Towards a fairer, data-informed reimbursement system for burn patients in England

Chimdimma Noelyn Onah[a], Richard Allmendinger[a], Julia Handl[a], Ken W. Dunn[b]

[a] *Alliance Manchester Business School, The University of Manchester, Oxford Road, Manchester, M13 9PL, Manchester, United Kingdom*
[b] *Department of Burns and Plastic Surgery, University Hospital South Manchester, Southmoor Road, wythenshawe, M23 9LT, Manchester, United Kingdom*

Abstract

We use quantitative data from burn patients in England to investigate whether a data-driven approach can lead to a fairer reimbursement system compared to the current expert-generated rule-based approach, also known as Health Resource Groups (HRGs). Currently, HRGs aim to group and calculate treatment costs of patients similar in terms of clinical characteristics and resource usage. However, research has shown that HRGs are insufficiently homogeneous in clinical characteristics and resource usage, especially for patients treated in highly specialised services, leading to an inequitable reimbursement of funds. Our analysis comprises a machine learning approach, and the results confirm that the proposed approach identifies groups with increased homogeneity compared to the current HRG groups. This implies that burn patients with, for example, similar injury profiles and an expected low cost of treatment are appropriately identified as members of the same group, allowing for the equal reimbursement of treatment costs regardless of the burn service attended. Our results show that the efficient utilisation of data, augmented with expert advice, can improve the reimbursement system to ensure an equitable and fair distribution of funds.

Keywords: payment system, health utilisation, explainable artificial intelligence, health resource groups, clustering, burn care

1. Introduction

In the United Kingdom (UK), the National Health Service (NHS) financing can be considered an instrument to implement change and improve the nation's health by focusing on improving outcomes. One of the payment systems currently adopted is the activity-based payment system. The activity-based payment system aims to pay healthcare providers for each patient seen or treated, considering the patient's expected healthcare need [1]. The expected healthcare need is based on historical data of similar patients.

The current activity-based payment system is tailored for different care types through Health Resource Groups (HRGs) to ensure a financing system that fits its purpose [2]. HRGs are decision rules generated with patient-level data and by experts to suit each diagnosis type. There are used to identify units of healthcare for which payment is made through the three building blocks shown in Figure 1 and elaborated in Section 2.

This paper focuses on improving the grouping logic used to create HRGs. The grouping rules are generated by transcribing expert advice into if-else rules meant to capture differing patient injury severity and Length of Stay (LOS). For HRGs to act as an activity-based payment system, each generated group should be a clinically meaningful group of diagnoses and interventions consuming similar NHS resource levels. However, given the use of expert transcribed if-else rules, the aim of having homogeneous groups in terms of resource usage might not be met due to:

- The methodology's dependence on expert advice carries the risk of ignoring less established factors that account for certain patient subgroups' case complexity.
- The methodology's group evaluation solely relies on LOS as a sole indicator of resource usage when it is known as an incomplete indicator of resource usage [3].

The points mentioned above may lead to significant disadvantages for specialised services. This is especially for small and highly specialised services that, by necessity, operate at very high expenditure [4, 5, 6]. Burn services are one such service – small, it relies on specialist equipment and intervention, deals with a variety of complex cases, and stays open regardless of the number of patients admitted with a minimum staff number constantly on the rota. The significant disadvantage is due to payment rate calculation

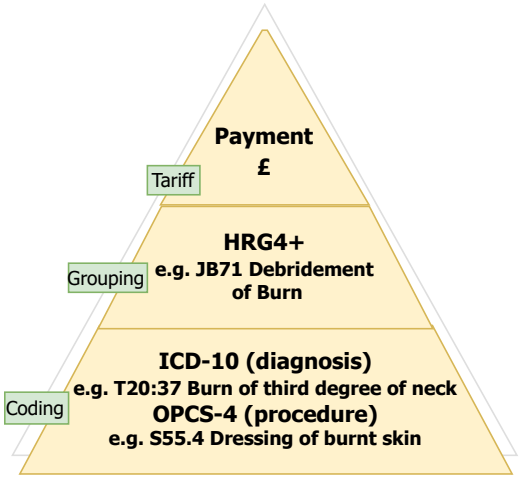

Figure 1: Building blocks for an activity-based payment system.

as an average of the cost of each HRG. This average cost calculation is, in essence, a generalisation that exacerbates the effect of any non-homogeneous groups. Therefore, there is a risk of inequitable and unfair reimbursement of funds when these HRGs are not sufficiently homogeneous. Our previous research found that using data-driven models improves the homogeneity of burn HRGs [7, 8]. Thus, our study uses burn services as a base case to re-develop one of the current payment system's building blocks - the Grouper. Here, we define a Grouper as the methodology developed to allocate a patient into suitable smaller groups.

Our proposed method aims to create a fair and equitable payment system by answering the following research questions.

1. Is there evidence of high variability in resource usage (LOS and cost) and patient injury severity (Total Burn Surface Area) in the current burn HRG groups?
2. Can the proposed Grouper be used to explore the trade-off between model complexity and the introduction of false incentives?
3. Drawing on established machine learning techniques, particularly the cost-sensitive Decision Tree, can we create a Grouper that provides higher homogeneity, meaning low variability in resource usage and patient injury severity?

Our study tackles these research questions by employing additional data on burn care, such as the inclusion of patient-level cost as an indicator of

resource usage. Previously, the direct consideration of patient-level costs for patient classification was not possible. This is due to the non-existence of an accepted costing methodology. However, the recent development of a suitable costing methodology [5] assists us in addressing this. We also adopt a data-driven approach to identify which information about patient characteristics, burn characteristics, and care provided can be employed to improve the Grouper whilst exploring the accuracy gap introduced in excluding potential false incentive contributors as independent variables. The data-driven approach was augmented with expert advice in deciding variables used in the target and feature space. Ultimately, our study aims to demonstrate how augmenting a data-driven approach with expert advice could improve the Grouper for a fairer financing system.

The remainder of this paper is arranged as follows. Section 2 provides the background. Section 3 describes the methods, including the study population and analytical pipeline adopted. In Section 4, the findings are introduced. Section 5 discusses the results of the adopted method, comparing these results with the current HRG. Finally, 6 concludes this paper and identifies the study's limitations and future directions.

2. Background

The NHS's activity-based system is tailored for different care types through HRGs. This tailoring is made of three building blocks, as described below.

1. *Coding:* This is the clinical coding of patient care using the Office of Population Censuses and Surveys (OPCS) procedure codes and the International Classification of Disease (ICD) to capture the diagnosis and interventions provided to a patient.
2. *Grouper:* The classifications generated using ICD and OPCS codes result in too many classes to form a practical basis for payment. The classes are further grouped into a currency unit known as HRGs. HRGs rely on expert advice to generate if-else rules on the identified OPCS and ICD codes and other patient-level data, such as age, length of stay, number of pre-existing conditions, and complications. These HRGs cover a spell of care from admission to discharge. HRGs can be referred to as currency. The term currency refers to the units of healthcare for which a payment is made.

3. *Tariff:* This is the fixed price for a given group to enable a payment system that does not transfer all cost burdens to the payer. The cost of treating a patient is typically calculated based on its group (HRG) membership average care cost. Though reimbursement is typically the average cost of all patients in an HRG, adjustments are made for unusually short or long stays and emergency use of specialised equipment. When adopted nationally, these average costs are known as Tariffs. They are further adjusted using the Market Forces Factor to account for varying costs in providing care in some parts of the country.

The overall aim of the current system is to ensure that resources needed for the comprehensive delivery of better care are available in all parts of the country. This approach, if implemented appropriately, will result in a more equitable and fair reimbursement system that encourages providers to compete on quality rather than price [9, 10]. This system would also act as an instrument to understand an organisation's activity regarding the type of patient seen and treatment delivered. Thus, it can be used to study the behavioural patterns within and between organisations, providing the opportunity to benchmark treatments and services. The described payment system avoids paying actual patient costs. As evidenced in [11], the optimal payment system should be neither fully prospective nor fully cost-based. The payment of actual patient costs transfers all cost burdens to the payer and, thus, insufficiently motivates profit-maximising providers to decrease costs. Therefore, the activity-based payment system is used to incentivise providers in maximising efficiency [12]. It supports patient, and commissioner choice [13] through the careful allocation of different forms of risk between commissioners and the care provider, where the commissioner bears the risk of unforeseen population growth and increased healthcare demands [14]. In comparison, the care provider bears the risk associated with patients using the service more than envisaged and potential failure to implement intended efficiencies.

Given the argument above that reimbursement systems are a tool used for influencing provider behaviour, in the same vein, these systems may influence the pattern of treatment provided. Marshall et al. [13] noted that all methods, including PbR, could have unintended adverse effects (known as false incentives). In particular, in [11, 15] it was noted that cost-based reimbursement could result in the false incentive of over providing services to all types of patients, whilst a prospective payment system exacerbates the incentive of

cost-based reimbursement for providers to compete in attracting low severity patients. Therefore, anticipation and mitigation through additional controls are needed to ensure the adopted payment system has its intended effect. PbR, a prospective payment system, is often blended with other methods and controls to mitigate adverse effects. Some of these controls include the provision of additional payment for extremely high-cost patients to protect the equity of access to care; auditing of patient codes to detect up-coding; the incentive to provide unwanted activity can be countered by paying lower (or even zero) prices for activity above a certain level, and the introduction of a pay-for-performance element to mitigate against risks of quality of care that may be faced when cost savings are sought. This paper explored the trade-off between model complexity and the inclusion of variables, such as cost and LOS, which may introduce increased false incentives.

2.1. Evolution of Health Resource Groups

Casemix as a discipline was invented in the 1970s to aid providers in understanding the type of care provided and the resource implications. This casemix can be defined as the hospitals' products requiring monitoring, budgeting, cost control, reimbursement and planning. The introduction of HRGs in the UK's NHS was motivated by the success of Diagnosis Related groups (DRGs) as a casemix system in the United States [16]. DRGs are a patient classification system of similar clinical diagnoses and procedures given.

The NHS, by 1994, had developed two variations of DRG called Health Resource Groups (HRG – Version 1 and 2, respectively). HRGs were aimed to be homogeneous regarding resource use and not the outcome, quality or appropriateness of care. The first adaption of DRGs to HRGs (Version 1) was undertaken by panels of clinicians for each diagnosis and published in 1992. Version 2 improved aspects of the methodology by providing an analysis of discharge-related data to professionally-led Clinical Working Groups (CWGs) [17]. The patient groups were subsequently defined based on the ICD, OPCS, age, and discharge status. Version 2 HRGs were found to explain a more significant proportion of the variation in the LOS of hospital inpatients than Version 1.

The process of refining Version 2 started in 1995. Again, professionally-led CWGs were convened and provided with the analysis of the LOS and volume of patients for each diagnosis (ICD) and procedure code (OPCS) within each HRG. A Reduction in Variance (RIV) analysis of all other variables in the

dataset was also undertaken for each HRG to determine which factors significantly impacted resource use. The refinement led to changes in accounting for the presence of multiple diagnoses, changes in ICD and OPCS coding, increasing the procedure hierarchy, and developing NHS-patient-relevant complications and comorbidities list instead of using those generated for the American DRG [18]. Due to the unavailability of patient-level cost, length of stay was utilised as a proxy for resource utilisation to test for group homogeneity.

Since their adoption, HRGs have been reviewed and enhanced annually to ensure that the classification keeps pace with clinical advancements. At the moment, HRG4+, an improvement on HRG4, is in use.

2.2. The Suitability of Health Resource Groups

The 2001 National Burn Care Review Report [6], a national strategy report on burn services, made over 140 recommendations, one of which is the need to create a burn service payment system that is both sustainable and accurate. This recommendation was founded on evidence that suggests the inaccuracy of the current HRG by failing to reflect the actual cost of caring for burn injuries. The inaccuracy starts from the clinical coding of burn injury incidence and treatment using ICD and OPCS, which does not accurately capture the heterogeneity in burn patients regarding their injury severity, comorbidities, and treatment requirements. These coding systems have been found to have severe limitations on their ability to adequately define acute patients, such as intensive and burn patients, especially those requiring surgery or those with complex burn injuries [19]. One other contributor to the inaccuracy is the non-inclusion of other elements of care, except for the acute care period, in the current HRG. This inaccuracy poses a risk of burn services being seriously underfunded due to the variability in the cost of individual treatment. Thus, adopting inappropriate HRG would have a substantial financial implication on burn services. To address existing inaccuracies in the coding system, we propose a holistic classifier for burn patients that accounts for all elements of care provided in burn services - outpatient, inpatient, rehabilitation, outreach, reconstructive surgery and support groups.

The most comprehensive research reviewing DRGs and their variants is called the EuroDRG. The research tests the DRGs' ability to explain hospital cost variations and LOS. The suitability of DRGs as a payment system was evaluated using ten different episodes of care as base cases across 10 European

countries, including England [20]. In most cases, DRGs with a high number of variables explain patient cost better. Performance measured on the ability of the DRGs to explain cost variations showed that England was one of the best-performing countries. As part of the EuroDRG research, the different DRGs (including the English HRG) in Europe were evaluated to examine how well they explain variations in costs or length of stay among patients [21, 3]. Both studies found that including patient characteristics and DRG labels as the independent variable increases the goodness of fit compared to a model with only DRG labels as the predictor of cost or LOS. This should not be the case if DRGs accurately reflect patient characteristics and has a strong relationship with cost. Specifically evaluating the Greek DRG, [22] found that including clinical severity in the classification improved group homogeneity in cost.

The EuroDRG, however, did not evaluate burn episodes of care, and so far, as previously stated, burn care providers (Trusts) have not adopted HRGs. Its non-adoption is primarily due to the complexities of categorising burn injuries into similar cost groups. Burn services have negotiated block contracts typically based on the historical cost of care, updated for cost inflation, efficiency savings, and innovation. This non-adoption is supported by research on HRG suitability for specialisms with similar characteristics to burn and, in some cases, research on burn care. Research on rehabilitation specialisms closely linked to burn services has highlighted some disadvantages of the casemix system [23]. This includes the lack of variables that reflect patient contexts, such as socioeconomic and mental state and family support.

Further research highlights the potential disadvantage for centres that provide complex care services by estimating the cost differentials associated with caring for patients that receive complex care and examining the extent to which complex care services are concentrated across hospitals and HRGs [24]. Using the 2013/2014 financial year data (reference cost), 6,797 burn patients were evaluated amongst other complex needs patients. The analysis found that certain hospitals would be penalised due to high-cost differentials related to burn care complexity. For burn HRGs, they found a 73 per cent higher cost for patients with complex needs than other patients in the same HRGs. For another hospital care setting, intensive care patients, the inappropriateness of HRGs to effectively classify patients due to the wide heterogeneity of patients within this diagnostic group was noted [19]. The research introduced the use of the Classification and Regression Tree (CART) [25] model to generate homogeneous groups in resource consumption. The

CART model's dependent variable was LOS, used as the measure of resource usage, and the independent variables included the reason for admission, age, and speciality of the referring team. This methodology leans away from diagnosis-based grouping, arguing that due to wide ranges of age, the severity of illness, comorbidities and diagnoses, a simple classification system that requires one predominant underlying condition will not account for the confounding factors. This study on grouping Intensive Care Unit (ICU) patients using a CART model successfully identified groups with a low degree of variation in LOS. In [26], a similar methodology was adopted for home-based care to reimburse the appropriate cost of care. The patient-level cost was used to measure resource usage and a CART model for grouping based on demographic, diagnosis, behavioural and care data.

A recent evaluation of the child burn HRGs compared to groups generated from an adjusted CART model found evidence of heterogeneity in child HRGs and showed increased homogeneity with the inclusion of patient-level costing in a data-driven methodology [8]. In evaluating the early application of DRGs in three English regions, it was noted that the absence of patient-level cost is a significant limitation in evaluating DRGs [27]. The evaluation also summarised the suitability criteria of any potential Grouper (i) it should have no technical issues (recording and coding patient data), (ii) it should be clinically acceptable (understanding the context of the group), and (iii) be statistically homogeneous in describing resource use (homogeneity requirement). These points motivate the adopted methodology of our paper. To meet the three suitability criteria whilst minimising the introduction of any unexpected false incentives, we propose a model that incorporates patient-level cost and includes other factors capturing burn injury severity, patient demography, socioeconomic status, mental status, pre-existing conditions, complications, and treatment (e.g. theatre visits).

3. Materials and Methods

Our research aims to develop a Grouper that generates burn patient groups that are explainable by minimizing heterogeneity in resource usage and patient injury severity. Our Grouper aims to use a data-driven approach to create homogeneous groups within the Burns Procedures and Disorders diagnosis group.

We aim to compare our generated groups to replicated HRG groups. This comparison is meaningful if the generated groups are of the same number as

the HRG groups and can be ranked similarly to HRGs, by injury severity (which we proxy with total burn surface area) and expected resource usage (proxied with LOS and cost). Thus the adopted approach deals with case labelling and segmentation using an unsupervised clustering method and then a decision tree to define these segments and identify the classification rules. The two-step data-driven approach is summarized below.

1. *Segmentation and Class Labelling:* A suitable Grouper aims to create groups that are homogeneous in Total Burn Surface Area (TBSA), cost and LOS but with the non-existence of meaningful ground truth (class label) that captures these three variables. The first step of our data-driven approach is engineering these class labels, ensuring the engineered classes are comparable to the current HRG. Thus, the chosen approach for engineering the target group should meet certain criteria. The ability to (i) control the number of groups, (ii) ensure the generation of approximately equal-sized groups, (iii) control homogeneity by ensuring the engineered target groups are biased towards resource usage and severity, and finally, (iv) ensure the generated groups are rankable by injury severity and expected resource usage. These considerations guided the adoption of k-means as the unsupervised segmentation model for the engineering of target groups.
2. *Classification Rules:* There is a need for data-driven models adopted in healthcare settings to be explainable (as opposed to being a black box), given their potential responsibility for human life. Thus, any adopted data model in healthcare should enable result tracing and transparency [28]. Furthermore, given that the present Grouper is rule-based and the health professionals are familiar with it, our paper adopts a decision tree model. A decision tree model allows for the production of meaningful and interpretable discrimination rules among clusters [29, 30]. In particular, we adopt a cost-sensitive decision tree model [31] to group patients. This supervised model allows the proposed Grouper to meet key considerations, including the ability to state the specific number of groups, ensure groupings reflect injury severity and resource usage and enable model explainability. Another key feature of the cost-sensitive decision tree is its ability to extract the key features needed for grouping. The cost-sensitive decision tree model allocates misclassification costs between groups to ensure that the predicted group is the same or in close proximity to the expected patient group. A cost-sensitive

decision tree model penalizes the model as the distance of the group predicted compared to the actual group increases [31]. This distance is determined by ranking groups based on the key variables. The average LOS, cost, and TBSA for each engineered target group are calculated to give the rank value, with the most severe being the group with the largest sum.

In summary, our data-driven Grouper can identify a rankable target space, has a penalization function that ensures little to no misclassification, and yields explainable groups. We shall refer to this data-driven Grouper as the DT Grouper and the identified patient groups as DT groups in the remainder of this paper. The DT Grouper is introduced in more detail in the following sections.

3.1. Study Population

This study uses anonymized patient-level data from all burn units in England. The data covers 2003 to 2019, comprising just over 50,000 admitted patients and was extracted from the international Burn Injury Database (iBID) [32]. The patient information is collected by clinicians and nurses, from the first contact with a burn service to rehabilitation and any late reconstruction procedure [33]. The dataset includes all types of burn and some non-burn injury patients that attended a burn service (e.g. patients experiencing skin loss not caused by a burn but needing treatment like burns patients). The dataset excludes burn injury patients that died before reaching a burn service and those with minor injuries cared for in the community and hospitals with no specialist burn services. With the focus on burn patients, we ensured patients with no TBSA or depth of burn recorded were removed as the absence of TBSA or depth of burn does not allow for grouping, as these are key features needed to identify a burn case. To ensure the right population of burn patients is analyzed, with guidance from the database administrators, patients with LOS greater than 360 or costs greater than £1m are considered extreme outliers and thus removed.

As highlighted by the 2001 National Burn Care Review Report, young patients are more likely to be susceptible to unforeseen complications, even with relatively basic burn injuries [6]. The report argues and mandates the need for distinct burn units for children and adults because of the further support needed by children from play specialists, instructors, family counsellors, and intense psycho-social care. The burn care pathway is therefore

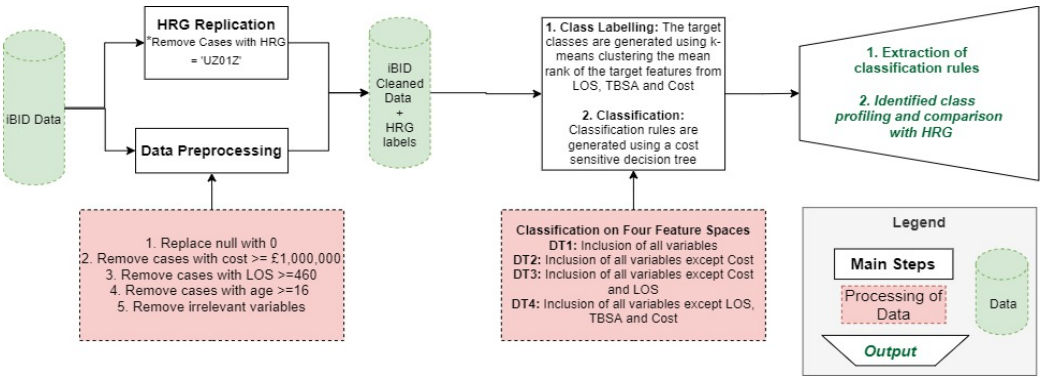

Figure 2: Analytical pipeline: data preprocessing, target and feature space identification, and model deployment.

made to treat children differently from adults. In a prior study on the segmentation of burn patients, further evidence was discovered that adult and paediatric burn patients have different resource requirements and are defined by different characteristics (such as the source of injury and presence of existing disorders) [7]. As a result, our research focuses on a specific population of burn patients, namely adult burn patients.

Our previous work considered the child burn population as a study population [8]. We developed and tested a data-driven Grouper against the current HRG Grouper with promising results [8]. This study builds on our previous work by using a different study population and includes methodological changes that reduce complexity. The analytical pipeline in the previous study relies on the individual LOS, TBSA and cost decision tree model to identify variables of importance; in contrast, this work includes all relevant variables, relying on one decision tree model to identify the feature space. In this study, we also compare multiple DT Groupers to identify the accuracy gap of excluding one or all of the three key variables - LOS, TBSA and cost.

3.2. Variables

This study aims to develop a DT Grouper by leveraging access to a wide range of information about burn patients admitted to any of the 23 burn services in England. The data includes patient-level costing, comorbidities, demographic and burn characteristics (see Table A.1 in the Appendix). These variables are collected on admission, during treatment (from patient records or linked data), or calculated on discharge. However, a key variable needed

for model comparison is the current HRG label of each burn patient, which is not recorded in the iBID database. Thus, the current HRG4+ label was added by replicating the methodology used by the NHS Digital Casemix Team (see Appendix B).

Direct consideration of patient-level costs for our model development is one of the significant contributions of this paper. The potential of this approach was pointed out in the NHS Costing Manual [34] but has not been implemented previously due to the complexity of developing a costing methodology that ensures the inclusion and appropriate apportionment of all elements contributing to the total cost of care. Here, we overcome this challenge by using the iBID patient-level costing methodology, developed to calculate the cost of care for all iBID recorded burn patients [5]. Patient-level cost is leveraged in this paper to identify the target and feature space - the independent variables - and check model suitability.

Prior to the model creation, irrelevant variables — administrative variables, variables with a large proportion of missing values, variables with unstructured data, variables with one unique value, and variables that are duplicates — were removed from the dataset. Also, all null values were replaced by 0. The assumption is that fields are left empty when the value is 0, no or not applicable (also adopted by the current HRG methodology) [35].

3.2.1. Class Labelling and the Identification of the Feature Space

The non-existence of class labels that captures TBSA, LOS and cost necessitate identifying the appropriate target space (class labels) for training the model and learning the classification rules. This need to generate class labels that capture TBSA, LOS and cost is because we need to describe the segmentation based on these variables. These labels were used as the three key factors in burn care, ensuring that the segments meet the homogeneity requirement (see Section 2). These factors were identified from previous analysis, expert advice and the literature and represent the aim of HRGs to

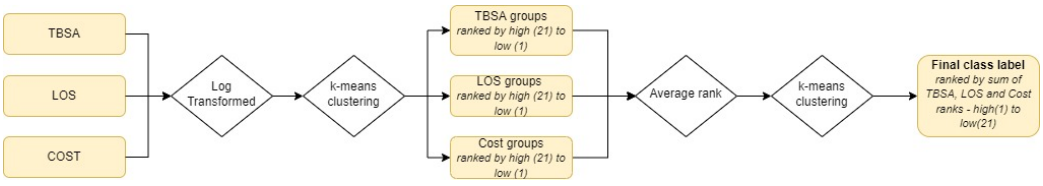

Figure 3: Summary of the class label creation process.

create clinically meaningful groups (measured by burn TBSA, a proxy for injury severity) that have similar resource usage (measured by LOS and cost of treatment).

The class label creation is summarised in Figure 3. The first step is the log transformation of the three factors to reduce skew, followed by k-means cluster analysis to segment each factor into 21 target groups. k-means clustering was selected over other potential classification methods, such as percentiles (equal spread), quantiles (equal range) or other clustering methods. Although using percentiles to assign group membership leads to a balanced sample size, it will not maximize homogeneity. The quantile method leads to highly imbalanced groups. On the contrary, k-means assigns patients to a group so that the sum of the squared distances between them and the cluster's centroid is as small as possible, thus ensuring high homogeneity. The identified groups were ranked from most severe/costly to least severe/cost to incorporate a cost-sensitive model.

Consistent with identifying the LOS, TBSA and cost group labels, the final class label was derived using k-means cluster analysis. The feature space for the k-means cluster analysis was computed by calculating the average rank across all LOS, TBSA and cost group labels. Adopting the average rank of each case allows identifying groups that do not overly focus on extremes of each factor but instead reflect a balance (average) across TBSA, cost and LOS.

As the adopted model's important characteristic is extracting key features and the classification rules, our analytical pipeline includes a decision tree model. This decision tree model uses all variables capturing burn injury severity, patient demography, socioeconomic status, mental status, pre-existing conditions, complications, and treatment as its feature space. However, to identify the accuracy gap of excluding one or all of the three key variables used in target space creation, a variation of the decision tree models (DT Models) is compared against the base - HRG - group.

1. DT1: Inclusion of all variables
2. DT2: Inclusion of all variables except cost
3. DT3: Inclusion of all variables except cost and LOS
4. DT4: Inclusion of all variables except LOS, TBSA and cost

The exploration of the feature space through the exclusion or inclusion of variables such as LOS and cost, which could be altered to gain higher reimbursement, i.e. false incentives to upcode, lends us the opportunity to

understand the accuracy gaps of excluding these variables. It also allows comparison with the HRG model, which does not include LOS (except for identifying the patient group for those who died or were discharged too early) and patient-level cost in its decision rules. In summary, burn HRGs are created using variables on the depth of burn, the body site of the burn, discharge destination, type of injury, source of injury, age, intubation, inhalation and presence of complications and comorbidities.

3.2.2. Replicating the current HRG

A patient's HRG is generated using Health Episode Statistics (HES), a database different from the iBID database and thus, the need to replicate the HRG Grouper to allow for the evaluation of the DT Grouper. This is due to non-access to the HRG labels in the iBID dataset and the inability (due to data privacy restrictions) to merge the iBID dataset to the HES database containing these labels. Thus, our analytical pipeline includes replicating HRG4+ using variables in the iBID database. Guidance from documentations on burn HRG and working collaboratively with NHS digital casemix experts to understand the grouping logic allowed us to perform this replication. The HRG methodology can be summarised as follows.

1. Divide patients into clinically meaningful chapters such as Chapter J for skin, breast and burns and sub-chapters such as JB representing burns. The chapter is determined by the dominant procedure, diagnosis, treatment or investigation code recorded for a patient using ICD and OPCS coding. The JB sub-chapter is generated if a burn or corrosion (ICD-10 rubrics T20-T32) diagnosis code is recorded for a patient.
2. Divide sub-chapters into smaller groups that reflect age, burn depth and severity.
3. Divide groups further to reflect the comorbidities and complications associated with the patients.

The reliability of this replication was evaluated by reviewing the characteristics of patients in each group against the description summarised in Table B.1. For example, the JB40A group profile includes adults with a total number of grafting operations greater or equal to 2, older patients with the highest escalation numbers for invasive ventilation, complication and comorbidities and face, hand and feet site of the burn. The replication is summarised in Figure 4 and detailed in Appendix B.

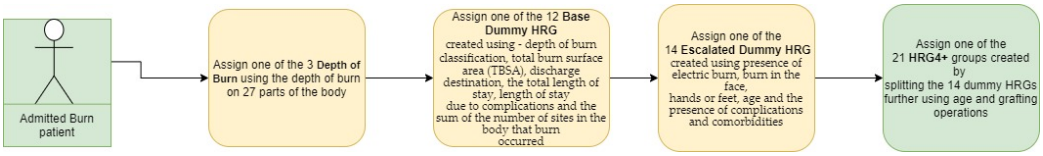

Figure 4: Summary of the HRG replication process.

3.3. Deployment and Evaluation of DT Grouper

This section describes the developed DT Grouper and introduces the contestant HRG Grouper. The 21 target groups for the decision tree models were engineered and identified using an unsupervised k-means algorithm as described in Section 3.2.1. The categorical independent features (13) were one-hot encoded, increasing the variable count from 240 to 462. The dataset is then randomly split into a train (75 per cent) set used for two-fold cross validation and a test (25 per cent) set.

The 21 groups of the train set are oversampled independently such that the sample size of the minority groups matches the majority, and at the end, all groups in the train set have the same number of samples. Table **??** shows, for the train set, the ratio of the majority group to all other groups and thus the amount of increase required. The ratio of the majority group to other groups in the test set matches that found in the train set.

The oversampling was implemented using the Synthetic Minority Oversampling Technique (SMOTE) [36]. The oversampling is needed to ensure the adopted model has enough samples to learn the decision boundary between all groups effectively.

The cost-sensitive decision tree model was implemented with the `mlr` [37] and `rpart` [38] packages in R. The classifier was customized to include our defined cost matrix as a measure such that the predicted probabilities for each class are adjusted by dividing them by the corresponding threshold value. Then the class with the highest adjusted probability is predicted. This class-dependent cost depends on the true and predicted class labels, as represented by the matrix shown in Figure 5 [39]. The misclassification cost

Table 1: Oversampling: Ratio of the majority group to other groups on the training data.

Group	1	2	3	4[a]	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
Ratio[b]	2.3	1.3	1.5	1	1.6	1.7	1.3	1.9	2.0	1.4	2.2	2.4	1.8	1.9	3.3	2.8	3.8	6.8	5.0	7.6	13.6
Size	1044	1926	1617	2447	1517	1406	1894	1287	1239	1745	1106	1029	1386	1257	749	866	651	360	489	322	180

[a] Group 4 is the majority group.
[b] The ratios represent the extent each group was increased to match the majority group.

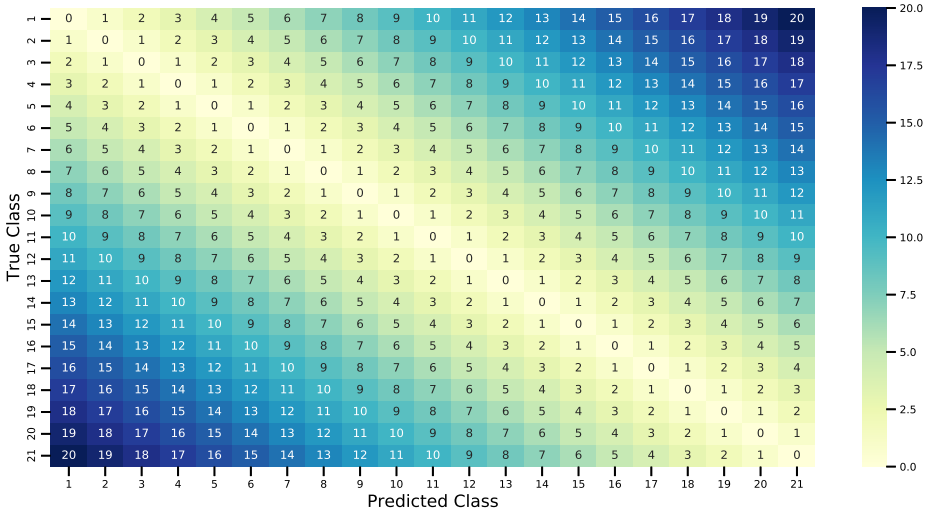

Figure 5: Cost matrix representing the relative penalty used in the cost-sensitive decision tree to ensure any misclassifications are close to the true group.

is relative, where the cost of predicting the correct class label is minimal (in this paper, zero), with increasing cost as the predicted class's proximity to the true class increases.

The values of the parameters, such as the maximum depth, minimum bucket, and complexity parameter, were set using a grid search to ensure that the model accurately learned the data. Two-fold cross validation of the train data (selected due to the relatively large size of the train dataset) was used to find the best value for key parameters. This includes the tree's maximum depth, minimum bucket, and complexity parameters with a search range of 5 to 10, 2 to 5 and 0.001 to 0.01, respectively. These parameters define the complexity of the classification model by defining the longest path (depth of decision tree), the minimum number of observations in each terminal node (leaf size), and providing a bound on tree construction. The identified optimal parameter values are 10, 4 and 0.001, respectively.

The model's output was evaluated using first a confusion matrix to evaluate the errors made by the decision tree model and the proximity of misclassification if any. Then to review the explainability and suitability of the adopted model, we present the decision tree rules generated to classify patients. Another evaluation step was comparing the groups generated from our DT Grouper and the existing HRG Grouper. This was done by visually

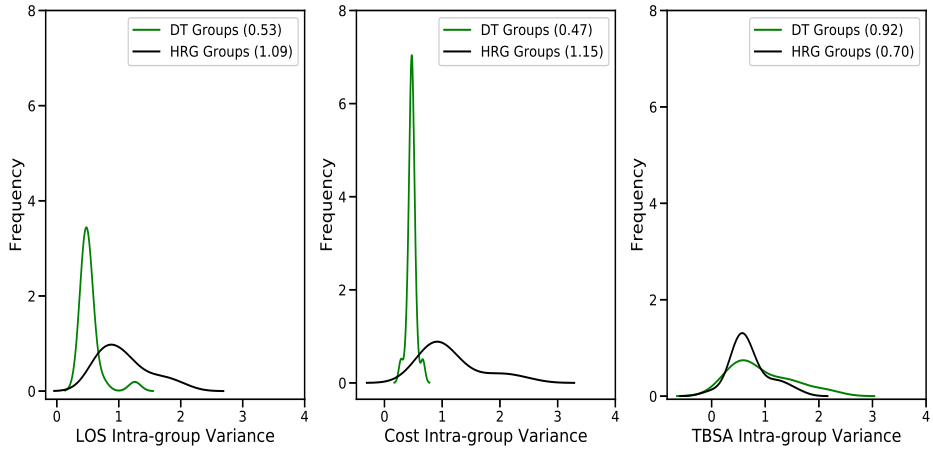

Figure 6: LOS, cost and TBSA intra-group variance of the HRG and DT Grouper (on the training data). The value in parenthesis in each legend is the average cost, LOS and TBSA intra-group variance for the DT and HRG groups.

evaluating the distribution of patients in each group by the three key factors — LOS, cost and TBSA. In addition, intra-group variances were calculated to better understand the spread of these critical factors in the new groups compared to HRGs.

4. Results

This section assesses the results of the data-driven model discussed in Section 3.

4.1. Homogeneity in HRG Compared to Proposed DT Grouper

As noted in Section 3, k-means and a cost-sensitive decision tree algorithm were adopted to develop the DT Grouper. The decision tree labels were created using three engineered target groups (TBSA, LOS and cost groups as shown in Figure 3) that reflect the key factors of interest. From Figure **??**, it can be observed that the engineering process described in Section 3 generates target groups with a decreasing LOS, cost and TBSA as injury severity (reflected by the final decision tree group) decreases. Thus, the final class labels identified are suitable target spaces for DT1, DT2, DT3 and DT4.

We consider the homogeneity of groups as a key performance indicator of a reliable Grouper. Mathematically, we capture this using intra-group

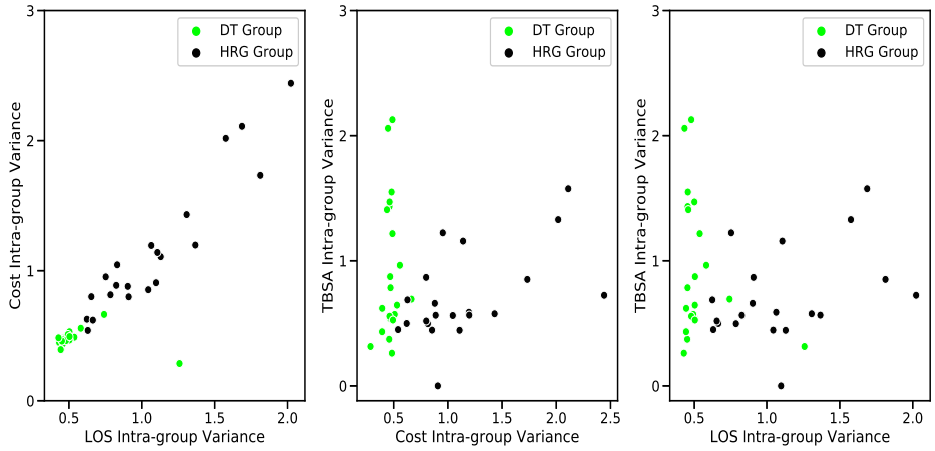

Figure 7: DT and HRG pairwise intra-group variances (on the training data). The scatter points show the value of LOS, cost and TBSA intra-group variances in relation to each other.

variance with respect to LOS, cost and TBSA, respectively. The DT groups generated from the DT Grouper are thus compared to the current HRG to evaluate the model's performance using the three key factors. Figure 6 shows within-group variation across patients in all 21 groups (both HRG and DT group) with respect to these three measures.

Our results show higher homogeneity for the DT groups compared to the HRG groups using LOS and cost, as seen in Figure 6. The cost and LOS intra-group variance is significantly lower in the DT groups than in the HRG groups. The narrower width (narrow range of values) and greater height (maximum frequency) observed for the DT groups' LOS and cost intra-group variance plots compared to the contestant HRG groups indicate that most DT groups have low intra-group variances, as depicted by the average values. We discovered that the cost intra-group variance (0.47) for DT groups is two times lower than the HRG groups' cost intra-group variance (1.15). The intra-group variance in LOS follows this pattern, with 0.53 for DT groups and 1.09 for HRG groups. When we compare the average injury severity (TBSA) intra-group variance of DT groups at 0.92 to that of HRG groups at 0.7, we observe that the HRG model performs better. This reflects the HRG's over-reliance on TBSA as the significant group allocator comes at the expense of missing other important factors. This trade-off reduces cost homogeneity in HRGs compared to the proposed DT group. However, it backs up our

strategy of identifying target groups based on three key factors rather than just one. In Figure 7, we show pairwise intra-group variances to understand the gain of non-reliance on TBSA as the most significant group allocator. When comparing the DT groups to the HRG groups (see Figure 7), we can see that where TBSA intra-group variance is high (depicting heterogeneity), homogeneity in LOS and cost is higher (lower intra-group variance).

In summary, these findings show that the engineered target groups have greater intra-group similarity than HRGs in at least two aspects. As a result, these target groups have been adopted for the remainder of the study.

4.2. Evaluation of the DT Grouper's Accuracy Gap and Homogeneity

The analysis of the train data suggests that the engineered target space meets the suitability criteria defined in [27] and summarised in Section 2.2. With the target space identified and established, we want to evaluate the performance of the four DT models compared to the HRG model using unseen data. This would allow us to compare the DT Grouper with the HRG and understand the accuracy gap among the four DT models due to the presence/absence of the key variables in the feature space. In addition, we can identify and compare the variables of importance in assigning patients to groups for each DT model. In particular, the evaluation of test data is needed to understand the model's performance on unseen patients. Thus, to check the ability of the DT Grouper to generalise, we analyse the test data with respect to resource usage (LOS and cost) and injury severity (TBSA).

When the classification rate of unseen patients (test data) is visualised using the confusion matrix, we see varying performance for the four DT models. For DT1, as seen in Figure 8, with the non-exclusion of any of the key variables, our proposed DT Grouper can allocate each case into one of the 21 groups with high accuracy. The confusion matrix reveals that the true group of the unseen patients is predicted for most patients (highest probabilities on the diagonal). Moreover, where the classification is wrong, the misclassified group is in close proximity to the true group. For most groups, misclassification is within two adjacent groups from the true group (i.e. we are within the same complexity regime). A similar pattern is seen in Figure 9, which shows the classification rate for DT2 - the DT model that excludes only cost. The highest probabilities are diagonal, though these are lower than that seen for DT1. Where misclassification occurs, these are closer to the true group than in DT1. For a model with the exclusion of the two resource-related variables - LOS and cost, as depicted in Figure 10, there is a

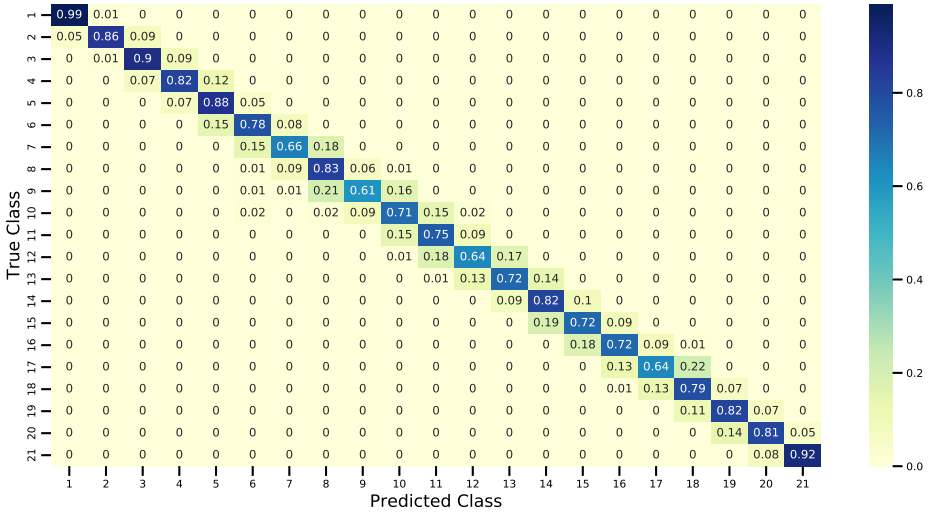

Figure 8: DT1 Confusion matrix represents penalisation's impact on the DT Groups, using the test data and all variables.

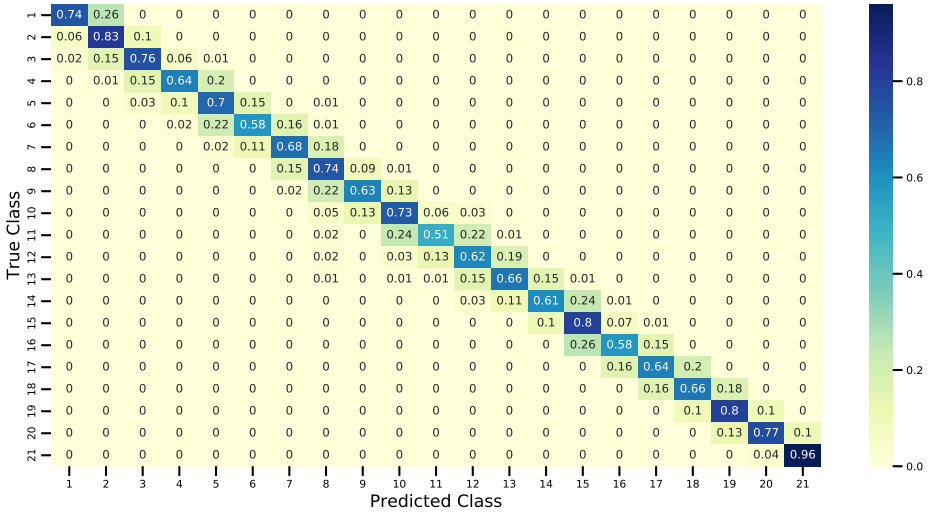

Figure 9: DT2 Confusion matrix represents penalisation's impact on the DT Groups, using the test data and excludes cost.

slight deterioration in the classification accuracy compared to DT1 and DT2. We see misclassification further away from the true class but no probability equal to 0 with a minimum of 11 per cent. In contrast, as depicted in Figure 11, the exclusion of the three key variables in DT4 hinders the model from

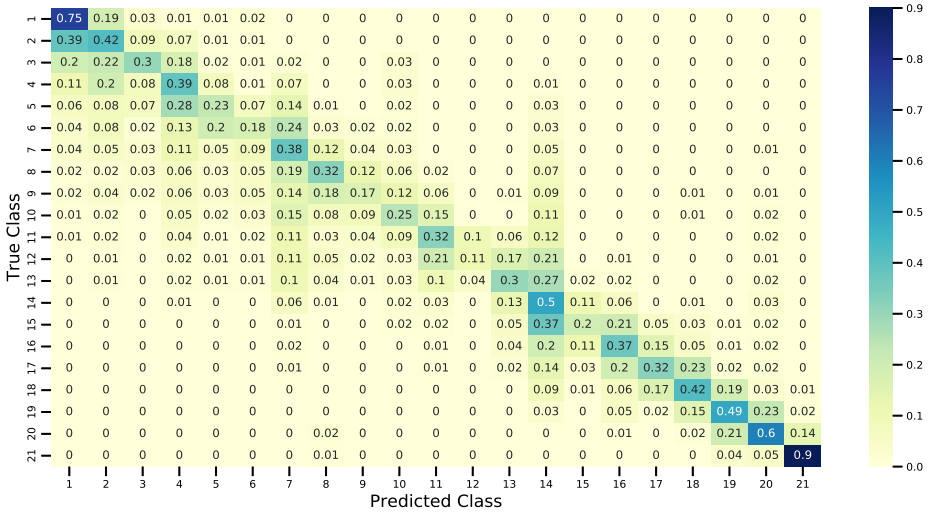

Figure 10: DT3 Confusion matrix represents penalisation's impact on the DT Groups, using the test data excludes LOS and cost.

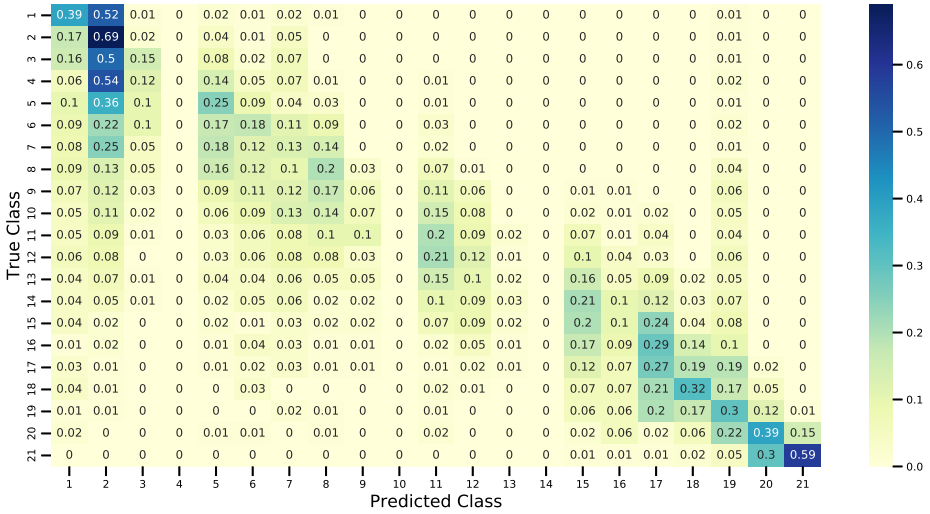

Figure 11: DT4 Confusion matrix represents penalisation's impact on the DT Groups, using the test data excludes LOS, TBSA and cost.

allocating cases into all 21 groups. Figure 11 shows that there are four groups with no case allocation, and the highest probabilities are not always found in the diagonals. Where misclassification occurs, it is not often to classes close

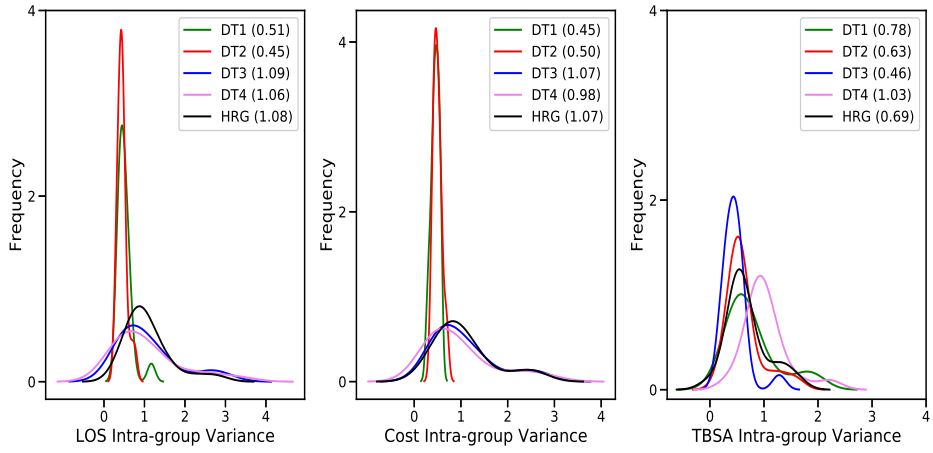

Figure 12: DT1, DT2, DT3, DT4 and HRG intra-group variance for LOS, cost, and TBSA, on test data. The value on each legend is the average LOS, cost and TBSA intra-group variance for the DT and HRG groups.

to the true class.

Finally, the results shown in the confusion matrices confirm that oversampling did not affect performance because there is no correlation between classes needing low/high oversampling (as shown in Table 1) versus prediction accuracy shown in Figures 8 to 11.

To further understand the differences in performance across the four DT models, we evaluate the homogeneity of the identified DT models compared to the current HRG model. The intra-group variance of the test data, as shown in Figure 12, has a similar shape to that seen in the train data. The short width and longer height observed in the DT1 and DT2's cost and LOS intra-group variance plots are similar to that observed in the train data. On average, a lower intra-group variance in the test data is observed compared to the training data. This is an artefact of the oversampling approach, which changes relative group sizes in the training population. Variance values are, therefore, not directly comparable across training and test data but are comparable across methods. Comparing intra-group variance across the five models, we see varying results. DT1 and DT2 have significantly lower average LOS and cost intra-group variances compared to the DT3, DT4 and the HRG model. For TBSA intra-group variances, DT4 is the worst performer, with an average of 1.02 compared to 0.46 for DT3 and 0.7 for the other three models.

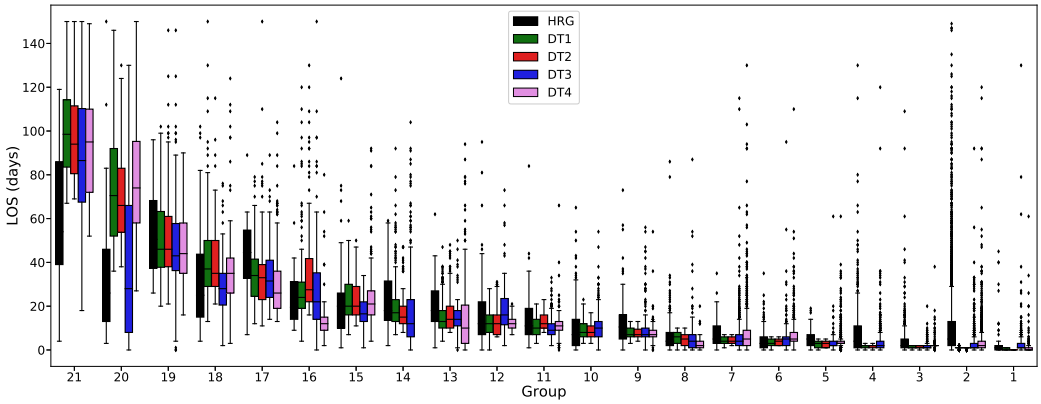

Figure 13: DT and HRG groups by LOS on test data. Ordered in decreasing order of injury severity.

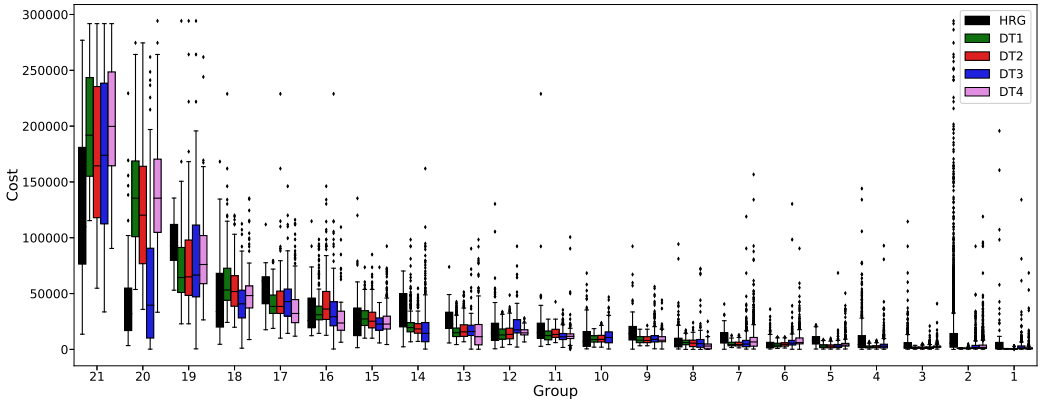

Figure 14: DT and HRG groups by COST on test data. Ordered in decreasing order of injury severity.

To further evaluate the difference in the homogeneity and separation of the HRG and DT groups on cost, LOS and TBSA, we show their distribution across all groups using the test data. This is illustrated in Figures 13 to 15. The DT groups are ordered by expected decreasing group complexity, i.e. ranked by the total of average TBSA, average LOS and average cost. Similarly, the HRG groups are ordered by known group severity (as summarised in Table B.1 in the Appendix). For the DT Grouper, as previously described, the groups were engineered such that it is expected that group 21 is the most severe case with high resource usage. For the HRG groups, these were created to inherently depict an order of injury severity (see Table B.1).

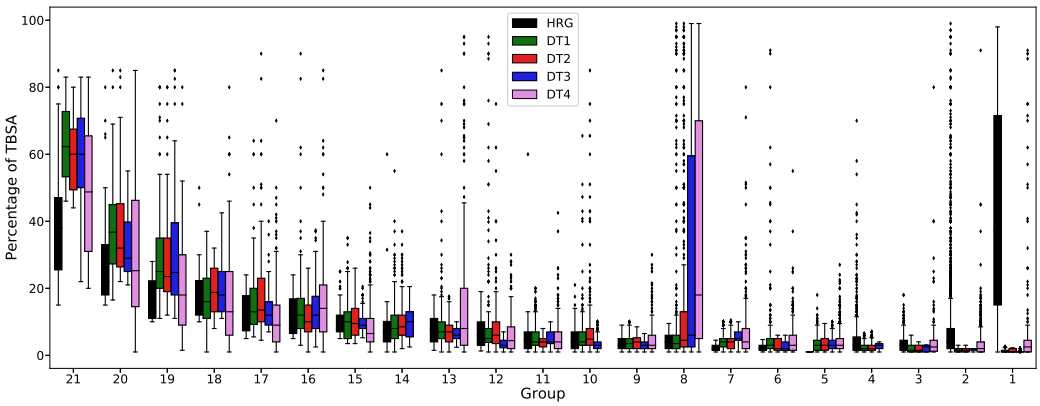

Figure 15: DT and HRG groups by TBSA on test data. Ordered in decreasing order of injury severity.

From Figures 13 and 14, we observe that for the DT groups, there is a steep descent of the box plots except for DT4, which sees slightly wider boxplots (higher variances) in group 13 and 7 due to 100 per cent misclassification in group 4, 10, 13 and 14. These wider boxplots are more pronounced for LOS compared to cost. For DT1 and DT2 specifically, we see a reduction in the average LOS and cost values as the complexity of each group reduces. The HRG groups, also ordered in decreasing order of injury severity, do not have a steep descent for both cost and LOS boxplots indicating a lower relationship between injury severity to LOS and cost.

Reviewing TBSA, wherein group 1, we see a very wide box for the HRG group; this is significantly smaller for the DT group. For DT1 and DT2, group 1, as expected, is the least resource-intensive group (in terms of LOS and cost) and has the least severe (TBSA) patients. Whilst for HRG group 1 (known as JB61A and described as the treatment of burn where patient transferred or died in 2 days or less - see Table B.1 in the Appendix), we see a wide spread of TBSA, low cost and LOS. The behaviour of HRG group 1 is expected, as patients with higher injury severity are more likely to die and spend less time in the hospital. Similar to HRG, DT3 and DT4's group 8 has a very wide box plot. The cluster profile indicates 98 per cent mortality, the highest average TBSA, lowest average cost and lowest LOS (minimum of 0 and maximum of 9).

Thus, unlike the DT1 and DT2, which should show a dramatic decline in LOS, cost, and TBSA as group complexity decreases, the HRG, DT4, and

to some extent, DT3 groups do not follow this trend. They have a lower degree of group separation. Also noted is a trend of narrower boxplots for DT groups compared to the HRG groups indicating higher homogeneity.

4.3. Profiles Identified by the DT Grouper: Grouping Rules

We now explore how the DT Grouper meets the suitability criteria for creating groups that can be explained. This is achieved by reviewing the classification rules of the four DT models. The classification tree depicts the criteria for categorising a case into a specific group and shows that the adopted model produces an explainable output.

For DT1, of the 243 variables included in the model, only three variables - the same as that used in creating the target space - were used for classification: LOS, cost and TBSA. Reviewing the predicted groups, as expected, we see that group 21 is made up of any patients with a cost greater than £99,000 and LOS greater than 61 and TBSA greater than 45, or those with LOS greater than 13 and TBSA greater than 35, or finally those with LOS greater than 130.9 and TBSA greater than 32.9. At the other extreme, group 1 are those with LOS less than 2 days and TBSA less than 1.5 or cost less than £581. Alternatively, LOS of less than 2 days and TBSA less than 3 and cost less than £543.

For DT2 of the 242 variables inputted into the model, only three variables were used for the classification: LOS, TBSA, and the total number of theatre visits. Reviewing the predicted groups, as expected, we see that group 21 is made up of the most severe patients: those with LOS greater than 39.9 and TBSA greater than 76.9, or those with LOS greater than 68 and TBSA greater than 44, or finally those with LOS greater than 131 and TBSA greater than 22. Conversely, group 1 has the least severe patients with LOS of less than 1, TBSA of less than 2 and total theatre visits less than 1.

For DT3, of all the 241 variables inputted into the model, 14 variables were used for the classification: bed ward days, TBSA, first operation after admission, the last operation before discharge, total theatre visits, age, presence of burn in the anterior chest, no existing disorders, the sum of escalation, source of injury (Bathing), Days from admission to the last operation, death, discharge destination, LOS in ICU. Reviewing the predicted groups, group 21 is made up of the most severe patients: those with LOS greater than 44 and days from admission to last operation greater than 45 or those with the probability of death greater than 0.64 and TBSA greater than 60 or those with TBSA less than 60 and last operation before discharge greater

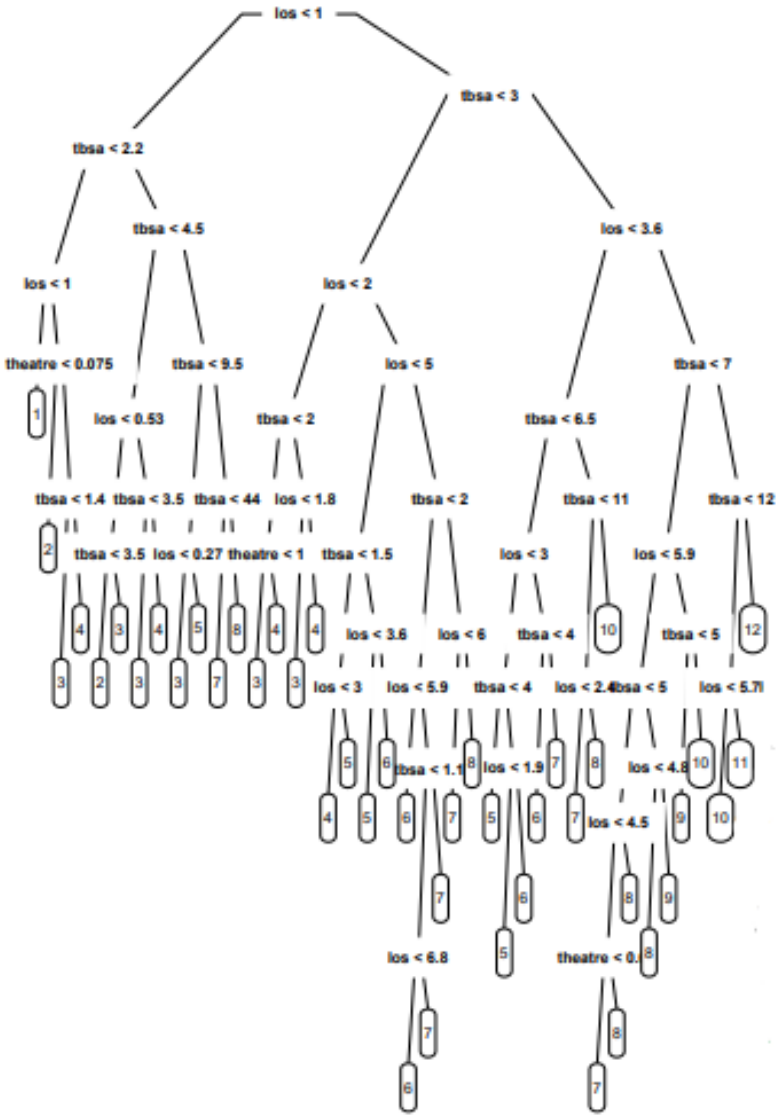

Figure 16: Sub-section of the classification tree showing explainable classification rules generated from DT2 using TBSA, total theatre visits and LOS.

than 49 or those with TBSA less than 60 and last operation before discharge less than 49 and discharge destination is unknown or those with days from admission to last operation greater than 32, TBSA greater than 21 and LOS in ICU greater than 46.

For DT4, of all the 240 variables inputted into the model, 21 (including six variables used in the creation of the sum of escalation - an input variable in the HRG Grouper) were used for the classification: total theatre visits, last operation before discharge, bed ward days, the sum of escalation, right upper arm, first operation after admission, depth at the right thigh, intubated, LOS in ICU, discharge destination (Home), comfort care only, presence of burn in the anterior chest, max org support, left shoulder depth, locality during the burn (street or roadway), activity during the burn (amusement or entertainment). Reviewing the predicted groups, as expected, we see that group 21 comprises the most severe patients: those with admission last operation greater than 44 and LOS in ICU greater than 22 or left shoulder greater than 1. Or those with admission last operation greater than 44, LOS in ICU greater than 6 and last operation before discharge greater than 66. Alternatively, those with admission last operation greater than 44 and LOS in ICU greater than 6, last operation before discharge less than 66, left thigh depth greater than 4.7 and probability of activity during burn being amusement or entertainment greater than 0.047. Alternatively, those with admission last operation greater than 44 and LOS in ICU greater than 6, last operation before discharge less than 66, left thigh depth less than 4.7 and locality during burn being in the street or roadway greater than 0.43. The other group of interest - group 8 - have patients with the last operation before discharge less than 2, bed ward days less than 5 and right upper arm greater than 0.16. Alternatively, it is bed ward days less than 0, the sum of escalation greater than 1, discharge destination is not home, and comfort care less than 0.55.

In contrast, the HRG Grouper uses 41 variables to determine the group membership of a patient. Regardless of this highlighted complexity in the HRG Grouper, the DT1 and DT2 perform better in terms of overall homogeneity.

5. Discussion

Due to the diversity and complexity of hospital care, creating reimbursement systems that ensure that the amount and quality of hospital treatment meet public needs while being cost-effective is difficult. The difficulty in designing a payment system is exacerbated by the additional need to ensure the minimisation of false incentives. Our research evaluated the suitability of the current reimbursement system in burn care whilst attempting to create

a more suitable reimbursement system. Currently, the NHS has adopted a system that generates patient groups using expert-generated decision rules as the basis of the reimbursement system. Our research presented multiple data-driven patient Groupers as an alternative patient Grouper. These were compared with the status quo. In particular, we examined the usefulness of expert-generated decision rules - HRG Grouper - in identifying groups of patients who are clinically similar and expend similar resources. Additionally, we evaluated the performance of the multiple DT Groupers to understand the accuracy gap introduced in excluding potential false incentive contributors as independent variables. The inclusion of certain variables in the Grouper, such as LOS and cost, which can be manipulated (directly or indirectly), may be gamed by care providers in increasing reimbursement amounts [11, 15].

For any adopted Grouper to be fit for purpose, it should create clinically acceptable groups with easy to understand the context and be statistically homogeneous in describing resource usage [27]. The Grouper should use routinely collected data and generate a manageable number of groups. Our proposed DT models were checked against these criteria and compared to the HRG.

We illustrate how LOS, TBSA and patient-level cost can be used to provide a more comprehensive definition of patient homogeneity. We then illustrate that if all variables, including LOS, TBSA and patient-level cost, are included in the feature space, these three are chosen as the significant variables. With the exclusion of cost (in line with the current HRG and as an attempt to minimise false incentives) from the feature space, only three variables, LOS, TBSA, and the total number of theatre visits, are needed to effectively group patients into homogeneous groups. On the other end, the exclusion of all three - LOS, TBSA and patient-level cost - leads to a more complex DT Grouper, which relies on 21 variables for the classification of patients.

Amongst the four DT models presented, DT2 is chosen as the DT Grouper for further comparison with the HRG. DT2 excludes cost in the feature space and still generates performance higher than the HRG, DT3 and DT4, and is often on par with DT1. DT1, though, with higher performance in terms of lower misclassification rate and higher homogeneity, increases the risk of introducing false incentives where Trusts could increase resources used in treating patients to enable the allocation of higher funds. On the other hand, DT4, excluding all three key variables, has a higher complexity, reduced homogeneity and increased misclassification. However, it could be argued that

including LOS in the feature space of DT2 could increase the risk of introducing false incentives, with LOS exclusion in DT3 leading to a performance on par with HRG or less. To increase reimbursement received, care providers may over-provide services, such as increasing days spent on admission, without necessarily adding any benefits to the patient. Nevertheless, we adopt DT2 given that LOS is included in the current HRG and LOS is not wholly influenced by care providers.

With our proposed Grouper's - DT2 - reliance on only three variables, the complexity of translating this Grouper into practice is minimised as the model relies on a smaller subset of data compared to the HRG, which relies on about 41 variables. Another advantage over the current Grouper is the reduced probability of missing variables. The HRG Grouper uses a higher number of variables, necessitating the creation of a 'catch all' group in which all patients with insufficient data for grouping are allocated and assigned a tariff of £0. In the case of potential missing variables in our three key variables, we advise adopting an imputation model [40]. A full evaluation that allows missing variables to be imputed using data from similar patients and, thus, potentially eliminating the requirement for a catch-all group is outside the scope of our study.

The identification of LOS and TBSA as the classification variables is very much in line with the current HRG Grouper. In the HRG Grouper, the creation of the first level groups (known as the dummy groups) mainly uses TBSA and, in some cases, LOS, depth of burn and discharge destination (see Figure B.2 in the Appendix). The proposed Grouper creates clinically similar groups that use similar levels of healthcare resources while being less complex than the HRG Grouper. Although the HRG Grouper yields clinically similar groups, as seen in Section 4, it performs significantly worse than the proposed Grouper in group homogeneity in LOS and cost of care.

The classification tree associated with the DT Grouper demonstrates that the created groups are simple to rank and characterise based on each group's injury severity and resource utilisation. This meets the suitability criteria related to creating clinically similar groups that use similar healthcare costs and thus allows the identified DT groups to serve one of the primary purposes of patient groups, providing the basis for comparative assessment [41]. The ability to characterise groups by cost and injury severity will support the assessment of each group of burn patients, its average cost and how it compares to the national average to identify outlier patients. Subsequently, this will allow clinical commissioners to use the actual patient-level cost to identify

which care providers, if any, over or underspend when treating patients. Results from the comparative assessment can be used to design policies to help ensure consistency in the quality of care provided and penalise overspenders.

The main purpose of HRGs is to inform the contracting process that assists clinical commissioners in the equitable and fair reimbursement to Trusts for the cost of providing care [41]. Thus, these groups are generated after care is delivered and can be developed using all variables collected during care. It also allows Trusts to carry out internal resource management, where the generated patient groups are used to examine the financial impact of projected changes in future patient volume and track actual vs expected spending. However, the use of patient groups for reimbursement purposes has not been adopted in burn care due to the current HRG groups' inability to reflect the cost of providing burn care. One of the key contributions of this study is incorporating patient-level cost in the model development and suitability check. This has not been previously done due to the non-availability of patient-level costs in the NHS. Resource usage homogeneity checks have been done using the length of stay and (or) reference costs. While reference cost is a top-down costing approach, patient-level costing amalgamates the top-down and bottom-up costing approaches. Patient-level costing enables detailed itemised costing of staff, equipment, drugs, consumables, and maintenance to be traced back to specific care areas [5]. Thus, enabling detailed budgetary analysis of the cost of operating a Trust. Our paper includes patient-level costs developed using an internal iBID patient-level costing methodology. In Section 3.2.1 we described how this is used to engineer target groups, and it is also used in Section 4 to evaluate the cost homogeneity of both the HRG and DT groups. The proposed data-driven DT Grouper results in higher homogeneity in patient-level cost than the current HRG Grouper. In Figure 12, we show two fold increase in cost intra-group variance when comparing the proposed DT groups to the current HRG groups, lending more evidence to the known inability of the current HRG groups to reflect cost accurately. These results also show that our proposed Grouper will aid in the equitable and fairer cost reimbursement to Trusts.

6. Conclusion

In this study, we evaluated the homogeneity of the current HRG on burn patients on resource usage and severity and compared this to our proposed DT Grouper. In addition, the accuracy gap generated by removing false

incentive contributors was analysed to understand our ability to create a DT Grouper that uses established machine learning techniques augmented with expert advice.

Our results show that the proposed DT Grouper (DT2) provides higher homogeneity in LOS, cost and TBSA compared to the current Grouper whilst minimising false incentive by excluding cost in the feature space. The proposed DT Grouper uses three key variables to effectively group burn patients, thus reducing model complexity compared to the HRG. A preliminary version of the proposed model presented and evaluated in this paper was previously tested on data from child burn patients [8]. The adult version includes methodological changes in the target and feature space engineering. Thus, with suitable changes, the proposed data-driven model can be adapted and deployed for other disease groups using data sets specific to the patients. The NHS can immediately adopt the DT grouper to reimburse the cost of providing burn care, given that the variables used are present in the iBID and HES databases.

In developing the cost matrix used in the model, we assumed a linear and monotonic penalty for misclassification. However, in reality, this might not be applicable. Thus, further work may be needed to customise the cost matrix to reflect a non-linear and non-monotonic penalty (e.g. actual monetary cost) for misclassification. The data-driven approach of the Grouper implies that any poor practices intrinsic and unchecked in the data would be reflected in the generated groups. However, it is worth noting that this limitation exists in the current Grouper and an attempt to manage this in our Grouper is through the use of multiple variables in the feature space.

Appendix A. Variables available in the iBID dataset

Type of Measure	Measures
Cost	Patient Level Cost
Circumstances of Burn	Living circumstances, Time from injury to admission, Injury type, Source of injury, Activity, Category, Living space, Locality, First aid received, Blast injury suspected, Body mass index, Core temperature on admission
Demographic Factors	Gender, Age, Race, Ethnic category, Immigrant generation.
Mental State	Pre-existing (Alcohol abuse, Substance abuse, Psychiatric disorder, Depression, Post traumatic stress disorder, Anxiety, Personality disorder, Mania, Schizophrenia, Self-harm, Attention deficit hyperactivity disorder, Eating disorder, Learning difficulty) Dementia, Suicide, Self harm, Mental state indicator.
Pre-existing Conditions	No pre-existing indicator, Non-insulin-dependent diabetes mellitus, Insulin-dependent diabetes mellitus, Epilepsy, Asthma, Upper respiratory tract infection, Urinary tract infection, Non-specific viral illness, Failure to thrive, Chronic obstructive pulmonary disease, Emphysema, Preadmission bacteria, Perivascular disease, Ischemic heart disease, Past myocardial infarction, Hypertension, Cardiac failure, Heart valve disorder, Cardiac dysrhythmia, Neoplasm, Metastatic disease, Past cerebrovascular, Hemiplegia, Past deep vein thrombosis, Poor Hearing or deafness, Poor eyesight or blind, Hepatic dysfunction, Renaldys function, Cerebral palsy, Paraplegia, Spina bifidia, Motor disability Immobility, Sensory disability, Multiple sclerosis, Pregnancy 1st semester, Pregnancy 2nd semester, Hepatitis B, Hepatitis C, Human immunodeficiency virus, Anaemia, Clotting disorder, Immunosuppression, Leukaemia lymphoma, Resus, Previous complications, Minor pre-existing disorder, Significance of pre-existing disorder and Count of pre-existing disorder.
Complications	Hyponatraemia litre-30mmol, No complications, Overwhelming wound infection, Myocardial infarction, pancreatitis, Deep Vein Thrombosis, pulmonary embolism, Cerebrovascular, Toxic shock like illness, Peripheral neuropathy, Encephalopathy, Ectopic calcification, Respiratory failure, Renal failure, Hepatic failure, Cardiac failure, Multi organ failure, Acute respiratory distress syndrome, Septicaemia, Disseminated Intravascular Coagulation, Small bowel ileus, Pseudo obstruction, Gastrointestinal bleed, Pneumonia.
Treatment	Intubated, Comfort care only, Maximum support, Selective digestive decontamination, Escharotomies, Days from admission to first operation, Days from last operation to discharge, Days from admission to last operation, Days from injury to first operation, Total thearte visits, Total grafting operations.
Outcome	Death, Discharge destination, Weight change at discharge, Total length of stay, Length of stay (in: ventilation, bed ward, high dependency unit, intensive care unit, complications and comorbidities), Days from injury to healing.
Injury Severity	Total burn surface area (TBSA), Per cent superficial burn, Per cent deep dermal burn, Acute injury indicator, Depth of burn (body, face, feet, face hand and feet, face, hands, feet and perineum, hands, hands and head, legs, upper limb, total), Charlson index, Frailty score, Presence of inhalation, Count of inhalation symptoms, Inhalation severity, Vocal cord oedema, ua oedema, c oedema, b oedema, Abdomen (presence of burn, depth, TBSA), Anterior chest (presence of burn, depth, TBSA), Buttocks (presence of burn, depth, TBSA), Face (presence of burn, depth, TBSA), Left foot dorsal (presence of burn, depth, TBSA), Left foot sole (presence of burn, depth, TBSA), Left forearm (presence of burn, depth, TBSA), Left hand dorsal (presence of burn, depth, TBSA), Left hand palmar (presence of burn, depth, TBSA), Left lower leg(presence of burn, depth, TBSA), Left shoulder (presence of burn, depth, TBSA), Left thigh (presence of burn, depth, TBSA), Left upper arm (presence of burn, depth, TBSA), Lumbar back (presence of burn, depth, TBSA), Neck (presence of burn, depth, TBSA), Perineum (presence of burn, depth, TBSA), Post chest (presence of burn, depth, TBSA), Right foot dorsal (presence of burn, depth, TBSA), Right foot sole (presence of burn, depth, TBSA), Right forearm (presence of burn, depth, TBSA), Right hand dorsal (presence of burn, depth, TBSA), Right hand palmar (presence of burn, depth, TBSA), Right lower leg (presence of burn, depth, TBSA), Right thigh (presence of burn, depth, TBSA), Right shoulder (presence of burn, depth, TBSA), Right upper arm (presence of burn, depth, TBSA), Scalp (presence of burn, depth, TBSA), Injury to admission delay, Core temperature.

Table A.1: List of iBID variables.

Appendix B. The current HRG4+ replication for burn patients using iBID variables

The first step was to identify the same iBID variables or a proxy to variables used in the HES dataset. After the identification of relevant iBID variables, the next step requires the creation of if-then rules, which includes four primary steps, as summarised below

1. Depth of Burn: The first step was to create a depth of burn classification. This uses the depth of burn recorded on 27 different parts of the body, where the depth of burn ranges from 1 (no burn) to 11 (full-thickness burn in the organs). The patients are then classified into three groups: 1st, 2nd/3rd Depth of burn or unspecified, as shown in Fig below.

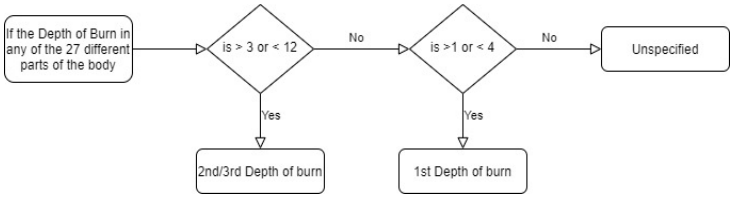

Figure B.1: Identification of the depth of burn classification.

2. Base dummy HRG: The next step is the creation of twelve base dummy HRGs using the depth of burn classification, total burn surface area (TBSA), discharge destination, the total length of stay, length of stay due to complications and the sum of the number of sites in the body that burn occurred (again 27 sites). The dummy HRG creation is summarised in the flow chart below, with progression on the flow chart indicating a more severe burn injury (JB97 are most severe burn injury)
3. Escalated Dummy HRG: The next step is to further account for complications and comorbidities identified to exacerbate the injury caused by burn. These factors are electric burn, burn in the face, hands or feet, escalation due to age and the presence of complications and cormobodities. Patients in lower Base Dummy HRGs identified above are escalated to more severe ones by first determining the escalation numbers as below
    - Electric burn escalation: This is escalated if the type of injury is electric and the source of injury is electrocution (domestic circuit, high tension or lightening).

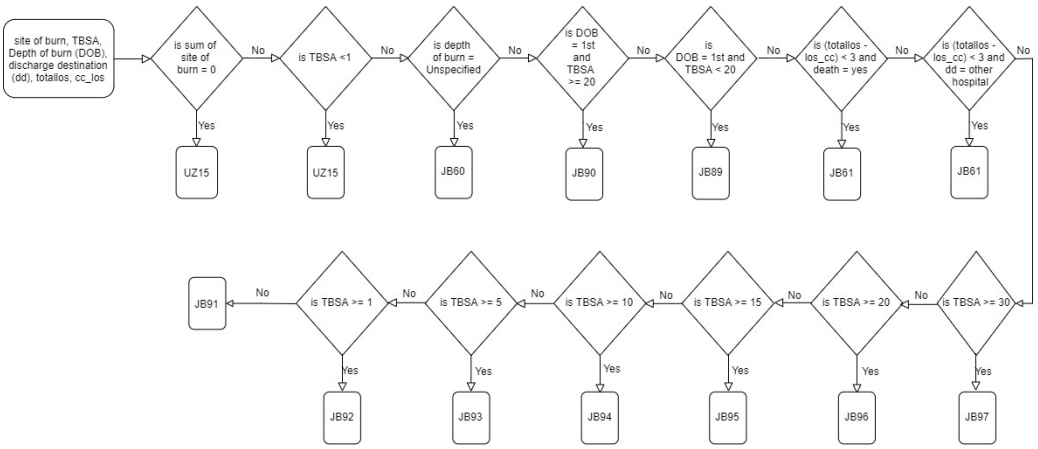

Figure B.2: Creation of base dummy HRGs.

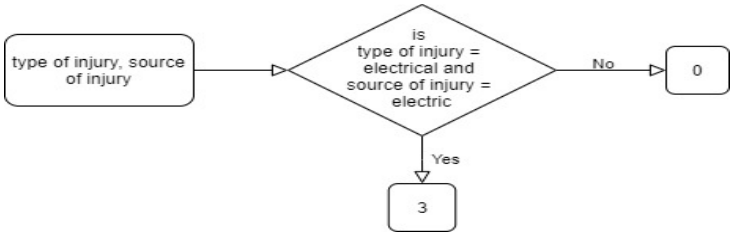

Figure B.3: Electric burn escalation.

- Age escalation: This is escalated to 2 if age is greater than or equals to 80, else it is 1 if age is greater than or equal to 60, otherwise no escalation (group 0)

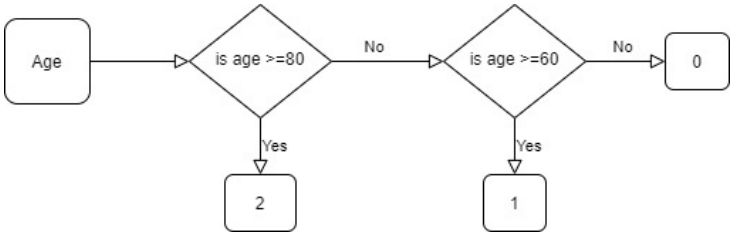

Figure B.4: Age escalation.

- Invasive ventilation escalation: An escalation score of 3 is assigned if a patient was intubated and had an inhalation severity score

greater than 1. The inhalation severity ranges in the order of severity from 1 (no inhalation severity) and 4 (severe inhalation severity).

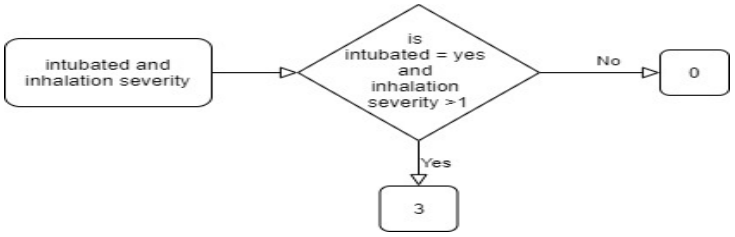

Figure B.5: Invasive ventilation escalation.

- Face hands feet escalation: For this escalation, three other measures are calculated to (a) Burn FNS: Sums up burn occurrence indicators in the face, neck or scalp. The generated value will range between 0 (no burn in any of the three sites) and 1 (occurrence in any or all of the three sites). (b) Burn Hand: Sums up the burn occurrence indicators in any of the four parts of the hand (right- or left- hand dorsal or right- or left-hand palmer). The generated value will range between 0 (no burn in any of the three sites) and 1 (occurrence in any or all of the four sites). (c) Burn Foot: Sums up the indicators of burn occurrence in any of the four parts of the foot (right or left foot dorsal or right or left foot sole). The generated value will range between 0 (no burn in any of the three sites) and 1 (occurrence in any or all of the four sites).

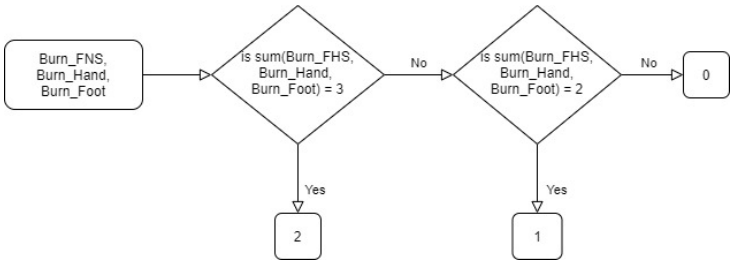

Figure B.6: Face, hands and feet escalation.

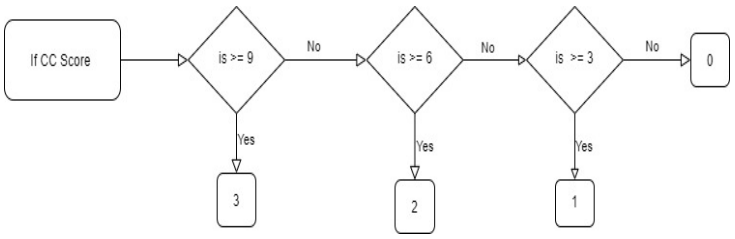

Figure B.7: Complications and Comorbidities escalation.

- Complication and Comorbidities (CC) escalation: A CC score is first calculated, which indicates the complications and comorbidities of patients. An ordered indicator of 2 or 1 for the presence of an existing disorder(s) or complication(s) (this is shown in Table B.1). The CC score is then generated by summing up the value assigned to complications and comorbidities present in the patient. These are then used to create an escalation, as shown in the flow chart below.

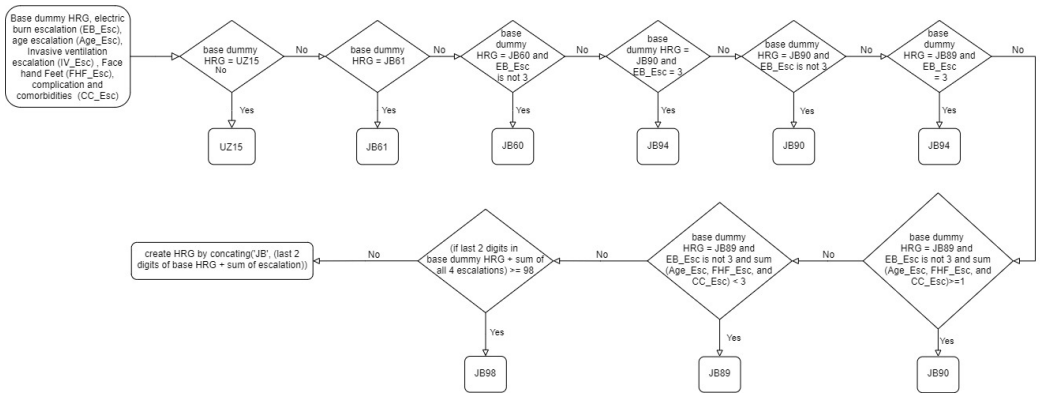

Figure B.8: The Escalation Logic.

With the five escalation variables described above implemented, the escalated dummy HRG is then created as depicted in flow chart below.

4. HRGs: With an escalated dummy HRG created to create homogenous groups in terms of complexity, a final split is done to account for kids and adults' different care pathway and the total grafting operations conducted. The penultimate step of creating the HRG4+ label for burn patients is shown in the flowchart below.

Dummy HRG	Description	Adult0	Adult1	Adult2
JB60	Treatment of Unspecified Degree of Burn	JB60A	JB60A	JB60A
JB61	Treatment of Burn where Patient Transferred or Died in 2 days or less	JB61A	JB61A	JB61A
JB70	Debridement of Burn	JB70A	JB70A	JB70A
JB71	Cleansing and Dressing of Burn	JB71A	JB71A	JB71A
JB89	Treatment of Bun, with Severity Score 1	JB49Z	JB49Z	JB49Z
JB90	Treatment of Burn, with Severity Score 2 (for 1st Degree Burns)	JB48B	JB48B	JB48B
JB91	Treatment of Burn, with Severity Score 2 (for 2nd and 3rd Degree Burns)	JB48B	JB48A	JB48A
JB92	Treatment of Burn, with Severity Score 3	JB47B	JB47A	JB47A
JB93	Treatment of Burn, with Severity Score 4	JB46B	JB46A	JB46A
JB94	Treatment of Burn, with Severity Score 5	JB45B	JB45A	JB43Z
JB95	Treatment of Burn, with Severity Score 6	JB44B	JB44A	JB43Z
JB96	Treatment of Burn, with Severity Score 7	JB42C	JB42B	JB42A
JB97	Treatment of Burn, with Severity Score 8-9	JB41B	JB41B	JB41A
JB98	Treatment of Burn, with Severity Score10+	JB40B	JB40B	JB40A

Table B.1: HRG4+ Creation Final Step to account for severity and patient outcome.

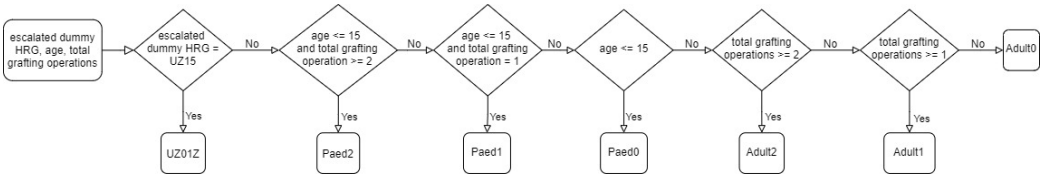

Figure B.9: The logic behind the final HRG label.

With these seven labels created, which represents three groups – adult, child and unclassed. The final step of creating HRG4+ uses the HRG4+ labels and the dummy HRG to create 36 groups using Table B.1 below. With these done, the final step includes the validation of HRGs created. This was done by sharing the summary statistics of the HRGs with the casemix experts, who cross-checked it with the summary statistics of HRGs created with HES data. This showed a similar distribution in each HRG.

Appendix C. Classification tree showing explainable classification rules generated from the costs

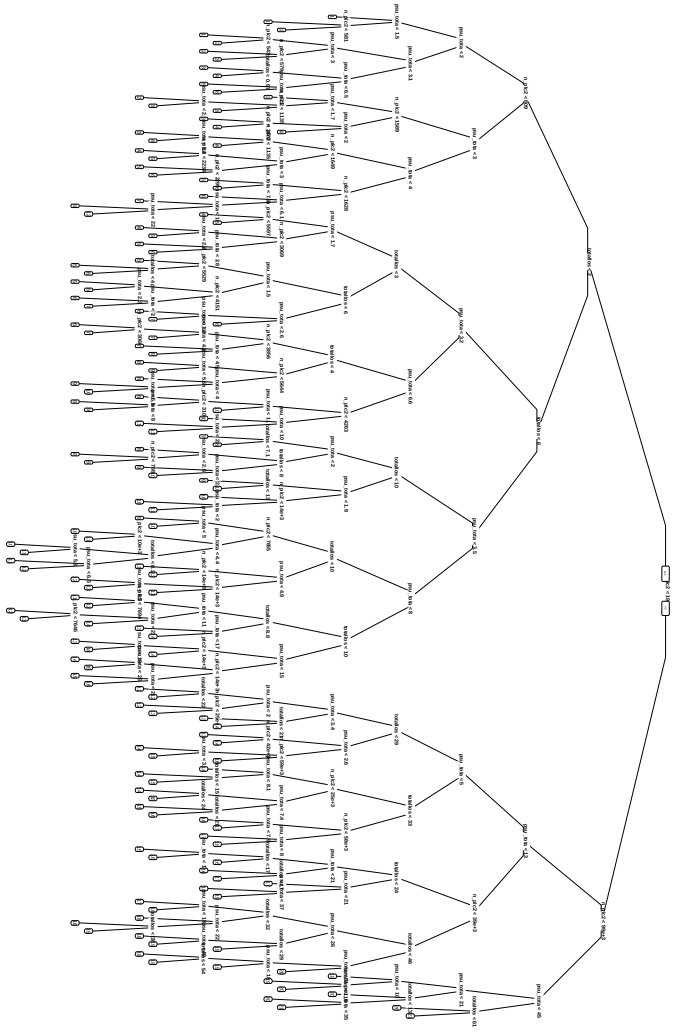

Figure C.1: DT1: Classification tree showing explainable classification rules generated from the cost-sensitive decision tree model.

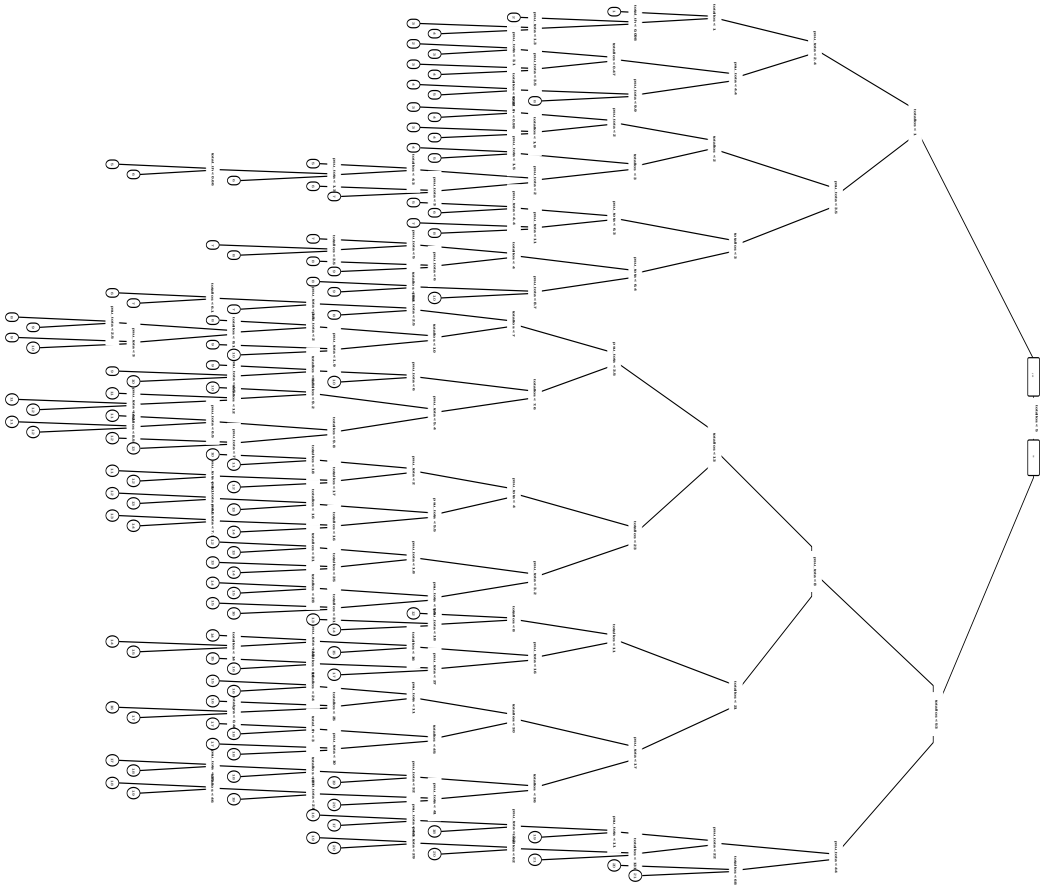

Figure C.2: DT2: Classification tree showing explainable classification rules generated from the cost-sensitive decision tree model.

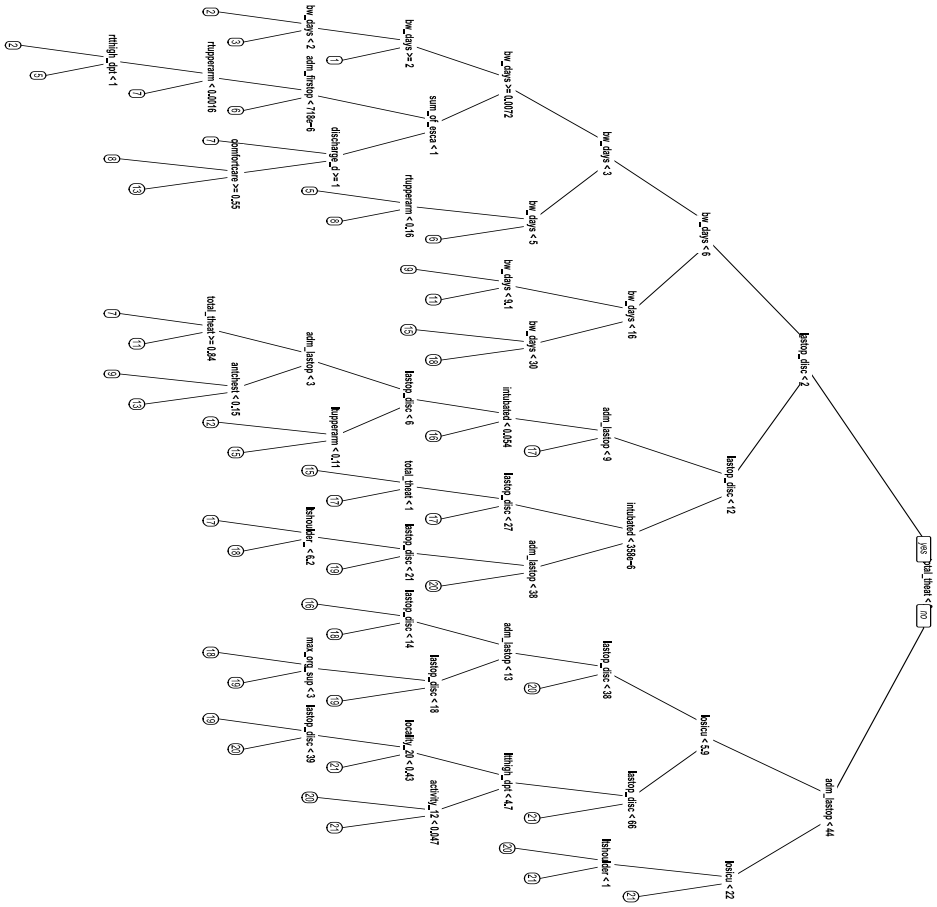

Figure C.3: DT3: Classification tree showing explainable classification rules generated from the cost-sensitive decision tree model.

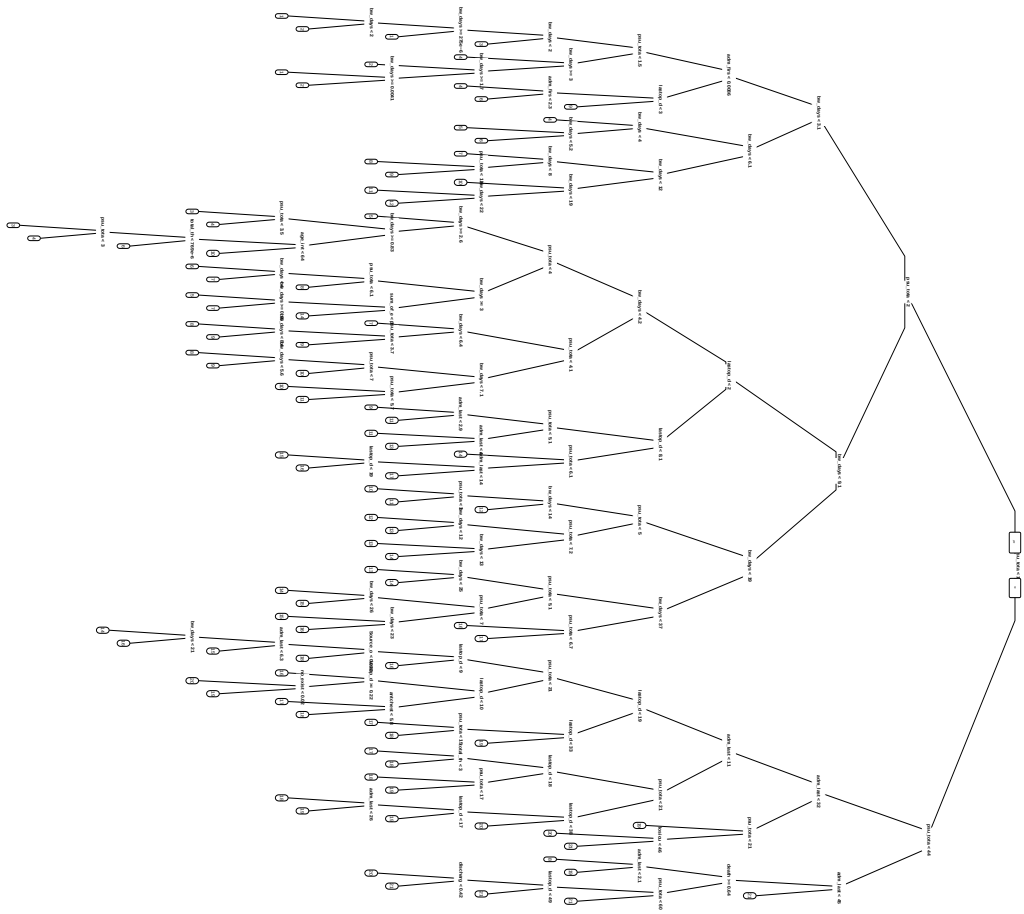

Figure C.4: DT4: Classification tree showing explainable classification rules generated from the cost-sensitive decision tree model.

References

[1] Department of Health Payment by Results team, Simple guide to payment by results, Tech. rep., Department of Health, Leeds (2011).

[2] Department of Health Payment by Results team, A simple guide to Payment by Results, Tech. rep., Department of Health, Leeds (11 2012). URL `www.dh.gov.uk/pbr`

[3] A. Street, C. Kobel, T. Renaud, J. Thuilliez, How Well Do Diagnosis-Related Groups Explain Variations In Costs Or Length Of Stay Among Patients And Across Hospitals? Methods For Analysing Routine Patient Data, Health Economics 21 (2012) 6–18.

[4] C. Bunch, Challenging times for specialist services, British Medical Journal 316 (7128) (1998) 378–379.

[5] R. T. Duncan, K. W. Dunn, Burn service costing using a mixed model methodology, Burns 46 (3) (2020) 520–530.

[6] National Burn Care Review Committee, National Burn Care Review Committee Report Standards and Strategy for Burn Care: A Review Of Burn Care In The British Isles, Tech. rep., National Burn Care Review Committee (2001).

[7] C. N. Onah, R. Allmendinger, J. Handl, P. Yiapanis, K. W. Dunn, A Clustering-Based Patient Grouper for Burn Care, in: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Vol. 11872 LNCS, Springer, Switzerland, 2019, pp. 123–131.

[8] C. N. Onah, R. Allmendinger Julia Handl, K. W. Dunn, R. Allmendinger, J. Handl, Towards a fairer reimbursement system for burn patients using cost-sensitive classification, in: Joint KDD 2021 Health Day and 2021 KDD Workshop on Applied Data Science for Healthcare: State of XAI and trustworthiness in Health, 2021.

[9] Monitor, Substantive guidance on the Procurement, Patient Choice and Competition Regulations (2013).
URL `www.monitor.gov.uk`

[10] NHS England and Monitor, How can the NHS payment system do more for patients?, Tech. rep. (2013).
URL https://www.gov.uk/government/publications/the-nhs-mandate

[11] R. P. Ellis, T. G. McGuire, Optimal payment systems for health services, Journal of Health Economics 9 (4) (1990) 375–396.
URL https://linkinghub.elsevier.com/retrieve/pii/016762969090001J

[12] S. Farrar, D. Yi, S. Boyle, Payment by results, in: New Labour's market reforms, The Kings Fund, 2011, Ch. 5, pp. 66–77.

[13] L. Marshall, A. Charlesworth, J. Hurst, The NHS payment system: evolving policy and emerging evidence, Tech. Rep. February, Nuffield Trust, London (2014).

[14] B. Collins, Payments and contracting for integrated care The false promise of the self-improving health system, Tech. Rep. March, The King's Fund, London (2019).

[15] R. P. Ellis, Creaming, skimping and dumping: provider competition on the intensive and extensive margins, Journal of Health Economics 17 (5) (1998) 537–555.
URL https://linkinghub.elsevier.com/retrieve/pii/S0167629697000428

[16] L. Vinet, A. Zhedanov, A 'missing' family of classical orthogonal polynomials, Journal of Physics A: Mathematical and Theoretical 44 (8) (2011) 085201.
URL https://iopscience.iop.org/article/10.1088/1751-8113/44/8/085201

[17] H. F. Sanderson, P. Anthony, L. M. Mountney, Healthcare Resource Groups–Version 2, Journal of Public Health 17 (3) (1995) 349–354.

[18] P. L. Benton, H. Evans, S. M. Light, L. M. Mountney, H. F. Sanderson, P. Anthony, The development of Healthcare Resource Groups - version 3, Journal of Public Health 20 (3) (1998) 351–358.

[19] S. Ridley, S. Jones, A. Shahani, W. Brampton, M. Nielsen, K. Rowan, Classification trees. A possible method for iso-resource grouping in intensive care., Anaesthesia 53 (9) (1998) 833–40.

[20] R. Busse, EuroDRG group, Do diagnosis-related groups explain variations in hospital costs and length of stay? Analyses from the EuroDRG project for 10 episodes of care across 10 European countries., Health economics 21 Suppl 2 (SUPPL. 2) (2012) 1–5.
URL http://www.ncbi.nlm.nih.gov/pubmed/22815107

[21] J. O'Reilly, L. Serdén, M. Talbäck, B. McCarthy, o. b. o. t. E. Group, Performance Of 10 European DRG Systems In Explaining Variation In Resource Utilisation In Inguinal Hernia Repair, Health Economics 21(S2) (2012) 89–101.

[22] A. Vozikis, S. Xesfingi, E. Moustaferi, T. Balbouzis, T. Rigatos, The DRG-Based Hospital Prospective Payment System in Greece: An Assessment of the Reimbursement Rates Using Clinical Severity Classification, Modern Economy 07 (13) (2016) 1584–1600.

[23] D. T. Wade, Editorial, Clinical Rehabilitation 13 (3) (1999) 183–185.

[24] C. Bojke, K. Grašič, A. Street, How should hospital reimbursement be refined to support concentration of complex care services?, Health Economics 27 (1) (2018) e26–e38.

[25] L. Breiman, J. H. Friedman, R. A. Olshen, C. J. Stone, Classification and regression trees, Classification and Regression Trees (2017) 1–358.

[26] M. E. Smith, C. R. Baker, L. G. Branch, R. C. Walls, R. M. Grimes, J. M. Karklins, M. Kashner, R. Burrage, A. Parks, P. Rogers, A. Saczuk, M. Wagster-Weare, Case-Mix Groups for VA Hospital-Based Home Care, Medical Care 30 (1) (1992) 1–16.

[27] H. F. Sanderson, A. Storey, D. Morris, R. A. McNay, M. P. Robson, J. Loeb, Evaluation of diagnosis-related groups in the National Health Service, Journal of Public Health 11 (4) (1989) 269–278.

[28] U. Pawar, D. O'Shea, S. Rea, R. O'Reilly, Explainable AI in Healthcare, in: 2020 International Conference on Cyber Situational Awareness, Data Analytics and Assessment, Cyber SA 2020, Institute of Electrical and Electronics Engineers Inc., Dublin, 2020, pp. 1–2.

[29] M. Michalopoulos, G. Tselentis, N. S. Thomaidis, G. D. Dounias, Decision Making Using Fuzzy C-means and Inductive Machine Learning for Managing Bank Branches Performance, 1999.

[30] I. Polaka, A. Borisov, Clustering-based Decision Tree Classifier Construction, Technological and Economic Development of Economy 16 (4) (2010) 765–781.
URL https://journals.vilniustech.lt/index.php/TEDE/article/view/5919

[31] P. Domingos, MetaCost: a general method for making classifiers cost-sensitive, in: U. M. Fayyad, S. Chaudhuri, Surajit Chaudhuri (Eds.), Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '99, ACM Press, New York, USA, 1999, pp. 155–164.

[32] International Burn Injury Database, International Burn Injury Database Annual Report (2014/15), Tech. rep., International Burn Injury Database (2014).

[33] N. Stylianou, I. Buchan, K. W. Dunn, A review of the international Burn Injury Database (iBID) for England and Wales: descriptive analysis of burn injuries 2003–2011, BMJ Open 5 (2) (2015) 1–10.

[34] Department of Health Payment by Results team, NHS Costing Manual, Tech. rep., Department of Health and Social Care (2012).

[35] NHS Digital, Chapter Summaries: HRG4+ 2020/21 Local Payment Grouper Version 2 (COVID-19), Tech. Rep. July, NHS Digital (2020).

[36] N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer, SMOTE: Synthetic Minority Over-sampling Technique, Journal of Artificial Intelligence Research 16 (2002) 321–357.

[37] B. Bischl, M. Lang, L. Kotthoff, J. Schiffner, J. Richter, E. Studerus, G. Casalicchio, Z. M. Jones, Mlr: Machine learning in R, Journal of Machine Learning Research 17 (170) (2016) 1–5.

[38] T. Therneau, E. Atkinson, Recursive Partitioning and Regression Trees (rpart) (4 2019).

[39] C. Elkan, The Foundations of Cost-Sensitive Learning, in: IJCAI'01: Proceedings of the 17th international joint conference on Artificial intelligence, Vol. 2, Morgan Kaufmann Publishers Inc., Seattle, WA, USA, 2001, pp. 973–978.

[40] B. Efron, Missing data, imputation, and the bootstrap, Journal of the American Statistical Association 89 (426) (1994) 463–475.

[41] A. Mason, P. Ward, A. D. Street, England : the Healthcare Resource Group system, in: R. Busse, A. Geissler, W. Quention, M. Wiley (Eds.), Diagnosis-Related Groups in Europe, European Observatory on Health Systems and Policies, Maidenhead: Open University Press, 2011, pp. 197–220.