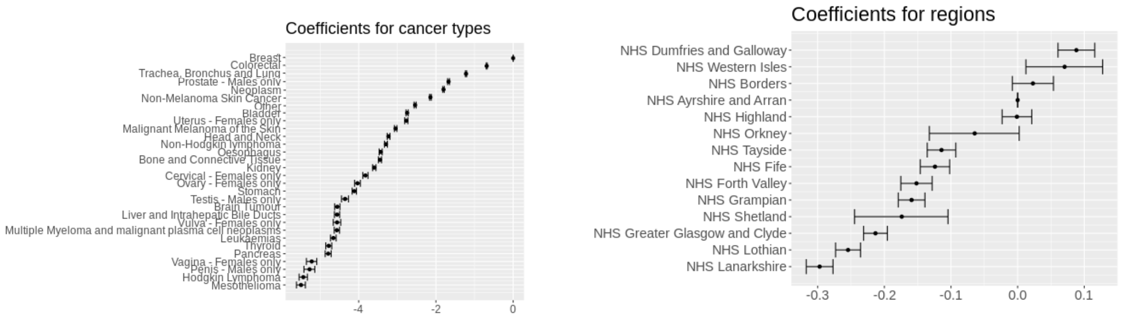

Figure 16: Coefficients for cancer types $\beta_{3,c}$ (left) and regions $\beta_{4,r}$ (right) of model 5.2.3 fitted to the cancer diagnosis data.

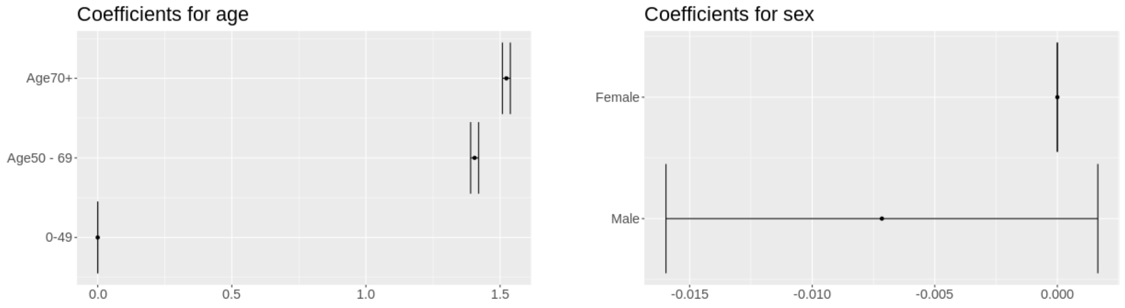

Figure 17: Coefficients for age $\beta_{6,a}$ (left) of model 5.2.3 and coefficients for sex $\beta_{5,s}$ of model 5.2.2 both fitted to the cancer diagnosis data.

Comparing these results to the equivalent figures for the 31 and 62-day referrals data (Figures 10 and 11), we notice remarkable similarities between the cancer types and regions that have higher rates. Moreover, this analysis shows the strength of age as a determinant for cancer diagnosis, which the 31 and 62 day data did not include due to data limitations.

5.3 Drivers for Breast and Lung Cancer

The results of the previous section highlight the significance of differences in rates of diagnosis between cancer types. We end this section by considering breast and lung cancer separately, analogous to our analysis of the 31 and 62 days standard. Table 6 shows the values and confidence intervals of the linear time and pandemic terms. These do not differ much from the values obtained fitting the model to data aggregated by cancer type, but we do notice that the value of γ_0 are more negative, especially for breast cancer. This suggests that the pandemic had a larger impact on breast cancer rates than lung cancer, which were reduced at the start of the pandemic more than the average for all cancer types.

	coefficient	value	confidence interval
Breast Cancer	β_1	-0.0021	$(-0.00371, -0.00127)$
	γ_0	-0.4715	$(-0.5463, -0.3968)$
	γ_1	0.01345	$(-0.01152, -0.01536)$
Lung Cancer	β_1	-0.001619	$(-0.003347, -0.001067)$
	γ_0	-0.3631	$(-0.4668, -0.2597)$
	γ_1	0.01836	$(-0.02034, -0.01739)$

Table 6: Time and pandemic coefficients and confidence intervals of model 5.2.3 fitted to breast and lung cancer diagnosis data. All coefficients were significant with with p-values smaller than $2e-12$.

Finally, figures 18 and 19 show the values and confidence intervals of the coefficients for region and age for breast and lung cancer, respectively.

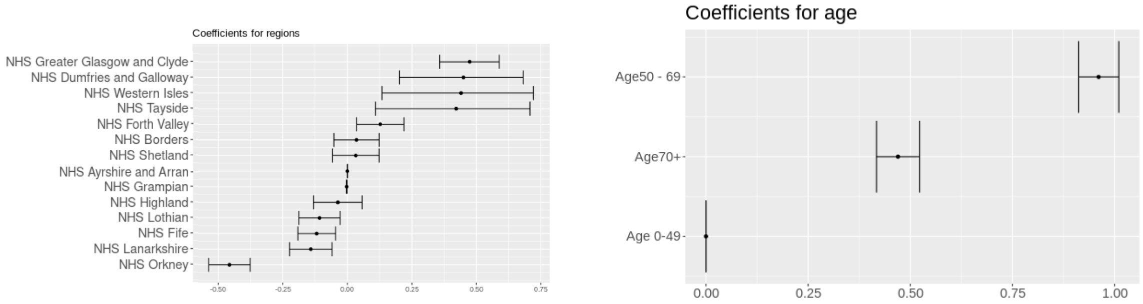

Figure 18: Coefficients for regions $\beta_{4,r}$ (left) and age $\beta_{6,a}$ (right) of model 5.2.3 fitted to the breast cancer diagnosis data.

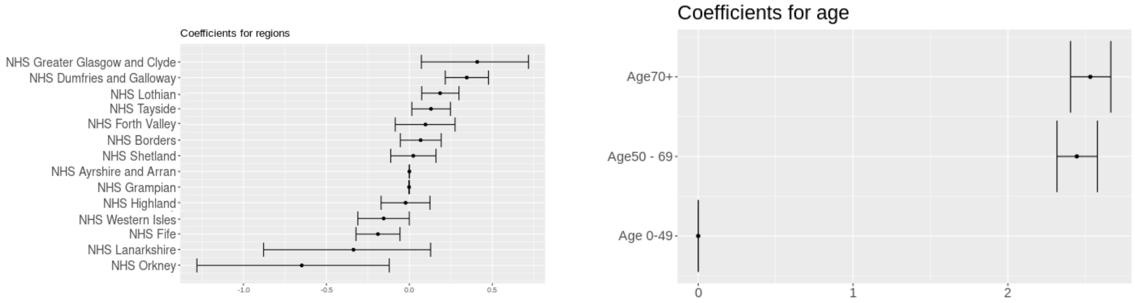

Figure 19: Coefficients for regions $\beta_{4,r}$ (left) and age $\beta_{6,a}$ (right) of model 5.2.3 fitted to the lung cancer diagnosis data.

We notice some remarkable differences between the coefficients for regions, which may be a consequence of the two types of cancer being influenced by different environmental and demographic factors. Breast cancer rates are highest for people between 50 and 70, whilst lung cancer has largest rates in the 70+ age group. This results is consistent with the literature [13]. Overall, this analysis, together with the results of Section 4.3 show the importance of treating different cancer types separately.

Part IV
Conclusions

"The essence of mathematics lies in its freedom."

Georg Cantor

Discussion of Results

Overall, we were able to show that there was a significant change point in the number of cancer referrals and diagnoses at the start of the pandemic (first quarter of 2020). Change point detection algorithms were insufficient to show this change point, but GLMs showed it for the number of cancer referrals for both the 31 and 62 day standard and for the number of cancer diagnoses. How this change point is best modelled is, however, to be discussed. For all datasets, the GLM that had the best fit to the data and lowest BIC is the most complex model with a new offset term at the start of the pandemic and a recovery term on the time trend. Thus after the pandemic, the number of cancer referrals and diagnoses declined within the next quarter and then increased on a steeper trend than before. Future work should consider the predictive power of the models suggested, although this might be compromised in the case of GLMs tailored to fit the influence of COVID-19. Once more data is available, it would therefore be interesting to evaluate the same models on their predictive power to see which model performs best in this regard.

PHS reports the percentage of treated cancer patients within the 31 or 62 day standard as a proxy for cancer waiting times. This seemed unaffected by COVID-19, as explained in the introduction. However, our analysis shows that there was a drop in numbers of referred cancer patients, it might be that the standards were met only because there were less patients to treat. However, our report cannot make any conclusions about the cancer waiting times. This is something to explore once more detailed data is available.

The rate of cancer referrals and diagnoses are determined by cancer type and region. Cancer type is the strongest determinant. This is in line with expectations, because cancer is very heterogenic as a disease, and different cancer types have different different features. This is reiterated by the fact that models for specific regions and cancer types give varying significant coefficients. This suggests that cancer type, and region, are important factors that significantly change the best fitted model, and that you lose important variations if you model aggregated data. Overall, the four cancer types that have the most impact on the number of referrals are: urological, breast, lung and colorectal cancer. Meanwhile ovarian and cervical cancer have the least impact on the number of referrals and diagnoses. Note that these differences include that some cancer types are more common than others.

Different regions also have a varying impact on the number of referrals and diagnoses, even if these variations are less strong in comparison with different cancer types. The Western Isles and Orkney HBs have a higher and lower impact on the number of referrals and diagnoses, which could be explained by the fact that they are relatively isolated from the rest of Scotland. It is, however, surprising that Lanarkshire, surrounded by regions with a higher impact such as Dumfries & Galloway and Greater Glasgow & Clyde, has a lower impact on the number of referrals and diagnoses. In the future, we suggest to take a more detailed look at a GLM for Lanarkshire specifically to explain this unexpected result.

In an exploration to find how different cancer types impacted the drop in the number of referrals during the pandemic, we noticed large confidence intervals that were most likely caused by a lack of data. There are only six time points after the start of COVID-19, and especially for less common cancer types, this made it difficult to make many conclusions. We did see that the number of referred patients with upper gastrointestinal cancer has been less affected by the pandemic, while number of referred patients with ovarian cancer have been more impacted by the pandemic.

For more detail on other cancer types and the differences between regions, we recommend to wait until more data is available for analysis. Taking a more detailed look at the number of referrals and diagnoses of breast and lung cancer specifically, showed that breast cancer is best modelled with a drop at the start of the pandemic, while lung cancer is not. This suggests that breast cancer was more impacted by COVID-19 than lung cancer. This reiterates the importance of looking at different cancer types separately, especially if more data is available as for more common cancer types as breast and lung caner.

An important difference between the referral and diagnosis data is that the diagnosis data also includes information on age and sex. As expected, the results show that age is a strong determinant, even stronger than region. Sex, however, is not a strong determinant with respect to the aggregated cancer types. Future research should consider the role of sex in specific cancer types.

Improvements to Public Health Measures and Data Collection

Across regions there is a varying impact on the number of referrals, even though it has been normalised by the population. It could be that the data is being recorded inconsistently across different regions. To make a stronger conclusion, it is necessary that NHS Scotland and PHS continue to use the 31 and 62 day standards to collect and present waiting times with a standardised application of the criteria. This has the potential to improve the quality of the data set, allowing statistical analysis to draw more reliable conclusions. Additionally, this would help in terms of modelling demand for the future in regions, which in turn could see the process of referrals to treat in a different region from that originally giving the referral becoming much more streamlined and consistent.

Furthermore, the 31 and 62 day standard data set is very limited in terms of the demographic characteristics of patients that might play important roles. In the case of cancer diagnosis data, our results highlight the importance of age as a determinant. It would be interesting to include demographic features into the analysis of the 31 and 62 day standard, for which data is recorded in hospitals but not not included in the open data platform for the 31 and 62 day standards. Enriching the open data platform with additional information would benefit future research. This could be done whilst maintaining anonymity of patients, for example following the format of the cancer diagnosis data.

Future Modelling

In terms of modelling strategies, there is a multitude of directions in which future research could be developed. Firstly, in terms of identifying a change point in the data sets corresponding with the pandemic, one could use **high dimensional change point detection** [21]. This would involve connecting a large number of interconnected data sets and running similar algorithms to those described in Section 3. Examining other waiting times data could be particularly useful for monitoring any changes which occur due to societal influence, for example a screening program for cancer.

Secondly, a similar analysis could be performed using **generalised additive model** (GAM) in place of GLMs. An advantage of GAM is that they do not incorporate time explicitly. Instead, they assume the linear response terms depend linearly on unknown smooth functions of some predictor variables, and then recover the time dependence within the system. Implementing GAMs would allow flexibility in creating a non-parametric relationship between the number of referrals and the time period but potentially could see a huge loss in terms of being able to narrow down what is happening in one select time period of historical data.

In this report, we have only focused on three data sets available by Public Health Scotland. However, the cancer waiting landscape might be better explained by additional data. For example, the total number of hospital beds might influence the capacity of the hospital and therefore its ability to treat cancer patients within the 31 or 62 day standard. Next to that, there are analogous types of data collected by other countries leading health bodies. It would be interesting to not only use the same methods in this report to analyse these data sets, but also to combine data sets to get a better picture of the interlinking data available. A model including more data sets could also include more important factors and therefore make more accurate predictions, allowing PHS to predict future demand and inform policy-making within NHS Scotland.

Appendices

A GLM Coefficients

This appendix includes all GLM coefficients for different models throughout the report. Note that $\beta^{(*)}$ means the coefficient β was not significant.

31 day referrals	coefficient	value	confidence interval $(2.5\%, 97.5\%)$	p-value	BIC
model 3.2.1	β_1	0.001562	$(0.001303431, 0.001819711)$	$< 2e-16$	1879.27
model 3.2.2	β_1	0.004066	$(3.049491e-03, 5.083217e-03)$	$4.61e-15$	1857.98
	β_2	-0.000066	$(-9.123126e-05, -3.979361e-05)$	$5.98e-07$	
model 3.2.3	β_1	0.002228	$(0.001910225, 0.002544998)$	$< 2e-16$	1832.86
	γ_0	-0.008914	$(-0.011391639, -0.006440445)$	$1.69e-12$	
model 3.2.4	β_1	0.002140	$(8.760938e-04, 3.404805e-03)$	0.000909	1836.51
	$\beta_2^{(*)}$	0.000003	$(-3.437254e-05, 3.965217e-05)$	0.888428	
	γ_0	-0.009097	$(-1.265463e-02, -5.540104e-03)$	$5.38e-07$	
model 3.2.5	β_1	0.003766	$(0.003434953, 0.004096536)$	$< 2e-16$	835.694
	γ_0	-0.296546	$(-0.315308665, -0.277830376)$	$< 2e-16$	
	γ_1	0.049672	$(0.045260123, 0.054085285)$	$< 2e-16$	
model 3.2.6	β_1	-0.002924	$(-0.0042165115, -0.0016301388)$	$9.36e-06$	729.878
	β_2	0.000207	$(0.0001682056, 0.0002455893)$	$< 2e-16$	
	γ_0	-0.3275	$(-0.3471114707, -0.3079262938)$	$< 2e-16$	
	γ_1	0.04165	$(0.0369885230, 0.0463079240)$	$< 2e-16$	

Table 7: Coefficients of different models for 31 day referrals for section 3.2.

62 day referrals	coefficient	value	confidence interval $(2.5\%, 97.5\%)$	p-value	BIC
model 3.2.1	β_1	0.006062	$(0.005577691, 0.006545496)$	$< 2e-16$	840.120
model 3.2.2	β_1	0.004066	$(2.151904e-03, 0.0059830861)$	$3.18e-05$	839.339
	β_2	0.000052	$(3.634094e-06, 0.0000995568)$	0.0349	
model 3.2.3	β_1	0.006168	$(0.005570462, 0.006766407)$	$< 2e-16$	843.429
	$\gamma_0^{(*)}$	-0.001360	$(-0.005841995, 0.003110133)$	0.551	
model 3.2.4	$\beta_1^{(*)}$	0.001799	$(-5.946579e-04, 0.0041957092)$	0.141026	833.4980
	β_2	0.000131	$(6.140107e-05, 0.0002005684)$	0.000224	
	γ_0	-0.0101950	$(-1.667789e-02, -0.0037132991)$	0.002052	
model 3.2.5	β_1	0.007627	$(0.007003263, 0.008251734)$	$< 2e-16$	601.387
	γ_0	-0.2666711	$(-0.300684259, -0.232810832)$	$< 2e-16$	
	γ_1	0.050600	$(0.042678222, 0.058525732)$	$< 2e-16$	
model 3.2.6	β_1	-0.0033165	$(-0.0057649541, -0.0008644362)$	0.00798	523.728
	β_2	0.0003357	$(0.0002629323, 0.0004084770)$	$< 2e-16$	
	γ_0	-0.315384	$(-0.3509311996, -0.2799677953)$	$< 2e-16$	
	γ_1	0.0376619	$(0.0292620303, 0.0460667410)$	$< 2e-16$	

Table 8: Coefficients of different models for 62 day referrals for section 3.2.

31 day referrals	coefficient	value	confidence interval $(2.5\%, 97.5\%)$	p-value	BIC
model 4.1.1	β_1	0.0016012	$(0.001343024, 0.0018593950)$	$< 2e-16$	54689.03
model 4.1.2	β_1	$4.008e-03$	$(2.991282e-03, 5.025614e-03)$	$1.14e-14$	54674.51
	β_2	$-6.294e-05$	$(-8.866774e-05, -3.722563e-05)$	$1.62e-06$	
model 4.1.3	β_1	0.002282	$(0.001964997, 0.0025999839)$	$< 2e-16$	54645.22
	γ_0	-0.009113	$(-0.011590176, -0.0066386229)$	$5.43e-13$	
model 4.1.4	β_1	$1.917e-03$	$(6.526079e-04, 3.181708e-03)$	0.00297	54653.40
	$\beta_2^{(*)}$	$1.106e-05$	$(-2.596156e-05, 4.806155e-05)$	0.55810	
	γ_0	$-9.876e-03$	$(-1.343259e-02, -6.319272e-03)$	$5.26e-08$	
model 4.1.5	β_1	0.0038119	$(0.003480981, 0.0041427747)$	$< 2e-16$	53664.28
	γ_0	-0.2949790	$(-0.313747501, -0.2762576298)$	$< 2e-16$	
	γ_1	0.0491720	$(0.044758340, 0.0535869194)$	$< 2e-16$	
model 4.1.6	β_1	$-3.133e-03$	$(-0.0044260670, -0.0018394747)$	$2.05e-06$	53554.83
	β_2	$2.148e-04$	$(0.0001760675, 0.0002534410)$	$< 2e-16$	
	γ_0	$-3.271e-01$	$(-0.3467038561, -0.3075116556)$	$< 2e-16$	
	γ_1	$4.084e-02$	$(0.0361839733, 0.0455062082)$	$< 2e-16$	

Table 9: Coefficients of different models for 31 day referrals for section 4.1.

62 day referrals	coefficient	value	confidence interval $(2.5\%, 97.5\%)$	p-value	BIC
model 4.1.1	β_1	0.0058459	$(0.005362238, 0.006329618)$	$< 2e-16$	33966.92
model 4.1.2	β_1	$4.225e-03$	$(2.309705e-03, 6.143930e-03)$	$1.56e-05$	33972.47
	$\beta_2^{(*)}$	$4.191e-05$	$(-6.105371e-06, 8.987178e-05)$	0.0869	
model 4.1.3	β_1	0.0060242	$(0.005426132, 0.006622378)$	$< 2e-16$	33974.41
	$\gamma_0^{(*)}$	-0.0022712	$(-0.006761063, 0.002207336)$	0.3208	
model 4.1.4	$\beta_1^{(*)}$	$1.843e-03$	$(-5.508759e-04, 0.0042411161)$	0.131565	33970.46
	β_2	$1.253e-04$	$(5.570707e-05, 0.0001948904)$	0.000416	
	γ_0	$-1.073e-02$	$(-1.721971e-02, -0.0042381830)$	0.001197	
model 4.1.5	β_1	0.0075826	$(0.006958306, 0.008207108)$	$< 2e-16$	33703.54
	γ_0	-0.2844607	$(-0.318525795, -0.250548434)$	$< 2e-16$	
	γ_1	0.0532337	$(0.045296133, 0.061175943)$	$< 2e-16$	
model 4.1.6	β_1	$-3.573e-03$	$(-0.0060220024, -0.0011212322)$	0.00426	33627.54
	β_2	$3.422e-04$	$(0.0002693608, 0.0004148734)$	$< 2e-16$	
	γ_0	$-3.341e-01$	$(-0.3696486930, -0.2985956790)$	$< 2e-16$	
	γ_1	$4.005e-02$	$(0.0316366765, 0.0484709246)$	$< 2e-16$	

Table 10: Coefficients of different models for 62 day referrals for section 4.1.

B PHS Data Sets

Link	Recorded data	Time (last)	Region	Other factors
Cancer waiting times	nr of referrals, nr treated within standard (31/ 62)	quarterly (2012Q1 → 2021Q3)	HB,HBT	cancer type
Covid	positive cases, cumulative cases, deaths, cumulative deaths, tests, hospital admissions, ICU admissions	daily (2020-02-28)	HB	age and sex, deprivation
Cancer incidence	incidence all ages, crude rate, european age-standardised, world age-standardised, standardised incidence ratio	annual (2019)	HB	diagnosed cancer site, sex
Cancer mortality	deaths all ages, crude rate, european age-standardised, world age-standardised, standardised mortality ratio	annual (2019)	HB	diagnosed cancer site, sex
Population estimates	nr of people per age	annual (2020)	HB	sex
A&E waiting times	nr of attendances, nr meeting target, attendance greater than 8 hrs, attendance greater than 12 hrs	monthly (2021-11)	HBT	treatment hospital, department type
Diagnostic waiting times	nr on waiting list, nr waiting over 4 weeks, nr waiting over 6 weeks	monthly (2021-09)	HBT	test type
Referral to treatment	within 18 weeks, over 18 weeks	monthly (2021-09)	HBT	speciality (all)

Cancelled operations	total cancelled, by patient reason, by clinical reason, by capacity reason, by other reason	monthly (2021-11)	HBT	-
Additions and removals waiting list	additions, removals, referred back to GP, transferred, treatment no longer required, other	quarterly (2021 Q3)	HBT	speciality, patient type
Hospital beds	available beds, occupied beds, daily avg	quarterly (2021 Q2)	HB	hospital, speciality
Mental health inpatient	admissions, discharges, stays, patients, hospital residents	financial year (2020/2021)	HBT	data source
Cancer diagnosis	number of cancer diagnosis	weekly (2019/2021)	HBT	cancer type, age group, sex

References

[1] ARIK, A., DODD, E., CAIRNS, A., AND STREFTARIS, G. Socioeconomic disparities in cancer incidence and mortality in england and the impact of age-at-diagnosis on cancer mortality. *PLoS One*, **16** (2021), e0253854. doi:10.1371/journal.pone.0253854.

[2] BARNDORFF-NIELSEN, O. *Introductory Theory of Exponential Families*, chap. 8, pp. 109–137. Springer, Cham, 1 edn. (2014). doi:10.1002/9781118857281.ch8.

[3] DING-GENG (DIN) CHEN, J. K. C. *Linear Regression*, chap. 1.5, pp. 12–14. Springer, Cham, 1 edn. (2021). doi:10.1007/978-3-030-67583-7.

[4] DOLL, R., PETO, R., BOREHAM, J., AND SUTHERLAND, I. Mortality in relation to smoking: 50 years' observations on male british doctors. *BMJ*, **328** (2004), 1519. doi:10.1136/bmj.38142.554479.AE.

[5] KENNETH. P. BURNHAM, D. R. A. Understanding aic and bic in model selection. *Sociological Methods and Research*, **33** (2004), 261. doi:10.1177/0049124104268644.

[6] MAFHAM, M. M., ET AL. Covid-19 pandemic and admission rates for and management of acute coronary syndromes in england. *The Lancet*, **396** (2020), 381. doi:10.1016/S0140-6736(20)31356-8.

[7] MORRIS, E. J., ET AL. Impact of the covid-19 pandemic on the detection and management of colorectal cancer in england: a population-based study. *The lancet Gastroenterology & hepatology*, **6** (2021), 199. doi:10.1016/S2468-1253(21)00005-4.

[8] NATIONAL HEALTH SERVICES. The nhs cancer plan, a plan for investment, a plan for reform (2000). [Online; accessed 2-February-2022]. Available from: https://www.thh.nhs.uk/documents/_Departments/Cancer/NHSCancerPlan.pdf.

[9] NHS SCOTLAND. Healthboard areas of nhs scotland (2021). Available from: https://www.scot.nhs.uk/mapofscotlandshowversion-2/.

[10] OSLER, W. The Principles and Practice of Medicine Designed for the Use of Practitioners and Students of Medicine. *Journal of the American Medical Association*, **106** (1936), 566. doi:10.1001/jama.1936.02770070062035.

[11] PAUL ROBACK, J. L. *Poisson Regression*. Chapman and Hall, 1 edn. (2020). doi:10.1002/9781118857281.ch8.

[12] PLATTO, S., WANG, Y., ZHOU, J., AND CARAFOLI, E. History of the covid-19 pandemic: Origin, explosion, worldwide spreading. *Biochemical and biophysical research communications*, **538** (2021), 14. doi:10.1016/j.bbrc.2020.10.087.

[13] PUBLIC HEALTH SCOTLAND. Cancer incidence in scotland (to december 2018) (2020). [Online; accessed 18-March-2022]. Available from: https://publichealthscotland.scot/media/3597/2020-04-28-cancer-incidence-report.pdf.

[14] PUBLIC HEALTH SCOTLAND. Cancer waiting times data definitions manual version 5.3 (2021). [Online; accessed 18-March-2022]. Available from: https://publichealthscotland.scot/media/2810/public-health-scotland-strategic-plan-2020-23.pdf.

[15] PUBLIC HEALTH SCOTLAND. Cancer waiting times data definitions manual version 5.3 (2021). [Online; accessed 18-March-2022]. Available from: https://www.isdscotland.org/Health-Topics/Waiting-Times/Cancer/Guidance/_docs/CWT-Data-and-Definitions-Manual-Version-5.3.pdf.

[16] PUBLIC HEALTH SCOTLAND. Cancer waiting times in nhs scotland, 1 july to 30 september 2021 (2021). Available from: https://www.publichealthscotland.scot/media/10777/2021-12-14-cwt-report.pdf.

[17] PUBLIC HEALTH SCOTLAND. About - scottish health and social care open data (2022). [Online; accessed 25-February-2022]. Available from: https://www.opendata.nhs.scot/about.

[18] SCOTTISH GOVERNMENT. Clinical review of cancer waiting times (cwt) standards in scotland (2018). [Online; accessed 09-March-2022]. Available from: https://www.gov.scot/publications/clinical-review-cancer-waiting-times-cwt-standards-scotland/documents/.

[19] SCOTTISH GOVERNMENT. Waiting times improvement plan (2018). [Online; accessed 18-March-2022]. Available from: https://www.gov.scot/binaries/content/documents/govscot/publications/strategy-plan/2018/10/waiting-times-improvement-plan/documents/waiting-times-improvement-plan/waiting-times-improvement-plan/govscot%3Adocument/00542255.pdf.

[20] SUD, A., ET AL. Effect of delays in the 2-week-wait cancer referral pathway during the covid-19 pandemic on cancer survival in the uk: a modelling study. *The Lancet Oncology*, **21** (2020), 1035. doi:10.1016/S1470-2045(20)30392-2.

[21] TENGYAO WANG, R. J. S. High dimensional change point estimation via sparse projection. *Journal of the Royal Statistical Society B*, **80** (2018), 57–83. doi:10.48550/arXiv.1606.06246.

[22] TRUONG, C., OUDRE, L., AND VAYATIS, N. Selective review of offline change point detection methods. *Signal Processing*, **167** (2020), 107299. doi:10.48550/arXiv.1801.00718.

[23] VAN DEN BURG, G. J. AND WILLIAMS, C. K. An evaluation of change point detection algorithms. *arXiv preprint*, (2020). doi:10.48550/arXiv.2003.06222.

[24] WILLAN, J., KING, A. J., JEFFERY, K., AND BIENZ, N. Challenges for nhs hospitals during covid-19 epidemic. *BMJ*, **368** (2020). doi:10.1136/bmj.m1117.