Supporting the development of an A&E model for Public Health Scotland

Toyo Sadare, Jake Skelton, Keren Tapper, Ross Walker

Supervised by
Dr Burak Buke, Dr Torben Sell

MAC-MIGS
University of Edinburgh and Heriot-Watt University

In Collaboration with

Hannah Jones, Ken Nicholson
Public Health Scotland

Spring 2024

Abstract

Public Health Scotland aim to produce a whole system model of the flow through the health and social care system to improve health and social care services in Scotland. To aid this, we look at different methods of modelling the number of people attending A&E, using open data.

The first method we look at is a Toy model that only assumes yearly periodicity by using regression to fit the data to a year-periodic sine. It is statistically the most accurate of the three models for predicting the number of people attending A&E per week at the small Health Board Trusts (as shown in the table below). This we generalise to a second, stochastic model that incorporates several periodic components in the framework of Poisson regression. This gives more accurate predictions, but can overfit, and assumes totally periodic A&E attendances. A trained Neural Network provides the best predictions for the larger Health Board Trusts, such as Greater Glasgow and Clyde, Lothian, and for the whole of Scotland, whereas the Toy model provides the best prediction for smaller HBTs such as Orkney.

Model	MSE_{Scot}	$\text{MSE}_{GG\&C}$	MSE_{Orkney}
Toy	1,705,123	133,602	253
FFT	517,776	68,295	443
NN	339,035	60,113	1,109

These three models all have benefits and drawbacks. Combining these models provides an accurate prediction for the number of A&E attendances in the future year with the most accuracy. Large HBTs should be modelled differently to the smaller HBTs and finding an appropriate ratio between models can provide a suitable model that incorporates these differences.

Authors Contributions

Jake, Keren, and Ross were equal contributors in the writing of the report. Jake wrote the section on the Fourier Series model and produced the map of Scotland with hospitals and HBTs. Keren wrote the Comparison, Conclusion, Data Analysis, and Toy Model Section. Ross wrote the Introduction, Machine Learning section, and Appendix A: Bayesian Neural Networks. Toyo produced the weekly interpolated data which was used in the models.

Contents

1. Introduction 4
2. Data Analysis 5
2.1. Deprivation 6
2.2. Age 6
2.3. Sex 8
2.4. Hospitals 8
2.5. Other events 8
3. Modelling 9
3.1. Toy Model 9
3.1.1. Monthly Model 10
3.1.2. Monthly Model Evaluation 13
3.1.3. Weekly Model Evaluation 14
3.2. Fourier Series Model 15
3.2.1. Training Procedure 17
3.2.2. Results 17
3.2.3. Model Evaluation 19
3.3. Machine Learning 20
3.3.1. Deep Neural Network Structure 21
3.3.2. Training Procedure 22
3.3.3. Model Evaluation 24
3.3.4. Next Steps for the Neural Network 26
4. Comparison 28
4.1. Accuracy 29
4.2. Simplicity 29
4.3. Future Predictions 30
4.4. Combining Models 30
5. Conclusion 31
References 33
Appendix A. Bayesian Neural Networks 36

List of Figures

1 Map of Health Board Trusts 4
2 Monthly A&E attendances by age group 7
3 Monthly Toy model fitted to 2018 proportions of attendances 11
4 Coefficients of Toy model calculated for 2018 against population size 12
5 Coefficients of Toy model calculated for 2018 against age and deprivation 13
6 Weekly Toy Model predictions 16
7 Predictions of the Fourier series model for small, medium, and large HBTs 18
8 Loss function convergence for ADAM and AMSGrad optimisation methods 23
9 Greater Glasgow and Clyde NN model prediction 24
10 Greater Glasgow and Clyde NN model predictions for Ages 25-39 25
11 Greater Glasgow and Clyde NN model predictions for Under 18s 25
12 Orkney NN model predictions 26
13 Whole Scotland NN model predictions 27
14 Comparing different models' predictions for 2023 28
15 Combining different models' predictions for 2023 31

List of Tables

1 HBT population deprivation, age, and size 6
2 RMSE of predictions using 2018 data 14
3 RMSE and MSE_{Scot} for 2023 predictions using different training years 15
4 MSE_{Scot} of Fourier series model with different numbers of modes 17
5 Periodic components of A&E attendance data learned by the Fourier series model 19
6 MSE by age group for the neural network model 26
7 MSE_{Scot} for different models in predictions for 2023 29
8 MSE for combined model predictions for 2023 30

List of Abbreviations

Health Service

A&E	Accident and Emergency Department
HBT	Health Board Trust
MIU	Minor Injuries Unit
NHS	National Health Service (Scotland)
PHS	Public Health Scotland

Modelling Tools

CI	Confidence Interval
DFT	Discrete Fourier Transform
DNN	Deep Neural Network
FFT	Fast Fourier Transform
GLM	Generalised Linear Model
MLE	Maximum-likelihood Estimator
MSE	Mean Squared Error
NN	Neural Network

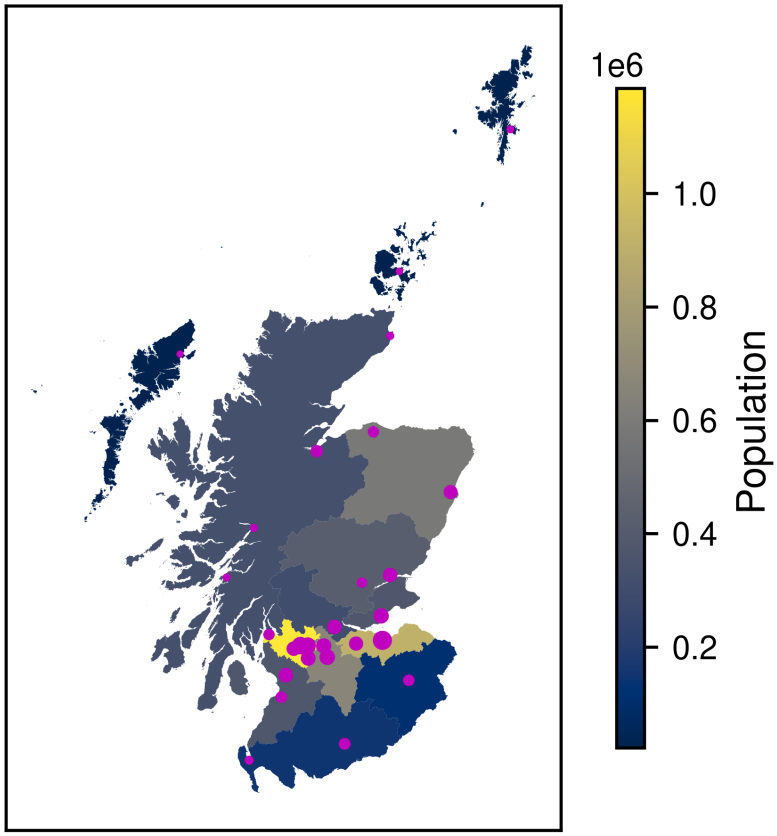


FIGURE 1. The Health Board Trusts and their population size are shown on a map of Scotland. Emergency departments and Minor Injury Units are shown with dots proportional in size to the number of A&E attendances per week.

1. INTRODUCTION

Public Health Scotland (PHS) is an NHS board that uses specialist knowledge and data analysis to influence health board planning processes across Scotland, with the aim of reducing health inequalities [1]. With this goal in mind, PHS are dedicating significant resources to building a mathematical model of patients through Scottish hospitals. Currently, they have an efficient model to predict hospital bed usage across different hospital departments, incorporating how long beds will be occupied given the number of people arriving at a hospital. Such a model is important in the everyday running of hospitals, given the immense strain on resources and bed numbers. The ability to accurately forward-plan bed usage is completely fundamental to ensuring resources are well distributed throughout the system. A key ambition in the future development of such models is to also predict daily attendances at each individual hospital, alongside being able to show predictions for the whole of Scotland.

Our aim is to build predictive models for the number of people attending A&E departments across Scotland. The dynamics of A&E departments are quite distinct from other hospital departments, having significantly more associated randomness and entirely different functions. Making use of publicly available data, we make use of a range of data-driven modelling methods each with a different level of complexity, accuracy, and computational expense. We present here three models, varying from the minimal parameter setting of regression to the large parameter setting of machine learning. The primary idea is to present models which can be used by PHS in implementation both individually and in combination. Throughout, we have a focus on future development work that could be conducted. A key focus is on model granularity - the more specific our models can be (while maintaining accuracy) the more useful our models are in application. Granularity in this context can mean predictions for shorter intervals of time, more

specific locations or containing additional breakdowns of the type of individuals predicted to attend.

Similarly to the models existing for other departments, the models presented in this report are made with the intention of being used for allocation of resources. In particular, the models provide expected attendance numbers to be used in decision making regarding bed availability and distribution of labour. Given this purpose, we discuss the level of uncertainty quantification in our results (through means of confidence intervals, for example). This is particularly important in our context, where the optimisation of resource allocation is fundamental to performance of the health system.

We initially provide our own exploratory data analysis, which motivates discussions of our model results. We describe which features we have access to and the basic trends observed in the data. We describe events which produce outliers in the data and discuss the significance of these outliers. We consider here the impact of the COVID pandemic upon A&E attendance trends throughout Scotland.

Our first model is a regression model, which attempts to identify and represent the periodic behaviours exhibited in the data. Certain aspects of the fitted model, such as additive in population size, are imposed to ensure physical realism of the model. This model is produced both at the weekly and monthly level and provides attendance predictions at the level of Health Board Trust. Our second model utilises the success of the Toy model's periodic assumptions to fit a truncated Fourier series to the data. This yields higher accuracy in general but this comes with a degree of overfitting. This model was able to output weekly level predictions on each HBT. Our final presented model is a Neural Network, which dramatically increases the model complexity. This model outputs predictions for each health board trust at a weekly level, broken down by age group. At a national level this model has the best performance, with no apparent sign of significant overfitting. This model is beaten by our previous models in areas with lower population size. The presented network shows strong results when trained using only post-COVID data and even without data standardisation. Hence, this model shows significant promise for the future as more data becomes available. The latter two models predictions can be aggregated to monthly predictions, if required.

We conclude with a thorough comparison of the presented models in both qualitative and quantitative terms. A linear combination of the models is considered, which is found using regression techniques. This combined model outperforms each individual model in terms of error, and can only improve with time as more usable data becomes available.

2. Data Analysis

Due to privacy policies, we are only able to access publicly available data. There are two main files that we are looking at: one has monthly data for the number of attendances in each Health Board Trust (HBT) and includes demographic data on attendees, the other has weekly data for the number of attendances in each hospital and includes the length of stay in A&E before either being discharged or moved to an appropriate ward. Each HBT has a name and a code beginning with "S080000". They all have different proportions of demographics attending A&E [2].

To be able to incorporate the different demographics attending A&E into the weekly model predictions, we found the proportion of each feature and split the number of weekly attendances appropriately. For example, if there were 10 females under 18, and with a deprivation level 1 attending A&E in Tayside out of 200 attendances in a given month, and we know that 20 people attended A&E in the first week of that month, we assumed that 1 under 18 female from deprivation level 1 attended. Naturally, this method means that in many cases half a person attended. PHS have access to weekly, and even daily, data that has this information recorded and can apply the modelling methods we use below to this data.

HBT	HBT Name	Normalised Deprivation	Normalised Age	Population Size
S08000015	Ayrshire and Arran	**2.484**	43.85	369,670
S08000016	Borders	3.651	53.91	115,270
S08000017	Dumfries and Galloway	3.077	47.97	148,790
S08000019	Forth Valley	3.250	48.52	306,070
S08000020	Grampian	3.894	47.70	584,550
S08000022	Highland	4.090	**62.92**	321,800
S08000024	Lothian	3.641	50.66	897,770
S08000025	Orkney	**4.277**	58.86	**22,190**
S08000026	Shetland	4.178	55.43	22,990
S08000028	Western Isles	3.551	58.38	26,830
S08000029	Fife	2.929	44.41	371,910
S08000030	Tayside	3.714	53.28	416,080
S08000031	Greater Glasgow and Clyde	2.608	**42.67**	**1,174,980**
S08000032	Lanarkshire	2.529	47.67	659,200

TABLE 1. HBT codes with names, deprivation coefficient, age coefficient, and population size in 2018. The maximum and minimum for each category are in bold. Orkney has both the smallest population and has the highest deprivation coefficient (is least deprived). Greater Glasgow and Clyde has the highest population age and population size.

2.1. **Deprivation.**

Each A&E attendee registered at a GP in Scotland has their deprivation level recorded on a scale from 1 to 5, where 1 is the most deprived. This is based on their home address. Scotland is split in 6,976 areas (data zones) which are ranked on the Scottish Index of Multiple Deprivation [3]. This ranking is based on income, employment, health, education, housing, access, and crime in that data zone, each weighted differently. In the health category of the deprivation ranking, factors include alcohol misuse related hospital stays; drug misuse related hospital stays; mortality; and Emergency hospital stays.

Naturally, one would expect an area with a low deprivation to have a high number of A&E attendance because of the influence of these factors.

For each HBT, we define a deprivation coefficient for people attending A&E by:

$$\mathrm{Dep}(H) = \sum_{d=1}^{5} p_H(d) \times d$$

where $p_H(d)$ is the proportion of people attending A&E from an area with deprivation level d attending A&E in a given year for a HBT, H.

HBTs vary greatly in deprivation level, as shown in Table 1. The most deprived areas are Ayrshire and Arran, Greater Glasgow and Clyde, and Lanarkshire, which are all in South West Scotland. The least deprived areas are Orkney, Shetland, and the Highlands.

2.2. **Age.**

In the monthly data, the age categories are:

(1) Under 18
(2) 18-24
(3) 25-39
(4) 40-64
(5) 65-74
(6) 75 plus

These age brackets are all of different ranges of ages. Part of this choice may be due to the typical causes of attendances for people within those brackets. For example, older people are more likely to attend A&E for complication from falls, and young people are the most likely to attend for sports related injuries [4].

To standardise, we choose the midpoint of each age bracket and choose 87.5 for the 75 plus bracket. As observed in Table 1, the normalised age in different areas is roughly correlated with deprivation. The lowest normalised age of A&E attendees is in Greater Glasgow and Clyde. This also corresponds to the lowest life expectancy in Scotland.

This may partly be due to the location of children's hospitals. There are children's hospitals in Aberdeen (Grampian), Glasgow (Greater Glasgow and Clyde), and Edinburgh (Lothian). It might also be due to the locations of large universities, the largest of which are located in the aforementioned three cities.

As shown in Figure 2, Under 18s also have different trends for attending A&E compared to other ages. These appear to be correlated with school holidays. Major school holidays fall in December, April, and August with minor ones in October, February, and May [5]. The Under 18 and 18-24 age categories are also much more likely to attend A&E for sports related injuries [6]. These injuries are often minor and don't require hospitalisation.

Similarly, older patients are more likely to attend A&E during the winter [7] and are often more likely to require hospitalisation. This is partly the cause of larger delays in A&E in winter rather than summer.

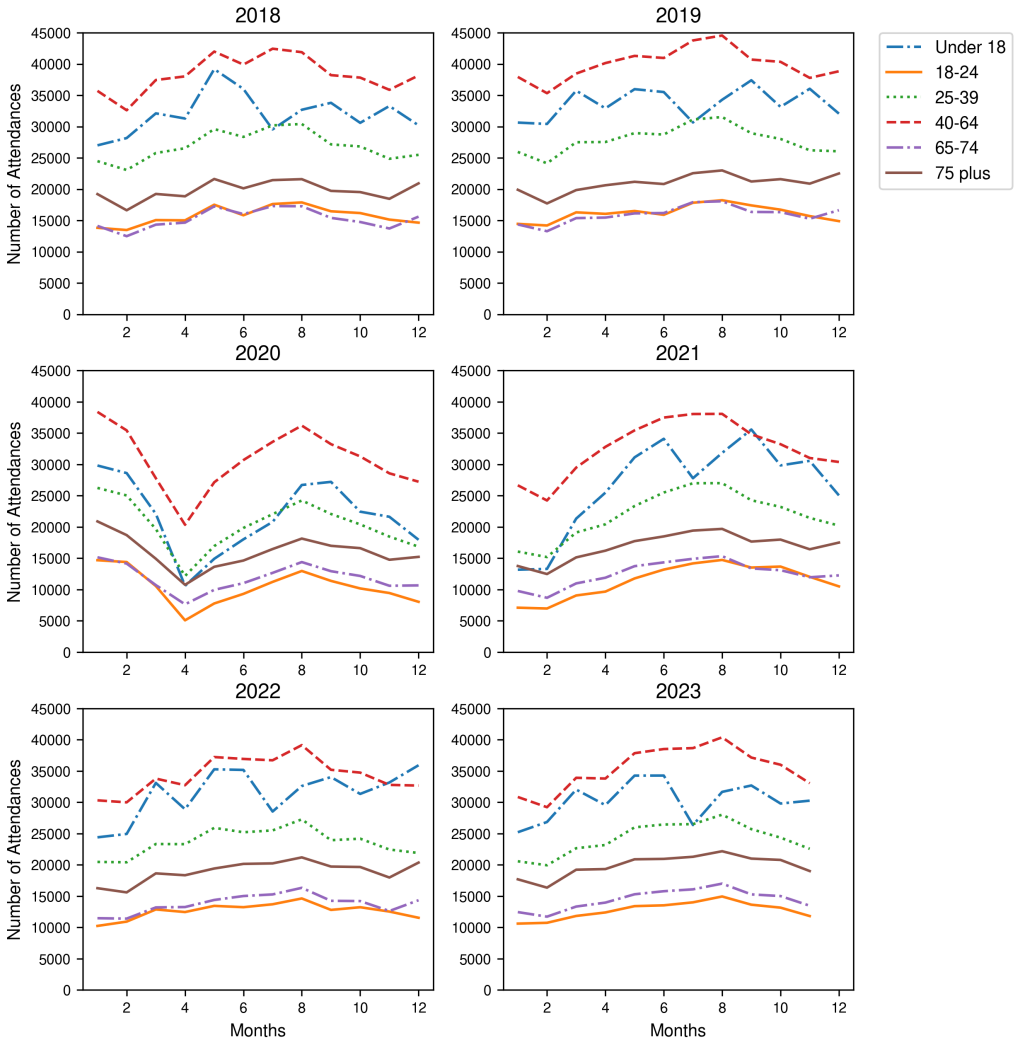


FIGURE 2. The monthly number of A&E attendances for each age category over the 6 years for which we have data. There was no published data for Dec 2023. Most people attending A&E are aged between 40 and 64. Under 18*s* have the largest variation. There are less people in the 18 – 24 age group attending A&E than any other age group.

2.3. **Sex.**

In Scotland, there were 2,728,000 women and 2,567,400 men in 2011 [8]. However, in almost all council areas and almost all months, more male patients attend A&E than female. Male and female people are more likely to attend A&E for different reasons. It has been observed that 90% of patients attending A&E for sports-related injuries are male [4]. The female life expectancy is also higher than the male life expectancy so we expect different proportions of males and females attending A&E around those ages.

2.4. **Hospitals.**

A&E attendees can be seen at either an Emergency Department or a Minor Injury Unit (or other). There is only hospital data available in weekly numbers of Emergency Department attendances. Some hospitals have much larger numbers of patients than others. The hospitals which see the largest number of patients are the Royal Infirmary of Edinburgh (S314H), Glasgow Royal Infirmary (G107H), and the Queen Elizabeth University Hospital (G405H). These receive a lot more patients that the smaller Island hospitals such as the Western Isles Hospital (W107H).

Some hospitals also contain Minor Injury Units, however the hospital location of Minor Injury Unit attendances is not recorded. It is available on monthly data for each HBT. Not all HBTs have Minor Injury Units. However, if there is a Minor Injury Unit, approximately 19.85% of A&E attendances are in the Minor Injury Unit, with a standard deviation of 5.62%. Looking at each HBT over 6 years, there is an average standard deviation of 2%.

We expect the number of A&E attendances to be dependent on the size of the hospitals and whether there is a Minor Injuries Unit. 88% of people in Scotland are within 30 minutes of an A&E department, so an area with a higher density of hospitals is more likely to have people attending A&E [9].

2.5. **Other events.**

Major events can have an effect on the number of people attending A&E. Black Swan events are events which have a major effect but are unexpected. Such events are very hard to predict ahead of time and can cause major errors in predictions [10]. The 2008 Financial Crash was a Black Swan event that caused an increase in A&E attendances, as well as a reduction in the NHS yearly budget increase [11] from 2010 when austerity was introduced.

A problem that we will have to face is that the patterns of A&E attendances changed greatly during COVID lockdowns in 2020 and 2021 (although it is argued that this is not a black swan event as a global pandemic was predicted [12, 13]). The dates of the COVID lockdowns, and thus the data which is most problematic, are as follows [14]:

- 2020
  - March to June: First National lockdown
  - November: Second National lockdown
- 2021
  - January to March: Third National lockdown

During these times, the number of people attending A&E reduced greatly, especially in April 2020. This is partly due to people questioning their need, and due to lack of interactions that might require emergency medical attention. The age group that saw the largest drop in attendances was the Under 18 category, as shown in Figure 2. A suggestion for the cause of this could be reduced amount of time playing sports or in playgrounds with other children. The age groups least affected by the COVID pandemic were the $65-74$ and 75 plus categories.

Before COVID, the number of people attending A&E each year was slowly increasing. However, post-COVID levels are still significantly below these levels. This makes future predictions difficult because the long term trends have changed.

Furthermore, in 2022 and 2023 (post-COVID), the data is a lot noisier; there is a greater variation in the number of people attending, even for two consecutive weeks. This means that

the error in a model is likely to be a lot higher for these years, and makes predictions less accurate. Part of the reason for this noise could be the British Medical Association junior doctor and doctors strikes. These were (somewhat) randomly dispersed through 2022 and the first half of 2023. During these times, people might have been less likely to attend A&E for more minor injuries so as not to overwhelm the system. People might also be more likely to not have their demographics during this time as well.

Other events that might drastically change the number of A&E attendances would be large storms, where there are floods, fallen trees, and high winds.

3. Modelling

We will now look at three different methods of predicting A&E attendances. Different modelling approaches give different understandings into trends [15]. For example, Neural Network and Time Series models predict solely using previous data. They are very accurate in their predictions but will not work if the dynamics of the system changes, for example a black swan event such as COVID, or an influx of immigration causing a change in the population demographics. Simpler models such as fast Fourier Transform models will give more insight in trends in the data at different levels, such as yearly periodicity and periodicity around school term time.

Different models might also be more accurate at different size of HBTs, or more accurate for the whole of Scotland. To allow comparison, we introduce the error estimator $\mathrm{MSE}_{\mathrm{Scot}}$, [15] which we define as follows:

$$\begin{aligned}
\mathrm{MSE}_{\mathrm{Scot}} &= \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2 \\
&= \frac{1}{N}\sum_{i=1}^{N}\left(\sum_{h=1}^{H}\hat{y}_i^h - \sum_{h=1}^{H}y_i^h\right)^2 \\
&= \frac{1}{N}\sum_{i=1}^{N}\left(\sum_{h=1}^{H}(\hat{y}_i^h - y_i^h)\right)^2 \\
&= \frac{1}{N}\sum_{i=1}^{N}\sum_{h,h'}^{H}\left(\hat{y}_i^h - y_i^h\right)\left(\hat{y}_i^{h'} - y_i^{h'}\right)
\end{aligned} \tag{1}$$

where N is the number of weeks in a year, H is the number of Health Board Trusts, ($H = 14$). $\hat{y}_i^h$ is the observed data and y_i^h is the calculated value for a given time i and HBT h.

We also expect that different models are better at picking up different trends. If we find that some models are much better at modelling different categories, it could be possible to combine them.

3.1. **Toy Model.**

A Toy Model is a model that has all the parameters stripped back as far as reasonably possible to be able to observe the most general trends expected in a model [16]. In this case, we will make the assumption that the only parameter affecting the number of people attending A&E is the time and size of the population. We will assume that this is year-periodic.

We start with a simple model which only considers the month or week, t. Given we want the model to be periodic in t, we expect that the number of people attending A&E at time t is given by

$$y(P,t) = A(P) + B(P)\sin(ct + d). \tag{2}$$

Here, A represents the monthly (or weekly) average number of people attending A&E in one period, B represents the amplitude of change, c and d are the period and phase shift, and P is the population size.

We want this to be additive, such that

$$\begin{aligned} y(P_1 + P_2, t) &= A(P_1 + P_2) + B(P_1 + P_2)\sin(ct + d) \\ &= (A(P_1) + A(P_2)) + (B(P_1) + B(P_2))\sin(ct + d) \\ &= A(P_1) + B(P_2)\sin(ct + d) + A(P_1) + B(P_2)\sin(ct + d) \\ &= y(P_1, t) + y(P_2, t) \end{aligned}$$

because we should be able to look at a group of HBTs and predict how many people will attend A&E. Therefore, to satisfy this condition, we require $A(P) = aP$ and $B(P) = bP$ for some a, b. We can now normalise to get that the proportion of people attending A&E is

$$n(t) = a + b\sin(ct + d) \tag{3}$$

where we want that $n(0) = n(p+1)$, where p is the number of time steps in a year (for a monthly model, $p = 12$; for a weekly model, $p = 52$ or 53). Because this is normalised and shouldn't depend on the population size, P,

$$\int_0^P n(t)dP = n(t)P = y(P, t).$$

The population of Scotland is currently increasing [8], although it is expected to begin decreasing in the next 10 years, and we expect that the number of people attending A&E will change proportionally. Therefore we don't expect the proportion to change as the years do, unless to show a change in the trend of people attending A&E, such as during the COVID pandemic.

This model assumes that every area in Scotland has people of the same demographics, with the same distance to A&E. This is not true and over simplifies the number of attendances, so we will assume that every HBT has a slightly different proportion of people attending A&E and increase the complexity of the model to try and take into account different demographics later. Take $n_H(t) = a_H + b_H\sin(c_H t + d_H)$ to be the proportion of people attending A&E at time t, in Health Board H.

3.1.1. *Monthly Model.*

Now, take the time step t to be months. To fit this model to our data points, we use regression to minimise the residuals. In other words,

$$(a_H, b_H, c_H, d_H) = \text{argmin}_{a_H, b_H, c_H, d_H} \quad \frac{1}{12}\sum_{t=1}^{12}(\hat{n}_H(t) - n_H(t))^2 \tag{4}$$

where $\hat{n}_H$ is our observed proportion of attendances for HBT H, and $a, b \in [0, 1]$, $c, d \in [0, 2\pi)$.

In Figure 3, we can observe that the proportion of people attending A&E is not the same in different Health Boards. Greater Glasgow and Clyde has more than double the proportion of people attending A&E in some months than Grampian.

The average root mean square error is given by:

$$\begin{aligned} \text{RMSE}_{av} &= \frac{1}{H}\sum_{h=1}^{H}\text{RMSE}_h \\ &= \frac{1}{H}\sum_{h=1}^{H}\sqrt{\frac{1}{N}\sum_{t=1}^{N}(\hat{n}_h(t) - n_h(t))^2} \end{aligned}$$

In 2018, the average RMSE for this model is 0.000842. In Greater Glasgow and Clyde (S0800031), the largest HBT, there are usually around 35,000 people attending A&E each

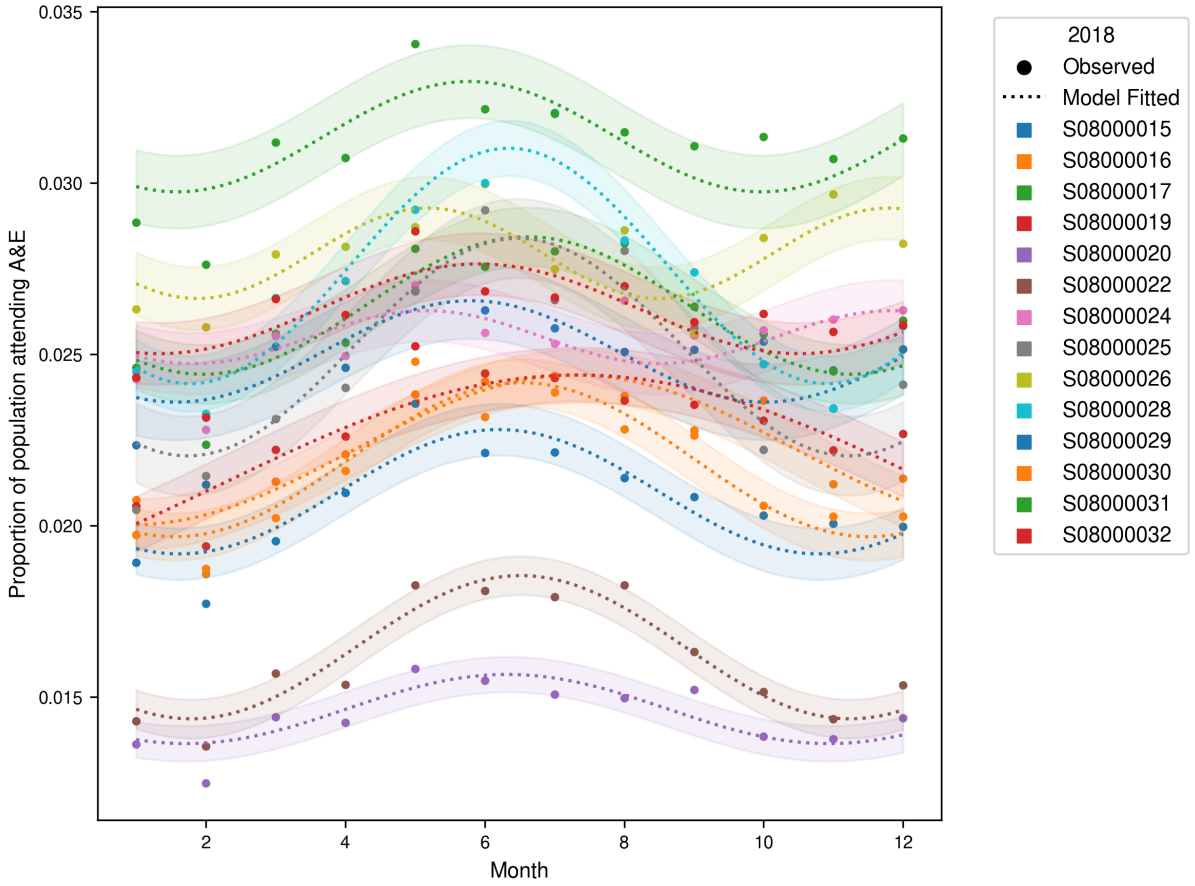


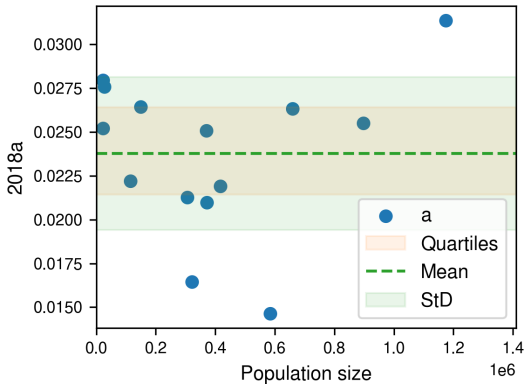
FIGURE 3. The Toy model is fitted (dotted line) to each HBT in 2018. HBT S08000031 (Greater Glasgow and Clyde) has the highest proportion of people attending A&E. The lowest proportions are HBT S08000020, 22 (Grampian and Highland). Grampian also has the lowest MSE.

month. The root mean square error is equivalent to 989 people in Greater Glasgow and Clyde. Therefore, there is a 68.3% chance of the number of people attending A&E, $\hat{n}_{31}$, satisfying $n_{31}(t) - 989 < \hat{n}_{31}(t) < n_{31}(t) + 989$. The smallest HBT, Orkney, only has a population of 22, 190 with an average monthly number of attendances of about 550 so the RMSE_{av} is equivalent to 19 people in Orkney.

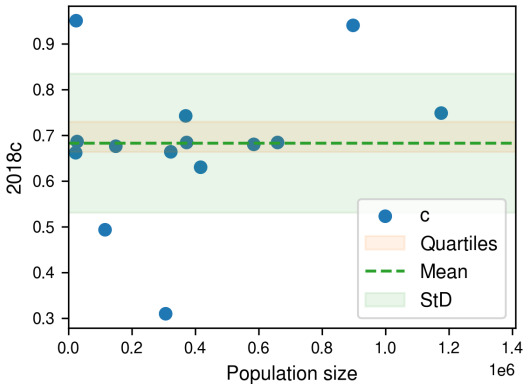
Fitting parameters alone does not provide a good indication of any physical trends or patterns that could be used to predict future A&E attendances. To improve this, we compare the parameters to the size of the population to determine if there is any correlation. Areas with higher populations are expected to have a higher proportion of A&E attendances as these populations are usually in densely populated cities with easy access to A&E departments [9]. Almost 90% of people in Scotland are within 30 minutes of an A&E or Minor Injuries Unit. Due to high demand for GP appointments, those with easy access to A&E departments are more likely to go to A&E than a GP.

However, in Figure 4, we can see that there isn't a strong correlation between population size and the coefficients. Applying linear regression does not significantly improve the error compared with the mean for a, c, d. We can improve the error on b by assuming that it is quadratic in population size. In 2018, the standard deviation of b from the mean is 0.000788 and by taking

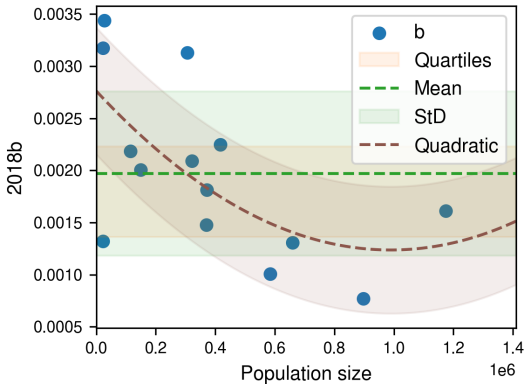
$$b(x) = (1.56 \times 10^{-15})x^2 - (3.08 \times 10^{-9})x + (2.76 \times 10^{-3})$$



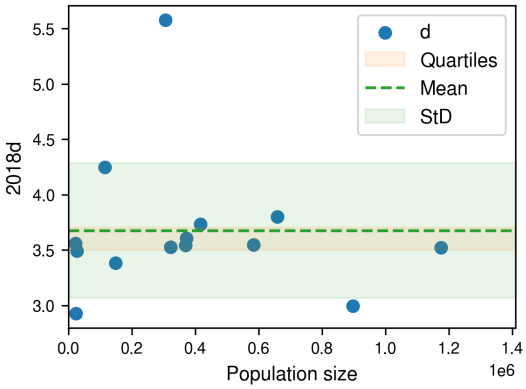
(A) a from 2018 against population



(B) c from 2018 against population



(C) b from 2018 against population



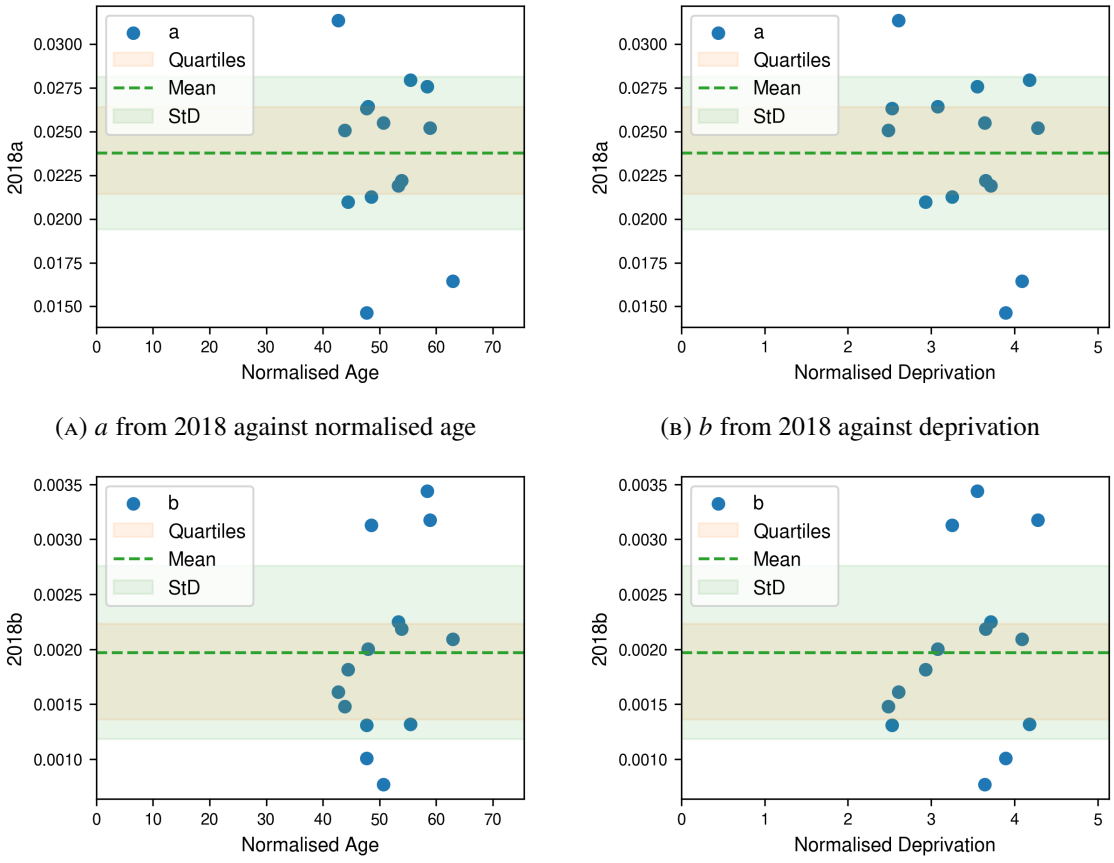
(D) d from 2018 against population

FIGURE 4. The calculated values of a,b,c,d in 2018 for each HBT with upper and lower quartiles (orange) and standard deviation (green) are plotted against population size. The quartiles are close to the mean for c and d. For b, a quadratic has also been fitted and RMSE has been plotted. RMSE is improved significantly by fitting to quadratic instead of the mean.

where x is the population size of the HBT, we get a RMSE of 0.000609. These coefficients are calculated using regression for each year.

It can also be expected that older people are more likely to go to A&E, especially during winter months [7]. Therefore, it is valid to assume that an area with more older people (or an area with a higher normalised age) will see a higher proportion of people attending A&E, especially during winter. We would expect to see a reduced amplitude b and a higher average a. It would be useful to find a relationship between age and the number of people attending A&E because in Scotland, the population is aging [8]. This means that the proportion of older people in the population is increasing. Understanding such a relationship would help predict how many people attend A&E once the proportions are significantly different.

However, as with comparing to population size, the values of a and b do not have a strong correlation with normalised age and normalised deprivation, shown in Figure 5. Other factors that could be affecting this but that we do not have the data to compare are tourism and hospital catchment areas. Specifically, 88% of the Scottish population are within 30 minutes of an Emergency Department or Minor Injuries Unit. People who are close to A&E departments are more likely to attend [9]. Thus, more densely populated areas are more likely to have a high level of attendances. Specifically, although Grampian has a large population and a major city (Aberdeen), much of the area is sparsely populated, which might explain the significantly lower proportion of A&E attendances, compared to areas with a similar population size but smaller area (such as Lothian).



(A) a from 2018 against normalised age

(B) b from 2018 against deprivation

(C) c from 2018 against normalised age

(D) d from 2018 against deprivation

FIGURE 5. Values of a and b for each HBT with upper and lower quartiles (orange) and standard deviation (green).

Areas that see a large amount of tourism only in the summer, such as Shetland, Western Isles, and Highlands, have some of the largest amplitudes. We know that this is due to some level of tourism as people not registered at a GP in Scotland do not have their demographic data recorded and in these areas we see a large rise in this number of patients with unrecorded data. Also, much of the population of Lanarkshire and the Forth Valley are close to hospitals in different HBTs (specifically Greater Glasgow and Clyde, and Lothian). Therefore, we expect many of them to attend those hospitals. We would also expect, for example, more farming related injuries in areas with more farmland. These factors, can be assumed to stay the same over different years.

To incorporate this, an assumption would be that the ranking of a stays the same and that there is some linear transformation happening due to environmental trends. For example, more people might attend A&E when there is the cost of living crisis and suffering from cold related ailments. Such trends are assumed to affect every area equally.

However, this is not the case either. Throughout the years the a neither has a constant order nor a linear function. Splitting the a values into small, medium, and large HBTs, we do not observe trends that are consistent, nor even that the order of a for a set of HBTs remains the same. Because of this, we do not have the additivity by population size and demographics that we hoped to have.

3.1.2. *Monthly Model Evaluation.*

In an attempt to predict future years, we can apply this model to later years and calculate the root mean square error. However, this is not very accurate, as shown in Table 2. Very quickly, this model becomes prohibitively inaccurate. This is partly due to the fact that the trends of

A&E attendances changed during the COVID pandemic and have not recovered to similar levels since then.

Year	RMSE_{av} fitted to year	RMSE_{av} against 2018 model	$\text{RMSE}_{\text{GG\&C}}$ against 2018 model
2019	0.000799	0.00162	1900
2022	0.00133	0.00214	2520
2023	0.00101	0.00277	3250

TABLE 2. The 2018 monthly model does not provide a good estimate for proportion of monthly attendances future years. The years 2020 and 2021 are removed as due to COVID, we do not expect them to have the same patterns as other years.

This accuracy of the predictions can be improved significantly by only looking one year into the future. Although looking far into the future can be useful to predict the demand for hospitals and if new hospitals need to be built to cope with demand, looking one year ahead can be useful in the short term for predicting problems in demand.

However, some HBTs for some years took too many iterations to find reasonable coefficients. For HBTs where this was the case, we took an average of "working" HBTs' c, d and re-calculated a, b with these as fixed coefficients. This did not greatly change the accuracy of the results. This is expected as because the area of Scotland is not large, we do not expect seasonal trends to be different in different areas. Countries such as Chile or Russia might have more of an effect from this averaging due having many different climates [17].

If we had a method of predicting a, we would be able to look further in the future, especially given there is not a lot of change in b, c, d throughout the years. If we had data that went back many years, we would be able to look at a longer time scale and find a linear regression for a, or some machine learning methods.

For each HBT, the proportion of the sex, deprivation, and hospital demographics out of the total number of people attending A&E does not change significantly throughout the year, with only minor long term changes due to changes in the population over the years. The one exception is due to the different trends in Under 18s and older people. Therefore, we can apply these proportions with some accuracy to achieve good models for how many of each demographic we expect to see attending A&E.

Specifically, over all HBTs, approximately 20% of A&E attendances attend a Minor Injury Unit (MIU) if there are MIUs in that HBT (see Section 2). Therefore, we can predict that the proportion of people attending a MIU is given by

$$n_H^{\text{MIU}}(t) = p_H^{\text{MIU}} a_H + p_H^{\text{MIU}} b_H \sin(c_H t + d_H)$$

where p_H^{MIU} is the proportion of A&E attendances that are at a Minor Injuries Unit. By definition, $p_H^{\text{MIU}} + p_H^{\text{ED}} = 1$, because people must attend either a MIU or Emergency Department. Similarly, we can calculate p_H^m and p_H^f (the male/female split). However, due to this being one of the demographics that is not recorded if a patient is not registered at a GP in Scotland, $p_H^m + p_H^f \leq 1$.

Due to there being biannual periodicity in the number of Under 18s attending A&E, it is not possible to use this model to pick up this pattern with just one sine function. This model is generalised in Section 3.2 and can incorporate these trends.

3.1.3. *Weekly Model Evaluation.*

The monthly model can also be converted into a weekly model to predict the number of people attending A&E in a given week. Finding a finer prediction is more useful for individual hospitals as there is a large number of people flowing through hospitals. The equation is now

given by:

$$y_h(t') = P_h a_h + P_h b_h \sin\left(\frac{52}{12}(c_h t' + d_h)\right) \tag{5}$$

where $y_h(t')$ is the number of people attending A&E in HBT h in week t'.

When attempting to fit the model directly to the weekly data, we see shorter periods for the sine wave. This is because it is picking up more trends in the data. For this toy model, we want to look only at the broadest trends. To prevent overfitting of the model in this way, we average the number of attendances over a month and calculate the coefficients from that data.

For the specific case of prediction 2023, this provides a good estimate with MSE_{Scot} = 1,333,000. This means that the model prediction is within 1,154 people per week in Scotland to the true number of attendances for 68% of the points. Given around 25,000 people attend A&E in Scotland every week, this is an error of approximately 5%.

Training Year	RMSE_{av}	MSE_{Scot}
2018	161.49	2,607,000
2019	217.68	7,428,000
2022	111.87	1,705,000
2023	96.73	1,333,000

TABLE 3. Using the model fitted for previous years (and fitted directly to 2023 itself) to predict the attendances in 2023, the most accurate prediction is from 2022.

The error is proportional to the size of the population, so a smaller HBT has a similar percentage error to a large HBT. This means that this model is suitably accurate for all sizes of HBT.

However for large populations, such as Greater Glasgow and Clyde, Lothian, and Grampian, there are trends that this model does not pick up on well. As shown in Figure 6, there is generally a large drop in the number of A&E attendances in the first few weeks, and also in the summer. To incorporate these biannual drops, we could include another sine function with a half-year period. When going into more detail, there are likely to be other periodic behaviours. Attempting to include all of these would increase the number of coefficients drastically and make the model more complicated, but could improve the accuracy significantly.

3.2. **Fourier Series Model.**

The success of fitting a single sinusoid above suggests the possibility of using a model with several periodic components, in the hope that the slightly-increased complexity will come with better predictive power.

To achieve this, we used the fact that our data comprise a series sampled at uniform time intervals, either weekly or monthly, and this is very convenient for the application of the discrete Fourier transform (DFT) and its algorithmic implementation, the fast Fourier transform (FFT) [18, 19].

The convention we use for the DFT is the following. If y_j, $j \in \{0, \ldots, N-1\}$ represent observations of a process at N uniformly-spaced times $t_j = j\Delta t$, then y_j can be represented by a Fourier series with spectral content in the interval $[\frac{2\pi}{N\Delta t}, \frac{\pi}{\Delta t}]$,

$$y_j = \frac{1}{N}\sum_{n=0}^{N-1} a_n e^{i2\pi nj/N} \tag{6}$$

$$a_n = \sum_{n=0}^{N-1} y_j e^{-i2\pi nj/N}. \tag{7}$$

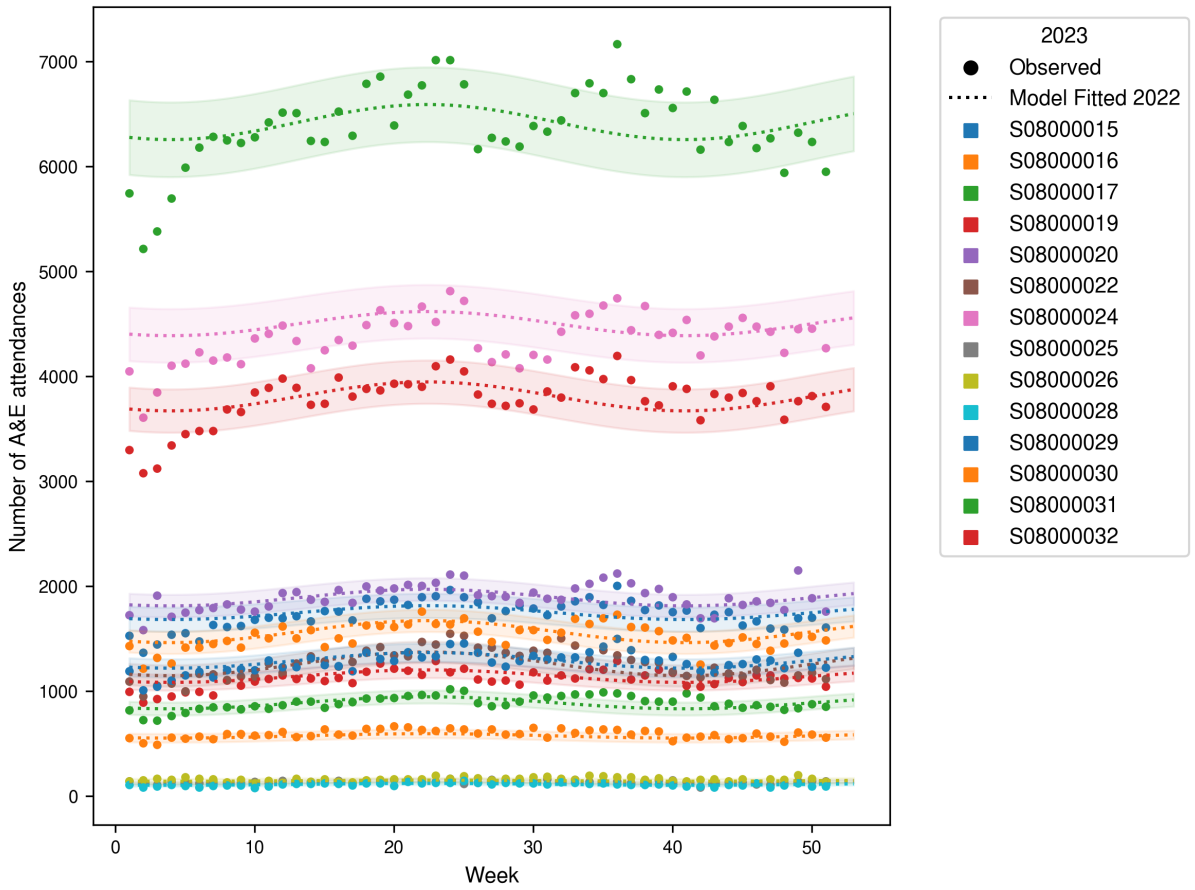


FIGURE 6. The dotted line is the model prediction of the number of people attending A&E for each HBT, trained on data from 2022. Shaded areas represent the RMSE from this model and the observed data (dots).

Using the fact that $a_{N-n} = a_n^*$ for the DFT of a real series, and insisting N be even, we can write the above in a form that extends it into continuous time

$$y(t) = \frac{2}{N} \sum_{n=0}^{N/2} |a_n| \cos\left(\frac{2\pi n}{N\Delta t} t + \arg a_n\right). \tag{8}$$

Applying the DFT to our data yields the relative importance of each (discrete) frequency component n in the form of the amplitude a_n. Because the DFT is invertible, using all of the frequency information from an earlier time period to forecast future data equates to assuming the observations in that period will repeat themselves exactly, which is certainly over-fitting.

To build a reduced model we can retain only a handful of the most significant frequency modes. For example, if we desire only M frequency components we retain the M largest a_n (in the sense of modulus) and set the rest to zero,

$$y(t) = \frac{2}{N} \sum_{m=1}^{M} |a_{n_m}| \cos(\omega_m t + \phi_m) \tag{9}$$

with $\omega_m = 2\pi n_m / N\Delta t$ and $\phi_m = \arg a_{n_m}$ for $m \in \{1, \ldots, M\}$ indexing the M most significant modes.

However, such a truncated cosine series is still entirely deterministic and assumes past observations will repeat themselves. To provide uncertainty quantification and to allow for, in principle, non-periodic terms and interactions with other covariates such as population demographics, we embed the above series into the linear predictors for a *generalised linear model*

(GLM) [20]. In particular, the counting nature of A&E arrivals suggests the use of a Poisson GLM [21].

In full then, our second model takes the following form

$$\begin{aligned} P(Y(t) = y) = p(y,t) &= \frac{e^{-\lambda(t)}\lambda(t)^y}{y!} \\ \text{with} \quad \ln\lambda(t) &= \ln\lambda_0 + \sum_{m=1}^{M} \beta_m \cos(\omega_m t + \phi_m) \\ &:= \ln\lambda_0 + \vec{\beta}\cdot\vec{F}(t) \end{aligned} \tag{10}$$

It is possible to place a confidence interval (CI) on the predictions of this model at each time t simply by using the quantile function for the Poisson distribution, available in standard software packages.

3.2.1. *Training Procedure.*

To fit the above model to a training dataset (y_i, t_i), we use a two-stage procedure that consists of first determining the dominant frequencies ω_m and phases ϕ_m of the sinusoidal components as outlined above, and then maximum-likelihood estimation (MLE) to fit the parameter vector β with the frequencies and phases fixed. Such a two-stage procedure has the advantage that we can use the FFT and its excellent $O(N \log_2 N)$ scaling [18, 19] where it naturally applies, but also, including the frequencies and phases within the MLE parameters would take the model outside of the scope of GLMs. In fact, attempting to fit M apparently identical cosine terms with MLE would lead to an under-determined problem.

In particular, the negative log-likelihood for eq. (10) is, excluding terms without $\vec{\beta}$,

$$l(\vec{\beta}) = -\sum_{i=0}^{N-1} \left(y_i \vec{\beta}\cdot\vec{F}(t_i) - \lambda_0 e^{\vec{\beta}\cdot\vec{F}(t_i)} \right) \tag{11}$$

which can be maximised straightforwardly using Newton's method. In passing, we note that the Hessian for the optimisation, the Fisher information matrix, is

$$H(\vec{\beta}) = \sum_{i=0}^{N-1} \lambda_0 \vec{F}(t_i) \otimes \vec{F}(t_i) e^{\vec{\beta}\cdot\vec{F}(t_i)}. \tag{12}$$

At the end of the MLE, we have the parameter vector that maximises eq. (11), $\hat{\vec{\beta}}$, and so the mean for the trained model, $\hat{\lambda}(t) = \lambda_0 e^{\hat{\vec{\beta}}\cdot\vec{F}(t)}$

3.2.2. *Results.*

The hyper-parameter M in eq. (10) determines the balance between the model failing to capture the structure of the data, and over-fitting to noise or spurious features. For an input of 52 data points representing a year of weekly data, from eq. (8) M can be as large as 26. To investigate which values of the hyper-parameter yield good results, we trained the model on data from 2016, and looked at the squared sum of residuals for prediction of 2017 data, following eq. (1).

M	3	4	5	6	10
MSE_{Scot}	1,011,000	993,200	992,400	1,012,000	1,095,000

TABLE 4. Accuracy of the Fourier series model with an increasing number of Fourier modes M included, quantified using a prediction for all of Scotland for 2017, trained on 2016 data.

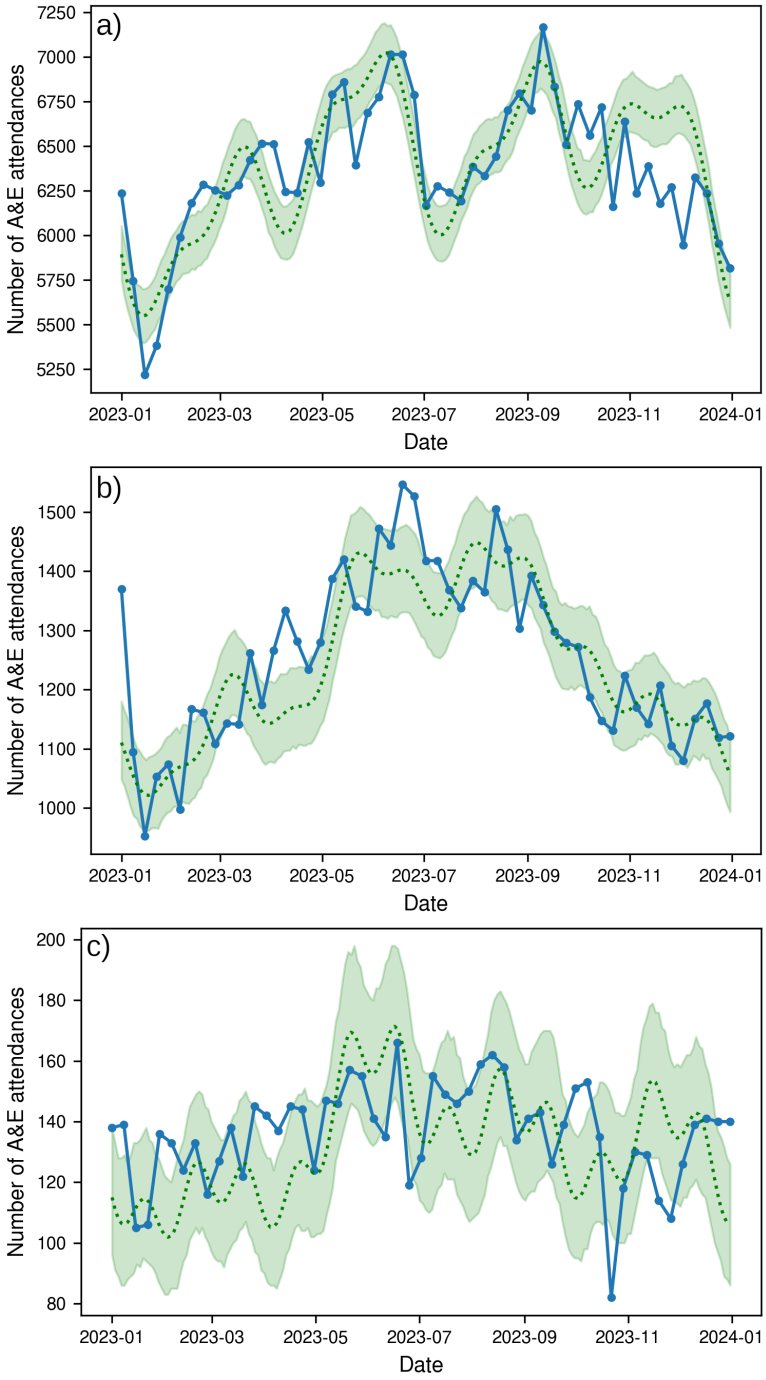


FIGURE 7. Predictions of the Fourier series model for a large, medium, and small HBT in the form of a) Greater Glasgow and Clyde, b) Highland, and c) Orkney respectively. All three cases represent a prediction for 2023 starting from data for 2022, with four modes included in the model. The blue dots indicate ground truth values for 2023, and are joined to aid the eye; the green dashed line is the model prediction, and the green shaded region is a 95% confidence interval.

As can be seen in Table 4, the variability in accuracy is quite small for reasonable choices of the number of modes, with only a 10% difference between the best and worst accuracies shown. As will be discussed later in this section, this lack of variability across Scotland masks starker effects for each HBT.

Unless otherwise stated, in the rest of this report, we use $M = 4$ modes, aiming to balance accuracy and over-parametrisation.

Figure 7 illustrates the performance of the model applied to three HBTs of decreasing size, namely Greater Glasgow and Clyde, Highland, and Orkney. Subjectively, the results perhaps appear best for the medium-size HBT. The four Fourier modes included in this modelling appear appropriate for Glasgow and Highland, but, in the case of Orkney, which has far smaller absolute numbers of A&E attendees, the Fourier model appears to over-fit the noise in the data. Putting a number to this, the MSE for the prediction shown in Figure 7 panel (c) is 446 with $M = 4$ modes, and 308 with $M = 2$, but these both compare poorly with the MSE of 253 achieved by the Toy model outlined above, with a single sinusoidal component.

The other factor in the prediction for this stochastic model is the CI. Because the variance for a Poisson distribution is also its mean, here $\lambda(t)$, the width of any CI will scale as $\sqrt{\lambda(t)}$, slower than the magnitude of the prediction itself. This behaviour is evident in Figure 7, where the CI looks subjectively too large for Orkney, and too small for Glasgow.

Another drawback of the CI, which is clearest looking at panel (a), is that it is narrowest where the prediction changes fastest, and widens when the rate-of-change of $\lambda(t)$ slows. In some sense this is the opposite of what one would desire for making practical recommendations, and so the exploration of other methods for CI evaluation that better account for temporal behaviour would be worthwhile.

As an example of the physical insight that can be provided by this model, despite its flexibility, Table 5 collects the periods of the four modes fit by the model to 2022 Glasgow data in order to produce Figure 7a and their associated weights. It is clear that the most significant periods in the data in this period are yearly variation and the 2nd, 4th, and 8th harmonics thereof. The relative explanatory power of these four modes is fairly uniform as indicated by the fact that their weights β_m are all the same order of magnitude. There are of course very many factors that determine A&E attendances, some of them discussed in Section 2, but it is not surprising to see that they can in part be summarised by trends that recur every year, every quarter, etc.

Mode, m	1	2	3	4
Period (weeks)	6.5	13.0	26.0	52.0
Weight, β_m	0.024	0.052	0.042	0.043

TABLE 5. Periodic components of A&E attendance data fit by the Fourier series model to weekly data from Greater Glasgow and Clyde in 2022, in the form of the periods of the four most significant modes found by discrete Fourier transform, and their associated weights for the Poisson model linear predictor.

3.2.3. *Model Evaluation.*

Before outlining our third model, here we shall discuss some of the shortcomings of the Fourier series model eq. (10), and how it might be extended.

Firstly, as presented here, the constant offset λ_0 of the prediction, or 'DC-mode', is simply taken to be the mean of the training data. This is fine if we expect the total annual A&E attendances to be unchanging year-on-year, but predictions may be more accurate if the model has the ability to incorporate year-on-year trends in the data, as might be expected due to population growth, for example. Initial experiments by ourselves incorporating a term of the form $\beta_0(t - t_0)$ to eq. (10) led to spuriously-large linear trends and poor results. It may be that the only flaw with this approach is training on a year or two of data, and that a longer input series would produce a more accurate trend, but this was unfeasible with the dataset used in this work, while excluding the COVID period. Alternatively, a more sophisticated approach to incorporating year-on-year trends maybe needed, such as predicting the total annual A&E admissions with a different, coarser model, then feeding this value in as λ_0.

As another possible extension, here the CI for the Poisson model prediction was taken as the simplest choice: the inverse cumulative distribution (or quantile function) is applied at each

time t. As mentioned in Section 3.2.2, the resulting CIs are somewhat unsatisfactory, being narrowest where the prediction changes fastest. Confidence intervals for Poisson process models and generalisations thereof, or *prediction intervals* to use the correct term, are an active area of research [22], and so there is scope to improve the intervals presented here. In particular, a better CI ought to account for the temporal behaviour of the model prediction.

A further possibility is that a Poisson distribution is inappropriate for the describing weekly A&E attendances, because the mean and variance of the true random variable are not equal at each time - known as *under-* or *over-dispersion* if the variance is smaller or larger than the mean, respectively. This is not trivial to test with the data arranged as a time series for each HBT as assumed in training the Fourier series model, but more sophisticated over-dispersion tests do exist for Poison GLMs [21]. It would certainly be worthwhile to apply these, and if eq. (10) is found to be inadequate, generalisations of the Poisson distribution should be explored [21, 22].

Finally, the Fourier series model is subject to many of the quirks associated with using the DFT [18]. As an obvious example, for data recorded over an interval of length T, the lowest frequency analysable by Fourier transform is $1/T$, so nothing can be said about a yearly pattern by looking at 11 months of data. Less intuitively, the procedure outlined in Section 3.2.1 is subject to the phenomenon of *spectral leakage* [18], which may lead to mis-identifying the most significant periodic components of an input dataset. This can be mitigated by using a window function, but probably the simplest solution is to exclusively use training sets that span a whole number of years (not necessarily aligned with calendar years), as then the spectral content of modes with frequency $f = n \cdot \text{years}^{-1}$ for $n \in \mathbb{N}$ is faithfully reproduced.

3.3. **Machine Learning.**

Many data-driven modelling techniques explicitly (or implicitly) assume a certain structure of the data itself. For example, in our first two models we assume there is some periodicity at multiple scales in the attendance rates. These assumptions may be relevant and/or useful, but they still mean imposing a certain structure upon the data which may not be entirely reliable. Models imposing these structures are called *parametric* [23]. This motivates our next approach - the construction of a Neural Network (NN).

While NNs do have parameters in the usual sense, there are so many of them that the model can be considered non-parametric ([23]). This means that NNs perform extremely well when you have lots of data and little prior knowledge [23, 24]. Machine learning has an extremely wide catalogue of applications, such as medical imaging [25], language processing [26] and weather prediction [27]. NNs take many general forms and are widely useful in their capability to pick up many non-trivial trends in data (which may not be easily observed through data exploration) [24]. Especially when there are many influencing factors across various data sets, detecting the important factors or key trends can be extremely challenging. In our case, there are many factors influencing A&E attendance which may have complicated interactions (seasonality, population age distributions, deprivation levels, population growth etc.). These interactions are intrinsic in the data, so a NN is able to effectively pick up the *influence* of these interactions on A&E attendance.

The key drawback of a NN is that it is a black-box model. In particular, we do not know which factors/interactions are contributing to our outputs - we simply are given an output for each input. The model is able to pick up the influence of interactions, but not the interactions themselves. Also, the parameter space is generally too large to be interpreted in a physically meaningful way. This makes uncertainty quantification (as used in the construction of confidence intervals) impossible. In medical related fields, a measure of uncertainty is absolutely vital in application. This issue is addressed by adding some probabilistic framework to the model. This is called a Bayesian Neural Network (BNN), which will output probability distributions rather than fixed

values. However BNNs struggle with scalability [28], and have significantly more complicated structures than standard NNs.

In this section we discuss our implementation of a NN on the interpolated weekly attendance data (containing demographic breakdowns). We are able to train a NN which outputs expected attendance numbers, split by age, for each week and each HBT. This is implemented in python, using the PyTorch package in Python [29]. We describe the training process in some generality and then evaluate the model. For the interested reader, we briefly discuss the ideas behind BNNs alongside their limitations in Appendix A.

3.3.1. *Deep Neural Network Structure.*

We consider a class of NN characterised by having multiple hidden layers, called a Deep Neural Network (DNN). In particular, we focus on fitting a function to the data which is the composition of affine linear maps and some generic non-linear maps. More formally, a DNN with L layers is a parametric mapping of the form

$$f_\theta(x) = \sigma_L \circ A_L \circ \sigma_{L-1} \circ A_{L-1} \circ \cdots \circ \sigma_1 \circ A_1(x)$$

where for $i = 1, \ldots, L$, each $A_i \colon \mathbb{R}^{n_i} \to \mathbb{R}^{n_{i+1}}$ is an affine linear mapping, each $\sigma_i \colon \mathbb{R}^{n_i} \to \mathbb{R}^{n_i}$ is some non-linear function called an *activation function* and θ is some set of parameters [30]. Here n_i represents the width of the i-th layer of the DNN. The activation functions are usually application specific and are fully chosen by the user, while the affine linear maps are determined by the parameters θ. In particular, the elements of the vectors and matrices making up the affine transformation are determined by θ. We are interested in determining an optimal choice of θ to minimise the error between the true values in the data and those predicted by f_θ. Such a function is called a Fully-connected Feed-forward Network or a Multi-layer Perceptron (MLP). Other classes of deep neural network exist and are commonly used, such as Convolutional Neural Networks [30], but a MLP is a simple class of model which is powerful in regression applications. Many MLP structures are known to satisfy the famous Universal Approximation Theorem, which tells us (loosely speaking) that some DNN can, under appropriate conditions, approximate any function to arbitrary precision [31]. Hence, if there is some functional relation describing our data we can approximate it by using a suitable DNN.

The activation functions used in MLPs are quite significant in determining the behaviour of the network. In particular, the shape of the activation function being used can be observed in the shape of the final network [32]. This can be attributed to all other contributions to the network being affine-linear, so the non-linearities are 'inherited' from the activations. This property (sometimes called inductive bias) is apparent when all activation functions are taken to be the same, as in [32]. Three common types of activation functions are ReLU, Sigmoid and hyperbolic tangent functions. These are given by $\mathrm{ReLU}(x) = \mathrm{Max}\{0, x\}$, $\sigma(x) = (1 + e^{-x})^{-1}$ and $\tanh(x) = (e^x - e^{-x})/(e^x + e^{-x})$ respectively, and are useful in a wide range of applications. In the multidimensional case, these functions are applied entry-wise to vector elements.

The process of finding a good choice of θ is also influenced by the choice of activation functions. In particular, some choices of σ_i's which are bounded can result in the training process completely failing, depending upon the type of data used for training. This is known as the exploding/vanishing gradients problem [33]. To avoid this we initially chose $\sigma_i = \mathrm{ReLU}$ for each i. In our final implementation however, we use the 'leaky ReLU' function, given by

$$\sigma(x) = \mathrm{LeakyReLU}(x) = \begin{cases} x, \text{ if } x > 0 \\ 0.01x, \text{ if } x \leq 0 \end{cases}$$

which just penalises non-positive values instead of setting them to be exactly zero. This improves the training time [34] and reduces the risk of the exploding/vanishing gradients problem [35] (which was observed during the original implementation). This also has the consequence that

our model has piece-wise linear properties, as is observed in our subsequent figures. This isn't particularly worrying or restrictive since we are dealing with discrete data.

We choose to have three hidden layers in our implementation, which represented a strong balance between expressibility and training tractability while also not producing significant over-fitting. Each of these layers was chosen to have a width of 32. Hence our particular model is of the form

$$f_\theta(x) = \sigma \circ A_3 \circ \sigma \circ A_2 \circ \sigma \circ A_1(x)$$

where $A_1 : \mathbb{R}^{16} \to \mathbb{R}^{32}$, $A_2 : \mathbb{R}^{32} \to \mathbb{R}^{32}$ and $A_3 : \mathbb{R}^{32} \to \mathbb{R}^7$ are the affine linear maps determined by θ. Here x represents an encoding of year, week and HBT. The year is inputted as an integer, the week is inputted as an integer between 1 and 366 (representing the day of the year on which the week ends) and the 14 HBTs are one-hot encoded. Hence $x \in \mathbb{R}^{16}$, with 13 of its entries guaranteed to be zero (due to the one-hot encoding). The data used is the interpolated weekly data which contains the demographic breakdowns by week. Then $f_\theta(x) \in \mathbb{R}^7$ represents the expected A&E attendances split by age group for the given HBT. We have 6 recorded age groups and an additional group representing data containing no age breakdown (representing tourist attendances, for example). We have 1798 parameters contained in θ, which provides us with a good range of model flexibility. Such a size of parameter space is actually very small in machine learning terms, with many modern models having millions to billions of parameters [36].

3.3.2. *Training Procedure.*

The 'training' of the model refers to the process of using our available data to find a good choice of parameters θ. This involves a high dimensional optimisation procedure, where we try to minimise some loss/error function which depends on θ. This function is constructed using some subset of the data, known as the *training data*. The other part of the data is known as the *test data*, which is used for the evaluation of the model performance on 'unseen' data. In general, this data split is done randomly to avoid biases and overfitting in the model. However, our data comes with a temporal structure which we can make use of. We take the first 80% of the data to be the training data and the remaining 20% as the test data. This means the test data can be considered as unseen future data, which is particularly useful when our end goal is future prediction. This method of splitting also ensures that all HBT's are equally represented in the data. When only using post-COVID weekly data, we have 1467 training points and 367 testing points. While this small amount of data is fairly limiting in terms of model potential, it has the implication that the model presented can only improve with time as more usable data becomes available. A model was also trained using COVID data, but the disruption to the data during this period completely threw off the predictions of the model for all times. The model attempted to understand the COVID attendance, which created periodic behaviours in the predictions that are non-existent in reality.

It is important to note that, as a consequence of only considering post-COVID data, we have more parameters than training data points. This turns out to not be as problematic as one would expect. In practice, many DNNs are overparametrised [37, 38] but this tends to add more flexibility to the model, allowing it to capture more sophisticated trends. One concern here is the opportunity for the model to overfit. Thankfully, networks trained using a stochastic optimisation algorithm (such as ours) do not exhibit overfitting even when extremely overparametrised [38]. Even in the general case overfitting does not pose a significant problem, which is a mysterious and useful trait of NNs [37].

Using the training data (of size $N = 1467$) we can construct a loss function. As is typical in regression tasks, our loss function is the mean squared error (MSE), defined as

$$L(\theta) = \frac{1}{N} \sum_{i=1}^{N} (y_i - f_\theta(x_i))^2 ,$$

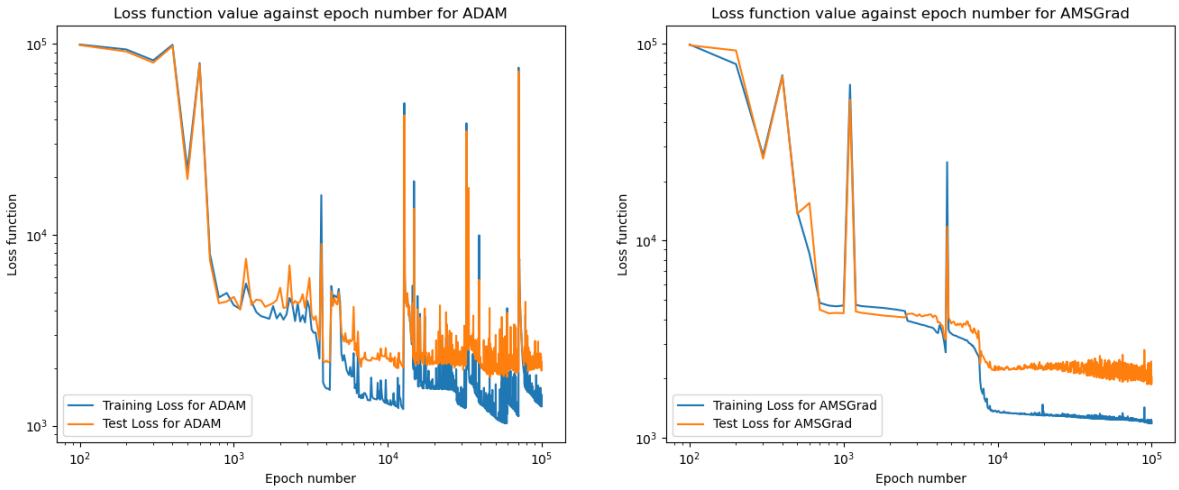


FIGURE 8. Loss function convergence for both ADAM (left) and AMSGrad (right) optimisation methods. Both axes are in logarithmic scale.

where (x_i, y_i) are our true inputs and outputs from the training set and $f_\theta(x_i)$ is the model prediction with parameters θ for input x_i. Since our model is multi-output, our training loss value is computed as the average of $L(\theta)$ for each age group. We do not represent this in the notation for the sake of simplicity. The optimal choice (with respect to L) of parameters is then given by

$$\theta^* = \operatorname{argmin}_{\theta \in \mathbb{R}^{1798}} \{L(\theta)\} .$$

Finding θ^* is extremely challenging in practice due to the high dimensionality of the problem. However, stochastic optimisation methods are effective in finding good estimates of θ^* [39–41]. Such methods are applied to the loss function (which itself is deterministic) by only considering the loss applied only to a subset of the training data, known as a *minibatch*. The minibatch is chosen at random, which is the cause of stochasticity. This reduces the computational cost of each optimisation step dramatically and proves to be effective in practice, especially in the high dimensional setting [41]. This also means, however, that each time the model is trained you will get different parameters. These (and other optimisation) algorithms are typically iterated over the training data many times (with each run-through of the data being called an 'epoch') until convergence is reached.

In our model implementation, we initially made use of a stochastic optimisation method known as ADAM (the name is not an exact acronym, but is derived from 'adaptive moment estimation') as introduced in [41]. While this yielded promising results, the loss function failed to reach a sufficiently steady state of convergence. The level of stochasticity was too significant, such that the value of the loss function changed too much between different training runs even with significant parameter tuning (meaning a good result relies primarily on luck). A variation of this method, known as AMSGrad [40, 41] (which was designed to directly address this issue), provided acceptable results. A comparison of the different training processes is provided in Figure 8. Large jumps are still observed in the training process when using the more stable AMSGrad method, but these are significantly less frequent and do not appear in the last 90% of the training times. For the first 10^4 epochs the value of the loss function for both the training and test data decreases significantly (besides some occasional explosions, which are quickly recovered from). The rest of the training process is significantly more stable, with test loss remaining stagnant and the training loss decreasing only marginally. Note here that the initial parameters for the optimisation procedure are chosen according to a uniform distribution, which in our case is $\mathcal{U}(-1/4, 1/4)$ (as chosen by the PyTorch package when dealing with linear layers with our given number of features) [29].

3.3.3. *Model Evaluation.*

Generally, the trained model performs well in larger populations and for more predictable age groups. In particular, age groups representing individuals who are over 18 are generally very well predicted by the model. Whereas the under 18 age group makes up for most of the error in the model, with the unknown age category coming second place in terms of error. For smaller populations this effect is exacerbated, where the model predicts completely unreasonably.

We briefly discuss the models performance in Glasgow, which has been notoriously difficult to model previously. Figure 9 shows said predictions over all age groups.

We manage to model the general trends in the data fairly well, predicting the dips in the data accurately. There does not appear to be any apparent overfitting. When we break this down into age groups we have, as discussed, differing levels of accuracy. Figure 10 shows the strong predictive power of the model for predicting A&E attendances for ages 25-39. The trend is obvious by inspection of the plot, but the model manages to find this and capture it well, with minimal overfitting. In contrast, the model performs very poorly for under 18's, as highlighted in Figure 11. However, in this case there is not any apparent trend at all in the data, and we can see that the model is struggling here. The significant attendance drops around the new year period are detected well, alongside the drop during the summer period (which can be attributed to school holidays). The performance of the model on the test data is also very poor, as shown in the region after the red line of the figure. This can be attributed to the combination of lack of data and the unpredictability of this age group. Glasgow contains the largest children's hospital in Scotland, which has the consequence that many attendances will be from surrounding HBTs for those requiring more specialist urgent treatments. Such cases are inherently more unpredictable, which adds additional difficulty to the task. As more data becomes available, the model should succeed in capturing the details more finely.

Furthermore, the model performs poorly in the low population setting. The model is trying to minimise the error in terms of number of individuals. Hence, finding a good fit for Glasgow (being off by order 10^1 to order 10^2) does not mean finding a good fit for areas with lower population such as Orkney. This is shown in Figure 12. Here the model predicts values which

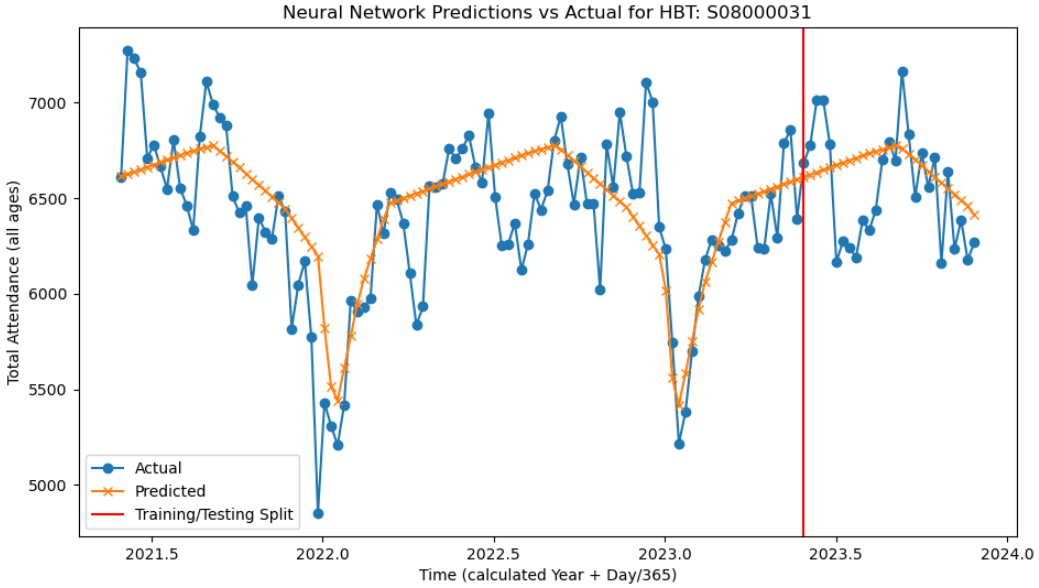


FIGURE 9. NN predictions for all age groups in Glasgow. Red line indicates the point at which the model is being tested upon unseen data. The predictions are found by summing each entry of the output $f_\theta(x) \in \mathbb{R}^7$.

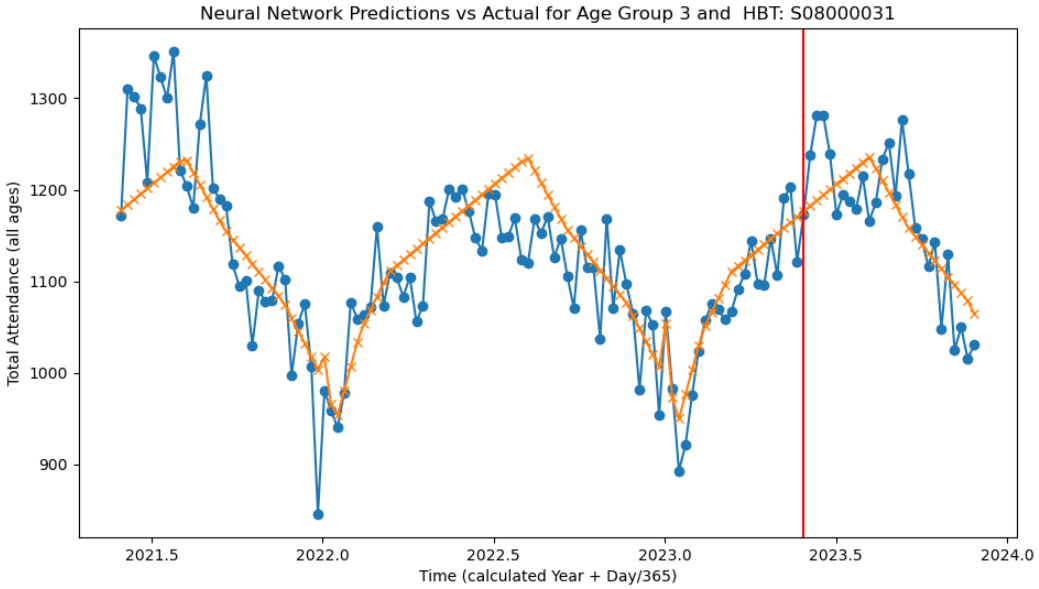


FIGURE 10. Glasgow model predictions for ages 25-39. Red line indicates the point at which the model is being tested upon unseen data.

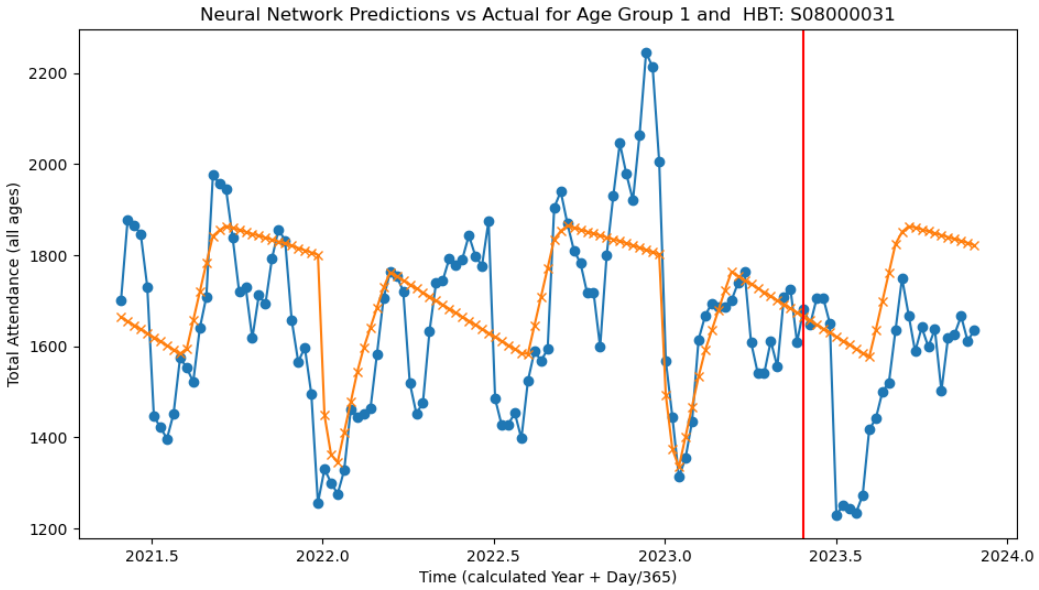


FIGURE 11. Glasgow model predictions for under 18s. Red line indicates the point at which the model is being tested upon unseen data.

are consistently far too high. We suppose that upon initialisation the model predicts at the global average, and then shifts up or down for each HBT, until the mean squared error for each HBT is sufficiently low. The final level of error is too high for the model to prioritise improving the fit on the areas with low population size. In general we expect that Glasgow, which has the highest population count, will be best predicted by the model for exactly this reason. This is of course just an educated guess. This related back to the key flaw of NNs; when we know the model is wrong we rarely know exactly why.

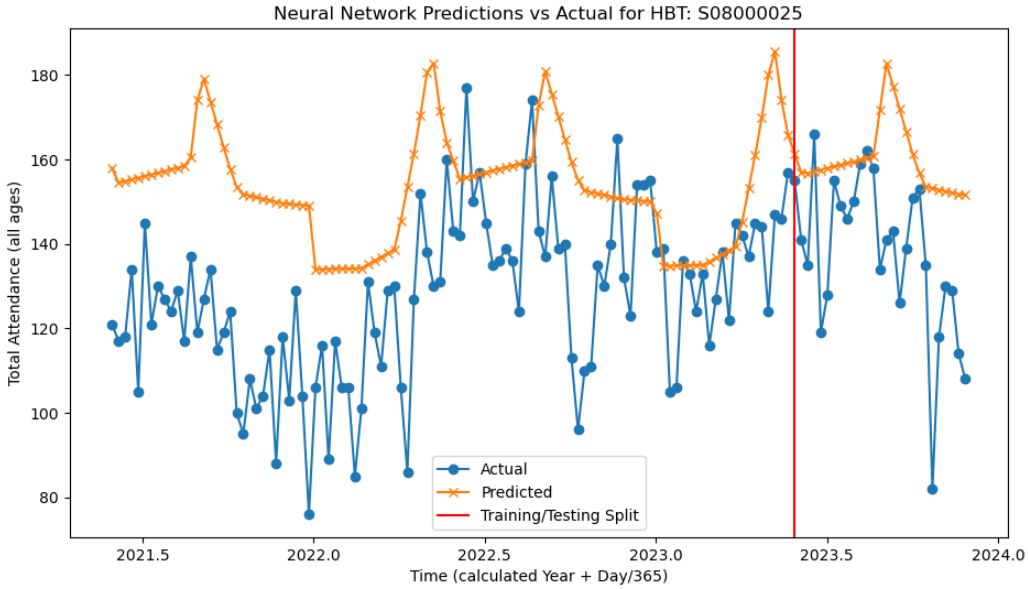


FIGURE 12. Orkney model predictions for all age groups. Red line indicates the point at which the model is being tested upon unseen data.

Age Group	Under 18	18-24	25-39	40-64	65-74	75 Plus	None
MSE	7540	359.8	492.8	720.7	147.2	406.9	1268

TABLE 6. MSE by age group for the neural network model across the entire test data set. Note that the test set contained information from every HBT.

The model statistics are also fairly strong in comparison to our previous models, as one would expect with the significantly higher degree of parametrisation. Additionally, we can obtain statistics by age group due to the models ability to predict for each age group. The statistics also align with our previous analysis that the Under 18 category is the most challenging to predict. The breakdown of the MSE for the test data is given in Table 6. This shows that the extreme majority of our uncertainty comes from the under 18 groups, contributing immensely to the average. The final mean squared error is then the sum of these quantities, given by MSE_l=10934.900. This corresponds to 7 times the final point of the right plot of Figure 8 for the test data. In particular, the loss function in our multi-output regression setting is given by the average MSE between each age group, so our final test MSE is simply the multiplication of this quantity with the number of age groups. This statistic is not directly comparable to the MSE from previous models, due to our model outputting the age breakdown of attendances. This is also due to the different size and structure of our test data, in comparison to our previous models. Thus, to compare with our other models we need to compute predictions for the total population (obtained by summing each entry of the vector output) and then compute the MSE directly with the true expected values and our computed values. Such value of the MSE is then directly comparable with the MSE_t statistic introduced at the beginning of this section. This value is significantly greater than what is computed above. For 2023, the MSE_{Scot} is 339035. The model predicts attendance at the Scotland-wide level well, with the general trend being followed remarkably well, as shown in Figure 13.

3.3.4. *Next Steps for the Neural Network.*

Overall, the NN model has strong predictive power in the high population setting and is able to identify and replicate the more clear trends in the data to high accuracy. The under-18 age group produces the majority of the uncertainty. Given the small amount of available data for

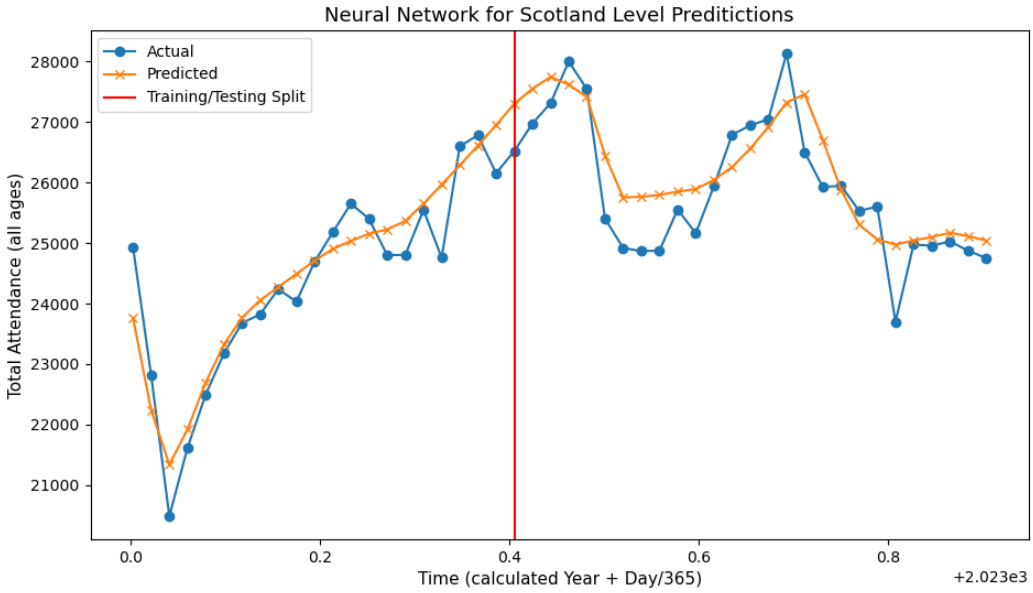


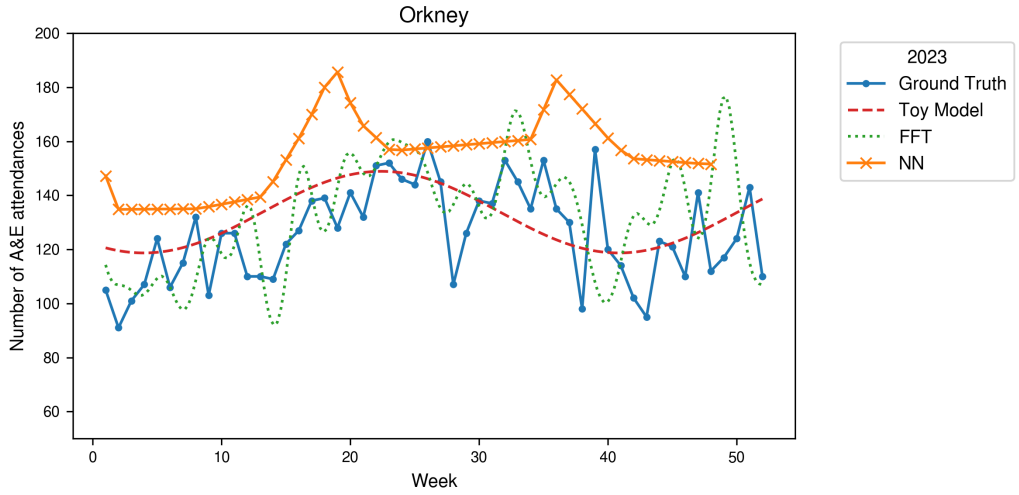
FIGURE 13. Whole of Scotland NN model predictions for all age groups and all HBTs for each week of 2023. Predictions are found by aggregating the age predictions for each HBT and then aggregating each HBT prediction. Red line indicates the point at which the model is being tested upon unseen data.

training, the performance of the model remains impressive on a Scotland wide level. In order to improve performance in the low population setting, the data would need to be normalised so variance is more uniform between HBTs. This ensures that the minimisation of the error is uniform between locations (by effectively minimising the percentage error rather than the absolute error). Data normalisation can also allow more freedom in the choice of activation functions when defining the model architecture. An interesting future work would be making use of more novel activation functions such as the 'snake' function (as in [32]) which are designed to pick up potential periodic behaviours. In some cases normalisation can also reduce risk of exploding/vanishing gradients [42].

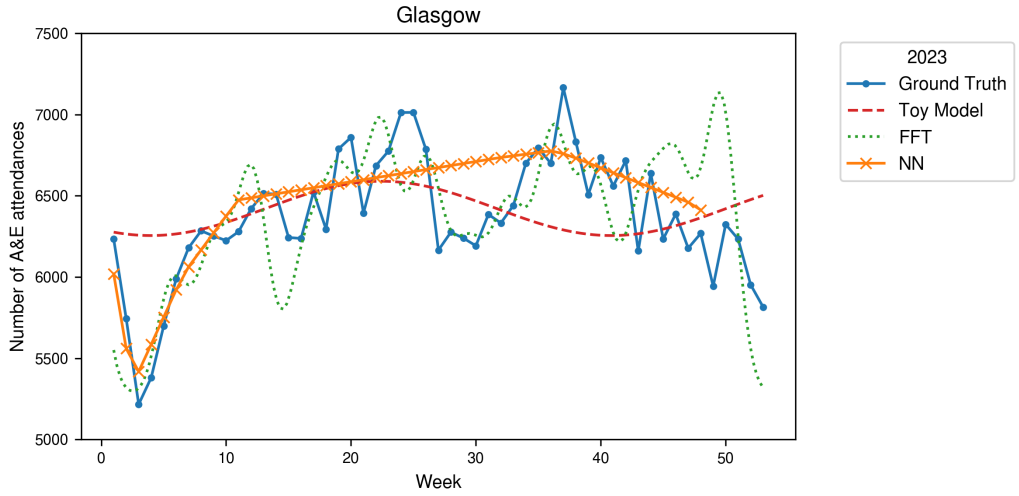
One could also experiment with making use of alternative loss functions. An interesting extension to this work would be to consider an asymmetrical loss function, which prioritises the model predicting a number too high rather than a number too low. Simple asymmetrical generalisations of the standard MSE loss function have been useful in applications and could be highly effective in our context [43].

Deep learning models have an unending appetite for data, so as more usable data becomes available with time the model can generally only improve. With the additional data, one could also add more hidden layers to increase model expressibility, which in turn could allow the model to detect some more complex trends which may appear in the Under 18s category. Additionally, the availability of more data could also allow for easy model generalisation into hospital level predictions.

As mentioned previously, the main downfall of a NN is the lack of explainability. Outputs come without any uncertainty of results, and we have seen in our discussion that they can be off quite significantly in their predictions. Hence, a recommended first step for further development would be to add a Bayesian framework to the model. We briefly outline the relevant theory in Appendix A. While dedicated computer packages for constructing BNNs exist in Python (such as Pyro, which is attached to the PyTorch extension as used in our NN implementation [44]), implementation is outside the scope of this project.



(A) S08000025 - Orkney, 2023



(B) S08000031 - Greater Glasgow and Clyde, 2023

FIGURE 14. The predictions provided by each of the models against the observed data for 2023 are plotted. The NN overestimates for Orkney but provides a good prediction for Greater Glasgow and Clyde that is not overfitted. The Toy model provides a good estimate for Orkney but does not pick up large peaks and troughs in the number of A&E attendances in Greater Glasgow and Clyde. FFT gives a very accurate prediction but appears to be overfitting.

4. COMPARISON

Each model has different benefits and problems. In this section we discuss these and compare the different models. Exploring different models is important because they can give different insights into the data, which may be useful for different clients [15]. To understand how the system is working as a whole, it is useful to be able to predict the number of attendances for all of Scotland. This can give an idea on how the NHS is performing in Scotland compared to England, Wales, and Northern Ireland. Hospital Management, however, may want to know how many people will attend on a given day to help with staffing and bed management.

There are different factors to what makes the best model. Some models may be more accurate, but aren't easily use-able or understandable for a client. Some models may provide less accurate results for weekly attendances but will be able to predict long term trends for the future. We will compare the models on short term accuracy, simplicity, and reliability of future predictions.

4.1. **Accuracy.**

Using MSE_{Scot} as defined previously (Equation 1), we can compare the accuracy of different models on a whole of Scotland scale. Using a standard MSE we can also compare the accuracy for different HBTs. Square rooting this number tells us how far away the prediction is from the real recorded number of A&E attendances.

In Table 7, we can see that Neural Networks are a lot more accurate on the whole of Scotland in predicting *A&E* attendances in 2023 and for very large HBTs. However, for HBTs with less data, such as the smallest HBT, Orkney, they provide a much less accurate result, as shown in Figure 14a.

For the smallest HBT, Orkney, the number of people attending A&E is more than 40 times smaller than the number of people attending in Greater Glasgow and Clyde. Because the population is so small, there is a lot more variance due to noise. There are some number of people that will likely come in at certain periods of the year, and then there are random accidents and emergencies that can't be easily predicted. In large populations, there are so many people that these random accidents don't have a large variance. However, for smaller health boards, these are more significant.

Model	MSE_{Scot}	$\text{MSE}_{\text{GG\&C}}$	$\text{MSE}_{\text{Orkney}}$
Toy	1,705,123	133,602	253
FFT	517,776	68,295	443
NN	339,035	60,113	1109

TABLE 7. NN is 2.2× more accurate at predicting attendances for Greater Glasgow and Clyde than the Toy model, but is 4.4× worse at predicting attendances for Orkney.

The Neural Network is inaccurate for small HBTs because for a much larger population, this amount of error is acceptable, and it does not differentiate between these errors. The FFT also has a high amount of error because, like the toy model, it uses predictions from the previous year. The predictions pick up the noise from previous years and carry them over. The Toy model is the most accurate because it picks up the general trends without the added noise.

For the small HBTs, Orkney, Shetland, and Western Isles, it is best to use the toy model and for the largest HBTs, Greater Glasgow and Clye, Lothian, and Grampian, it is best to use the NN. However, as mentioned in Section 3.3, the NN might provide better results for small HBTs if the data is normalised first or if the data for each HBT is trained individually.

4.2. **Simplicity.**

These models will be used by clients and to understand whether the predictions are reliable, they should be easy to understand. NN are a black box and are difficult to understand. It is not possible to use this model to investigate what could happen if there was a change in the dynamics, such as what occurred during the COVID pandemic. The Toy model and FFT model are a lot easier to understand as each component can be understood as a physical trend, especially with lower dimensions in the FFT.

The amount of data needed to provide an accurate model is also important because when new hospitals are built, there are new trends in the number of attendances and these should be quickly and accurately picked up on. The Toy model needs only one year previously to predict with some level of accuracy. However, Neural Networks work best with many years of data. Due to COVID, we are ignoring 2020 and 2021, so the NN is only trained on one year of data to predict 2023. In a few years, this will increase in accuracy as it will be able to predict yearly changes, such as population increases. With finer data, such as daily data, there will be many more data points so the predictions will improve in accuracy.

4.3. **Future Predictions.**

Predicting the number of people who will attend A&E far in the future is very difficult as there are many factors which we cannot predict. There are also many trends which are very slow, such as a yearly increase in the proportion of people attending A&E and an aging population [8]. It is important to prepare for this as building hospitals and training doctors takes many years.

The Toy model and FFT model become prohibitively inaccurate when predicting many years into the future. The proportion of people attending A&E was increasing pre-COVID and after a drop post-COVID appears to be again. These models do not pick up on this so need to be tuned on previous years' data, where the changes are less significant.

When trained on enough data, the NN model can pick up on the trends happening over many years. This means that, when is has more data to learn from and has appropriate scaling applied for the HBT population, it will produce the most accurate predictions for future years. One of the trends that it picks up on are the fact the population is aging. In other words, the proportion of people in older age categories is increasing in the country [8]. The NN predicts that the number of people in these older categories attending A&E will increase at a faster rate than those in younger categories. The other models need parameters to be adjusted to accurately model this.

4.4. **Combining Models.**

As seen in Table 7, different models perform differently in different areas, depending on the population size. The NN model is very accurate on a large population, whereas the Toy model provides the best prediction on the smallest population.

To provide the best prediction, we can combine these two models linearly by

$$y_h(t) = \alpha_h^N f_h^N(x,t) + \alpha_h^T f_h^T(x,t) \tag{13}$$

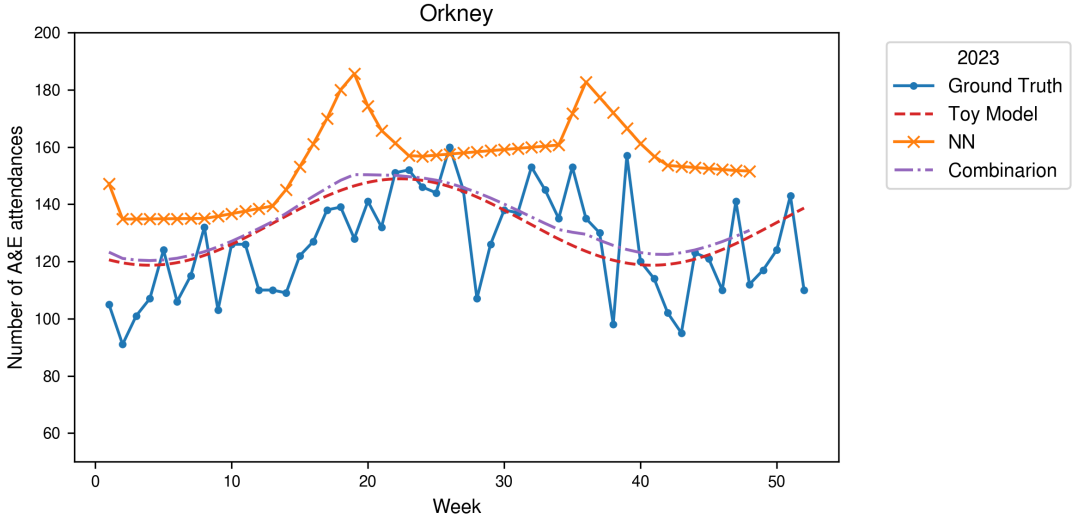
where $f^N(x,t)$ is the prediction from the NN model, $f^T(x,t)$ is the prediction from the Toy model. Here $\alpha^N, \alpha^T \in [0,1]$ are such that $\alpha^N + \alpha^T = 1$. The values are calculated using regression from the previous year. Combining these models enforces the periodicity and simplicity from the Toy model whilst also allowing the model to learn from previous data and change as appropriate [45]. It also further prevents the risk of the NN model overfitting to the data, especially with few data points. It also "chooses" the most accurate model from the previous year and weights it accordingly.

Because these values are fitted to data observed in the previous year, it will not always provide the best MSE because different models might perform differently in different years. However, as shown in Table 8, the combined MSE is very close to the minimum MSE of the other models. This slight loss in accuracy however is still better than having to decide which model is the best.

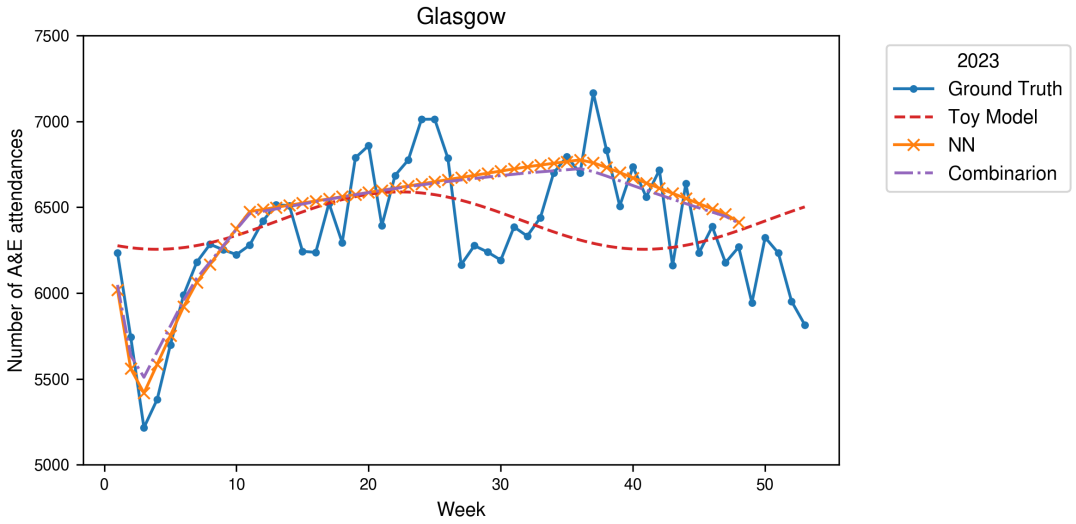
As shown in Figure 15, this combination "chooses" the NN model to predict Glasgow and the Toy model to predict Orkney. To investigate further, it might be possible to find a smooth function against population size to calculate the value α_h^N (and hence α_h^S). This would reduce the expense of the model as it would not need to calculate more parameters. It would also be useful to include the FFT model in this, perhaps with less modes, instead of the Toy model to pick up more smooth periodic patterns.

Model	$\text{MSE}_{\text{GG\&C}}$	$\text{MSE}_{\text{Orkney}}$
Toy	133,602	253
NN	60,113	11,089
Combined	580,821	269

TABLE 8. Combining Models improves the MSE by 3% in Greater Glasgow and Clyde from the NN model, and 97% in Orkney.



(A) S08000025 - Orkney, 2023 with $\alpha^N = 0.1$, $\alpha^S = 0.9$



(B) S08000031 - Greater Glasgow and Clyde,2023 with $\alpha^N = 0.9$, $\alpha^S = 0.1$

FIGURE 15. Combining the Toy model and NN model chooses the most accurate model. For Orkney, MSE = 268 (4× better than just the NN prediction) and for Greater Glasgow and Clyde, MSE = 58082 (3% more accurate than the NN prediction)

5. CONCLUSION

In this report, we have looked at different modelling approaches to predict the number of A&E attendances in Scotland. This will aid PHS in producing a model to show the flow of patients through hospitals and in predicting bed occupancy.

The different approaches we have looked at are a Toy model which only assumes that the number of attendances is 1-year periodic (i.e. more patients visit in the summer than the winter and this trend happens every year); a Fast Fourier Transform model, which picks up periodic behaviours at different time periods; and a Neural Network trained on previous years attendances.

The Toy model tells us that in general, we expect to see more people attending A&E in the summer months. This suggests that certain injuries (an assumption would be that these include heatstroke, hay fever, and broken bones [4]) are more prevalent during those times. Access to data of the reason for A&E attendances could inform how often to restock equipment and medicines required to look after such ailments.

The Neural Network gives the most accurate future predictions and has the power to be even more accurate with more access to past data. It predicts that the number of people attending A&E will increase significantly over the next 10 years. This is very useful as it gives the NHS time to prepare.

Combining these models provides a more accurate model over all HBTs as the NN is not very accurate for small HBTs but provides a very good prediction for large HBTs and similarly, the toy model provides a good estimate for small HBTs. Combining them with an appropriate ratio reduces the error. However, this ratio can only be used to predict one year ahead, like the FFT and Toy models. These models do not pick up on trends that happen on the year-scale so cannot predict far into the future.

To further this research, we would like to be able to predict A&E attendances at individual hospitals on a daily, or even hourly level. This data is not publicly accessible though. Having a much finer prediction would allow PHS to predict wait times using queuing theory [46]. It would also help predict better times for shift changes, as this should occur at the least busy times. However, such fine predictions would also have a much higher error, especially in HBTs with low populations. As shown in Section 3.1.3, it is easy to transform from different time periods with appropriate scaling, but can be harder to fit given an increased number of points and random noise.

Hospital-level predictions also require more data access, as the weekly data published only includes Emergency Departments and not Minor Injury Units and other A&E affiliated departments. Each of these models, when provided with enough training data where necessary, should be able to give predictions for A&E attendances at specific hospitals. However, due to the population size being significantly smaller, we expect the data to be rather noisy with lots of random variations. Because of this, the predictions will be less accurate and a confidence interval should be given, rather than an expected value.

With more access to data, the number of people of different demographics attending A&E could also be accurately predicted, as shown in Section 3.3. Section 3.3 uses interpolated data, but it could easily be trained on real data. By looking at data on typical causes for A&E attendances, it should also be possible to use all the methods to predict why people will attend A&E in a given time period.

References

1. *Public Health Scotland* https://publichealthscotland.scot/our-organisation/about-public-health-scotland/supporting-whole-system-approaches/applying-a-whole-system-approach/ (2024).
2. *Hospital Codes - Scottish Health and Social Care Open Data* https://www.opendata.nhs.scot/dataset/hospital-codes (2024).
3. *Scottish Index of Multiple Deprivation 2020* http://www.gov.scot/collections/scottish-index-of-multiple-deprivation-2020/.
4. Boyce, S. H. Review of sports injuries presenting to an accident and emergency department. en. *Emergency Medicine Journal* **21,** 704–706. ISSN: 1472-0205, 1472-0213. doi:10.1136/emj.2002.002873. https://emj.bmj.com/lookup/doi/10.1136/emj.2002.002873 (2024) (Nov. 2004).
5. *School Term Dates* en. Oct. 2015. https://www.glasgow.gov.uk/index.aspx?articleid=17024 (2024).
6. Kirkwood, G., Hughes, T. C. & Pollock, A. M. Results on sports-related injuries in children from NHS emergency care dataset Oxfordshire pilot: an ecological study. en. *Journal of the Royal Society of Medicine* **112,** 109–118. ISSN: 0141-0768. doi:10.1177/0141076818808430. https://doi.org/10.1177/0141076818808430 (2024) (Mar. 2019).
7. A&E waiting times. en. *Nuffield Trust*. https://www.nuffieldtrust.org.uk/resource/a-e-waiting-times (2024).
8. *Scotland's Census: Population* en. https://www.scotlandscensus.gov.uk/census-results/at-a-glance/population/ (2024).
9. Giebel, C. *et al.* What are the social predictors of accident and emergency attendance in disadvantaged neighbourhoods? Results from a cross-sectional household health survey in the north west of England. *BMJ Open* **9.** ISSN: 2044-6055. doi:10.1136/bmjopen-2018-022820. https://bmjopen.bmj.com/content/9/1/e022820 (2019).
10. Aven, T. On the meaning of a black swan in a risk context. *Safety Science* **57,** 44–51. ISSN: 0925-7535. doi:https://doi.org/10.1016/j.ssci.2013.01.016. https://www.sciencedirect.com/science/article/pii/S0925753513000301 (2013).
11. Kerasidou, A. & Kingori, P. Austerity measures and the transforming role of A&E professionals in a weakening welfare system. *PLoS ONE* **14,** e0212314. ISSN: 1932-6203. doi:10.1371/journal.pone.0212314. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6373963/ (2024) (Feb. 2019).
12. Avishai, B. The Pandemic Isn't a Black Swan but a Portent of a More Fragile Global System. en-US. *The New Yorker*. ISSN: 0028-792X. https://www.newyorker.com/news/daily-comment/the-pandemic-isnt-a-black-swan-but-a-portent-of-a-more-fragile-global-system (2024) (Apr. 2020).
13. Antipova, T. *Coronavirus Pandemic as Black Swan Event* in *Integrated Science in Digital Age 2020* (ed Antipova, T.) (Springer International Publishing, Cham, 2021), 356–366. ISBN: 978-3-030-49264-9.
14. Baker, C., Kirk-Wade, E., Brown, J. & Barber, S. Coronavirus: A history of English lockdown laws. https://commonslibrary.parliament.uk/research-briefings/cbp-9068/ (2024) (Jan. 2024).
15. Tedeschi, L. O. Assessment of the adequacy of mathematical models. *Agricultural Systems* **89,** 225–247. ISSN: 0308-521X. doi:10.1016/j.agsy.2005.11.004. https://www.sciencedirect.com/science/article/pii/S0308521X05002568 (2024) (Sept. 2006).

16. Redish, E. Using Math in Physics: 4. Toy models. *The Physics Teacher* **59,** 683–688. ISSN: 0031-921X. doi:10.1119/5.0025840. eprint: https://pubs.aip.org/aapt/pte/article-pdf/59/9/683/9874510/683_1_online.pdf. https://doi.org/10.1119/5.0025840 (Dec. 2021).
17. Haynes, R. The geographical distribution of mortality by cause in Chile. *Social Science & Medicine* **17,** 355–364. ISSN: 0277-9536. doi:https://doi.org/10.1016/0277-9536(83)90238-1. https://www.sciencedirect.com/science/article/pii/0277953683902381 (1983).
18. Brigham, E. O. *The Fast Fourier Transform: An Introduction to Its Theory and Application* (Prentice Hall, Nov. 1973).
19. Press, W. H., Teukolsky, S. A., Vetterling, W. T. & Flannery, B. P. *Numerical Recipes : The Art of Scientific Computing* ISBN: 0-521-88068-8 (Cambridge University Press, Cambridge ; New York, N.Y., 2007).
20. Dobson, A. J. & Barnett, A. G. *An Introduction to Generalized Linear Models* 4th ed. ISBN: 978-1-315-18278-0. doi:10.1201/9781315182780 (Chapman and Hall/CRC, New York, Apr. 2018).
21. Hilbe, J. M. *Modeling Count Data* ISBN: 978-1-107-02833-3. doi:10.1017/CBO9781139236065. (2024) (Cambridge University Press, Cambridge, 2014).
22. Kim, T., Lieberman, B., Luta, G. & Peña, E. A. Prediction Intervals for Poisson-based Regression Models. *WIREs Computational Statistics* **14,** e1568. ISSN: 1939-0068. doi:10.1002/wics.1568. (2024) (2022).
23. Russell, S. & Norvig, P. *Artificial Intelligence: A Modern Approach* 3rd ed. (Prentice Hall, 2010).
24. Alzubaidi, L. *et al.* Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. *Journal of Big Data* **8.** https://api.semanticscholar.org/CorpusID:232434552 (2021).
25. Mall, P. K. *et al.* A comprehensive review of deep neural networks for medical image processing: Recent developments and future opportunities. *Healthcare Analytics* **4,** 100216. ISSN: 2772-4425. doi:https://doi.org/10.1016/j.health.2023.100216. https://www.sciencedirect.com/science/article/pii/S2772442523000837 (2023).
26. Ma, Q. *Natural language processing with neural networks* in *Language Engineering Conference, 2002. Proceedings* (2002), 45–56. doi:10.1109/LEC.2002.1182290.
27. Han, J. M., Ang, Y. Q., Malkawi, A. & Samuelson, H. W. Using recurrent neural networks for localized weather prediction with combined use of public airport data and on-site measurements. *Building and Environment* **192,** 107601. ISSN: 0360-1323. doi:https://doi.org/10.1016/j.buildenv.2021.107601. https://www.sciencedirect.com/science/article/pii/S0360132321000160 (2021).
28. Dusenberry, M. W. *et al.* Efficient and Scalable Bayesian Neural Nets with Rank-1 Factors. *CoRR* **abs/2005.07186.** arXiv: 2005.07186. https://arxiv.org/abs/2005.07186 (2020).
29. Paszke, A. *et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library* 2019. arXiv: 1912.01703 [cs.LG].
30. Murphy, K. P. *Probabilistic Machine Learning: An introduction* (MIT Press, 2022).
31. Hornik, K., Stinchcombe, M. & White, H. Multilayer feedforward networks are universal approximators. *Neural Networks* **2,** 359–366. ISSN: 0893-6080. doi:https://doi.org/10.1016/0893-6080(89)90020-8. https://www.sciencedirect.com/science/article/pii/0893608089900208 (1989).
32. Ziyin, L., Hartwig, T. & Ueda, M. Neural Networks Fail to Learn Periodic Functions and How to Fix It. arXiv: 2006.08195 [cs.LG] (2020).

33. Paik, I. & Choi, J. *The Disharmony between BN and ReLU Causes Gradient Explosion, but is Offset by the Correlation between Activations* 2023. arXiv: 2304.11692 [cs.LG].
34. Zhang, G. & Li, H. *Effectiveness of Scaled Exponentially-Regularized Linear Units (SERLUs)* 2018. arXiv: 1807.10117 [cs.LG].
35. Mastromichalakis, S. *ALReLU: A different approach on Leaky ReLU activation function to improve Neural Networks Performance* 2021. arXiv: 2012.07564 [cs.LG].
36. Ray, P. P. ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems* **3,** 121–154. ISSN: 2667-3452. doi:https://doi.org/10.1016/j.iotcps.2023.04.003. https://www.sciencedirect.com/science/article/pii/S266734522300024X (2023).
37. Allen-Zhu, Z., Li, Y. & Liang, Y. *Learning and Generalization in Overparameterized Neural Networks, Going Beyond Two Layers* 2020. arXiv: 1811.04918 [cs.LG].
38. Li, Y. & Liang, Y. *Learning Overparameterized Neural Networks via Stochastic Gradient Descent on Structured Data* 2019. arXiv: 1808.01204 [cs.LG].
39. Kingma, D. P. & Ba, J. *Adam: A Method for Stochastic Optimization* 2017. arXiv: 1412.6980 [cs.LG].
40. Tran, P. T. & Phong, L. T. On the Convergence Proof of AMSGrad and a New Version. *IEEE Access* **7,** 61706–61716. ISSN: 2169-3536. doi:10.1109/access.2019.2916341. http://dx.doi.org/10.1109/ACCESS.2019.2916341 (2019).
41. Reddi, S. J., Kale, S. & Kumar, S. *On the Convergence of Adam and Beyond* in *International Conference on Learning Representations* (2018). https://openreview.net/forum?id=ryQu7f-RZ.
42. De, S. & Smith, S. L. *Batch Normalization Biases Residual Blocks Towards the Identity Function in Deep Networks* 2020. arXiv: 2002.10444 [cs.LG].
43. Rengasamy, D., Rothwell, B. & Figueredo, G. P. *Asymmetric Loss Functions for Deep Learning Early Predictions of Remaining Useful Life in Aerospace Gas Turbine Engines* in *2020 International Joint Conference on Neural Networks (IJCNN)* (2020), 1–7. doi:10.1109/IJCNN48605.2020.9207051.
44. Bingham, E. *et al. Pyro: Deep Universal Probabilistic Programming* 2018. arXiv: 1810.09538 [cs.LG].
45. Meade, N. & Islam, T. Technological Forecasting—Model Selection, Model Stability, and Combining Models. *Management Science* **44,** 1115–1130. doi:10.1287/mnsc.44.8.1115. eprint: https://doi.org/10.1287/mnsc.44.8.1115. https://doi.org/10.1287/mnsc.44.8.1115 (1998).
46. Silvester, K., Lendon, R., Bevan, H., Steyn, R. & Walley, P. Reducing waiting times in the NHS: is lack of capacity the problem?: Clinician in Management. *Clinician in Management* **12,** 105–109. ISSN: 09655751. https://search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=15289290&site=ehost-live (2024) (June 2004).
47. Goan, E. & Fookes, C. in *Lecture Notes in Mathematics* 45–87 (Springer International Publishing, 2020). ISBN: 9783030425531. doi:10.1007/978-3-030-42553-1_3. http://dx.doi.org/10.1007/978-3-030-42553-1_3.
48. Dashti, M. & Stuart, A. M. *The Bayesian Approach To Inverse Problems* 2015. arXiv: 1302.6989 [math.PR].
49. Klebanov, I., Sikorski, A., Schütte, C. & Röblitz, S. *Objective Priors in the Empirical Bayes Framework* 2020. arXiv: 1612.00064 [stat.ME].
50. Uhlig, H. On Jeffreys Prior When Using the Exact Likelihood Function. *Econometric Theory* **10,** 633–644. ISSN: 02664666, 14694360. http://www.jstor.org/stable/3532553 (2024) (1994).

Appendix A. Bayesian Neural Networks

Standard NNs are useful as function approximators, due to the aforementioned Universal Approximation Theorem. This leads to high model accuracy, but in many applications (including our own) we require a level of uncertainty quantification. BNNs are a generalisation of the classical NN structure where our parameters are now considered to be random variables. So instead of learning the parameters of the model, we aim to learn the distributions of the parameters conditioned on the training data using statistical techniques [28, 47]. In this appendix we aim to present the key ideas behind BNNs and highlight the key challenges one may face in application.

Let $\mathcal{D} = \{x_i : i = 1, \dots, N\}$ be the input training data set. Suppose we have chosen a structure of a neural network, such as a MLP, given by f_θ. Now θ consists of random variables rather than fixed parameters. The task of finding the distribution of these model parameters can be thought of as a classical inverse problem, in which Bayesian methods are typically employed [48]. More concretely, we aim to find the distribution of our parameters when observing the data $\mathcal{D}$, which we call the *posterior distribution* $\pi(\theta|\mathcal{D})$. This can be found by first considering the joint distribution function

$$p(\theta, \mathcal{D}) = p(\theta)p(\mathcal{D}|\theta),$$

where $p(\theta)$ is the *prior distribution* and $p(\mathcal{D}|\theta)$ is the *likelihood*. These two distributions represent the previous knowledge we have about our parameters and the likelihood of observing the current data for a given set of parameters respectively. In many regression tasks, the likelihood is taken as

$$p(\mathcal{D}|\theta) = \mathcal{N}\left(f_\theta(\mathcal{D}), \sigma^2\right),$$

which under the assumption of independence of each point in the data yields the neat expression

$$p(\mathcal{D}|\theta) = \prod_{i=1}^{N} \mathcal{N}\left(f_\theta(x_i), \sigma^2\right).$$

In our case, the assumption of independence between data points means assuming that A&E attendance numbers in one HBT do not impact the numbers in another HBT, and that each week the attendance numbers are independent of one another (e.g. having one extremely busy week does not mean the next week will be extremely busy). The explicit form of the likelihood is generally not known, so inference techniques are required to compute distribution statistics (such as the σ^2 terms appearing above) [47].

The choice of prior distribution is usually quite heuristic. There are a variety of methods to pick your prior distribution, depending on the level of background knowledge. If one expects the parameters to follow a certain distribution or to be a of a certain scale, then the prior may be chosen to represent this. For example, if you expect your parameters to be very small then you could take a Gaussian distribution centered at zero with small variance to be your prior. Typically Gaussian distributions produce more tractable problems [28]. Empirical methods choose the prior distribution based upon the data itself [28, 49]. This makes the distinction between the prior and the likelihood less clear, but can significantly improve model fit. An example would be picking the prior to be a Gaussian distribution centered around the parameters obtained from training a standard NN of the same structure. However, such methods come with a significant risk of overfitting to the training data [28, 49]. There are priors which impose very little assumptions about the parameter space, such as (the interestingly named) 'Jeffrey's Prior', which do not favour any particular parameter scales [50]. These can be very useful when generating completely new models in new and emerging fields, where you do not have any background knowledge on the parameters. Generally, one needs to find the correct prior which encapsulates your knowledge of the parameter space while ensuring computational tractability (as more complicated priors, which may be more accurate, could produce infeasible problems) [28].

Bayes Theorem [28, 47, 48] relates our posterior, prior and likelihood according to the rule

$$\pi(\theta|\mathcal{D}) = \frac{p(\theta)p(\mathcal{D}|\theta)}{\int p(\theta)p(\mathcal{D}|\theta)d\theta}.$$

The integral term appearing in the denominator is a normalisation term, ensuring that π indeed has the properties of a distribution function (i.e. it integrates to 1 over the whole space of possible parameters). Computing this term is very difficult in practice, and typically one has to result to numerical methods such as Markov Chain Monte Carlo [47]. After the process of finding a good choice of prior and likelihood distribution and computing (an approximation of) the integral of the joint distribution of, we have full knowledge of our desired posterior distribution. The predictions of the model can then be framed in probabilistic language. For example, the exact prediction of a model for input $x_i \in \mathcal{D}$ can be phrased as the expectation over the posterior

$$\mathbb{E}_\pi\left(f_\theta(x_i)\right) = \int f_\theta(x_i)\pi(\theta|\mathcal{D})d\theta,$$

where the integral is taken over the whole possible parameter space [47]. All other typical predictive qualities can be be phrased in a similar form - allowing for the computation of variances and confidence intervals for each output. While it is a great step forward that we now have an explicit form for these quantities, these are still very high dimensional integrals which may need to be computed over arbitrarily large domains. This is the biggest downside of BNNs, reflecting upon why they aren't nearly as commonly seen in practice - they scale very poorly [28]. Hence, this framework cannot be applied in a computationally feasible way to NN structures with more than a few hidden layers without a *significant* amount of luck.

Alas, the cost of considering a more simplistic model may still be worth it when given the ability to gain a reasonable level of trust in these models.