COVID-19 Incidence and Health Burden Modelling at Finer Geographies Using Online Symptoms Search Data

Masashi Asai*

MSc Data Science and Machine Learning

Supervisor: Dr. Vasileios Lampos

Submission date: 12 September 2022

Acknowledgement

I would like to thank Dr.Vasileios Lampos for his supervision in the project. It was a very tough challenge to conduct the project and write the dissertation mostly remotely since I could not stay in the UK in this summer because of my job, but thanks to the kind and detailed guidance of Dr.Lampos, I was able to carry out the research and write my master's thesis as shown here. I would like to take this opportunity to express my sincere gratitude.

I would also like to thank my employer, the Japanese government and the Ministry of Internal Affairs and Communications, for giving me the opportunity to study abroad, and my supervisors and colleagues at work for their understanding of this challenge. Lastly, I would also like to show my sincere gratitude to my family for their constant dedication and support. Without their warm support, I would not be where I am today, and I would not have been able to complete this thesis. I am grateful for all support I have had in my whole life.

Candidate Number: VRMM7

Abstract

This paper analyses how online COVID-19 related symptom search data can contribute to understanding and predicting trends of COVID-19 pandemic through the use of time- and geographically-extended datasets in the UK from Google's online database. Analysis is carried out by applying both unsupervised and supervised models to both country-level data and regional-level data. The analysis using the unsupervised model confirm that, similar to the results of previous studies, the search data correlate with and are ahead of clinical data such as the number of daily confirmed cases, indicating that the data can be used to understand infection trends. In addition, an analysis of the cross-regional relationship between search data and clinical data using regional data show that search data in the leading regions of the spread of infection has a certain pattern to infection trends in the lagging regions, and can be useful leading information. Furthermore, an advanced linear regression model with an improved training process to allow the model to cope with temporal changes in the relationships between variables and LSTM are built for forecasting. The results show that the models' predictions improved with the introduction of search data, again demonstrating the potential of search data as a leading indicator in forecasting. Finally, the performance of the LSTM on regional data with a large number of samples significantly outperformed the linear regression model, indicating the future potential of utilising the RNN architecture in this field.

Contents

1 Introduction

1.1 Motivation

Having an appropriate outlook on infectious disease incidence at an early stage is a significant issue for implementing effective public health measures and the optimal allocation of health care resources. In particular, when responding to a highly uncertain unprecedented event such as the recent COVID-19 pandemic, it is essential to identify early warning signals of the spread of infection and health care burdens from miscellaneous information. In this respect, it is a positive development that in recent years, various forms of data on people's behaviour both offline and online have been collected and made available to the public. It is hoped that these various types of information accumulated can be utilised to provide more helpful information for the future outlook of the pandemic.

In this paper, we analyse how data on people's online symptom-related searches during a pandemic can contribute to understanding and forecasting the spread of infectious diseases and health care consequences, using the most recent data since the beginning of the COVID-19 pandemic in the UK. Online search data are expected to serve as promising leading indicators of infection status and thereby to play an important role in early warning dissemination, rapid public health response and optimal allocation of healthcare resources. Indeed, several studies have already been conducted in this research area, demonstrating its potential (see next chapter for details). In this regard, it would still be meaningful to do the research based on the most recent developments, given that the pandemic started more than two years ago and more data has accumulated over that time. Furthermore, it is noteworthy that Google has started to publish anonymised online symptom-related search data of some countries. Those data allow access to a granular regional level (although there are certain restrictions on its accessibility) and open the possibility of analysing them from a new perspective and obtaining new implications in this research field. This study aims to add further contributions to the existing research by utilising this latest dataset.

1.2 Analytical Approach

The specific approach to the analysis is to build both unsupervised and supervised models, taking into account the advantages of the dataset, and to analyse in detail the relationship between online search data and key public health statistics, such as the number of new

confirmed cases and new deaths. For the unsupervised model, the relationship between an aggregated search data related to COVID-19 and the clinical data (e.g., the number of new confirmed cases and new deaths) is analysed through correlation analysis. The aggregated series of search data is derived by aggregating search data of symptoms related to COVID-19 using heuristically specified weights for them. The study will also check how the results change at different stages of the pandemic by varying the period of analysis, as there is an accumulation of data over a longer period of time. In addition, since the precedence of search data with respect to clinical data is particularly important, and therefore more detailed temporal patterns of correlation will be analysed.

While the framework for these analyses is basically targeting the entire country-level data, we also take advantage of the availability of more detailed regional data to see what patterns in the relationship between search data and clinical data emerge between regions where the pandemic trend was leading and lagging. In this respect, it would be interesting if online search trends in the leading regions have some implication for the lagging regions, then it may be possible to design and forecast prospects using search data in leading regions for having an outlook of infection trends in other regions. As we will see later, the results of the analysis need to be interpreted with a certain degree of caution due to the lack of complete data availability for regional data, but as previous studies on online data, including search data, during the pandemic have generally focused on country-level data, this study on the interrelationships between many regions within a country can provide new insights.

Analysis with supervised models is conducted from two main perspectives. One is the regression analysis with linear regression models and the assessment of the goodness of fit of the models. In other words, the impact of individual search data related to COVID-19 on the trends of COVID-19 clinical data and the explanatory power of these search data are analysed by linear regression models (with and without regularisation). This analysis allows us to assess the impact of individual search data in a multivariate manner without specifying weights in advance. In addition, as with unsupervised model, to shed light on temporal relationships between variables, the regression analysis is conducted by shifting the clinical data and see how the results change accordingly.

The second supervised modelling is to build prediction models using search data. There

have been many studies on the prediction of infectious disease outbreaks using search data, and various analyses of COVID-19 have been conducted since the beginning of the pandemic. In this analysis, while also taking into account those findings, we analyse what findings can be obtained by extending the data set over time and geographically. Specifically, firstly, based also on the results of the linear regression analysis above, a forecasting analysis using a simple linear regression model is carried out under the assumption that the relationship between search and case data is stable through time. Then, based on the results, we introduce more powerful forecasting models. One is a model that improves the training process of the linear regression model, allowing parameters to be re-trained as new data becomes available, thereby allowing them to cope with changes over time, and the other is a model with the Long Short Term Memory (LSTM), an advanced architecture of the recurrent neural network. For these models, two models that models with only autoregressive terms as features and models that also include search data are built, and their performance is compared to examine how search data contributes to prediction. In particular, neural networks, despite their high expressive power, were considered difficult to handle in small samples, such as modelling new infectious diseases, due to the risk of over-fitting. How this issue is (or is not) resolved by expanding the sample size utilising the regional data is one of the main interest of this analysis.

1.3 Structure of the Paper

The structure of this paper is as follows. First, in the next chapter (Chapter 2) introduces existing research relevant to this study including studies on statistical predictive modelling of infectious diseases, as well as trends in predictive modelling of COVID-19 pandemics and the contribution of online search record. Chapter 3 provides a detailed description of the overall methodology of this project and introduces the datasets analysed. Chapter 4 presents the results of the analysis and discusses the interpretation of the results.The limitation of this study is discussed in this chapter as well. Finally, Chapter 5 summarises the results of this analysis and discuss future remaining works, and then concludes this paper. All analysis in this paper has been done in Python utilising machine learning packages of Tensorflow/Keras and skit-learn. These codes are accessible in an auther's online repository[1]

[1]https://github.com/MasashiA/covid_modelling

2 Related Works

2.1 Overview of Forecasting Modeling of Infectious Diseases

Following the recent outbreak of an unprecedented scale of pandemic, various studies have been conducted from its inception to the present, with the aim of providing an early insight into current and future trends regarding infection and its medical consequences. The analytical approaches to forecasting related to infectious diseases can be broadly divided into two main types of analysis: analysis based on mathematical and mechanical models of the propagation process of infectious diseases (simultaneous differential equation models are commonly used), and analysis based on statistical and machine learning forecasting models. The former approach is to build a model that deterministically/stochastically describes the process of epidemic, as typified by the SIR model[1], and has a relatively long history in the field of research on predicting the trend of infectious diseases[2]. This approach has been refined and evolved over the course of its history and now forms a single research area, know as theoretical epidemiology.

On the other hand, in the field of statistics and machine learning[3], which is the subject of this analysis, various studies have been actively conducted in recent years, combining the knowledge of traditional time series analysis with the recent significant development of data science and machine learning (including deep learning) fields[4]. Forecasts of COVID-19 trends can be made for a variety of time periods, regions and indicators, and there are countless number of data that can be used as feature values in forecasting models. In this respect, this research approach is an area with potential for further development, not only through the refinement of its models and analytical methods, but also through the recent development of the collection and publication of medical statistics[5].

Although this is a diverse research area, the following section provides an overview of those that are particularly relevant to this study. First, key studies on COVID-19 forecasting models are presented, followed by an overview of disease prediction studies

[2]This area of research is said to have originated with the work of Kermack and McKendrick[1] in 1927.

[3]Its history is also said to date back to 1931[2], making it a research area with a long history similar to that of mathematical modelling.

[4]Diverse analyses have been conducted using machine learning not only to predict future infection trends, but also to improve the diagnostic accuracy of coronas, drug discovery, and so on[3].

[5]In this regard, it is interesting to note that a number of public and private research institutions successively published various statistics and forecasts as the pandemic progressed.

using online data, including search data which are conducted since before the COVID-19 pandemic.

2.2 Statistical Forecasting Models of COVID-19

Statistical modelling for predicting infectious diseases has been studied in a variety of subjects, including influenza-like illness (ILI)[4][5], Malaria[6] and Zika fever[7], in which a variety of methods have been tested. For example, for ILI which is the most widely studied subject, has been studied using traditional time series models such as the ARIMA model[8], as well as machine learning methods such as the random forest[9] and neural networks[10][11]. In addition to improvements to models, there have been active and flourishing attempts to use a variety of data for input data, including hospital visit records[12] and (as will be seen in more detail in a later section) data obtained online such as data from social media and search engines which are available relatively quickly.

Against this background, various studies have been conducted immediately after the outbreak of the pandemic in order to predict its trends using statistical models. This pandemic was characterised by a global outbreak occurring at almost the same time, and studies have been carried out in various countries and regions and inter-comparisons have been made[6]. For example, Kumar et.al.[14] and Sahai et.al.[15] used an type of ARIMA model to predict infection trends in several countries where infection was widespread at the begining of the pandemic, as well as rustam et.al.[16] applies standard machine learning models such as linear regression, lasso regression, support vector machine (SVM) and exponential smoothing to the cross-country database[7] to see the performance of each model for prediction and reports on the effectiveness of the exponential smoothing.

As for the models themselves, while taking into account the accumulation of previous research, there are also many studies that seek to actively utilise advanced models of neural networks and many of which have demonstrated that these models can show higher performance than traditional statistical time series models or simple non-linear machine learning models[17][18]. In particular, in terms of recurrent neural network (RNN), a type of the model is also used in this study, for example, Zeroual et.al.[18] has applied various RNNs, including LSTM[19] and other advanced models, to data from six countries (Italy,

[6] As a survey, Shinde et.al.[13] provides useful summary.

[7] Which is obtained from the Center for Systems Science and Engineering, Johns Hopkins University

Spain, France, China, USA and Australia) and reported their usefulness in forecasting. On the other hand, advaced RNNs and other advanced neural networks are known to be prone to overfitting due to their high expressive power[20], and it is necessary to devise the model architecture or training process to prevent overfitting as much as possible, especially in relatively small sample time series data, such as infectious diseases[18].

2.3 Statistical Forecasting Using Online Data

Records of people's behaviour on social media and the history of searches on search engines are widely used today to understand and predict various socio-economic situations, not only because they enable large-scale data to be obtained cheaply, but also because they allow monitoring people's behaviour and situations in real time[21]. For example, the use of social media and online search data to predict the popularity of certain products is a very well-known example[22]. Online data also play an important role in economic forecasting and stock price prediction todayl[23].

This is of course no exception for forecasting of infectious diseases, and efforts to utilise such online data for forecasting of trends infectious diseases have been widely conducted in the past. For example, there have been several analyses using social media such as Twitter and other social media, as typified by Lampos and Cristianini[24], most of which have reported on the usefulness of such data[25][26]. There have also been several analyses of trends of search engines, and, as with social media, their potential for understanding and predicting the spread of infection are now widely recognised[26][27][28]. For example, Polgreen et.al.[27] has analysed influenza-related search data (which are not symptoms-related search) on Yahoo to demonstrate that they are strongly related to the actual state of infection by using a linear regression models. Dugas et.al.[28] also utilised Google search data to construct a generalised linear model and a generalised ARIMA (GARIMA) model and showed that the data can contribute to influenza prediction.

Regarding the COVID-19 pandemic, several studies using online search data have already been conducted. For example, Lampos et.al.[29] demonstrats that search data on COVID-19-related symptoms correlated strongly with clinical indicators (the number of new confirmed cases and deaths) of COVID-19 by an analysis using Google search trends in eight countries in the first few months of the pandemic. They also introduced time-series forecasting models to show these symptom-related search data help to improve the predic-

tion performance of these indicators. Higgins et.al.[30] also utilised data from Google and Baidu on the earliest stages of the infection pandemic and similarly analysed correlations with confirmed cases and deaths in various countries, and reports that there are strong correlations between some COVID-19 related symptom searches and the actual medical indicators. On the other hand, the development of neural networks for the prediction of COVID-19 using search data has not been done to date, as far as it can be found.

3 Methods

This chapter describes the dataset and methodology of this study. As mentioned in the opening chapter, the aim of this research is to analyse the relationship between people's online symptom-related search trends and the COVID-19 incidence, and how the online search data can help to predict key clinical statistics of COVID-19. The methods for the analysis are broadly divided into unsupervised model analysis and supervised model analysis, and the research focuses on UK data, both at the entire country level and regional level. As the general framework of this analysis is common for both country-level data and for regional-level data, the following section will primarily focus on explaining the general framework of analysis not depending on datasets, while explaining dataset-specific analysis, dataset-specific model adjustments and dataset-specific matters to be noted as necessary. Therefore, unless otherwise stated, the analytical framework described below applies in common to both datasets.

3.1 Dataset

The data used in this analysis broadly consist of two types of data; i) COVID-19 related clinical statistics and ii) symptom-related online search data, each obtained from different online sources. Data are retrieved in a daily format, both aggregated data for a country as a whole and data subdivided by local authority level. The concept of regional division is discussed later, but as the way of division differs somewhat between the clinical dataset and the online search dataset, minimal adjustments have been made in a way that is consistent with each other.

3.1.1 Clinical Data

COVID-19 related clinical data such as the number of daily new confirmed cases were collected from a dashboard of the UK government[8] which provides a variety of COVID-19 related statistics, including the number of daily/weekly confirmed cases, deaths, testings, and vaccinations, etc. Some of these data are available not only for an entire country, but also for smaller units, such as at the national or local authority level. This analysis focused on the number of new confirmed cases, new deaths and new hospital admissions, which are general indicators of the spread of the disease and the health burdens it creates; for the country as a whole, all of these data are used, as they are all available, and for

[8]https://coronavirus.data.gov.uk/

regional data, only new confirmed cases and new deaths are adopted, as data on hospital admissions are not available for many regions. Data were collected on a daily basis from the beginning of the pandemic to 19 May 2022, the latest available date.

On the dashboard, several types of data are accessible as the indicator of the number of new cases depending on the method used to allocate the date but, for this analysis, data named "New cases by specimen date" which with dates identified by "the date the sample was taken from the person being tested."[9] were used, as it is not affected by shipping and testing procedures. The same applies to the number of new deaths and several data have been published, but it was decided to use "Deaths within 28 days of positive test by date of death" as the data are available in a wider range of regions. As for the hospital admissions, there is only one type of data available, so we adopted it[10].

Here, it should be pointed out that none of these clinical data is ideal as ground-truth data, when we apply machine learning models. In particular, it should be borne in mind that data on the number of confirmed cases may deviate to a considerable extent from the actual situation of infection in a country or region. This is due to constraints on testing capacity in a broad sense[11]. It is therefore assumed that published new confirmed cases always tended to be undercounted relative to the actual number of infections, and that the extent of this was constantly changing over time, depending on testing capacity and other factors.

The same applies to the number of hospital admissions, particularly at the beginning of the pandemic, when access to hospital facilities was partly restricted due to a lack of medical resources, and there are not always completely objective indicators or other criteria for whether a person should be admitted or not. Therefore, the number of hospital admissions is not necessarily the best grand-truth indicator, as it does not fully reflect the number of people with a certain medical condition at a particular point in time. In this respect, the number of deaths is less prone to definitional fluctuations and statistical

[9] https://coronavirus.data.gov.uk/metrics/doc/newCasesBySpecimenDate

[10] For the hospital admissions, data at the very beginning of the pandemic are not accessible, as the statistics started in late March 2020. On the other hand, because data are available for most of the period under this analysis and, as will be seen later, the data is smoothed by applying a sort of moving average before the applying the model, no special interpolations or adjustments were made for the data.

[11] If, in order to obtain completely correct ground-truth data, tests would have to be carried out every day on every citizen or at least every infected person, assuming that no false positives or false negatives would occur there, which would be unrealistic.

recording problems than other indicators, and is the most suitable as ground-truth data. However, even so, given that the duration of life extension and fatality rates fluctuate with the state of healthcare resources, developments in coping therapies and treatments, and the spread of vaccines, it might not necessarily be stable through time.

Given these issues, some researchers tried to estimate the number of infection more closely to the actual situation[**?**][31]. However, despite the issues, these publicly available clinical data contain important information on infection trends and their impact, and it remains an interesting question to investigate what patterns of relationships exist between these data and online searches. Therefore, in this study, we will use the above clinical data without any processing. In any case, this issue will be discussed again in the following chapter.

3.1.2 Online Search Data

For the online search data, data was collected from an online database provided by Google[12]. It also publishes a detailed document containing a method of collection and processing of the data[32]. The dataset covers six other English speaking countries including the United States and the United Kingdom, for the period from 2017 to the present both daily and weekly. It provides the relative frequency of Google online searches for keywords related to symptoms, medical conditions, illnesses and changes in physical condition, i.e, it is provided as a form of a proportion of the volume of searches containing a specific symptom-related keyword to the volume of all searches. There are more than 400 keywords in total covered by the dataset, and for some countries, data are available not only at the country level, but also at the level of regions in a country. For the UK, which is the focus of this study, data are available not only for the country as a whole, but also at the upper local authority level and, for some local authorities, at a more detailed local authority level (i.e., data in three levels of subgroups are available).

On the other hand, the dataset does not necessarily provide completely accurate data in their full form, and a certain amount of processing and data masking is applied due to considerations for the privacy of the users. Specifically, artificial noise is added to the individual data to avoid identification of an individual person, and daily/weekly thresholds are set for the number of searches for each keyword, and if the number of searches for a

[12]https://blog.google/technology/health/using-symptoms-search-trends-inform-covid-19-research/

keyword in a region is smaller than the threshold, the data is not provided. This means that data may not be accessible especially for some minor keywords in less populated areas[13]. Each data set is then min-max normalised between 0 and 100 for each region.

3.1.3 Selection of Regions

In selecting the regions, first, a basic list was created based on the classification of local authority according to ISO standards. Next, from that list, regions were identified for which data on both new confirmed cases and new deaths were available. Then, search data corresponding to those regions were collected from a subset of the online search database ("sub_region_1" or "sub_region_2" in the database) to construct the dataset for analysis. The regional dataset thus created contains a total of 189 regions. A list of regions is included in the appendix.

3.2 Unsupervised Model

For the unsupervised model, correlation analysis is conducted in accordance with the method of previous research[29], using a set of series of data for each clinical data and a representative series of search frequency derived by aggregating multiple search data relevant to COVID-19.

3.2.1 Correlation Analysis

3.2.1.1 Calculation of Search Score

In deriving the aggregated series of COVID-19 related online searches(hereafter we call this data series "search score") from several search data series, the following steps are taken based on the methods of the previous study[29]; i) select data corresponding to symptoms related to COVID-19 from the search dataset and determine the corresponding weights for each symptom; ii) calculate the weighted average using the weights and each search data on a daily basis to obtain an aggregated series, or the search score; iii) adjust the obtained search scores to remove noise components as far as possible. These steps are described in turn below.

First, in selecting categories of symptoms, not only must the categories be relevant to COVID-19, but they must also be given a reasonable weighting for those categories. In

[13]More details on the search data used in this analysis are reviewed in the later section.

this regard, we make use of the results of a symptom survey by the Office of National Statistics (ONS) which is conducted to COVID-19 positive persons on a regular basis since December 2020, asking about symptoms they have experienced. In the survey, COVID-19 positive persons are asked whether they have experienced any of the pre-defined categories of symptoms that are considered to be associated with COVID-19, and the results are provided in the form of a summary of the percentage of experience of each symptom. Seventeen survey results have been published as of now (the end of May 2021), and as the categories have remained constant, we adopt those categories for our analysis. For weights, it was decided to use the arithmetic mean of the percentage of symptoms experienced across all survey results[14]. The ONS symptom categories, corresponding search keywords and corresponding weights are shown in the table 1. A brief summary of these search data by dataset is also given in table in the appendix. Here, in addition to the ONS based symptom categories and weights, we also introduce symptom categories and weights based on reported results from clinical studies at the beginning of the pandemic[33] (so-called FF100), which were used by the previous study[29]], to check robustness.

Table 1: Symptom category and correponding search keywords and probabilities

Symptom Category	**Search Keyword**	**Percentage**
Cough	Cough	0.418
Headache	Headache	0.376
Fatigue weakness	Fatigue, Weakness	0.373
Sore throat	Sore throat	0.321
Muscle ache myalgia	Myalgia	0.257
Fever	Fever	0.244
Loss of taste	Ageusia	0.188
Loss of smell	Anosmia	0.180
Shortness of breath	Shortness of breath	0.143
Nausea vomiting	Nausea, Vomiting	0.095
Abdominal pain	Abdominal pain	0.073
Diarrhea	Diarrhea	0.068

Note: symptom categories are from ONS survey and all corresponding keywords for each category are listed, separated by commas.

[14]For the weights, some symptoms have a variation in the proportion of experience over time, but it was decided to assume a constant value throughout the period for simplicity of the model.

Next, based on each series of symptom searches and weights thus obtained, a daily weighted average is calculated, after taking a moving average for each series over the past seven days, including the present day, with the aim of smoothing out the day-to-day variation (variation from the day of the week). Here, for search data that belongs to the same symptom categories in the ONS survey, the respective series are added together before the moving average is taken. For the regional data, some of the retrieved data are not available due to processing applied for privacy reasons, as discussed in the dataset section above. All of those data are padded with zeros before taking the moving average.

3.2.1.2 Mitigation of Media Effect

The search score thus obtained, however, is considered to include not only the searches done by actual infected people, but also those from information gathering not attributable to the infection itself. In this respect, the latter is considered to be noise when capturing the actual clinical situation from the scores. Therefore, adjustments should be made to the series to remove as much of the latter noise as possible. Here, based on a method in the previous research[29], we introduce a simple time-series model to detect and remove the impact coming from media coverage[15].

Firstly, we set a fair assumption on the online symptoms search as the equation 1 where g represents the search score derived above which is corresponding to COVID-19 related symptom search frequency, and g can be interpreted to consists of searched caused by infections (g_p) and searched caused by concern or information gathering (g_c), so

$$g = g_p + g_c \tag{1}$$

Then, the equation can be simplified by introducing a constant value $\gamma \in \{0, 1\}$ such that $g_p = \gamma g$ and $g_c = (1 - \gamma)g$. In this way, our problem is embodied in the question of how to approximate γ. Here we have daily data on how much proportion of news articles in the UK are about COVID-19[16] so it can be represented by $m \in \{0, 1\}$ at any given date. By using this m, we formulate two time-series models to approximate γ. The first model is a

[15]The following explanation is largely based on the description of a method from the previous study[29]

[16]This data was obtained from Media Cloud, as in the previous study

AR model with I lag:

$$\arg\min_{\mathbf{w},b_1} \frac{1}{N}\sum_{t=1}^{N}(g_t - \sum_{i=1}^{I} w_i g_{t-i} - b_1)^2 \tag{2}$$

and the second model is a modified AR model added by J lag m including the present day's data:

$$\arg\min_{\mathbf{w},\mathbf{v},b_2} \frac{1}{N}\sum_{t=1}^{N}(g_t - \sum_{i=1}^{I} w_i g_{t-i} - \sum_{j=0}^{J} v_j m_{t-j} - b_2)^2 \tag{3}$$

By comparing the performance of these two models, we can see how much impact the proportion of media articles has to predict the search at each time t[17]. For the specification of the models, to avoid complexity $I = J = 3$ is employed.

First, each model is trained by the data of the previous N days (N is set to 90 due to the sample size being more generous than in previous studies), and then the next day's data is predicted and the absolute error between the predicted value and the real data is calculated. This process is repeated for all days by moving the training data and the prediction target forward by one day each time. By doing so, the absolute error of each model can be obtained for all days of forecasting. Here, we denote the error of the first model at any time as ϵ_1 and the error of the second model as ϵ_2. If $\epsilon_1 > \epsilon_2$, then media article data (m) is considered to contribute to the improvement of the model's predictions, and vice versa. Based on this reasoning, when $\epsilon_2/\epsilon_1 > 0$ we set γ as 1 because m has no impact, otherwise we set γ as ϵ_2/ϵ_1. Finally, the series of γ is smoothed by applying a moving average over the past seven days, including the present day.

3.2.1.3 Division of Period

In conducting correlation analyses, it is also necessary to take into account that the relationship between the search score and clinical data changes over time (i.e., it varies according to the stage of the pandemic). For example, it has been noted that the actual number of new cases was not adequately captured due to the lack of testing capacity, and it is also assumed that there was a larger gap in the time between infection, onset and confirmation of infection for some time at the beginning of the pandemic[18]. With regard

[17] This is based on the concept of the Granger causality test.

[18] Statistics on the government's dashboard show that the number of the conducted test was reached to the limit of the testing capacity during the second wave of the pandemic

to the number of deaths, it is expected that at the beginning of the pandemic, the medical system and resources were not sufficiently arranged for this unprecedented event and there were no established methods of treatment as well, so the fatality rate was likely to have been higher at that time. Similarly, for the number of the hospital admissions, it is expected that the lead time to admission to hospital would have been longer during the period when medical resources were more constrained. Furthermore, more generally, it can also be expected that at the beginning of the pandemic, the uncertainty of the phenomenon led to a more proactive attitude towards information gathering by those suspected to be infected and more pronounced search behaviour for COVID-19-related symptoms online, while in the later stages of the pandemic, which were more recent, the relationship somewhat diminished, partly because the intrinsic uncertainty had also diminished[19]. In light of this issue, the correlation analysis will also be carried out by splitting the period of time of the dataset, depending on the stage of progression of the pandemic, so that it captures changes through time.

Although there is no uniquely determinable method for dividing the period for the analysis, for the purpose of ensuring a sufficient sample size and avoiding undue complexity, it was decided to split out the period corresponding to the first wave of the pandemic, which is assumed to have been the most uncertain and for which testing and medical resources were not sufficient, as the first period[20]. As certain definition is given for the timing of the first wave in ONS documents[21], with it ended at the end of May 2020, 30 June, exactly one month after the date was chosen as the cut-off date. The remaining period was then divided into two periods, based on the early spring of 2021, when the outbreak had temporarily settled down (in particular, the date of the split is set to 31 March). This period also coincides with the time when the second vaccine dose coverage has just started to increase rapidly, so it can be expected to capture the changes caused by this factor. The start of the analysis period was set for 13 February 2020, 30 days back from the date when the cumulative number of confirmed cases for the UK exceeded 1,000, taking into account the possibility that in the initial stages of the pandemic, the small number of actual cases

[19]Also, as a more general thought, the longer the observation period, the more likely the situation of other infectious diseases and other epidemics will become noise of symptom searches.

[20]Many existing researches have generally covered the period of the first wave, which has the advantage of making comparisons easier[29][30]

[21]https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/conditionsanddiseases/articles/coronaviruscovid19infectionsurveytechnicalarticle/wavesandlagsofcovid19inenglandjune2021#waves-of-covid-19

themselves may not have appeared as a significant change in online searches. The figure **??** shows visual images of period division of each clinical indicators.

3.2.1.4 Model Application

Given the above preparation, the correlation between the respective daily clinical data (confirmed cases, deaths and hospital admissions) and the daily search scores will be calculated to see how well they are linked. For the clinical data, a moving average is taken over the past seven days, including the present day, with the aim of smoothing out the day-to-day variation (variation from the day of the week). Also, in order to clarify their temporal structure, the clinical data will be shifted forward and backwards over a certain range of days to see how their correlations change according to the shift. As is intuitively understandable, if online search behaviour is linked to clinical data, it is probable that it occurs prior to the clinical statistics. This is because it is assumed that things will proceed in the following order: after some change in physical condition, searches are done, followed by a certain period of time, testing and confirmation of infection, (and hospitalisation depending on the condition, etc.)[22]. Indeed, the fact that online search precedes clinical statistics has been confirmed in the results of previous studies of ILI and COVID-19 outbreak[34][29].

3.2.2 Regional Analysis

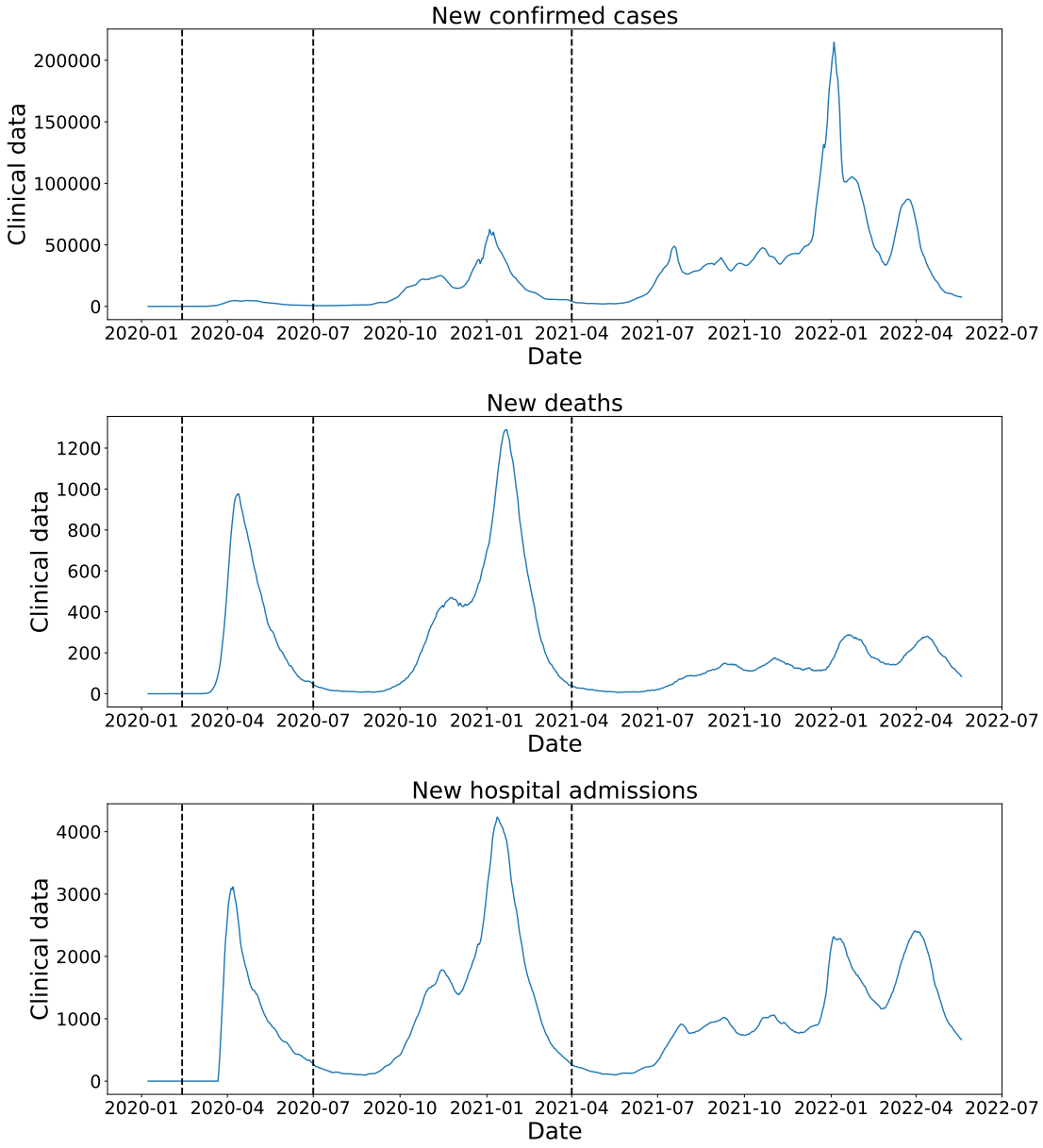
For the regional data, the correlation analysis described above is carried out as the first analysis. In other words, using the dataset from the clinical and search data created for each region, search scores are derived for each region, and using the daily percentage of news articles (albeit only the county level data is available), the mitigation of media effects which is the same as above, is applied on the search scores for each region to calculate adjusted search scores. In this way, the same correlation analysis can be performed for each region.

3.2.2.1 Interregional Correlation Analysis

Next, the regional analysis is taken further to analyse patterns in the inter-regional relationship between online searches and clinical data among regions that are ahead of and behind

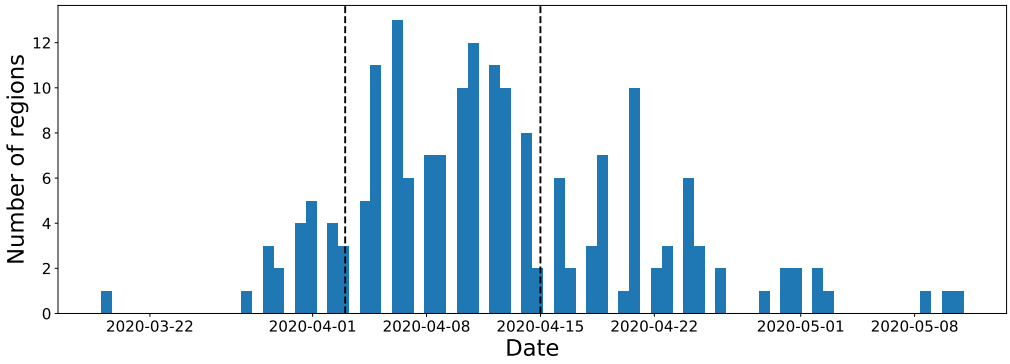
[22]However, this relationship may not always be clearly observable, as symptoms can persist even after confirmation of infection and therefore a certain persistence in search behaviour is assumed. It should also be noted that, given that there is a certain interval between infection and disease onset, this relationship is reversed in the case of pre-onset detection of infection by testing due to contact with an infected person.

Figure 1: Diagram of division of period



infection trends, utilising the framework of the above correlation analysis. Specifically, in this analysis, particular attention will be paid to the first wave of the pandemic, where temporal differences in the spread of infection between regions are easier to detect, to see what temporal relationships exist between search data and clinical data across groups of regions classified according to chronological differences in the spread of infection. Although there are various possible approaches to regional classification, we have decided to keep it simple and classify the regions according to when the ratio of cumulative cases to the population exceeded a certain threshold[23]. The figure 2 shows a histogram created based on the dates identified when the cumulative number of confirmed cases exceeded one per 10,000 population in each of the 189 regions in our dataset (The histograms are cut off at 15 May 2020 for reasons of visibility).

Figure 2: Histograms according to the progression of infection



This shows that the pattern of infection spread (expansion in the number of regions above the threshold) appears to be somewhat smooth, but there are apparent timing for the trend to drop at 3th and 15th of April (vertical lines are depicted in the graph)[24]. In addition, As the temporal proximity of the infection situation in the analysed areas is important for this analysis, areas that exceeded the threshold after 15 May 2020 were excluded from the analysis as outliers[25]. Thus, although somewhat ad hoc, the regions are

[23]As discussed above, data on the number of new cases has a potential problem as an objective indicator, but as it is used in this case for comparison within a single country, this problem should be mitigated to some extent.

[24]Although this regional classification is a sort of retrospective in the sense that the data actually cannot be observed in real time, we have decided to use this simple method as a starting point of the future work.

[25]Histograms covering all time periods are presented in the appendix.

divided into a total of three groups based on these two points[26].

Based on this preparation, the search scores and clinical data for each region are used to analyse the pattern of correlations between online searches and spread of infection between three regions with different infection trends over time. In other words, the correlation between the search scores of each region of the leading group and the clinical data of each region of the lagging group will be calculated and analysed to see what kind of relationship can be observed[27]. We will also see the temporal structure of the correlations by shifting the clinical data in line with the previous analyses. If the analysis confirms that online search trends in the leading regions show a significant pattern with respect to the lagging regions, it will indicate that search data can be important information not only within a particular region, but also for having an outlook of infection trends in other regions. This could provide the basis for further research.

3.3 Supervised Model

In addition to the unsupervised models described above, supervised machine learning models will be built to provide a more detailed analysis of the relationship between online search data and clinical data and its contribution to predictions. The analysis with supervised models can be broadly divided into two analyses: i) analysing the relationship between search data and clinical data using linear regression models, and ii) building machine learning prediction models utilizing search data and analysing the performance of the models.

3.3.1 Linear Regression Analysis (Goodness-of-Fit Analysis)

The unsupervised model analysis above aims to examine what patterns of correlation online search behaviour exhibit with key clinical indicators. However, due to the nature of unsupervised models, the weights for calculating search scores need to be determined heuristically in advance, and the weights may not perfectly reflect the real impact of each symptom-related search. To address this issue, a linear regression model is applied. From

[26]The numbers of regions in each group are then 23, 102, and 56, respectively.

[27]The analysis is done by averaging out all correlations within a group. For example, if the number of regions in the leading group is $n1$ and the number of regions in the lagging group is $n2$, a total of $n1 \times n2$ correlations will be obtained, and by calculating the average (and the standard deviation) of these $n1 \times n2$ correlations, the relationship between groups is observed.

the regression analysis, it is possible to verify the extent to which the respective search data are more related to the clinical data.

The data to be analysed is entire country data and two types of linear regression models are introduced. The first is a simple classical linear regression model. The input data are the daily original unaggregated search frequency data corresponding to the ONS stat's symptom category and news ratio, and the ground-truth data are the corresponding daily data of cases, deaths, and hospital admissions. That is, the input data is described as $\mathbf{X} \in \mathbb{R}_{+}^{T\times(S+1)}$ where T is the number of days in the input and S is the number of symptoms which is added by a single series of the news ratio. In this case, $T =$ and $S = 15$. The corresponding clinical series of daily new cases, new deaths, and new admissions are $\mathbf{y}_c \in \mathbb{R}_{+}^{T}$. Therefore the model solves the following minimisation task, where $b \in \mathbb{R}$ is a corresponding bias term and $\mathbf{w} \in \mathbb{R}^{S+1}$ is a corresponding weight vector The model is descrived as:

$$\arg\min_{\mathbf{w},b} \|\mathbf{y}_c - \mathbf{w}\mathbf{X} - b\|_2^2 \tag{4}$$

The second model is a regularised regression model introducing regularisation terms to the above simple regression model in order to mitigate the over-fitting. Specifically, we introduce the elastic net[35] that has both L1 and L2 regularisers, and which can be interpreted as a generalisation of Lasso and Ridge regressions (when the hyper-parameters are appropriately tuned). In general, it is known to out-perform Lasso and Ridge regressions in many cases by managing that L1 and L2 regularisation terms parallelly. Input and output data are the same as in the regular linear regression model above. Then, the optimisation problem is characterised as:

$$\arg\min_{\mathbf{w},b} \ (\|\mathbf{y}_c - \mathbf{w}\mathbf{X} - b\|_2^2 + \lambda_1\|\mathbf{w}\|_1 + \lambda_2\|\mathbf{w}\|_2^2) \tag{5}$$

where λ_1 and $\lambda_2 \in \mathbb{R}_{++}$ are L1 and L2 regularisation parameters. For the hyper-parameter tuning of the elastic net model (i.e., λ_1 and λ_2), candidate values are chosen in a certain range of numbers for the sum of the two regularisation parameters and the ratio of each, and cross-validation is conducted in the form of a grid search to select the best performing model parameters.

Before applying each model, as in the unsupervised model, a moving average is taken for both the explanatory and explained variables over the past seven days, including the

present day. In addition, every variable is normalised between 0 and 1 range by the min-imax normalisation[28]. These models are applied to the data for the entire period and periods split in the same manner as in the unsupervised analysis to obtain the coefficients (weights) of the respective search data, news data, and intercept. Then, the goodness-of-fits of the models are checked by adjusted R-squared. In addition, as with the unsupervised model, the temporal relationship of the search data with the clinical indicators shall also be analysed in terms of how the goodness-of-fit changes as the respective clinical data are shifted forwards and backwards.

3.3.2 Forecasting Models

3.3.2.1 Naive Forecasting Model

For the forecasting model, we begin the analysis with a simple model with very naive assumptions. The model's setting is the same as the elastic net introduced above (in eq.5), and the data is split into training and test data to check the performance of the model. The division of the data is set to be the same as the division of the three periods defined in the unsupervised model's section. We test a model that is trained with the data in the first period and predicts data in the second period, and a model that is trained with the data in the first and second periods and predicts data in the third period[29]. The hyper-parameters of the elastic net are identified by cross-validation during the training/validation phase, as in the goodness-of-fit analysis above, and the performance of the model is assessed by mean absolute error (MAE) and correlation. Note that the predictions of the linear regression model can be negative values, so if a negative value is predicted, the prediction is replaced by zero before the model is evaluated[30].

Given the possibility that the relationship between clinical data and online search activity may not be constant over time, and that the number of infections may be biased due to changes in testing capacity, and that incidence and the fatality rate may change

[28]By mapping the features to positive values, it is easier to understand the impact of each feature on the clinical data from the sign of the regression coefficient.

[29]That is, the first model trains on data from 13 February 2020 to 15 July 2020 and predicts data from 16 July 2020 to 31 March 2021, while the second model trains on data from 13 February 2020 to 31 March 2021 and predicts data from 1 April 2021 to 19 May 2022.

[30]It could be considered other methods such as: taking logs for both explanatory and explained variables (so-called logit regression)[34] to ensure that the predictive value is not negative, but the range of clinical data in this analysis is wide and that model was found not to work well. Therefore, it was decided to use the simpler method described above.

with the spread of vaccines, the model, which assumes constant parameters over time, may remain too naive with very poor prediction performance. However, it is still an interesting starting point for identifying the contribution of search data and the changes depending on the stage of the pandemic.

3.3.2.2 Linear Retraining Model

Using the above model as a starting point, here, we build more practical models. First, we introduce a linear regression model with a retraining feature in order to deal with inter-temporal changes in the relationships between clinical data and search data. The basis of the model is the elastic net, like above, but the training process is improved to retrain the model each time new data is obtained, so that the model can better adjust changes through time. In other words, every period, the model is trained on the latest chunk of data available at that point in time, and the model makes predictions on current or future clinical data[31].

In building the model, it is assumed that there is a time lag in the latest available clinical data (prediction target), which is a more realistic assumption for practical disease forecasting models. In order to quantitatively verify how search data contributes to prediction, the model is constructed as both an autoregressive type model that includes previous values of its target value but not search data as feature values, and a model in which search and news data are added as features to this autoregressive model. The structure of the autoregressive type model at any given time t is described as below:

$$\hat{y}_{t+F} = \hat{a}_t + \sum_{p=0}^{P} \hat{w}_p^t y_{t-L-p} \tag{6}$$

where $\hat{y}_t$ is a scalar prediction of clinical data at time t, F represents how far into the future the forecast is[32], $\hat{a}_t$ is a bias term at time t estimated from the training, P is a depth of auto-regressive lag, $\hat{w}_t^p$ is a trained weight of p lagged input at time t [33], y_t is an observable input clinical data, and L is the lag based on the constraints of thr real-time availability

[31] Therefore, the former case corresponds to nowcastig and the latter case corresponds to forecasting.

[32] For example, when $F = 0$, the model gives nowcasting and when $F = 10$, it predicts 10 days future value.

[33] This means t is a superscript, not a power term.

of clinical data. Similarly, we formulate the second model structure as equation:

$$\hat{y}_{t+F} = \hat{b}_t + \sum_{p=0}^{P} \hat{w}_p^t y_{t-L-p} + \sum_{q=0}^{Q} \hat{\mathbf{h}}_p^{t\top} \mathbf{x}_{t-p} \tag{7}$$

where $\hat{b}_t$ and $\hat{\mathbf{h}}_p^t \in \mathbb{R}^{S+1}$ are estimated intercept and weight vector for p lagged input at time t respectively, and $\mathbf{x}_{t-p} \in \mathbb{R}_+^{S+1}$ is a vector of p lagged search data and new data at time t[34].

For the analysis of regional data, the dimensions of the input and output data are expanded by the number of regions and are trained and predicted with a single model at a time[35]. These models are trained with the most recently available N days data, and forecast the target value each day. In this analysis, N is set to 10 and, L is set to 7. The autoregressive lags are set to 2 (i.e., $P = Q = 2$) for the purpose of avoiding model complexity. For the evaluation of model performance, MAE and the correlation coefficient are used as in the naive model. The treatment for negative predictions was the same as for the naive model above, which is a simple substitution by zero.

In addition to this model, a simple persistent model is also introduced as a baseline which gives exactly the same value as the most recently available clinical data as a prediction[36] (i.e., $\hat{y}_{t+F} = y_{t-L}$).

3.3.2.3 Recurrent Neural Network (LSTM)

Another analysis is also carried out using LSTM, an advanced RNN, which is designed to learn both long-term and short-term relationships in a time series data in a more sophisticated manner. In particular, for data such as infectious diseases, which are persistent yet fluctuating in a complex way, LSTM is expected to learn complex non-linear dependent relationships between variables, which cannot be learnt by normal linear models. On the other hand, it is generally known that neural networks are at risk of over-fitting due to their high expressive power and would not perform well when the number of samples is small. In this respect, the sample size would not be large enough for the country-level data,

[34]It is clear from the setting of the model that it is assumed that search data is available in real time.

[35]Another option would be to train as many models as the number of regions, but taking into account its high computational costs and from the point of view of comparison with the neural networks introduced later, it was decided to carry out the analysis with a single model.

[36]Although the model has a very simple structure, it is a suitable baseline, as it is not uncommon for these simple models to perform better than complex predictive models in time series data with persistent nature, such as data on infectious diseases.

but for regional data, since data from 189 regions can be used in parallel, it is interesting to see how this sample size expansion enhances the performance of the model.

In training the neural network, the input data is made as similar as possible to the linear regression model above in order to compare the models' performance. Therefore, two types of models are prepared: a model consisting solely of autoregressive features with two periods of lags, and a model with autoregressive and search data (and news ratios) with two periods of lags. The models are trained with a past 10-day chunk of data at once. In order to control the model complexity, the model consists of only two layers: one LSTM layer and one dense layer[37]. Models are trained for 30 epochs each, during which the accuracy is checked each time using validation data which is randomly chosen from the training data, and the model with the best performance is adopted. The models are tuned for the number of output dimensions in the LSTM layer and the percentage of drop-outs, and the combination with the best performance on the validation data is chosen. The activation function is Relu, the optimiser is set to Adam, and the loss metric is mean squared error.

LSTM differs from the linear retraining model described in that it is not a model that makes predictions on a time-to-time basis, so it is necessary to split the training and test data in order to verify the performance of the model. In this analysis, in order to ensure sufficient training samples, the test data was set to the last six months of the data[38] and the model was trained with all data up to that point. In terms of model comparison, performance metrics are calculated for the linear retraining model and the persistent model for the same time period[39]. As before, the performance metrics are MAE and correlation. The treatment for negative predictions was the same as for the naive model and the linear retraining model.

[37]It is considered to be composed of three layers if a drop-out feature which is introduced to mitigate over-fitting is counted as one layer.

[38]Therefore, it is from 20th November 2021 to 19th May 2022.

[39]To see the robustness of the results, we also evaluate the models with different test data.

4 Results and Discussion

4.1 Unsupervised Model

4.1.1 Country Analysis

We start by looking at the results of the unsupervised model for the country data. The results are shown in the table **??**. Overall, it can be seen that the search scores and the respective clinical data have a certain degree of correlation, although there is some variation depending on the clinical data and the period covered. For each clinical data, the correlations (maximum correlations) for the number of new deaths and new hospital admissions, which are expected to have relatively reliable as ground-truth data, are relatively higher than for the number of new confirmed cases, and is an expected result. Looking at the temporal structure of the search score and clinical data (results with shifting clinical data), the correlation is maximised when the clinical data is shifted a certain degree into the past[40], both with and without adjustment of media effects, suggesting that the search score emerges in advance to the clinical data[41]. This is consistent with the intuition discussed in the previous chapter and the results of previous studies[29][30]. On the other hand, while the degree of precedence is expected to increase in the order of the number of new cases, new hospitalisations, and new deaths, the results are not very consistent with this expectation. That is, such a relationship is observed relatively clearly between the number of deaths and hospital admissions, but only a very vague and even opposite relationship is observed for the number of new cases. It can be pointed out that this may stem from the deficiencies as ground-truth data that the number of new cases inherently suffers from, as pointed out above.

[40]The clinical data is shifted to maximise the correlation within a range of 40 days in the past and 10 days in the future (range of 50 days in total), so the maximum correlation, in this case, means maximum within the range of this 50 days (not the global maximum).

[41]Only for new admissions in the third period, the clinical data is ahead of the search score (the none-adjusted one only), but it should be noted that the data on hospital admissions are based on reporting dates, so they tend to lag slightly behind the actual situation in nature.

Table 2: Correlation analysis: Country data

	Correl. without shifting		**Max correl. with shifting**	
	w/o media adj.	w/ media adj.	w/o media adj.	w/ media adj.
New Cases				
entire period	0.227	0.229	0.228 [-3]	0.247 [-40]
13/02/2020 - 30/06/2020	-0.191	-0.207	0.835 [-35]	0.854 [-35]
01/07/2020 - 31/03/2021	0.288	0.417	0.310 [-3]	0.451 [-21]
01/04/2021 - 19/05/2022	0.349	0.307	0.351 [-2]	0.347 [-40]
New Deaths				
entire period	-0.095	-0.106	0.325 [-26]	0.320 [-27]
13/02/2020 - 30/06/2020	-0.150	-0.164	0.883 [-26]	0.886 [-26]
01/07/2020 - 31/03/2021	0.121	0.230	0.227 [-32]	0.399 [-34]
01/04/2021 - 19/05/2022	0.631	0.554	0.669 [-13]	0.601 [-16]
New Admissions				
entire period	0.065	0.050	0.365 [-19]	0.358 [-20]
13/02/2020 - 30/06/2020	-0.093	-0.103	0.843 [-20]	0.835 [-20]
01/07/2020 - 31/03/2021	0.193	0.316	0.274 [-26]	0.434 [-29]
01/04/2021 - 19/05/2022	0.598	0.504	0.600 [+2]	0.509 [-6]

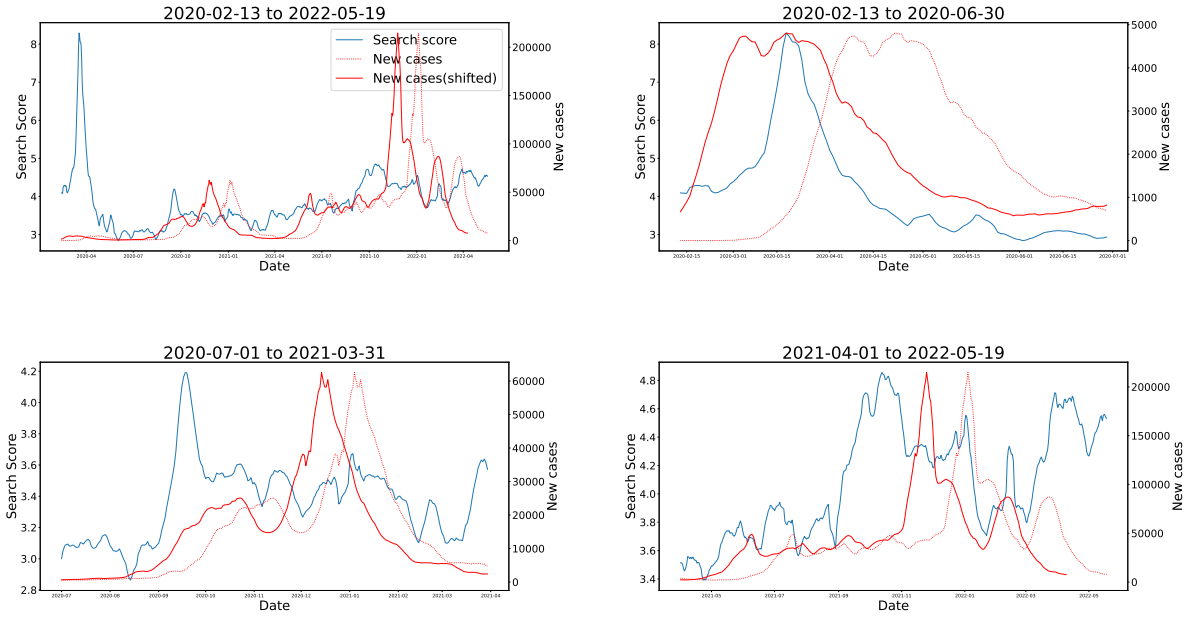
Note: correlations between the search scores and the respective clinical data by time period. The left column shows the correlation coefficients calculated without shifting the clinical data, while the right column shows the results of shifting the clinical data to maximise the correlation within a range of 40 days in the past and 10 days in the future (range of 50 days in total). The numbers in brackets correspond to the number of days shifted, with negative (positive) values indicating that the clinical data has been shifted into the past (future), meaning that the data lags (precedes) the search score in terms of correlation.

The results are also somewhat difficult to interpret in terms of the effect of mitigating the impact of media coverage. In the second period, there is generally a considerable effect of improving the correlation, and improvement in the correlation for the number of new confirmed cases observed, but for other periods and for two other indicators, there is a little effect or, conversely, a slight decrease in the correlation for some of them. Furthermore, with regard to the temporal structure a reasonable temporal structure is maintained for new deaths and new hospital admissions but, for the number of cases, the shifting days lag further behind the other two and do not show a reasonable temporal relationship. This result is somewhat contradictory to the finding of the previous study, but it can be possible that the search score used in this analysis only aggregated only from symptom-related search keywords and, unlike the previous study[29],

does not include common search terms such as "covid", which are considered more sensitive to the media article[36], and thus the impact of adjusting media effects may not be as clear as in the study[42].

For the results by periods, from the figures 3-5, it appears that as time passes since the start of the pandemic, there is a gradual loss of a clear relationship between the two series. However, this trend is not intuitively surprising, given that once some time has passed since the start of the pandemic and uncertainty has decreased, search behaviour may become more moderate and the effects of other factors may become more pronounced.

Figure 3: Search score and new confirmed cases

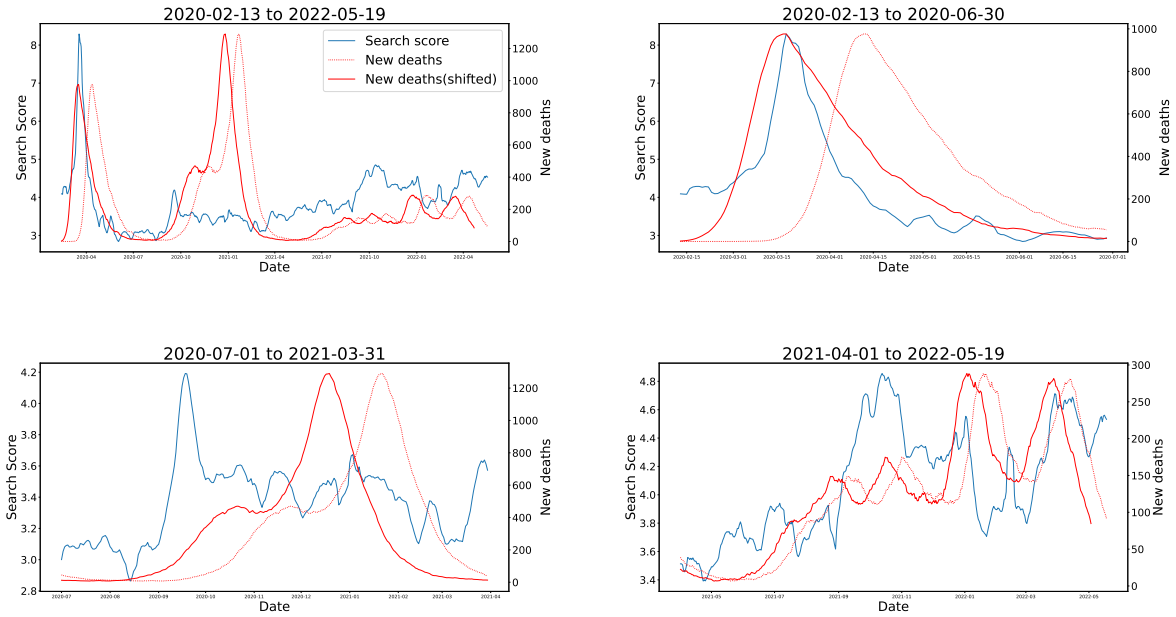


Note: the blue line, red line, and dotted red line represent the adjusted search score, shifted clinical data (with the maximum correlation), and original clinical data, respectively. The top left shows correlations of the entire period, the top right shows the period1 (13/02/2020-30/06/2020), the bottom left shows the period2 (01/07/2020-31/03/2021), and the bottom left shows the period3 (01/04/2021-19/05/2022).

Indeed, the correlation in the first period, which is the beginning of the pandemic, is quite high, and the correlation is smaller in period 2 and 3. Regarding this, it is somewhat puzzling that the correlation is particularly low in the second period for new admissions and new deaths, which are considered relatively reliable as grand-truth data, unlike the number of new cases, which is

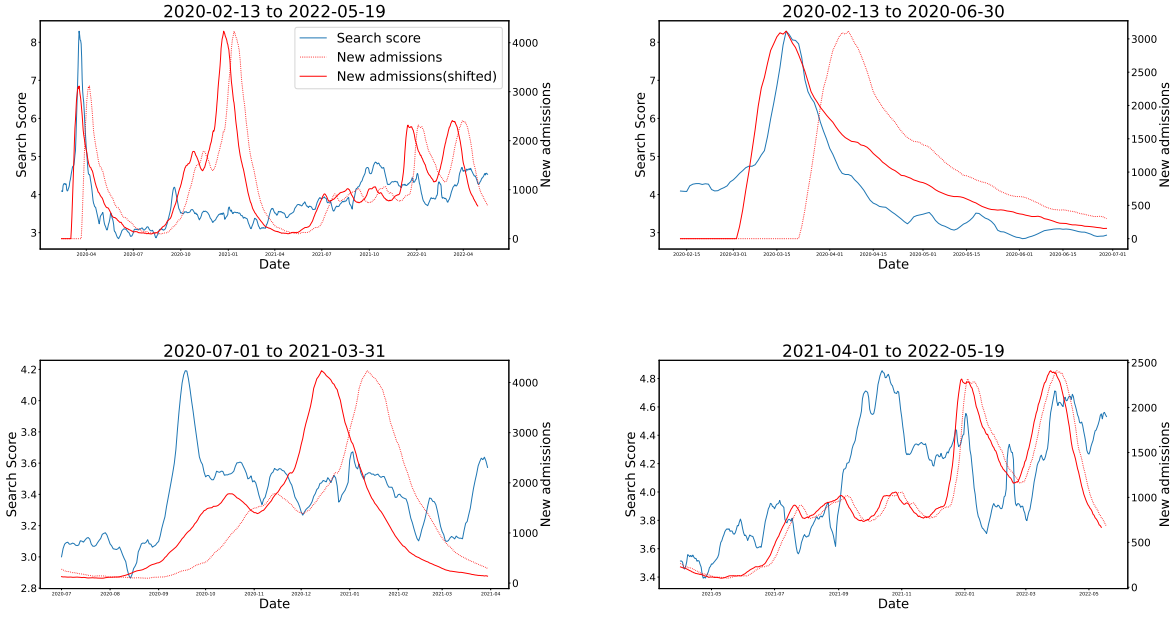
[42]There may also be room for the future work to optimise the structure (e.g., training period and lags) of the time-series model for measuring γ in media effects to observe changes in results.

Figure 4: Search score and new deaths



Note: see the note in the figure 3

Figure 5: Search score and admissions



Note: see the note in the figure 3

largely influenced by the limitations of testing capacity[43]. In particular, in the second phase, as can be seen from the bottom left graphs in the figure 3-5, there is even a pattern where search scores are inversely moving with clinical data. It is not easy to give a clear interpretation of this result, but this may be due to the limitations of the unsupervised model. For instance, there is no guarantee that the symptom categories and weights selected for calculating the search score are optimal in capturing the trends of COVID-19 incidence. For instance, Common symptoms such as fatigue, fever, or muscle ache are reported as the main symptoms of COVID-19 in both the ONS survey and FF100, but these are fairly common symptoms that may arise from other diseases and other lifestyle factors, would generate a bias in the search score. In addition, the assumption that they do not change over time may be too strong. Even if the choice of symptom categories and weights is correct at a given point in time, if it is not constant throughout the analysis period, there will be periods when the correct results are obtained and periods when they are not. These points will be discussed again in the light of the results of the supervised model later on. Very similar results were also confirmed in the analysis of the FF100 based symptom categories and weights, which was conducted to check the robustness of the results (see the appendix).

This is followed by a more detailed look at the temporal relationship between the correlation of individual search data and clinical data. The figures 6-8 depict the temporal relationships between clinical data and each search data that consists of the search score in more detail. The vertical axis represents the correlation coefficient, while the horizontal axis represents how much the clinical data is shifted in the range of 40 days past and 10 days future. In other words, A negative (positive) horizontal axis means that the clinical data is shifted into the past (future), and the vertical axis at that location represents the correlation between the shifted clinical and search data. The markers on each line indicate which point is the point of maximum correlation for each series[44].

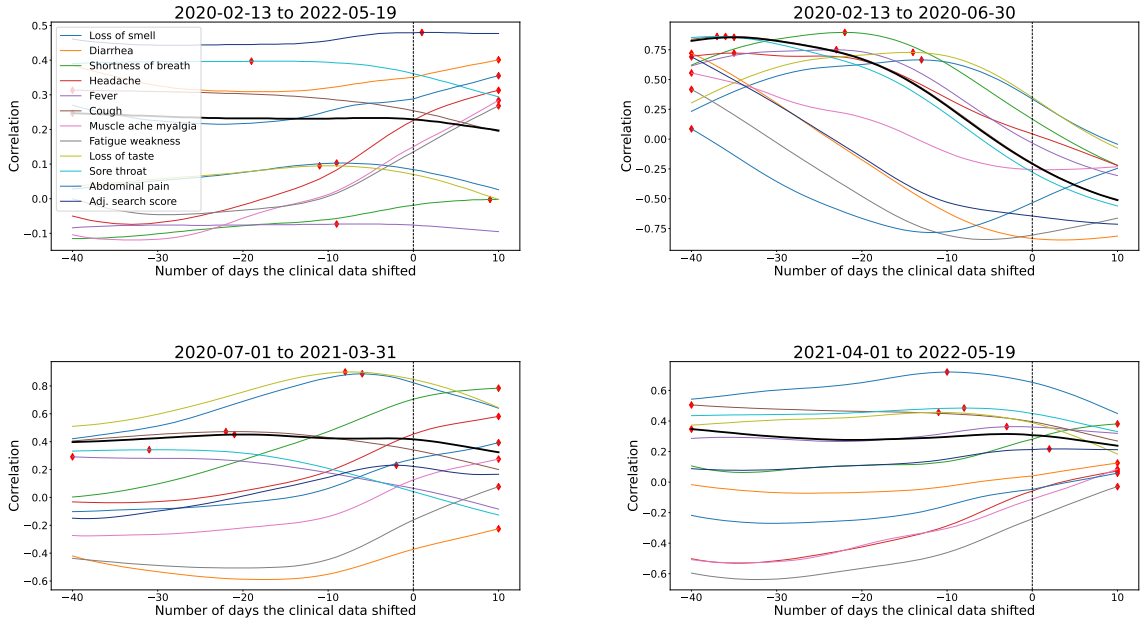
These show that the temporal structure with the clinical data differs from one individual search data, and that the pattern is not stable over time[45]. In terms of a comparison of the results for the entire period (top left in the figures) per data, it can be said that for the number of new deaths and new hospital admissions, many of these search data can be a useful leading indicator, but this is not necessarily the case for the number of new cases. This, again, illustrates

[43]Even so, it has been pointed out above that none of them is entirely reliable as the ground-truth.

[44]For example, for the search score (black line) in top left graph in the figure 7, it can be read that it is maximum at the point when the data is shifted 26 days in the past (as x-axis indicates minus 26).

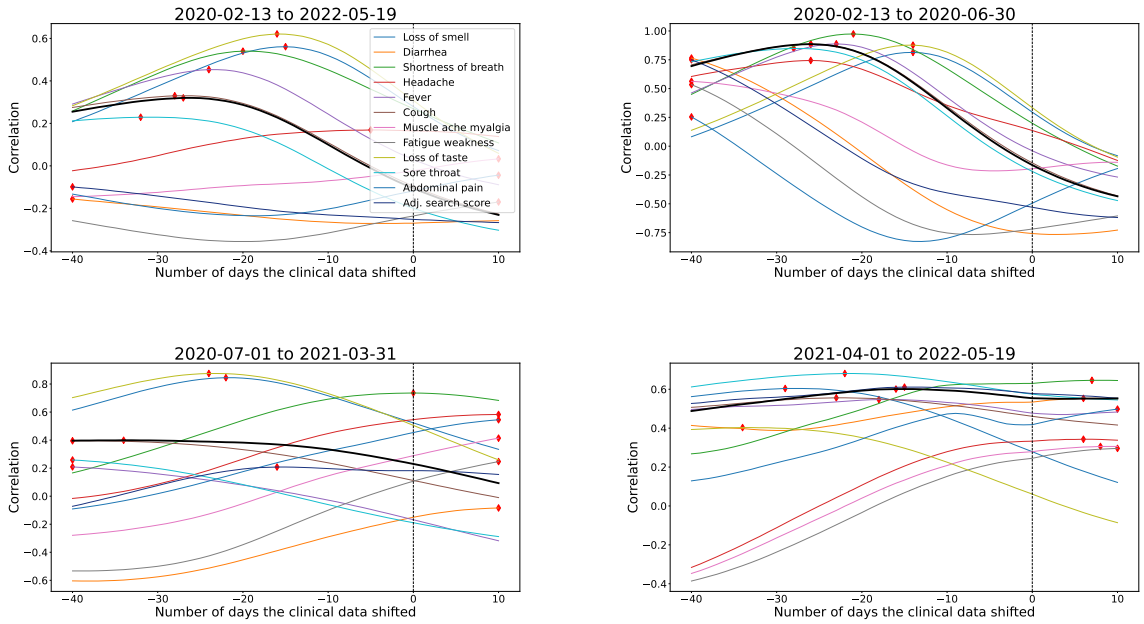
[45]Only loss of smell and loss of taste have relatively high correlations and stable precedence in all divided periods.

Figure 6: Correlation between new confirmed cases and each clinical data



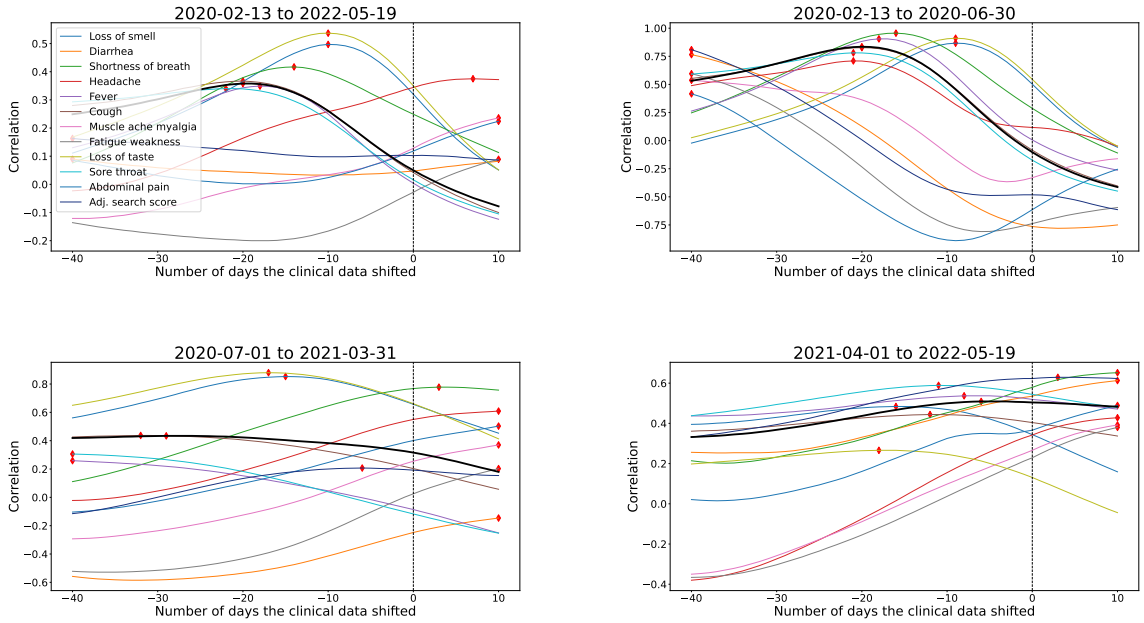
Note: graphical representation of the correlation between each search data and clinical data according to the number of days the case data was shifted. The top left shows correlations of the entire period, the top right shows the period1 (13/02/2020-30/06/2020), the bottom left shows the period2 (01/07/2020-31/03/2021), and the bottom left shows the period3 (01/04/2021-19/05/2022). If the x-axis takes a negative (positive) value, the clinical data has shifted into the past (future). The red markers represent the maximum correlation of the respective data.

Figure 7: Correlation between new deaths and each clinical data



Note: see the note in figure 6

Figure 8: Correlation between new admissions and each clinical data



Note: see the note in the figure 6

the difficulty of handling data on the number of new cases. In terms of time periods, many of the search data are reasonably ahead in the first period[46], but the relationship becomes blurred as it moves into the second and third periods. In particular, in the second period, many of the data show that the period of the shift with the highest correlation is stuck at both ends of the graph, indicating that it does not provide much useful information. Similar results were broadly confirmed in the analysis by the aggregated search score above, but when looking at individual symptom search data, it can be seen in a clearer way that some symptom searches contain useful information and others do not, and that they vary from period to period.

More to the point, this change over time may not only be due to factors such as other infections, but may also imply that the patterns of useful signals coming from each symptom have changed over time, as mentioned above. In particular, the difficulty in identifying patterns of correlation in the second period, which was a relatively early stage of the pandemic, may reflect this change. That is, it can be pointed out that the period corresponding to the second period, which included both the second and third waves of infection, may have seen a significant change in the relationship between the symptom profile and the actual spread of infection. This point is also discussed further in later sections, based on the results of the supervised model.

4.1.2 Regional Analysis

The above analysis for country data was then applied to regional data. The results of the correlation analysis are shown in the table and are generally similar to those for country-level data. The square brackets in the table show the standard deviation of the data by region, and the results show that the variation in the data by region is fairly small, confirming that the tendency identified in the country-level data has a certain regional generality.

[46]It can be said as reasonable the correlation peaks at least at the point where the x-axis is within about minus 30 days and then shows a gradual decline.

Table 3: Correlation analysis: Regional data

	Correl. without shifting		**Max correl. with shifting**	
	w/o min. media	w/ min. media	w/o min. media	w/ min. media
New Cases				
entire period	0.235	0.262	0.244 [-7.0]	0.274 [-10.8]
	(0.048)	(0.055)	(0.052)	(0.058)
01/07/2020 - 31/03/2021	-0.137	-0.161	0.808 [-30.8]	0.809 [-31.4]
	(0.274)	(0.277)	(0.155)	(0.158)
01/07/2020 - 31/03/2021	0.302	0.349	0.411 [-16.9]	0.443 [-16.6]
	(0.161)	(0.164)	(0.175)	(0.162)
01/04/2021 - 19/05/2022	0.343	0.344	0.372 [-8.9]	0.383 [-13.1]
	(0.093)	(0.101)	(0.101)	(0.107)
New Deaths				
entire period	-0.051	-0.057	0.335 [-28.2]	0.327 [-28.8]
	(0.113)	(0.113)	(0.145)	(0.140)
01/07/2020 - 31/03/2021	-0.143	-0.165	0.841 [-28.6]	0.840 [-29.2]
	(0.200)	(0.202)	(0.106)	(0.103)
01/07/2020 - 31/03/2021	0.121	0.164	0.302 [-22.9]	0.345 [-23.6]
	(0.152)	(0.157)	(0.176)	(0.168)
01/04/2021 - 19/05/2022	0.407	0.395	0.511 [-10.6]	0.499 [-10.3]
	(0.155)	(0.151)	(0.110)	(0.109)

Note: correlations between the search scores and the respective clinical data by time period. The left column shows the correlation coefficients calculated without shifting the clinical data, while the right column shows the results of shifting the clinical data to maximise the correlation within a range of 40 days in the past and 10 days in the future (range of 50 days in total). The numbers in square brackets correspond to the standard deviations of correlations over the regions. The numbers in round brackets represents the mean number of days shifted over the regions.

Next we look at the results of the inter-regional correlation analysis between regions based on the classification according to infection status. The results are shown in the table 4 and 5. The leftmost column represents the base groups of regions for comparison and the top row represents the group of target regions. The clinical data of regions in the target group are shifted back in time by the number of days listed in the leftmost column, and the correlations are calculated between the search scores of regions in the base group. The tables show mean correlations and their standard deviations in brackets.

Table 4: Interregional correlation: New cases

	Group1	**Group2**	**Group3**
Group1			
No shift	0.260 (0.187)	-0.157 (0.228)	-0.135 (0.229)
7-days back	0.584 (0.166)	0.142 (0.268)	0.166 (0.267)
14-days back	0.814 (0.129)	0.450 (0.260)	0.471 (0.257)
21-days back	0.820 (0.103)	0.637 (0.220)	0.650 (0.215)
28-days back	0.687 (0.111)	0.698 (0.192)	0.700 (0.189)
Group2			
No shift		-0.320 (0.151)	-0.338 (0.15)
7-days back		-0.106 (0.196)	-0.127 (0.195)
14-days back		0.151 (0.232)	0.130 (0.233)
21-days back		0.382 (0.242)	0.363 (0.244)
28-days back		0.574 (0.226)	0.561 (0.230)
Group3			
No shift			-0.356 (0.145)
7-days back			-0.154 (0.189)
14-days back			0.097 (0.228)
21-days back			0.326 (0.244)
28-days back			0.527 (0.236)

Note: mean correlations between the search scores and clinical data with and without shifting. The numbers in brackets show the standard deviations of correlations. The leftmost column represents the base groups of regions for comparison and the top row represents the group of target regions. The clinical data of regions in the target group are shifted back in time by the number of days listed in the leftmost column, and the correlations are calculated between the search scores of regions in the base group. 6

As can be seen from the table, clinical data from regions where the infection situation occurred late are strongly linked, with a lag, to the search scores from regions where the infection occurred ahead. For example, for new infections, the correlation of group1's search scores with group1's clinical data (shown in the top left) shows that the correlation becomes stronger as the clinical data is shifted back, peaking at around 21 days back. Similar to this, group1's search scores show a tendency for the correlation of group2's clinical data (shown in the top middle) to become stronger as the clinical data is shifted back in, but with further lags. This result can be seen in both clinical data and among all groups, suggesting that search data from regions where infections

occur ahead may be useful for understanding infection trends in other regions[47].

Table 5: Interregional correlation: New deaths

	Group1	**Group2**	**Group3**
Group1			
No shift	0.088 (0.143)	-0.163 (0.168)	-0.145 (0.169)
7-days back	0.417 (0.172)	0.115 (0.207)	0.137 (0.207)
14-days back	0.738 (0.161)	0.465 (0.215)	0.485 (0.213)
21-days back	0.882 (0.085)	0.725 (0.167)	0.739 (0.163)
28-days back	0.771 (0.074)	0.782 (0.113)	0.784 (0.111)
Group2			
No shift		-0.244 (0.145)	-0.262 (0.144)
7-days back		0.005 (0.187)	-0.016 (0.187)
14-days back		0.312 (0.217)	0.291 (0.217)
21-days back		0.570 (0.225)	0.554 (0.227)
28-days back		0.705 (0.189)	0.697 (0.193)
Group3			
No shift			-0.286 (0.138)
7-days back			-0.050 (0.180)
14-days back			0.256 (0.213)
21-days back			0.521 (0.228)
28-days back			0.673 (0.198)

Note: see the note in table 6

country, this results is predictable to some extent because the search behaviour and infection situation across regions would may show similar patterns with regional time differences. What is interesting in this respect, however, is that the correlation between search scores in the preceding regions and clinical data in the lagging regions generally tends to be higher than the correlation between search scores and clinical data within the same group. Further in-depth analysis of this result is needed, but given that it holds for all comparisons, it is possible that the search behaviour of the preceding regions may contain more important information for understanding trends in infection.

[47]For the third group, the comparison with the first group shows slightly less lag than for the second group, but the difference is slight and it can be assumed that the results will vary somewhat depending on how the groups are divided.

4.1.3 Summary of Results of Unsupervised Analysis

For the unsupervised model, a correlation analysis was conducted between a single aggregated series of search data pertaining to COVID-19-related symptoms (i.e., search score) and some clinical data on COVID-19 (number of new confirmed infections, new deaths and new hospital admissions), based on the methodology of previous study[29]. The results showed that search score has a certain correlation with clinical data and that they tend to emerge prior to clinical data, confirming once again that search data on infectious diseases can be a leading indicator for understanding infection trends and the prospects of their medical consequences. On the other hand, the correlations and their temporal structure varied widely across the clinical data and across the time periods analysed, and some did not appear to provide much useful information. To a certain extent, this may stem from the limitations of the unsupervised model of heuristic selection and assignment of features and parameters (weights), and may also be influenced by the assumption of constant weights through time in the present analysis. In this respect, it may be natural to assume that the appropriate weights have changed over time in the first place, as variations by analysis period were also observed in the correlation analysis for individual search data.

For these, it would be useful to utilise other clinical categories and weights to further look at changes in outcomes, and varying the weights through time could also be considered. In this regard, the ONS survey used in this analysis regularly publishes new data on symptom profile, and this would be one way to make use of it. We also attempted to improve the unsupervised model by mitigating the effect of media coverage on the search score, but did not see a strong improvement. There is still room for improvement in this respect, e.g. by changing the scope of the search data or by adjusting the hyperparameters of the model for detecting the media effect, which are also tasks for the future.

The unsupervised analysis utilising regional datasets first confirmed that search data can be useful in understanding infection trends at the regional level as well, by applying the same correlation analysis as above. In addition, the correlation analysis between the search scores of the regions ahead in infection status and the clinical data of the other regions clearly showed a tendency for that clinical data to emerge with lag in the search scores of preceding regions. This may indicate the potential of search data from other regions in understanding and predicting infection situation. In this analysis, the analysis was ended here, but there is room for further deepening this analysis in the future. For example, this analysis used a fairly simple method of classifying regions according to the degree of infection prevalence, but one future direction

would be to improve this method to reflect the actual real-time prevalence situation. Also, although the present analysis was conducted from a macro perspective, dividing an entire country into three large groups and making comparisons between the groups, it may be necessary to incorporate regional adjacencies into the analysis in some way, given the reality of the transmission of infectious diseases.

4.2 Supervised Model

4.2.1 Linear Regression Analysis (Goodness-of-Fit Analysis)

The results of the unsupervised model confirm that the unsupervised model using the online search data is an effective tool in capturing trends of clinical data. On the other hand, the degree of correlation and the degree of precedence also depended to a large extent on the clinical data and the selection of the period, which revealed certain limitations of the unsupervised model. In light of these issues, the following section looks at the results of the linear regression analysis as the first part of the supervised analyses. As described in detail in the methodology chapter, the regression analysis is performed using the original series of search data, which are not aggregated, and news data and see the impact of individual symptom-related search on each clinical data in a multivariate manner[48]. (As related to this analysis, a cross-correlation table among each explanatory variable is presented in the appendix.)

The table 6 shows the results of regression analysis for new confirmed cases. First, it can be seen that the estimated coefficients for both the classical linear regression model (denoted as CLR in the table) and the elastic net (denoted as Elastic in the table), are quite different from those calculated based on the ONS survey. For instance, while "Cough", "Fatigue" and "Weakness" are assigned relatively high values in the ONS based weights, they contribute negatively or zero in the regression analysis. "Headache" and "Myalgia" are also assigned much lower weights in both models than in the ONS based weights. These differences may be due to the fact that these symptoms are fairly common, so searches for this are likely to have been done more often by other diseases than COVID-19. This result is consistent with the implications of the previous study[29]. As for the news data, it shows a negative relationship with new cases which is an expected outcome.

[48]The impact of each search data could be seen to a certain extent in the graphical analysis of the unsupervised model explained above, but that analysis only focused on one-on-one correlations between individual search data and clinical data. In this respect, regression analysis is expected to capture more detailed relationships, as the impact of other search data can be controlled through multivariate analysis.

Table 6: Regression coefficients: New confirmed cases

	CLR	**Elastic**	**(Ref.)ONS**
const.	0.232	0.138	-
Abdominal pain	-0.137	0	0.073
Ageusia	-0.695	0	0.188
Anosmia	0.730	0.123	0.180
Cough	-1.075	0	0.418
Diarrhea	-0.665	-0.222	0.068
Fatigue	-0.658	-0.506	0.373
Fever	-0.001	0	0.244
Headache	0.167	0.118	0.376
Myalgia	0.126	0.104	0.257
Nausea	0.619	0.527	0.095
Shortness of breath	0.452	0	0.143
Sore throat	0.780	0.084	0.321
Vomiting	0.345	0.058	0.095
Weakness	-0.048	-0.071	0.373
News	-0.418	-0.304	-

Note: the table shows the regression coefficients of each regression model by symptom and news ratio. The column named "CLR" represents regression coefficient with the classical linear regression model without regularisers, the column named "Elastic" represents coefficient with the elastic net model, and the column named "ONS" shows weights for each corresponding symptom category from ONS's survey as reference.

Furthermore, a comparison of the classical regression model with the model with regularisation (in the figure 6) shows that the regularised model became somewhat sparse due to the effect of regularisation and also shows differences with regard to the magnitude of the individual coefficients. For example, "Ageusia" and "Anosmia" had relatively large positive and negative values without regularisation, respectively, but in the elastic net, "Ageusia" was compressed to zero and "Anosmia" to a smaller value. This may be due to the fact that the search data for these two symptoms have a very high (almost 1) correlation[49], and this was appropriately adjusted by regularisation[50]. Other coefficients for "Shortness of breath" and "Sore throat" are also compressed to a very small value or zero by regularisation.

[49](see the diagram of cross-correlation in the appendix)

[50]They are highly correlated most probably because the same person was experiencing these two symptoms at the same time in most cases.

Table 7: Regression coefficients: New deaths

	CLR	**Elastic**	**(Ref.)ONS**
const.	0.219	0.108	-
Abdominal pain	0.948	0.602	0.073
Ageusia	0.128	0.122	0.188
Anosmia	-0.268	-0.031	0.180
Cough	-0.164	-0.116	0.418
Diarrhea	0.009	0	0.068
Fatigue	-1.220	-0.795	0.373
Fever	-0.936	-0.563	0.244
Headache	0.711	0.694	0.376
Myalgia	-0.055	-0.041	0.257
Nausea	-0.108	0.01	0.095
Shortness of breath	1.453	0.870	0.143
Sore throat	-0.636	-0.534	0.321
Vomiting	0.08	0.052	0.095
Weakness	-0.077	-0.189	0.373
News	-0.108	0.021	-

Note: see the note in table 6

For the data on deaths, it can be seen that most of the coefficients remain non-zero even the model is regularised, and that the overall patterns of coefficients are similar between the two models, not like the table 6. Symptoms with a positive impact are "Abdominal pain", "Ageusia", "Headache" and "Shortness of breath", which are different from the results for newly confirmed cases[51]. The contribution of the news data is also very small. This may be because, although the increase in media coverage has the effect of increasing searches that are not due to infections, it is also linked to the trend of the spread of infection itself, which makes its contribution somewhat ambiguous in relation to the number of deaths, where there is a larger time lag.

[51]For Ageusia, as noted above, there is a very strong correlation with Anosmia, so it can be interpreted as these two as practically the same.

Table 8: Regression coefficients: New hospital admissions

	CLR	**Elastic**	**(Ref.)ONS**
const.	0.282	0.122	-
Abdominal pain	0.524	0.326	0.073
Ageusia	0.325	0.324	0.188
Anosmia	-0.054	0.113	0.180
Cough	-1.183	-0.402	0.418
Diarrhea	-0.158	-0.037	0.068
Fatigue	-1.096	-0.773	0.373
Fever	-0.537	-0.383	0.244
Headache	0.830	0.774	0.376
Myalgia	-0.037	0	0.257
Nausea	0.227	0.310	0.095
Shortness of breath	1.398	0.726	0.143
Sore throat	0.014	-0.350	0.321
Vomiting	0.248	0.160	0.095
Weakness	-0.229	-0.274	0.373
News	-0.350	-0.144	-

Note: see the note in tabel 6

For the number of new admissions, the sign and magnitude of the coefficients show a relatively similar pattern to the results for the number of deaths. The symptom variables with positive contributions are “Abdominal pain”, “Ageusia”, “Anosmia”, “Headache”, “Nausea”, “Shortness of breath” and “Vomiting”. Interestingly, the variables with a positive contribution to new admissions are just the union set of variables contributing positively to new deaths and new cases (though these two sets are fairly different). This may suggest searches for each of the symptoms are related to the severity of the condition. In other words, there may be a tendency for symptoms that are more related to the newly confirmed cases rather than the new deaths, to not necessarily indicate the severity of public health conditions, even if they indicate the spread of disease (and vice versa). In the public health response, it may need to look more closely at trends in the search for symptoms related more to deaths.

Next, using the linear regression model without regularisation, we look at the goodness-of-fit between the model estimates and each clinical data and its temporal structure. The basic idea of the analysis is similar to that of correlation analysis with the unsupervised model, where clinical

data are shifted by a certain width to find the shifting of best fit. The results are summarised in the table 9[52].

Table 9: Goodness-of-fit: Adj. R-squared of linear regression model

	$\mathbf{R^2}$ **w/o shift**	**Best** $\mathbf{R^2}$ **w/ shift**
New cases		
entire period	0.521	0.631 [-15]
period1	0.964	0.990 [-17]
period2	0.912	0.934 [-10]
period3	0.839	0.860 [-8]
period1+2	0.693	0.700 [-14]
New deaths		
entire period	0.600	0.679 [-13]
period1	0.973	0.996 [-19]
period2	0.792	0.909 [-25]
period3	0.821	0.882 [-25]
period1+2	0.729	0.753 [-13]
New admissions		
entire period	0.559	0.606 [-8]
period1	0.983	0.996 [-13]
period2	0.849	0.913 [-16]
period3	0.833	0.856 [-13]
period1+2	0.755	0.761 [-3]

Note: adjusted R-squared with the unregularised linear regression model. The right column shows the results of shifting the clinical data to maximise the adjusted R-squared within a range of 40 days in the past and 10 days in the future. The number brackets corresponds to the number of days shifted, with negative (positive) values indicating that the clinical data has been shifted into the past (future). Period 1 corresponds to the period 13/02/2020 to 30/06/2020, period 2 corresponds to 01/07/2020 to 31/03/2021, and period 3 corresponds to 01/04/2021 to 19/05/2022, respectively.

First, it can be seen that, in general, the R-squared are quite high, and this is particularly true for the data per split time period. The temporal structure also appears to be more reasonable than in the correlation analysis with the unsupervised model. The data on deaths and hospital admissions are in an intuitively interpretable order, as in the correlation analysis, and the number

[52]Corresponding graphs are also available in the appendix

of optimal shifting for the new cases in periods 2 and 3 are more reasonable, as they are shorter than that of the other 2 clinical data. However, as the length of the period becomes longer, like period 1+2 and the entire period, the good-of-fit decreases and the reasonable temporal structure tends to break. This may reflect the fact that the relationships between variables change over time and a simple single model does not capture these transitions well. This point will be further analysed later.

4.2.2 Forecasting Models

4.2.2.1 Naive Forecasting Model

In the following, we will look first at the results of the naive forecasting model. For the explanatory variables of the model, based on the results of the above goodness-of-fit analysis, the data which are shifted by the number of days with the best fit are used[53].

Table 10: Naive forecasting model with optimal shifting

	New Cases		**New Deaths**		**New Admissions**	
	CLR	Elastic	CLR	Elastic	CLR	Elastic
Period2						
MAE	14685.7	14685.8	235.9	266.0	782.9	872.7
Correl.	0.240	0.241	0.690	0.724	0.796	0.785
Period3						
MAE	42589.8	35575.1	223.3	267.7	749.5	667.5
Correl.	0.279	0.503	0.332	0.570	0.165	0.455

Note: mean absolute error and correlation between truth values and the predicted values. The above model is trained with period 1 (13/02/2020-30/06/2020) and predicted period 2 (01/07/2020-31/03/2021) and the bottom model is trained with both period 1 and 2 (13/02/2020-31/03/2021) and predicted period 3 (01/04/2021-19/05/2022). "CLR" means the classical linear regression model and "Elastic" means ElasticNet.

The results show that the performance of the predictions are poor when looking at MAE. On the other hand, looking at the correlations, it can be seen that some models (especially the elastic net) exceed 0.7, indicating that certain patterns could be extracted. However, when compared to the results in the goodness-of-fit table, it is clear that the performance against the test data deteriorated significantly for the training data. This may be due to over-fitting or structural changes between periods, or both. So models with more sophisticated structure are needed to make better predictions.

[53]The number of days are changed to the most appropriate one for each period.

Figure 9: Prediction of naive forecasting model: New cases

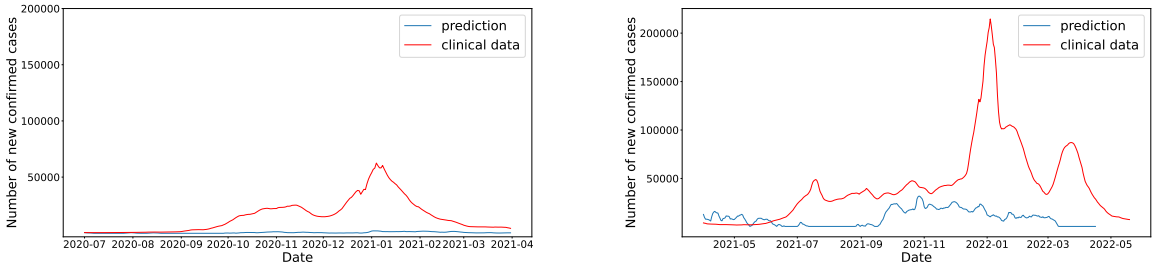


Figure 10: Prediction of naive forecasting model: New deaths

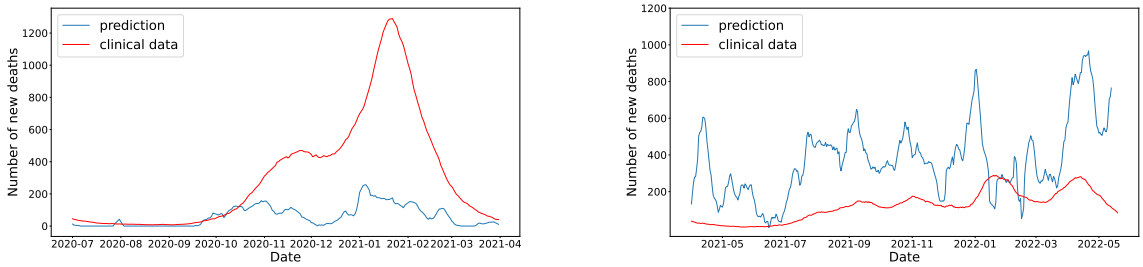
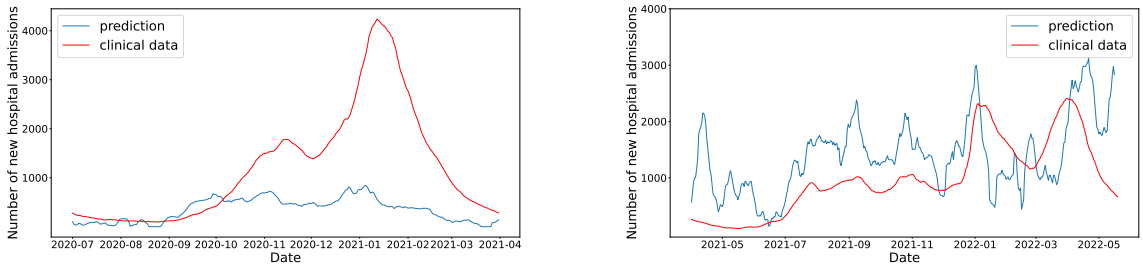


Figure 11: Prediction of naive forecasting model: New hospital admissions



Note: the left figure is predictions of a model trained with period 1 and the right figure is predictions of a model trained with period 1 and 2.

4.2.2.2 Linear Retraining Model

Then, we look at the results of applying the linear retraining model to the country data as a more full-fledged forecasting model analysis. This model is trained with a small chunk of data in the past and makes a single nowcasting or forecasting, and the same process is repeated in the next period by shifting the training set and forecast data by one period. Thus, forecasts are made every period. Table 11 shows the respective results for the analysis: the simple persistent model as a baseline is shown as PER, the model with only autoregressive terms without search data is shown as LRM-A, and the model with search/news data added to it is represented as LRM-S. In the following, given the above results, we will examine the results of the forecasting model for the number of deaths, which can be more reliable ground-truth data (the results for other clinical data are presented in the appendix.).

Table 11: Linear Retraining model: New deaths

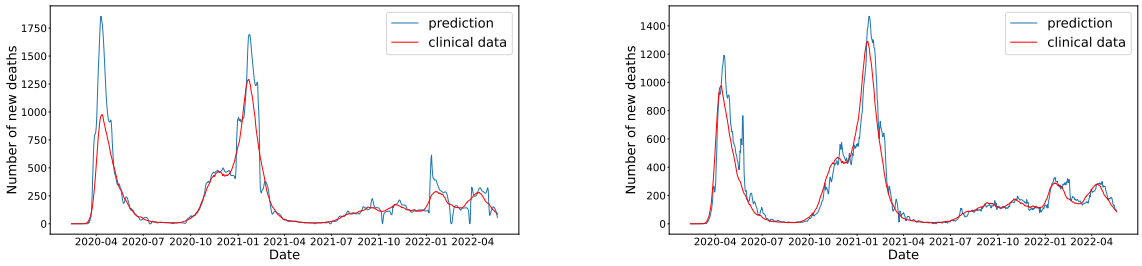
	PER	**LRM-A**	**LRM-S**
Nowcasting			
MAE	46.1 (70.5)	51.3 (108.5)	44.8 (68.0)
Correl.	0.949	0.952	0.964
7-day Forecasting			
MAE	87.7 (131.2)	83.1 (253.5)	49.4 (76.9)
Correl.	0.821	0.852	0.956
14-day Forecasting			
MAE (std.)	123.7 (178.6)	122.0 (566.0)	53.8 (133.7)
Correl.	0.661	0.661	0.906

Note: mean absolute error and correlation between truth values and the predicted values. PER is the persistent model, LRM-A is the linear retraining model with autoregressive features, and LRM-S is the model with autoregressive features and search/news data. The numbers in the brackets are standard deviations of each MAE. The prediction covers entire period in country.

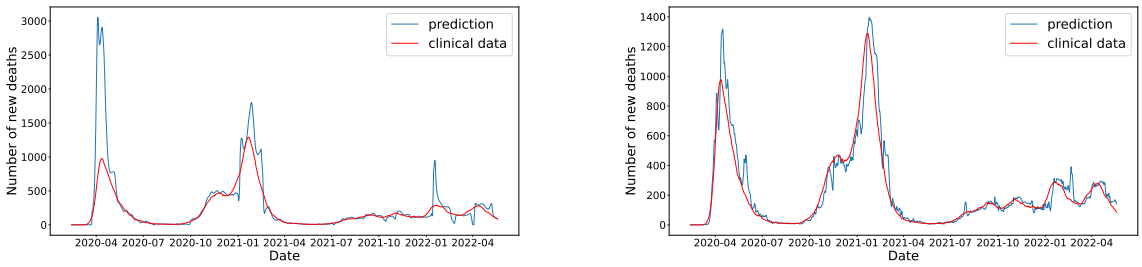
The performance of the LRM-S model outperforms the baseline (PER) and LRM-A results on both indicators, indicating not only that the model has a certain level of accuracy in forecasting, but also that the search data contribute significantly to the accuracy of the forecasts. It is also interesting to note that for PER and LRM-A, the accuracy deteriorates significantly as the period of forecasting becomes more distant, whereas for LRM-S, there is only a little deterioration as the period is extended. This can be interpreted as being due to the precedence of search data over clinical data, which has been confirmed in previous analyses, and supports the potential of search data for medium-term forecasting accuracy.

Figure 12: Prediction of linear retraining model: New deaths

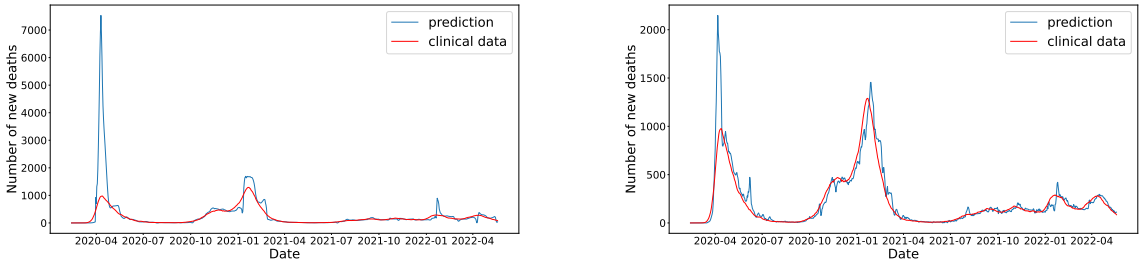
Nowcasting (Right:LRM-A / Left:LRM-S)



7-day forecasting (Right:LRM-A / Left:LRM-S)



14-day forecasting (Right:LRM-A / Left:LRM-S)



Note: the left figure is of the linear retraining model with autoregressive features (LRM-A) and the right figure is of the model with autoregressive and search/news features (LRM-S).

The difference in accuracy becomes clearer when looking at the respective graphical results (figure 12) for LRM-A and LRM-S. LRM-A learns only from past trends, so its accuracy deteriorates significantly when the trend of clinical data change, whereas for LRM-S it appears to be somewhat better at capturing these changes. This would be due to the fact that the search data serve as leading indicators and fit the model well to changes in the trend. However, there are some issues with the smoothness of the forecast series, as there are sudden spikes at certain points of the series. There is also a tendency for the forecasts to lag slightly behind the real values as the forecasting period lengthens. How to improve these would be a challenge in further brushing up the model.

4.2.2.3 Model Comparison with LSTM: Country Data

In addition to the above analysis, the model is further compared with the LSTM, which, as mentioned above, is a non-linear regression model with high expressive power for time series data, but is generally prone to over-fitting and it is desirable to prepare a sufficient amount of samples for training. Therefore, as the model is expected to show its full potential in the following regional analyses, it is considered appropriate to understand the analysis of country data as a benchmark. As explained in the methodology chapter, for this analysis, the period of the forecast (performance evaluation) is limited to the latest six months, as the main focus is on comparison with the neural network[54].

[54]Thus, for the linear retraining model, a subset of the results of the predictions in the previous analysis coincide with the results of this section.

Table 12: Country data forecasting: New deaths (pred. latest 6 months)

	PER	**LRM-A**	**LRM-S**	**LSTM-A**	**LSTM-S**
Nowcast					
MAE	25.8 (18.5)	49.1 (61.4)	31.4 (25.7)	27.9 (23.8)	62.5 (45.5)
Correl.	0.853	0.763	0.844	0.815	0.266
7-day					
MAE	48.3 (32.9)	67.0 (111.8)	35.7 (35.0)	45.5 (25.1)	70.5 (49.1)
Correl.	0.494	0.606	0.73	0.553	0.184
14-day					
MAE	66.6 (43.5)	59.2 (95.4)	31.2 (34.0)	72.2 (40.6)	103.1 (63.2)
Correl.	0.062	0.532	0.75	0.039	0.102

Note: mean absolute error and correlation between truth values and the predicted values. PER is the persistent model, LRM-A and LSTM-A are the linear retraining model and LSTM with autoregressive features respectively, and LRM-S and LSTM-S are the models with autoregressive features and search/news data. The numbers in the brackets are standard deviations of each MAE. The prediction covers the latest six months (20/11/2021-19/05/2022).

The results table shows that for LRM, the performance of the model including the search data (LRM-S) was significantly better than the baseline for both 7-day and 14-day forecasts, although the nowcasting was inferior to the persistent model[55]. Comparisons with LRM-A also show that the introduction of search data has a significant impact on forecasting improvement. As for the LSTM, as one might expect, none of the performance is good. The only one, nowcasting with only autoregressive terms, has relatively good accuracy, but this is worse than the performance of the persistence model, making it difficult to accept that the model is finding meaningful relationships in the data. In addition, and in contrast to the LRM, the accuracy of the predictions deteriorate as a result of the introduction of the search data. This may be a result of over-fitting due to the addition of features[56].

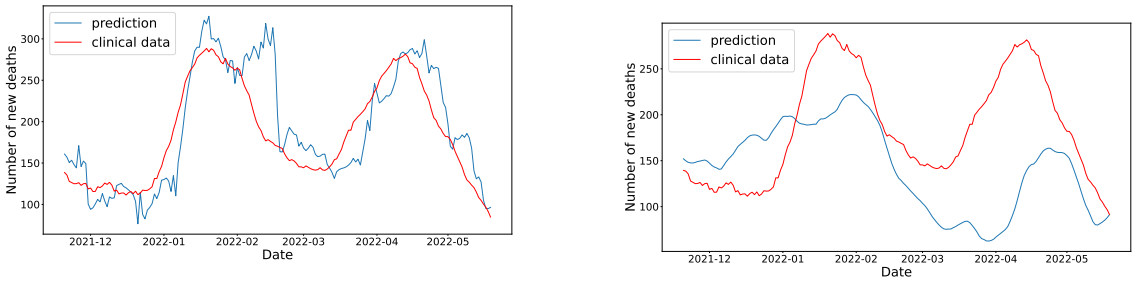
On the other hand, the graph 13 shows that LSTM predictions are much smoother. This is thought to be a result of the LSTM's ability to make forecasts by appropriately retaining both long-term and short-term dependent structure of data. This is not the case in linear regression models with a simple structure that reflect sensitively to fluctuation of the present data. This

[55]Here, it should be noted that, although the nowcasting results did not perform well, the forecasting period in this analysis was a time when fluctuations of clinical data were relatively moderate and therefore the performance of the persistent model was particularly likely to be enhanced.

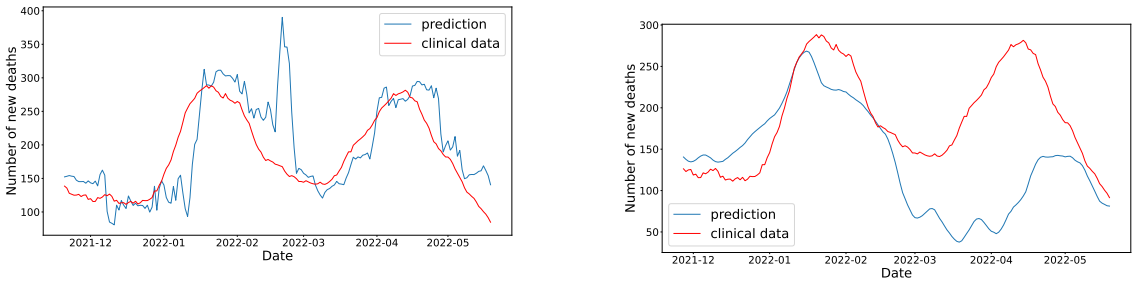
[56]As mentioned above, to limit over-fitting, drop-outs were introduced and the model that were most accurate against the validation data during the training process were adopted.

Figure 13: Model Comparison: New deaths of country data

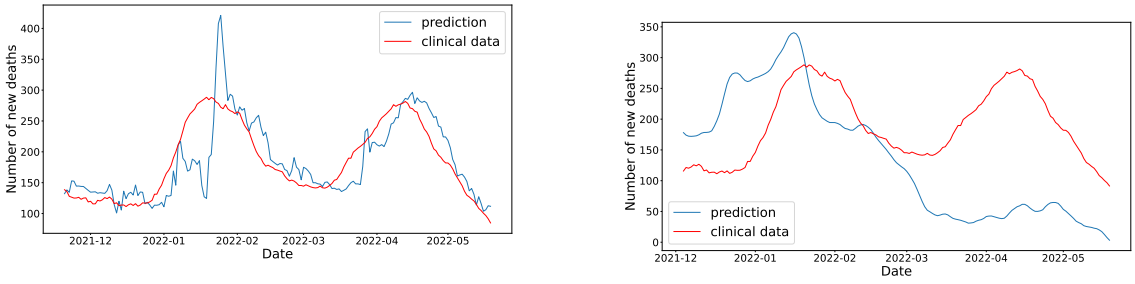
Nowcasting (Right:LRM-S / Left:LSTM-S)



7-day forecasting (Right:LRM-S / Left:LSTM-S))



14-day forecasting (Right:LRM-S / Left:LSTM-S))



Note: the left figure is of the linear retraining model with autoregressive and search/news features (LRM-S) and the left figure is of LSTM with autoregressive and search/news features (LSTM-S).

can be said to accurately capture the characteristics of a series with persistence, such as trends in infectious diseases, and represents the potential of LSTM, which cannot be captured by MAE alone.

4.2.2.4 Model Comparison with LSTM: Regional Data

Finally, we look at the results of applying the forecasting models to the regional data. The results are shown in the table 13, where it is clear that the prediction accuracy of the LSTM-S outperforms that of the PER and LRM in all prediction periods[57]. It can also be seen that the standard deviations of MAE by region are also small (in brackets are the averaged standard deviations of the MAE calculated by region).

Table 13: Reginal data forecasting: New deaths (pred. latest 6 months)

	PER	**LRM-A**	**LRM-S**	**LSTM-A**	**LSTM-S**
Nowcast					
m-MAE	0.398 (0.392)	0.395 (0.384)	0.372 (0.370)	0.403 (0.329)	0.368 (0.302)
m-Correl.	0.379	0.201	0.335	0.380	0.385
7-day					
m-MAE	0.464 (0.470)	0.687 (0.700)	0.686 (0.700)	0.482 (0.393)	0.419 (0.326)
m-Correl.	0.253	0.277	0.279	0.272	0.259
14-day					
m-MAE	0.537 (0.550)	0.687 (0.700)	0.685 (0.701)	0.549 (0.440)	0.474 (0.385)
m-Correl.	0.067	0.277	0.282	0.118	0.085

Note: m-MAE and m-Correl. represent the mean of MAE and the mean of correlation of each region, respectively. The numbers in the brackets are the mean of standard deviations of each MAE of each region. See the note in table 12 for other notes.

This is in line with some expected results and demonstrates the superiority of the recurrent neural network on time-series data sets. In the model with the search data (i.e., models with "-S"), the LRM showed little improvement in forecasting accuracy, whereas the LSTM showed relatively large improvement. Here it should be noted the fact that the regional search data is not a complete dataset in that full search data are not available in some regions or data below a certain threshold are cut off to zero. That is, a prediction model with a simple structure like the LRM-S, even with improvements in the training algorithms, may not be able to extract

[57]These results may appear relatively small in value because they are simple averages of MAEs by region, but if, for example, the MAEs of all regions are simply added up rather than averaged out and converted to units roughly corresponding to the country level, m-MAE of 0.5 correspond to MAE of 94.5.

meaningful information from those incomplete data. In contrast, LSTM-S appears to be able to successfully extract useful information for prediction even from imperfect data due to its more sophisticated learning process and expressive power, as its accuracy is improved by introducing search data. The similar results can be confirmed by results with a forecasting period of the latest 12 months for the robustness chuck (the result table is in the appendix).

4.2.3 Summary of Results of Supervised Analysis

For the supervised model, a linear regression analysis (with and without regularisation) was first conducted utilising the unaggregated search and clinical data. The results showed that the regression coefficients estimated by the model differed significantly from the weights of the unsupervised model, supporting inferences based on the results of the unsupervised analysis. Those regression coefficients differed depending on the clinical data, in particular, the number of new infections and the number of new deaths differed in terms of the high-impact symptoms. Given these results, public health policy makers may need to be more careful in searching for symptoms that are more strongly associated with new deaths.

The goodness-of-fit analysis (period-by-period) using this linear regression model showed that its R-squared was considerably high and that the temporal relationship between the explanatory variables (search data) and the explained variables (clinical data) was not only consistent with the results of the unsupervised analysis, but also showed a more interpretable relationship. On the other hand, extending the time period over which the model was applied not only worsened the goodness-of-fit, but also blurred the temporal structure between the variables. This problem was also seen in the results of the forecasting model using the linear regression model (i.e., the naive forecasting model), which showed very poor performance for forecasts on data from a different time period than the training sample.

In view of this problem, a linear regression model with an improved training process (i.e., the linear retraining model) was introduced to allow the model to cope with changes in the relationship between variables over time. Two models were constructed, one with only an autoregressive terms and one with an autoregressive terms plus search data, and their results were compared with those of a very simple persistent model as a baseline. As a result, the accuracy of the linear retraining model with search data in nowcasting and 7 and 14 days forecasting outperformed the baseline model and the model with autoregressive terms only, for country-level data. This again confirms that the relationships between variables are changing over time and that improvements in the training process can be effective in addressing this issue. In addition, the improvement in accuracy

prediction with the introduction of search data was very significant, demonstrating the potential of search data as a leading indicator in forecasting. On the other hand, it can be pointed out that, due to the characteristics of the linear regression model, the series of predicted values is not as smooth as it should be, and how to deal with this may be an issue.

Finally, a comparison of the performance of the linear retraining model and LSTM on regional data was carried out: although LSTM is known for its advanced architecture and potentially high performance in forecasting time-series data, their complex structure and high expressive power make it challenging to deal with the risk of overfitting. In this study, we analysed how LSTM performs with search data, taking advantage of the large number of training samples in the geographically extended dataset. As a result, it was confirmed that LSTM outperforms both the baseline model and the linear retraining model, showing the potential of utilising RNN architecture together with search data in this field.

5 Conclusion

This paper analysed how online search data on infectious disease symptoms can contribute to understanding and predicting trends of COVID-19, through the use of time- and geographically-extended datasets in the UK. The dataset consists of both country-level data and regional-level data covering 189 regions. The analysis has been conducted from various perspectives, utilising both unsupervised and supervised models. The main findings are summarised below.

Analyses using unsupervised modelling showed that, in line with the results of previous studies, online COVID-19 related symptom search data related correlate with and preceded clinical data such as the number of new confirmed cases, new deaths and new hospital admissions. In order to look at the cross-regional relationship between search data and clinical data, we also analysed the correlation between search data in regions where the spread of infection was leading and clinical data in regions where the spread was lagging, utilising regional-level data. Results showed that strong correlations and common patterns were observed between them, indicating that search data in the leading regions can provide useful prior information on infection trends in the lagging regions. This regional analysis provides a starting point for further analysis and further understanding of the characteristics of search data, taking into account other perspectives, as discussed in the paper.

Linear regression analysis was also applied as a supervised analysis to further look at the relationship between the search and clinical data. The results showed that important symptom search data differed according to the clinical data. This may provide useful information for public health policy makers to understand the severity of the situation. The supervised analysis also suggested that the correlation and temporal relationship between search data and clinical data changed through time, as partly implied by the results of the correlation analysis as well.

Finally, based on these findings, an advanced linear regression model with an improved training process to allow the model to cope with temporal changes in the relationships between variables and an advanced RNN, LSTM, were introduced as prediction models. The results of using these to predict clinical data in the short-term future (nowcasting and 7 days and 14 days forecasting) showed that for the linear regression model, search data improved the model's predictions in every prediction periods. This again demonstrates the superiority of search data as a leading indicator in forecasting. In addition, although the LSTM did not perform well on country-level data with a limited number of samples, it significantly outperformed the linear regression model on region-specific data, where the sufficient number of samples. The results showed the future

potential of utilising RNN architecture together with search data in this field. However, as this analysis used an LSTM with a very simple structure, designing optimised models in the prediction of infectious disease using search data will be a future work.

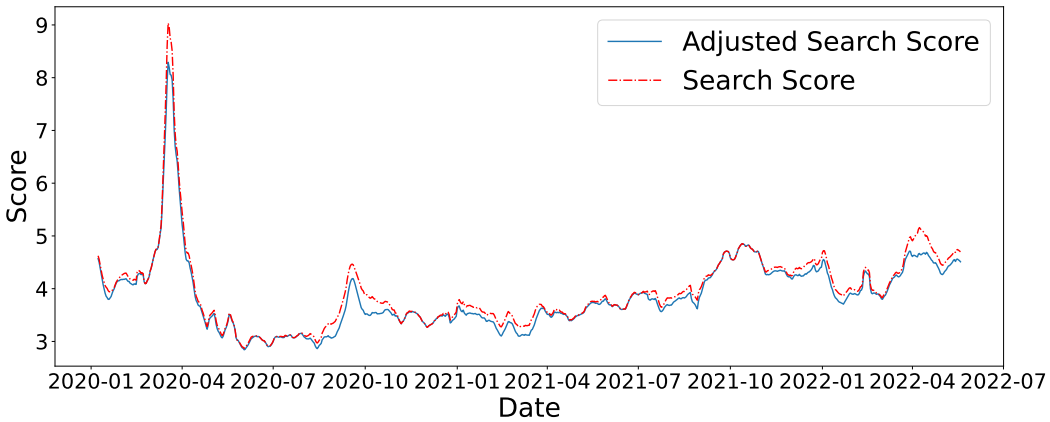
Appendix

Summary of Regional Search Data

	% of null	# of unabailable region
Abdominal pain	1.4	0
Ageusia	95.6	163
Anosmia	91.1	151
Cough	1.5	0
Diarrhea	1.6	0
Fatigue	1.2	0
Fever	1.7	0
Headache	1.6	0
Myalgia	32.1	45
Nausea	3.3	0
Shortness of breath	16.2	12
Sore throat	3.4	0
Vomiting	2.0	0
Weakness	43.6	60

Note: the left columns shows the percentage of null (the unavailable data) to the total number of data for each symptom. The right shows the number of regions where none of the data is available for each symptom.

Adjusted Search Score



Correlation Table with FF100 weights

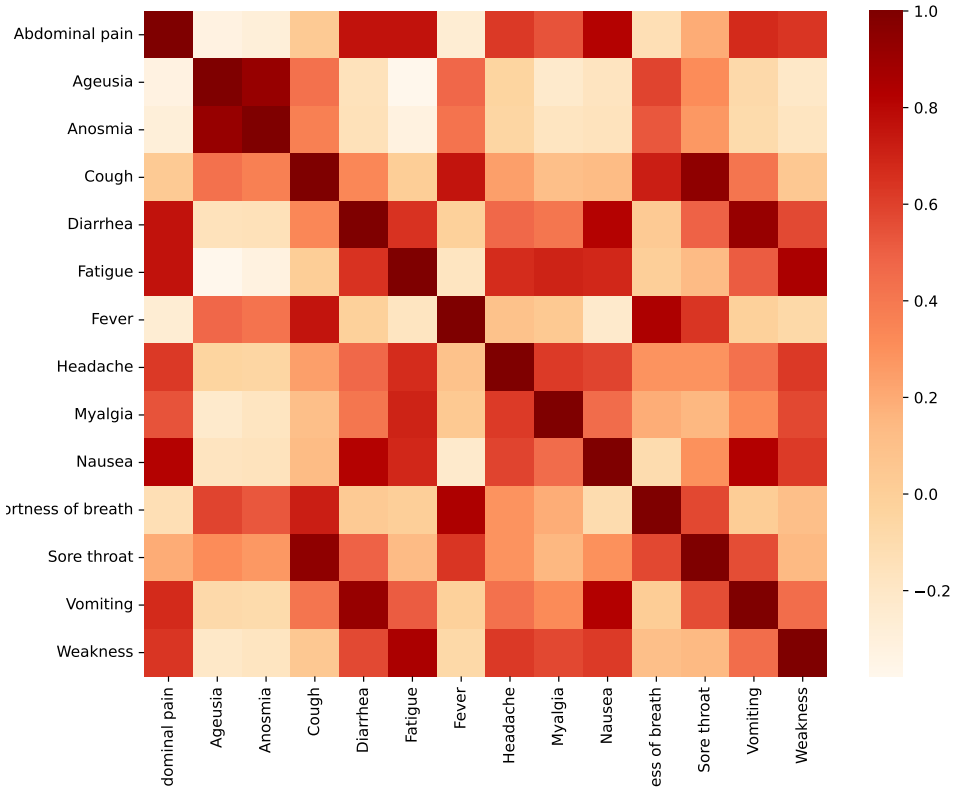
Table 14: Correlation analysis with FF100 weights: Country data

	Correl. without shifting		**Max correl. with shifting**	
	w/o media adj.	w/ media adj.	w/o media adj.	w/ media adj.
New Cases				
entire period	0.183	0.183	0.184 [-2]	0.199 [-40]
13/02/2020 - 30/06/2020	-0.170	-0.173	0.822 [-34]	0.844 [-34]
01/07/2020 - 31/03/2021	0.248	0.351	0.277 [-22]	0.411 [-22]
01/04/2021 - 19/05/2022	0.323	0.270	0.351 [-1]	0.320 [-40]
New Deaths				
entire period	-0.087	-0.102	0.334 [-26]	0.327 [-26]
13/02/2020 - 30/06/2020	-0.135	-0.135	0.888 [-26]	0.895 [-25]
01/07/2020 - 31/03/2021	0.079	0.137	0.190 [-40]	0.355 [-40]
01/04/2021 - 19/05/2022	0.625	0.554	0.658 [-11]	0.597 [-14]
New Admissions				
entire period	0.052	0.036	0.356 [-19]	0.346 [-20]
13/02/2020 - 30/06/2020	-0.081	-0.075	0.857 [-19]	0.853 [-20]
01/07/2020 - 31/03/2021	0.150	0.230	0.237 [-29]	0.388 [-32]
01/04/2021 - 19/05/2022	0.598	0.512	0.602 [+2]	0.513[-3]

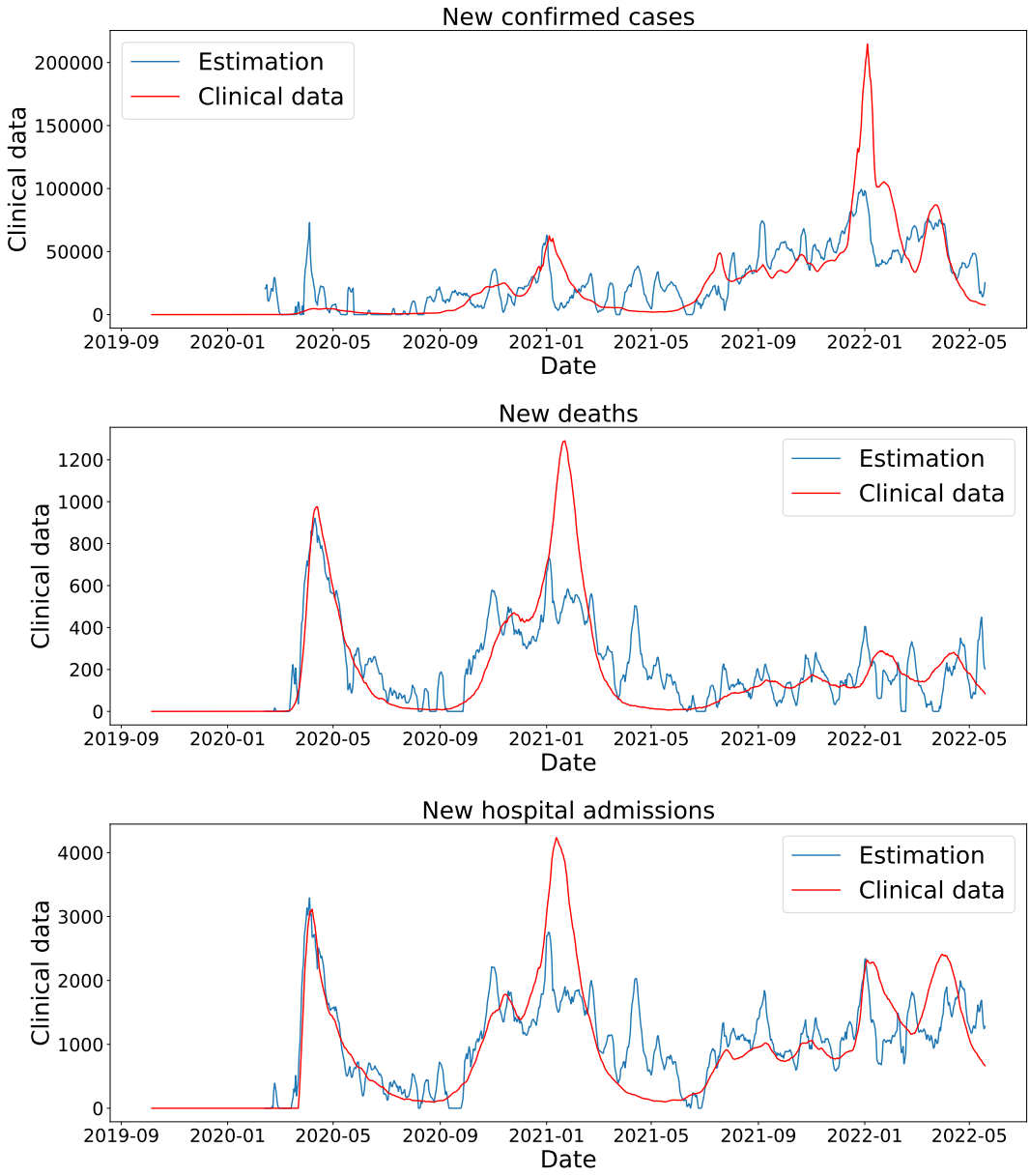
Note: see the note in the table 2

Cross-correlation of Search Data

Figure 14: Heat map of cross-correlation



Goodness-of-Fit of Linear Regression Model



Country Data Forecasting

Table 15: Linear Retraining model: New cases

	PER	**LRM-A**	**LRM-S**
Nowcasting			
MAE	6136.7	12200.6	7080.6
Correl.	0.928	0.858	0.878
7-day			
MAE	10878.7	11785.5	7112.2
Correl.	0.806	0.883	0.901
14-day			
MAE	14305.2	13470.6	7981.7
Correl.	0.693	0.724	0.893

Note: see the note in table 11

Table 16: Linear Retraining model: New admissions

	PER	**LRM-A**	**LRM-S**
Nowcasting			
MAE	188.9	305.5	192.6
Correl.	0.936	0.799	0.931
7-day			
MAE	358.0	320.2	209.3
Correl.	0.795	0.767	0.922
14-day			
MAE	497.0	271.8	204.4
Correl.	0.641	0.869	0.878

Note: see the note in table **??**

Regional Data Forecasting

Table 17: Reginal data forecasting: New deaths (pred. latest 12 months)

	PER	**LRM-A**	**LRM-S**	**LSTM-A**	**LSTM-S**
Nowcast					
m-MAE	0.316 (0.317)	0.312 (0.339)	0.295 (0.326)	0.318 (0.301)	0.303 (0.289)
m-Correl.	0.544	0.363	0.335	0.533	0.569
7-day					
m-MAE	0.357 (0.367)	0.495 (0.596)	0.494 (0.596)	0.377 (0.358)	0.335 (0.296)
m-Correl.	0.467	0.533	0.534	0.430	0.480
14-day					
m-MAE	0.401 (0.417)	0.495 (0.596)	0.494 (0.596)	0.415 (0.389)	0.399 (0.342)
m-Correl.	0.363	0.533	0.535	0.349	0.282

Note: see the note in table 13.

List of Regions

Region Names
Aberdeen City, Aberdeenshire, Angus, Antrim and Newtownabbey, Ards and North Down, Argyll and Bute, Armagh City, Banbridge and Craigavon, Barking and Dagenham, Barnet, Barnsley, Bath and North East Somerset, Bedford, Belfast City, Bexley, Birmingham, Blackburn with Darwen, Blackpool, Bolton, Bournemouth, Christchurch and Poole, Bracknell Forest, Bradford, Brent, Brighton and Hove, City of Bristol, Bromley, Buckinghamshire, Bury, Calderdale, Cambridgeshire, Camden, Causeway Coast and Glens, Central Bedfordshire, Cheshire East, Cheshire West and Chester, Clackmannanshire, Cornwall, Coventry, Croydon, Cumbria, Darlington, Derby, Derbyshire, Derry and Strabane, Devon, Doncaster, Dorset, Dudley, Dumfries and Galloway, Dundee City, Durham, County, Ealing, East Ayrshire, East Dunbartonshire, East Lothian, East Renfrewshire, East Riding of Yorkshire, East Sussex, City of Edinburgh, Eilean Siar, Enfield, Essex, Falkirk, Fermanagh and Omagh, Fife, Gateshead, Glasgow City, Gloucestershire, Greenwich, Hackney, Halton, Hammersmith and Fulham, Hampshire, Haringey, Harrow, Hartlepool, Havering, Herefordshire, Hertfordshire, Highland, Hillingdon, Hounslow, Inverclyde, Isle of Wight, Islington, Kensington and Chelsea, Kent, Kingston upon Hull, Kingston upon Thames, Kirklees, Knowsley, Lambeth, Lancashire, Leeds, Leicester, Leicestershire, Lewisham, Lincolnshire, Lisburn and Castlereagh, Liverpool, Luton, Manchester, Medway, Merton, Mid and East Antrim, Mid-Ulster, Middlesbrough, Midlothian, Milton Keynes, Moray, Newcastle upon Tyne, Newham, Newry, Mourne and Down, Norfolk, North Ayrshire, North East Lincolnshire, North Lanarkshire, North Lincolnshire, North Somerset, North Tyneside, North Yorkshire, Northamptonshire, Northumberland, Nottingham, Nottinghamshire, Oldham, Orkney Islands, Oxfordshire, Perth and Kinross, Peterborough, Plymouth, Portsmouth, Reading, Redbridge, Redcar and Cleveland, Renfrewshire, Richmond upon Thames, Rochdale, Rotherham, Rutland, Salford, Sandwell, Scottish Borders, Sefton, Sheffield, Shetland Islands, Shropshire, Slough, Solihull, Somerset, South Ayrshire, South Gloucestershire, South Lanarkshire, South Tyneside, Southampton, Southend-on-Sea, Southwark, St. Helens, Stirling, Stockport, Stockton-on-Tees, Suffolk, Sunderland, Surrey, Sutton, Swindon, Tameside, Thurrock, Torbay, Tower Hamlets, Trafford, Wakefield, Walsall, Waltham Forest, Wandsworth, Warrington, Warwickshire, West Berkshire, West Dunbartonshire, West Lothian, West Sussex, Westminster, Wigan, Wiltshire, Windsor and Maidenhead, Wirral, Wokingham, Wolverhampton, Worcestershire, York

References

[1] William Ogilvy Kermack and Anderson G McKendrick. A contribution to the mathematical theory of epidemics. *Proceedings of the royal society of london. Series A, Containing papers of a mathematical and physical character*, 115(772):700–721, 1927.

[2] Walter Andrew Shewhart. *Economic control of quality of manufactured product*. Macmillan And Co Ltd, London, 1931.

[3] Firuz Kamalov, Aswani Cherukuri, Hana Sulieman, Fadi Thabtah, and Akbar Hossain. Machine learning applications for covid-19: a state-of-the-art review. *arXiv preprint arXiv:2101.07824*, 2021.

[4] Jeffrey Shaman and Alicia Karspeck. Forecasting seasonal outbreaks of influenza. *Proceedings of the National Academy of Sciences*, 109(50):20425–20430, 2012.

[5] Jeffrey Shaman, Alicia Karspeck, Wan Yang, James Tamerius, and Marc Lipsitch. Real-time influenza forecasts during the 2012–2013 season. *Nature communications*, 4(1):1–10, 2013.

[6] You Won Lee, Jae Woo Choi, and Eun-Hee Shin. Machine learning model for predicting malaria using clinical information. *Computers in Biology and Medicine*, 129:104151, 2021.

[7] Mahmood Akhtar, Moritz UG Kraemer, and Lauren M Gardner. A dynamic neural network model for predicting risk of zika in real time. *BMC medicine*, 17(1):1–16, 2019.

[8] Sasikiran Kandula and Jeffrey Shaman. Near-term forecasts of influenza-like illness: An evaluation of autoregressive time series approaches. *Epidemics*, 27:41–51, 2019.

[9] Michael J Kane, Natalie Price, Matthew Scotch, and Peter Rabinowitz. Comparison of arima and random forest time series models for prediction of avian influenza h5n1 outbreaks. *BMC bioinformatics*, 15(1):1–9, 2014.

[10] Yuexin Wu, Yiming Yang, Hiroshi Nishiura, and Masaya Saitoh. Deep learning for epidemiological predictions. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 1085–1088, 2018.

[11] Siva R Venna, Amirhossein Tavanaei, Raju N Gottumukkala, Vijay V Raghavan, Anthony S Maida, and Stephen Nichols. A novel data-driven model for real-time influenza forecasting. *Ieee Access*, 7:7691–7701, 2018.

[12] Shihao Yang, Mauricio Santillana, John S Brownstein, Josh Gray, Stewart Richardson, and SC Kou. Using electronic health records and internet search information for accurate influenza forecasting. *BMC infectious diseases*, 17(1):1–9, 2017.

[13] Gitanjali R Shinde, Asmita B Kalamkar, Parikshit N Mahalle, Nilanjan Dey, Jyotismita Chaki, and Aboul Ella Hassanien. Forecasting models for coronavirus disease (covid-19): a survey of the state-of-the-art. *SN Computer Science*, 1(4):1–15, 2020.

[14] Pavan Kumar, Himangshu Kalita, Shashikanta Patairiya, Yagya Datt Sharma, Chintan Nanda, Meenu Rani, Jamal Rahmani, and Akshaya Srikanth Bhagavathula. Forecasting the dynamics of covid-19 pandemic in top 15 countries in april 2020: Arima model with machine learning approach. *MedRxiv*, 2020.

[15] Alok Kumar Sahai, Namita Rath, Vishal Sood, and Manvendra Pratap Singh. Arima modelling & forecasting of covid-19 in top five affected countries. *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, 14(5):1419–1427, 2020.

[16] Furqan Rustam, Aijaz Ahmad Reshi, Arif Mehmood, Saleem Ullah, Byung-Won On, Waqar Aslam, and Gyu Sang Choi. Covid-19 future forecasting using supervised machine learning models. *IEEE access*, 8:101489–101499, 2020.

[17] Michał Wieczorek, Jakub Siłka, and Marcin Woźniak. Neural network powered covid-19 spread forecasting model. *Chaos, Solitons & Fractals*, 140:110203, 2020.

[18] Abdelhafid Zeroual, Fouzi Harrou, Abdelkader Dairi, and Ying Sun. Deep learning methods for forecasting covid-19 time-series data: A comparative study. *Chaos, Solitons & Fractals*, 140:110121, 2020.

[19] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[20] Hasim Sak, Andrew W Senior, and Françoise Beaufays. Long short-term memory recurrent neural network architectures for large scale acoustic modeling. 2014.

[21] Hyunyoung Choi and Hal Varian. Predicting the present with google trends. *Economic record*, 88:2–9, 2012.

[22] Lisette De Vries, Sonja Gensler, and Peter SH Leeflang. Popularity of brand posts on brand fan pages: An investigation of the effects of social media marketing. *Journal of interactive marketing*, 26(2):83–91, 2012.

[23] Nikos Askitas and Klaus F Zimmermann. Google econometrics and unemployment forecasting. 2009.

[24] Vasileios Lampos and Nello Cristianini. Nowcasting events from the social web with statistical learning. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 3(4):1–22, 2012.

[25] Mauricio Santillana, André T Nguyen, Mark Dredze, Michael J Paul, Elaine O Nsoesie, and John S Brownstein. Combining search, social media, and traditional data sources to improve influenza surveillance. *PLoS computational biology*, 11(10):e1004513, 2015.

[26] Edward Velasco, Tumacha Agheneza, Kerstin Denecke, Goeran Kirchner, and Tim Eckmanns. Social media and internet-based data in global systems for public health surveillance: a systematic review. *The Milbank Quarterly*, 92(1):7–33, 2014.

[27] Philip M Polgreen, Yiling Chen, David M Pennock, Forrest D Nelson, and Robert A Weinstein. Using internet searches for influenza surveillance. *Clinical infectious diseases*, 47(11):1443–1448, 2008.

[28] Andrea Freyer Dugas, Mehdi Jalalpour, Yulia Gel, Scott Levin, Fred Torcaso, Takeru Igusa, and Richard E Rothman. Influenza forecasting with google flu trends. *PloS one*, 8(2):e56176, 2013.

[29] Vasileios Lampos, Maimuna S Majumder, Elad Yom-Tov, Michael Edelstein, Simon Moura, Yohhei Hamada, Molebogeng X Rangaka, Rachel A McKendry, and Ingemar J Cox. Tracking covid-19 using online search. *NPJ digital medicine*, 4(1):1–11, 2021.

[30] Thomas S Higgins, Arthur W Wu, Dhruv Sharma, Elisa A Illing, Kolin Rubel, Jonathan Y Ting, and Snot Force Alliance. Correlations of online search engine trends with coronavirus disease (covid-19) incidence: infodemiology study. *JMIR public health and surveillance*, 6(2):e19702, 2020.

[31] Hien Lau, Tanja Khosrawipour, Piotr Kocbach, Hirohito Ichii, Jacek Bania, and Veria Khosrawipour. Evaluating the massive underreporting and undertesting of covid-19 cases in multiple global epicenters. *Pulmonology*, 27(2):110–115, 2021.

[32] Shailesh Bavadekar, Andrew Dai, John Davis, Damien Desfontaines, Ilya Eckstein, Katie Everett, Alex Fabrikant, Gerardo Flores, Evgeniy Gabrilovich, Krishna Gadepalli, et al. Google covid-19 search trends symptoms dataset: Anonymization process description (version 1.0). *arXiv preprint arXiv:2009.01265*, 2020.

[33] Nicola L Boddington, Andre Charlett, Suzanne Elgohari, Jemma L Walker, Helen I McDonald, Chloe Byers, Laura Coughlan, Tatiana Garcia Vilaplana, Rosie Whillock, Mary Sinnathamby, et al. Covid-19 in great britain: epidemiological and clinical characteristics of the first few hundred (ff100) cases: a descriptive case series and case control analysis. *MedRxiv*, 2020.

[34] Vasileios Lampos, Andrew C Miller, Steve Crossan, and Christian Stefansen. Advances in nowcasting influenza-like illness rates using search query logs. *Scientific reports*, 5(1):1–10, 2015.

[35] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, 67(2):301–320, 2005.

[36] Tomasz Szmuda, Shan Ali, Tarjei Vevang Hetzger, Philip Rosvall, and Paweł Słoniewski. Are online searches for the novel coronavirus (covid-19) related to media or epidemiology? a cross-sectional study. *International Journal of Infectious Diseases*, 97:386–390, 2020.