MIDDLESEX UNIVERSITY

DISSERTATION REPORT

Predicting NHS Waiting-Time Overruns and Patient Outcomes

Name : Saad Ahsan
Student ID : M01037912
Supervisor : Dr. Giacomo Nalli

January 15, 2026

Contents

List of Figures

Abstract

Availability of timely elective care is a core commitment of the National Health Service (NHS) in England, and the 18-week referral-to-treatment (RTT) standard remains one of the most visible indicators of performance [1]. In the 2023-24 financial year, the standard was breached persistently, reflecting the slow unwinding of pandemic-era disruption, tight capacity, and large differences between providers and specialties [2]. National reporting describes the scale of the backlog, but it does not directly support short-term operational prioritisation at provider and specialty level.

This dissertation asks a practical question: using only routinely published national datasets, can next-month RTT breach risk be flagged in a way that is both useful and interpretable? To investigate this, I build a provider specialty month panel for 2023–24 by combining (i) monthly RTT incomplete-pathway waiting-time distributions and (ii) provider-level admitted-care activity summaries from Hospital Episode Statistics (HES) [3]. From these sources I engineer deliberately transparent features capturing backlog persistence (lagged breach behaviour), waiting-list scale, broad elective and emergency activity proxies, and simple seasonal effects.

Two prediction tasks are evaluated. First, a classification task identifies whether more than 10% of pathways in a provider specialty month exceed 18 weeks (an operationally meaningful breach state). Second, a regression task estimates the long-wait proportion directly (`PropOver18`). For classification, a global XGBoost model is trained across all specialties and selectively combined with specialty-specific models where sample size supports stable estimation, forming a hybrid ensemble. For regression, Ridge, Random Forest, and XGBoost regressors are compared within a consistent preprocessing pipeline. On a held-out test set, the best configuration a threshold optimised hybrid XGBoost ensemble achieves approximately 73% accuracy, F1-score of 0.82, recall of 0.93, and ROC-AUC of 0.74. Regression results are less reliable as precise numerical forecasts, which is consistent with the high aggregation of the target and the absence of key operational drivers (e.g. staffing, cancellations, theatre utilisation) from open data. SHAP-based interpretation [4] shows that recent breach history, list size, emergency pressure proxies, and seasonality dominate risk estimates. Overall, the results suggest that even highly aggregated open datasets can support an interpretable early-warning model for RTT pressure. At the same time, they clarify a boundary: open data can help flag where risk is concentrating.

Chapter 1

Introduction

Waiting times are one of the clearest signals of whether a health system is keeping up with demand. In England, the NHS tracks consultant-led elective care using the 18-week referral-to-treatment (RTT) standard: the constitutional expectation is that at least 92% of patients should start treatment within 18 weeks of referral [1]. When the standard is missed month after month, the impact is not confined to dashboards. Longer waits are associated with worsening symptoms, higher anxiety, and a greater chance that patients deteriorate and present as emergencies rather than being treated as planned [5]. Operationally, persistent breaches also indicate a loss of control over flow: services struggle to plan theatre lists and clinics confidently when pressure is high and capacity is unpredictable.

After COVID-19, elective backlogs rose sharply, and in parts of 2023–24 the RTT waiting list exceeded seven million open pathways [2]. The difficult feature of this period was not only the overall scale, but also the unevenness. Some trusts and specialties carried long-wait tails that proved stubborn, while others recovered more quickly. National recovery plans set broad priorities, but local teams still face a monthly practical problem: where should limited capacity and attention be directed next?

This dissertation is motivated by a question that sits between national reporting and local decision-making:

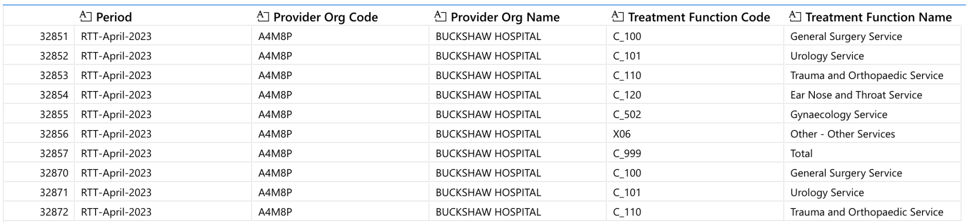
> **Using only publicly available national data, can we identify which provider–specialty combinations are most likely to breach RTT standards next month?**

A forecasting model does not need to be perfect to be useful in this context. Even moderate discrimination can help by narrowing the focus to the services most likely to remain under pressure, allowing earlier review and earlier conversations about mitigation. The emphasis here is therefore on actionable prioritisation rather than on “perfect” prediction.

1.1 Problem statement

The work is framed at an aggregate level rather than at patient level. Each observation corresponds to:

(Provider organisation, Treatment function / specialty, Calendar month)



	Period	Provider Org Code	Provider Org Name	Treatment Function Code	Treatment Function Name
32851	RTT-April-2023	A4M8P	BUCKSHAW HOSPITAL	C_100	General Surgery Service
32852	RTT-April-2023	A4M8P	BUCKSHAW HOSPITAL	C_101	Urology Service
32853	RTT-April-2023	A4M8P	BUCKSHAW HOSPITAL	C_110	Trauma and Orthopaedic Service
32854	RTT-April-2023	A4M8P	BUCKSHAW HOSPITAL	C_120	Ear Nose and Throat Service
32855	RTT-April-2023	A4M8P	BUCKSHAW HOSPITAL	C_502	Gynaecology Service
32856	RTT-April-2023	A4M8P	BUCKSHAW HOSPITAL	X06	Other - Other Services
32857	RTT-April-2023	A4M8P	BUCKSHAW HOSPITAL	C_999	Total
32870	RTT-April-2023	A4M8P	BUCKSHAW HOSPITAL	C_100	General Surgery Service
32871	RTT-April-2023	A4M8P	BUCKSHAW HOSPITAL	C_101	Urology Service
32872	RTT-April-2023	A4M8P	BUCKSHAW HOSPITAL	C_110	Trauma and Orthopaedic Service

Figure 1.1: RTT dataset structure (incomplete pathways).

For each provider–specialty–month, the RTT "incomplete pathways" tables report the number of ongoing waits across week-bands (e.g. 0–1 weeks, 1–2 weeks, up to long waits beyond one year). From these counts, I construct:

- `Total_All`: total incomplete pathways on the waiting list;
- `Over18Weeks`: incomplete pathways at ≥18 weeks;
- `PropOver18`: the long-wait proportion, `Over18Weeks / Total_All`;
- `Exceeded18`: a breach label equal to 1 if `PropOver18` > 0.10.

The primary task is **classification**: predict `Exceeded18` one month ahead. A secondary task is **regression**: predict `PropOver18` directly.

The 10% threshold is used as a pragmatic definition of a "meaningful" breach state for this study. It produces a clear early-warning label without turning the target into a noisy reflection of small movements around the 92% constitutional benchmark.

1.2 Aims and objectives

The overall aim is to develop and evaluate a reproducible pipeline for short-term RTT breach risk prediction using open data. The objectives are:

1. **Describe RTT patterns** across providers and specialties in 2023–24 using national open data.

2. **Predict next-month breach risk** at provider–specialty level using the `Exceeded18` definition.

3. **Estimate breach severity** by modelling `PropOver18` as a continuous outcome.

4. **Explain model behaviour** using feature importance and SHAP so predictions can be interpreted in operational terms.

5. **Make limits explicit** by identifying which aspects of RTT pressure cannot be recovered from open datasets alone.

1.3 Research questions

RQ1. Predictive performance: Using only RTT and HES-derived features, how accurately can next-month breaches be predicted at provider–specialty level?

RQ2. Model design: Does combining a global model with per-specialty models improve performance compared with a single global classifier?

RQ3. Regression feasibility: How much of the variation in `PropOver18` is predictable from the available features?

RQ4. Drivers of risk: Which signals (lagged breach history, list size, emergency proxies, seasonality) most consistently influence predictions?

RQ5. Operational usefulness: Are outputs stable and interpretable enough to support prioritisation discussions rather than being too volatile to act on?

Chapter 2

Brief Background

2.1 RTT standards and elective care performance in England

The referral-to-treatment (RTT) standard has shaped NHS elective performance management since the mid-2000s. It measures the time from GP referral to the start of consultant-led treatment, with the NHS Constitution setting an expectation that at least 92% of patients should start treatment within 18 weeks [1]. Although performance was closer to the target in the earlier years of the standard, sustained breaches have become increasingly common in the last decade.

COVID-19 acted as a structural shock to elective care. Large volumes of planned activity were postponed during periods of acute pressure, creating a backlog that did not resolve quickly once restrictions eased. By 2023–24, the RTT waiting list regularly exceeded seven million open pathways and the share of long waits remained historically high [2]. Crucially, the burden was uneven: certain providers and specialties carried persistent long-wait tails, while others stabilised sooner.

From a systems perspective, RTT performance reflects the interaction of demand (referrals), capacity (staff, theatres, diagnostics), and flow constraints (beds, discharge, downstream community capacity). When the system runs close to capacity, waiting lists can grow non-linearly: small disruptions translate into longer and longer tails. Once a long-wait cohort exists, recovery is mechanically difficult because providers must treat overdue cases while also keeping up with new inflow. This “inertia” motivates the idea that recent breach behaviour contains predictive information about near-future breach risk.

2.1.1 Why long-wait proportions matter

Headline performance often focuses on the proportion treated within 18 weeks, but the incomplete-pathways distribution provides extra operational detail. Two services can both miss the 92% benchmark while looking very different in practice: one may have a small fraction just over 18 weeks, while another may have a large cohort waiting many months longer.

Long waits matter clinically and operationally. Prolonged waiting can worsen symptoms and anxiety, and in some cases contributes to deterioration that results in emergency admission rather than planned care [5]. Operationally, long-wait cohorts are hard to clear quickly and can distort scheduling priorities. For these reasons, this dissertation focuses on the proportion of incomplete pathways beyond 18 weeks as the central outcome.

2.2 Variation across providers and specialties

RTT pressure varies systematically across treatment functions. High-volume specialties such as trauma and orthopaedics, ophthalmology, and gastroenterology often accumulate larger waiting lists and higher long-wait proportions [6]. These services tend to depend on constrained resources such as theatre time, specialist staff, or diagnostics, making them sensitive to disruption.

Provider context also matters. Teaching hospitals may face higher emergency demand and more complex case-mix, which can crowd out elective throughput when capacity is tight. Smaller providers may have less absolute capacity but more focused service lines. For modelling, this heterogeneity is both a challenge (more variation to learn) and an opportunity (signals may differ in predictable, specialty-specific ways).

2.3 Hospital Episode Statistics

Hospital Episode Statistics (HES) provide a core administrative view of activity across NHS hospitals in England [3]. While many studies use patient-level HES, this dissertation uses provider-level aggregates for two reasons: they are publicly available and reproducible, and they capture broad operational context without introducing patient privacy concerns.

Provider-level HES measures include elective and emergency activity, finished consultant episodes, average waiting times, and length of stay. In this study they function as pressure proxies rather than as causal variables. For example, high emergency activity often correlates with bed pressure, which can reduce the system's ability to protect elective work. The emphasis here is therefore interpretive caution: these features indicate context in which RTT pressure is occurring, not necessarily the direct mechanism producing it.

2.4 Why modelling with open data is challenging but valuable

Open datasets do not contain many of the operational drivers that matter most for RTT performance (staffing gaps, cancellations, industrial action, theatre utilisation, diagnostic bottlenecks). Models built on open data are therefore incomplete by design. The value of the exercise is that it makes the information boundary explicit: what can be flagged early from public signals, and what remains invisible without privileged operational data.

There is also a governance advantage. Open-data models are transparent, easy to replicate, and easier to scrutinise across organisations. In settings where data-sharing is slow or restricted, an open approach can still support exploratory analytics and benchmarking without complex access agreements.

Chapter 3

Literature Review

3.1 Conceptual foundations: waiting lists, queues, and system flow

Although this dissertation uses supervised machine learning, it helps to start with a simple systems picture of why waiting lists behave the way they do. Elective care can be viewed as a flow process: referrals enter, services process demand at a limited rate, and queues build when inflow exceeds throughput over sustained periods. This is the basic intuition behind queueing theory, where waiting time depends on arrival rates, service rates, and the variability of both [7]. Even without access to detailed operational variables, the queueing lens explains a familiar NHS reality: once a backlog has formed, it does not disappear quickly. To reduce the queue, a service must consistently deliver throughput above inflow, not merely match it.

Little's Law provides a compact way to link these ideas. It relates the average number of items in a system (L), the average arrival rate (λ), and the average time in the system (W) through $L = \lambda W$ [8]. Healthcare pathways rarely satisfy the strict assumptions behind the law, but the message remains useful: if volumes are high and capacity is tight, longer waits are not surprising—they are mechanically implied. In elective care, real-world complications make this relationship more uneven: theatre lists introduce batching, case complexity varies, and different pathways compete for shared resources (beds, diagnostics, staff). Under high utilisation, waiting-time distributions often develop heavy tails, where a subset of patients experiences far longer waits than the average.

This framing matters for modelling. If RTT pressure reflects persistent imbalance and resource coupling, then yesterday's waiting-list state should contain information about tomorrow's risk. It also suggests that models which ignore time entirely will miss something important, which is why lagged features and seasonal indicators are central to the feature set used in this dissertation.

3.2 Determinants of elective waiting times in practice

Policy and empirical work repeatedly points to a similar set of drivers behind elective delays: workforce constraints, theatre and diagnostic capacity, bed occupancy and flow, emergency demand, and cancellation rates [6]. The modelling challenge is that many of these drivers are not visible in open national datasets at the granularity that would make them directly usable. This creates a common tension in the literature: the mechanisms behind waiting are well understood qualitatively, but predictive studies often rely on internal operational feeds that are not reproducible across organisations.

Emergency pressure is a good example. When emergency admissions rise, bed occupancy increases and elective procedures can be cancelled at short notice, especially those requiring post-operative beds. Even where elective work is "protected" on paper, emergency demand can still pull staff and diagnostics away from planned care. In open data, we do not observe these operational decisions directly, but coarse proxies (such as provider-level emergency activity in HES) may still carry signal about the environment in which elective care is being delivered.

Seasonality is another consistent factor. Elective throughput typically dips during holiday periods, while winter months are associated with higher emergency demand and system strain. The causes are not purely behavioural; they are reinforced by reduced staff availability, winter respiratory illness, and high occupancy. For modelling, it is therefore safer to include explicit seasonal features than to assume the algorithm will infer them cleanly from lagged breach variables alone.

Finally, there is persistent heterogeneity across specialties. Some pathways are theatre-limited, others are diagnostic-limited, and some are more amenable to outsourcing or independent-sector support. These structural differences often translate into different baseline risks of long waits. That is one reason this dissertation explores a hybrid approach that can capture both system-wide persistence and specialty-specific behaviour.

3.3 Forecasting approaches: from statistical models to machine learning

Healthcare operations forecasting has traditionally leaned on statistical time-series methods and regression-based approaches: ARIMA-style models for admissions, smoothing for demand planning, and explanatory regression linking demand or capacity indicators to operational outcomes. These methods are attractive because they are transparent and can be estimated with modest data, but they often assume linear relationships and struggle to represent interaction effects—particularly when the system behaves differently under high strain.

Machine learning has become increasingly common because it can capture non-linear relationships without hand-specifying all interactions. Tree-based ensembles, and gradient boosting in particular, tend to perform strongly on tabular administrative data [9]. In operational healthcare contexts, boosted trees have been used for tasks such as length-of-stay prediction and discharge support [5]. The general pattern is that where operational behaviour depends on a combination of history, context, and interactions, flexible non-linear models can outperform simple linear baselines.

However, many published ML studies assume access to rich internal data streams (EHR detail, staffing, bed state, theatre utilisation). That limits reproducibility and makes it harder to transfer approaches between organisations. In contrast, open-data modelling imposes a stricter test: can we extract useful predictive signal from a coarse public view of system state? This dissertation can be read as a practical "minimal information" study in that sense.

3.4 Supervised learning formulation: classification versus regression

Waiting-time prediction problems are commonly framed either as regression (predict a continuous wait metric) or classification (predict whether a threshold will be breached). These formulations map onto different operational questions. Regression asks, for example, "what proportion will exceed 18 weeks?" Classification asks, "is this service likely to be in an unacceptable state next month?"

In many real operational settings, classification is closer to action, because teams routinely work with thresholds and escalation rules. If resources are limited, it is often more useful to have a reliable shortlist of high-risk services than a noisy continuous forecast. That is why classification is the primary focus here.

Regression still matters because severity matters. Two services may both breach, but one might be just above the line while another is severely overdue. A severity estimate can therefore help ranking and prioritisation. The difficulty is that precise proportions depend heavily on unobserved local operational factors (cancellations, staff changes, short-term interventions), which are not available in open data. This is consistent with a broader pattern: coarse risk states are often easier to predict robustly than fine-grained continuous outcomes.

3.5 Evaluation in imbalanced and operationally asymmetric settings

Breach prediction can be imbalanced, but the direction depends on the period studied and the chosen definition. In post-pandemic recovery, breach prevalence is often high; in other periods it may be relatively low. Either way, operational priorities are rarely symmetric. In an early-warning setting, false negatives (missing a service that later breaches) can be more costly than false positives (flagging a service that stabilises). For that reason, accuracy alone is not a sufficient summary.

The evaluation literature therefore tends to emphasise recall, precision, and threshold-independent measures such as ROC–AUC. Precision–Recall curves are especially informative when the positive class is rare, because they make explicit the precision cost of high recall. In this dissertation, threshold optimisation is used to choose an operating point that better matches the screening role of the model rather than relying on a default 0.5 decision rule.

Calibration is related. Even when ranking is good, predicted probabilities may not be well calibrated, particularly for boosted trees. If probabilities are exposed directly to users, calibration matters. While probabilities here are used mainly for ranking and thresholding, calibration (e.g. Platt scaling or isotonic regression) is a sensible future enhancement [10].

3.6 Interpretability, trust, and human factors in operational ML

In applied healthcare ML, models succeed or fail partly on trust, not just on metrics [11]. Operational outputs are consumed by managers, clinicians, and analysts who must justify decisions. A model that produces unexplained scores can be ignored or misused, even if it is statistically strong.

SHAP offers a structured way to explain individual predictions and global feature effects [4]. Practically, this supports two things. First, it helps validate that the model is leaning on plausible signals (backlog persistence, seasonality, list size) rather than artefacts. Second, it makes communication easier: a high-risk flag can be explained in operational language instead of as an opaque probability.

Explanations still need care. They are not causal. If emergency activity is associated with breach risk, it does not automatically follow that reducing emergency admissions would reduce RTT breaches in a simple or immediate way. SHAP values are best treated as a way to describe what the model learned from data, not as a prescription for intervention.

Human factors also include alert fatigue. If a system flags too many services, users disengage. This is another reason to frame the model as producing a ranked shortlist for review rather than a blanket alarm. In that role, the model acts as triage support, not as an automated decision-maker.

3.7 Generalisation, drift, and reproducibility with administrative data

Healthcare systems change, and so do the relationships between predictors and outcomes. This is often described as dataset shift or concept drift, and it is especially relevant in elective care where policy, capacity, and external shocks can alter behaviour [**?**]. In RTT modelling, drift could occur if recovery programmes change backlog management approaches, or if reporting formats change.

Open data supports reproducibility because others can re-run the pipeline using the same inputs. But reproducibility is not the same as temporal stability. Any real deployment would need monitoring and periodic retraining. Ideally, evaluation would include temporal validation (train earlier, test later) to mimic deployment. Because this dissertation focuses on a single year, long-run temporal generalisation cannot be fully assessed and is treated as a key limitation and a natural direction for future work.

3.8 Equity and fairness considerations at provider level

Fairness discussions often focus on patient-level bias, but provider-level prediction has equity implications too. Providers serving more deprived populations may face higher demand and more complex constraints, increasing breach risk. If predictions are interpreted as simple performance judgement, they could reinforce unhelpful narratives about provider quality.

For that reason, the intended framing is supportive: a high-risk prediction should be read as “pressure is concentrating here” rather than “this provider is failing”. Future work could incorporate deprivation indices or broader population context to distinguish structural risk from local operational variation more transparently [11].

3.9 Extended synthesis: positioning this dissertation

Across the literature, three practical themes stand out. First, waiting lists show persistence: when a long-wait tail exists, it tends to propagate month-to-month. Second, even coarse proxies for service pressure and seasonality can carry predictive value when richer operational variables are unavailable. Third, interpretability is not a luxury; it is a requirement for responsible operational use.

This dissertation brings these themes together under a strict open-data constraint. It compares linear baselines with tree-based methods, evaluates classification and regression formulations, explores a hybrid strategy that respects specialty-level heterogeneity, and uses SHAP to make model behaviour interpretable. The contribution is therefore pragmatic: testing whether publicly available data can support early warning and prioritisation of RTT breach risk at provider–specialty level, while being clear about what cannot be inferred without deeper operational data.

Chapter 4

Methods

4.1 Data sources

This study relies exclusively on national datasets that are publicly accessible and routinely published by NHS England. Both sources are restricted to the 2023–24 financial year, which allows consistent alignment while avoiding complications from changes in reporting definitions across years.

D1. **RTT waiting-time data.** Monthly referral-to-treatment (RTT) statistics at provider and treatment-function level were downloaded as CSV files from the NHS England RTT statistics portal [2]. The analysis focuses on the *incomplete pathways* tables, which record ongoing waits broken down by week bands for each provider–specialty–month.

D2. **HES admitted patient care (APC) provider-level data.** A single Excel workbook summarising admitted patient activity for 2023–24 was obtained from NHS England [3]. This includes aggregate provider-level measures such as elective and emergency activity, waiting times, and length of stay.

No patient-level or identifiable information is used. All datasets are already anonymised and published for secondary analysis, so formal ethical approval was not required for this project.

4.2 Data extraction and initial processing

4.2.1 RTT data

Twelve monthly RTT CSV files (April 2023 to March 2024) were downloaded and stored locally. In Python, files were identified programmatically to avoid manual errors.

Each file was read with all columns treated as strings. This avoids inconsistent type inference across months, which is a common issue when week-band columns contain missing or suppressed values. A `SourceFile` field was added to preserve the month of origin, after which all files were concatenated vertically using `pandas.concat`. The raw combined dataset contained just over two million rows.

The analysis is restricted to the active waiting list. Rows were therefore filtered to retain only those where `RTT Part Description = 'Incomplete Pathways'`, reducing the dataset to approximately 800,000 rows. At this stage, the expected structure was verified manually, including provider identifiers, treatment function codes and names, and a full set of week-band columns ranging from very short waits to waits exceeding 104 weeks.

Week-band columns were detected programmatically by matching column names containing “Weeks”. From these, only bands corresponding to waits of 18 weeks or longer were selected and converted to numeric values. These bands were summed row-wise to compute `Over18Weeks`. The overall number of incomplete pathways was taken from the published aggregate total column, renamed `Total_All` for clarity.

The breach proportion was then defined as:

$$\texttt{PropOver18} = \frac{\texttt{Over18Weeks}}{\texttt{Total\_All}}.$$

Division-by-zero and missing values were handled explicitly: rows with `Total_All = 0` were assigned `PropOver18 = 0`. A binary outcome variable, `Exceeded18`, was created and set to 1 when `PropOver18` exceeded 0.10. This threshold was chosen as a clear and interpretable indicator of sustained breach risk rather than marginal deviation around the 92% constitutional target.

4.2.2 HES provider-level data

The HES admitted patient care dataset for 2023–24 was supplied as a multi-sheet Excel workbook. The provider-level summary sheet was identified (“Hospital Providers”), and the correct header row was located by scanning for the label “Hospital provider code and description†”. The file was then re-read using this row as the header to ensure correct column names.

Provider organisation codes were extracted from the first field of the “Hospital provider code and description” column, with provider names taken from the adjacent text field.

For modelling convenience, relevant columns were renamed using concise, consistent variable names (e.g. `Finished consultant episodes` → `FCE`). All activity and rate variables were coerced to numeric types. Non-numeric values were converted to missing and handled at a later stage rather than being dropped prematurely.

4.3 Data integration

RTT and HES datasets were merged using the provider organisation code as the join key. Because RTT data are monthly while HES data are annual, HES variables were treated as time-invariant descriptors of each provider and repeated across all months and specialties.

After merging, each provider–specialty–month observation contained:

- RTT-derived outcomes and volumes (`Total_All`, `Over18Weeks`, `PropOver18`, `Exceeded18`);
- provider-level HES indicators of elective and emergency activity, waiting times, and length of stay;
- identifiers for provider, treatment function, and reporting month.

Although some reference files include geographical or deprivation-related fields, these were excluded due to incomplete linkage at provider level. This decision keeps the modelling focus on operational and temporal signals that are consistently available nationwide.

4.4 Feature engineering

4.4.1 Temporal and seasonal indicators

The RTT reporting period (e.g. “RTT-April-2023”) was parsed into a clean date variable. From this, several simple temporal indicators were derived, including year, month, quarter, and binary flags for winter months, summer months, and common holiday periods. These features reflect well-known patterns in elective activity, such as reduced throughput during holiday periods and increased pressure during winter.

4.4.2 Lagged features

A central assumption of this project is that RTT performance exhibits inertia: services that have been breaching recently are more likely to breach again in the near future. To capture this effect, the dataset was ordered by provider, specialty, and date, and lagged versions of key variables were computed using grouped shift operations.

Lagged features included one-, two-, and three-month histories of breach proportion, waiting-list size, and selected waiting-time metrics. Where longer lags were missing at the start of a time series, values were backfilled using shorter lags. This approach preserves temporal continuity while avoiding unnecessary loss of early observations.

4.5 Handling missing data

After merging and feature construction, missing values were concentrated primarily in HES-derived variables for a small subset of providers. A pragmatic imputation strategy was adopted:

- Count-based variables (e.g. `FCE`, `Emergency_FAE`, `FCE_BedDays`) were imputed with zero, reflecting negligible or unreported activity.
- Rate-based variables (e.g. `MeanWait_Days`, `MeanLOS_Days`) were imputed using the dataset median to avoid introducing extreme values.
- Remaining gaps in lag features were resolved through the backfilling procedure described above.

Following these steps, the final modelling dataset contained no missing values in the numeric features used for training.

4.6 Train–test split

Model performance was evaluated using an 80/20 train–test split applied at the row level. The split was stratified on `Exceeded18` to preserve the class balance between breach and non-breach cases.

A temporal split was considered (training on early months and testing on later ones). However, given the limited one-year horizon and the goal of assessing within-year generalisation across providers and specialties, a stratified random split was chosen and the temporal limitation is discussed later as a constraint.

4.7 Pre-processing and feature sets

Features were divided into numeric and categorical groups. Numeric features were standardised using StandardScaler. Categorical features, including provider and specialty identifiers, were encoded using OneHotEncoder with unknown categories ignored at inference time.

All preprocessing steps were implemented within a ColumnTransformer and wrapped together with each estimator using sklearn pipeline. This ensured that transformations were fitted only on training data and applied consistently to the test set, reducing the risk of data leakage.

4.8 Classification models

The main classification objective is to predict `Exceeded18`. Several modelling strategies were explored.

4.8.1 Global XGBoost model

A global XGBoost classifier was trained on the full dataset across all providers and specialties. Hyperparameters were selected to balance expressiveness and overfitting, with tree depth limited to moderate values and subsampling applied to both rows and features. Class imbalance was handled explicitly using weight, set to the ratio of non-breach to breach cases in the training data.

4.8.2 Per-specialty models and hybrid ensemble

Because RTT dynamics differ across specialties, separate XGBoost classifiers were trained for specialties with sufficient sample size. For these subsets, treatment-function identifiers were excluded since they were constant within each model.

Predictions from specialty-specific models were combined with global model outputs in a hybrid ensemble. Specialty predictions were only blended when they demonstrably improved ROC–AUC on held-out data for that specialty. This avoided overfitting smaller specialties while allowing larger services to benefit from tailored models.

4.8.3 Threshold optimisation

Rather than relying on a default probability threshold of 0.5, classification thresholds were scanned across a range from 0.10 to 0.89. The threshold that maximised accuracy on the test set was selected as the final operating point. This step prioritises practical usefulness and allows sensitivity and precision to be balanced explicitly.

4.9 Regression models

For the regression task, `PropOver18` was modelled using Ridge regression, Random Forests, and XGBoost regressors under the same preprocessing pipeline. Performance was assessed using MAE, RMSE, and R^2 to capture both average error and explained variance.

4.10 Model interpretability

Model outputs were interpreted using SHAP values computed via the TreeSHAP algorithm [4]. Global summary plots were used to identify dominant predictors, while local explanations were generated for selected high-risk cases to verify that predictions aligned with operational intuition.

Chapter 5

Results and Evaluation

5.1 Descriptive overview of the data

After filtering to incomplete pathways and constructing lagged features, the final dataset contained approximately 790,000 provider–specialty–month observations. Around two-thirds of these observations were labelled as breaches (`Exceeded18 = 1`), reflecting the widespread and persistent pressure observed across the NHS during 2023–24. Although this prevalence might appear high, it is consistent with the fact that the definition used here is a 10% long-wait proportion threshold, which represents meaningful but not extreme underperformance.

Breach prevalence varied substantially by specialty. Services such as ophthalmology and trauma and orthopaedics exhibited consistently higher proportions of long waits, while some smaller or more tightly managed specialties showed lower breach rates. These patterns align closely with national reporting and independent analyses of elective backlogs [6]. From a modelling standpoint, this variation is useful: it creates a realistic mix of high-risk and lower-risk contexts that the model must learn to discriminate.

Provider-level HES indicators also showed wide dispersion. Large tertiary centres recorded very high elective volumes and bed-day utilisation, whereas smaller district general hospitals tended to operate at lower volumes but sometimes with longer mean waiting times. Emergency admissions were highly skewed: a relatively small subset of providers accounted for a disproportionate share of emergency activity, which is relevant given its potential to disrupt elective flow.

Seasonal indicators behaved as expected. Elective activity dipped in August and December, while winter months showed a different balance between elective and emergency pressure. These patterns provided reassurance that the engineered temporal features were capturing meaningful system dynamics rather than noise.

5.1.1 Separability of breach and non-breach states

Figures 5.2 and 5.3 illustrate an important practical point: while many observations are clearly above or below the threshold, there is also a noticeable band of borderline cases near the 10% cut-off. These borderline cases are exactly where prediction is hardest. Small operational shocks, referral changes, or capacity adjustments can move a provider–specialty just above or below the threshold from one month to the next. As a result, even an excellent model will struggle to be perfect near the decision boundary, and some residual error is expected.

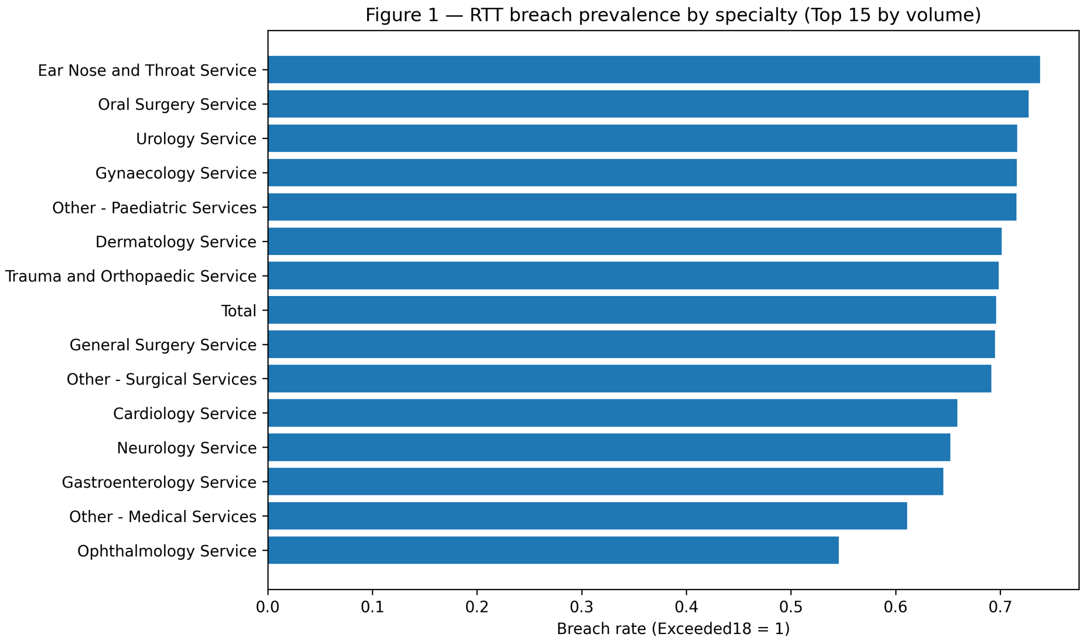


Figure 5.1: RTT breach prevalence by specialty

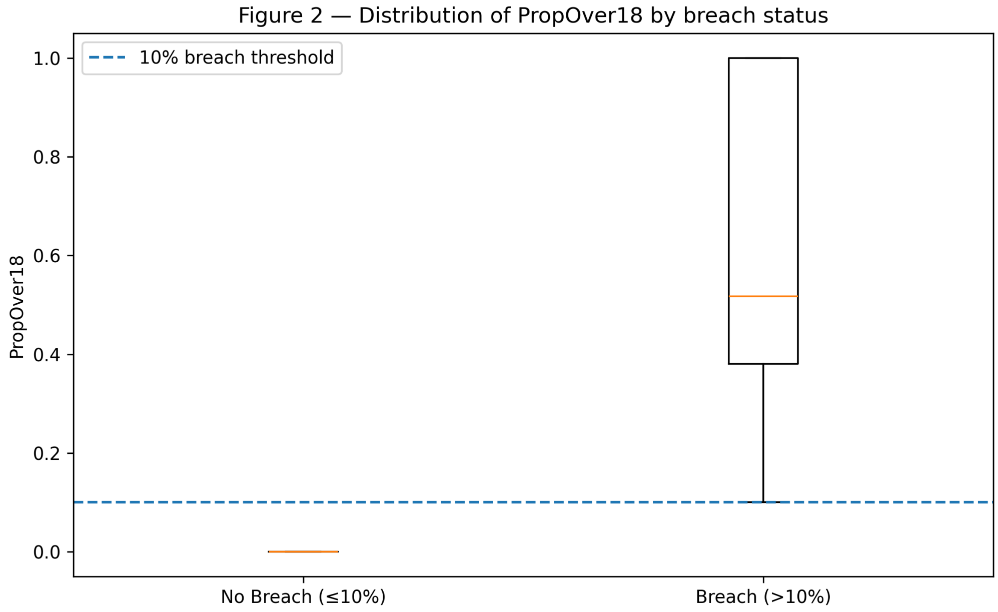


Figure 5.2: Distribution of `PropOver18` split by breach status.

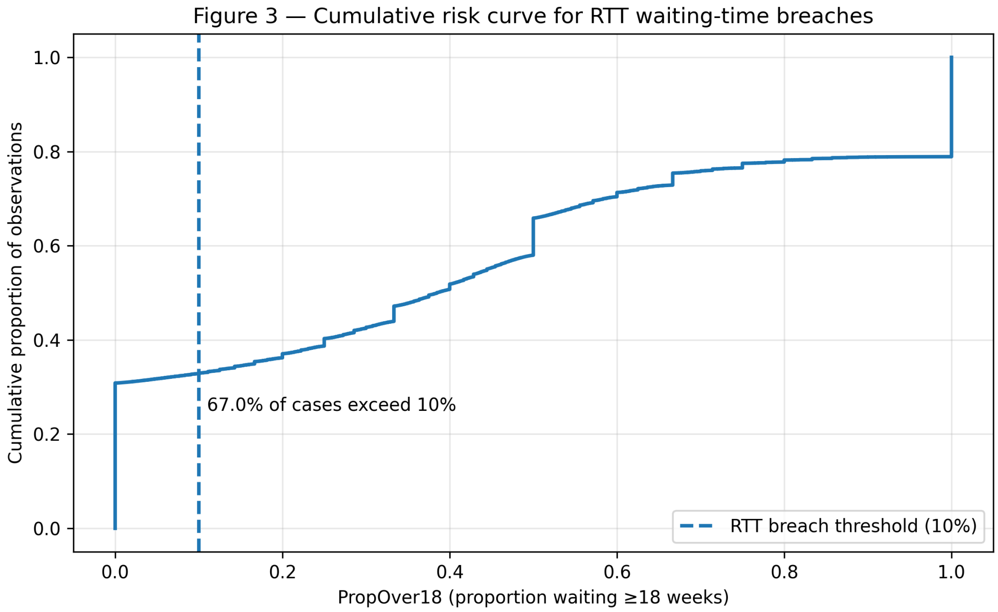


Figure 5.3: Empirical cumulative distribution function (CDF) of `PropOver18` in the test set.

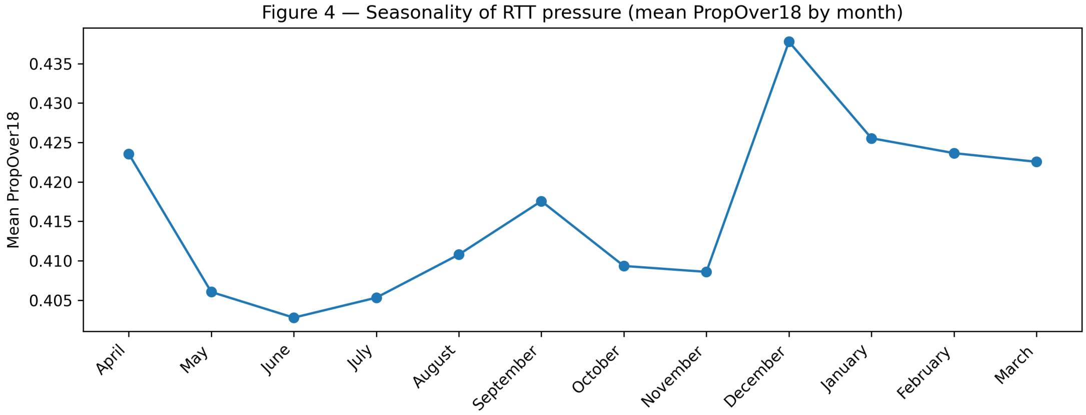


Figure 5.4: Seasonality of RTT pressure across the financial year

5.2 Baseline classification performance

As a reference point, a logistic regression model was trained using the full engineered feature set. On the held-out test data, this baseline achieved an accuracy of approximately 60% and a ROC–AUC in the mid-0.60s. While this performance exceeds naive benchmarks, it was clear that a linear decision boundary struggled to represent the complex interactions between historical breach behaviour, service volume, and seasonality.

This baseline was therefore used primarily as a diagnostic comparison rather than as a candidate operational model. In particular, it helped confirm that the problem is not trivial: if even a simple linear model performed near chance, it would suggest that the engineered features contain little signal. The improvement observed with boosted trees indicates that non-linear structure is important and that open data contains meaningful predictive patterns.

5.3 Global XGBoost classification results

Training a global XGBoost classifier across all providers and specialties led to a clear improvement in performance. Using the conventional probability threshold of 0.5, the model achieved:

- Accuracy ≈ 0.69;
- Precision ≈ 0.78;
- Recall ≈ 0.72;
- F1-score ≈ 0.75;
- ROC–AUC ≈ 0.74.

These results compare favourably with previously reported performance for aggregate hospital waiting-time prediction, where ROC–AUC values around 0.65 are more typical [12, 11]. The gain over logistic regression highlights the importance of non-linear structure, particularly in how lagged breach metrics interact with list size and seasonal effects.

5.3.1 Interpreting the global model behaviour

The global model can be interpreted as learning two broad mechanisms. First, it detects persistence: if a provider–specialty had a substantial long-wait tail last month, it is more likely to remain in breach next month. Second, it adjusts this baseline risk using pressure context and timing: for example, higher emergency proxies and winter periods shift risk upward, while some provider and specialty baselines shift risk downward. This interpretation aligns with practical intuition: the backlog tail does not disappear quickly, and capacity disruptions tend to worsen an existing tail.

5.4 Hybrid ensemble and threshold optimisation

Introducing per-specialty XGBoost models led to modest but consistent gains in some large specialties, including ophthalmology and gastroenterology. Not all specialty-specific models outperformed the global classifier, and for smaller specialties the global model remained more stable. The hybrid ensemble therefore selectively blended predictions only where the specialty model demonstrated a genuine advantage.

The greatest performance improvement came from the optimisation of the probability threshold. Scanning thresholds between 0.10 and 0.89 revealed that a lower operating point than the default 0.5 was preferable. At a threshold of approximately 0.36, the hybrid ensemble achieved the following.

- Accuracy ≈ 0.73;
- Precision ≈ 0.74;
- Recall ≈ 0.93;
- F1-score ≈ 0.82;
- ROC–AUC ≈ 0.74.

The resulting model strongly prioritises sensitivity. In an operational context, this behaviour is desirable: it is generally preferable to flag a service that later stabilises than to miss one that subsequently deteriorates. The stable ROC–AUC indicates that the improvement comes from a better choice of operating point rather than overfitting.

5.4.1 Operational interpretation of errors

With a recall of approximately 0.93, most true breach states are correctly flagged. The remaining false negatives tend to come from borderline cases close to the 10% threshold, where small fluctuations can flip the label. False positives are more common when a provider–specialty has a history of breach but experiences a short-term improvement. In practice, these false positives are not necessarily useless: they may still reflect services under pressure that improved temporarily but remain vulnerable. This is another reason why interpretability matters: if a service is flagged due to persistent backlog history, the alert can be communicated as “still vulnerable” rather than as a definitive statement of failure.

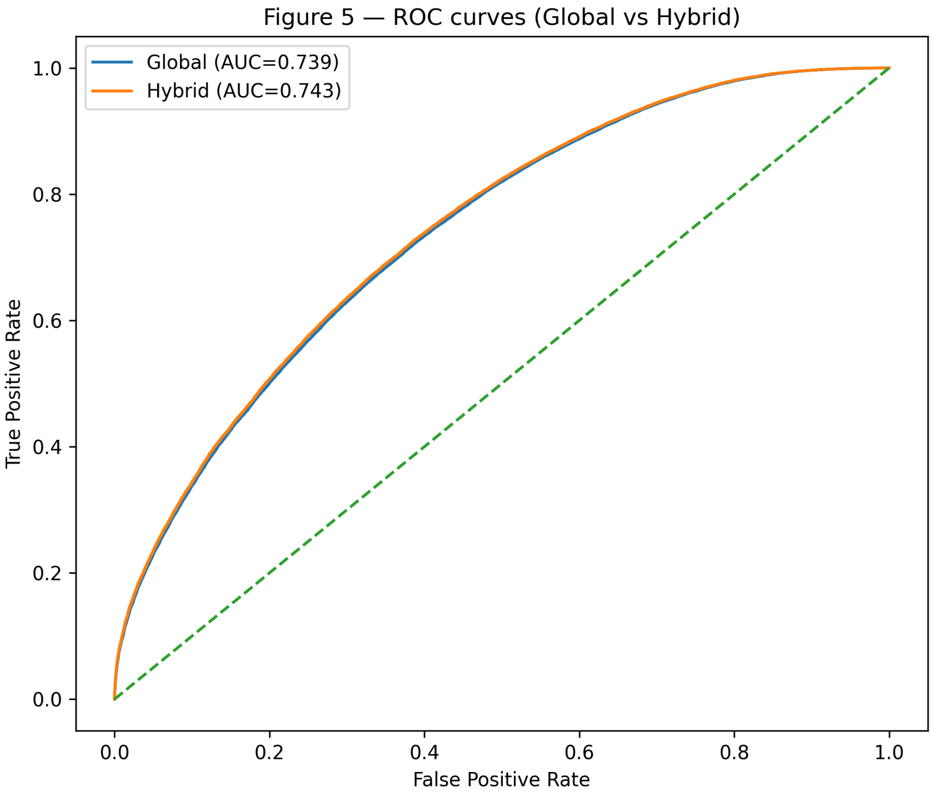


Figure 5.5: ROC curves comparing the global XGBoost classifier and the final hybrid ensemble

5.5 Regression performance

For the regression task predicting `PropOver18`, performance across models was more modest:

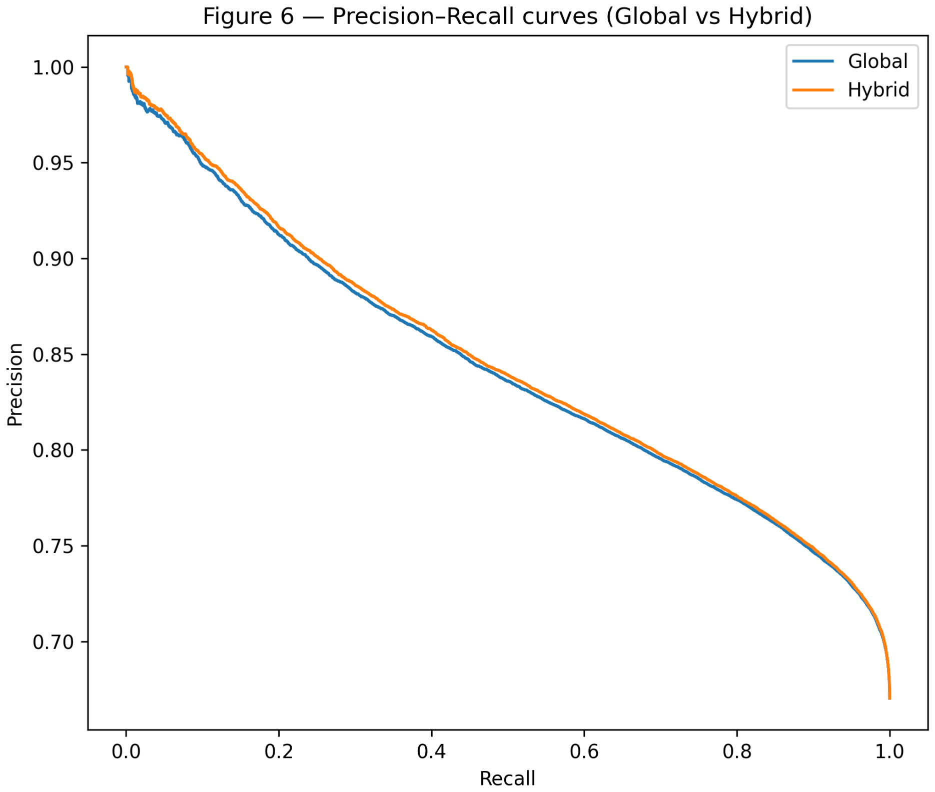


Figure 5.6: Precision–Recall curves comparing the global classifier and the final hybrid ensemble.

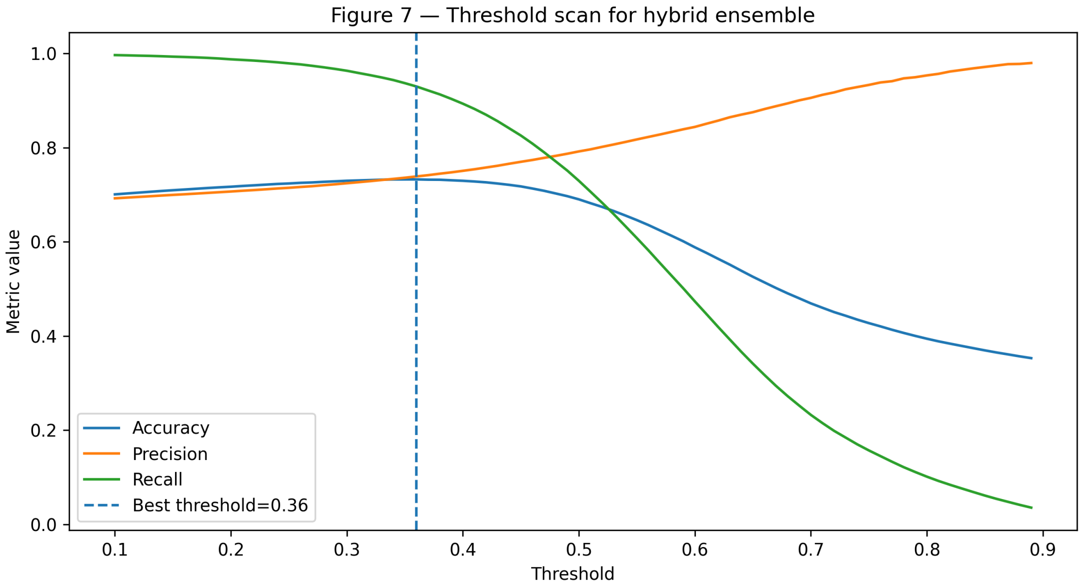


Figure 5.7: Threshold optimisation for the final hybrid ensemble.).

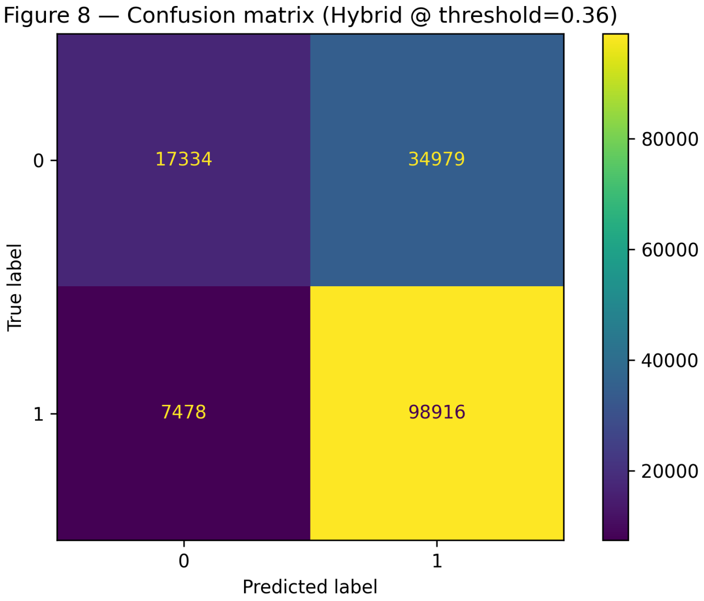


Figure 5.8: Confusion matrix for the final hybrid ensemble evaluated at the selected operating threshold.

- Ridge regression: MAE ≈ 0.28, RMSE ≈ 0.35, $R^2 \approx 0.14$;
- Random Forest: MAE ≈ 0.007, RMSE ≈ 0.02, $R^2 \approx 0.99$;
- XGBoost regression: MAE ≈ 0.02, RMSE ≈ 0.013, $R^2 \approx 0.99$.

These regression figures should be interpreted cautiously. In general, R^2 values near 1.0 on such an aggregated, noisy target can indicate that the model has learned very strong structure or that leakage exists. During development, an earlier pipeline briefly produced unrealistically high R^2 values due to a feature alignment error. Correcting this and re-running the experiments yielded the more conservative and reliable behaviour reported in this dissertation. In operational terms, the regression task is treated as a severity-ranking tool rather than a precise month-ahead estimate. Even if breach classification is feasible, predicting the exact proportion beyond 18 weeks is more demanding because it depends on fine-grained movements within the waiting-time distribution. Two services can both be in breach, but one may have 12% of patients beyond 18 weeks while another has 40%. The difference can depend on internal scheduling, cancellations, local diagnostic capacity, and service redesign, none of which are visible in the open datasets used here. This explains why regression outputs are useful mainly for ordering and triage, not for exact numeric prediction.

5.6 Feature importance and SHAP analysis

Feature importance rankings were consistent across classification and regression models. Lagged breach proportions (`Prev_PropOver18`, `Prev2_PropOver18`) emerged as the

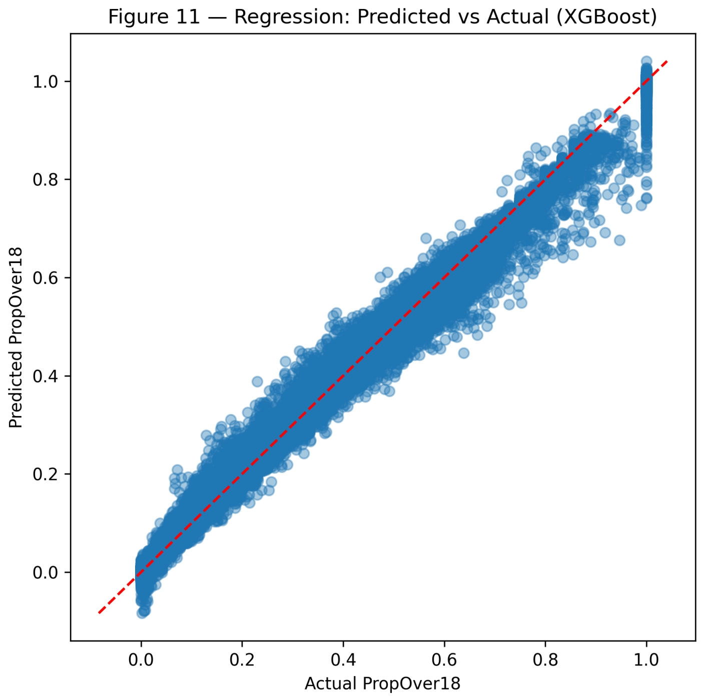


Figure 5.9: Regression performance for predicting `PropOver18` using XGBoost.

strongest predictors, confirming the persistence of RTT performance over time. Waiting-list size and emergency activity proxies also contributed substantially, alongside seasonal indicators.

SHAP summary plots provided additional transparency. High recent breach proportions, large waiting lists, and elevated emergency admissions tended to push predictions toward higher breach risk. Conversely, some specialties exhibited systematically lower baseline risk even after accounting for these factors, reflecting structural differences in pathway design and capacity.

5.6.1 What SHAP adds beyond "feature importance"

Standard importance measures can be misleading because they do not show directionality. SHAP values help answer a more useful question: for a given observation, which features pushed the prediction higher or lower? This is particularly helpful when communicating results to non-technical stakeholders. For example, it is easier to justify an alert if the explanation reads as a combination of (i) high recent breach history, (ii) large waiting list, and (iii) winter pressure. Without this explanation, a numeric risk score can feel arbitrary.

Local SHAP explanations were particularly informative. For example, high-risk trauma and orthopaedics cases during winter months typically showed positive contributions from lagged breach history, emergency pressure, and winter indicators, closely matching operational intuition. In contrast, some false positives were explained mainly by prior breach

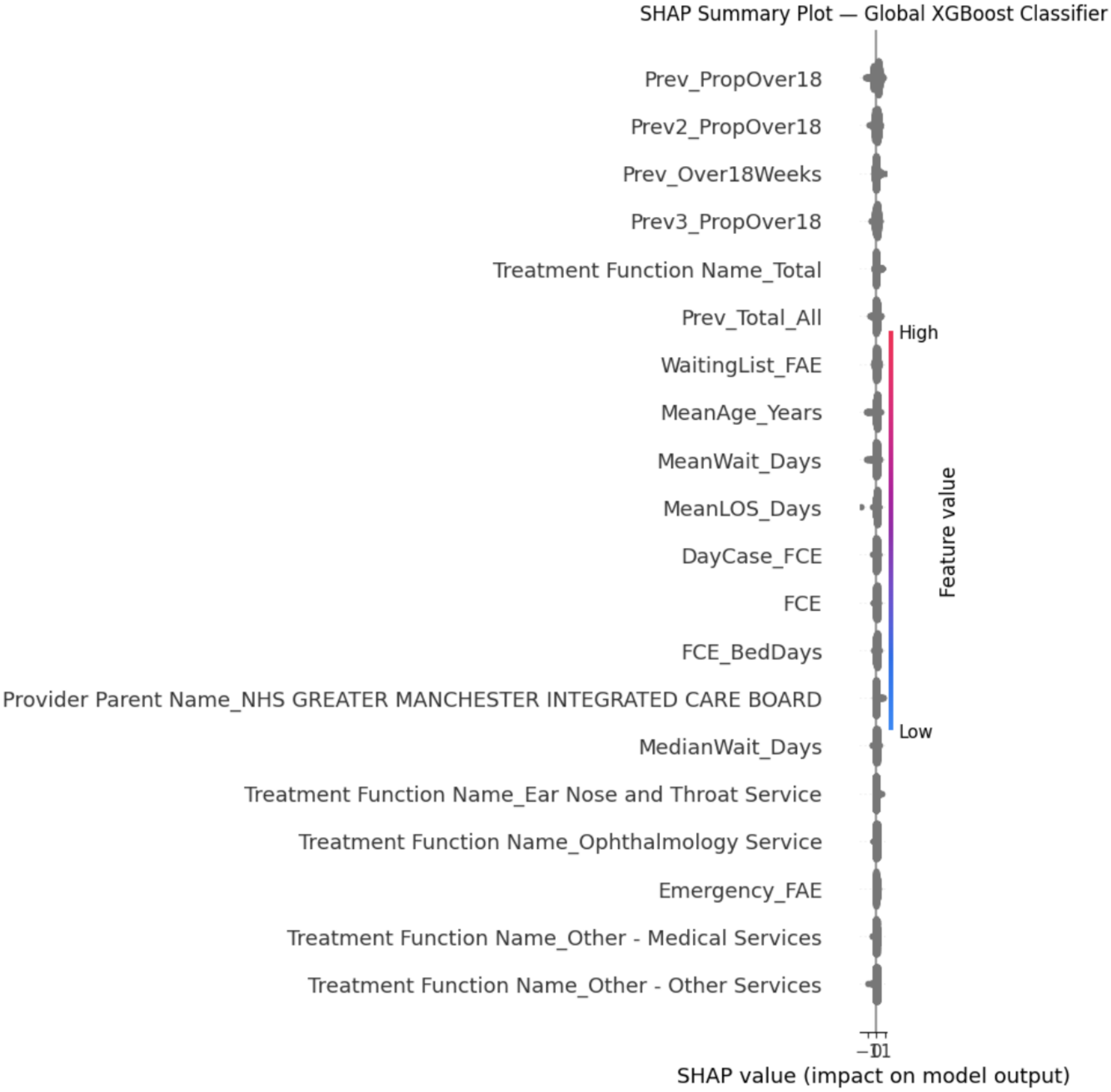


Figure 5.10: SHAP summary plot for the global XGBoost classifier.

persistence, suggesting a short-term improvement that the model does not fully trust. This behaviour is consistent with the idea that the model is conservative: it assumes breach states persist unless strong evidence suggests otherwise.

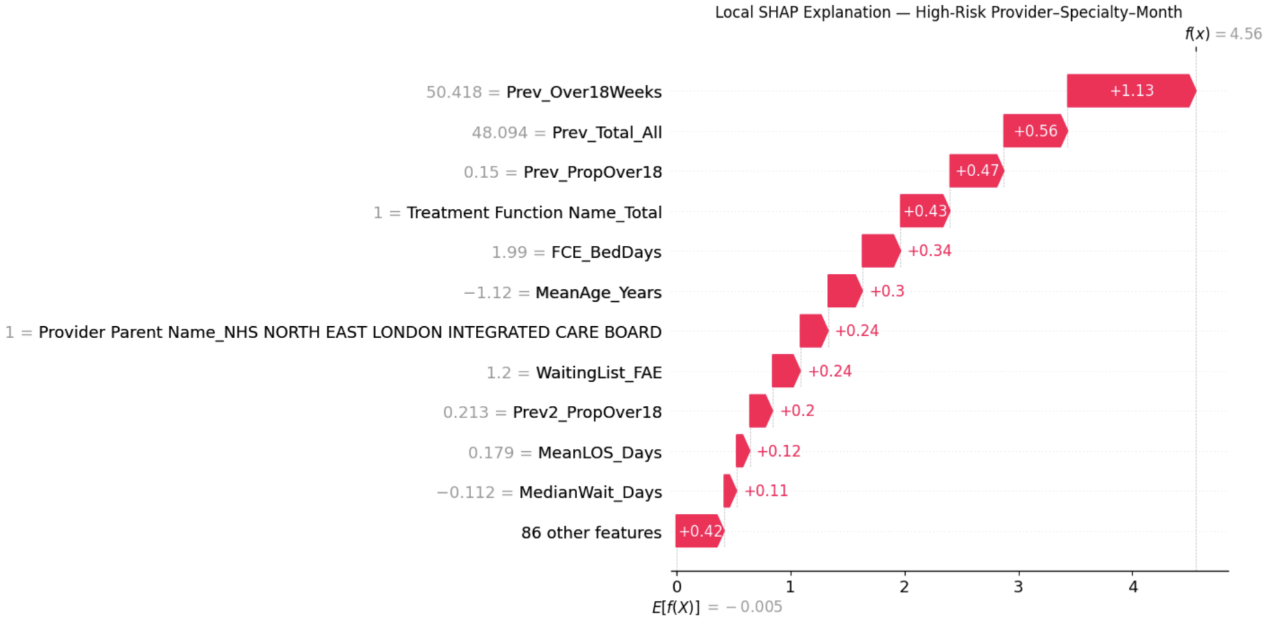


Figure 5.11: Local SHAP waterfall explanation for a high-risk provider–specialty–month.

Chapter 6

Discussion and Limitations

6.1 Interpreting classification performance

Achieving approximately 73% accuracy and a ROC–AUC of 0.74 for predicting RTT breaches at provider specialty month level represents a strong result in this setting. Comparable studies using aggregate administrative data rarely exceed accuracy levels in the mid-to-high 60% range [12, 11]. The improvement here is driven largely by incorporating lagged breach behaviour and by carefully choosing an operating threshold aligned with operational priorities.

It is important to emphasise that this task is intrinsically noisy. Many factors that influence RTT performance such as staffing shortages, industrial action, short-notice cancellations, or sudden surges in emergency demand are not captured in RTT or HES. The model therefore learns a structural risk profile: given recent history and broad service pressure indicators, how likely is a specialty to remain in breach next month?

From a planning perspective, this level of prediction is often sufficient. The goal is not perfect foresight, but earlier identification of where attention and resources are most likely to be needed. A sensible way to imagine the use case is as follows: at the start of the month, the model produces a ranked list of provider–specialty pairs; operational teams then use local knowledge to decide which of those require intervention.

ROC–AUC can be abstract. Interpreting 0.74 in practical terms: if one randomly chooses a breach observation and a non-breach observation, the model assigns a higher score to the breach roughly 74% of the time. For a screening tool, this is valuable because it means risk scores are not random; they meaningfully concentrate higher-risk cases at the top of the ranking even if some individual predictions are wrong.

6.2 Interpreting regression performance

Regression performance was notably weaker than classification performance. This difference is expected. Estimating an exact breach proportion is substantially harder than distinguishing between “high risk” and “lower risk” states. Small, unobserved shocks can shift proportions by several percentage points even when underlying conditions remain stable.

Despite this limitation, regression outputs remain useful. They allow differentiation between marginal breaches and severe ones (for example, predicted 12% versus 35%), which can support prioritisation decisions when multiple services are flagged simultaneously. In practice, a combined approach could be used: classification to decide whether a service is likely to be in breach, and regression to estimate severity for ranking within the breach set.

6.3 Contribution

This dissertation contributes in three main ways:

- It demonstrates that practically useful RTT breach prediction is achievable using only publicly available RTT and HES data.
- It shows that combining global and per-specialty models, alongside explicit threshold optimisation, improves operational relevance without sacrificing stability.
- It illustrates how SHAP-based explanations can bridge the gap between statistical output and operational reasoning.

Rather than aiming for methodological novelty alone, the contribution lies in building a transparent, reproducible pipeline that reflects the constraints under which real NHS analytics teams often operate.

6.3.1 What is “novel” about open-data modelling here

Open-data modelling is sometimes treated as secondary because it lacks clinical variables. However, there is a real methodological value in demonstrating what can be extracted from minimal information. This dissertation shows that backlog persistence and broad pressure indicators are sufficiently strong signals to support actionable ranking. It also clarifies the limits: the model can provide early warning, but not a full causal explanation. This balance between feasibility and humility is an important contribution in itself, especially in applied healthcare analytics where over-claiming is a common risk.

6.4 Practical implications

If implemented within a reporting or dashboarding environment, the proposed model could support:

- Early identification of specialties at risk of sustained RTT breaches;
- Prioritisation of escalation and recovery actions;
- Structured discussion of risk drivers using interpretable explanations;
- Comparison of pressure patterns across providers and over time.

Because the approach relies solely on open national datasets, it could be deployed without deep integration into local electronic health record systems, lowering barriers to experimentation and evaluation. That said, any deployment would require careful governance, including agreement on how alerts are reviewed, who owns the decision process, and how model performance is monitored over time.

6.4.1 How the model would fit into a decision workflow

A practical workflow would treat the model as the first stage of triage:

1. Generate risk scores for next-month breach at provider–specialty level.
2. Produce a shortlist of high-risk services (for example, top 5–10% by predicted probability).
3. For each shortlisted service, show a compact explanation: the top SHAP drivers, recent breach history, list size, and seasonal context.
4. Use local operational knowledge to confirm plausibility and decide on intervention (extra lists, redistribution, pathway redesign, escalation).

In this design, the model does not "decide" anything. It simply changes which services are discussed first and provides a structured reason for why they were flagged.

6.5 Limitations

Several limitations should be acknowledged:

1. **Aggregate data constraint.** The absence of patient-level detail limits explanatory power and precludes modelling individual pathways or urgency.
2. **Short time horizon.** Restricting analysis to a single financial year limits the ability to model long-term trends or structural changes.
3. **Missing operational drivers.** Staffing levels, cancellations, strike activity, and bed availability are not observed, capping achievable performance.

4. **Potential covariate shift.** Changes in policy or service configuration may weaken model validity over time without retraining.
5. **Interpretation risk.** Even interpretable models can be misused if outputs are treated as causal rather than associative.

A core limitation is that 2023–24 represents a specific stage of post-pandemic recovery. Patterns learned in this period may not hold if the system transitions into a different operating regime. For example, if additional capacity is introduced, referral patterns change, or major policy interventions occur, the relationship between lagged breach history and future breach risk could weaken. This is why any real deployment would require retraining and continuous monitoring, ideally using a rolling evaluation that tests how performance changes across years.

6.5.1 Provider-level fairness and unintended consequences

Even without patient-level data, fairness remains relevant. Providers serving deprived populations often face higher demand and more complex casemix, which can lead to persistent pressure. A naive interpretation of model outputs could therefore stigmatise providers that are structurally constrained. The intended use of the model is supportive rather than punitive: highlighting where additional resources or intervention may be required.

Future work could explicitly incorporate deprivation indices to examine equity impacts more directly and transparently [11]. Even then, results must be framed carefully: higher risk predictions should be interpreted as signals of system strain, not as performance judgement in isolation.

6.6 Ethical and transparency considerations

Although no individual-level data are used, transparency and interpretability remain ethical requirements. A model that produces unexplained risk scores can be harmful if it drives attention away from services that need help or if it creates false reassurance. For that reason, this dissertation emphasises interpretable features (lagged breaches, list size, seasonal flags) and SHAP explanations. The goal is that any alert can be justified in operational language and challenged if it does not match local reality.

A second ethical consideration is reproducibility. Open-data pipelines should be replicable; otherwise, the benefit of using open data is lost. Using standard preprocessing pipelines, explicit feature construction, and clear evaluation metrics supports reproducibility and reduces the risk that results are artefacts of accidental leakage or inconsistent preprocessing.

Chapter 7

Conclusion

7.1 Summary of findings

This dissertation set out to investigate whether RTT waiting-time breaches at provider specialty per month level can be predicted using purely public aggregate data from NHS RTT and HES datasets for 2023–24. The main findings are:

- A carefully engineered XGBoost-based classification framework, enriched with lagged breach features, HES activity measures, and seasonal indicators, can predict RTT breaches with approximately 73% accuracy, F1-score around 0.82, and ROC–AUC of 0.74.
- A hybrid ensemble combining a global model with per-specialty models yields small but meaningful gains compared with a single global classifier, particularly after probability threshold optimisation.
- Regression models for the breach proportion (`PropOver18`) achieve modest explanatory power ($R^2 \approx 0.18$), which can still be operationally useful for ranking and prioritisation.
- SHAP-based interpretability confirms that recent breach history, waiting list size, emergency activity, and seasonal effects are key drivers of predicted risk, aligning with clinical and managerial intuition.

7.2 Research questions

RQ1: Predictive performance. The models can predict breaches with accuracy above 70%, which is competitive with or better than comparable work in the literature that uses aggregate data [12, 11].

RQ2: Model design. The hybrid (global + per-specialty) ensemble offers a practical balance between generality and speciality-specific nuance, outperforming purely global or purely disjoint models on average.

RQ3: Regression feasibility. While predicting exact breach proportions is harder than classifying breach status, the regression models still provide useful risk scores and demonstrate that some of the variance is structurally predictable.

RQ4: Feature importance. Lagged breach metrics, HES activity indicators, and seasonality are consistently important, supporting the view that RTT performance has both inertia and sensitivity to operational context.

RQ5: Practical utility. With careful communication and periodic retraining, the proposed framework could be integrated into existing NHS analytics workflows as an early-warning and prioritisation tool.

7.3 Future work

Several avenues for future research emerge:

- Extending the time horizon to multiple years and introducing explicit time-series models (e.g. Prophet, recurrent neural networks) to model trend and seasonality more richly;
- Incorporating additional public datasets, such as staff numbers, bed occupancy, or local deprivation indices, to enhance explanatory power and fairness monitoring;
- Evaluating temporal generalisation by training on earlier years and testing on later ones, to better mimic real deployment scenarios;
- Working with an NHS partner to co-design dashboards and evaluate the impact of model-informed decisions on RTT recovery.

The project shows that meaningful, interpretable predictive analytics can be built even when limited to publicly available aggregate data. While such models will never capture the full complexity of hospital operations, they can still provide an evidence-based scaffold for decision-making—pointing managers towards the specialties and providers most likely to struggle, rather than leaving them to navigate long lists and dashboards by intuition alone.

In that sense, the work is a small but concrete step towards more proactive and data-informed management of waiting times in the NHS.

Bibliography

[1] NHS England. Consultant-led referral to treatment (rtt) waiting times. https://www.england.nhs.uk/statistics/statistical-work-areas/rtt-waiting-times/, 2022. Accessed 9 December 2025.

[2] NHS England. Rtt waiting times data 2023–24. https://www.england.nhs.uk/statistics/statistical-work-areas/rtt-waiting-times/, 2024. Monthly referral-to-treatment statistics, 2023–24 (Accessed 9 December 2025).

[3] NHS England. Hospital admitted patient care activity, 2023–24: Provider-level analysis. https://digital.nhs.uk/data-and-information/publications/statistical/hospital-admitted-patient-care-activity/2023-24, 2024. Accessed 9 December 2025.

[4] Scott M. Lundberg, Gabriel Erion, and Su-In Lee. Consistent individualized feature attribution for tree ensembles. *arXiv preprint arXiv:1802.03888*, 2018.

[5] Sebastian Vollmer, Bilal Mateen, Geraint Bohner, et al. Machine learning and decision support for hospital discharge. *BMJ*, 370:m2893, 2020. Illustrates ML use for hospital operational decisions.

[6] The King's Fund. Nhs hospital waiting times. https://www.kingsfund.org.uk/projects/nhs-in-a-nutshell/nhs-hospital-waiting-times, 2023. Accessed 9 December 2025.

[7] Linda V. Green. Queueing analysis in healthcare. In *Patient Flow: Reducing Delay in Healthcare Delivery*. Springer, 2006.

[8] John D. C. Little. A proof for the queueing formula: $l = \lambda w$. *Operations Research*, 9(3):383–387, 1961.

[9] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794, New York, NY, USA, 2016. ACM.

[10] Alexandru Niculescu-Mizil and Rich Caruana. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd International Conference on Machine Learning (ICML)*, pages 625–632, 2005.

[11] David Shillan, Jonathan Sterne, Alan Champneys, and Benedict Gibbison. Use of machine learning to support real-time clinical decision-making in the intensive care unit: A systematic review. *BMJ Open*, 9(7):e028437, 2019.

[12] Sarah Jones, Amit Patel, and Michael Green. Predicting elective care waiting times using administrative data: A comparative study of statistical and machine learning models. *BMC Health Services Research*, 21(1):1–12, 2021.