NCIP Portal
Online Portal for Consultant Learning

NHS

Risk adjustment of surgical outcomes in the National Consultant Information Programme (NCIP)

A technical report including a proposed vision for a Minimum Viable Product and approach to release for Urology

Christian Moroy, Edge Health

David Meester, Edge Health

Tom Michaelis, Edge Health

28/02/2021

Abstract

The following paper outlines a methodology and approach for risk adjustment of outcome data in the National Consultant Information Portal, at consultant and provider level for Urology. We also outline the pathway to release within the Urology specialty for a private beta. Approach and methodology were developed together with clinicians and experts in the field as well as aligned closely with the NCIP engineering team to enable implementation into the portal once user facing development resource is secured. In particular, we developed models for 30-day emergency readmission and 90-day mortality rates within the Urology specialty. Resulting risk-adjusted rates are presented in funnel plots. Models use 3 years of historical data for each NCIP procedure. We control for patient characteristics, time of year, clinical risk factors and social risk factors. To ensure clinician buy-in, statistical robustness, and limit any reputational risk to NCIP, we recommend a minimum number of procedures per consultant before consultants are risk adjusted. Model performance is very good for prediction of 90-day mortality, model performance is less accurate for models of 30-day readmission as this is a much harder to predict variable. Performance for both is similar to the existing Urology literature. The developed methodology is limited to HES data for now. Improvements may be achieved by including further data sources, such as clinical registries. While we have developed the methodology to be scalable to all NCIP specialties, some recalibration and optimisation will be needed for each specialty and procedure.

Executive Summary

This report develops a methodology for risk adjustment for the National Consultant Information Programme, at consultant (and provider) level. Risk adjustment is important for NCIP, as taking account of risk factors prevents incorrect or unfair comparisons between consultants and providers. This allows fairer quality measurement, further supporting the goals of the programme. We describe a minimum viable NCIP risk adjustment approach (product) for the Urology speciality. We outline the approach to release within the Urology specialty for a private beta.

In line with the NCIP tool, we developed risk adjustment models for 30-day emergency readmission and 90-day mortality rates within the Urology specialty. We presented risk-adjusted rates in funnel plots, with appropriate control limits. Models are based on 3 years of historical data, and specific to each NCIP procedure. Different procedures carry different risks for the patients, and only at procedure level do clinicians truly believe that the approach can add value. This approach is also recommended in the Urology literature.

We included patient characteristics, time of year, clinical risk factors and social risk factors based on outputs from workshops held with our NCIP clinical leads.

Model performance is excellent for prediction of 90-day mortality, but lower for models of 30-day readmission. This is expected based on the Urology literature and the difficulty in predicting a readmission event [5, 27]. For 30-day emergency readmission, the ranking of consultants before and after risk adjustment was highly similar; the rank correlation was 0.959, out of a maximum of 1. Rates were adjusted between 0.32 to 2.49 times the observed England rate, on average. For 90-day mortality, this rank correlation was 0.894 out of 1. This is similar to other large-scale risk adjustment models [7].

We highlight that some consultants have not undertaken enough procedures to be risk-adjusted. To decide how many procedures are “enough” to enable risk-adjustment, we provide a recommendation that balances coverage with statistical validity.

The developed methodology is limited to HES data. Risk adjustment may be improved with further data sources, such as clinical registries.

We have developed this methodology to be directly and easily scalable to other NCIP specialties. The methodology is easily extended to other metrics, but the risk factors used may need limited calibration and optimisation when studying other metrics. The control limits and prediction models can be implemented in the NCIP tool with sufficient engineering and front-end resource; the models have low computing times, and risk adjustments can be aggregated at consultant level in a similar approach to current metrics.

All our models that we develop at consultant level automatically work at provider level as well, given the set-up of the NCIP tool.

Table of Contents

Tables and Figures

1. Introduction

1.1 Background

Risk adjustment in the context of surgical outcome reporting aims to provide a fairer comparison of outcome measurements between consultants, providers, or other units of analysis. This entails adjusting outcomes for risk factors beyond the control of the care provider, such as demographic, clinical, and socioeconomic factors [1]. The performance or quality of surgeons or providers can then be compared if they hypothetically had treated identical patients, thus allowing to draw conclusions about the quality of care alone [2].

The National Consultant Information Programme (NCIP) enables consultants (surgeons in the first instance) in England to view their patient level outcome data. Currently it presents unadjusted outcome measures to consultants ("crude rates"). Crude rates are presented in boxplots as well as funnel plots, centred around the England median. However, variation in case-mix and social factors across England inevitably partly drive care outcomes. Hence, risk adjustment is important for NCIP, as taking account of risk factors prevents incorrect or unfair comparisons between consultants and providers. This allows fairer quality measurement, further supporting the goals of the programme [3].

The Urology specialty is the vanguard specialty for NCIP. Development of Urology dashboards and metrics is the furthest advanced, hence in the first instance risk adjustment models are tested and applied to the Urology procedures and metrics. This report is based on tests and recommendations for this specialty. Despite this, the approach for the other NCIP specialties will be able to draw heavily on the findings and approach documented in this report.

1.2 Current literature

In the wider literature, urological risk adjustment models are of particular importance to cystectomy [4] or nephrectomy [5, 6]. Risk adjustment models have also been developed for the Urology specialty as a whole [7]. The American College of Surgeons National Surgical Quality Improvement Program (ACS NSQIP) risk calculator has been validated for risk adjustment on specific Urology procedures, but results have shown low performance and poor prediction [5], resulting in calls for procedure specific models [8]. Within the NHS, the summary hospital level mortality indicator (SHMI) or hospital standardised mortality ratio (HSMR) [9] may be used to risk adjust Urology mortality, but these models were never developed for risk adjustment at consultant level.

1.3 Aims

This report develops a methodology for risk adjustment in the National Consultant Information Programme, at consultant (and provider) level. We describe a minimum viable NCIP risk adjustment approach (product) for the Urology speciality. We outline the theoretical basis and motivation for the approach taken. We present the results of the risk adjustment, its strengths, and limitations. We outline the approach to release within the Urology specialty for a private beta.

Currently, the NCIP tool presents funnel plots of outcome rates. The most frequently presented ones are

a) 30-day emergency readmission; and
b) 90-day mortality rates

The aims of the project were to risk adjust commonly presented metrics within the Urology specialty, therefore;

- We develop risk adjustment models for 30-day emergency readmission and 90-day mortality rates;
- We present risk-adjusted rates in funnel plots, with appropriate control limits for internal use, in line with the current presentation of funnel plots for outcome measures; and
- We identify the appropriate approach for release to the NCIP tool.

We have identified and validated the proposed approach in close collaboration with clinicians, such as Mr Andrew Dickinson, NCIP Clinical Lead for Urology, statisticians, such as Dr William Gray, Senior Researcher at GIRFT, and risk adjustment experts, such as Professor Mark de Belder, *Chair of NCAP Operational and Methodology Group and consultant cardiologist,* and Professor Jan van der Meulen, professor of Clinical Epidemiology at the London School of Hygiene and Tropical Medicine.

2. Methods

2.1 Population

We developed risk adjustment approaches (referring to the algorithm for obtaining the final risk adjustment rates in this report) for 30-day emergency readmission and 90-day mortality outcomes as these are key outcomes for Urologists.

Models are based on 3 years of historical data. Workshops with clinical experts suggested that three years gets the trade-off right: longer time periods carry the risk that clinical practice changes, while shorter time periods carry the risk that numbers are too low, leading to lower model performance. The 3-year timeframe also aligns with the NCIP portal display of data making it easier to deploy models to the portal without changing the NCIP way of working. The results in this report are based on data between April 2017 and March 2020.

We use procedure-specific data for risk adjustment, hence we fit risk adjustment models separately, for each procedure. This is important to enable a risk adjustment model to capture the risks for a particular procedure accurately. Not every risk-adjustment model is at procedure level. There again is a trade-off here. Models at specialty level allow more robust models as there is more data than at procedure level. However, different procedures do carry different risks for the patients and only at procedure level do clinicians truly believe that the approach can add value.

Our sample in each risk adjustment model therefore consists of all patients undergoing the NCIP ***procedure*** of interest ***within a three-year time period***. This is consistent with the literature [8] and enables buy-in from clinicians. Further, this optimises amongst various trade-offs and ensures high validity of the NCIP risk adjustment approach, thus safeguarding NCIP reputation amongst clinicians.[1]

2.2 Risk factors

Clinically significant risk factors were identified for our risk adjustment models in workshops with the Urology clinical leads. Key risk factors focussed on patient demographics, clinical risk factors and comorbidities, social risk factors, and NCIP procedure-specific subgroups, presented in Table 2.1. After several workshops and a variety of empirical tests (more detail

[1] Further, NCIP presents outcome rates separately for each procedure. If risk adjustment models are more general, using all Urology data, risk adjustments in a particularly low-risk procedure might be too high, because they are adjusted to the overall Urology risk. Using procedure specific data allows calibration of adjustments to the observed procedure-specific outcomes, in contrast to a model for all Urology patients. Although methods for recalibration exist, we prefer this simpler approach, which ensures correct calibration for each procedure.

in Appendix A1, B6) the final model includes the risk factors outlined in Table 2.1[2]. We note the following considerations for this list of risk factors:

- We do not include obesity, alcohol and tobacco related diagnoses. They have clinical significance, but they are poorly coded and do not increase model performance[3] (Appendix B6).
- All models include social risk factors and ethnicity. The literature remains divided on this topic, as adjustment may result in, and signal acceptance, of suboptimal care for disadvantaged groups. However, social risk factors are crucial for validity of the developed models, as preliminary results have shown, and absence of social risk factors may encourage under delivery of care, increasing health disparities. We held several workshops on this topic and on balance believe that the conceptual (e.g., [4]) and empirical basis for including these is stronger than for excluding.
- Ethnicities were grouped into white, black, Asian, other and unknown (not reported) ethnicities. Unknown ethnicities were included as a separate category. This represents a large group with significantly lower mortality/readmission rates (Appendix A1).
- NCIP has previously identified clinically relevant subgroups of patients within each procedure with clinicians. These subcategories contain valuable information, indicate different levels of risk, and are highly important to clinicians. This may include diagnoses or surgical approaches enhancing surgical risk. For example, neuropathy strongly increases surgical complexity in patients undergoing percutaneous nephrolithotomy [10]. We included NCIP procedure, diagnosis and approach subgroups, where identified for the procedure of interest, in all models.
- Most clinical risk factors are based on ICD-10 coding of HES data. We therefore included the mean trust coding depth of ICD-10 codes in models, as previously defined by NHS Summary Hospital-level Mortality Indicator. This adjusts risk estimates so that patients don't get a higher risk because their provider "codes" more of their diagnoses.

We provide further details on the risk factors included in Appendix A2. NCIP uses Hospital Episode Statistics (HES), hence we have limited ourselves to risk factors available from HES data. Clinical presentation, such as tumour staging, may not be appropriately accounted for. We note that this may cause biases to the risk adjustment. Patients with a worse clinical presentation than represented by their HES data may not be presented accurately, for example. This may also reduce the predictive accuracy of the model. We discuss this limitation further in Section 4.

We have excluded individuals with missing data on the index of multiple deprivation, age and sex, or implausible values corresponding to data entry errors. While removing these patients did not impact the representativeness of the sample, further investigation of missing data is recommended. Data cleaning steps are described in detail in Appendix A1.

[2] We recommend inclusion of palliative care in models of mortality, but not in models of readmission.
[3] In the first instance, we did not include these risk factors in our conceptual model. We list them in this table because they are included in further sensitivity analyses. All sensitivity analyses used out-of-sample, cross validated model performance to prevent overfitting.

Table 2.1: Overview of identified risk factors and their descriptions

Category	Variable name	Description
Demographics	Age	Age of patient in years
	Gender	Sex of patient (Male/Female)
	Ethnicity	Ethnicity of patient: White/White British, Black/Black British, Asian/Asian British, Mixed/Other and Unknown (Not provided). Reference group: Asian.
Time	Quarter of year	Quarter of episode start date in calendar year (Jan-Mar, Apr-Jun, Jul-Sep, Oct-Dec).
Clinical	Emergency admission	Emergency admission for current spell (yes/no)
	Charlson score	Charlson comorbidity score, see [12] for units and construction
	Number of emergency admissions in prev. year	Total number of emergency admissions in the 365 days before the current episode.
	HFRS frailty band	Hospital Frailty Risk Score (HFRS) and subsequent categorisation (none, mild, moderate, severe), see [13]
	Palliative care	Patient is on palliative care, following ICD-10 codes Z515, Z518.
	Obesity	Diagnosis of obesity, following ICD-10 code E66.
	Tobacco-related diagnosis	Nicotine dependence, following ICD-10 code F17
	Alcohol-related diagnosis	Alcohol dependence, following ICD-10 code F10
	Coding depth	Mean number of secondary diagnoses, per provider, see [11]
Social	IMD score	Combined score for the Index of Multiple Deprivation (IMD).
	Urbanicity	Indicates if patient is living in a city and town (Urban), compared to Rural areas (town and fringe, villages, hamlets and isolated dwellings) [14] (categories: Urban, Rural).
	GP to patient ratio	Care provision, as defined by the number of GPs to patients in the individuals' GP practice during the midpoint of the 3-year timeframe. Linked to national data for the general practice workforce as of March 2018 (the time-period studied) [15].
NCIP categories	NCIP Procedure subgroup	Subgroups of NCIP procedure diagnoses, indicating strong heterogeneity between procedures carried out.
	NCIP Diagnosis subgroup	Diagnoses identified by NCIP that are relevant to the procedure of interest. For example, there is increased surgical difficulty to a percutaneous nephrolithotomy with neuropathic patients.
	NCIP Approach subgroup	Approach, as identified by NCIP (Open/Robotic/Minimal access), where applicable.

Note: Variable categories, names and descriptions are presented. Age and quadratic age are included to account for non-linear effects. Male, Emergency (EM) admission variables are included only where applicable. NCIP subcategories, such as procedure, diagnosis and approach subgroups are used when relevant to the procedure, as identified by clinical leads and analysts, and as displayed on the NCIP dashboard.

2.3 Variable selection

For the Urology specialty, 27 procedure-specific models are to be considered for addition to the tool. We use the clinically relevant variables, identified in Table 2.1, in the final models. We do not allow for automated variable selection as previous literature shows an automated variable selection procedure is likely to be of little influence on HES data [16, 17] and is often seen as a "data driven" rather than "clinically driven" approach. In sensitivity analyses we do investigate inclusion of additional variables and alternative categorisations.

2.4 Risk prediction

Once risk factors are identified, the risk adjustment methodology consist of 3 steps[4]:

1. Prediction of risk for each individual patient
2. Calculation of multiplication factors for each consultant (Standardised Mortality Ratios (SMRs)). These factors indicate the degree of risk adjustment for each consultant, using the predicted scores.
3. Calculation of risk adjusted rates for each consultant. This adjustment indicates performance compared to the England average.

For risk predictions, we used multivariable logistic regression models to predict risk at patient level for each procedure. We assessed model performance for discrimination (indicating how well the model distinguishes high-risk patients from low-risk patients), and calibration (indicating how well the model predicted events correspond to the observed events) using four different methods:

- Discrimination was assessed using the c-statistic. This statistic ranges between 0 and 1, where higher scores are better. A random model achieves a c-statistic of 0.5.
- We assessed calibration using:
    - the Brier score, which ranges between 0 and 1, where lower scores are better; and random models achieve a score of 0.25.
    - A visual inspection of observed and expected rates in model predicted risk deciles; and
    - the Hosmer-Lemeshow (HL) test.

Performance metrics are described in more detail in Appendix A3. We assess performance in a test set, using data from April 2020 to March 2021[5], further detail is provided in Appendix A4.

2.5 Risk adjustment

For risk adjustment, we aggregate patient level predictions of risk at consultant or provider level (unit of analysis) to perform indirect standardisation[4]. First, we compute Observed (O) and Expected cases (E) for each unit of analysis from the risk predictions. This provides Standardized Readmission and Mortality Ratios (SMRs and SRRs)[6], indicating the degree of risk adjustment. Next, SMRs/SRRs were multiplied by the observed rate to create risk-adjusted rates that are directly interpretable[7].

[4] This method uses indirect standardisation of rates, through logistic regression. We recommend indirect standardisation, compared to direct standardisation, as it allows for adjusting multiple variables. [CMS]
[5] Performance metrics based on the 2017-2020 sample may overestimate performance as the model may specifically be optimised (overfitted) to that period. We therefore present performance in an unseen test set.
[6] The SMR (for mortality) and SRR (for readmission) is defined as O/E.
[7] Thus, it follows that the Risk Adjusted Rate = O/E * Observed England rate = SMR * Observed England rate.

2.6 Presentation of risk adjusted rates

To align with the current presentation of rates in the NCIP tool, risk adjusted rates were presented in funnel plots[8]. Moreover, there is clinician support for the use of funnel plots [18]. We centre rates around the observed England rate (benchmark). We recommend the usage of approximate binomial control limits (at 95% and 99.8% level). Again, there is a trade-off here and disagreement in the literature. These limits can be seen as quite narrow as they do not consider structural differences between providers (over-dispersion), causing many outliers [19]. However, there is controversy in the literature around the use of "over-dispersed" control limits [20] as these may artificially widen limits and thus "let clinicians off the hook" if the over-dispersion assumption is not justified. Given that there is no obvious right answer from the literature, and after holding workshops with experts in the field, we recommend the usage of "traditional" approximate binomial control limits to avoid controversy. However, we note that these control limits should guide a discussion for clinical improvement; they should not be interpreted as strict boundaries[9]. The approach used here is in line with guidance on the detection and management of outliers, produced by the National Clinical Audit Advisory Group (NCAAG) [21]. The 95% level indicates "alert", the 99.8% level indicates "alarm"[10].

For some procedures, risk-adjusted outcome rates contain too much statistical uncertainty for presentation. This especially occurs in rare procedures with small numbers and low outcome rates, such as mortality outcomes. We calculated the minimal number of procedures required to obtain enough statistical certainty (also called "power")[11]. Presentation of risk-adjusted rates was then limited to units of analysis with sufficient procedures in the data, setting a lower bound.

The minimum number was determined separately for 30-day readmission and 90-day mortality, but used observed rates across all Urology procedures to maintain consistency of presented rates within NCIP[12]. Concerns about small sample sizes per unit of analysis are not new [22]. In taking this approach, we align with the most recent research on this topic [23] and with previous literature [24]. The approach ensures that all consultants are presented reliable rates, which is of high importance to ensure reliability during appraisal and re-validation and keep consultant buy-in. Moreover, it maintains statistical validity.

However, we recognise that setting a lower bound also limits the number of consultants that are suitable for risk adjustment. We therefore present 3 different scenarios for consideration:

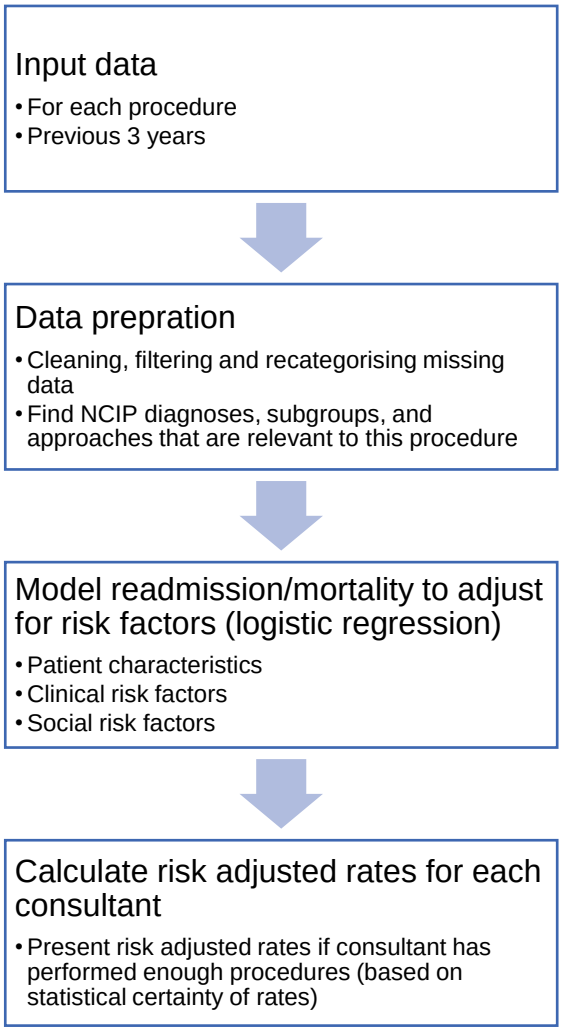
1. An approach with maximum consultant/procedure coverage, but with higher risk to NCIP, as risk-adjusted rates are unstable and less statistically valid for consultants with few procedures[13].
2. A mixture of both statistical validity and high consultant coverage. This approach presents a minimum bound that optimises consultant coverage but maintains the NCIP reputation for consultants[14].

[8] It is worth noting that funnel plots are currently not well signposted in the tool. Use of the functionality may be optimised by displaying risk-adjustment on box plots or improving the UI to make the funnel plots more easily accessible (or both).
[9] We refer the interested reader to Appendix B8 for a further comparison.
[10] An alternative option would be to avoid control-limits altogether, if desirable.
[11] Here, we aimed for 80% power to detect potential outliers with a 2.5 times increase in the risk adjusted outcome rate, compared to the observed population rate (1-sample, 1-sided), using the normal approximated binomial distribution.
[12] An exception is made for Nephrostomy, Cystectomy and other procedures with a high mortality rate (> 2%).
[13] This approach uses the lower bound of method 1., divided by 5. This amounts to a lower bound of 10 (50) procedures for emergency readmission (mortality)
[14] This approach uses the lower bound of method 1., divided by 2. This amounts to a lower bound of 25 (125) procedures for emergency readmission (mortality)

3. A statistically valid lower bound on the sample size per consultant (presented above). This achieves maximum statistical validity, but limits the coverage of consultants, as many consultants do not perform the minimum number of procedures.

The overall process of risk adjustment is summarised in Figure 2.1.

Figure 2.1: Steps needed for risk adjustment



3. Results

3.1 Sample sizes

Sample sizes for each procedure-specific risk adjustment model, including the number of cases per procedure, are presented in Appendix B1 (3-year period, April 2017-March of 2020). Descriptive statistics on all risk factors, by 30-day readmission or 90-day mortality, are presented in Appendix B2. Risk adjusted rates vary strongly across procedures.

We have identified 3 approaches for handling consultants with a lower number of procedures, balancing statistical validity and consultant coverage. In approach 3, we determine the minimum number of procedures required per unit of analysis for sufficient statistical certainty (detailed in Table B.4). For mortality, consultants need to have performed 250 procedures before risk adjust ments are statistically stable. (Table 3.1).

This approach highly limits the number of consultants for which risk adjustment takes place, as seen in Table 3.1. For risk adjustment of 90-day mortality, only 9 out of 27 procedures therefore have consultants with sufficient procedures carried out[15], with, on average, only 7% of total consultants available for risk adjustment.

To balance the statistical validity and coverage of the number of procedures, approaches 1 (maximum coverage) and 2 (balanced statistical validity – coverage) are also presented. They allow for presenting more consultants, but lose statistical validity for consultants with less procedures.

Table 3.1: Number of procedures and consultants available for risk adjustment, by approach

Nr of procedures required per procedure (lower bound)	**Approach 1**	**Approach 2**	**Approach 3**
30-day Emergency Readmission	10	25	50
90-day mortality	50	125	250
Procedures with more than 5 consultants:	**Approach 1**	**Approach 2**	**Approach 3**
30-day Emergency Readmission	26	25	20
90-day mortality	21	14	9
Avg. percentage of consultants with risk adjustment:	**Approach 1**	**Approach 2**	**Approach 3**
30-day Emergency Readmission	75.6%	43.7%	30.7%
90-day mortality	34.5%	18.3%	7.0%

Note: Lower bounds for the number of procedures performed per consultant are presented, by approach. The resulting number of dashboards and percentage of consultants with risk adjustment is presented. Percentage of consultants with risk adjustment represents average percentage of consultants with the number of procedures performed above the lower bound, over all procedures.

[15] Here, we define a procedure with enough consultants as a procedure with more than 5 consultants available for risk adjustment.

For Approach 3, Table 3.2 outlines the procedures selected for inclusion in the tool by outcome, and the number of consultants presented:

Table 3.2: List of procedures, consultants suitable for presentation of risk adjusted rates, and total consultants in NCIP

Procedures	Consultants presented for readmission	Consultants presented for mortality	Total consultants in NCIP	Suitable for risk adjustment
Insertion of ureteric stent | age 17+	470	138	1422	Yes
Ureteroscopy | age 17+	397	15	1005	Yes
Endoscopic resection of lesion of bladder (TURBT) | age 17+ | elective	524	25	1024	Yes
Percutaneous nephrolithotomy (PCNL) | age 17+ | elective	214	7	551	Yes
Prostate biopsy | age 17+ | elective	468	121	794	Yes
Male bladder outflow obstruction surgery | age 17+ | elective	551	24	1034	Yes
Circumcision | age 17+ | elective	436	≤5	1032	Readmission only
Hydrocoele | age 17+ | elective	20	≤5	817	Readmission only
Extracorporeal shock wave lithotripsy of calculus (ESWL) | age 17+ | elective	226	66	490	Yes
Urethrotomy | age 17+ | elective	18	≤5	722	Readmission only
Prostatectomy for cancer | age 17+ | elective	135	23	180	Yes
Nephrectomy for cancer | age 17+ | elective	116	≤5	301	Readmission only
Cystectomy for malignant neoplasms of the bladder | age 17+ | elective	40	6	129	Yes
Urethral dilatation, male | age 17+ | elective	68	≤5	912	Readmission only
Urethral dilatation, female | age 17+ | elective	128	≤5	844	Readmission only
Injection of bulking agents | age 17+ | elective	38	≤5	283	Readmission only
Peyronie's surgery | age 17+ | elective	7	≤5	78	Readmission only
Sacral nerve stimulation for urinary conditions | age 17+ | elective	8	≤5	75	Readmission only
Insertion of penile prosthesis | age 17+ | elective	6	≤5	39	Readmission only
Urethroplasty for stricture, male | age 17+ | elective	11	≤5	36	Readmission only
Nephrostomy | age 17+	≤5	≤5	514	No
Nephrectomy for benign disease | age 17+ | elective	≤5	≤5	217	No
Litholapaxy for bladder stones | age 17+ | elective	≤5	≤5	709	No
Nephroureterectomy for cancer | age 17+ | elective	≤5	≤5	188	No
Insertion of artificial sphincter | age 17+ | elective	≤5	≤5	42	No
Cystectomy for benign disease | age 17+ | elective	≤5	≤5	20	No
Vaginal fistula | age 17+ | elective	≤5	≤5	18	No

Note: *Number of consultants available for risk adjustment, meeting the lower bound set in Approach 3, and total number of consultants is presented, by NCIP Urology procedure. A procedure is available for risk adjustment if more than 5 consultants satisfy the lower bound.*

In Approach 3, we note that the number of consultants presented is especially low for mortality. The low outcome rates in the Urology specialty and low volumes for some procedures prevents statistically reliable presentation. We believe that the statistical reliability is necessary to ensure clinician buy-in and maintain the NCIP reputation: otherwise, some consultants with low procedure numbers will be presented unstable rates. Approach 2 balances statistical validity and consultant coverage. This approach may therefore be similarly suitable to maintain the NCIP reputation, but also has both statistical and coverage advantages. Not all programs for risk adjustment set a lower bound. The National Joint Registry (NJR), for example, does not specify a minimum bound on the number of procedures [25]. Approach 1, for instance, could therefore also be considered.

3.2 Performance of predictive models

We predict patient risk of readmission and mortality for relevant procedures. For emergency readmission, 20 procedures were risk adjusted. For mortality, 9 procedures with sufficient data were risk adjusted, as seen in Table 3.1. The c-statistic, ranging from 0 to 1, measures model performance. Following the literature, a performance (c-statistic) above 0.9 is widely considered as “excellent”, between 0.8-0.9 is considered good, between 0.6-0.8 is considered fair, and below 0.6 is generally considered poor [26].

Model performance is summarised in Table 3.3. Over procedure specific models, this table presented the maximum and minimum performance achieved, as well as the mean and median performance over all procedures. Moreover, it presents the number of procedures achieving a performance above set levels of the c-statistic.

For mortality, model performance was excellent, following the literature, as indicated by the high c-statistics, above 0.8 (or 0.9). For the test set, most procedures showed obtained a performance above 0.7, and 56% of procedures achieving a c-statistic above 0.8. At 0.874, models for Ureteroscopy achieved the highest discrimination. Prostatectomy for cancer achieved the lowest discrimination, with a c-statistic of 0.459, due to an extremely small sample size[16]. Model performance for each procedure is presented in detail in Appendix B.4.

On the other hand, for emergency readmission, we therefore classify model performance (discrimination) unfortunately as relatively poor, following the risk prediction literature. The maximum performance was 0.810, for Sacral nerve stimulation for urinary conditions, with the lowest performance for risk adjustment models of Urethroplasty for stricture, at 0.464. The median performance was 0.613 in the test sets, over all procedure-specific models, and 65% of procedures achieved a c-statistic above 0.6.

This was as expected. First, emergency readmission remains a difficult outcome to predict [27]. Second, predictions are at NCIP procedure level, with relatively small datasets. Third, the definitions of NCIP procedures aim at eliminating large variations, hence patients are similar in risk due to the use of diagnoses, procedures, and patient characteristics of in NCIP procedure definitions.

Nonetheless, this only indicates discrimination, not calibration (indicating if the observed and expected events are similar), which is shown in Appendix B5-6. In all models, calibration was excellent[17]. We use calibration to decide if a model can be used; good calibration is essential for risk-adjustment estimates to be of high quality[18].

[16] In a further step, procedures with small sample sizes should be excluded.

[17] Brier scores were low; average: 0.027 for mortality, and 0.083 for readmission, and HL-test p-values were frequently above 0.05.

[18] Without proper calibration, risk adjustment models might structurally adjust rates too much (too high/low) for some individuals

Discrimination indicates how much of the variation in rates is adjusted for, and therefore indicates how good the risk adjustment models are, once used. Poorly discriminating models[19] , such as the models for emergency readmission, can therefore still adjust for the observed risk factors. The subsequent risk adjusted rates will therefore still be adjusted for the risk factors in our model, but some unexplained variation remains[20].

Table 3.3: Model performance (discrimination) for mortality and emergency readmission in all procedure specific models.

Mortality (90-days)	**In-sample performance**	**Test performance[21]**
Median	0.811 out of 1	0.805 out of 1
Min.	0.479[22] out of 1	0.459 out of 1
Max.	0.906 out of 1	0.874 out of 1
Mean	0.777 out of 1	0.765 out of 1
Number of procedures with C-statistic > 0.6	8 (89%)	8 (89%)
Number of procedures with C-statistic > 0.7	8 (89%)	8 (89%)
Number of procedures with C-statistic > 0.8	5 (56%)	5 (56%)
Number of procedures with C-statistic > 0.9	1 (11%)	0 (0%)
Emergency readmission (30 days)	**In-sample performance**	**Test performance**
Median	0.599 out of 1	0.613 out of 1
Min.	0.502 out of 1	0.464 out of 1
Max.	0.715 out of 1	0.810 out of 1
Mean	0.598 out of 1	0.613 out of 1
Number of procedures with C-statistic > 0.6	10 (50%)	13 (65%)
Number of procedures with C-statistic > 0.7	1 (5%)	1 (5%)
Number of procedures with C-statistic > 0.8	0 (0%)	1 (5%)
Number of procedures with C-statistic > 0.9	0 (0%)	0 (0%)

Note: Model performance for the minimum and maximum performing model out of all procedure-specific models are presented. The median and mean performance are presented over all procedures. Following the literature, a c-statistic above 0.9 is widely considered as "excellent", between 0.8 - 0.9 is considered good, between 0.6 - 0.8 is considered fair, and below 0.6 is generally considered poor.

3.3 Sensitivity Analyses

In sensitivity analyses, we considered the inclusion of additional variables, alternative adjustments for age (categorisation in 10-year age bands or quantiles) and Charlson score (individual Charlson comorbidities, quantiles and "0", "0-5", "more than 5" categories, as in the SHMI indicator). This did not impact the model results (Appendix B6).

[19] *Unobserved variation,* such as clinical presentation, is large in these models. The final risk adjusted rates therefore still contain some variation that is not entirely due to consultant performance. However, the *observed variation*, clinical, social, and patient-level risk factors, can still be adjusted for.

[20] To ensure high quality risk adjustment, it is advisable to include a minimum discrimination too, where a c-statistic of 0.58 is likely to balance coverage and quality for the Urology specialty. However, it is hard to specify a single cut-off; this is again a trade-off; and no cut-off is perfect.

[21] Test performance best indicates the future performance; it uses data the model has not yet seen. In-sample performance indicates the cross-validated performance of data within the sample (see Appendix A4).

[22] Note that the automated nature of the procedure leaves some c-statistics below 0.50, due to extremely small datasets. These procedures should be excluded, following the c-statistic cut-off.

3.4 Risk adjustment

Next, we performed indirect standardisation to construct risk-adjusted outcome rates using the model predictions above. As discussed, we only present risk-adjusted rates with enough statistical certainty, following Table B.4. Hence, risk-adjustment models of mortality exclude a considerable number of consultants not achieving the minimum number of procedures.

Of particular importance to NCIP and its consultants is the question: ***"How much do the outcome rates change after risk adjustment?"***

- Before risk adjustment, 30-day emergency readmission rates ranged between 2.58% and 24.64%, on average. After risk adjustment, rates were adjusted between 0.32 to 2.49 times the observed England rate, on average[23]. The ranking of consultants before and after risk adjustment was highly similar; the rank correlation was 0.939, out of a maximum of 1[24].
- For 90-day mortality, we observed similar risk adjustments. After risk adjustment, rates were adjusted between 0.60 to 4.37 times the observed England rate, on average. The ranking of consultants before and after risk adjustment showed little change, with a rank correlation of 0.894 out of 1. This is similar to other large-scale risk adjustment models. [7]
- We present a full overview of the adjustments in boxplots by procedure in Appendix B5.

3.5 Social risk factors

We compared the inclusion and exclusion of social risk factors (Ethnicity, deprivation, urbanicity), to assess the effect of social risk factors on the risk adjusted rates. Detailed results are provided in Appendix B.7. Overall, results were not strongly impacted by social risk factors.

The inclusion of social risk factors resulted in highly similar risk-adjusted rates to the rates without social risk factors (correlation of 0.991, out of 1). A hypothetical rank order of consultants also did not change considerably (rank correlation: 0.937). This was similar in mortality and readmission risk adjustment, and similar for all procedures. In 17 out of 25 procedures, no consultants changed their outlier status between inclusion and exclusion of social risk factors. We conclude that the inclusion of social risk factors is not of major importance towards the risk adjusted rates, but both the literature and empirical results do show that these variables are relevant to care in Urology. We therefore include them in the final model.

[23] This represents the average adjustment over all procedures. For some procedures, there are more (less) adjustments.
[24] Spearman rank correlation, to compare the ranking of consultants

Case study: Male bladder outflow obstruction surgery

Male bladder outflow obstruction surgery achieved the median performance for risk adjustment of readmission. To illustrate the risk adjusted rates, we present funnel plots for risk adjusted 30-day readmission rates before and after Male bladder outflow obstruction surgery. Outcome rates before risk adjustment are presented in Figure 3.1, while the final risk-adjusted rates are presented in Figure 3.2.

Emergency readmission was significantly related to almost all risk factors included in the model, except ethnicity and deprivation (Appendix C). The predictive model had a c-statistic of 0.60 and brier score of 0.08. This implies that performance was good for NCIP, because a significant part of the variation is outcome rates could be risk adjusted. Calibration of the model was also sufficient, indicating that the observed and expected rates were similar.

The figures show that most consultants remain within the control limits, indicating performance as expected under statistical uncertainty.

For male bladder outflow obstruction, 16 consultants were above the upper control limit before risk adjustment (Figure 3.1). After risk adjustment, this reduces to 11 (Figure 3.2). Further, 3 consultants are below the control limits, out of a total of 551.

Figure 3.1: Unadjusted funnel plot for Male bladder outflow obstruction surgery

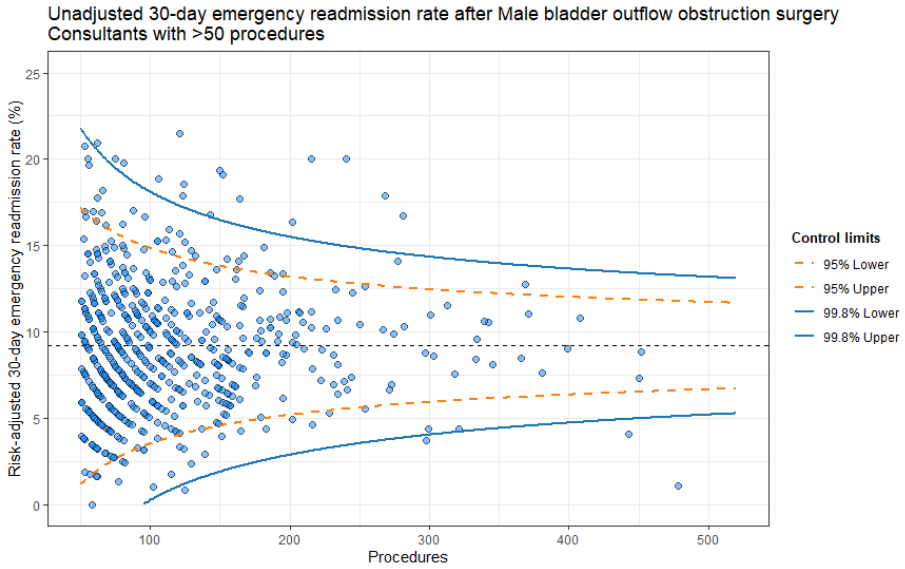
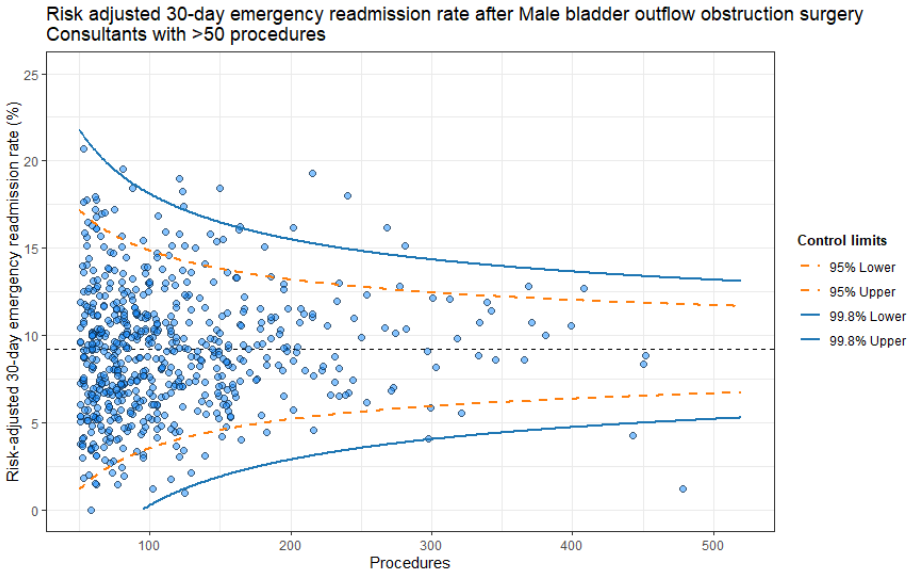


Figure 3.2: Risk adjusted funnel plot for Male bladder outflow obstruction surgery



Case study: Cystectomy

Variation in mortality rates is of particular concern to Cystectomy. We therefore present the risk adjustments for 90-day mortality after “Cystectomy for malignant neoplasms of the bladder.

In predicting cystectomy mortality, model performance was good, as the c-statistic was 0.74, with a Brier score of 0.02. This implies that the model could distinguish mortality events well, and observed and predicted rates were highly similar. The mortality rates before risk adjustment are presented in Figure 3.3. The risk adjusted rates are presented in Figure 3.4.

Before risk adjustment, one consultant performed within the “alert” control limit.

After risk adjustment, no consultants are outside either of the control limits, but one consultant remains inside the “alert” limit, out of a total of 6. This total number of consultants is likely to increase when alternative approaches for the minimum sample size are used. For the presented consultants, we observe that some are risk adjusted downwards strongly, towards the observed England rate.

A full overview of performance and consultants outside control limits for each procedure is given in Appendix B.4.

Figure 3.3: Funnel plot before risk adjustment (unadjusted) for “Cystectomy for malignant neoplasms of the bladder”

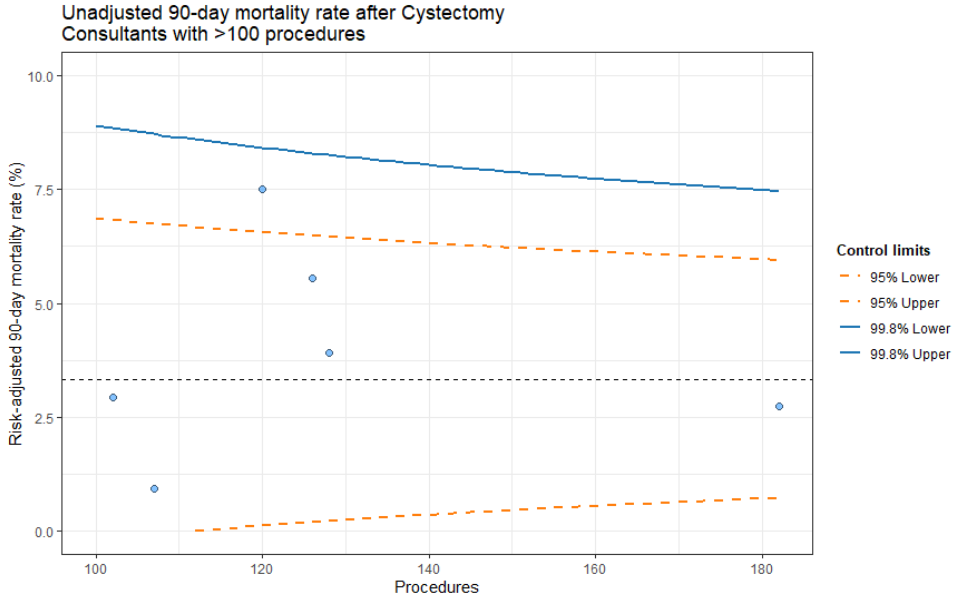
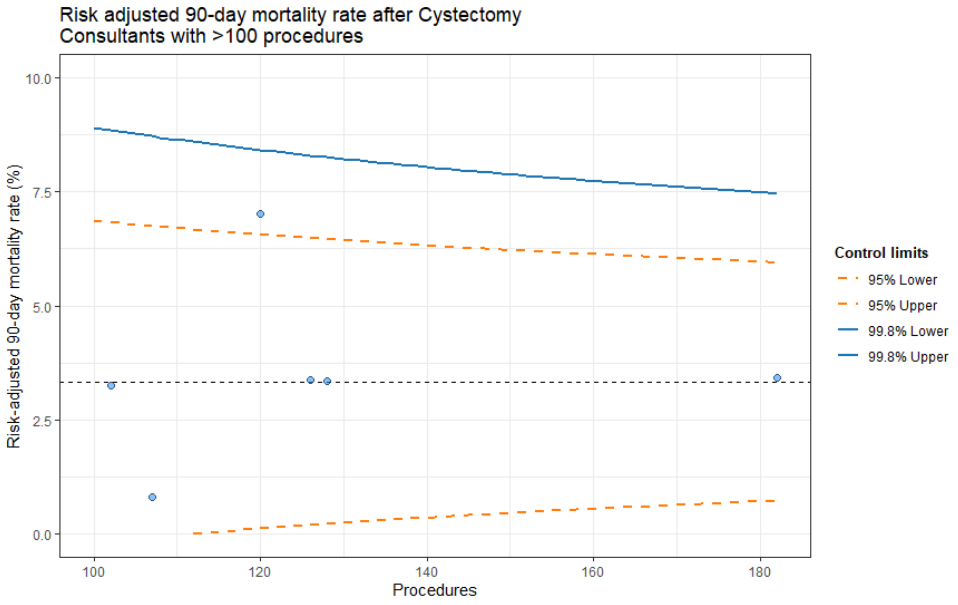


Figure 3.4: Funnel plot after risk adjustment, for “Cystectomy for malignant neoplasms of the bladder”



4. Discussion

4.1 Summary

We performed risk adjustment for all NCIP procedures. First, we selected a clinically significant list of risk factors, including social risk factors, for prediction of readmission and mortality, with appropriate categorisations. Next, we performed indirect standardisation for consultants with a minimum number of procedures performed, to obtain statistically relevant risk adjusted rates at consultant level.

4.2 Comparison of performance to literature

Our models of 90-day mortality were often superior to the existing Urology literature. For instance, BAUS risk analysis achieves a c-statistic of 0.796 for nephrectomy [6], while it was 0.8251 in our model. NSQIP achieves 0.861 over all Urology procedures [28], while our median performance was 0.805 in individual Urology procedures – a much more difficult predictive task. The Summary Hospital Level Mortality Indicator, measuring performance by diagnosis, achieves c-statistics of approximately 0.92 for cancer of the bladder, kidney, and prostate [29]. Our models perform at approximately 0.75 for cystectomy but it is unclear how predictions at procedure and diagnosis level may relate[25].

Models of 30-day emergency readmission performed similarly to the Urology literature which unfortunately is not great. BAUS achieved 0.66 in Nephrectomy complications, while our models achieve 0.60 in Nephrectomy readmission, a more difficult endpoint for prediction. NSQIP achieves 0.51 in predicting complications of partial Nephrectomy. For cystectomy, we achieved a c-statistic of 0.55. While this is not high, cystectomy readmissions remain poorly understood [30], and are hence difficult to predict. A low discrimination does not prevent risk adjustment, but only results in a small amount of variation to be adjusted for.

[25] These are internal model fits, while our model fits are on validation sets. Hence, our model is also a better assessment of predictive performance, and may predict worse for that reason.

4.3 Limitations

4.3.1 Biases due to data availability

Several data sources are not available, leading to potential biases. There is no tumour staging data, for example. We recognise that the model therefore does not completely adjust for clinical presentation. Patients with a poor clinical presentation may appear healthier in the risk adjustment models than expected from ICD-10 codes. In a worst-case scenario, consultants may recognise that patients with poor clinical presentation due to advanced tumour staging (which is currently uncaptured in the HES data) are assigned a low expected risk, while these are high-risk patients in reality. This may result in under delivery of care. While this is highly unlikely, further data sources are required to prevent these biases.

4.3.2 Model performance

Model performance was high for mortality outcomes. However, model performance, in terms of discrimination, was relatively low for emergency readmission. This is a difficult endpoint to predict in the Urology specialty. The use of additional data sources can further increase the predictive performance. Moreover, clinician workshops indicate that when the endpoint is "Emergency readmission with 1 or more days length of stay", this may increase performance, while also being more relevant towards quality measurement. Trusts are increasingly coding routine readmissions as emergency readmissions with 0 days length of stay, hence the data quality may improve when excluding these episodes.

4.3.3 Small numbers

NCIP is unique in presenting risk adjustment at consultant level. Only few consultants (Table 3.1) have performed enough procedures for rates to be statistically stable. We recommend a lower bound on the number of procedures because:

- Rates have little statistical value when a consultant has performed few procedures
- Rates are unstable and may strongly deviate from the England average when a consultant has performed only few procedures
- Unstable rates may show wide variation in care, but the low statistical value implies variations have little meaning
- This may result in unwarranted stress for consultants; risk adjustment could indicate poor performance when this is not statistically meaningful.

Therefore, unlike other risk-adjustment indices in England (SHMI, HSMI, etc.), NCIP must carefully consider its boundary for presenting risk adjusted rates. After holding several workshops with experts and clinicians we advise the use of a conservative power calculation, as detailed in the results. This limits the presentation of risk-adjusted rates to a subset of consultants, but it ensures they are statistically valuable, limits the risk of putting consultants under undue pressure and maintains integrity of the tool. However, to increase coverage, NCIP may decide to balance the statistical validity and consultants available for risk adjustment, by using Approach 2, for example.

4.4 Next Steps and options for further improvement

Going forward, several steps could be taken to improve the model:

- Additional data sources may be considered to improve the performance of the predictive models. Data on medications prescribed, access to care and care delivery, further indices of socioeconomic status and health behaviours may be included. Tumour staging, an important variable often present in clinical audit data, is similarly not available.

- Multilevel logistic regression (with random effects/intercepts, hierarchical logistic regression), or empirical Bayes may be considered as an alternative method to logistic regression. This has several advantages. It considers the clustering of the data, as observations are likely similar at provider/hospital level. This may improve predictive accuracy, although the literature shows only modest improvements [31]. However, rates may be more stable for consultants with a small number of procedures [32] and therefore this method could be especially valuable to solve the small numbers problem that NCIP has [33].

- Outliers may be a result of true variation between hospitals or consultants. If this is not considered, a too high number of providers and consultants are flagged as outliers. Fewer outliers may be flagged by appropriately adjusting control limits for this overdispersion.

- A more sophisticated variable selection procedure may be considered to improve performance within individual procedures. Each model may, for example, be further optimised to include the optimal selection of variables for that procedure. Previous literature shows this may provide small improvements which may result in valuable improvements to readmission models, whose current performance is poor.

- Data cleaning steps may be further improved. For example, to account for unknown gender/sex and ages[26].

- Finally, we recommend a development log for future requests and suggestions on risk adjustment, with responses from the NCIP team, available to all consultants and providers as a page on NCIP. The current risk adjustment model is not perfect and should continuously be further developed as the literature and clinical practice changes. This page can reflect the development. For example, the SHMI is continuously developed based on provider input in a similar, publicly available development log [34]. NCIP may consider a similar page, to be presented to consultants.

In addition to model developments and before expanding to other specialties, metrics and releasing the model on the tool more work will need to be done to validate the findings with clinicians, test how best to explain and illustrate the concept on the tool and ensure buy-in. Feedback so far has been positive and risk adjustment as a concept as well as our models in particular have been well received. The ultimate gold standard for a model of this type is publication in a peer reviewed journal. This would be further enhanced by analysing variation of risk adjusted rates across specialties or procedures.

4.5 Implementation in NCIP and approach to release

A plan for release into the NCIP tool has been worked through together with the NCIP Engineering team. The current method, using logistic regression, works especially well for NCIP. First, because predictions are made at a patient level. Risk-adjusted metric can then be calculated and aggregated analogous to the current metric calculation. Second, control limits in NCIP are currently already approximate binomial. Finally, the current outcome measures used, 90-day mortality and 30-day emergency readmission, are already in the raw HES data, and therefore do not need to be constructed first, allowing risk adjustment to take place at the same stage as other metric calculation.

The steps to take to integrate our risk adjustment models are therefore as follows:

- Add additional columns to raw HES data, corresponding to the risk factors used for risk adjustment. These need to be constructed in the NCIP tables for future use.

[26] Here, it should be noted that NCIP as a whole discards patients with unknown/unspecified genders for some gender-specific procedures.

- Next, in Azure data factory, a pipeline can be added to perform risk adjustment. A R (or Python) data brick can be added to the processing to fit logistic regression models to each procedure. Outcomes will be the patient-level scores and predictions. This step is in line with the current migration of data processing in UDAL. Training of each timeframe specific model can be performed alongside processing. Alternative technical options should be explored with the data engineering team to assess performance, cost, and ease of maintenance.

- In a next step, patient-level scores of expected readmission and mortality are aggregated to consultant or provider level. Observed/Expected ratios are calculated. This processing step is similar to the current calculation of (over time) metrics in NCIP, and hence doesn't require a change to the release process.

- The aggregated rates are then presented in Funnel plots. Current funnel plot limits are similar to the one used in this report. Therefore, no changes are required to the plots[27] once metrics have been computed.

- The model runtime is low on the sample data when data columns are prespecified. Risk adjustment takes approximately 5 minutes per 3-year time period. Performance will need to be benchmarked again once the model has been deployed to the cloud and integrated into the data processing pipeline.

However, there are also several challenges to overcome to a swift implementation:

- NCIP currently allows selection of 3-year timeframes. New risk adjustment models need to be calculated for each timeframe, as models are based on 3 years of data.

    - Solution: Multiple metrics, corresponding to each timeframe, need to be constructed. This requires fitting multiple models. For example, presentation of data between April 2017 and April 2020 requires April 2017 to April 2020 risk adjustment, while presentation of data between July 2017 and July 2020 requires a new risk adjustment model, fitted to that specific data. Metrics must be separately calculated and stored.

- Rates must only be presented when consultants reach the minimum number of procedures (this minimum number, e.g., 50, for readmission, is different from the current NCIP minimum). This is particularly challenging. Metrics must only be aggregated for certain consultants.

    - Solution: This is not currently built into NCIP, and must either be changed in front-end development, or consultants must be filtered out in metric calculation to prevent aggregation for those consultants. Alternatively, the consultants shown may be controlled using a modified version of the "show peers" functionality, currently in NCIP.

- When smaller timeframes are selected (e.g., 1 year), models are still based on the 3-year timeframe, but the aggregation (indirect standardisation) of scores is based on a smaller set of data. Consequently, less consultants reach the threshold of the minimum number of procedures.

    - Solution: the exploration of hierarchical models may prevent the small numbers problem to some extent.

[27] In exploratory phases, documentation on funnel plot limits was not consistent. The current code needs to be verified; such that 95% and 99.8% limits are used (this is indicated in NCIP, but documentation indicates the calculation may be wrong; suggesting 99.9% limits are used).

- Blocker: Back-end and front-end development is required to add a “risk adjust” button allowing to “toggle” between risk adjusted and non risk adjusted funnel plots.

4.6 Scalability

The methods in this report, including the risk factors selected, are directly scalable to other specialties.

In the Urology specialty, some outcome rates were low, causing challenges with the sample sizes required per consultant for statistical validity. This is often a trade-off: specialities and procedures with low mortality rates will have larger sample sizes, and vice versa. Therefore, the challenges in the Urology specialty are likely to be similar to other specialties. For example, in neurosurgery, the current procedures have a 30-day readmission rate above 10%. For these procedures, risk adjustment may be easier and require less procedures per consultant, but sample sizes per consultant are lower. On the other hand, some specialties, such as ENT, have much lower rates (e.g., 6% 30-day readmission for ENT), but consultants perform a high volume of procedures, enabling more consultants to be risk adjusted. The current approaches for handling small numbers are therefore directly applicable to other specialties.

For some specialties, risk adjustment models might not perform well, due to low calibration or discrimination. Alternatively, associations between risk factors and outcomes may not be as expected. We therefore recommend setting a lower bound on the performance metrics, that ensures consistency across all specialities while maintaining scalability.

We have ensured that the methods and risk factors in this report can easily be used for other performance metrics, such as rate of length of stay > 1 day, for example. However, the risk factors and categorisation may need to be optimised and adapted towards each metric, although this requires only minimal efforts.

4.7 Conclusion

We performed risk adjustment of 30-day emergency readmission for 20 NCIP procedures and 90-day mortality, for 9 NCIP procedures, in the Urology specialty. We advise a 3-year timeframe, binomial control limits, and the usage of social risk factors. Because risk adjustment is performed at consultant level, we note that a minimum number of procedures performed must be reached before risk adjusted rates are presented, where we advise balancing statistical validity and consultant coverage. This is required to maintain statistical stability of the estimates.

References

[1] Centers for Medicare & Medicaid Services, “Risk adjustment in Quality Measurement: Supplemental Material to the CMS MMS Blueprint,” 2021.

[2] National Quality Forum. Risk adjustment for socioeconomic status or other sociodemographic factors. 2014. [cited 15 dec 2021].

[3] House of Commons. Report of the Independent Inquiry into the Issues raised by Paterson (HC 31). London: the Stationary Office; 2020.

[4] Odisho AY, Etzioni R, Gore JL. Beyond classic risk adjustment: socioeconomic status and hospital performance in urologic oncology surgery. Cancer. 2018 Aug 15;124(16):3372-80.

[5] Winoker JS, Paulucci DJ, Anastos H, Waingankar N, Abaza R, Eun DD, et al. Predicting complications following robot-assisted partial nephrectomy with the ACS NSQIP® universal surgical risk calculator. The Journal of Urology. 2017 Oct;198(4):803-9.

[6] British Association of Urological Surgeons. How we do risk analysis [internet]. 2021. [cited 15 dec 2021] Available from: https://www.baus.org.uk/patients/surgical_outcomes/how_we_do_risk_analysis.aspx.

[7] Daley J, Khuri SF, Henderson W, Hur K, Gibbs JO, Barbour G, et al., “Risk adjustment of the postoperative morbidity rate for the comparative assessment of the quality of surgical care: results of the National Veterans Affairs Surgical Risk Study,” *Journal of the American College of Surgeons,* vol. 185, no. 4, pp. 328-340, 1997.

[8] Sokoloff MH. Editorial Comment. J Urol. 2017 Oct; 198(4):809. doi: 10.1016/j.juro.2017.04.117.

[9] CHKS. Hospital mortality measures: Comparison of HSMR, SHMI and RAMI [internet]. 2018. [cited 15 dec 2021]..

[10] Symons S, Biyani CS, Bhargava S, Irvine HC, Ellingham J, Cartledge J, Lloyd SN, Joyce AD, Browning AJ. Challenge of percutaneous nephrolithotomy in patients with spinal neuropathy. International journal of Urology. 2006 Jul;13(7):874-9.

[11] Clinical Indicators Team, NHS Digital. Mean depth of coding for provider spells with an elective admission method: Indicator specification. 2019; 1.3.

[12] Sundararajan V, Henderson T, Perry C, Muggivan A, Quan H, Ghali WA. New ICD-10 version of the Charlson comorbidity index predicted in-hospital mortality. Journal of clinical epidemiology. 2004 Dec 1;57(12):1288-94.

[13] Gilbert T, Neuburger J, Kraindler J, Keeble E, Smith P, Ariti C, Arora S, Street A, Parker S, Roberts HC, Bardsley M. Development and validation of a Hospital Frailty Risk Score focusing on older people in acute care settings using electronic hospital.

[14] Government Statistical Service. The 2011 Rural-Urban Classification for Output Areas in England . 2011.

[15] NHS Digital. General Practice Workforce, March 2018. [internet]. 2018.

[16] Bottle A, Gaudoin R, Goudie R, Jones S, Aylin P., Can valid and practical risk-prediction or casemix adjustment models, including adjustment for comorbidity, be generated from English hospital administrative data (Hospital Episode Statistics)? A national observational study., , 2015.

[17] Jen MH, Bottle A, Kirkwood G, Johnston R, Aylin P. The performance of automated case-mix adjustment regression model building methods in a health outcome prediction setting. Health care management science. 2011 Sep 1;14(3):267-78.

[18] Department of Health. Patient Reported Outcome Measures (PROMs) in England: A Methodology for Identifying Potential Outliers. 2012.

[19] Spiegelhalter DJ. Funnel plots for comparing institutional performance. Statistics in medicine. 2005 Apr 30;24(8):1185-202.

[20] Campbell MJ, Jacques RM, Fotheringham J, Pearson T, Maheswaran R, Nicholl J. An evaluation of the summary hospital mortality index. The University of Sheffield, School of Health and Related Research. 2011;81.

[21] Department of Health, Healthcare Quality Improvement Partnership. Detection and management of outliers: Guidance prepared by National Clinical Audit Advisory Group . 2011.

[22] Verburg IW, Holman R, Peek N, Abu-Hanna A, de Keizer NF. Guidelines on constructing funnel plots for quality indicators: a case study on mortality in intensive care unit patients. Statistical methods in medical research. 2018 Nov;27(11):3350-66.

[23] Austin PC, Reeves MJ. Effect of provider volume on the accuracy of hospital report cards: a Monte Carlo study. Circulation: Cardiovascular Quality and Outcomes. 2014 Mar;7(2):299-305.

[24] Dimick JB, Welch HG, Birkmeyer JD. Surgical mortality as an indicator of hospital quality: the problem with small sample size. Jama. 2004 Aug 18;292(7):847-51.

[25] National Joint Registry Centre. NJR Adopted Statistical Methodology for Potential Outlier Identification. 2008.

[26] Austin PC, Reeves MJ. The relationship between the C-statistic of a risk-adjustment model and the accuracy of hospital report cards: a Monte Carlo Study. Medical care. 2013 Mar;51(3):275.

[27] Artetxe A, Beristain A, Grana M. Predictive models for hospital readmission risk: A systematic review of methods. Computer methods and programs in biomedicine. 2018 Oct 1;164:49-64.

[28] Henderson WG, Khuri, SF. Risk Adjustment. In: Penson DF, Wei, JT, editors. Clinical Research for Surgeons. Totowa, Nj: Humana Press Inc; 2006. pp 105-119.

[29] NHS Digital. SHMI statistical model data [internet]. 2020. [cited 15 dec 2021].

[30] N Krishnan, B Li, BL Jacobs, SN Ambani, T Borza, C He, BK Hollenbeck, T Morgan, KS Hafez, AZ Weizer, JS Montgomery., “The fate of radical cystectomy patients after

hospital discharge: understanding the black box of the pre-readmission interval," *European Urology Focus,* vol. 4, no. 5, pp. 711-7, 2018.

[31] Cohen ME, Dimick JB, Bilimoria KY, Ko CY, Richards K, Hall BL., "Risk adjustment in the American College of Surgeons National Surgical Quality Improvement Program: a comparison of logistic versus hierarchical modeling," *. Journal of the American College of Surgeons,* vol. 209, no. 6, pp. 687-93, 2009.

[32] Clark DE, Hannan EL, Wu C. Predicting risk-adjusted mortality for trauma patients: logistic versus multilevel logistic models. Journal of the American College of Surgeons. 2010 Aug 1;211(2):224-31.

[33] Arling G, Lewis T, Kane RL, Mueller C, Flood S. Improving quality assessment through multilevel modeling: the case of nursing home compare. Health services research. 2007 Jun;42(3p1):1177-99.

[34] NHS Digital, Analytical Services Team. Summary Hospital-level Mortality Indicator (SHMI): Methodology development log. 2021.

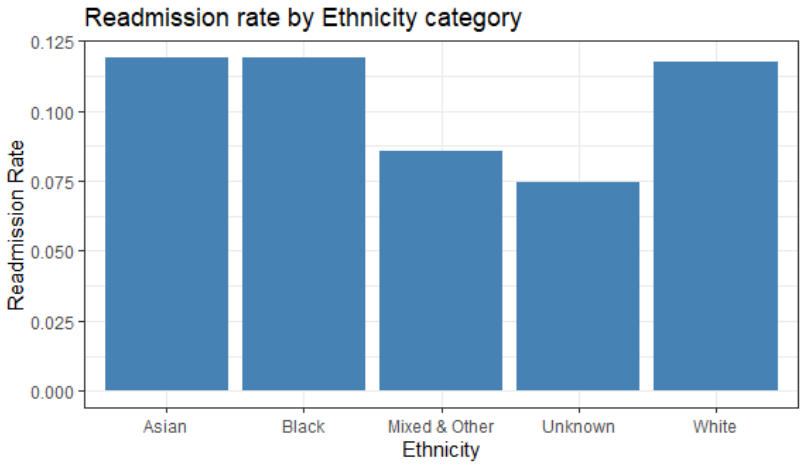
Appendix A. Methods

Appendix A1: Data Cleaning

In data cleaning of HES data, we have taken the following steps:

- Only patients with male and female gender were included.[28] Further investigation is required to prevent this (see Section 4.4, Next Steps).
- Patients with unknown frailty scores or IMD scores were excluded.
- Only patients under 120 years of age were included. Older patients were implausible, and unknown ages (coded as 999) were excluded.
- Ethnicity categories were grouped into Asian, Black, White, Other/mixed, and unknown categories. Across all Urology procedures, admission rates were lower for mixed & other, and unknown ethnicities (**Figure A.1**). Individuals with unknown ethnicities are likely to be younger, more urban, and are more likely to have mixed and other ethnicities, hence their readmission rates are lower. They are a distinct group, and therefore treated as a separate category.

Figure A.1: Readmission rate by ethnicity category



- In models comparing the inclusion of GP/Patient ratios, patients with unknown GP practices or unavailable GP staffing data were excluded.
- The urbanicity (rural-urban indicator) was grouped into Urban (city and town) and rural (town and less populated areas). Unknown urbanicity statuses were excluded.
- We removed further outliers: procedures with a length of stay of more than 1000 days or less than 0, patients with more than 34 emergency admissions in the previous year.

Appendix A2: Risk factors

A squared age term was included to account for non-linear effects. The use of 5, or 10-year age bands, or other methods, show a higher risk of overfitting when using small, consultant-level, procedure-specific datasets. We provide a sensitivity analysis for alternative methods of age categorisation in Appendix B.

Clinical risk factors include the continuous Charlson comorbidity score. This is in contrast with other risk adjustment models, such as the SHMI, which use categories of Charlson

[28] The inclusion of a separate category for unknown genders results in overfitting in procedure-specific models; the procedure specific sample sizes are small, and unknown genders have a low frequency. Inclusion in either male or female genders appears inappropriate. Currently, NCIP already filters these genders in gender-specific dashboards.

scores. First, preliminary analyses show a risk of overfitting when using individual Charlson comorbidities in models. Moreover, the Charlson score is validated for longer term mortality outcomes, but not readmission or 90-day mortality. We provide a sensitivity analysis in Appendix B studying the use of individual Charlson comorbidities, quantiles and clinical grouping of Charlson scores.

Appendix A3: Performance metrics for logistic regression models

First, the c-statistic, or area under the receiver operator curve (ROC), is used to assess discrimination in the predicted probabilities. In the sample, there are two groups: “Readmission” and “No readmission”. For each patient, we predict a probability of readmission. If we choose one random patient with readmission and one random patient without readmission, we aim to have a higher probability of readmission predicted for the patient with a readmission than the patient without a readmission. The c-statistic represents the probability that this happens. The c-statistic ranges between 0 and 1. If we were to predict scores randomly, the c-statistic would be 0.5; when picking a patient with readmission and without readmission, a random prediction would correctly predict 50% of the time. [20]

Second, the Brier score is used to assess calibration. If most readmissions are assigned a high score, and most non-readmissions are assigned a low score, model calibration is good, and brier scores are low. The brier score is therefore a measure of how close the predicted scores are to the observed status, “Readmission”, or “No readmission” [1]. Brier scores range between 0 and 1, but a random model achieves a brier score of 0.25. Lower scores indicate better model calibration.

Further, calibration charts can be presented to assess the calibration of the model. This chart divides up the models’ predicted scores into 10 categories, deciles. In the lowest deciles, where the model predicted scores are low, we aim to observe a low number of readmissions. In the highest deciles, where model predicted scores are high, we aim to observe the highest number of readmissions. The calibration chart compares, in each risk decile, the observed readmissions and predicted number of readmissions. We aim to have similar predicted readmissions and observed readmissions in each category if calibration is good [1].

A formal test comparing observed and expected number of readmissions in each group is the Hosmer-Lemeshow (HL) test. This test is frequently used for assessing calibration of risk adjustment models [1]. The HL-test formally tests if the observed and expected number of readmissions are as expected under statistical uncertainty. If the p-value of this test is lower than 0.05, the null hypothesis is rejected, and model calibration is considered poor. However, it is advisable to follow calibration charts, as the HL-test rejects frequently; even a small inconsistency can result in low p-values. Inspection of the calibration charts is therefore advisable if HL-tests show poor calibration; calibration may not be poor.

Appendix A4: Calculation of performance metrics

If all data is used towards fitting the model, the selected risk factors or variable categorisation may be optimed specifically to the dataset and timeframe studied. Future predictions may then not be as accurate as expected from the historical data, as models were “overfitted” to follow the “quirks” in the historical data.

The results presented in Section 3 are therefore based on unseen data. This test set was 2020/2021 data, while all other models used 3-year cross-validation for 2017/2018, 2018/2019 and 2019/2020 financial years (April 2017 – March 2020).

All model selection stages and sensitivity analyses used cross-validated c-statistics over the 3-year dataset. Cross validation is a statistical procedure, in which the model is fitted multiple times to a smaller dataset. Subsequent performance is then measured on a specific part of the data that the model has not seen. This allows us to evaluate the model performance without the model being overfitted to a specific dataset. 5-fold cross validation was used, implying that models were fitted 5 times, in which 80% of the data was used for testing, and 20% of the data was used for performance measurement.

Appendix B. Results

Appendix B1: Sample sizes per procedure

Table B.1: Sample sizes, readmission rates and mortality rates by NCIP procedure

NCIP Procedure	Sample	Readmissions (30 day)	Readmission rate	Mortality (90 days)	Mortality rate
Circumcision | age 17+ | elective	56,563	2,425	4.29%	102	0.18%
Cystectomy for benign disease | age 17+ | elective	522	125	23.95%	5	0.96%
Cystectomy for malignant neoplasms of the bladder | age 17+ | elective	5,601	1,380	24.64%	196	3.50%
Endoscopic resection of lesion of bladder (TURBT) | age 17+ | elective	69,675	6,292	9.03%	1,471	2.11%
Extracorporeal shock wave lithotripsy of calculus (ESWL) | age 17+ | elective	58,740	2,642	4.50%	66	0.11%
Hydrocoele | age 17+ | elective	17,282	1,747	10.11%	34	0.20%
Injection of bulking agents | age 17+ | elective	8,694	175	2.01%	12	0.14%
Insertion of artificial sphincter | age 17+ | elective	1,053	64	6.08%	2	0.19%
Insertion of penile prosthesis | age 17+ | elective	1,442	121	8.39%	2	0.14%
Insertion of ureteric stent | age 17+	70,069	9,274	13.24%	2,835	4.05%
Litholapaxy for bladder stones | age 17+ | elective	11,876	985	8.29%	128	1.08%
Male bladder outflow obstruction surgery | age 17+ | elective	73,613	6,720	9.13%	380	0.52%
Nephrectomy for benign disease | age 17+ | elective	4,257	532	12.50%	35	0.82%
Nephrectomy for cancer | age 17+ | elective	17,022	1,788	10.50%	161	0.95%
Nephrostomy | age 17+	11,158	2,845	25.50%	2,595	23.26%
Nephroureterectomy for cancer | age 17+ | elective	3,241	424	13.08%	58	1.79%
Percutaneous nephrolithotomy (PCNL) | age 17+ | elective	33,766	3,338	9.89%	149	0.44%
Peyronie's surgery | age 17+ | elective	1,978	54	2.73%	0	0%
Prostate biopsy | age 17+ | elective	110,986	3,656	3.29%	198	0.18%
Prostatectomy for cancer | age 17+ | elective	26,118	2,449	9.38%	24	0.09%
Sacral nerve stimulation for urinary conditions | age 17+ | elective	1,990	91	4.57%	0	0%
Ureteroscopy | age 17+	62,202	6,554	10.54%	731	1.18%
Urethral dilatation, female | age 17+ | elective	26,733	691	2.58%	38	0.14%
Urethral dilatation, male | age 17+ | elective	23,711	1,115	4.70%	152	0.64%
Urethroplasty for stricture, male | age 17+ | elective	1,843	162	8.79%	0	0%
Urethrotomy | age 17+ | elective	14,332	888	6.20%	70	0.49%
Vaginal fistula | age 17+ | elective	682	109	15.98%	1	0.15%
Urology (all)	715,149	56,646	7.92%	9,445	1.32%

Note: Sample sizes, readmission rate and mortality rate for 2017-2020 data are presented, by NCIP procedure, before data cleaning.

Appendix B2: Descriptive Statistics

Table B.2: Descriptive statistics, 30-day readmission, by readmission status and overall.

	No readmission (N=790224)	**Readmission** (N=69757)	**Overall** (N=859981)
Age (years)			
Median [IQR]	65.0 [53.0, 73.0]	67.0 [53.0, 76.0]	65.0 [53.0, 74.0]
Sex			
Female	177,535 (22.5%)	18,245 (26.2%)	15,780 (22.8%)
Male	612,689 (77.5%)	51,512 (73.8%)	664,201 (77.2%)
Ethnicity			
Asian or Asian British	28,832 (3.6%)	2,869 (4.1%)	31,701 (3.7%)
Black or Black British	14,778 (1.9%)	1,201 (1.7%)	15,979 (1.9%)
Mixed, other	106,271 (13.4%)	6,685 (9.6%)	112,956 (13.1%)
Unknown	31,948 (4.0%)	1,766 (2.5%)	33,714 (3.9%)
White	608,395 (77.0%)	57,236 (82.1%)	665,631 (77.4%)
Quarter of calendar year			
1	196,990 (24.9%)	16,430 (23.6%)	213,420 (24.8%)
2	181,447 (23.0%)	16,343 (23.4%)	197,790 (23.0%)
3	203,836 (25.8%)	18,918 (27.1%)	222,754 (25.9%)
4	207,951 (26.3%)	18,066 (25.9%)	226,017 (26.3%)
Emergency adm.			
Emergency	731,679 (92.6%)	56,235 (80.6%)	787,914 (91.6%)
No emergency	58,545 (7.4%)	13,522 (19.4%)	72,067 (8.4%)
Charlson Score			
Median [IQR]	1.00 [0, 2.00]	1.00 [0, 3.00]	1.00 [0, 2.00]
N. emergency adm.			
Median [IQR]	0 [0, 1.00]	0 [0, 2.00]	0 [0, 1.00]
HFRS Frailty Band			
Mild	167,799 (21.2%)	12,521 (17.9%)	180,320 (21.0%)
Moderate	106,579 (13.5%)	13,876 (19.9%)	120,455 (14.0%)
None	488,896 (61.9%)	36,948 (53.0%)	525,844 (61.1%)
Severe	26,950 (3.4%)	6,412 (9.2%)	33,362 (3.9%)
Palliative			
Yes	787,463 (99.7%)	69,010 (98.9%)	856,473 (99.6%)
No	2,761 (0.3%)	747 (1.1%)	3,508 (0.4%)
Obesity			
Yes	703,719 (89.1%)	61,568 (88.3%)	765,287 (89.0%)
No	86,505 (10.9%)	8,189 (11.7%)	94,694 (11.0%)
Alcohol			
Yes	786,450 (99.5%)	69,353 (99.4%)	855,803 (99.5%)
No	3,774 (0.5%)	404 (0.6%)	4,178 (0.5%)
Tobacco			
Yes	718,451 (90.9%)	62,674 (89.8%)	781,125 (90.8%)
No	71,773 (9.1%)	7083 (10.2%)	78,856 (9.2%)
IMD deprivation score			
Median [IQR]	16.2 [9.28, 27.5]	17.1 [9.72, 29.3]	16.3 [9.31, 27.7]
Urbanicity			
Rural	179,313 (22.7%)	15,337 (22.0%)	194,650 (22.6%)
Urban	610,911 (77.3%)	54,420 (78.0%)	665,331 (77.4%)
GP/patient ratio			
Median [IQR] (/10000)	5.03 [4.06, 6.14]	5.00 [4.02, 6.12]	5.02 [4.05, 6.14]

Note: Descriptive characteristics, by variable and readmission status. Normally distributed variables are presented as Mean (Std. Deviation), Non-normally distributed variables are presented as Median (Interquartile Range) (IQR). Categorical variables are presented as N (%), by category. 2017-2021 data (includes both training and test sets)

Table B.3: Descriptive statistics, 90-day mortality, by mortality status and overall.

	No mortality (N=847504)	**Mortality** (N=12477)	**Overall** (N=859981)
Age (years)			
Median [IQR]	65.0 [53.0, 73.0]	76.0 [67.0, 82.0]	65.0 [53.0, 74.0]
Sex			
Female	191,117 (22.6%)	4,663 (37.4%)	195,780 (22.8%)
Male	656,387 (77.4%)	7,814 (62.6%)	664,201 (77.2%)
Ethnicity			
Asian or Asian British	31,436 (3.7%)	265 (2.1%)	31,701 (3.7%)
Black or Black British	15,794 (1.9%)	185 (1.5%)	15,979 (1.9%)
Mixed, other	111,985 (13.2%)	971 (7.8%)	112,956 (13.1%)
Unknown	33,427 (3.9%)	287 (2.3%)	33,714 (3.9%)
White	654,862 (77.3%)	10,769 (86.3%)	665,631 (77.4%)
Quarter of calendar year			
1	210,361 (24.8%)	3,059 (24.5%)	213,420 (24.8%)
2	195,013 (23.0%)	2,777 (22.3%)	197,790 (23.0%)
3	219,518 (25.9%)	3,236 (25.9%)	222,754 (25.9%)
4	222,612 (26.3%)	3,405 (27.3%)	226,017 (26.3%)
Emergency adm.			
Emergency	781,492 (92.2%)	6,422 (51.5%)	787,914 (91.6%)
No emergency	66,012 (7.8%)	6,055 (48.5%)	72,067 (8.4%)
Charlson Score			
Median [IQR]	1.00 [0, 2.00]	4.00 [2.00, 8.00]	1.00 [0, 2.00]
N. emergency adm.			
Median [IQR]	0 [0, 1.00]	2.00 [0, 4.00]	0 [0, 1.00]
HFRS Frailty Band			
Mild	178,776 (21.1%)	1,544 (12.4%)	180,320 (21.0%)
Moderate	116,266 (13.7%)	4,189 (33.6%)	120,455 (14.0%)
None	522,908 (61.7%)	2,936 (23.5%)	525,844 (61.1%)
Severe	29,554 (3.5%)	3,808 (30.5%)	33,362 (3.9%)
Palliative			
Yes	846,130 (99.8%)	10,343 (82.9%)	856,473 (99.6%)
No	1,374 (0.2%)	2,134 (17.1%)	3,508 (0.4%)
Obesity			
Yes	753,618 (88.9%)	11,669 (93.5%)	765,287 (89.0%)
No	93,886 (11.1%)	808 (6.5%)	94,694 (11.0%)
Alcohol			
Yes	843,415 (99.5%)	12,388 (99.3%)	855,803 (99.5%)
No	4,089 (0.5%)	89 (0.7%)	4,178 (0.5%)
Tobacco			
Yes	769,602 (90.8%)	11,523 (92.4%)	781,125 (90.8%)
No	77,902 (9.2%)	954 (7.6%)	78,856 (9.2%)
IMD deprivation score			
Median [IQR]	16.3 [9.31, 27.7]	16.5 [9.64, 28.0]	16.3 [9.31, 27.7]
Urbanicity			
Rural	191,733 (22.6%)	2,917 (23.4%)	194,650 (22.6%)
Urban	655,771 (77.4%)	9,560 (76.6%)	665,331 (77.4%)
GP/patient ratio			
Median [IQR] (/10000)	5.02 [4.05, 6.13]	5.04 [4.07, 6.14]	5.02 [4.05, 6.14]

Note: Descriptive characteristics, by variable and mortality status. Normally distributed variables are presented as Mean (Std. Deviation), Non-normally distributed variables are presented as Median (Interquartile Range) (IQR). Categorical variables are presented as N (%), by category. 2017-2021 data (includes both training and test sets).

Appendix B3: Minimum sample size per consultant/provider required for presentation

Table B.4: Minimum sample sizes, following power calculations, for readmission or mortality risk adjustment

Procedure	Rate	Sample size (per unit of analysis)	Sample size used (rounded to nearest 50)
Readmission (Urology)	0.0792	43	50
Mortality (Urology)	0.0132	273	250
Mortality (Cystectomy)	0.0350	106	100
Mortality (Nephrostomy)	0.2326	11	50
Mortality (Insertion of ureteric stent)	0.0405	91	100
Mortality (Nephroureterectomy)	0.0179	214	200
Mortality (TURBT)	0.0211	180	200

Note: Rate and minimum sample size required are presented by mortality, readmission, and procedures with a high mortality rate (above 1.5%). Minimum sample sizes are based on 80% power to detect a 2.5 times increase in the risk adjusted rate, compared to the observed England rate, at 5% significance level.

Appendix B.4. Model performance by NCIP procedure

Table B.5: Model performance, consultants presented and consultants outside control limits, by NCIP procedure (30-day readmission)

NCIP Procedures – 30-day readmission risk adjustment	C-statistic	Brier score	HL-test (p-value)	Consultants suitable for presentation	Consultants below lower limit	Consultants above upper limit
Insertion of ureteric stent | age 17+	0.648172	0.112569	0.079459	470	5	6
Ureteroscopy | age 17+	0.605328	0.093744	0.380123	397	10	6
Endoscopic resection of lesion of bladder (TURBT) | age 17+ | elective	0.640601	0.082375	0.006079	524	4	0
Percutaneous nephrolithotomy (PCNL) | age 17+ | elective	0.580217	0.088579	0.776208	214	5	3
Prostate biopsy | age 17+ | elective	0.615153	0.024157	0.13194	468	12	3
Male bladder outflow obstruction surgery | age 17+ | elective	0.604192	0.082419	0.036026	551	11	3
Circumcision | age 17+ | elective	0.612007	0.046605	0.215993	436	14	0
Hydrocoele | age 17+ | elective	0.623814	0.091779	0.590229	20	0	0
Extracorporeal shock wave lithotripsy of calculus (ESWL) | age 17+ | elective	0.613944	0.043614	0.145709	226	9	2
Urethrotomy | age 17+ | elective	0.646386	0.054627	0.418508	18	1	0
Prostatectomy for cancer | age 17+ | elective	0.590234	0.079221	0.009512	135	13	7
Nephrectomy for cancer | age 17+ | elective	0.599765	0.090281	0.237207	116	2	0
Cystectomy for malignant neoplasms of the bladder | age 17+ | elective	0.550554	0.165318	0.700673	40	3	1
Urethral dilatation, male | age 17+ | elective	0.684297	0.047735	0.003194	68	3	0
Urethral dilatation, female | age 17+ | elective	0.654248	0.032156	0.152873	128	2	0
Injection of bulking agents | age 17+ | elective	0.621458	0.028853	0.638548	38	1	0
Peyronie's surgery | age 17+ | elective	0.578125	0.042145	0.055766	7	0	0

Sacral nerve stimulation for urinary conditions | age 17+ | elective	0.810084	0.028307	0.9099	8	0	0
Insertion of penile prosthesis | age 17+ | elective	0.523022	0.053301	0.743376	6	0	0
Urethroplasty for stricture, male | age 17+ | elective	0.46441	0.080226	0.459794	11	0	0

Note: C-statistic (0 - 1, higher is better), Brier score (0 - 1, lower is better), HL-test p-value (0 - 1, above 0.05 indicates good calibration), Consultants suitable for presentation (following Approach 3 minimum sample sizes) and consultants outside 99.8% "alarm" control limits are presented by NCIP procedure, for 30-day readmission.

Table B.6: Model performance, consultants presented and consultants outside control limits, by NCIP procedure (90-day mortality)

NCIP Procedures – 30-day readmission risk adjustment	C-statistic	Brier score	HL-test (p-value)	Consultants suitable for presentation	Consultants below lower limit	Consultants above upper limit
Endoscopic resection of lesion of bladder (TURBT) | age 17+ | elective	0.7666	0.0220	0.0441	25	0	0
Extracorporeal shock wave lithotripsy of calculus (ESWL) | age 17+ | elective	0.8050	0.0020	0.5970	66	2	0
Insertion of ureteric stent | age 17+	0.8643	0.0356	0.0000	138	1	0
Male bladder outflow obstruction surgery | age 17+ | elective	0.8093	0.0055	0.3479	24	1	0
Percutaneous nephrolithotomy (PCNL) | age 17+ | elective	0.8354	0.0068	0.3128	7	0	0
Prostate biopsy | age 17+ | elective	0.7305	0.0024	0.3801	121	4	0
Prostatectomy for cancer | age 17+ | elective	0.4588	0.0009	0.6760	23	1	0
Ureteroscopy | age 17+	0.8737	0.0128	0.0680	15	0	0
Cystectomy for malignant neoplasms of the bladder | age 17+ | elective	0.7414	0.0249	0.9991	6	0	0

Note: C-statistic (0 - 1, higher is better), Brier score (0 - 1, lower is better), HL-test p-value (0 - 1, above 0.05 indicates good calibration), Consultants suitable for presentation (following Approach 3 minimum sample sizes) and consultants outside 99.8% "alarm" control limits are presented by NCIP procedure, for 30-day readmission.

Appendix B.5. Risk adjustments (SMRs, SRRs) by NCIP procedure

Figure B.1: Boxplot of risk adjustment SRRs, by NCIP procedure

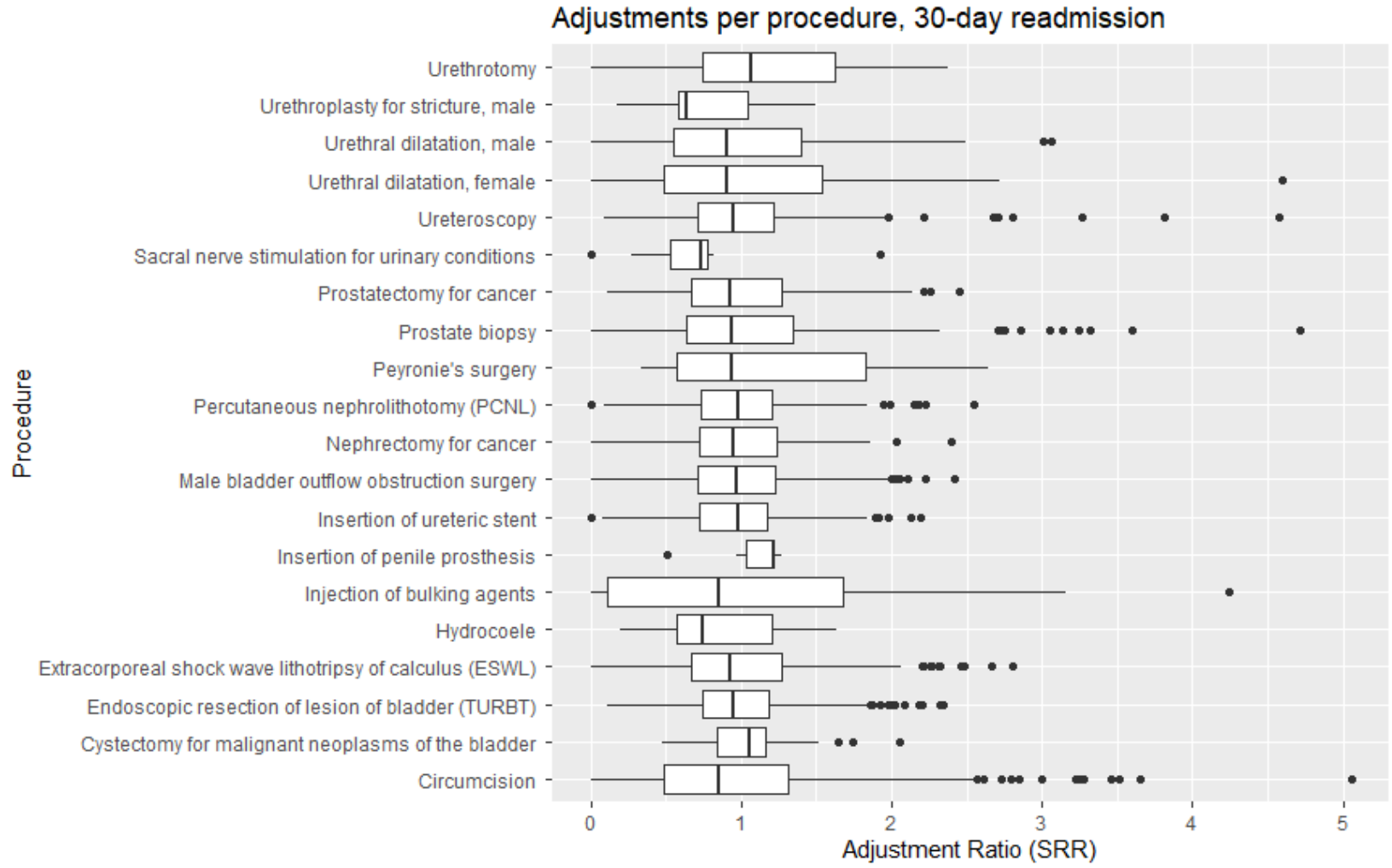
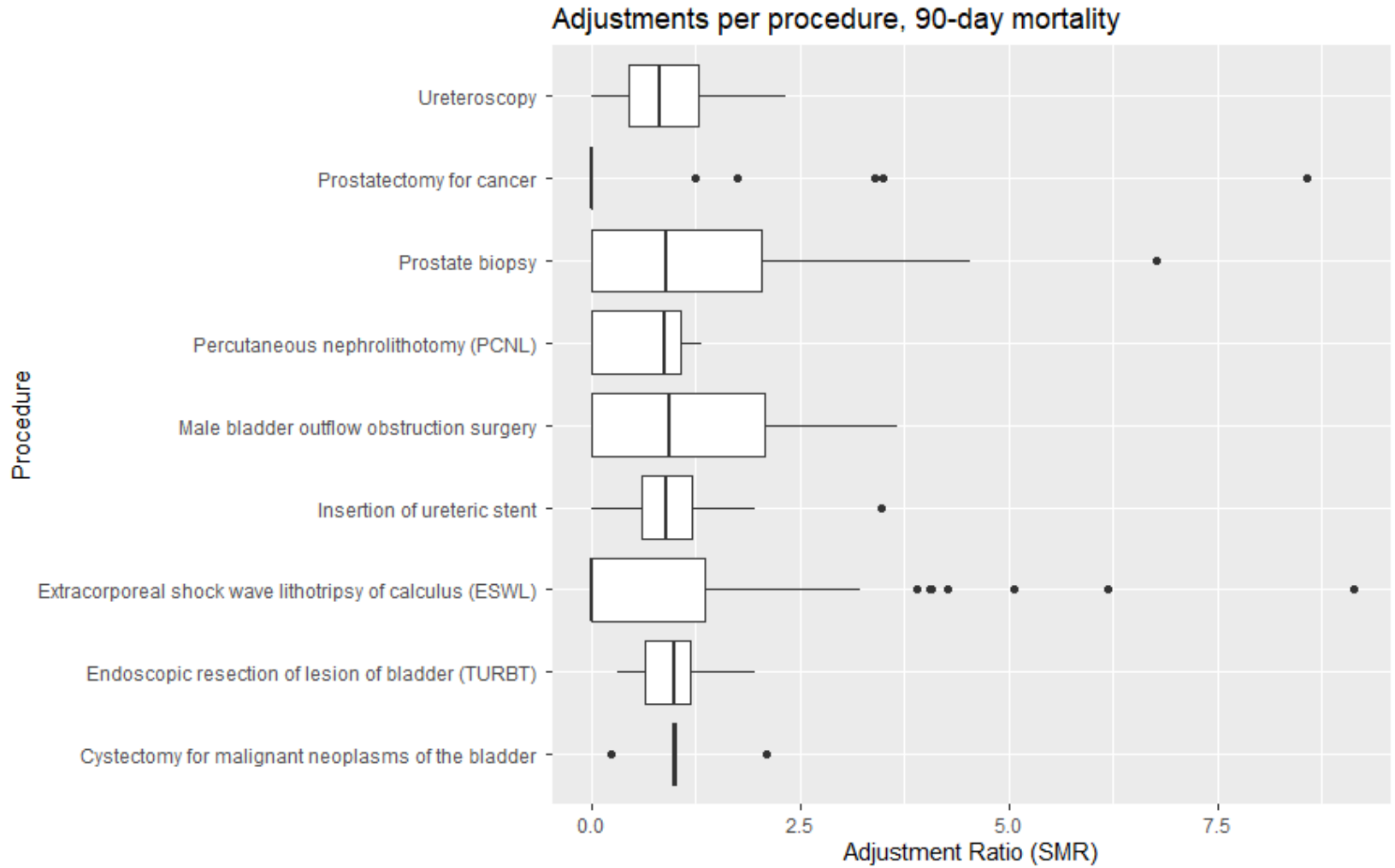


Figure B.2: Boxplot of risk adjustment SMRs, by NCIP procedure



Appendix B.6: Sensitivity Analyses

In a sensitivity analysis, alternative categorisations for the age variable were considered. Here, we compare the usage of age and squared age (final model), age quantiles, and 10-year age bands (18-29,30-39,40-49,…,90+). We present the performance over all procedures by presenting the median, minimum, maximum and mean c-statistics achieved in all procedures, obtained from 5-fold cross-validation within the sample. There appears to be no increase in performance when using 10-year age band or quartiles.

Table B.7: Sensitivity Analysis: Alternative categorisations of the age variable (30-day readmission)

C-statistic	Final model	Age: 10-year bands	Age: quartiles
Median	0.599	0.598	0.599
Min.	0.502	0.506	0.515
Max.	0.715	0.700	0.702
Mean	0.598	0.597	0.597
C>0.6	10 (50%)	10 (50%)	10 (50%)
C>0.7	1 (5%)	1 (5%)	1 (5%)
C>0.8	0 (0%)	0 (0%)	0 (0%)
C>0.9	0 (0%)	0 (0%)	0 (0%)

Note: Model performance for the minimum performing model and maximum performing model out of all procedure-specific models are presented. The median and mean performance are presented over all procedures. Following the literature, a c-statistic above 0.9 is widely considered as "excellent", a performance between 0.8 - 0.9 is considered good, a performance between 0.6 - 0.8 is considered fair, and a performance below 0.6 is generally considered poor.

Alternative Charlson categorisations were similarly considered, comparing mean, minimum, maximum, and median 5-fold cross-validated c-statistics over all procedures. We compare the usage of the continuous Charlson score, individual Charlson comorbidities as indicator variables, quantiles, and the SHMI categorisation: '0', '0-5', 'more than 5'.

Table B.8: Sensitivity Analysis: Alternative categorisations of the Charlson score variable (30-day readmission)

	Charlson Continuous	Charlson comorbidities	Charlson SHMI	Charlson Quantiles
Median	0.599	0.599	0.597	0.598
Min.	0.502	0.46	0.526	0.498
Max.	0.715	0.707	0.712	0.712
Mean	0.598	0.598	0.603	0.601
C>0.6	10 (50%)	8 (40%)	10 (50%)	10 (50%)
C>0.7	1 (5%)	1 (5%)	1 (5%)	1 (5%)
C>0.8	0 (0%)	0 (0%)	0 (0%)	0 (0%)
C>0.9	0 (0%)	0 (0%)	0 (0%)	0 (0%)

Note: Model performance for the minimum performing model and maximum performing model out of all procedure-specific models are presented. The median and mean performance are presented over all procedures. Following the literature, a c-statistic above 0.9 is widely considered as "excellent", a performance between 0.8 - 0.9 is considered good, a performance between 0.6 - 0.8 is considered fair, and a performance below 0.6 is generally considered poor.

For mortality, a similar pattern is observed. There is no improvement in median or mean performance when using an alternative categorisation for the Charlson score. Hence, we maintain the usage of continuous Charlson scores.

Table B.9: Sensitivity Analysis: Alternative categorisations of the Charlson score variable (90-day mortality)

	Charlson Continuous	Charlson comorbidities	Charlson SHMI	Charlson Quantiles
Median	0.811	0.804	0.799	0.804
Min.	0.479	0.354	0.446	0.354
Max.	0.906	0.903	0.900	0.903
Mean	0.777	0.756	0.765	0.756
C>0.6	8 (89%)	8 (89%)	8 (89%)	8 (89%)
C>0.7	8 (89%)	7 (78%)	7 (78%)	7 (78%)
C>0.8	5 (56%)	5 (56%)	4 (44%)	4 (44%)
C>0.9	1 (11%)	1 (11%)	0 (0%)	0 (0%)

Note: Model performance for the minimum performing model and maximum performing model out of all procedure-specific models are presented. The median and mean performance are presented over all procedures. Following the literature, a c-statistic above 0.9 is widely considered as "excellent", a performance between 0.8 - 0.9 is considered good, a performance between 0.6 - 0.8 is considered fair, and a performance below 0.6 is generally considered poor.

Finally, a sensitivity analysis was conducted to examine the inclusion of additional variables. Risk factors were based on ICD-10 codes for "nicotine dependence" (F17), "alcohol dependence" (F10) and "obesity" (E66). The GP to patient ratio was obtained from the GP practice of a patient, linked to national data for the general practice workforce as of March 2018 (the time-period studied). The results showed no improvement when including any of the potential variables.

Table B.10: Sensitivity Analysis: Inclusion of additional variables (30-day readmission)

	Base model	+Tobacco	+Alcohol	+Obesity	+GP/Patient ratio
Median	0.599	0.599	0.599	0.599	0.597
Min.	0.502	0.500	0.507	0.494	0.522
Max.	0.715	0.714	0.713	0.715	0.718
Mean	0.598	0.598	0.598	0.597	0.601
C>0.6	10 (50%)	10 (50%)	10 (50%)	10 (50%)	9 (45%)
C>0.7	1 (5%)	1 (5%)	1 (5%)	1 5%)	1 (5%)
C>0.8	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
C>0.9	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)

Note: Model performance for the minimum performing model and maximum performing model out of all procedure-specific models are presented. The median and mean performance are presented over all procedures. Following the literature, a c-statistic above 0.9 is widely considered as "excellent", a performance between 0.8 - 0.9 is considered good, a performance between 0.6 - 0.8 is considered fair, and a performance below 0.6 is generally considered poor.

In models of mortality, palliative care (Z515, Z518) improved the cross-validated performance, increasing the mean and median c-statistic, although to a small extent. Moreover, it improved the performance in specific procedures. Therefore, we included palliative care in the final model.

Table B.11: Sensitivity Analysis: Inclusion of additional variables (90-day mortality)

	Base model	Tobacco	Alcohol	Obesity	GP/Patient ratio	Palliative Care
Median	0.809	0.809	0.807	0.809	0.797	0.811
Min.	0.492	0.482	0.496	0.494	0.566	0.479
Max.	0.901	0.901	0.900	0.901	0.901	0.906
Mean	0.775	0.773	0.775	0.776	0.782	0.777
C>0.6	8 (89%)	8 (89%)	8 (89%)	8 (89%)	8 (89%)	8 (89%)
C>0.7	7 (78%)	7 (78%)	7 (78%)	7 (78%)	7 (78%)	8 (89%)
C>0.8	5 (56%)	5 (56%)	5 (56%)	5 (56%)	4 (44%)	5 (44%)
C>0.9	1 (11%)	1 (11%)	1 (11%)	1 (11%)	1 (11%)	1 (11%)

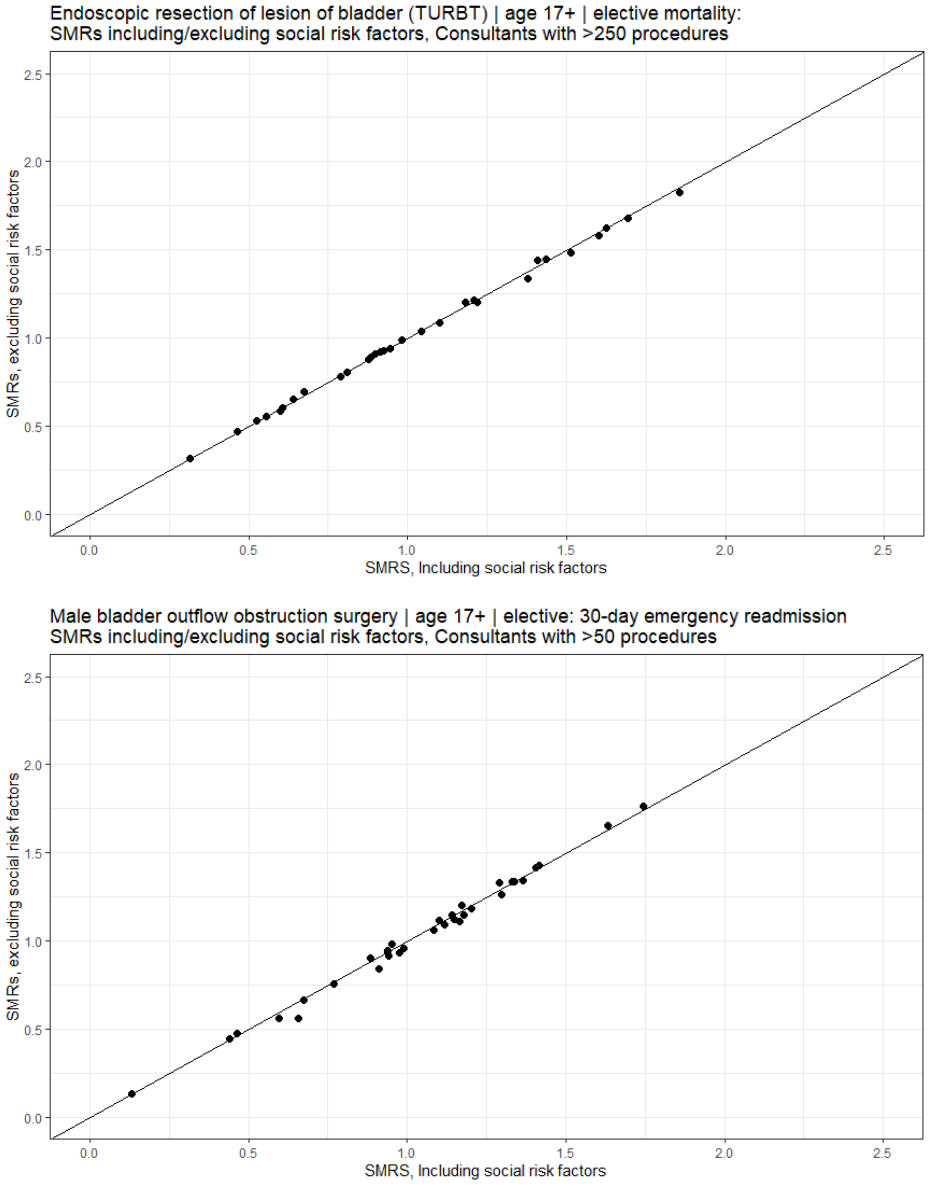
Note: Model performance for the minimum performing model and maximum performing model out of all procedure-specific models are presented. The median and mean performance are presented over all procedures. Following the literature, a c-statistic above 0.9 is widely considered as "excellent", a performance between 0.8 - 0.9 is considered good, a performance between 0.6 - 0.8 is considered fair, and a performance below 0.6 is generally considered poor.

Appendix B.7. Social risk factors

In all procedures, for emergency readmission, there was a high correlation between risk-adjusted rates (average: 0.991, min: 0.946, max: 0.999), as well as consultant rankings (Kendall's tau, average: 0.937, min: 0.714, max: 1.0). When considering outliers, in 17 out of 25 procedures, no consultants changed their outlier status between inclusion and exclusion of social risk factors. For mortality, similar results were found. Correlation between risk-adjusted rates was equally high in all procedures (mean: 0.997, mean for ranking: 0.986).

To illustrate this, the risk adjusted rates with and without social risk factors were plotted against each other in Figure B.3, for TURBT (mortality) and Male bladder outflow obstruction surgery (readmission). These plots compare the degree of risk adjustment for individual consultants with social risk factors (X-axis) and without social risk factors (Y-axis). As most consultants in this plot remain close to the diagonal line, the risk adjustments with and without social risk factors show great similarity.

Figure B.3: Comparison of SMRs, SRRs, with and without adjustment for social risk factors

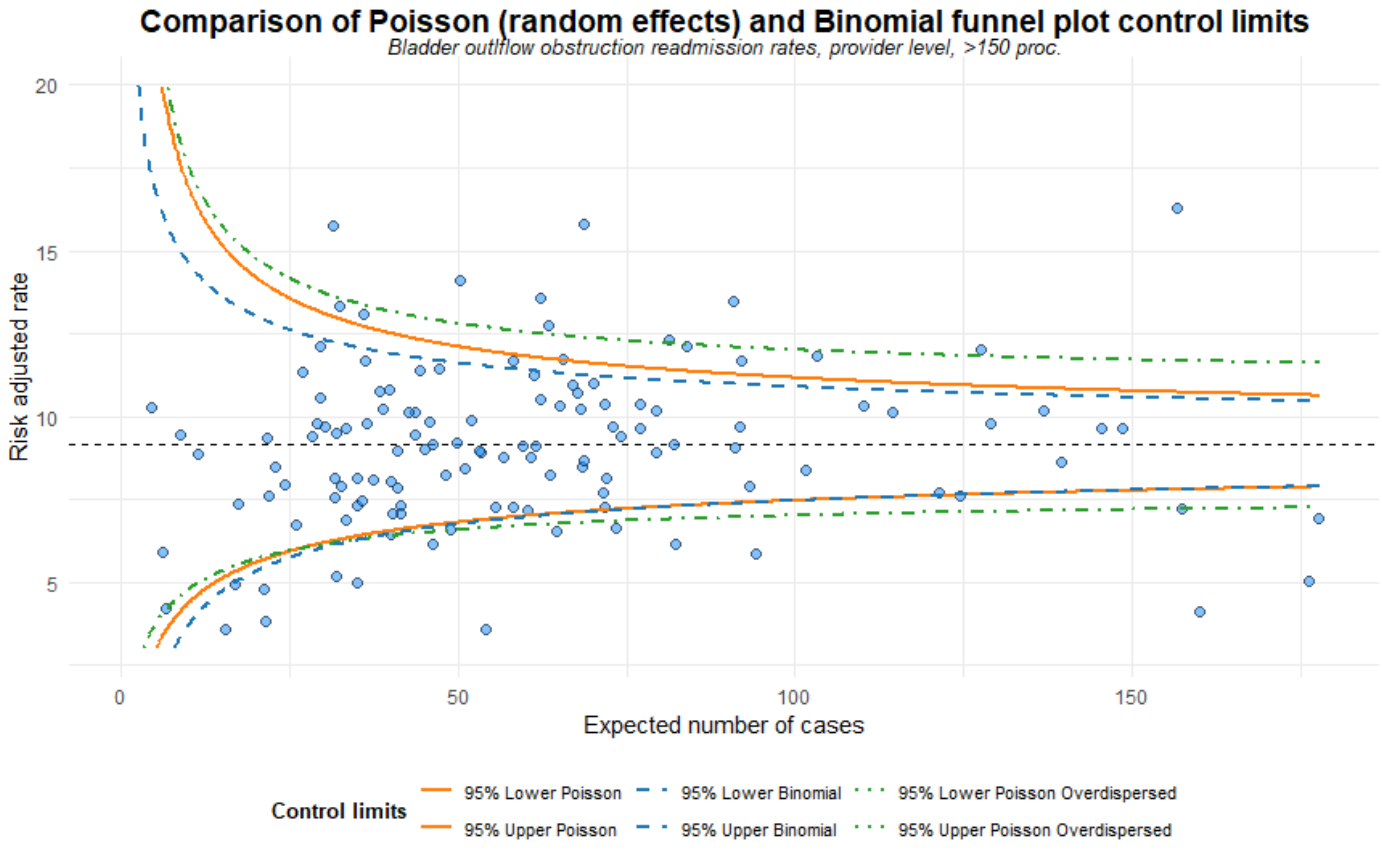


Appendix B.8 Funnel plot control limits

In a sensitivity analysis, alternative limits were presented for risk-adjustment of male bladder outflow obstruction procedures. Here, adjustments for overdispersion indicated a lower number of outliers. However, we advise the use of binomial control limits to maintain simplicity and consistency across all procedures.

Figure B.4 shows the risk adjusted rates of Male bladder outflow obstruction surgery in a funnel plot with binomial, random effect control limits (Poisson), and random effect control limits adjusted for overdispersion. When adjusting for overdispersion, control limits are wider. However, we aim for a simplistic method across all procedures. The use of binomial control limits ensures limits are constructed similarly across all procedures. Moreover, adjustment for overdispersion remains controversial, as it is unclear to what account heterogeneity between consultants should be controlled for [19,20]. On the other hand, the binomial control limits are frequently recommended in the literature, and therefore we recommend the use of traditional control limits [22].

Figure B.4: Comparison of funnel plot control limits: approximate binomial, Poisson, over-dispersed Poisson



Appendix C. Case studies

The below table presents the coefficients and results of logistic regression models for male bladder outflow obstruction surgery (30-day readmission) and Cystectomy for malignant neoplasms of the bladder (90-day mortality). Overall, relation between risk factors and outcomes were as expected. There is no significant association between deprivation/urbanicity and male bladder outflow obstruction 30-day readmission, but all other risk factors showed strong, significant associations.

Table C.1: Results of logistic regression, male bladder outflow obstruction, 30-day emergency readmission

Emergency Readmission 30-days *Male Bladder outflow obstruction surgery*	
Risk factor	**Coefficient (Standard error)**
Age	−0.060*** (0.016)
Age^2	0.001*** (0.0001)
Charlson Score	0.044*** (0.008)
Procedure subgroup: Bladder neck incision (BNI)	−0.671*** (0.149)
Procedure subgroup: Endoscopic resection of prostate (TURP)	−0.564*** (0.136)
Procedure subgroup: Laser prostatectomy	−0.570*** (0.140)
Procedure subgroup: Urolift	−0.841*** (0.154)
Quarter 2	0.078* (0.041)
Quarter 3	0.117*** (0.041)
Quarter 4	0.069* (0.041)
N. of Emergency admissions	0.089*** (0.007)
HFRS Band: None	−0.164*** (0.039)
HFRS Band: Moderate	0.101*** (0.035)
HFRS Band: Severe	0.289*** (0.061)
Coding Depth	−0.038** (0.019)
IMD Score	0.001 (0.001)
Ethnicity Category: Black or Black British	−0.253** (0.127)
Ethnicity Category: Mixed,other	−0.356*** (0.083)
Ethnicity Category: Unknown	−0.587*** (0.125)
Ethnicity Category: White	−0.152** (0.070)
Urbanicity: Urban	0.050 (0.032)
Constant	−0.034 (0.608)
Observations	67,915

Note: Logistic regression results for 30-day emergency readmission presented. Coefficients, standard error (in parentheses) and significance levels are shown, indicated by asterisks. (∗ p<0.10 ∗∗ p<0.05 ∗∗∗ p<0.01).

Table C.2: Results of logistic regression, Cystectomy, 90-day mortality

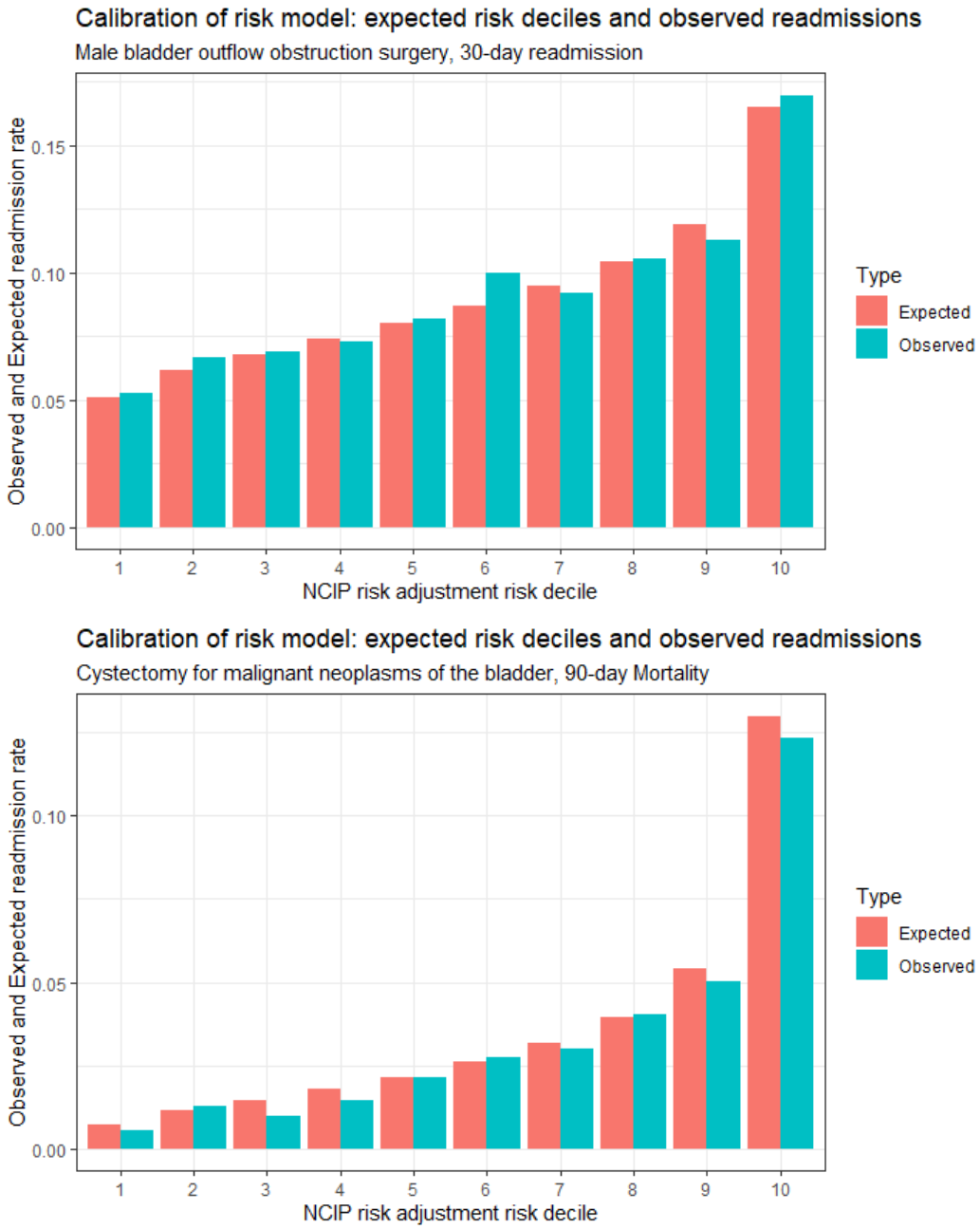
90-day mortality	
Cystectomy for malignant neoplasms of the bladder	
Risk factor	**Coefficient (Standard error)**
Age	−0.184*** (0.067)
Age^2	0.002*** (0.001)
Charlson Score	0.165*** (0.027)
Procedure subgroup: with nephroureterectomy	−0.644 (0.787)
Procedure subgroup: without nephroureterectomy or nephrectomy	−0.939 (0.647)
Approach: Other minimal access	−0.535 (0.466)
Approach: Robotic	0.152 (0.157)
Male	−0.003 (0.174)
Quarter 2	−0.296 (0.207)
Quarter 3	−0.126 (0.199)
Quarter 4	−0.588*** (0.224)
N of emergency admissions	0.049 (0.041)
HFRS Band Moderate	0.427** (0.209)
HFRS Band None	0.043 (0.279)
HFRS Band Severe	0.872*** (0.309)
Coding depth	−0.165* (0.090)
Palliative care	2.877*** (0.322)
Ethnicity Category: Black or Black British	14.222 (466.617)
Ethnicity Category: Mixed,other	14.121 (466.616)
Ethnicity Category: Unknown	14.563 (466.616)
Ethnicity Category: White	14.342 (466.616)
IMD score	0.005 (0.005)
Urbanicity: Urban	−0.285* (0.172)
Constant	−12.257 (466.622)
Observations	5,575

Note: Logistic regression results for 30-day emergency readmission presented. Coefficients, standard error (in parentheses) and significance levels are shown, indicated by asterisks. (p<0.10 ** p<0.05 *** p<0.01).*

Calibration

We observe that the model's predicted number of deaths/readmissions is similar to the observed number of deaths/readmissions (Figure C.1) in each risk decile. The risk deciles are constructed from the predicted model probabilities. This implies good calibration, and is in line with the high p-values from the HL-test.

Figure C.1: Risk decile plots, showing expected and observed number of events by risk decile.



Appendix D. Discussion

We note that risk adjustment traditionally does not use procedure data. A shortcoming of presenting procedure-specific outcome measurements is that consultants/providers have direct influence over the delivery of care and procedures chosen. Hence, consultant or provider decisions may bias the presented outcome measures. For example, if a certain approach is carried out more often than usual by a specific provider, this may favourably impact risk-adjustments.

This is highly unlikely to occur. Clinical specialty associations ensure good clinical (coding) practice is maintained, and should continue to do so. One way to address this further is by using the risk-adjusted model in conjunction with the volume metrics presented in NCIP. Both above average volumes for high-risk procedures, as well as risk-adjusted outcome measures should be considered. Further, in the risk adjustment model we naturally assume that consultants choose the correct procedure for each patient.