Preoperative risk prediction tools that predict morbidity risk in adults undergoing surgery: An Evidence Review

Authors: Alesha Wale[1], Toby Ayres[1], Salina Khatoon[1], Amy Fox-McNally[1], Claire Morgan[1], Helen Morgan[1], Hannah Shaw[1], Jacob Davies[2], Rhiannon Tudor Edwards[2], Claire Dunstan[3], Adrian Edwards[4], Alison Cooper[4], Ruth Lewis[5,6]

1 Public Health Wales Evidence Service, Cardiff, United Kingdom
2 Centre for Health Economics and Medicines Evaluation, Bangor University, United Kingdom
3 Clinical Implementation Network, NHS Performance and Improvement, Cardiff and Vale University Health Board, United Kingdom
4 Health and Care Research Wales Evidence Centre, Cardiff University, United Kingdom
5 Health and Care Research Wales Evidence Centre, Bangor University, United Kingdom
6 Bangor Institute for Medical and Health Research, Bangor University, United Kingdom

Abstract:
Risk prediction tools play a critical role in preoperative care by estimating the likelihood of adverse outcomes, including mortality, morbidity, and postoperative complications. In low-risk surgical settings such as surgical hubs, accurate risk prediction is particularly valuable. The aim of this review was to identify and map the evidence for 14 validated pre-operative surgical risk prediction tools currently used in Wales within any elective, or non-emergency surgical setting, and to provide a more in-depth look at the findings for a selection of tools deemed to be the most applicable on a population level to the context of surgical hubs.

Included studies were published between 1999 and 2024. No evidence was found for two of the risk prediction tools however, a total of 118 studies were identified across 12 risk prediction tools. None of the evidence found was looking at the predictive ability of risk prediction tools for selecting patients suitable for surgical hubs. The tools were used across a range of surgical specialties and measured composite complications, individual complications, and healthcare utilisation and recovery measures. No risk prediction tool adequately predicted complications across all surgical specialties. Among the included studies, there was considerable heterogeneity in which surgical specialties the risk prediction tools were used for, how complications were defined, and which measures were used to determine a tool's predictive ability. This makes direct comparisons very challenging.

Four tools were selected as being potentially the most impactful at a population level for a more in-depth look at the findings: ACS NSQIP, P-POSSUM, RCRI, ASA classification system. A total of 76 studies were identified across these 4 tools. Key findings for the four risk prediction tools of interest are described. Overall, no one tool was identified that adequately predicted complications across all surgical specialties. The predictive ability of the tools varied across different surgical specialties.

Further research using consistent methods is needed to better understand the predictive ability of risk prediction tools and allow a robust evaluation. Given no single risk prediction tool adequately predicted complications across all surgical specialties, it may be likely that some tools are better suited for specific surgery types or that a combination of risk prediction tools may be needed to adequately assess an individual's level of risk.

Funding statement: The authors and their Institutions were funded for this work by the Health and Care Research Wales Evidence Centre, itself funded by Health and Care Research Wales on behalf of Welsh Government.

Health and Care Research Wales Evidence Centre

Canolfan Dystiolaeth Ymchwil Iechyd a Gofal Cymru

Preoperative risk prediction tools that predict morbidity risk in adults undergoing surgery: An Evidence Review

June 2025

Ymchwil Iechyd a Gofal Cymru
Health and Care Research Wales

Ariennir gan Lywodraeth Cymru
Funded by Welsh Government

GIG CYMRU NHS WALES
Iechyd Cyhoeddus Cymru
Public Health Wales

Sefydliad Ymchwil Iechyd a Meddygol Bangor
Prifysgol Bangor University
Bangor Institute for Health & Medical Research

Canolfan Economeg Iechyd a Gwerthuso Meddyginiaethau
CHEME
Centre for Health Economics and Medicines Evaluation

Report Contributors

Review Team
Alesha Wale, Toby Ayres, Salina Khatoon, Amy Fox-McNally, Claire Morgan, Helen Morgan, Hannah Shaw, Public Health Wales Evidence Service

Economic Considerations
Jacob Rees Davies and Rhiannon Tudor Edwards, Centre for Health Economics and Medicines Evaluation (CHEME), Bangor University

Methodological Advice
Ruth Lewis, Health and Care Research Wales Evidence Centre, and Bangor Institute for Medical and Health Research, Bangor University, United Kingdom

Evidence Centre Team
Adrian Edwards, Ruth Lewis, Alison Cooper, Elizabeth Doe and Micaela Gal involved in Stakeholder engagement, review of report and editing.

Public Partner
Praveena Pemmasani

Stakeholders
Clinical Implementation Network
Dr Claire Dunstan - National Clinical Lead for Anaesthetics and Peri-operative care
Dr Catherine Cromey - Deputy South Wales - Anaesthetics and Peri-operative Care
Dr Linda Warnock - Deputy North Wales - Anaesthetics and Peri-operative Care
Meredith Graham - Project Manager, Anaesthetics and Peri-operative care

Clinical Implementation Network (CIN), Strategic Programme for Planned Care and Recovery, NHS Performance and Improvement.

Evidence need submitted to the Evidence Centre: 22nd April 2024

Initial Stakeholder Consultation Meeting: 24th October 2024

Final Report issued: June 2025

The review should be cited as: Health and Care Research Wales Evidence Centre. Preoperative risk prediction tools that predict morbidity risk in adults undergoing surgery: An Evidence Review (RR0037). June 2025.

Disclaimer: The views expressed in this publication are those of the authors, not necessarily Health and Care Research Wales. The Health and Care Research Wales Evidence Centre and authors of this work declare that they have no conflict of interest.

Preoperative risk prediction tools that predict morbidity risk in adults undergoing surgery: An Evidence Review

EXECUTIVE SUMMARY

Report number RR0037 (June 2025)

What are Evidence Reviews?

This evidence review includes a Rapid Evidence Map (REM) and an in-depth summary. The REM describes the evidence base and the in-depth summary focusses on the validity of a sub-set of studies. It has utilised abbreviated systematic mapping and scoping review methods to provide a description of the nature, characteristics and volume of the available evidence for a particular policy domain or research question, and then a detailed summary of a subset of this evidence.

Who is this summary for?

Planned Care Wales

Background / Aim of Rapid Evidence Map

Risk prediction tools play a critical role in preoperative care by estimating the likelihood of adverse outcomes, including mortality, morbidity, and postoperative complications. In low-risk surgical settings such as surgical hubs, which typically focus on high-volume, low-complexity procedures, accurate risk prediction is particularly valuable. The aim of this review was to identify and map the evidence for 14 validated pre-operative surgical risk prediction tools currently used in Wales within any elective, or non-emergency surgical setting, and to provide a more in-depth look at the findings for a selection of tools deemed to be the most applicable on a population level to the context of surgical hubs. The initial list of prediction tools used in Wales was identified by the stakeholders, who also informed the selection of tools for a more-in depth summary based on the findings of the initial evidence map.

Results

Recency of the evidence base

- This review included evidence available up until December 2024. Included studies were published between 1999 and 2024.

Extent of the evidence base

- A total of 118 studies were identified across 12 risk prediction tools.
- No evidence was identified assessing the predictive ability of two of the tools: the Carlisle Risk Calculator and the NELA PRS.
- No evidence was found looking at the predictive ability of risk prediction tools for selecting patients suitable for surgical hubs.
- The tools were used across a range of surgical specialties including: general; mixed; orthopaedic; cardiothoracic; urology; vascular; neurosurgery; plastic; gynaecology; ENT; urogynaecology and oral and maxillofacial surgery.
- Risk prediction tools measured: Composite (grouped) complications (e.g. any complications, severe complications, morbidity), individual complications (e.g. pneumonia, surgical site infection), and healthcare utilisation and recovery measures (e.g. readmission, length of stay, return to the operating room).
- No risk prediction tool adequately predicted complications across all surgical specialties.
- There is considerable heterogeneity among the included studies in which surgical specialties the risk prediction tools are being used for, how complications are defined, and which measures are used to determine a tool's predictive ability. This makes direct comparisons very challenging.

- Four tools were selected as being potentially the most impactful at a population level for a more in-depth look at the findings: ACS NSQIP, P-POSSUM, RCRI, ASA classification system.
- A total of 76 studies were identified across the 4 risk prediction tools (ACS NSQIP n=40; RCRI n=16; ASA classification system n=13; P-POSSUM n=7).

Key findings for the four risk prediction tools of interest

- Overall, no one tool was identified that adequately predicted complications across all surgical specialties. The predictive ability of the tools varied across different surgical specialties. The findings for the different surgical specialities may be limited due to a very small evidence base available for each surgical type.
- **ACS NSQIP was found to have a poor predictive ability for composite complications** across studies. There is limited evidence to suggest that **ACS NSQIP had an excellent predictive ability** for complications after mixed surgery; **a fair predictive ability** after thoracic or plastic surgery; **a poor predictive ability** after neurosurgery, gynaecology or general surgery; **and a very poor predictive ability** after orthopaedic, urology or vascular surgery.
- **P-POSSUM was found to have a poor predictive ability for composite complications** across studies. There is very limited evidence to suggest that **P-POSSUM had a fair predictive ability** for complications after ENT surgery; **a poor predictive ability** after general surgery; **and a very poor predictive ability** after gynaecology surgery.
- **The RCRI was found to have a fair predictive ability for composite complications** across studies. There is very limited evidence to suggest that **the RCRI had a fair predictive ability** for complications after vascular, orthopaedic, or mixed surgery, and **a poor predictive ability** after urology surgery.
- **The ASA classification system was found to have a poor predictive ability for composite complications** across studies. There is very limited evidence to suggest **the ASA classification system had a fair predictive ability** for complications after mixed surgery; **a poor predictive ability** after general or orthopaedic surgery; and **a very poor predictive ability** for vascular surgery and urology surgery.
- The evidence directly comparing risk prediction tools appears to be mixed.

Summary of the evidence gaps

- No quality appraisal of included studies was conducted and therefore we cannot report the quality of the included studies.
- No evidence was found looking at the use of risk prediction tools for identifying patients suitable for treatment in surgical hubs.
- Further research using consistent methods is needed to better understand the predictive ability of risk prediction tools.

Implications and next steps

- No risk prediction tool adequately predicted complications across all surgical specialties, as such it may be likely that some tools are better suited for specific surgery types or that a combination of risk prediction tools may be needed to adequately assess an individual's level of risk.
- The heterogeneity among the included studies makes direct comparisons very challenging even when looking at the evidence available for each tool individually, further research should ensure consistent methods are used to assess the predictive ability of risk prediction tools and allow a robust evaluation.

Economic considerations

- Despite not being within the scope of this review, there is a recognised research gap regarding economic evaluations of risk prediction tools.
- Future research into risk prediction tools should incorporate health economic evaluations, considering not only individual risks but associated cost implications.

TABLE OF CONTENTS

TABLE OF CONTENTS 6
1. BACKGROUND 11
1.1 Who is this Evidence Review for? 11
1.2 Background and purpose of this Evidence Review 11
1.3 NICE guidance on preoperative risk stratification tools 12
2. Summary of the evidence base included in the Rapid Evidence Map 12
2.1 Risk prediction tools included in the Rapid Evidence Map 13
2.2 Overview of the available evidence on external validation 15
2.3 Narrative Summary of the evidence base identified for each risk prediction tool 17
2.3.1 ACS NSQIP 17
2.3.2 CPET 18
2.3.3 RCRI 18
2.3.4 POSSUM 19
2.3.5 ASA classification system 19
2.3.6 Apfel Score (PONV) 20
2.3.7 P-POSSUM 20
2.3.8 NRS-2002 21
2.3.9 ARISCAT 21
2.3.10 CFS 21
2.3.11 SORT 22
2.3.12 DASI 22
3. Summary of the findings for ACS NSQIP, P-POSSUM, RCRI and the ASA classification system 24
3.1 Studies comparing the predictive ability of multiple risk prediction tools 25
3.2 ACS NSQIP findings 28
3.2.1 ACS NSQIP Discrimination findings 28
3.2.2 ACS NSQIP Calibration findings 30
3.2.3 ACS NSQIP Accuracy findings 31
3.2.4 ACS NSQIP Healthcare utilisation and recovery outcomes 34
3.2.5 Bottom line summary for ACS NSQIP 34
3.3 P-POSSUM findings 35
3.3.1 P-POSSUM Discrimination findings 35
3.3.2 P-POSSUM Calibration findings 36
3.3.3 P-POSSUM Accuracy findings 37

3.3.4 P-POSSUM healthcare utilisation and recovery outcomes 37
3.3.5 Bottom line summary for P-POSSUM 37
3.4 RCRI findings 38
3.4.1 RCRI Discrimination findings 38
3.4.2 RCRI Calibration findings 39
3.4.3 RCRI Accuracy findings 39
3.4.4 RCRI Healthcare utilisation and recovery outcomes 40
3.4.5 Bottom line summary for RCRI 40
3.5 ASA classification system findings 41
3.5.1 ASA classification system Discrimination findings 41
3.5.2 ASA classification system Calibration findings 42
3.5.3 ASA classification system Accuracy findings 42
3.5.4. ASA classification system Healthcare utilisation and recovery outcomes 42
3.5.5. Bottom line summary for ASA classification system. 43
4. DISCUSSION 44
4.1 The evidence base 44
4.2 Findings for ACS NSQIP, P-POSSUM, RCRI and the ASA classification system 44
4.3 Limitations of the available evidence 45
4.4 Summary of the Evidence gaps 46
4.5 Strengths and limitations of this Evidence Review 46
4.6 Implications and next steps 47
4.7 Economic considerations* 47
5. REFERENCES 48
6. EVIDENCE REVIEW METHODS 50
6.1 Eligibility criteria 50
6.2 Literature search 51
6.3 Study selection process 51
6.4 Data extraction and coding/charting 51
6.5 Assessment of methodological quality 52
7. EVIDENCE 52
7.1 Search results and study selection 52
7.2 Data extraction Tables 53
7.3 Information available on request 99
8. ADDITIONAL INFORMATION 99
8.1 Conflicts of interest 99
8.2 Acknowledgements 99
9. APPENDIX 100

9.1 Appendix 1 Data extraction table of studies that include mortality 100
9.2 Appendix 2. Modified versions of surgical risk prediction tools 104
9.3 Appendix 3. Breakdown of all outcomes reported across studies by tool and surgical specialty. 107
9.4 Appendix 4. Medline search strategies 125

Abbreviations

Acronym	Full Description
ACS NSQIP	The American College of Surgeons National Surgical Quality Improvement Program
ARISCAT	Assess Respiratory Risk in Surgical Patients in Catalonia
ASA	The American Society of Anesthesiologists Physical Status Classification System
AUC	Area Under the Curve
CFS	Clinical Frailty Scale
CPET	Cardiopulmonary exercise testing
DASI	Duke Activity Status Index
ECG	Electrocardiogram
ENT	Ear, Nose, and Throat
GCS	Glasgow Coma Scale
NELA-PRS	National Emergency Laparotomy Audit Parsimonious Risk Score
NICE	National Institute for Health and Care Excellence
NRS-2002	Nutritional Risk Screening 2002
O/E	Observed/Expected
POSSUM	Physiological and Operative Severity Score for the enUmeration of Mortality and morbidity
P-POSSUM	Portsmouth Physiological and Operative Severity Score for the enUmeration of Mortality and morbidity
PONV	Apfel Score for Postoperative Nausea and Vomiting
RCRI	Revised Cardiac Risk Index for Pre-Operative Risk
REM	Rapid Evidence Map
ROC	Receiver Operating Characteristics
SORT	Surgical Outcome Risk Tool
UK	United Kingdom
USA	United States of America

Glossary

Term	Definition
Accuracy	Accuracy of risk prediction tools is commonly assessed in the literature using the 'Brier Score'. This is a simultaneous measure of calibration and discrimination, reported as a score between 0 and 1. A score of 0 indicates no difference between the outcome predicted by the tool and actual outcome, thus indicating the best possible result. A score of 1 indicates that the test did not predict the outcome.
Calibration	Calibration assesses the agreement between the predicted and observed outcomes, typically presented graphically as observed risks versus predicted risks or an observed/expected (O/E) ratio (Collins et al., 2014). An O/E ratio of 1 implies the tool accurately predicted complications, a ratio below 1 suggests the tool overpredicted complications, whereas a ratio over 1 suggests the tool underpredicted complications (Hammond et al., 2021). Following the approach utilised by NICE, an O/E ratio of between 0.9 to1.1 would be considered a fair level of calibration.
Composite outcomes	Rather than reporting individual outcomes, some studies combined two or more outcomes into a single measure. For example, 'All complications' could include, pneumonia, sepsis, infection etc. Where outcomes are combined, this is known as a composite outcome.
Discrimination	Discrimination measures how well a tool differentiates between patients who do and do not experience an event, quantified using the area under the receiver operating characteristic (ROC) curve (AUC) or c-statistic. Within the literature, the following c-statistic scores relate to a tool's performance: • 90% or greater -considered an excellent level of discriminative ability, • 80% or greater is considered a good level of discriminative ability, • 70% or greater is considered a fair level of discriminative ability, • 60% or greater is considered a poor level of discriminative ability, • 50% or greater is considered a very poor level of discriminative ability (or no better than chance) (NICE 2020; Çorbacıoğlu and Aksel, 2023).
External validation	In this context external validation refers to the evaluation of a risk prediction tool's predictive performance in a dataset that was not used to develop the tool, which is known as internal validation (Collins et al., 2014).
Median	The median is the middle value in a group of numbers ranked in order of size; in this way it is not sensitive to outliers or skewed data.
Mixed surgery	In this context mixed surgery refers a study that included a dataset or sample with a range of surgeries across more than one surgical specialty
Rapid Evidence Map	Rapid Evidence Maps (REMs) use abbreviated systematic mapping or scoping review methods to provide a description of the nature, characteristics and volume of the available evidence for a particular policy domain or research question. The methods have been developed as part of the Health and Care Research Wales Evidence Centre Collaboration.
Risk prediction tools	Tools, usually a set of clinical or personal information that is used to calculate the estimated likelihood of adverse outcomes, including mortality, morbidity, and postoperative complications.

1. BACKGROUND

1.1 Who is this Evidence Review for?

This Evidence Review was conducted as part of the Health and Care Research Wales Evidence Centre Work Programme. The review question was suggested by Planned Care Wales.

1.2 Background and purpose of this Evidence Review

Risk prediction tools play a critical role in preoperative care by estimating the likelihood of adverse outcomes, including mortality, morbidity, and postoperative complications (The National Institute for Health and Care Excellence [NICE], 2020). These tools support clinicians in making informed decisions about a patient's overall suitability for surgery and identifying the need for enhanced postoperative care. However, despite their widespread use, there is significant variation in how these tools are applied across different disciplines and surgical settings, with no standardised approach being adopted (Pradhan et al. 2022).

In low-risk surgical settings such as surgical hubs, which typically focus on high-volume, low-complexity procedures, accurate risk prediction is particularly valuable. It helps ensure that patients selected for these settings can safely benefit from treatment while maintaining the efficiency and safety standards required for such facilities. A wide range of risk prediction tools are used for selecting patients who can safely be treated at a surgical hub. However, while the tools have been found to be accurate and reliable, given that there is no standardised approach to selecting risk prediction tools, their selection may not be evidence based and it can be unclear which tools should be used. Some of the tools used are not designed for use in surgery and others may not be best-suited for the low-risk surgical settings.

To ensure a risk prediction tool will be effective it is essential to refer to studies that evaluate its performance using new datasets, and not the dataset that was used to develop the model (referred to as *external validation studies*) (Collins et al., 2014). Assessing the performance of a risk prediction tool in other datasets allows evaluation of the transferability of the tool across different cohorts to examine how well it performs (Collins et al., 2014). External validation studies evaluate both *discrimination* and *calibration* to determine a tool's performance (how accurately it predicts a risk) (Collins et al., 2014). Discrimination measures how well the tool differentiates between patients who do and do not experience an event, quantified using the area under the receiver operating characteristic (ROC) curve (AUC) or c-statistic. Calibration assesses the agreement between the predicted and observed outcomes, typically presented graphically as observed risks versus predicted risks or in tabular format (Collins et al., 2014). Some studies also report the accuracy of the tool using the '*Brier Score'*. The Brier score is a simultaneous measure of calibration and discrimination (Alzahrani et al., 2020).

A preliminary review of studies reporting on the external validation of risk prediction tools for assessing the risk of post-operative morbidity or mortality in any surgical settings identified a large volume of studies and a very complex picture of the evidence base. As such, it was decided that an evidence review would be conducted which included an initial rapid evidence map (REM) to understand this complex evidence landscape, and an in-depth summary of the findings for a selection of the risk prediction tools. The REM provides a description of the intended purpose of each tool, the context within which they have been developed and used, and the amount of external validation studies available. The findings of this initial map were

then used to select a manageable sub-set of relevant risk prediction tools for a more in-depth evaluation of their validity.

There is a need to identify which validated risk prediction tools are most suitable for use/application to the Welsh population, and which ones are best for selecting patients that can safely go to surgical hubs. A review of external validation studies of risk prediction tools could help decision-makers when selecting which tool(s) to use for assessing patients' risk of adverse outcomes, and could help to optimise resource allocation or candidacy for surgery in low-risk settings such as surgical hubs. As surgical hubs are a relatively new concept, and can be limited to specific surgical specialties, little evidence exists around the use of risk prediction tools within these settings. Therefore, the aims of this review were:

- To identify and map the evidence for validated pre-operative surgical risk prediction tools currently used in Wales in any elective, or non-emergency surgical settings (See Section 2).

- To provide a more in-depth summary of the findings for a selection of the tools deemed to be the most applicable on a population level (See Section 3).

This will allow a clearer picture of which tools may be most suitable for use within low-risk surgical settings, such as surgical hubs. As surgical hubs are intended as low-risk settings, this review focuses on external validation studies of risk prediction models for assessing the risk of **preoperative morbidity and complications**.

1.3 NICE guidance on preoperative risk stratification tools

A recent review conducted by NICE (2020) sought to explore which of three risk prediction tools (P-POSSUM, SORT, or ACS NSQIP) could best identify the risk of mortality and morbidity in adults undergoing surgery (NICE 2020). While separate findings were reported for morbidity and mortality, some of the included studies reporting morbidity also considered mortality as a complication. As surgical hubs are low-risk settings, our review focusses on complications only and not mortality. However, our review compliments the NICE evidence review by focussing on complications that do not include mortality, thus making it more applicable to low-risk settings, and by updating the evidence base to include more recently published studies. The studies included in the NICE review were screened for inclusion in this review, and any studies that reported complications separately to mortality were included in our review.

Whilst we recognise mortality can occur as a post-operative complication, the focus of our review is on morbidity only given the likelihood of death in low-risk surgical settings should be significantly lower than other settings. However, in order to better reflect the entire evidence base, a summary of studies included in the NICE review that were identified during screening that included mortality and complications in one composite outcome will be provided to ensure the evidence base as a whole is represented (See Appendix 1).

2. Summary of the evidence base included in the Rapid Evidence Map

The eligibility criteria used to select relevant validation studies for inclusion in this review and a description of the methods used to conduct the overall review are described in Section 6. This section starts with a description of the risk prediction tools included in the evidence map. Table 1 provides details about the original purpose of the tools, their uses and the outputs they provide. This is followed by a detailed description of the evidence base identified across all risk prediction tools. Table 2 includes how many studies were identified for each tool, the overall sample sizes used across studies, any specific population group being used to validate the tool, the number of studies identified for each surgical specialty, and the number of studies reporting on the different outcomes.

2.1 Risk prediction tools included in the Rapid Evidence Map

The evidence map focuses on 14 risk prediction tools currently utilised within the surgical setting across Wales. These were identified and provided by the stakeholders and include the following tools:

- The American College of Surgeons National Surgical Quality Improvement Program (ACS NSQIP)
- Cardiopulmonary exercise testing (CPET)
- Revised Cardiac Risk Index for Pre-Operative Risk (RCRI)
- The American Society of Anesthesiologists Physical Status (ASA) Classification System
- Apfel score (PONV)
- Portsmouth Physiological and Operative Severity Score for the enUmeration of Mortality and morbidity (P-POSSUM)
- Physiological and Operative Severity Score for the enUmeration of Mortality and morbidity (POSSUM)
- Nutritional Risk Screening 2002 (NRS-2002)
- Assess Respiratory Risk in Surgical Patients in Catalonia (ARISCAT)
- Clinical Frailty Scale (CFS)
- Surgical Outcome Risk Tool (SORT)
- Duke Activity Status Index (DASI)
- Carlisle Risk Calculator
- National Emergency Laparotomy Audit Parsimonious Risk Score (NELA-PRS)

We initially set out to identify evidence in relation to the external validation of 13 risk prediction tools provided by the stakeholders. However, we identified external validation studies for a further relevant risk prediction tool, POSSUM, which is an earlier version of the P-POSSUM tool (identified by the stakeholders). It is very similar but uses a slightly different formula to calculate risk. The POSSUM was therefore also included, which brought the total number of risk prediction tools included in this evidence map to 14. However, we did not identify evidence relating to the external validation of two of these risk prediction tools: the Carlisle Risk Calculator and the National Emergency Laparotomy Audit Parsimonious Risk Score. **Table 1** provides details about the 12 individual tools included in this evidence map, their purpose, who the tools are completed by, the surgical disciplines they are used in, and how their findings are provided.

Table 1. Summary of risk prediction tools

Tool name	Purpose*	Who is it completed by	Surgical disciplines	Output*
ACS NSQIP	Estimates the chance of an unfavourable outcome (such as a complication or death) after surgery through a series of questions relating to patient severity and surgical information. The ACS NSQIP surgical risk calculator is a decision-support tool based on reliable multi-institutional clinical data, which can be used to estimate the risks of most operations.	Clinician	Any procedure, in most surgical subspecialties,	Individual risk scores for up to 19 different outcomes within 30-days following surgery, including a predicted length of stay.
POSSUM	Estimates morbidity and mortality for general surgery patients through a series of questions around patient physiological score which looks at age, cardiac and respiratory function, GCS, blood work, ECG, and operative severity which interrogates the procedure being undertaken.	Clinician	Emergency and elective general surgical procedures	A single overall score for predicted morbidity and a single overall score for predicted mortality.
P-POSSUM	POSSUM-P is very similar to the POSSUM tool but uses a slightly different formula to calculate risk.	Clinician	Same as POSSUM	Same as POSSUM
RCRI	Estimates risk of cardiac complications after non-cardiac surgery in patients ≥45 years old (or 18-44 years old with significant cardiovascular disease). Six questions focus on surgical risk, history of heart and cerebrovascular disease, preoperative treatment with insulin and preoperative creatinine levels.	Clinician	Elective non-cardiac surgery or urgent/semi-urgent (non-emergent) non-cardiac surgery.	A single overall score for predicted postoperative cardiac complications
ARISCAT	Predicts risk of pulmonary complications after surgery, including respiratory failure. Seven questions focussed on oxygen saturation levels, respiratory infections, anaemia, and surgery information.	Clinician	Surgery performed under general, neuraxial, or regional anaesthesia	A single overall score for predicted postoperative pulmonary complications
CPET	Provides an objective assessment of exercise capacity preoperatively and identifies the causes of exercise limitation.	Clinician	Abdominal, colorectal, urological, hepatobiliary, liver, bariatric, vascular, thoracic, oesophageal gastric.	Scores for VO^2 peak and anaerobic threshold
DASI	Measures functional capacity and assesses aspects of quality of life by asking 12 questions. Noted not to be as accurate as objective measures such as exercise stress testing.	Patients self-administer	Not specified but used during preoperative assessments for surgery and before and during exercise programmes	A single overall score for functional capacity
ASA Classification system	Classifies health of patients prior to surgery. Should not be used alone to determine appropriateness of surgery, but as part of a comprehensive preoperative evaluation. Other evaluations to be considered prior to surgery may include functional capacity, classification of surgery type, and other chronic conditions. The tool asks two questions relating to overall health and if surgery is an emergency procedure.	Clinician	Not specified however it should not be used alone to predict risk of morbidity.	Provides an ASA classification (I-VI) where a higher grade suggests higher operative risk
PONV	Predicts risk of postoperative nausea and vomiting (PONV) through four questions looking at gender, smoking status, history of motion sickness or PONV, and use of postoperative opioids.	Clinician	Any surgery performed under general anaesthesia	Provides an overall score for the risk of experiencing PONV
SORT	Provides an estimate of the risk of death within 30 days of an operation through a series of questions looking at: the procedure; ASA classification; clinical urgency based on the NCEPOD classification of intervention (2004); and patient age. This tool also requires a clinical risk assessment to estimate 30-day mortality (ideally made by a senior clinician in the multi-disciplinary peri-operative care team).	Clinician	Surgery not specified, but for use in adults only	Provides an overall risk prediction score for mortality
NRS-2002	Predicts risk of malnutrition in hospitalised patients through four questions relating to weight, diet and if an ICU patient.	Clinician	N/A not originally designed for surgery	Provides an overall score for malnutritional risk
CFS	Launched during the COVID-19 crisis, it measures frailty to predict survival in medium and long term (mortality and need for institutional care) through a 9-point scale. .	Clinician	N/A not originally designed for surgery	Provides an overall frailty score

*See Section 2.2 for more details on what the tools took into consideration as part of their risk calculation and how risk is determined.
Abbreviations: ACS NSQIP - The American College of Surgeons National Surgical Quality Improvement Program Tool; CPET - Cardiopulmonary exercise testing; RCRI - The Revised Cardiac Risk Index; POSSUM- The Physiological and Operative Severity Score for the enUmeration of Mortality and Morbidity; ASA Classification system - The American Society of Anesthesiologists Physical Status Classification System; PONV - Postoperative Nausea and Vomiting; P-POSSUM - The Portsmouth Physiological and Operative Severity Score for the enUmeration of Mortality and Morbidity; NRS-2002 - The Nutritional Risk Screening 2002; ARISCAT - The Assess Respiratory Risk in Surgical Patients in Catalonia; CFS - The Clinical Frailty Scale; SORT - The Surgical Outcome Risk Tool; DASI - The Duke Activity Status Index; GCS - Glasgow Coma Scale; ECG - Electrocardiogram; ICU – Intensive care unit.

2.2 Overview of the available evidence on external validation

A summary of the available evidence underpinning each tool is provided in Table 2. A total of 118 validation studies (85 retrospective and 33 prospective) met our inclusion criteria. The summary table (Table 2) shows how many studies were identified for each tool, the overall sample sizes used across studies, any specific population group being used to validate the tool, the number of studies identified for each surgical specialty, and the number of studies reporting on the different outcomes. All included studies conducted external validation evaluation of surgical risk prediction tools in adults undergoing non-emergency surgery, assessing their ability to predict complications.

Our included studies examined risk prediction tools currently used in Wales. In order of the number of included studies, these were:

- The American College of Surgeons National Surgical Quality Improvement Program (ACS NSQIP) (n=40)
- Cardiopulmonary exercise testing (CPET) (n=17)
- Revised Cardiac Risk Index for Pre-Operative Risk (RCRI) (n=16)
- Physiological and Operative Severity Score for the enUmeration of Mortality and morbidity (POSSUM) (n=14)
- Society of Anesthesiologists Physical Status (ASA) Classification System (n=13),
- Apfel score (n=9)
- Portsmouth Physiological and Operative Severity Score for the enUmeration of Mortality and morbidity (P-POSSUM) (n=7)
- Nutritional Risk Screening 2002 (NRS-2002) (n=5)
- Assess Respiratory Risk in Surgical Patients in Catalonia (ARISCAT) (n=4)
- Clinical Frailty Scale (CFS) (n=4)
- Surgical Outcome Risk Tool (SORT) (n=2)
- Duke Activity Status Index (DASI) (n=1).

Thirteen studies assessed multiple surgical risk prediction tools of interest: 12 studies assessed two tools; and one study assessed three tools. With the exception of the POSSUM and P-POSSUM tools, studies that assessed modifications of the surgical risk prediction tools were excluded. Tool modifications can be used to tailor risk prediction tools to specific surgical disciplines or medical conditions and can improve its' predictive ability and strengthen the validity of the tool (Hageman et al., 2023). However, the inclusion of these studies would have complicated the evidence base further, so for simplicity these were excluded. Despite excluding external validation studies of tool modifications from our review, in order to reflect the totality of the evidence base for non-specific tools used within surgical settings in Wales, those that were identified during our screening process are referenced in Appendix 2.

The included validation studies were from a range of countries including; USA (n=39), UK (n=16), China (n=12), Germany (n=6), Canada (n=5), Turkey (n=5), Italy (n=4), South Korea (n=4), Australia (n=2), India (n=2), Romania (n=2), Taiwan (n=2), and one study each from Africa, Austria, Denmark, France, Ireland, Japan, Latin America, Netherlands, New Zealand, Philippines, Russia, Sri Lanka, Sweden, Tanzania and one study included data from Germany and Finland combined. The country was not reported in four studies. The sample sizes of included studies varied, ranging from 29 to 447,352. Studies assessing the RCRI had the

highest number of overall participants (n=742,794), whereas SORT had the lowest number of participants (n=445).

External validity was assessed in patients undergoing various types of surgeries. Where possible, surgery types were categorised using the surgical specialties described by the Royal College of Surgeons of England (2024). However, an additional surgical specialty was also included to cover gynaecological surgeries and a 'mixed surgeries' category for studies that included a mixed sample of different surgery specialties. More information on the methods used to data extract, code and chart information, can be found in Section 6.4.

Most of the studies included patients undergoing general surgeries (n=43), followed by mixed surgeries (n=23), orthopaedic (n=16), cardiothoracic (n=9), urology (n=8), vascular (n=6), neurosurgery (n=5), plastic (n=2), gynaecology (n=2), ENT (n=2), urogynaecology (n=1), and oral and maxillofacial (n=1). While the majority of included studies included adults aged 18 years and over, a total of 18 studies validated the risk prediction tools in specific age groups (≥25yrs n=1; ≥40yrs n=1; ≥45yrs n=1; ≥50yrs n=1; ≥60yrs n=2; ≥65yrs n=6; ≥70yrs n=3; ≥75yrs n=1; ≥80yrs n=1).

Of the 16 UK external validation studies, the following risk prediction tools were assessed: DASI, P-POSSUM, POSSUM, CPET, and ACS NSQIP. The most commonly assessed tool was CPET (n=12 studies), followed by POSSUM (n=2). DASI, P-POSSUM, and ACS NSQIP were each assessed in one study.

Predicted outcomes of interest were categorised into complications, and healthcare utilisation and recovery. Where possible, individual categories for cardiac complications, non-cardiac complications, and postoperative nausea and vomiting were used. However, this was not always possible due to differences between tools and reporting in studies. Complications included composite (grouped outcomes e.g. any complications, severe complications) and individual outcomes (e.g. pneumonia, surgical site infection). Following the approach taken by the NICE review, morbidity was included as a composite complication (NICE 2020). Healthcare utilisation and recovery included outcomes such as readmissions, length of stay, return to the operating room and more. Each relevant outcome reported by the studies are detailed in Section 2.3.

2.3 Narrative Summary of the evidence base identified for each risk prediction tool

The sub-sections below outline a brief summary of each surgical risk prediction tool. They provide an overview of the tools, their uses and details the number of studies included that evaluated its external validity, the study region, participant information, surgical disciplines, and what outcomes were assessed. The summaries provide additional information to Tables 1 and 2.

2.3.1 ACS NSQIP

The American College of Surgeons (ACS) National Surgical Quality Improvement Program (NSQIP) was developed in 1994 to estimate the chances of unfavourable outcomes for veterans attending surgery. It is suitable for any procedure, in most surgical subspecialties. Risk is determined through a series of questions relating to patient condition, severity and information around the surgery being conducted and individual risk scores for up to 19 different outcomes are reported included a predicted length of stay (American College of Surgeons National Surgical Quality Improvement Program, 2025).

A total of 40 studies externally validated the ACS NSQIP tool. These were published between 2016 and 2024. A wide range of countries were used including USA (n=24), Italy (n=3), Canada (n=2), South Korea (n=2), Australia (n=1), China (n=1), Latin America (n=1), Netherlands (n=1), New Zealand (n=1), Philippines (n=1), Sri Lanka (n=1), Turkey (n=1), and the UK (n=1). The sample sizes of included studies varied, ranging from 29 to 18,078. A total of 46,802 participants were included across all the studies. Most of the included studies included patients aged 18 or over (n=37), however, three studies focussed on specific age groups (≥60 n=1; ≥70 n=1; ≥80 n=1).

The ACS NSQIP was assessed across nine surgical specialties. General surgeries were most commonly studied (n=14), followed by orthopaedic (n=7), urology, (n=4) neurosurgery (n=4), mixed (n=3), plastic (n=2), thoracic (n=2), vascular (n=2), urogynaecology (n=1), and gynaecology (n=1). While all studies assessed the predictive ability of ACS NSQIP, some studies also compared the predictive ability of ACS NSQIP to other risk prediction tools of interest including the P-POSSUM (n=3) and RCRI (n=2).

Included studies assessed the ACS NSQIP's ability to predict a broad range of outcomes including both composite and individual complications, as well as health care utilisation and recovery outcomes. This included any complication (n=24), serious complications (n=24), major complications (n=1), postoperative complications (n=1), cardiac complications (n=13), pneumonia (n=20), surgical site infection (n=19), urinary tract infections (n=19), venous thromboembolism (n=19), sepsis (n=4), renal complications (n=17), systemic complications (n=1), thyroidectomy complications (n=1). readmission (n=21), discharge to rehabilitation/skilled nursing facility or location other than home (n=18), return to operating room (n=14), length of stay (n=11), and reoperation (n=6) were assessed as health care utilisation and recovery.

2.3.2 CPET

The Cardiopulmonary exercise testing (CPET) is an objective preoperative assessment of exercise capacity and aims to identify the causes of exercise limitation. The test itself consists of four main phases: rest, unloaded cycling, ramp exercise, and recovery. The dynamic metabolic challenge imposed by peri-operative CPET provides an objective means of evaluating exercise capacity. It can be used to evaluate chronic comorbidities and may enable identification of new pathology that requires treatment, optimization, or both preoperatively. It is used for a variety of different surgeries including abdominal, colorectal, urological, hepatobiliary, liver, bariatric, vascular, thoracic and oesophageal-gastric and provides scores for VO2 peak and anaerobic threshold (Levett et al., 2018).

A total of 17 studies evaluated the predictive ability of the CPET tool. Studies were published between 2008 and 2024 and were from the UK (n=12), USA (n=1), Sweden (n=1), China (n=1), Russia (n=1), and Australia (n=1). The sample sizes of included studies varied, ranging from 32 to 703. A total of 3,057 participants were included across all the studies. Most of the included studies included patients aged 18 or older, however three studies focused on specific age groups (≥40 (n=1), ≥65 or younger with a comorbidity (n=1), ≥70 or younger with a comorbidity (n=1)).

Studies assessed CPET across six surgical specialties. General surgeries were the most commonly studied (n=11), followed by cardiothoracic (n=2), ENT (n=1), mixed (n=1), urology (n=1) and vascular (n=1). One study compared the predictive ability of CPET to the ASA classification system.

Studies assessed the predictive ability of CPET predominantly for complications, with one study evaluating its ability to predict critical care unit admissions (n=1). Complication outcomes were all complications (n=3), major complications (n=1), cardiopulmonary complications (n=2), pulmonary complications (n=1), postoperative morbidity (n=5), cardiovascular complications (n=2), cardiopulmonary morbidity (n=1), postoperative complications (n=2).

2.3.3 RCRI

The Revised Cardiac Risk Index for Pre-Operative Risk (RCRI) is used in patients ≥45 years old (or 18 to 44 years old with significant cardiovascular disease) undergoing elective non-cardiac surgery or urgent/semi-urgent (non-emergent) non-cardiac surgery (MDCALC, 2024). It predicts based on six risk factors: high-risk surgery, history of ischemic heart diseases, history of congestive heart failure, history of cerebrovascular disease, pre-operative treatment with insulin and pre-operative creatinine >2 mg/dL / 176.8 µmol/L and provides a single overall score for predicted postoperative cardiac complications (Lee et al.,1999).

A total of 16 studies evaluated the predictive ability of the RCRI tool. Studies were published between 2010 and 2024. Study regions included USA (n=9), Austria (n=1), Canada (n=1), Denmark (n=1), Germany (n=1), Tanzania (n=1), the Philippines (n=1), and a dataset of multiple African countries (n=1). The sample sizes of included studies varied, ranging from 225 to 447,352. A total of 743,218 participants were included across all the studies. Most of the included studies included patients aged 18 or older, however, five studies focused on specific age groups (≥25 n=1; ≥45 n=1; ≥50 n=1; ≥65 n=2).

The RCRI was assessed across four surgical specialties. Mixed surgeries were the most commonly studied (n=9), followed by orthopaedic (n=4), vascular (n=2) and urology (n=1).

While all studies assessed the predictive ability of RCRI, some studies also compared the predictive ability of RCRI to other risk prediction tools of interest including the ASA classification system (n=3) and ACS NSQIP (n=3).

Although the RCRI typically predicts the risk of major cardiac events, identified studies assessed its ability to predict a range of both cardiac and non-cardiac outcomes. Outcomes included major cardiac complications (n=8), adverse cardiac event (n=3), myocardial infarction or cardiac arrest within 30 days (n=2), postoperative cardiac morbidity (n=1), pulmonary complications (n=1), non-cardiac complications (n=1), and a composite morbidity endpoint consisting of cardiac and non-cardiac complications (n=1).

2.3.4 POSSUM

The Physiological and Operative Severity Score for the enUmeration of Mortality and Morbidity (POSSUM) estimates morbidity and mortality for emergency and general surgery patients through a series of questions around patient physiological score which looks at age, cardiac and respiratory function, GCS, blood work, ECG, and operative severity, which interrogates the procedure being undertaken (MDCALC, 2024; Copeland et al.,1991). It provides a single overall score for predicted morbidity and a single overall score for predicted mortality.

A total of 14 studies evaluated the predictive ability of the POSSUM tool. Studies were published between 2003 and 2024. Study regions included China (n=5), UK (n=2), Korea (n=2), Germany (n=1), Romania (n=1). The region of three studies was unclear. The sample sizes of included studies varied, ranging from 100 to 1,262. A total of 5,653 participants were included across all the studies. Most of the included studies included patients aged 18 or older, however, one study focused on patients over 60 years.

POSSUM was assessed across six surgical specialties categories. General surgeries were the most commonly studied (n=9), followed by mixed (n=1), neurosurgery (n=1), orthopaedic (n=1), vascular (n=1) and cardiothoracic (n=1). One study compared the predictive ability of POSSUM to P-POSSUM and SORT.

Outcomes of interest were grouped as complications. These included morbidity or postoperative complications (n=14), pulmonary complications (n=1), cardiovascular complications (n=1), infectious complications (n=1) and nonfatal complications (n=1). No healthcare utilisation and recovery outcomes were assessed.

2.3.5 ASA classification system

The American Society of Anesthesiologists Physical Status (ASA) Classification System assesses a patients' pre-anaesthesia co-morbidities. While it should not be used alone, when used with other factors the classification system can be used to predict peri-operative risk and provides an ASA classification (I-VI) where a higher grade suggests higher operative risk (Saklad, 1941; American Society of Anesthesiologists, 2020).

A total of 13 studies assessed the ability of ASA classification system to accurately predict complications after surgery. Studies were published between 2006 and 2024. Study regions included the USA (n=9), Canada (n=1), Germany (n=1), Italy (n=1), and Tanzania (n=1). A total of 584,743 participants were included across all the studies, with sample sizes ranging from 32 to 52,066. The majority of studies included patients aged 18 or over but one study focused on patients aged 65 or older. While all studies assessed the predictive ability of the ASA classification system, some studies compared this to other risk prediction tools of interest

including the RCRI (n=3), ARISCAT (n=1), CPET (n=1), and POSSUM (n=1) (See comparison of tools section, Section 3.1).

Orthopaedic surgeries were most commonly studied (n=6), followed by studies that included a mix of surgery types (n=3), vascular surgeries (n=2), urology surgeries (n=1), and general surgeries (n=1). Outcomes of interest included a range of composite complications (grouped complications) such as, complications, cardiac complications, and pulmonary complications. A number of healthcare utilisation and recovery outcomes including, readmission, length of stay, and discharge to higher level of care or discharge not to home, were also reported.

2.3.6 Apfel Score (PONV)

The Apfel score predicts and provides an overall score for the risk of experiencing postoperative nausea and vomiting (PONV) through four questions looking at gender, smoking status, history of motion sickness or PONV, and use of postoperative opioids among patient's undergoing general anaesthesia (Apfel et al.,1999).

A total of nine studies evaluated the ability of the Apfel score to predict postoperative nausea and vomiting. Studies were published between 1999 and 2024. Study regions included Germany (n=2), China (n=2), Germany & Finland (n=1), Turkey (n=1), France (n=1), Japan (n=1), and Taiwan (n=1). The sample sizes of included studies varied, ranging from 100 to 2,702. A total of 8,804 participants were included across all the studies and all studies included adults aged 18 or over.

Studies were included across three of the eleven surgical specialties with mixed surgeries being the most commonly studied (n=6), followed by general (n=2), and cardiothoracic (n=1). As the Apfel score only predicts the risk of postoperative nausea and vomiting no other outcomes were reported.

2.3.7 P-POSSUM

The Portsmouth Physiological and Operative Severity Score for the enUmeration of Mortality and Morbidity (P-POSSUM) is a modification of the POSSUM system, that uses 12 physiological score parameters and 6 operation severity parameters. It was developed to adjust the logistic regression analysis used in POSSUM scoring to better predict mortality (Prytherch et al.,1998). However, P-POSSUM still provides a single overall score for predicted morbidity and a single overall score for predicted mortality.

A total of seven studies assessed the ability of the P-POSSUM to accurately predict complications after surgery. Studies were published between 2018 and 2024. Study regions included Turkey (n=2), China (n=1), Ireland (n=1), Italy (n=1), Romania (n=1), and the UK (n=1). A total of 1,668 participants were included across studies, with sample sizes ranging from 67 to 567. All studies included adults aged 18 or over. Four studies also compared the predictive ability of P-POSSUM to other risk prediction tools of interest (POSSUM and SORT n=1 and ACS NSQIP n=3).

The seven included studies included three different surgery types, these were general surgeries (n=5), ENT (n=1) and gynaecology surgeries (n=1). Outcomes included mostly composite complications. This included, any complications (n=2), severe complications (n=1), morbidity (n=4) and pancreaticoduodenectomy morbidity (n=1).

2.3.8 NRS-2002

The Nutritional Risk Screening 2002 (NRS-2002) predicts risk of malnutrition in hospitalised patients through four screening questions relating to weight, diet and patient ICU status and provides an overall score for malnutritional risk (Kondrup et al., 2003). This risk prediction tool does not appear to be specific to surgery.

A total of five studies evaluated the predictive ability of the NRS-2002. Studies were published between 2012 and 2022. Study regions included Germany (n=2), China (n=1), Turkey (n=1), and India (n=1). The sample sizes of included studies varied, ranging from 219 to 1,193. A total of 2,973 participants were included across all the studies and all included adults aged 18 or over.

Samples included in studies underwent mixed surgeries (n=3), and cardiothoracic surgery (n=2). Included studies assessed the NRS-2002's ability to predict a range of outcomes. These included complications (n=1), postoperative complications (n=2), postoperative pulmonary complications (n=1), anastomotic leakage (n=1), infectious complications (n=1), overall complications (n=1) and major complications (n=1). Healthcare utilisation and recovery outcomes included ICU Stay >4 Days (n=1), ICU Stay >5 Days (n=1), readmission to the ICU (n=1), and delayed hospital discharge (n=1).

2.3.9 ARISCAT

The Assess Respiratory Risk in Surgical Patients in Catalonia (ARISCAT) predicts risk of pulmonary complications after surgery, including respiratory failure. This tool evaluates based on seven risk factors: oxygen saturation levels, respiratory infections, age, anaemia and surgery information such as surgical incision, duration of surgery and emergency procedure. It was developed for patients undergoing general, neuraxial or regional anaesthesia and provides a single overall score for predicted postoperative pulmonary complications (Canet et al., 2010).

A total of four studies assessed the ability of the ARISCAT tool to accurately predict complications after surgery. Studies were published between 2020 and 2024. Study regions included India (n=1), Italy (n=1), USA (n=1) and it was unclear in one study. A total of 3,136 participants were included across studies, with sample sizes ranging from 105 to 2,104. Three studies included patients aged 18 or older while one study focused on patients aged 65 or older. While all studies assessed the predictive ability of the ARISCAT tool, one study also compared the predictive ability of ARISCAT to the ASA classification system.

Studies included general surgeries (n=2), oral and maxillofacial surgeries (n=1) and cardiothoracic (n=1). The outcome of interest in all studies was postoperative pulmonary complications (n=3).

2.3.10 CFS

The Clinical Frailty Scale (CFS) was originally developed to summarise the overall level of fitness or frailty of an older adult. It measures frailty to predict survival, medium- and long-term outcomes (mortality and need for institutional care) through an inclusive 9-point scale and provides an overall frailty score. Higher scores indicate greater risk. It focuses on items including mobility, balance, use of walking aids, and the ability to eat, dress, shop, cook, and bank (Rockwood et al., 2020). This risk prediction tool does not appear to be specific to surgery.

A total of four studies evaluated the predictive ability of the CFS. Studies were published between 2020 and 2023. Study regions included China (n=2), Canada (n=1), and Taiwan (n=1). The sample sizes of included studies varied, ranging from 82 to 27,027. A total of 27,497 participants were included across all the studies. All four studies focused on specific age groups (≥65 n=3; ≥75 n=1).

Studies mainly included mixed surgeries (n=3) but also urology (n=1). Outcomes of interest included complications: major complication (n=1) and postoperative complications (n=1); and healthcare utilisation and recovery: prolonged hospital stays (10+ days) (n=1), unplanned hospital readmission (30 days) (n=1), long-term hospitalisation (90 days) (n=1), and long-term care admission (1 year) (n=1).

2.3.11 SORT

The Surgical Outcome Risk Tool (SORT) comprises of six variables (ASA classification system grade, urgency of surgery, high-risk surgical specialty, surgical severity, cancer and age 65 years or over) and provides an overall risk prediction score of death within 30 days of surgery. It was developed for adults undergoing surgery (Protopapa et al., 2014).

A total of two studies assessed the ability of the SORT tool to accurately predict complications after surgery. Studies were published between 2023 and 2024. Study regions included China (n=1) and Turkey (n=1). A total of 445 participants were included across studies, with sample sizes ranging from 96 to 349. All studies included adults aged 18 or over. While both studies assessed the predictive ability of the SORT tool, one study compared the predictive ability of SORT to other risk prediction tools of interest including POSSUM and P-POSSUM.

Studies included general (n=1) and gynaecology surgeries (n=1). Outcomes of interest included complications; early complication (n=1) and healthcare utilisation & recovery; postoperative admission to ICU (n=1).

2.3.12 DASI

The Duke Activity Status Index (DASI) is a 12-item, self-administered questionnaire that measures functional capacity and assesses quality of life developed for use during preoperative assessments (Hlatky et al., 1989). It provides a single overall score for functional capacity.

One study was identified that assessed the ability of DASI to accurately predict complications after surgery. The study was published in 2024 and was conducted in the UK assessing 4,199 participants aged 70 years or above. Surgeries included a mix of non-cardiac surgeries. Outcomes of interest included complications, more specifically the need for a blood transfusion after surgery.

Table 2: Summary of the available external validation studies for each risk prediction tool

(For a detailed description of the individual outcomes reported across studies see the risk prediction tool summaries in Section 2.2).

Tool (No. of studies) Surgical disciplines	**Sample size (range)**	**Predicted risk**	**Surgical Specialty** for which external valuation studies were conducted										
			General	Mixed	Urology	Plastic	ENT	Gynaecology	Neurosurgery	Orthopaedic	Vascular	Oral & Maxillofacial	Cardiothoracic
ACS NSQIP (40) Any procedure, in most surgical subspecialties	46,802 (29-18,078) Aged ≥ 60y (n=1) Aged ≥ 70y (n=1) Aged ≥ 80y (n=1)	Complications [1]	14	3	4	2		2	4	7	2		2
		Healthcare Utilisation & Recovery[2]	10	2	4	2		1	3	4	1		2
CPET (17) Abdominal, colorectal, urological, hepatobiliary, liver, bariatric, vascular, thoracic, oesophageal gastric.	3057 (32-703) Aged ≥40 (n=1) Aged ≥65 (n=1)* Aged ≥70 (n =1)*		General	Mixed	Urology	Plastic	ENT	Gynaecology	Neurosurgery	Orthopaedic	Vascular	Oral & Maxillofacial	Cardiothoracic
		Complications [1]	11	1			1				1		2
		Healthcare Utilisation & Recovery[2]			1								
RCRI (16) Elective non-cardiac surgery or urgent/semi-urgent (non-emergent) non-cardiac surgery.	743,218 (225- 447352) Aged ≥25 (n=1) Aged ≥45 (n=1) Aged ≥50 (n=1) Aged ≥65 (n=2)		General	Mixed	Urology	Plastic	ENT	Gynaecology	Neurosurgery	Orthopaedic	Vascular	Oral & Maxillofacial	Cardiothoracic
		Cardiac Complications		8	1					3	2		
		Non-cardiac Complications		1						1			
POSSUM (14) Emergency and elective general surgical procedures.	5,653 (100-1262) Aged >60 (n=1)		General	Mixed	Urology	Plastic	ENT	Gynaecology	Neurosurgery	Orthopaedic	Vascular	Oral & Maxillofacial	Cardiothoracic
		Complications [1]	9	1					1	1	1		1
		Healthcare Utilisation & Recovery[2]											
ASA Classification system (13) Not specified	584,743 (32-52,066) Aged ≥65 (n=1)		General	Mixed	Urology	Plastic	ENT	Gynaecology	Neurosurgery	Orthopaedic	Vascular	Oral & Maxillofacial	Cardiothoracic
		Complications [1]	1		1					4	2		
		Cardiac complications		2						1			
		Pulmonary complications		2						1			
		Healthcare Utilisation & Recovery[2]			1					4			
Apfel Score (PONV) (9) Any surgery performed under general anaesthesia.	8804 (100-2702)		General	Mixed	Urology	Plastic	ENT	Gynaecology	Neurosurgery	Orthopaedic	Vascular	Oral & Maxillofacial	Cardiothoracic
		Postoperative nausea and vomiting	2	6									1
P-POSSUM (7) Emergency and elective general surgical procedures.	1,668 (67-567)		General	Mixed	Urology	Plastic	ENT	Gynaecology	Neurosurgery	Orthopaedic	Vascular	Oral & Maxillofacial	Cardiothoracic
		Complications [1]	5				1	1					
		Healthcare Utilisation & Recovery[2]											
NRS-2002 (5) N/A not originally designed for surgery.	2973 (219 - 1193)		General	Mixed	Urology	Plastic	ENT	Gynaecology	Neurosurgery	Orthopaedic	Vascular	Oral & Maxillofacial	Cardiothoracic
		Complications [1]	3										1
		Healthcare Utilisation & Recovery[2]											2
ARISCAT (4) Surgery performed under general, neuraxial, or regional anaesthesia	3136 (105 -2104) Aged ≥65 (n=1)		General	Mixed	Urology	Plastic	ENT	Gynaecology	Neurosurgery	Orthopaedic	Vascular	Oral & Maxillofacial	Cardiothoracic
		Complications [1]	2									1	1
		Healthcare Utilisation & Recovery[2]											
CFS (4) N/A not originally designed for surgery.	27497 (82-27,027) Aged ≥65 (n=3) Aged ≥75 (n=1)		General	Mixed	Urology	Plastic	ENT	Gynaecology	Neurosurgery	Orthopaedic	Vascular	Oral & Maxillofacial	Cardiothoracic
		Complications [1]		1	1								
		Healthcare Utilisation & Recovery[2]		2									
SORT (2) Not specified	445 (96-349)		General	Mixed	Urology	Plastic	ENT	Gynaecology	Neurosurgery	Orthopaedic	Vascular	Oral & Maxillofacial	Cardiothoracic
		Complications [1]	1										
		Healthcare Utilisation & Recovery[2]						1					
DASI (1) Not specified	4199 Aged ≥70 (n=1)		General	Mixed	Urology	Plastic	ENT	Gynaecology	Neurosurgery	Orthopaedic	Vascular	Oral & Maxillofacial	Cardiothoracic
		Complications [1]		1									
		Healthcare Utilisation & Recovery[2]											

Studies validating tools in specific population groups are detailed in the sample size (range) column.

[1] Complications include composite (e.g. any complications, severe complications, morbidity) and individual (e.g. pneumonia, surgical site infection) outcomes.

[2] Healthcare utilisation and recovery include outcomes such as readmissions, length of stay, return to the operating room and more.

*May also include younger participants if they have a comorbidity

Abbreviations: ACS NSQIP - The American College of Surgeons National Surgical Quality Improvement Program Tool; CPET - Cardiopulmonary exercise testing; RCRI - The Revised Cardiac Risk Index; POSSUM- The Physiological and Operative Severity Score for the enUmeration of Mortality and Morbidity; ASA Classification system - The American Society of Anesthesiologists Physical Status Classification System; PONV - Postoperative Nausea and Vomiting; P-POSSUM - The Portsmouth Physiological and Operative Severity Score for the enUmeration of Mortality and Morbidity; NRS-2002 - The Nutritional Risk Screening 2002; ARISCAT - The Assess Respiratory Risk in Surgical Patients in Catalonia; CFS - The Clinical Frailty Scale; SORT - The Surgical Outcome Risk Tool; DASI - The Duke Activity Status Index.

3. Summary of the findings for ACS NSQIP, P-POSSUM, RCRI and the ASA classification system

It became clear that not all of the risk prediction tools were specifically developed for surgery and therefore may not be appropriate for the identification of patients suitable for treatment in low-risk surgical settings such as surgical hubs. Given the large evidence base identified, this summary of findings focusses on four of the risk prediction tools (ACS NSQIP, RCRI, P-POSSUM and the ASA classification system). These tools were selected as being potentially the most impactful at a population level and P-POSSUM was selected over POSSUM for inclusion in the summary as it is the most recent version of the tool. Following the approach utilised by the NICE review (NICE 2020), composite outcomes were used to compare the predictive ability of each tool. Studies that report mortality as part of their composite complications were not included to allow for a more direct comparison. Some studies reported multiple composite outcomes, in this case the largest composite (including the most complications but not including mortality) was taken for the results. A full breakdown of the different composite complications and individual complications reported, along with the findings have been tabulated for each tool in Appendix 3. In addition, where more than one study was reported for a surgical specialty, the findings have also been tabulated in Appendix 3 for each tool.

This section begins with a summary of the findings from the studies that directly compared multiple risk prediction tools (Section 3.1). This is followed by a breakdown of the overall findings for each of the tools of interest (Section 3.2 – 3.5). This includes a look at any findings related to the discrimination, calibration and accuracy of the risk prediction tools, along with findings in relation to the different surgical specialties and a summary of any healthcare utilisation and recovery outcomes that were reported.

Discrimination measures how well a tool differentiates between patients who do and do not experience an event, quantified using the area under the receiver operating characteristic (ROC) curve (AUC) or c-statistic. Within the literature, a c-statistic of 90% or greater is typically considered an excellent level of discriminative ability, a c-statistic of 80% or greater is considered a good level of discriminative ability, a c-statistic of 70% or greater is considered a fair level of discriminative ability, a c-statistic of 60% or greater is considered a poor level of discriminative ability and a c-statistic of 50% or greater is considered a very poor level of discriminative ability (or no better than chance) (NICE 2020; Çorbacıoğlu and Aksel, 2023).

Calibration assesses the agreement between the predicted and observed outcomes, typically presented graphically as observed risks versus predicted risks or in tabular format (Collins et al., 2014). Following the approach utilised by the NICE review, an observed/expected (O/E) ratio of 0.9-1.1 would be considered a fair level of calibration.

In the literature, the accuracy of the risk prediction tools was assessed using the 'Brier Score'. The Brier score is a simultaneous measure of calibration and discrimination. It is reported as a score between 0 and 1. A score of 0 indicates no difference between the predicted and actual outcome, thus indicating the best possible test result. A score of 1 indicates that the test did not predict the outcome. The Brier score is compared with a Brier score cut-off, which is partially based on the incidence in the sample, and a score above the cut-off is considered not useful (Alzahrani et al., 2020). Brier score cut-offs are dependent on the datasets of

individual studies, as the value of the Brier score depends on both the prevalence of the event in the data and the performance of the model (Huang et al., 2021). As such, direct comparisons across studies may not be appropriate, for the purposes of this in-depth summary, the range of the Brier scores reported across studies will be highlighted.

3.1 Studies comparing the predictive ability of multiple risk prediction tools

A total of 9 studies compared three of the four risk prediction tools of interest (See Table 3 below). Three studies compared the predictive ability of the ASA classification system to the RCRI, three studies compared ACS NSQIP with P-POSSUM, and three studies compared ACS NSQIP with the RCRI tool. These assessed each tools' predictive ability using a variety of outcomes in several different surgical disciplines.

ASA and RCRI

When comparing the predictive ability of the ASA classification system and the RCRI the findings were mixed. Of the three studies comparing the ASA classification system and the RCRI, none reported any calibration or accuracy findings, only reporting the discriminative ability of the tools (c-statistics). One study (Chrisant et al., 2024) reported both tools to have a fair discriminative ability in predicting cardiac and pulmonary complications in elective non-cardiothoracic surgery. One study (Bronheim et al., 2018) reported the ASA classification system to have a fair discriminative ability whereas the RCRI was found to have a poor discriminative ability to predict a composite of non-cardiac complications after posterior lumbar decompression. However, a second study from the same author (Bronheim et al.,2019), found the RCRI to have a good discriminative ability for predicting myocardial infarction and cardiac arrest requiring cardiopulmonary resuscitation (CPR), whereas the ASA classification system was found to have a fair and poor discriminative ability, respectively.

ACS NSQIP and P-POSSUM

When comparing the predictive ability of ACS NSQIP and P-POSSUM the findings were mixed. One study reported a poor discriminative ability and inaccuracy for both the P-POSSUM tool and ACS NSQIP after retroperitoneal sarcoma surgery (Angelucci et al., 2024). However, ACS NSQIP was found to have a fair discriminative ability compared to P-POSSUM's poor discriminative ability in another study after gynaecological surgery (Karabulut et al., 2024). Calibration was not reported in either study. Lastly, one study found ACS NSQIP to have a poor discriminative ability compared to P-POSSUM's very poor discriminative ability for predicting complications after hepatobiliary surgery (Sevinyan et al., 2024), however, the calibration was slightly better in the P-POSSUM tool compared with ACS NSQIP for morbidity.

ACS NSQIP and RCRI

Finally, three studies compared the predictive ability of ACS NSQIP and the RCRI and the findings were mixed. One study in four cardiac surgeries (Cohn and Ros, 2018) found good discriminative ability for 'All cardiac complications in-hospital' for both tools. In the same study, the ACS NSQIP tool had a fair discriminative ability in 'Major cardiac complications 30 day' compared to poor discriminative ability for RCRI for the same outcome. For the outcome of 'All cardiac complications 30 day' the ACS NSQIP tool had good discriminative ability compared to a fair discriminative ability in the RCRI tool. This study did not report calibration or accuracy. In the second study comparing ACS NSQIP with the RCRI in vascular surgery (Moses et al., 2019), only calibration was reported. For the outcome of 'Adverse Cardiac

Events' the ACS NSQIP tool reported a ratio of 1.60 of observed to predicted adverse events; the RCRI tool reported a ratio of 2.23 for the same outcome suggesting both tools underpredicted adverse events. In a third study in mixed surgeries (Yap et al., 2018) both tools had excellent discriminative ability for the outcome of 'Cardiac complication/ MACE (major adverse cardiac event)'. This study also reported excellent calibration in the ACS NSQIP tool for the same outcome, while calibration scores were not reported for the RCRI tool.

Table 3. Results of studies comparing multiple tools

Outcome	Reference	Surgery	Risk Prediction Tol				P value
			ASA	P-POSSUM	ACS NSQIP	RCRI	
Morbidity	Karabulut et al., (2024)	General		AUC ▼ Brier Score 0.2567 O/E event ratio 1.48	AUC ◄► Brier Score 0.2156 O/E event ratio 1.36.		ACS NSQIP: 0.186 P-POSSUM: 0.352
Morbidity	Sevinyan et al., (2024)	Gynaecology		AUC ▼ Brier score 0.1359 O0E event ratio 0.183	AUC ▼ Brier score 0.1827 O/E event ratio 0.136		P-POSSUM's overestimation of morbidity was statistically significant (0.018)
Any complication	Angelucci et al., (2024)	General		AUC ▼ Brier Score 0.229	AUC ▼ Brier Score 0.231		-
Severe complication	Angelucci et al., (2024)	General		AUC ▼ Brier Score 0.205	AUC ▼ Brier Score 0.206		-
Any complication	Bronheim et al., (2018)	Spinal	AUC ◄►			AUC ▼	<0.001
Cardiac complications*	Bronheim et al., (2019)	Spinal	AUC ◄►			AUC ▲	0.0004
Cardiac complications	Christant et al., (2024)	Mixed	AUC ◄►			AUC ◄►	0.817
Pulmonary complications	Christant et al., (2024)	Mixed	AUC ◄►			AUC ◄►	0.469
Major cardiac complications 30 day	Cohn and Ros (2018)	Mixed			AUC ◄►	AUC ▼	-
All cardiac complications in-hospital	Cohn and Ros (2018)	Mixed			AUC ▲	AUC ▲	-
All cardiac complications 30 day	Cohn and Ros (2018)	Mixed			AUC ▲	AUC ◄►	-
Adverse cardiac events	Moses (2019)	Vascular			O/E event ratio 1.60	O/E event ratio of 2.23	-
Cardiac complication/ MACE (major adverse cardiac event)	Yap et al., (2018)	Mixed			AUC ▲ Brier Score 0.08	AUC ▲	-

Effect direction: ▲=Good to excellent (80%+) ◄►= Fair (70-79%) ▼=Poor (50-60%)

* Denotes the mean of two individual outcomes reported in a study. These were cardiac arrest requiring CPR and myocardial infarction
O/E: Observed/Expected event ratio
-: information not provided by study

3.2 ACS NSQIP findings

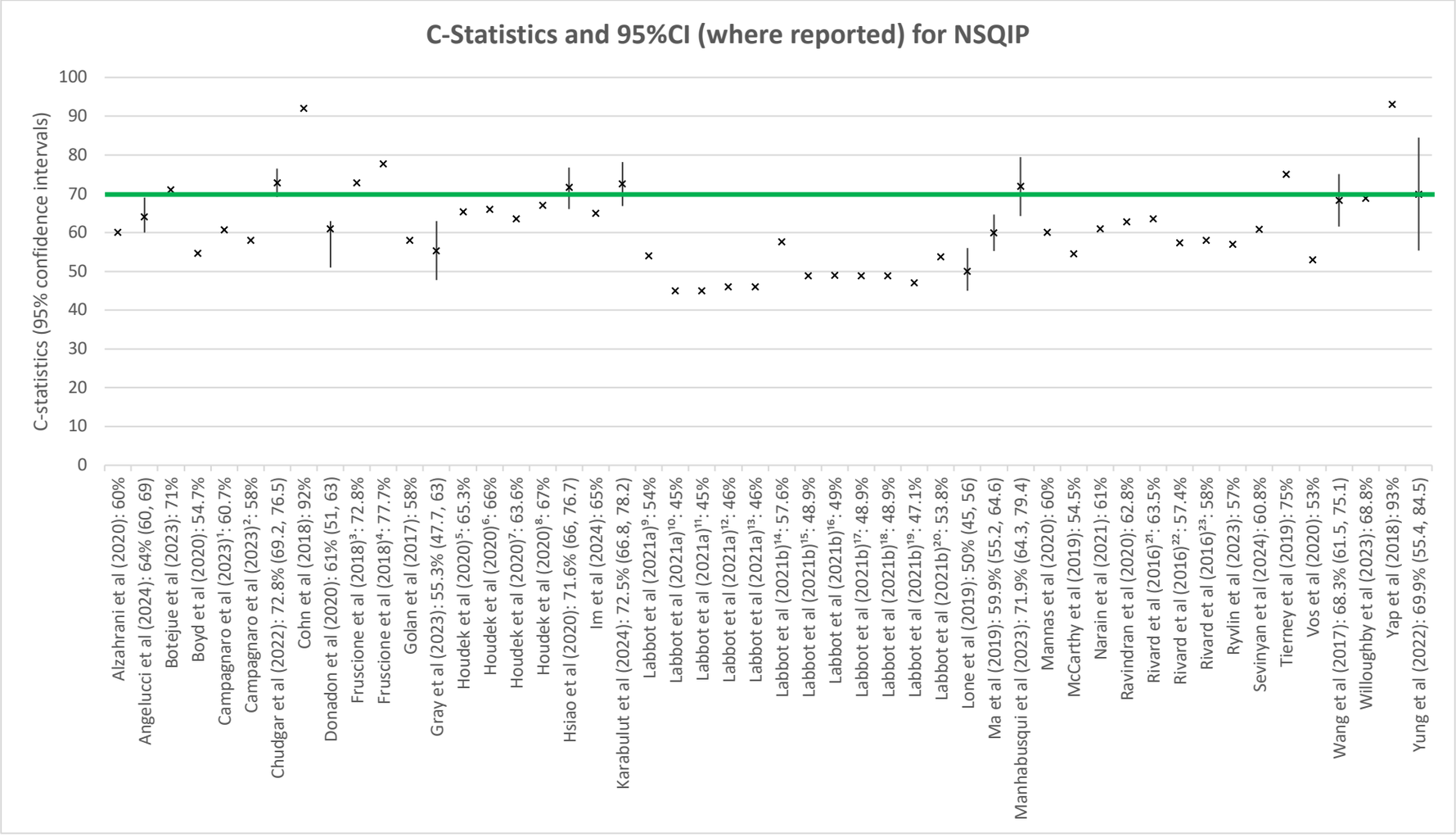
3.2.1 ACS NSQIP Discrimination findings

A total of 33 studies reported c-statistics on composite complications for ACS NSQIP. Two studies (Chudgar et al., 2021; Chudgar et al., 2022) appear to use the same data so have only been reported once. Six studies reported multiple c-statistics for different surgeries (Campagnaro et al., 2023; Fruscione et al., 2018; Houdek et al., 2020; Labbot et al., 2021a; Labbot et al., 2021b; Rivard et al., 2016), details of the different surgeries reported can be seen in Figure 1 which plots the c-statistics taken from each paper. **Overall, c-statistics ranged from 45% to 93% with a median c-statistic of 60.35% suggesting ACS NSQIP had a poor predictive ability for composite complications across all 33 studies.**

The majority of studies were conducted in populations receiving general surgeries (n=12), followed by orthopaedic surgeries (n=6), neurosurgery (n=4), urology (n=3), mixed surgeries (n=3) and plastic surgeries (n=2). One study included gynaecology surgeries, one study included vascular surgeries, and one study included thoracic surgeries. The findings by surgery type can be seen in Table 4.

Table 4. C-statistic results for ACS NSQIP by surgical specialty

Surgical specialty	No of studies	c-statistic range	Median c-statistic	Discriminative ability
General surgery	12	53% - 77.7%	63.15%	Poor
Orthopaedic surgery	6	45% - 71.9%	49.5%	Very poor
Neurosurgery	4	57% - 68.3%	65.3%	Poor
Urology surgery	3	50% - 60%	58%%	Very poor
Mixed surgery	3	54.7% - 93%	92%	Excellent
Plastic surgery	2	69.9% - 75%	72.45%	Fair
Thoracic surgery	1	N/A	72.8%	Fair
Vascular surgery	1	N/A	59.9%	Very poor
Gynaecology surgery	1	N/A	60.8%	Poor



- The bold line represents the cut off for a fair discriminative ability, anything above this point is considered fair or better. [1]Colon resection, [2]Liver only resection, [3]General surgery cholecystectomy, [4]Hepatopancreatobiliary cholecystectomy, [5]Excision of presacral or sacrococcygeal tumor, [6]Laminectomy with exploration and/or decompression of spinal cord and/or cauda equina, without facetectomy, foraminotomy or discectomy (eg, spinal stenosis), 1 or 2 vertebral segments; sacral. [7]Laminectomy for biopsy/excision of intraspinal neoplasm; extradural, sacral. [8]Transperitoneal or retroperitoneal vertebral corpectomy, intradural, lumbar or sacral, for excision of an intraspinal lesion of one vertebral segment. [9]Under Excision Procedures on the Femur and Knee Joint. [10]Arthroplasty, knee, condyle and plateau. [11]Revision of total knee arthroplasty, with or without allograft, 1 component. [12]Revision of total knee arthroplasty, with or without allograft, femoral and entire tibial component. [13]Repair, Revision, and/or Reconstruction Procedures on the Femur [Thigh Region] and Knee Joint. [14]Hemiarthroplasty, [15]Total hip, [16]Conversion to total hip, [17]Revision of total hip, [18]Revision acetabulum, [19]Revision femur, [20]Excision tumour. [21]Gynecological oncology, [22]Tumour debulking, [23]Bowel resection.

Figure 1. ACS NSQIP C-statistic Plot

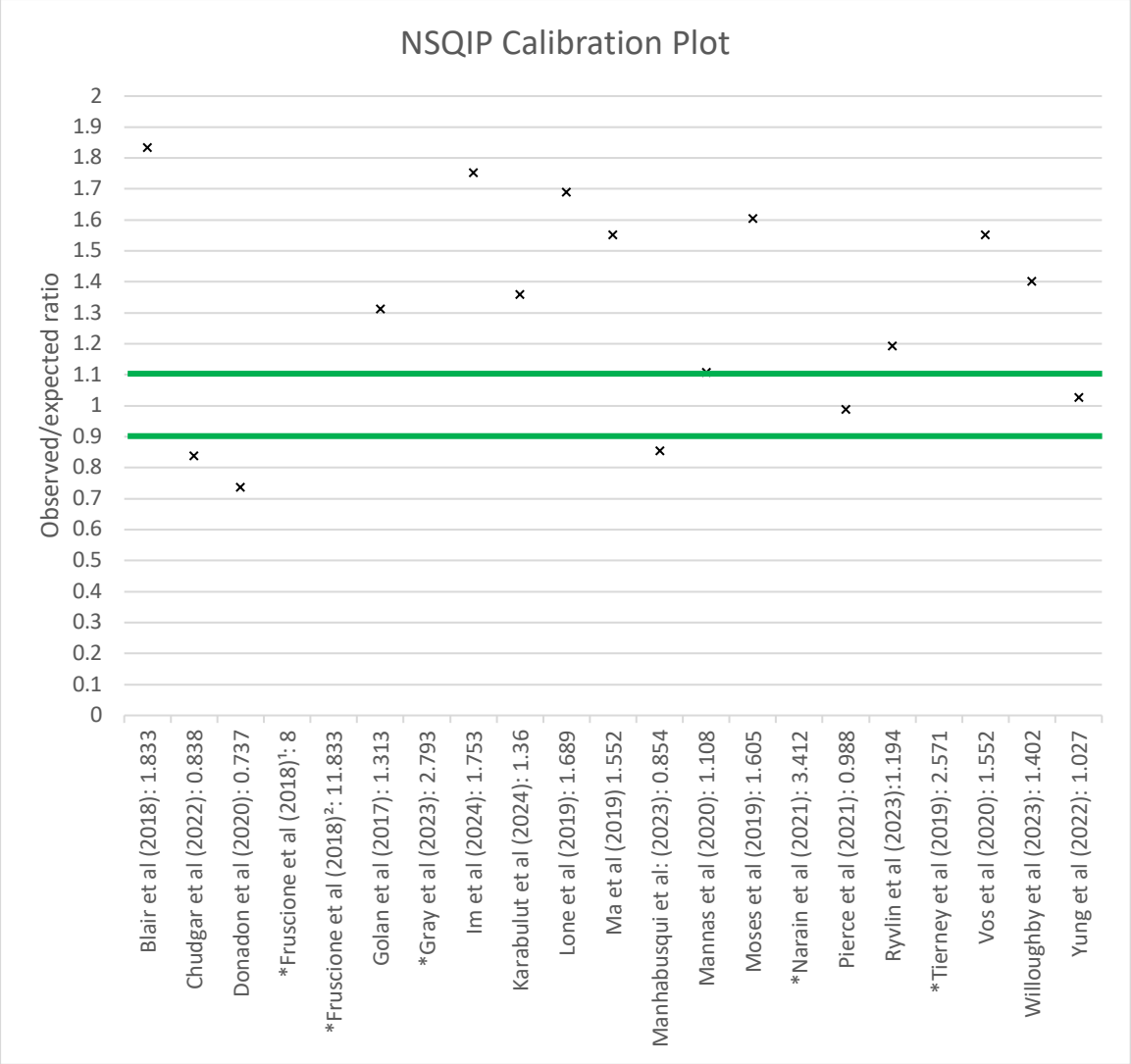
3.2.2 ACS NSQIP Calibration findings

A total of 21 studies reported the number of observed complications (O) and the number of complications predicted (E) by ACS NSQIP (O/E ratio). Two studies (Chudgar et al., 2021; Chudgar et al., 2022) appear to have used the same data, so are only been reported once below. One study reported multiple O/E ratios for different surgeries (Fruscione et al., 2018) details of the different surgeries reported can be seen in Figure 2 which plots the O/E ratios taken from each study The O/E ratio ranged from 0.737 to 11.833. **Overall, the median O/E ratio for ACS NSQIP was 1.552.**

The majority of studies were conducted in populations receiving general surgery (n=5), followed by orthopaedic surgery (n-4), urology surgery (n=4), neurosurgery (n=2), vascular surgery (n=2), plastic surgery (n=2), and one study included thoracic surgery. The calibration findings for ACS NSQIP by surgical specialty are displayed in Table 5.

Table 5. ACS NSQIP calibration findings by surgical specialty

Surgical specialty	No of studies	O/E ratio range	Median O/E ratio
General surgery	5	0.737 - 11.833	2.173
Orthopaedic surgery	4	0.854 - 3.412	1.195
Urology surgery	4	1.108 - 1.833	1.501
Neurosurgery	2	1.194 - 1.753	1.474
Vascular surgery	2	1.552 - 1.605	1.579
Plastic surgery	2	1.027 - 2.571	1.799
Thoracic surgery	1	N./A	0.838



- The bold lines represent the cut off for a fair calibration ratio, any data point between these lines is considered fair or better. *Studies that report an O/E ratio of 2 or higher are not plotted and show considerable underprediction of complications [1]Colon [1]General surgery cholecystectomy, [2]Hepatopancreatobiliary cholecystectomy.

Figure 2. ACS NSQIP calibration plot

3.2.3 ACS NSQIP Accuracy findings

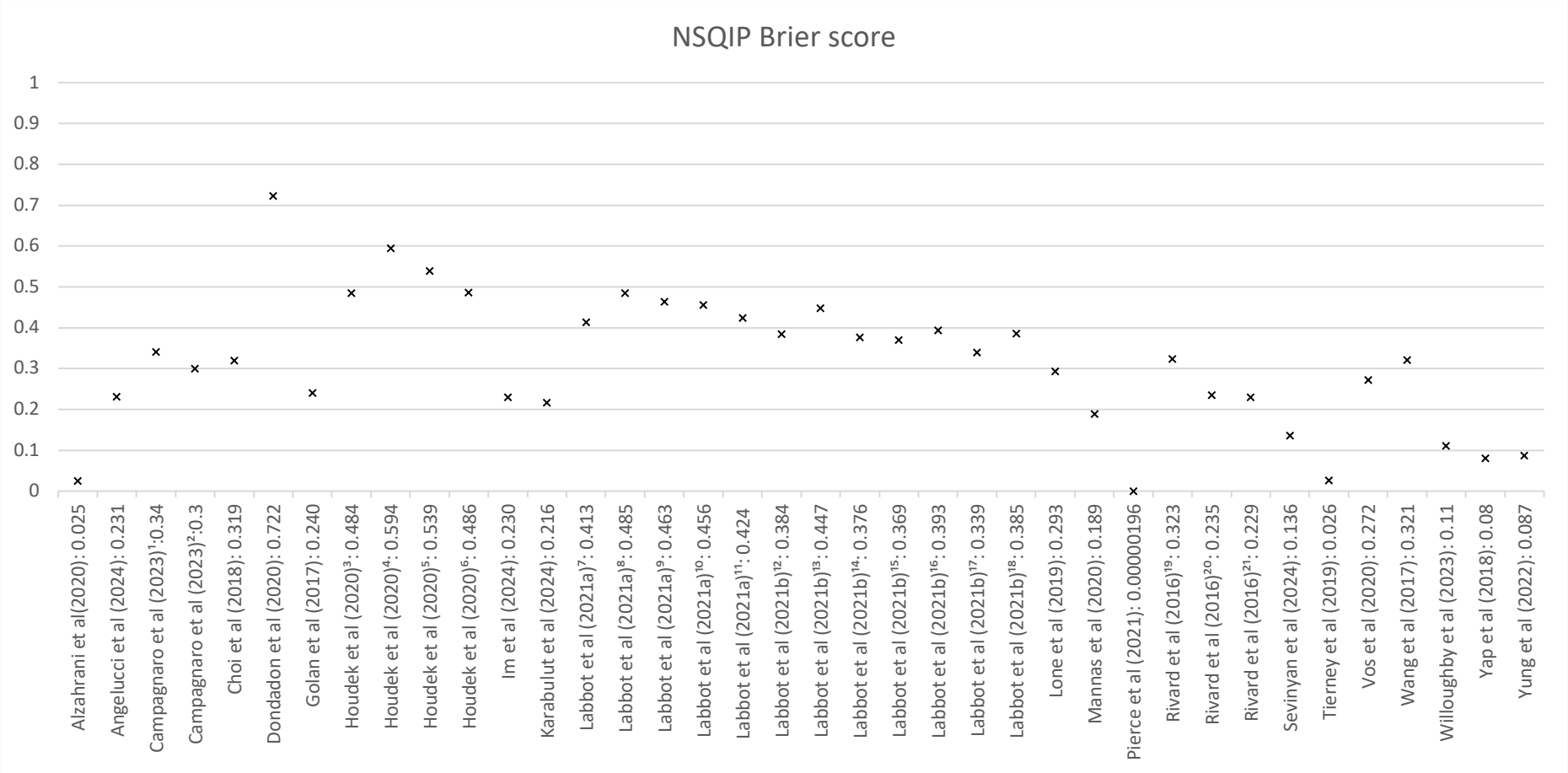
A total of 22 studies assessing ACS NSQIP reported a Brier score for composite complications (See Figure 3). **Overall, the Brier scores for ACS NSQIP ranged from 0.00000196 to 0.722.**

The majority of studies were conducted in populations receiving general surgery (n=8), followed by orthopaedic surgery (n=4), neurosurgery (n=3), urology surgery (n=3), plastic surgery (n=2), one study included mixed surgery, and one study included gynaecology surgery. Accuracy findings by surgery type can be seen in Table 6.

Table 6. ACS NSQIP Brier results by surgical specialty

Surgical specialty	No of studies	Brier score range
General surgery	8	0.025 - 0.722
Orthopaedic surgery	4	0.00000196 - 0.485

Neurosurgery	3	0.230 - 0.594
Urology surgery	3	0.189 - 0.293
Plastic surgery	2	0.026 - 0.087
Mixed surgery	1	0.08
Gynaecology surgery	1	0.136



[1]Colon resection. [2]Liver only resection. [3]excision of presacral or sacrococcygeal tumour. [4]laminectomy with exploration and/or decompression of spinal cord and/or cauda equina, without facetectomy, foraminotomy or discectomy (e.g., spinal stenosis), 1 or 2 vertebral segments; sacral. [5]laminectomy for biopsy/excision of intraspinal neoplasm; extradural, sacral. [6]transperitoneal or retroperitoneal vertebral corpectomy, intradural, lumbar or sacral, for excision of an intraspinal lesion of one vertebral segment. [7]Under Excision Procedures on the Femur and Knee Joint, [8]Arthroplasty, knee, condyle and plateau), [9]Revision of total knee arthroplasty, with or without allograft, 1 component, [10]Revision of total knee arthroplasty, with or without allograft, femoral and entire tibial component ,[11]Repair, Revision, and/or Reconstruction Procedures on the Femur [Thigh Region] and Knee Joint. [12]Hemiarthroplasty, [13]Total hip, [14]Conversion to total hip, [15]Revision of total hip, [16]Revision acetabulum, [17]Revision femur, [18]Excision tumour hip. [19]Gynecological oncology, [20]Tumour debulking, [21]Bowel resection.

Figure 3. ACS NSQIP Brier score plot

3.2.4 ACS NSQIP Healthcare utilisation and recovery outcomes

A total of 29 of the 40 studies assessing the predictive ability of ACS NSQIP reported five healthcare utilisation and recovery outcomes. The findings for all healthcare utilisation and recovery outcomes can be seen in Appendix 3. This included readmission, return to operating room, length of stay, discharge to a facility other than home and adverse discharge. The studies included all surgical specialities, apart from ENT and oral & Maxillofacial surgery. While discrimination, calibration and accuracy scores were reported, they were not consistently reported by each study, with the c-statistic being the most commonly reported measure. The findings varied considerably across studies with wide ranges being reported however, overall ACS NSQIP was found to be very poor at predicting readmission (59% across 24 studies); return to operating room (57.8% across 25 studies) and length of stay (56% across 7 studies). ACS NSQIP was found to have a fair discriminative ability for discharge to a facility other than home across 17 studies (ranging from 58.5% to 90% with a median c-statistic of 70%) and a fair calibration ratio for discharge to a facility other than home (ranging from 0.129 to 3.630, with a median O/E ratio of 0.938) however, the large ranges highlight these findings were not consistently reported across studies and further evidence would be needed.

3.2.5 Bottom line summary for ACS NSQIP

There is evidence to suggest that when looking at composite complications ACS NSQIP had a poor discriminative ability across 33 studies overall (median c-statistic of 60.35%). When looking at the discriminative ability of ACS NSQIP by surgical specialty the c-stats varied, ranging from very poor (orthopaedic, urology and vascular surgery) to excellent (mixed surgery). However, the findings for the individual surgical specialties are limited and showed wide ranges across studies, reducing the confidence in the findings. As such, the findings should be interpreted with caution.

Calibration findings suggest ACS NSQIP under predicted complications overall across 21 studies (median O/E ratio of 1.552). ACS NSQIP was found to under predict complications across all surgical specialties assessed apart from thoracic surgery where it was found to over predict complications (O/E ratio of 0.838), however this finding was only reported by one study and so further evidence would be needed to confirm this. When looking at accuracy the Brier scores for ACS NSQIP ranged from 0.00000196 to 0.722 across 22 studies.

The findings for ACS NSQIP's ability to predict healthcare utilisation and recovery outcomes also varied considerably across studies with wide ranges being reported, suggesting further evidence is needed. However overall, there is evidence to suggest ACS NSQIP was very poor at predicting readmission; return to operating room and length of stay and fair at predicting discharge not to home.

The evidence directly comparing the predictive ability of ACS NSQIP to P-POSSUM or the RCRI was limited, and the findings appear to be mixed, suggesting further evidence is needed.

3.3 P-POSSUM findings

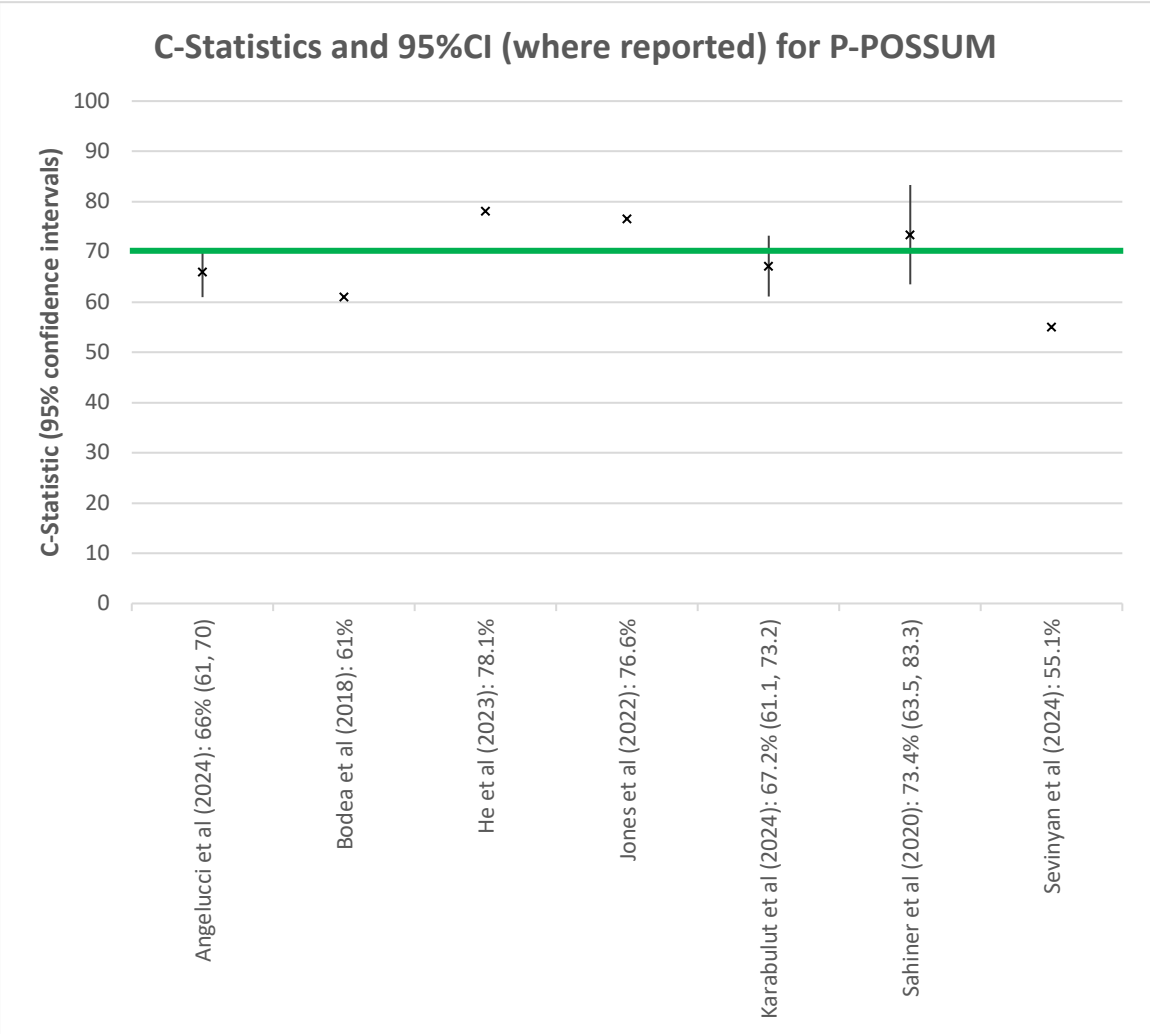
3.3.1 P-POSSUM Discrimination findings

A total of 7 studies reported c-statistics for the P-POSSUM on composite complications (See Figure 4). **Overall c-statistics ranged from 55.1%-78.1% with a median c-statistic of 67.2% suggesting P-POSSUM had a poor predictive ability for composite complications across all seven studies.**

The majority of studies were conducted in populations receiving general surgeries (n=5), one study included ENT surgeries, and one study included gynaecology surgeries. The findings by surgery type can be seen in Table 7.

Table 7. P-POSSUM c-statistics by surgical specialty

Surgical specialty	No of studies	c-statistic range	Median c-statistic	Discriminative ability
General surgery	5	61% - 78.1%	67.2%	Poor
Gynaecology surgery	1	N/A	55.1	Very poor
ENT surgery	1	N/A	76.6%	Fair



- The bold line represents the cut off for a fair discriminative ability, anything above this point is considered fair or better.

Figure 4. P-POSSUM C-Statistics plot

3.3.2 P-POSSUM Calibration findings

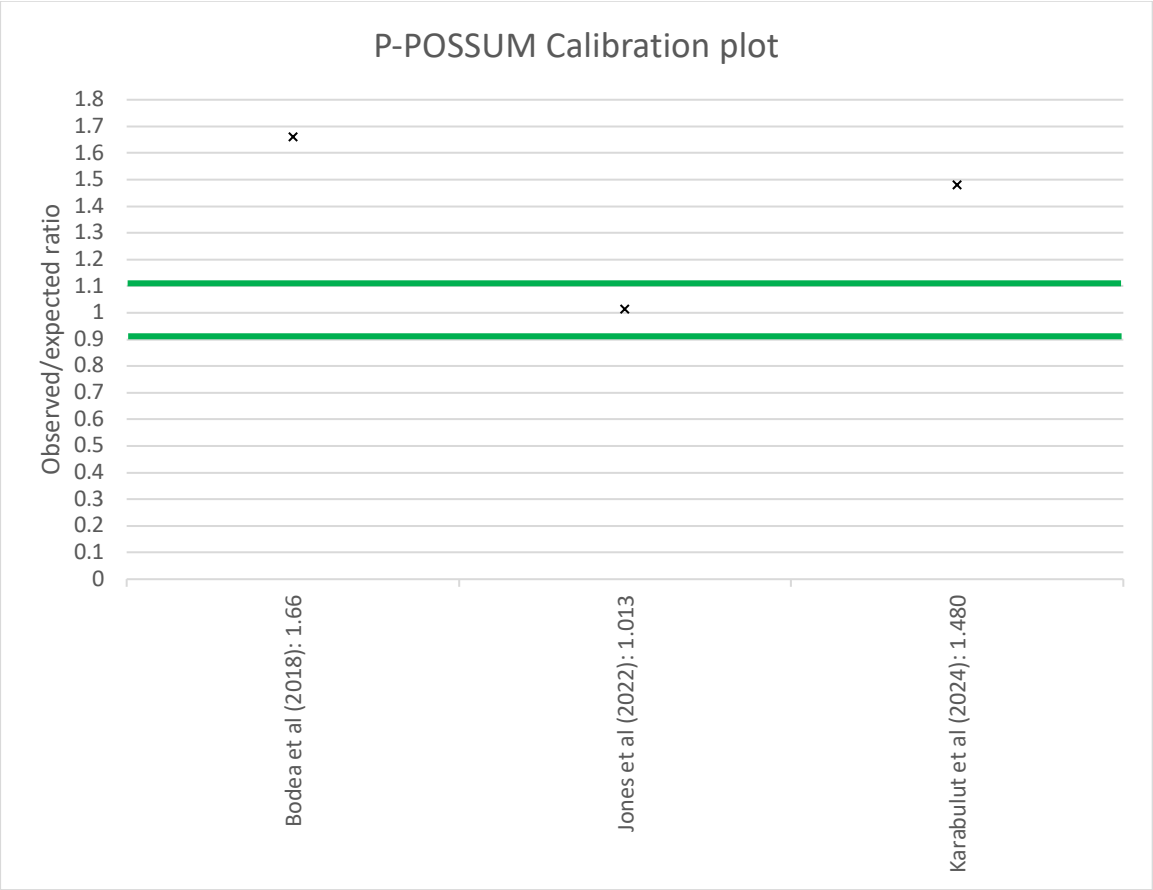
A total of three studies reported the number of observed complications and the number of complications predicted by P-POSSUM (O/E ratio) (See Figure 5). **Overall, the median O/E ratio of P-POSSUM for composite complications was 1.480 across all three studies.**

Two of the studies included populations receiving general surgery and one study included ENT surgery. The P-POSSUM calibration findings by surgical specialties can be seen in Table 8.

Table 8. P-POSSUM calibration findings by surgical specialty

Surgical specialty	No of studies	O/E ratio range	Median O/E ratio
General surgery	2	1.48 - 1.66	1.57
ENT surgery	1	N/A	1.013*

*Fair level of calibration.



- The bold lines represent the cut off for a fair calibration ratio, any data point between these lines is considered fair or better.

Figure 5. P-POSSUM calibration plot

3.3.3 P-POSSUM Accuracy findings

A total of three studies assessing P-POSSUM reported a Brier score for composite complications (See Figure 6). **Overall, the Brier scores for P-POSSUM ranged from 0.183 – 0.257 across all three studies.**

Two of the studies included populations receiving general surgery and one study included gynaecology surgery. Accuracy findings by surgery type can be seen in Table 9:

Table 9. P-Possum Brier score by surgical specialty

Surgical specialty	No of studies	Brier score range
General surgery	2	0.229 - 0.257
Gynaecology surgery	1	0.183

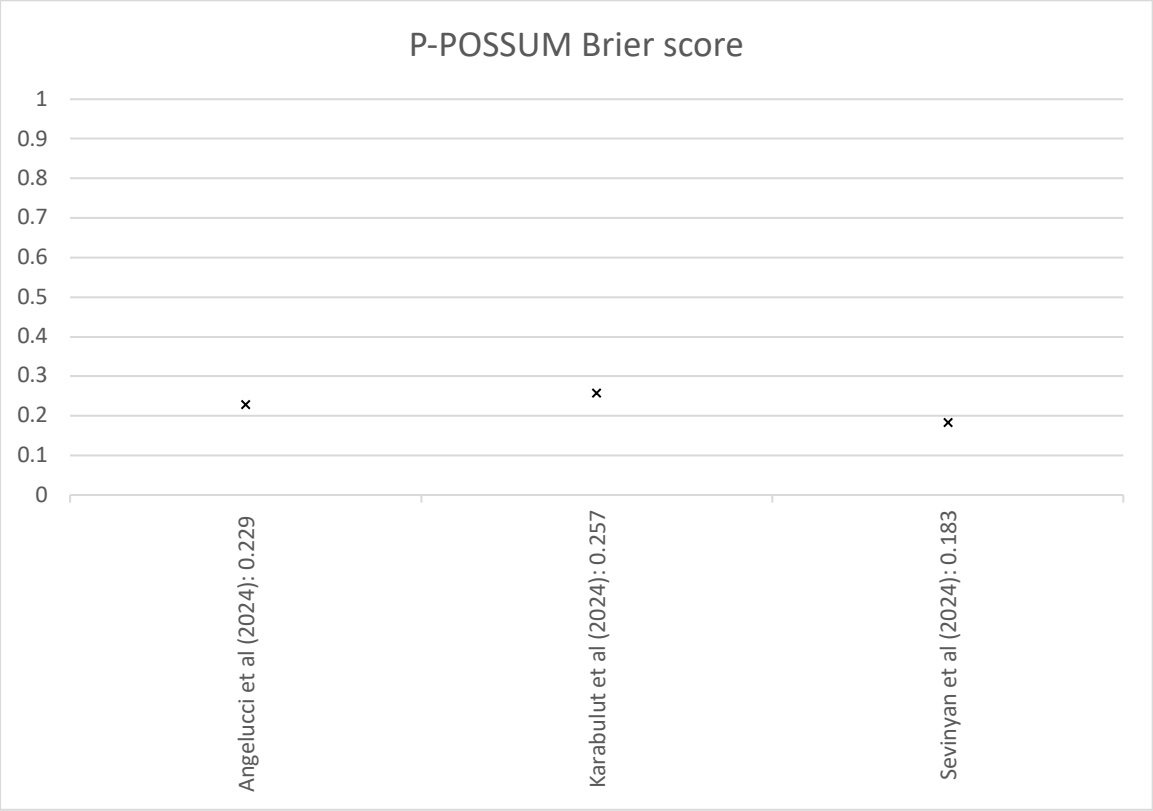


Figure 6. P-POSSUM Brier score plot

3.3.4 P-POSSUM healthcare utilisation and recovery outcomes

No studies assessing the predictive ability of P-POSSUM reported any healthcare utilisation and recovery outcomes.

3.3.5 Bottom line summary for P-POSSUM

There is evidence to suggest that when looking at composite complications P-POSSUM had a poor discriminative ability across seven studies (median c-statistic 67.2%). When looking at the discriminative ability of P-POSSUM by surgical specialty the c-stats varied, ranging from

very poor in one study after gynaecology surgery to fair in one study after ENT surgery. However, the findings for the individual surgical specialties are very limited, reducing the confidence in the findings, and further evidence is needed. Calibration findings suggest P-POSSUM under predicted complications across three studies overall (with a median O/E ratio of 1.480). P-POSSUM was found to under predict complications in general surgery across two studies but was found to have a fair calibration score in one study after ENT surgery (1.013). In terms of accuracy the Brier scores for P-POSSUM ranged from 0.183 – 0.257 across three studies.

No studies assessing the predictive ability of P-POSSUM reported any healthcare utilisation and recovery outcomes. The evidence directly comparing the predictive ability of P-POSSUM compared to ACS NSQIP was limited and the findings appear to be mixed, suggesting further evidence is needed.

3.4 RCRI findings

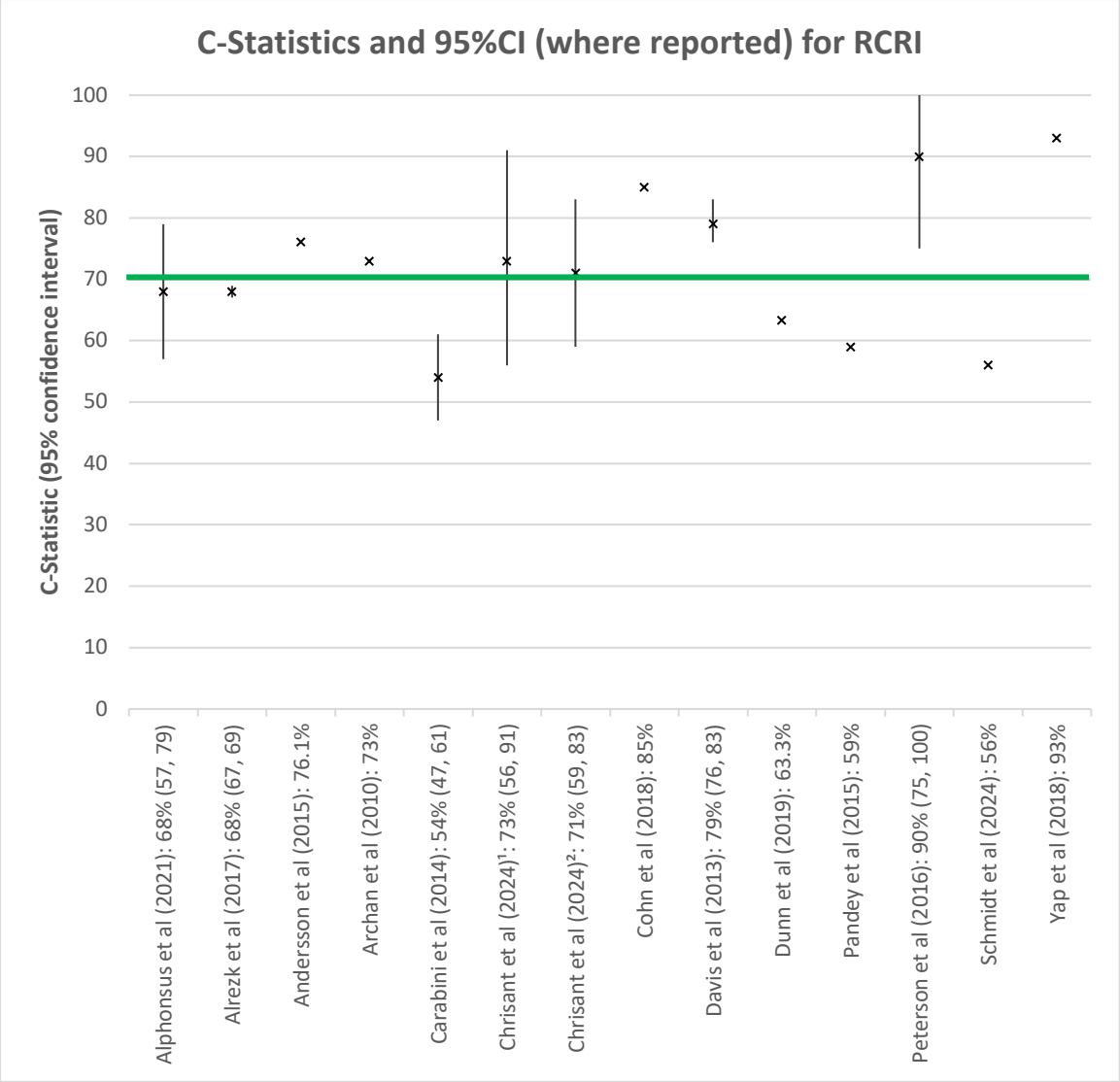
3.4.1 RCRI Discrimination findings

A total of 13 studies reported c-statistics on composite complications for the RCRI tool (See Figure 7). One study reported c-statistics for the predictive ability of RCRI for cardiac complications and pulmonary complications (Chrisant et al, 2024). **C-statistics ranged from 54%-93% with a median c-statistic of 72% suggesting the RCRI had a fair predictive ability for composite complications across all 13 studies.**

The majority of studies were conducted in populations receiving mixed surgeries (n=9), two studies included orthopaedic surgeries, one study included vascular surgeries, and one study included urology surgeries. Findings by surgery type can be seen in Table 10.

Table 10. RCRI c-statistics by surgical specialty

Surgical specialty	No of studies	c-statistic range	Median c-statistic	Discriminative ability
Mixed	9	56% - 93%	72%	Fair
Orthopaedic	2	54% - 90%	72%	Fair
Vascular	1	N/A	73%	Fair
Urology surgery	1	N/A	63.3%	Poor



- The bold line represents the cut off for a fair discriminative ability, anything above this point is considered fair or better. [1]Cardiac complications, [2]Pulmonary complications

Figure 7. RCRI C-Statistics plot

3.4.2 RCRI Calibration findings

Only one study assessing the RCRI reported O/E ratios (Moses et al., 2019). When looking at the RCRI'S ability to predict adverse cardiac events in vascular surgeries, **the O/E ratio was 2.23.**

Table 11 RCRI calibration findings by surgical specialty

Surgical specialty	No of studies	O/E ratio range	Median O/E ratio
Vascular surgery	1	N/A	2.23

3.4.3 RCRI Accuracy findings

None of the 15 studies assessing the RCRI reported a Brier score.

3.4.4 RCRI Healthcare utilisation and recovery outcomes

A total of two of the 16 studies assessing the predictive ability of RCRI reported on two healthcare utilisation and recovery outcome (readmission and reoperation) after orthopaedic surgery (Bronheim et al., 2018) or mixed surgeries (Schmidt et al., 2024). Only c-statistics were reported. The RCRI was found to have a poor discriminative ability for readmission (45.7% to 83.5%, with a median c-statistic of 64.6%) in two studies and good discriminative ability for reoperation in one study (c-statistic 85%). However as only one or two studies reported these outcomes, further evidence would be needed to draw firm conclusions.

3.4.5 Bottom line summary for RCRI

When looking at composite complications, there is evidence to suggest the RCRI had a fair discriminative ability across 13 studies (median c-statistic of 72%). When looking at the discriminative ability of the RCRI by surgical specialty the c-stats varied, ranging from poor in one study after urology surgery to fair after mixed (n=9), orthopaedic (n=2) and vascular surgery (n=1). However, the findings for the individual surgical specialties are limited and showed wide ranges across studies, reducing the confidence in the findings. Calibration findings suggest the RCRI under predicted complications in the one study that reported this outcome (O/E ratio of 2.23). None of the studies assessing RCRI reported accuracy (Brier) scores.

The findings for the RCRI's ability to predict healthcare utilisation and recovery outcomes was also very limited, suggesting further evidence is needed. However, there is evidence to suggest the RCRI had a poor discriminative ability for predicting readmission and a good discriminative ability for predicting reoperation.

The evidence directly comparing the predictive ability of the RCRI to the ASA classification system or to ACS NSQIP was limited and the findings appear to be mixed, suggesting further evidence is needed.

3.5 ASA classification system findings

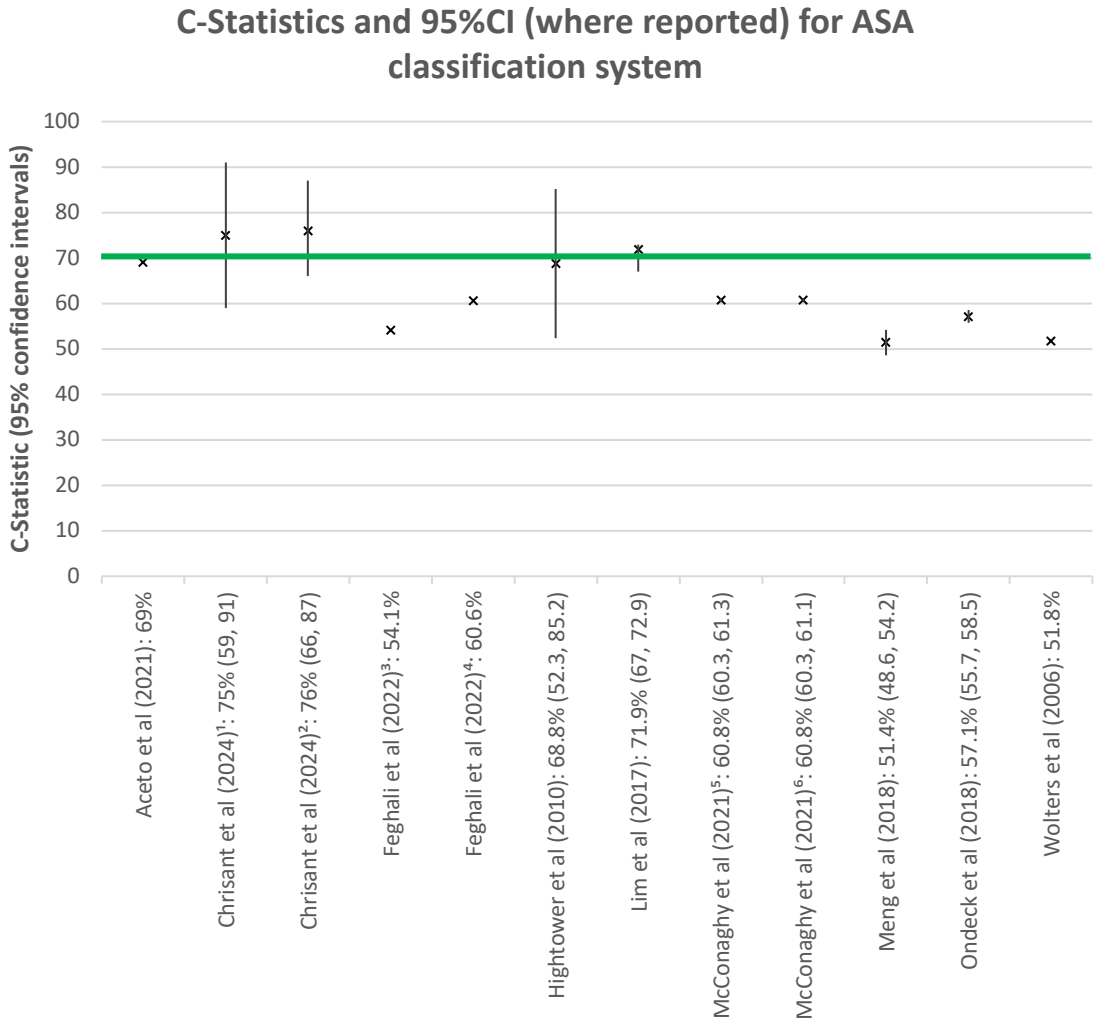
3.5.1 ASA classification system Discrimination findings

A total of nine studies reported c-statistics on composite complications for the ASA classification system (See Figure 8). One study reported c-statistics for the predictive ability of the ASA classification system for cardiac complications and pulmonary complications (Chrisant et al., 2024). One study reported c-statistics for the predictive ability of the ASA classification system for a surgical cohort and an endovascular cohort (Feghali et al., 2022), and one study reported c-statistics for patients receiving total hip arthroplasty, and total knee arthroplasty (McConaghy et al., 2021). **C-statistics ranged from 51.4%-77% with a median c-statistic of 60.8% suggesting the ASA had a poor predictive ability for composite complications across all nine studies.**

The majority of studies were conducted in populations receiving orthopaedic surgeries (n=3), two studies included mixed surgeries, two studies included vascular surgeries, one study included general surgeries, and one study included urology surgeries. The findings by surgery type can be seen in Table 12.

Table 12. ASA classification system c-statistics by surgical specialty

Surgical specialty	No of studies	c-statistic range	Median c-statistic	Discriminative ability
Orthopaedic surgery	3	57.1% - 71.9%	60.8%	Poor
Mixed	2	69% - 76%	75%	Fair
Vascular surgery	2	51.8% - 60.6%	54.1%	Very poor
Urology surgery	1	N/A	51.4%	Very poor
General	1	N/A	68.8%	Poor



- The bold line represents the cut off for a fair discriminative ability, anything above this point is considered fair or better. [1]Cardiac complications, [2]Pulmonary complications, [3]Patients who underwent clip placement, [4]Patients who underwent endovascular therapy, [5]Total Hip Arthroplasty, [6]Total Knee Arthroplasty.

Figure 8. ASA classification system C-Statistics plot

3.5.2 ASA classification system Calibration findings

None of the 13 studies assessing the ASA classification system reported O/E ratios.

3.5.3 ASA classification system Accuracy findings

None of the 13 studies assessing the ASA classification system reported a Brier score.

3.5.4. ASA classification system Healthcare utilisation and recovery outcomes

A total of five of the 13 studies assessing the predictive ability of the ASA classification system reported on four healthcare utilisation and recovery outcomes. This included readmission, reoperation, length of stay, extended length of stay, and discharge not to home. The studies reporting these outcomes only included orthopaedic surgeries (n=4) or urology surgeries (n=1). Only c-statistics were reported and can be seen in Appendix 3. It is important to note the definition of extended length of stay varied slightly across the studies. Extended length of

stay was defined as a hospital stay greater than the 75th percentile, or greater than 3 days (Fu et al., 2018), a hospital stay greater than one day (McConaghy et al., 2021), a hospital stay greater than the 75th percentile (Meng et al., 2018) or a hospital stay greater than or equal to the 75th percentile (Ondeck et al., 2018).

While only a limited number of studies reported each outcome, the evidence suggests that overall, the ASA classification system has a very poor discriminative ability for length of stay (c-statistics of 53.6% from 1 study), and extended length of stay (ranging from 56.1%-63%, with a median c-statistic of 57.2% across three studies). A poor discriminative ability for predicting discharge not to home (ranging from 59%-64%, with a median c-statistic of 63.1% across 4 studies), and an excellent discriminative ability for predicting readmission (90.6% from 1 study) and reoperation (86.6% from one study). However, as each outcome was only reported by a limited number of studies further evidence is needed. Overall, the evidence for the predictive ability of the ASA classification system to assess healthcare utilisation and recovery outcomes was very mixed.

3.5.5. Bottom line summary for ASA classification system.

When looking at composite complications, there is evidence to suggest the ASA classification system had a poor discriminative ability across nine studies (median c-statistic of 60.8%). When looking at the discriminative ability of the ASA classification system by surgical specialty the c-stats varied, ranging from very poor after vascular (n=2) or urology surgery (n=1) to fair after mixed surgeries (n=2). However, the findings for the individual surgical specialties are limited and showed wide ranges across studies, reducing the confidence in the findings. None of the studies assessing the ASA classification system reported calibration (O/E ratio) or accuracy (Brier) scores.

The findings for the ASA classification system's ability to predict healthcare utilisation and recovery outcomes was also very limited, suggesting further evidence is needed. However, overall, there is evidence to suggest the ASA classification system had a very poor discriminative ability for length of stay, or extended length of stay, a poor discriminative ability for predicting discharge not to home, and an excellent discriminative ability to predict readmission or reoperation.

The evidence directly comparing the predictive ability of the ASA classification system compared to the RCRI was limited and the findings appear to be mixed suggesting further evidence is needed.

4. DISCUSSION

4.1 The evidence base

This review set out to identify and map the evidence for externally validated pre-operative surgical risk prediction tools currently used in Wales for identifying patients suitable for surgery in low-risk regional surgical settings, such as surgical hubs. Accurate risk prediction is particularly valuable in these settings as they help clinicians to identify which patients can safely benefit from treatment while maintaining the efficiency and safety standards required for such facilities. This work builds on the NICE (2020) evidence review exploring which of the P-POSSUM, SORT, or ACS NSQIP risk stratification tools could best identify the risk of mortality and morbidity of adults undergoing surgery by examining more risk stratification tools. However this review focused on post-surgical complications and healthcare utilisation and recovery outcomes, which are likely to be priority considerations in low-risk regional surgical settings, such as surgical hubs.

A total of 118 studies were included in the rapid evidence map across 12 risk prediction tools. The most commonly studied risk prediction tool was ACS NSQIP (n=40), with the least common being the Carlisle Risk Calculator and the National Emergency Laparotomy Audit Parsimonious Risk Score (NELA PRS) where no relevant evidence was identified. Of the three risk prediction tools that were also included in the NICE evidence review (NICE, 2020), our evidence map identified a total of 51 studies meeting our eligibility criteria. Although not examined, it is likely some of these studies are included in both reviews.

The tools have been developed and used for a wide variety of surgical types, with some not specifically developed for surgery (NRS-2002 and CFS). All but one tool was designed to be carried out by clinicians. The exception to this is the DASI tool, which is a self-administered questionnaire, completed by patients. The SORT risk prediction tool and the ACS NSQIP tool is designed to be used in conjunction with the ASA (or a modified version) classification system, as it is included as part of the risk calculation. The ASA classification risk prediction tool should be considered alongside evaluations to determine appropriateness of surgery and should not be used alone. The ARISCAT score, CPET, POSSUM, P-POSSUM and RCRI tools utilise clinical findings in their assessment of pre-surgical risk while the others assess risk based on lifestyle questions or exercise tests. A further challenge relates to differences in assessed outcomes. Some tools predict specific complications (e.g., postoperative nausea and vomiting), while others estimate overall morbidity or surgical fitness (e.g., ASA classification system, CPET). Varying risk definitions further complicate direct comparisons. While a large evidence base exists for these surgical risk prediction tools, given the variability in terms of surgery and outcomes assessed, it is unlikely that any one tool would be suitable for use across all populations and surgery types within a surgical hub setting.

4.2 Findings for ACS NSQIP, P-POSSUM, RCRI and the ASA classification system

This review also set out to provide a more in-depth summary of the findings for a selection of the tools deemed to be the most applicable on a population level. When looking at the findings for ACS NSQIP, P-POSSUM, the RCRI and the ASA classification system in more detail, the

performance results appeared to be mixed. There was considerable heterogeneity amongst the included studies which limits direct comparisons and reduces the confidence in the overall findings. The evidence available for each tool shows a variation in how well the tool performed, while ACS NSQIP, P-POSSUM and the ASA classification system were all found to have an overall poor predictive ability across all studies, the RCRI was found to have a fair predictive ability across studies. This could suggest that the RCRI was the most effective at predicting complications across a range of surgical disciplines, however, the tools had been assessed in different surgical specialties, with ACS NSQIP being assessed across the largest number of surgical specialities (n=9), followed by the ASA classification system (n=5), the RCRI (n=4) and P-POSSUM (n=3) which could skew the findings if the tools are found to be more effective for predicting morbidity in certain surgical specialties than others. The evidence supports this as the findings for each tool show a considerable range in predictive ability depending on surgical specialty, which suggests the overall scores may not adequately reflect the predictive ability of each tool. P-POSSUM and the ASA classification system were found to range from very poor to fair when the findings were split by surgical specialty, the RCRI was found to range from poor to fair depending on the surgical specialty and ACS NSQIP was the only risk prediction tool that was found to range from very poor (orthopaedic, urology and vascular surgery) to excellent (mixed surgery) depending on the surgical specialty. While the evidence suggests that certain tools may be more effective for specific surgical specialties or even specific outcomes, the evidence base available for each surgical specialty varied between tools and was often very limited. As such, further evidence would be needed to draw firm conclusions, and the results should be interpreted with caution.

4.3 Limitations of the available evidence

The evidence highlights a large variation in what the risk prediction tools are predicting within the literature which makes synthesis of included studies very challenging. Some studies provided composite outcomes, incorporating multiple single outcomes, and some studies reported findings for individual outcomes. The studies reporting composite outcomes provide a larger range of outcomes which may make the findings more generalisable. However, the findings may be skewed if a risk prediction tool is found to be extremely good or extremely bad at predicting one of the outcomes included in the composite. For example, ‘cardiac complications may include both ‘cardiac arrest requiring CPR’ and ‘myocardial infarction’; the risk prediction tool’s performance may vary between the two individual outcomes which make up the composite. Some composite complications were also less frequently reported in the literature, such as ‘pulmonary complications’. The studies that reported individual outcomes highlight which specific outcomes the risk prediction tools may be better suited at identifying. Thus, the limitation is that the evidence base may be limited by a smaller number of studies reporting findings for each outcome, which will reduce the overall certainty of the findings.

A limitation of the evidence is the variance in the statistical reporting of outcome results. Most studies reported the discriminative ability of the risk prediction tool using the ‘area under the curve’ (AUC) and a combination of discrimination and calibration using the ‘Brier score’. However, some studies only reported calibration findings, and a smaller number of studies reported only the Brier score. For this reason, comparison of predictive ability across tools was challenging. In addition, although the Brier score is considered the most comprehensive score of accuracy, Brier scores should not be directly compared across studies using different patient data, as the Brier score depends on both the prevalence of the event in the data and the performance of the tool. Thus, overall comparison of a tool’s Brier score was not possible, although Brier scores for individual studies have been reported.

Some tools were well represented in the literature by a larger number of studies, such as ACS NSQIP (n=40), whereas other tools such as P-POSSUM had a smaller number of studies (n=7). Broader surgery types, such as 'general' (n=43) and 'mixed surgeries' (n=23) were more frequently represented in the literature than other surgery types, again reducing the overall certainty of the findings for the tools with fewer studies and in a less common surgical category. Most studies were conducted in the USA which may limit the generalisability of the findings to the Welsh context.

4.4 Summary of the Evidence gaps

While overall a large evidence base was identified, a number of risk prediction tools had very minimal evidence available with the NRS-2002, ARISCAT, CFS, SORT and DASI all being assessed in five studies or less. No evidence was identified assessing the predictive ability of the Carlisle Risk Calculator or the NELA PRS. In addition, no evidence was identified looking at the use of risk prediction tools when identifying patients suitable for treatment in surgical hubs.

4.5 Strengths and limitations of this Evidence Review

The studies included in this review were identified through an extensive search of electronic databases, trial registries, grey literature sites and through citation tracking. As the search identified a very large number of studies, the review can be considered a good reflection of the overall evidence base for this topic. However, there is a possibility that additional publications may have been missed or we may have introduced some biases to this review. As no date or country limits were set, it is possible that some of the evidence gained could be outdated (for example if the tool had been updated overtime or if surgical practices have changed, making them safer), and as a wide range of countries were included it is unclear how generalisable the findings would be to the UK context. No quality appraisal of included studies was conducted and therefore we cannot report the quality of the included studies.

Although it has been impossible to directly compare some accuracy measures (such as the Brier score) across the same risk prediction tool, the in-depth summary does compare the discriminative ability and the calibration of the four tools of interest using the AUC and the O/E ratios (where reported), producing median values which give an indication of the overall strength of the tool's performance. Given the complexity of the evidence base, the map is organised into aspects which would be most useful to clinicians looking to choose a tool for use within a low-risk surgical setting, looking at outcomes of composite complications and overall morbidity.

While the review is focussed on evidence relating solely to morbidity outcomes, we have also included separately evidence in which mortality is included within a composite morbidity outcome, to recognise the frequency in which mortality is included in these composite outcomes (Appendix 1).

In addition, the review excluded studies assessing the external validation of modification of the 13 risk predictions tools of interest. It is common for risk prediction tools to undergo modification in order to make them better at predicting risk in specific situations or for specific groups. However, those identified during the screening process have been collated and referenced in Appendix 2 in order to better reflect the totality of the evidence base.

4.6 Implications and next steps

This review provides a summary of 12 risk prediction tools currently used in surgical disciplines across Wales. It can be used to inform practice in low-risk surgical settings in Wales to help clinicians, practitioners and other stakeholders in deciding which tools are most appropriate to use for different surgery types by giving an indication of the predictive ability of four discrete tools: ACS NSQIP, ASA, RCRI and P-POSSUM for composite and morbidity outcomes. However, several limitations were identified, such as inconsistent reporting methods and heterogeneity across the studies, and the variation in the amount of evidence available for each tool. In addition, no quality appraisal of the included studies was conducted and as such the findings should be interpreted with caution. While it is clear that no risk prediction tool adequately predicted complications across all surgical specialties, it may be likely that some tools are better suited for specific surgery types or that a combination of risk prediction tools may be needed to adequately assess an individual's level of risk.

This review has identified evidence gaps as no external validation studies were retrieved for the Carlisle Risk Calculator and the National Emergency Laparotomy Audit Parsimonious Risk Score (NELA PRS), suggesting further research is needed for these tools. As considerable limitations were identified limiting the comparability between studies, further research should ensure risk prediction tools are assessed in a consistent way to allow for direct comparisons.

4.7 Economic considerations*

- Future research into risk prediction tools should incorporate health economic evaluations, to provide consideration of individual risk as well as associated health and social care resource use costs. NICE state that risk prediction tools are freely available, and therefore do not have an associated cost to use. However, they do require some time to complete, but this is typically less than 5 minutes during a preoperative assessment NICE (2020).
- Economic evaluation/impact evaluations of risk prediction tools are a known evidence gap, as described by a systematic review of Health Economic Impact Evaluations of Risk Prediction Models (van Giessen et al., 2017). This review considered any study where a clinical risk prediction model was evaluated by health economic evaluation. Further, an evidence review of preoperative risk stratification tools conducted by NICE also identified no relevant economic evaluations NICE (2020).
- Model-based health economic evaluations may be an appropriate method of conducting health economic analysis of risk prediction tools as it can account for long-term health and cost outcomes. Further, it can be applied across cohorts of different surgical specialties (van Giessen et al., 2017).

**This section has been completed by the Centre for Health Economics & Medicines Evaluation (CHEME), Bangor University*

5. REFERENCES

American College of Surgeons National Surgical Quality Improvement Program (2025). ACS NSQIP Surgical Risk Calculator. Available at: https://riskcalculator.facs.org/RiskCalculator/

American Society of Anesthesiologists (2020) Statement on ASA Physical Status Classification System. Available at: https://www.asahq.org/standards-and-practice-parameters/statement-on-asa-physical-status-classification-system

Apfel, C. C., Läärä, E., Koivuranta, M., et al. (1999). A simplified risk score for predicting postoperative nausea and vomiting: conclusions from cross-validations between two centers. Anesthesiology, 91(3), 693–700. https://doi.org/10.1097/00000542-199909000-00022

Canet, J., Gallart, L., Gomar, C., et al. (2010). Prediction of postoperative pulmonary complications in a population-based surgical cohort. Anesthesiology, 113(6), 1338–1350. Available at: https://doi.org/10.1097/ALN.0b013e3181fc6e0a

Collins, G. S., de Groot, J. A., Dutton, S., et al. (2014). External validation of multivariable prediction models: a systematic review of methodological conduct and reporting. BMC medical research methodology, 14, 1-11. Available at: https://link.springer.com/article/10.1186/1471-2288-14-40

Copeland, G. P., Jones, D., & Walters, M. (1991). POSSUM: a scoring system for surgical audit. The British journal of surgery, 78(3), 355–360. https://doi.org/10.1002/bjs.1800780327

Çorbacıoğlu, Ş. K., & Aksel, G. (2023). Receiver operating characteristic curve analysis in diagnostic accuracy studies: A guide to interpreting the area under the curve value. Turkish Journal of Emergency Medicine, 23(4),195. Available at: https://pmc.ncbi.nlm.nih.gov/articles/PMC10664195/

Dash, K., Goodacre, S., & Sutton, L. (2022). Composite outcomes in clinical prediction modeling: are we trying to predict apples and oranges?. Annals of Emergency Medicine, 80(1), 12-19. Available at: https://www.sciencedirect.com/science/article/pii/S0196064422001020

Hageman, S. H., Petitjean, C., Pennells, L., et al. (2023). Improving 10-year cardiovascular risk prediction in apparently healthy people: flexible addition of risk modifiers on top of SCORE2. European Journal of Preventive Cardiology, 30(15), 1705-1714. Available at: https://academic.oup.com/eurjpc/article/30/15/1705/7188647

Hlatky, M. A., Boineau, R. E., Higginbotham, M. B., et al. (1989). A brief self-administered questionnaire to determine functional capacity (the Duke Activity Status Index). The American journal of cardiology, 64(10), pp.651-654. Available at: https://pubmed.ncbi.nlm.nih.gov/2782256/

Huang, C., Li, S. X., Caraballo, C., et al. (2021). Performance metrics for the comparative analysis of clinical risk prediction models employing machine learning. Circulation: Cardiovascular Quality and Outcomes, 14(10), e007526. Available at: https://pubmed.ncbi.nlm.nih.gov/34601947/

Kondrup, J., Rasmussen, H. H., Hamberg, O., et al. (2003). Nutritional risk screening (NRS 2002): a new method based on an analysis of controlled clinical trials. Clinical nutrition (Edinburgh, Scotland), 22(3), 321–336. Available at: https://doi.org/10.1016/s0261-5614(02)00214-5

Lee, T. H., Marcantonio, E. R., Mangione, C. M., et al. (1999). Derivation and prospective validation of a simple index for prediction of cardiac risk of major noncardiac surgery. Circulation, 100(10), 1043–1049. Available at: https://www.ahajournals.org/doi/10.1161/01.CIR.100.10.1043

Levett, D. Z. H., Jack, S., Swart, M., et al. (2018). Perioperative cardiopulmonary exercise testing (CPET): consensus clinical guidelines on indications, organization, conduct, and physiological interpretation. British journal of anaesthesia, 120(3), 484–500. Available at: https://doi.org/10.1016/j.bja.2017.10.020

MDCALC (2024). All Calculators. Available at: https://www.mdcalc.com/

NICE (2020) Perioperative care in adults [C] Evidence review for preoperative risk stratification tools. Available at: https://www.nice.org.uk/guidance/ng180/evidence/c-preoperative-risk-stratification-tools-pdf-8833151056

Pradhan, N., Dyas, A. R., Bronsert, M. R., et al. (2022). Attitudes about use of preoperative risk assessment tools: a survey of surgeons and surgical residents in an academic health system. Patient Safety in Surgery, 16(1), 13. Available at: https://link.springer.com/content/pdf/10.1186/s13037-022-00320-1.pdf

Protopapa, K. L., Simpson, J. C., Smith, N. C. E., et al. (2014). Development and validation of the surgical outcome risk tool (SORT). Journal of British Surgery, 101(13), 1774-1783. Available at: https://academic.oup.com/bjs/article/101/13/1774/6137960?login=true

Prytherch, D. R., Whiteley, M. S., Higgins, B., et al. (1998). POSSUM and Portsmouth POSSUM for predicting mortality. Physiological and Operative Severity Score for the enUmeration of Mortality and morbidity. *The British journal of surgery*, *85*(9), 1217–1220. Available at: https://doi.org/10.1046/j.1365-2168.1998.00840.x

Rockwood, K., & Theou, O. (2020). Using the Clinical Frailty Scale in Allocating Scarce Health Care Resources. Canadian geriatrics journal: CGJ, 23(3), 210–215. Available at: https://doi.org/10.5770/cgj.23.463

Royal College of Surgeons of England. (2024). Surgical specialties. Available at: https://www.rcseng.ac.uk/news-and-events/media-centre/media-background-briefings-and-statistics/

Saklad, M. (1941, May). Grading of patients for surgical procedures. In The Journal of the American Society of Anesthesiologists (Vol. 2, No. 3, pp. 281-284). The American Society of Anesthesiologists. Available at: https://journals.lww.com/anesthesiology/citation/1941/05000/grading_of_patients_for_surgical_procedures.4.aspx

van Giessen, A., Peters, J., Wilcher, B., Hyde, C., Moons, C., de Wit, A., & Koffijberg, E. (2017). Systematic review of health economic impact evaluations of risk prediction models: stop developing, start evaluating. Value in health, 20(4), 718-726. Available at: https://www.sciencedirect.com/science/article/pii/S1098301516300328

6. EVIDENCE REVIEW METHODS

6.1 Eligibility criteria

Table 13. Eligibility criteria

	Inclusion criteria	Exclusion criteria
Population	Adults scheduled for surgery	Children scheduled for surgery Emergency surgery
Index	Risk prediction tools commonly used in Wales: • ARISCAT score (for Postoperative Pulmonary Complications) • ASA Physical Status/ ASA Classification • Carlisle Risk/Carlisle Calculator • CFS (Clinical Frailty Scale, also known as Rockwood) • CPET (Cardiopulmonary exercise testing) • DASI (Duke Activity Status Index) • NELA PRS (National Emergency Laparotomy Audit (Parsimonious Risk Score) • NRS-2002 (Nutrition Risk Screening 2002) • ACS NSQIP (National Surgical Quality Improvement Program) universal surgical risk calculator) • P-POSSUM score (Physiological and Operative Severity Score for the enUmeration of Mortality and Morbidity) • PONV (Apfel Score for Postoperative Nausea and Vomiting) • RCRI (Revised Cardiac Risk Index for Pre-Operative Risk) • SORT (Surgical Outcome Risk Tool)	Other risk prediction tools
Comparator	No predefined comparator	
Outcome(s)	• Morbidity • Complications	
Study type	Primary studies including prospective or retrospective data that conduct external validation evaluation	Model development and internal validation studies
Timing	Prognostic tools used to determine intra-operative and post-operative risk	
Setting	To determine an individuals' surgical risk	
Other Study Considerations		

6.2 Literature search

All searches were conducted between 9th Dec -17th Dec 2024. The search strategies used for MEDLINE can be seen in Appendix 4. A search of the following databases and resources was conducted to identify published and ongoing primary studies:

- Medline
- Embase
- CINAHL
- Cochrane Central Register of Controlled Trials (CENTRAL)
- Clinicaltrials.gov
- WHO International Trials Registry Platform
- Scopus
- Google Scholar

Several websites and specialist sources were also searched, including:

- Royal College of Anaesthetists
- Royal College of Surgeons
- The Royal College of Surgeons of England
- NHS Scotland
- NHS England
- NIHR Public Health Research
- NICE
- The Health Foundation
- Centre for Peri-operative Care
- Association of Anaesthetists
- American Society of Anesthesiologists (ASA)
- GIRFT- Getting It Right First Time Home - Getting It Right First Time - GIRFT
- EPPI centre

6.3 Study selection process

The searches yielded a total of 10,089 records. Records were imported into an Endnote database library and duplicates were removed. After deduplication, a total of 5,106 records were screened at title and abstract. Title and abstracts were screened by one reviewer in Rayyan, with around 10% being assessed by another reviewer to ensure consistency and minimise bias. Any uncertainty was discussed within the review team. The full texts were screened by two reviewers in duplicate, if disagreements arose, these were discussed, and a third reviewer was consulted to make a final inclusion decision. After full text screening a total of 118 records met the inclusion criteria for the REM.

6.4 Data extraction and coding/charting

Data extraction was conducted by one reviewer and consistency checked by another reviewer while creating the evidence map and summaries for each risk prediction tool. Extracted information included the study's country, study aim, design, population, surgery type, surgery category, sample size, outcomes of interest, and reported outcomes.

The evidence map employed a structured coding approach. Studies were first grouped by the surgical risk prediction tool assessed and then by surgical specialty, following Royal College of Surgeons categorisations. Studies using data from a combination of different surgical specialties were categorised as 'Mixed' and where categorisation was not possible or unclear, reviewers used information from the study to code the surgical specialty. Outcomes were coded into broad categories, including complications and healthcare utilisation and recovery. Complications encompassed any reported complications, while healthcare utilisation included

measures such as length of stay, readmission, and discharge to higher-level care. Where specified, complications were further classified into subcategories such as cardiac, non-cardiac, and pulmonary complications to accurately represent a tools focus. Once all studies had been grouped and assigned into categories, they were charted onto the evidence map accordingly.

6.5 Assessment of methodological quality

None of the studies included in this review were assessed for methodological quality.

7. EVIDENCE

7.1 Search results and study selection

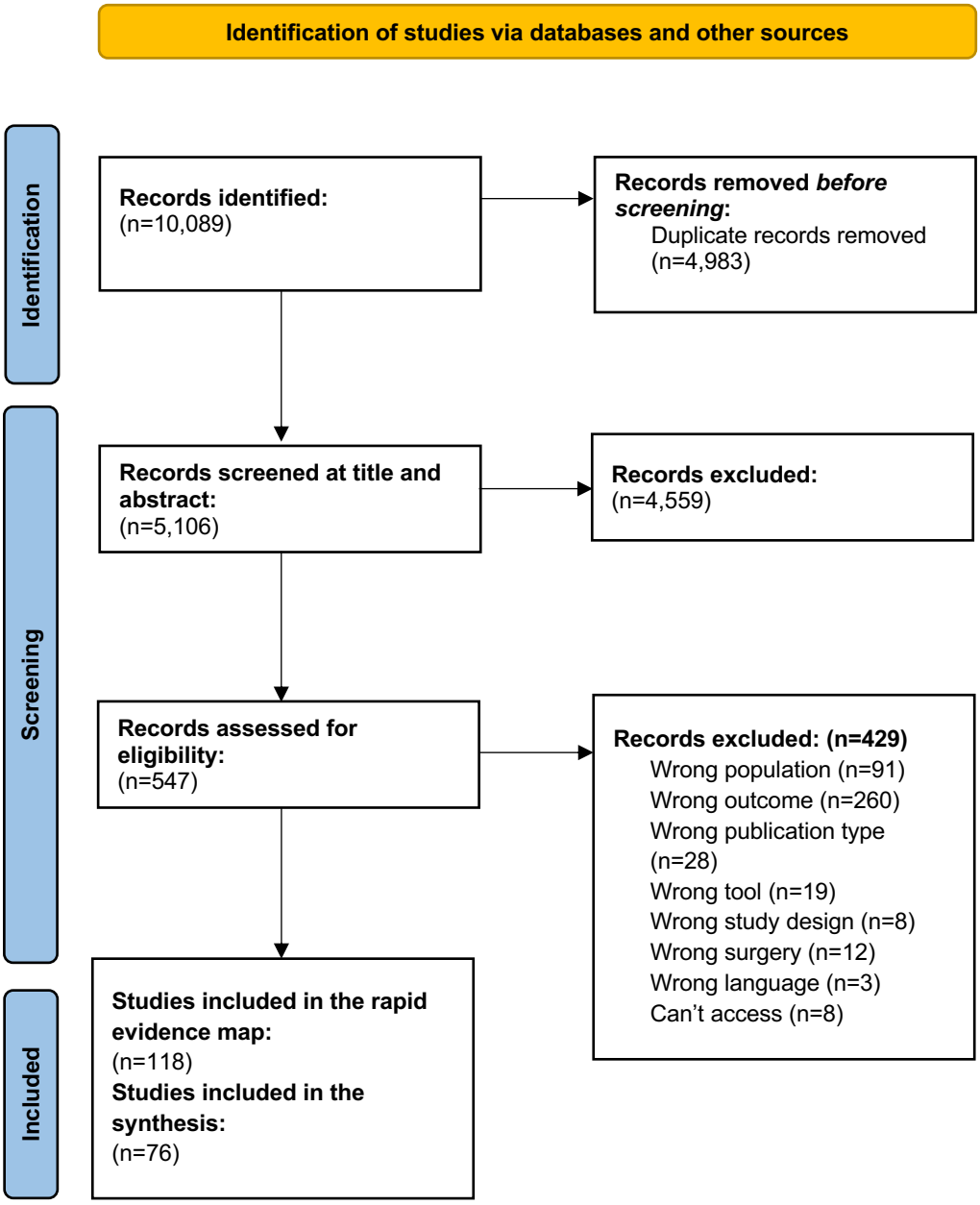


Figure 9. PRISMA Flow Diagram

7.2 Data extraction Tables

Table 14. Data Extraction Tables: ACS NSQIP

REF Country Sample size	Aim and design	Population, surgery and (surgical specialty category)	Validation measures	Outcomes	C-statistic	O/E ratio	Brier score	Comments
Alzahrani et al (2020) Republic of Korea 207	To examine the ability of the ACS NSQIP risk calculator to estimate short-term surgical outcomes among Korean patients with gastric cancer, treated with a minimally invasive approach. Retrospective	Patients undergoing Laparoscopic gastrectomy (General surgery)	AUC Brier score	Serious complications Any complications Pneumonia Surgical site infection Readmission Return to operating room	0.600 0.570 0.650 0.600 0.570 0.610	-	0.025 0.015 0.103 0.089 0.032 0.069	The C-statistics (AUC) were poor (<0.7) for all the outcomes. The Brier scores were contradictory to the AUC results. They showed a considerably impressive values by approaching 0.0 for all of the outcomes (Brier score range, 0.01–0.06), except for any complication (Brier score, 0.15). The best score result was demonstrated for pneumonia (0.01) and all results were below the cut-off, which indicates its validity
Angelucci et al (2024) Italy 567	To investigate the accuracy of ACS-NSQIP and P-POSSUM risk calculators in predicting postoperative outcomes for patients undergoing retroperitoneal sarcoma surgery. Retrospective	Adult patients (≥18), undergoing elective comprehensive resection for primary or persistent adult-type retroperitoneal sarcoma (RPS) with ASA score I–IV (General surgery	AUC, Brier scores	Any complication Severe complication Readmission Reoperation Sepsis	0.640 (0.60–0.69) 0.610 (0.56–0.66) 0.510 (0.39–0.63) 0.590 (0.52-0.66) 0.560 (0.46-0.65)	-	0.231 0.206 0.049 0.104 0.066	Accuracy was assessed by Brier Score and area under the curve (AUC). Severe complications occurred after 30th postoperative day in 3.5% cases. ACS-NSQIP predicted below-average complication for 65.1%, average for 16.9%, and above-average for 18% of patients. The accuracy of both

REF Country Sample size	Aim and design	Population, surgery and (surgical specialty category)	Validation measures	Outcomes	C-statistic	O/E ratio	Brier score	Comments
								ACS-NSQIP and P-POSSUM in predicting postoperative outcomes for patients undergoing retroperitoneal sarcoma surgery was found to be low, with an AUC < 0.7.
Blair et al (2018) USA 470	To evaluate the accuracy of the American College of Surgeons NSQIP Surgical Risk Calculator for predicting risk-adjusted 30-day outcomes for patients undergoing partial nephrectomy for renal cell carcinoma. Retrospective	Patients with tumours undergoing Partial Nephrectomy (PN). (Urology surgery)	O:E	Severe complications Overall complications Cardiac event Pneumonia Surgical site infection Urinary tract infection Venous thromboembolism Acute renal failure Return to OR Discharge to rehab Length of stay (days)	-	0.791 1.833 1.349 1.174 1.679 2.411 1.063 0.895 1.324 0.990 1.148	-	The NSQIP calculator consistently provided discordant results from that which was observed. The twofold underestimation of overall complications for open PN (11.94 vs. 23.44%, p < 0.001) and MIPN (6.93 vs. 11.49%, p < 0.001) groups were most notable. In contrast, the calculator again overestimated serious complications as seen in the MIPN cohort (5.01 vs. 2.68%, p < 0.001). Significant under- and over-estimations persisted, with clinically significant differences in overall complications for pure laparoscopic (6.88 vs. 15.32%, p < 0.001) and robotic-assisted (6.97 vs. 8.67%, p < 0.001) PN and a nearly threefold overestimation of serious complications in the laparoscopic

REF Country Sample size	Aim and design	Population, surgery and (surgical specialty category)	Validation measures	Outcomes	C-statistic	O/E ratio	Brier score	Comments
								cohort (5.05 vs. 1.80%, p < 0.001).
Botejue et al (2023) Sri Lanka 126	The aim of the study was to correlate the predicted outcomes of the calculator with the actual need for organ support and identify a reliable cut-off point to determine which patients require post-operative organ support. Prospective	Patients undergoing elective major general surgery (general surgery)	AUC Youden Index method and validated the results using the cut-off scores	Serious complications	0.710	-	-	The percentage risk of death did not show a significant correlation with the need for organ support (p = 0.13). This suggests that the serious risk has the ability to distinguish between patients who require organ support and those who do not. The mean predicted percentage risk of serious complications, for the group that did not require any organ support was 10.5% and the group requiring 1 or more organ support was 18.1%. The standard error was 0.49 (p= 0.001). A receiver-operating characteristic (ROC) curve gave an area under the curve of 0.71. The percentage risk of serious complications calculated by the ACS-NSQIP surgical risk calculator has a strong positive correlation with the need for postoperative organ support.

REF Country Sample size	Aim and design	Population, surgery and (surgical specialty category)	Validation measures	Outcomes	C-statistic	O/E ratio	Brier score	Comments
Boyd et al (2020) USA 731	To determine the correlation between the predicted risk for perioperative complications, readmission, reoperation, length of stay, and discharge to skilled nursing or rehabilitation facility using the ACS NSQIP surgical risk calculator with actual outcomes in patients undergoing pelvic reconstructive and incontinence surgery. Retrospective	All women 18 years or older undergoing surgery for Pelvic organ prolapse (POP) and/or incontinence by all routes were included. Vaginal, Laparoscopic, Abdominal, Obliterative and other (urethral diverticulum excision, fistula repair, or vaginal cyst removal). (Mixed surgery)	C-Statistic (AUC)	Any serious complication Any complication Pneumonia Cardiac Superficial skin infection Urinary tract infection Venous thromboembolism Renal failure Readmission Return to OR Discharge to SNF or rehabilitation	0.529 0.547 0.629 0.650 0.732 0.556 0.683 NG 0.512 0.519 0.848	-	-	C-statistics for the risk calculator were poor for all categories with the exception of superficial skin infection (c-statistic = 0.732, P = 0.001) and post–acute care discharge (c-statistic = 0.848, P = 0.007). These findings did not correlate to results of the BS, which showed the calculator best predicted renal failure (BS = 0.002), cardiac events (BS = 0.037), pneumonia (BS = 0.037), and death (BS = 0.048). Additionally, the categories with the lowest BS also had the lowest number of total events for the cohort The ACS NSQIP surgical risk calculator is an overall poor predictor of actual outcomes in a sample of patients who underwent pelvic reconstructive surgery, perhaps because of low prevalence of serious events
Campagnaro et al (2023) Italy	To evaluate the performance of the American College of Surgeons National Surgical Quality	Patients undergoing liver resection for CRLM or simultaneous liver-colon resection for	AUC Brier score	Liver surgeon adjustment score (SAS) SAS-1: Overall complications Severe complications	 0.619 0.572	-	 0.380 0.100	Two types of surgery in study: 1. liver +simultaneous colon and 2. liver surgery, both for colorectal

REF Country Sample size	Aim and design	Population, surgery and (surgical specialty category)	Validation measures	Outcomes	C-statistic	O/E ratio	Brier score	Comments
410	Improvement Program surgical risk calculator (ACS-NSQIP SRC) for patients undergoing liver resection for colorectal liver metastases or simultaneous liver and colic resection for metastatic colorectal cancer. Prospective	metastatic CRC in two high-volume Italian institutions (General surgery)		Cardiac complications Pneumonia Liver SAS-2: Overall complications Severe complications Cardiac complications Pneumonia Liver SAS-3: Overall complications Severe complications Cardiac complications Pneumonia Colon: Overall complications Severe complications Cardiac complications Pneumonia Liver only resections: Overall complications Severe complications Cardiac complications Pneumonia	0.720 0.677 0.613 0.606 0.740 0.654 0.620 0.673 0.702 0.640 0.607 0.688 0.714 0.671 0.580 0.556 0.667 0.543		0.040 0.170 0.340 0.110 0.040 0.170 0.320 0.110 0.040 0.170 0.340 0.100 0.040 0.170 0.300 0.100 0.030 0.150	cancer (CRC). An AUC ≥ 0.7 shows acceptable discrimination; a Brier score next to 0 means the prediction tool has good calibration. ACS-NSQIP SRC showed good predicting capabilities only for 1 out of 5 evaluated outcomes; therefore, it is not a reliable tool for patients undergoing liver surgery for colorectal liver metastases (CRLM). The NSQIP tool underestimated the incidence of overall complications, pneumonia, cardiac complications, and the length of hospital stay.
Choi et al (2019) Republic of Korea 199	To validate the ACS NSQIP surgical risk calculator in its ability to predict surgical complication and to test the potential feasibility of the ACS NSQIP surgical risk calculator to predict long-term oncologic outcomes of patients with resected	Patients underwent Pancreaticoduodenectomy (PD) or pylorus-preserving pancreaticoduodenectomy (PPPD) for pancreatic head cancer. (General surgery)	Brier Score	Serious complication Any complication Pneumonia Cardiac complication Superficial skin infection Urinary tract infection Venous thromboembolism Renal failure	-

REF Country Sample size	Aim and design	Population, surgery and (surgical specialty category)	Validation measures	Outcomes	C-statistic	O/E ratio	Brier score	Comments
	pancreatic head cancer by their calculated serious complication rate. Retrospective			Return to OR Discharge to NH or rehabilitation			0.000 0.005	
Cohn and Ros (2018) USA 663	This study compared the Revised Cardiac Risk Index (RCRI) and National Surgical Quality Improvement Program (NSQIP) risk calculators and a reconstructed Revised Cardiac Risk Index in predicting postoperative cardiac complications, both during hospitalisation and 30 days after operation, in a patient cohort who underwent select surgical procedures in various risk categories. Retrospective	A patient cohort who underwent select surgical procedures in various risk categories. Non-cardiac surgery (15.7% high risk; 62.9% Intermediate risk; 21.4% low risk). (Mixed surgery)	AUC	All cardiac complications (in hospital) All cardiac complications (-30 days) Major cardiac complications	0.920 0.890 0.770	-	-	The mean scores ± SD (predicted % chance of having a cardiac event) for ACS-SRC was 0.33 ± 0.58. The ROC curves predicting MCC-30d were not discriminative for ACS-SRC. The performance of ACS-SRC for predicting major cardiac events within 30 days after surgery remained discriminative (moderate to excellent) regardless of the group.
Chudgar et al (2022) USA 2514	To externally evaluate the performance of the NSQIP SRC for patients undergoing pulmonary resection. Retrospective	Patients undergoing pulmonary resection. (Thoracic surgery)	C-index O:E ratio Calibration curves	Serious complication Any complication Pneumonia Cardiac complication Surgical site infection Urinary tract infection	0.734 (0.696-0.771) 0.728 (0.692-0.765) 0.715 (0.654-0.776) 0.821 (0.698-0.945) 0.741 (0.636-0.845) 0.703	0.804 0.838 0.935 0.571 0.833 0.667	-	The C-index was excellent for cardiac complication (0.821). The C-indices were good for serious complication (0.734) and any complication (0.728). Aside from renal failure, discharge to a nursing or rehabilitation facility,

REF Country Sample size	Aim and design	Population, surgery and (surgical specialty category)	Validation measures	Outcomes	C-statistic	O/E ratio	Brier score	Comments
					(0.605-0.801)			and sepsis, good calibration is also noted for the remaining outcomes.
				Venous thromboembolism	0.792 (0.672-0.912)	0.545		
				Renal failure	0.798 (0.738-0.859)	3.000		
				Readmission	0.644 (0.594-0.694)	0.761		
				Return to operating theatre	0.771 (0.710-0.832)	0.679		
				Discharge to nursing home or rehab	0.706 (0.613-0.798)	0.262		
				Sepsis	0.707 (0.572-0.841)	0.417		
Donadon et al (2020) Italy 950	To assess the ACS-NSQIP calculator's ability to predict complications, mortality and length of stay in patients undergoing hepatectomy for liver tumors. Retrospective	Patients undergoing Hepatectomy for liver tumours (General)	C-statistic Brier score	Complications	0.610 (0.51 – 0.63)			The performance of the calculator was tested by using c-statistic and Brier score. C-statistic and Brier scores showed low performance of the calculator.
				Serious complication		0.737	0.722	
				Any complication		2.060	0.688	
				Pneumonia		2.301	0.966	
				Cardiac event		1.483	0.011	
				Site infection		0.273	0.222	
				Urinary tract infection		0.473	0.013	
				Venous thromboembolism		0.336	0.967	
				Renal failure		0.866	0.010	
				Readmission		0.042	0.934	
				Reoperation		1.784	0.011	
				Length of stay (days)		1.878	0.431	
Fruscione et al (2018) USA 295	To assess the predictive capacity of the NSQIP Surgical Risk Calculator at Carolinas Medical Center, both general surgery and	Patients who underwent cholecystectomy between 2008 and 2016 and were deemed too high risk for acute care general	AUC	GS Cholecystectomy: Serious complication	0.720	1.290	-	NSQIP's predictive ability was compared with a new predictive model, the SRC at Carolina Medical Center (CMC). 169 patients with 'difficult
				30-day readmission	0.688	1.400		
				Cardiac complication	0.728	8.000		
				HPB Cholecystectomy: Serious complication	0.756	2.029		

REF Country Sample size	Aim and design	Population, surgery and (surgical specialty category)	Validation measures	Outcomes	C-statistic	O/E ratio	Brier score	Comments
	hepatopancreatobiliary cholecystectomy procedures over an 8-year period were utilized to construct an algorithm to more precisely predict outcomes for patients undergoing a cholecystectomy procedure. Retrospective	surgery (GS) and had surgery performed by the Division of Hepatopancreatobiliary Surgery (HPB). (General)		Discharge to nursing or rehab Facility Renal failure 30-day readmission Cardiac complication	0.741 0.836 0.570 0.777	1.442 4.000 1.824 11.833		cholecystectomy' under HPB were matched with 126 patients who underwent 'routine cholecystectomy' under GS. Although the SRC predictions closely approximated observed incidence of certain outcomes for GS cholecystectomies, outcomes for higher risk HPB cholecystectomies were significantly underestimated.
Golan et al (2017) USA 954	To evaluate the accuracy of the ACS-NSQIP surgical risk calculator in patients undergoing radical cystectomy with urinary diversion. Retrospective	Patients undergoing radical cystectomy(RC) with urinary diversion. Radical cystectomy with Ileal conduit or orthotropic neobladder. (Urology surgery)	Brier score AUC	Any complication Serious complication Pneumonia Cardiac Surgical site infection Acute urinary tract infection Venous thromboembolism Renal failure Readmission Return to OR Discharge to rehab	0.580 0.570 0.590 0.690 0.570 0.580 0.540 0.620 0.600 0.590 0.750	1.313 1.365 0.880 1.412 1.092 1.752 1.703 5.250 1.191 1.196 0.938	0.240 0.240 0.020 0.020 0.110 0.150 0.060 0.100 0.180 0.050 0.090	Analysis of the overall cohort revealed BS greater than 0.01 for all outcomes, indicating inadequate predictive accuracy. A high BS of 0.24 was observed for serious complications indicating poor performance. There were some exceptions in the ONB subgroup: cardiac complications BS was 0.008. Receiver operating characteristic curve analysis of the entire cohort revealed that nearly all models had a calculated AUC of 0.6 or less, suggesting poor accuracy in predicting

REF Country Sample size	Aim and design	Population, surgery and (surgical specialty category)	Validation measures	Outcomes	C-statistic	O/E ratio	Brier score	Comments
								most potential complications. The most accurately predicted outcomes with moderate discrimination of their models were cardiac (AUC =0.69).
Gray et al (2023) USA 240	To externally validate the ability of the NSQIP SRC and frailty index to predict postoperative complications in a large series of consecutive patients undergoing esophagectomy. Retrospective	Patients who underwent open or minimally invasive esophagectomy conducted using laparoscopy or thoracoscopy, with or without robotic assistance at a high-volume cancer center. All patients had a diagnosis of esophageal neoplasm. (General surgery)	C-index Calibration curves	Any complication Serious complication Pneumonia Cardiac complication Surgical site infection Urinary tract infection Venous thromboembolism Renal failure Readmission Return to operating room Discharge to a facility other than home Length of stay, median (IQR), days	0.553 (0.477-0.630) 0.554 (0.474-0.634) 0.658 (0.553-0.762) — 0.477 (0.349-0.605) 0.599 (0.392-0.805) 0.585 (0.392-0.779) 0.636 (0.536-0.737) 0.625 (0.527-0.723) 0.533 (0.433-0.633) 0.728 (0.555-0.902) -	2.793 2.658 2.541 0.455 1.667 4.737 2.250 20.000 3.130 3.485 0.769 0.917	-	Performance was evaluated using concordance index (C-index) and calibration curves. The SRC (NSQIP) did not identify risk of complications in the entire cohort (C-index, 0.553); patients undergoing open esophagectomy (C-index, 0.569); or patients undergoing minimally invasive esophagectomy (C-index, 0.542). Calibration curves showed general underestimation. Although the risk calculator correctly estimated a high aggregate risk of complications in the overall cohort (mean predicted probability of any complication was 33% versus an observed 39% complication rate; mean

REF Country Sample size	Aim and design	Population, surgery and (surgical specialty category)	Validation measures	Outcomes	C-statistic	O/E ratio	Brier score	Comments
								predicted probability of serious complication was 30% versus an observed 33% serious complication rate), there was poor concordance between predicted and observed complication rates for the majority of specific endpoints. Observed and predicted differences as well as calibration curves showed adequate correlation for only three of the twelve metrics measured, specifically: serious complications, pneumonia, and venous thromboembolism
Houdek et al (2020) Canada 65	To evaluate the accuracy of the ACS-NSQIP risk calculator for patients undergoing an en-bloc sacrectomy for chordoma, a procedure which is known for a high complication rate. Retrospective	Patients who underwent sacrectomy Excision of presacral/sacral tumor), laminectomy of sacral vertebrae, laminectomy for biopsy/excision of sacral neoplasm and sacral vertebral corpectomy for intraspinal lesion (Neurosurgery)	AUC Brier score	Complications (49215 excision of presacral or sacrococcygeal tumor) Complications (63011 (laminectomy with exploration and/or decompression of spinal cord and/or cauda equina, without facetectomy, foraminotomy or discectomy (e.g., spinal stenosis), 1 or 2 vertebral segments; sacral)	0.653 0.660	-	0.484 0.594 0.539	The ACS-NSQIP calculator was a poor predictor of complications and was marginally better than a coin flip in its ability to predict complications following sacrectomy for chordoma.

REF Country Sample size	Aim and design	Population, surgery and (surgical specialty category)	Validation measures	Outcomes	C-statistic	O/E ratio	Brier score	Comments
				Complications (63728 (laminectomy for biopsy/excision of intraspinal neoplasm; extradural, sacral) Complications (63307 (transperitoneal or retroperitoneal vertebral corpectomy, intradural, lumbar or sacral, for excision of an intraspinal lesion of one vertebral segment)>	0.636 0.670		0.486	
Hsiao et al (2022) USA (Data from the NSQIP thyroidectomy module) 18,078	To compare the performance of the ACS-NSQIP SRC to other classical machine learning algorithms trained on NSQIP data, and to demonstrate challenges and strategies in predicting such rare events. Retrospective	Patients undergoing thyroidectomy. (General)	AUC	Systemic complications	0.716 (0.660–0.767)	-	-	In the thyroidectomy patient population, the AUC of the ACS-NSQIP morbidity estimate appeared somewhat lower than the figures produced by validation in different cohorts. Using the SRC as a classifier where intervention occurs above a certain validated threshold, rather than citing the numeric estimates of complication risk, should be considered in low-risk patients.
Im et al (2024) USA	To examine the ability of the American College of Surgeons National Surgical Quality	Patients undergoing spinal deformity surgery.	AUC Brier score	Serious complication Any complication Pneumonia Surgical site infection	0.620 0.650 0.750 0.610	0.965 1.753 1.563 2.686	0.130 0.230 0.050 0.090	The incidence of serious complications was similar between the predicted and observed

REF Country Sample size	Aim and design	Population, surgery and (surgical specialty category)	Validation measures	Outcomes	C-statistic	O/E ratio	Brier score	Comments
159	Improvement Program Surgical Risk Calculator (ACS NSQIP SRC) to accurately predict postoperative outcomes after spinal fusion surgery for patients suffering from adult spinal deformity at a tertiary referral center. Retrospective	(Neurosurgery)		Sepsis Urinary tract infection Reoperation Readmission Discharge to nursing home or rehab Length of stay	0.560 0.560 0.600 0.630 0.590 -	1.462 1.267 2.204 1.533 1.202 0.922	0.040 0.040 0.110 0.120 0.260 -	groups (observed = 16.4% vs predicted = 17.0%, AUC = 0.62, Brier score = 0.13); the β coefficient was nonsignificant, and the AUC was <0.8. Any complication was underpredicted by ACS-NSQIP SRC: (observed 33.3% vs 19%, AUC = 0.65, Brier = 0.23). The ACS Risk Calculator was unreliable in our cohort, leaving many predictive models just marginally better than chance alone.
Karabulut et al (2024) Turkey 300	To investigate whether these two risk calculators (ACS surgical risk calculator and P-POSSUM) accurately reflect actual mortality and morbidity outcomes when retrospectively assessed based on preoperative risk scores, to compare their predictive superiority against each other, and to assess their applicability to major hepatobiliary surgery Retrospective	Patients undergoing major hepatobiliary surgeries between August 2016 and December 2021. Major hepatopancreaticobiliary surgery, pancreaticoduodenectomy, pylorus-preserving pancreaticoduodenectomy, total pancreatectomy, distal pancreatectomy, hepatectomy (right, left, partial,	C-statistic O/E, Brier	Morbidity M1 Severe complication Pneumonia Cardiac complication Surgical site infection Urinary tract infection Venous thromboembolism Renal failure Readmission Reoperation Palliative care Sepsis	0.725 (0.668–0.782)	1.360	0.216 0.092 0.002 0.0004 0.047 0.0005 0.001 0.0002 0.026 0.003 0.306 0.007	Two ACS-NSQIP models are used in this paper: M1 and M2.

REF Country Sample size	Aim and design	Population, surgery and (surgical specialty category)	Validation measures	Outcomes	C-statistic	O/E ratio	Brier score	Comments
		trisegmentectomy), and hepaticojejunostomy surgeries (General surgery)						
Labbot et al (2021a) USA 56	To evaluate the utility of the ACS-NSQIP calculator to predict the risk of complications in patients undergoing reconstruction with a distal femoral replacement, which is known to be associate with a high risk of complications. Retrospective	Patients undergoing distal femur replacement (Orthopaedic surgery)	AUC Brier score	Complications (27365 (Under Excision Procedures on the Femur and Knee Joint) Complications (27447 (Arthroplasty, knee, condyle and plateau) Complications (27486 (Revision of total knee arthroplasty, with or without allograft, 1 component) Complications (27487 (Revision of total knee arthroplasty, with or without allograft, femoral and entire tibial component) Complications (27488 (Repair, Revision, and/or Reconstruction Procedures on the Femur [Thigh Region] and Knee Joint).	0.540 0.450 0.450 0.460 0.460	-	0.413 0.485 0.463 0.456 0.424	Based on receiver-operative curve (ROC) analysis, the use of the ACS-NSQIP score were poor predictors of complications based of the CPT codes: (27365, AUC 0.54); (27447, AUC 0.45); (27486, AUC 0.45); (27487, AUC 0.46); and (27488, AUC 0.46). An ideal model would have an AUC of 1.00, while a model of random chance, ‘‘flip-of-a-coin,” would be an AUC of 0.50. In addition, when performing Brier-score based analysis, the ACS-NSQIP score again demonstrated poor predictive value based off CPT code: 27365 (Brier’s score 0.4127, p < 0.001); 27447 (Brier’s score 0.4848, p < 0.001); 27486 (Brier’s score 0.4629, p < 0.001); 27487 (Brier’s score 0.4564, p < 0.001); and 27488

REF Country Sample size	Aim and design	Population, surgery and (surgical specialty category)	Validation measures	Outcomes	C-statistic	O/E ratio	Brier score	Comments
								(Brier's score 0.4243, p < 0.001), resulting in an average Brier score of 0.4363.
Labbot et al (2021b) USA 103	To evaluate the accuracy of the ACS-NSQIP risk calculator for patients undergoing proximal femoral replacements in the oncologic setting. Retrospective	Patients undergoing proximal femur replacement. (Orthopaedic)	AUC Brier score	Complications (27125 hemiarthroplasty) Complications (27130 total hip) Complications (27132 conversion to total hip) Complications (27134 revision of total hip) Complications (27137 revision acetabulum) Complications (27138 revision femur) Complications (27365 excision tumor hip).	0.576 0.489 0.490 0.489 0.489 0.471 0.538	-	0.348 0.447 0.376 0.369 0.393 0.339 0.385	Based on receiver-operative curve (ROC) analysis, the use of the ACS-NSQIP score were poor predictors of complications based of the CPT codes: (27125, AUC 0.576); (27130, AUC 0.489); (27132, AUC 0.490); (27134, AUC 00.489); (27137, AUC 0.489); (27138, AUC 0.471); and (27365, AUC 0.538). An ideal model would have an AUC of 1.00, while a model of random chance, "flip-of-a-coin," would be an AUC of 0.50. Furthermore, when performing Brier-score based analysis, the ACS- NSQIP score again demonstrated poor predictive value based on CPT code: 27125 (Brier's score 0.348, p < 0.001); 27130 (Brier's score 0.447, p < 0.001); 27132 (Brier's score 0.376, p < 0.001); 27134 (Brier's score 0.369, p < 0.001); 27137

REF Country Sample size	Aim and design	Population, surgery and (surgical specialty category)	Validation measures	Outcomes	C-statistic	O/E ratio	Brier score	Comments
								(Brier's score 0.393, p < 0.001); 27138 (Brier's score 0.339, p < 0.001), and 27365 (Brier's score 0.385, p < 0.001), resulting in an average Brier score of 0.380 (p < 0.001).
Lone et al (2019) USA 462	To evaluate the accuracy of the ACS NSQIP surgical risk calculator in the patients who underwent robot-assisted radical cystectomy. Retrospective	Patients who underwent Radical cystectomy and reconstructed with the ileal conduit and neobladder (RARC) (Urology surgery)	AUC Brier score	Any complication Serious complication Renal insufficiency Pneumonia Surgical site infection Acute urinary tract infection Venous Thromboembolism Cardiac complications Re-admission Return to OR	0.500 (0.45-0.56) 0.530 (0.45- 0.61) 0.640 (0.54-0.73) 0.590 (0.48-0.69) 0.480 (0.43 – 0.53) 0.610 (0.54-0.59) 0.460 (0.33-0.60) 0.590 (0.46-0.71) 0.550 (0.47-0.63) 0.580 (0.47-0.68)	1.689 0.440 2.667 1.750 1.750 2.600 3.250 2.000 0.667 1.400	0.293 0.122 0.075 0.061 0.349 0.138 0.038 0.038 0.111 0.067	The ACS NSQIP calculator demonstrated low accuracy in predicting postoperative outcomes after RARC. The calculated AUC was low (<0.8) for all outcomes. The AUC was 0.50 (95% CI 0.45–0.56) for overall complications and 0.53 (95% CI 0.45–0.61) for serious complications. For cardiac complications, the calculated AUC was 0.59 (95% CI 0.46–0.71). None of the Brier scores (BS) was <0.01, which would have indicated good predictive performance.
Ma et al (2019) USA 554	To evaluate the accuracy of the Surgical Risk Calculator (SRC) of the ACS NSQIP in predicting head and neck microvascular	Patients who received free tissue transfer for head and neck specific indications. 425 myocutaneous, 134 osseous (84	AUC Brier Score	Any complication Serious complication Pneumonia Cardiac complications	0.599 (0.552-0.646) 0.588 (0.538-0.637) 0.661 (0.561-0.760) 0.601	1.552 1.618 1.057 1.795	-	All SRC risk estimates were lower than the actual prevalence of outcome, except death and return to OR. Serious complication and cardiac

REF Country Sample size	Aim and design	Population, surgery and (surgical specialty category)	Validation measures	Outcomes	C-statistic	O/E ratio	Brier score	Comments
	reconstruction outcomes. Prospective	fibula, 47 scapula, and 3 iliac crest), and 2 omental free flaps. (Vascular surgery)		Surgical site infection Urinary tract infection Venous Thromboembolism Renal complications Return to OR Rehabilitation placement	(0.450-0.753) 0.482 (0.408-0.556) 0.728 (0.612-0.845) 0.545 (0.451-0.638) 0.480 (0.306-0.653) 0.535 (0.481-0.589) 0.693 (0.638-0.748)	2.724 0.983 4.805 1.81 1.884 1.058		complication had statistically significant different mean estimated risks among those who developed the complication versus those who did not. While Brier scores ranged from 0.281 to 0.004, all but 2 primary outcomes (UTI and renal) had Brier scores >.0.01 (i.e., not strongly predictive). All primary outcomes evaluated had AUC values < or = to 0.80, ranging from 0.480 to 0.728.
Manhabusqui et al (2023) Brazil 879	To determine how length of stay and complication rates changed over the past 10 years, in comparison to values estimated by the ACS-NSQIP surgical risk calculator. The secondary goals were to determine preoperative patient factors associated with complications, which could lead some clinicians to reconsider their discharge criteria. Retrospective	Patients who underwent primary elective total hip arthroplasty (THA) over 10 years. (Orthopaedic surgery)	ROC curve	Severe complication Any complication surgical site infection urinary tract infection acute renal failure pneumonia readmission venous thromboembolism	0.707 (0.603–0.812) 0.719 (0.643–0.794) 0.494 (0.211–0.778) 0.754 (0.602–0.906) 0.958 (0.933–0.982) 0.954 (0.897–1.000) 0.701 (0.573–0.828) 0.846 (0.809–0.884)	0.572 0.854 0.241 1.356 2.000 1.333 0.084 0.682	-	The reliability of the risk calculator was fair for predicting serious complications (AUC 0.71) or any complication (AUC 0.72)

REF Country Sample size	Aim and design	Population, surgery and (surgical specialty category)	Validation measures	Outcomes	C-statistic	O/E ratio	Brier score	Comments
Mannas et al (2020) Canada 29	To evaluate the predictive value of the ACS NSQIP universal surgical risk calculator in our patients who underwent radical cystectomy. Retrospective	Patients who underwent a Radical cystectomy (RC) for genitourinary cancer without significant deviation from NSQIP surgery codes 51590, 51595, and 51596 (Urology)	AUC Brier score	Serious complication Any complication Pneumonia Cardiac complications Surgical site infection Urinary tract infection venous thromboembolism renal failure 30-day readmission Unplanned to OR Discharged to rehab LOS > 7 days Clavien–Dindo serious (grade 3–5) complication Clavien–Dindo any (grade 1–5) complication	0.600 0.600 0.750 0.800 0.590 0.640 0.610 0.820 0.520 0.520 0.690 0.590 0.560 0.610	1.086 1.108 1.722 0.955 0.977 0.647 1.818 3.222 1.296 0.841 3.630 3.630 - -	0.175 0.189 0.017 0.021 0.111 0.078 0.022 0.009 0.119 0.059 0.034 0.419 0.121 0.361	The predicted rates of complications using the NSQIP USRC were fairly similar to the observed rates of complications, but the model discrimination at the level of the individual patient was poor.
McCarthy et al (2019) USA 641	To assess the American College of Surgeons National Surgery Quality Improvement Program (ACS NSQIP) Risk Calculator's ability to predict 30-day complications after spine surgery. Retrospective	Patients who underwent primary lumbar and cervical fusions. (Orthopaedic surgery)	AUC	Any complication SNF/rehab admit Serious complication venous thromboembolism Surgical site infection Readmit Cardiac Renal Return to OR Pneumonia	0.545 0.673 0.549 0.537 0.649 0.571 0.585 0.776 0.578 0.850	-	-	The Risk Calculator risk estimates significantly predicted (P < 0.001) ''any complication''. 'Any complication'' met the criteria for c > 0.80 for acceptable concordance. Logistical regression results for anterior and posterior patients demonstrated that Risk Calculator predictions were better in the anterior group for ''Any complication'' and ''Serious complication'' than in the posterior

REF Country Sample size	Aim and design	Population, surgery and (surgical specialty category)	Validation measures	Outcomes	C-statistic	O/E ratio	Brier score	Comments
								group. Although complications occurred at a significantly higher rate in the posterior group, the ability for the Risk Calculator to accurately predict complications within this group was worse when compared to the anterior group.
Moses et al (2019) USA 856	To compare and further validate online calculators (ACS-NSQIP and RCRI) with actual adverse cardiac outcomes at a single institution. Retrospective	All patients from January 2011 through December 2015 on the vascular surgical service. Vascular operations: carotid endarterectomy (CEA), infrainguinal lower extremity bypass, open abdominal aortic aneurysm (AAA) repair, and endovascular aneurysm repair (EVAR). (Vascular surgery)	O/E	Adverse cardiac event		1.605		Four different cardiac surgical procedures were tested: carotid endarterectomy (CEA); infrainguinal bypasses; abdominal aortic aneurysm (AAA) repair and endovascular aneurysm repair (EVAR). Pooled data for the entire group documented that ACEs were underpredicted by NSQIP (P = .0055)
Narain et al (2021) USA 253	To determine the effectiveness of the American College of Surgeons National Surgical Quality Improvement Program (ACS NSQIP) surgical	Patients undergoing Anterior lumbar interbody fusion (ALIF) at a single institution. (Orthopaedic surgery)	AUC	Serious complication Any complication Pneumonia Surgical site infection Urinary tract infection venous thromboembolism	0.600 0.610 0.610 0.700 0.570 0.660	3.667 3.412	-	The ACS NSQIP surgical risk calculator was not an adequate predictive tool for aggregate complications (AUC < 0.7)

REF Country Sample size	Aim and design	Population, surgery and (surgical specialty category)	Validation measures	Outcomes	C-statistic	O/E ratio	Brier score	Comments
	risk calculator in the prediction of complications after anterior lumbar interbody fusion. Retrospective			Long length of stay Acute renal injury Readmission Reoperation Adverse discharge	0.530 0.810 0.580 0.530 0.710	1.607		
Pierce et al (2021) USA (Data from the NSQIP) 9143	To calculate the risk for postoperative complications and mortality after corrective surgery of adult spinal deformity patients using the American College of Surgeons National Surgical Quality Improvement Program (ACS-NSQIP) surgical risk calculator (SRC). Retrospective	NSQIP database patients undergoing elective Adult spinal deformity corrective surgery (ASD). (Orthopaedic surgery)	AUC Brier score	Serious complication Any complication Pneumonia Cardiac complication Surgical site infection Urinary tract infection venous thromboembolism Renal failure Readmission Return to OR Discharge to nursing or rehabilitation Sepsis	-	0.964 0.988 1.967 2.647 0.882 1.598 1.484 0.870 1.160 0.974 0.129 1.406	0.0000 1444 0.0000 0196 0.0001 3924 0.0000 3136 0.0000 0576 0.0001 0201 0.0000 5625 0.0000 0009 0.0000 8281) 0.0000 01 0.0034 2225 0.0000 2704	The Brier max is the value between the predicted and observed percentages. Predicted rates via the ACS-NSQIP calculator of any 30-day postoperative complications ranged from 2.8% to 18.5% across CPT codes, where the actual rate in the cohort was 11.4%, and demonstrated good predictive performance via Brier score (0.000002, max: 0.101), as all other outcomes were assessed.
Ravindran et al (2020) USA 100	To assess its applicability in a series of patients undergoing an Ivor Lewis esophagectomy.	Consecutive patients who underwent an Ivor Lewis esophagectomy	AUC	Serious complication Any complication Pneumonia Cardiac Surgical site infection	0.608 0.628 0.652 0.684 0.845	-	-	The c-statistic was generated using logistic regression and was used to assess the capability of the NSQIP

REF Country Sample size	Aim and design	Population, surgery and (surgical specialty category)	Validation measures	Outcomes	C-statistic	O/E ratio	Brier score	Comments
	Retrospective	(General surgery)		Urinary tract infection venous thromboembolism Renal failure Readmission Return to OR Discharge to Rehab	- 0.204 0.631 0.767 0.584 0.608			calculator based on our data set. A model is considered reasonable if its c-statistic is higher than 0.7 and good if >0.8. The results of this study suggest that the ACS NSQIP surgical risk calculator has limited utility as a risk stratification tool for patients undergoing Ivor Lewis esophagectomy. Importantly, although the calculator may predict risk of death and surgical site of infection with acceptable accuracy, most outcome variables generated by the calculator correlated poorly with clinically observed incidence rates.
Rivard et al (2016) USA 1094	To evaluate the ability of the American College of Surgeons (ACS) National Surgical Quality Improvement Program (NSQIP) surgical risk calculator to predict complications in gynecologic oncology patients undergoing laparotomy. Retrospective	Patients who underwent laparotomy on the gynecologic oncology service at a single academic hospital (General surgery)	AUC Brier score	Gynecological oncology patients: Serious complication Any complication Pneumonia Cardiac Surgical site infection Urinary tract infection venous thromboembolism Renal failure Tumour debulking:	 0.644 0.635 0.591 0.708 0.625 0.619 0.655 0.752		 0.148 0.323 0.034 0.011 0.126 0.075 0.018 0.015	The c-statistic and Brier score were used to calculate the prediction capability of the risk calculator. The calculator did not accurately predict most complications (except death, cardiac complications and renal failure). Higher calculated risk scores were associated

REF Country Sample size	Aim and design	Population, surgery and (surgical specialty category)	Validation measures	Outcomes	C-statistic	O/E ratio	Brier score	Comments
				Serious complication Any complication Surgical site infection Urinary tract infection Bowel resection: Serious complication Any complication Surgical site infection Urinary tract infection	0.598 0.574 0.546 0.549 0.614 0.580 0.536 0.703		0.232 0.235 0.134 0.165 0.224 0.229 0.124 0.161	with increased risk of actual complications for all events (p <0.05 in all logistic regression models)
Ryvlin et al (2023) USA 153	To assess the American College of Surgeons National Surgical Quality Improvement Program (ACS-NSQIP) surgical risk calculator performance in patients undergoing surgery for metastatic spine disease. Retrospective	Patients undergoing surgery for metastatic spine disease. Corpectomy, laminectomy (Neurosurgery)	AUC	30 day major complication Corpectomy Laminectomy	0.570 0.555 0.623	1.194 0.950 1.636	-	Poor 30-day major complication discrimination was seen in all procedural cohorts, including overall (AUC = 0.570)
Sevinyan et al (2024) UK 153	To evaluate the accuracy and reliability of the P-POSSUM and ACS-NSQIP surgical risk calculators in predicting postoperative complications in gynaecological–oncological robotic surgery. Retrospective	Patients who had undergone Da Vinci-assisted RS for suspected or confirmed gynaecological malignancies. Gynaecological–oncological robotic surgery (Gynaecology surgery)	AUC, Brier	Morbidity venous thromboembolism Pneumonia Readmission UTI	0.608 0.793 0.657 0.587 0.515	-	0.136 0.013 0.020 0.044 0.063	ACS-NSQIP risk prediction was most accurate for VTE (AUC)-0.793) and pneumonia (AUC-0.657) and it showed 90% accuracy in prediction of five major complications (Brier score 0.01). When comparing the morbidity prediction ability of P-POSSUM and ACS-NSQIP surgical risk calculators,

REF Country Sample size	Aim and design	Population, surgery and (surgical specialty category)	Validation measures	Outcomes	C-statistic	O/E ratio	Brier score	Comments
								ACS-NSQIP showed a better predictive value than P-POSSUM with AUC 0.608 against 0.551 (Figure 1). ACS-NSQIP was found to statistically significantly predict postoperative complications better than P-POSSUM in patients with ASA class I and II.
Tierney et al (2019) USA 320	To determine the accuracy of the American College of Surgeons National Surgical Quality Improvement Program (ACS NSQIP) Surgical Risk Calculator in estimating length of hospital stay and risk of postoperative complications after free tissue transfer surgery. Retrospective	Patients who underwent free flap reconstructions Anterolateral thigh (ALT) flap, fibula free flap (FFF), and radial forearm free flap (RFFF) reconstruction. (Plastic surgery)	AUC Brier	Pneumonia Cardiac complication venous thromboembolism Additional operation within 30 days Skilled nursing facility after discharge	0.750 0.750 0.590 0.500 0.700	0.804 2.571 2.558 0.801 1.841	0.026 0.026 0.032 0.128 0.187	C statistics showed good fit in ALT flap cardiac complications. The NSQIP SRC predictive values approach significance for cardiac complications. The NSQIP SRC fails to accurately forecast any single categorical outcome addressed in this study as measured by the Brier score.
Van der Hulst et al (2022) Netherlands 682	To externally validated the ACS NSQIP surgical risk calculator in a cohort of older patients (≥70 years) who underwent elective colorectal cancer surgery in the Netherlands. Retrospective	Patients ≥70 years undergoing elective non-metastatic Colorectal cancer surgery (CRC) surgery. (General)	Calibration plots AUC	Anastomotic leakage Return to OR Pneumonia Readmission Discharge not to home	0.570 (0.47-0.67) 0.550 (0.47- 0.60) 0.750 (0.67 -0.83) 0.590 (0.50 -0.68) 0.700 (0.62 -0.78)	1.850 2.400 2.444 0.907 0.625	-	We were unable to determine the validity of the ACS NSQIP on other relevant outcomes, e.g., 'any complications', 'serious complications,' and 'cardiac complications,' because the definitions of these complications in the ACS NSQIP

REF Country Sample size	Aim and design	Population, surgery and (surgical specialty category)	Validation measures	Outcomes	C-statistic	O/E ratio	Brier score	Comments
								dataset did not match the definitions in the DCRA dataset or on geriatric outcomes (e.g., functional decline) as we were unable to collect these outcomes from the medical record. The ACS NSQIP surgical risk calculator had a poor individual risk prediction (calibration) for all outcomes and only a fair discriminative ability (discrimination) to predict pneumonia and discharge not to home. The calculator might be considered to identify patients at high risk of pneumonia and discharge not to home to initiate additional preoperative interventions.
Vos et al (2020) USA 452	To assess the predictive performance of the ACS-NSQIP Risk Calculator for adverse events after total gastrectomy for gastric cancer in terms of discrimination and goodness of fit. The secondary aim was to identify the adverse outcomes that the	Patients with gastric cancer who underwent total gastrectomy with curative intent at Memorial Sloan Kettering Cancer Center (General)	AUC Brier scores Hosmer-Lemeshow p value	Any complication Pneumonia Cardiac complication Surgical site infection Urinary tract infection venous thromboembolism renal failure readmission return to operating room	0.530 0.490 0.830 0.600 0.670 0.570 0.790 0.440 0.580	1.552 0.757 2.500 1.563 1.313 2.095 0.833 1.333 0.659	0.272 0.052 0.019 0.195 0.010 0.043 0.015 0.140 0.055	Predictions for adverse outcomes were compared with observed outcomes by Brier scores, c-statistics, and Hosmer-Lemeshow p value. The Hosmer-Lemeshow p value statistic evaluates differences in the probability of observed and predicted

REF Country Sample size	Aim and design	Population, surgery and (surgical specialty category)	Validation measures	Outcomes	C-statistic	O/E ratio	Brier score	Comments
	calculator accurately predicts and for which it is therefore clinically applicable Retrospective			discharge to nursing or rehabilitation facility length of stay (day) Longer length of stay	0.900 - 0.560	0.181 1.122 1.091	0.031 - 0.235	events across deciles of increasing predicted risk. A Hosmer-Lemeshow p value < 0.05 leads to rejection of the null hypothesis that the model is well calibrated. For adverse outcomes after total gastrectomy with curative intent in gastric cancer patients, performance of the ACS-NSQIP Risk Calculator is variable. Its predictive performance is best for cardiac complications, renal failure, death, and discharge to nursing or rehabilitation facility.
Wang et al (2017) China 242	To evaluate the predictive value of the ACSNSQIP calculator in geriatric patients undergoing lumbar surgery. Retrospective	Geriatric patients (>60) who underwent lumbar surgery (Conventional decompressive laminectomy without fusion). (Neurosurgery)	AUC Brier score Hosmer-Lemeshow p value	Serious complications Any complication Surgical site infection Pneumonia Cardiac complication Urinary tract infection Venous Thrombosis Renal failure	0.666 (0.59 - 0.738) 0.683 (0.615 -0.751) 0.427 (0.362 -0.493) 0.691 (0.612 -0.769) 0.648 (0.538 -0.759) 0.648 (0.494 -0.714) 0.432 (0.222 -0.642) 0.825	-	0.241 0.321	The predictive value of the ACS-NSQIP model was assessed using the Hosmer–Lemeshow test, Brier score (B), and receiver operating characteristics (ROC, also referred C-statistic) curve analysis. Observed and predicted incidence of postoperative complications was 43.8% and 13.7% (±5.9%) (P<.01), respectively. The

REF Country Sample size	Aim and design	Population, surgery and (surgical specialty category)	Validation measures	Outcomes	C-statistic	O/E ratio	Brier score	Comments
				Readmission Reoperation	(0.641 -1.000) 0.843 (0.733 -0.952) 0.311 (0.179 -0.443)			Hosmer–Lemeshow test demonstrated adequate predictive accuracy of the ACS-NSQIP model for all complications. However, Brier score showed that the ACS-NSQIP model could not accurately predict risk of all (B=0.321) or serious (B=0.241) complications; this was supported by ROC curve analysis.
Wherley et al (2020) USA 354	To evaluate the accuracy of the American College of Surgeons National Surgery Quality Improvement Program (ACS NSQIP) surgical risk calculator in predicting postoperative complications in patients undergoing pelvic organ prolapse surgery. Retrospective	Patients who underwent surgery for pelvic organ prolapse. Apical prolapse repair surgery included sacrocolpopexy, sacrohysteropexy, uterosacral ligament suspension, sacrospinous ligament suspension, iliococcygeus suspension, and colpocleisis (Urogynecology surgery)	AUC Brier score	Urinary tract infection - primary prolapse Urinary tract infection - High risk Surgical site infection - primary prolapse Surgical site infection - high risk Readmission Nonhome discharge Reoperation Pneumonia Renal failure venous thromboembolism Sepsis Cardiac complication	0.589 0.547 0.592 0.588	4.351 10.818 1.667 1.571 0.786 5.500 8.000 2.000 1.000 1.500	0.150 0.151 0.116 0.112	AUC ROC curves for each of the primary and secondary outcomes depict the low predictive performance of the calculator.

REF Country Sample size	Aim and design	Population, surgery and (surgical specialty category)	Validation measures	Outcomes	C-statistic	O/E ratio	Brier score	Comments
Willoughby et al (2023) New Zealand 210	To validate the predictive accuracy of both the SpineSage and ACS-NSQIP surgical risk calculators in patients over the age of 80 years, undergoing spine surgery for any reason. Retrospective	Patients over the age of 80 years, undergoing spinal surgery for any reason. (Orthopaedic surgery)	AUC Brier score	Major medical complications Overall medical complications Risk group <10% Risk group 10 – 20% Risk group >20%	0.675 0.688 - - -	1.333 1.402 1.000 1.359 1.247	0.100 0.110 - - -	The ACS surgical risk calculator significantly distinguished between those who developed complications and those who did not. The mean calculated risk of experiencing any complication in the group of patients that had at least one complication was 16.3%, compared with 11.2% in the group that did not have any complications (P = 0.001). The mean calculated risk of a serious complication occurring in the group of patients with at least one serious complication was 14.5% compared with 10.3% in those that did not experience any serious complication (P = .004)
Yap et al (2018) Philippines 424	To validate the ACS-NSQIP calculator in a Filipino population and compare its predictive ability with the Revised Cardiac Risk Index. Prospective	All patients aged 19 years and older admitted from January 2016 to March 2017, referred for preoperative evaluation and cardiopulmonary risk stratification before non-cardiac surgery.	AUC/ Brier	MA Cardiac Event Morbidity Pneumonia Surgical site infection Urinary tract infection Venous thromboembolism Renal failure Return to operating room	0.930 0.880 0.930 0.690 0.650 0.630 0.760 0.780	-	0.080 0.160 0.120 0.320 0.230 0.090 0.060 0.220	Primary outcome: MACE—cardiac arrest, acute myocardial infarction (ST elevation or non-elevation acute coronary syndrome), heart failure. Secondary outcomes: Morbidity—any postoperative

REF Country Sample size	Aim and design	Population, surgery and (surgical specialty category)	Validation measures	Outcomes	C-statistic	O/E ratio	Brier score	Comments
		Open, laparoscopic and percutaneous abdominal surgeries, anorectal surgeries, breast surgeries, thyroid surgeries, head and neck surgeries, orthopaedic surgeries, urologic surgeries, excision and incision biopsies of superficial masses, wound debridement, vascular surgeries, and neurosurgical procedures (Mixed surgery)						complication. The discriminative ability of the ACS NSQIP Surgical Risk Calculator was considered excellent if the AUC was 0.90 to 1.0, good if 0.80 to 0.89, fair if 0.70 to 0.79, poor if 0.60 to 0.69, and fail if 0.50 to 0.59.
Yung et al (2022) Australia 200	To assess the validity of the SRC in patients undergoing microsurgical free flap reconstruction at an Australian tertiary referral centre. Retrospective	Patients undergoing microsurgical free flap reconstruction at an Australian tertiary referral centre. (Plastic surgery)	Brier score ROC AUC Hosmer–Lemeshow test.	Serious complication Any complication Pneumonia Surgical site infection Urinary tract infection Readmission Return to OR Discharge to nursing or rehabilitation Sepsis	0.539 (0.457-0.622) 0.699 (0.554-0.845) 0.913 (0.846-0.981) 0.679 (0.574-0.784) 0.841 (0.672-0.920) 0.591 (0.426-0.756) 0.493 (0.377-0.610) 0.585 (0.452-0.718) 0.795	0.901 1.027 0.778 0.859 35.211 0.698 1.187 0.705 0.581	0.273 0.087 0.019 0.128 0.025 0.048 0.129 0.084 0.016	Notably, for “any complication”, the Brier score was small, and the Hosmer–Lemeshow test indicated good calibration, but the ROC AUC showed poor discrimination. This suggests that the SRC is better calibrated for any complication for the ablative component compared to the reconstructive component. In patients undergoing microsurgical free flap

REF Country Sample size	Aim and design	Population, surgery and (surgical specialty category)	Validation measures	Outcomes	C-statistic	O/E ratio	Brier score	Comments
					(0.592-0.999)			reconstruction following complex head and neck surgery, the SRC has poor predictive performance for postoperative morbidity.

Table 15. Data Extraction Tables: ASA classification system

REF Country Sample size	Aim and design	Population, surgery and (surgical specialty category)	Validation measures	Outcomes	C-statistic	O/E ratio	Brier score	Comments
Aceto et al 2021 Italy 105	The primary aim of this prospective cohort study was to evaluate the usefulness of the modified Frailty Index (mFI) score to predict postoperative pulmonary complications (PPCs) in elderly patients undergoing major open abdominal surgery. The secondary purpose was to compare the prediction power of mFI, Ariscat (Assess Respiratory Risk in Surgical Patients in Catalonia), and American Society physical status classification (ASA) scores. Prospective	Patients aged ≥65 years undergoing Open major upper elective (partial/total colectomy; Hartmann's procedure; total/partial gastrectomy; liver resection; and pancreatic-duodenectomy) and lower (nephrectomy, prostatectomy or hysterectomy) abdominal surgery. (Mixed surgery)	ROC/AUC	Postoperative pulmonary complications	0.690	-	-	

REF Country Sample size	Aim and design	Population, surgery and (surgical specialty category)	Validation measures	Outcomes	C-statistic	O/E ratio	Brier score	Comments
Bronheim et al 2018 USA 52066	To determine the ability of RCRI to predict non-cardiac adverse events after Posterior Lumbar Decompression (PLD) Retrospective	Adults (≥18 years), undergoing Posterior Lumbar Decompression (PLD) between 2006 and 2014 (Orthopaedic)	AUC	Unplanned intubation Pulmonary embolism Ventilated >48 hours Acute renal failure Cerebrovascular accident/stroke with neurologic deficit Coma >24 hours Sepsis Septic shock Reoperation Superficial Surgical site infection (SSI) Deep incisional SSI Organ space SSI Wound dehiscence Pneumonia Progressive renal insufficiency Urinary tract infection Peripheral nerve injury Bleeding transfusions Deep vein thrombosis/thrombophlebitis Readmission-	0.741 (0.736 – 0.745) 0.812(0.808 – 0.816) 0.742 (0.73.8 – 0.747) 0.789 (0.784 – 0.793) 0.838 (0.835 – 0.842) 0.653 (0.648 – 0.658) 0.905 (0.902 – 0.908) 0.759 (0.755 – 0.764) 0.866 (0.862 – 0.870) 0.842 (0.839 – 0.845) 0.950 (0.948 – 0.953) 0.776 (0.772 – 0.780) 0.792 (0.788 – 0.796) 0.824 (0.820 – 0.828) 0.814 (0.810 – 0.819) 0.825 (0.821 – 0.828) 0.513 (0.507 – 0.518) 0.800 (0.796 – 0.804) 0.782 (0.778 – 0.787) 0.906 (0.903 – 0.909)	-	-	ASA status had a fair discriminative ability (AUC = 0.770) for a composite of any non-cardiac complication.
Bronheim et al 2019 USA 52066	To determine the ability of Revised Cardiac Risk Index (RCRI) to predict adverse cardiac events following posterior lumbar decompression Retrospective	Adult (≥18 years), undergoing Posterior Lumbar Decompression PLD between 2006 and 2015 (Orthopaedic surgery)	AUC	Cardiac arrest requiring CPR Myocardial Infarction	0.674 (0.669 – 0.678) 0.799 (0.795 – 0.803)	-	-	ASA had discriminative abilities of ‘‘fair’’ for MI (AUC = 0.799) and ‘‘poor’’ for cardiac arrest requiring CPR (AUC = 0.674)
Chrisant et al 2024	To assess the accuracy of RCRI compared to the ASA-PS classification	Patients ≥18 admitted for elective	AUC	Pulmonary complications	0.760 (0.66 – 0.87) 0.750 (0.59 – 0.91)			There was no significant difference in the predictive ability

REF Country Sample size	Aim and design	Population, surgery and (surgical specialty category)	Validation measures	Outcomes	C-statistic	O/E ratio	Brier score	Comments
Tanzania 184	system in preoperative risk assessment for pulmonary and cardiac problems among non-cardiothoracic surgery patients Prospective	non-cardiothoracic surgery (Mixed)		Cardiac complications)				of the two tools. Both RCRI and ASA-PS classification systems were equally effective in predicting these complications.
Feghali et al (2022) USA The institutional database included 275 patients (88 clip placement, 187 endovascular treatment). AND 1047 patients who underwent clip placement were included in the NSQIP database.	To validate the utility of mFI-5 for predicting endovascular and microsurgical treatment outcomes in patients with unruptured aneurysms. Retrospective	Patients with unruptured aneurysm who were treated with clip placement or endovascular therapy between January 2017 and December 2018 from the institutional database and NSQIP database (Vascular surgery)	AUC C-statistic	Major complications for clip placement Major complications for endovascular therapy	0.541 0.606			Goodness of fit was indicated by a non-significant p value determined with the Hosmer-Lemeshow test. The AUC of ASA was 0.606 (p = 0.219).

REF Country Sample size	Aim and design	Population, surgery and (surgical specialty category)	Validation measures	Outcomes	C-statistic	O/E ratio	Brier score	Comments
Fu et al (2018) USA 10,527	Research questions: (1) Which demographic/anthropometric variable among age, sex, and body mass index (BMI) has the best discriminative ability as measured by receiver operating characteristics (ROC) in its association with adverse events following Total Shoulder Arthroplasty? (2) Which comorbidity index, among the American Society of Anesthesiologists (ASA) classification, the mCCI, or the mFI, has the best ROC in its association with adverse events following Total Shoulder Arthroplasty? Retrospective	Patients who underwent Total shoulder arthroplasty (TSA) from 2005 to 2015 were identified from the National Surgical Quality Improvement Program (NSQIP) (Orthopaedic surgery)	AUC	Extended length of stay Discharge to higher level of care	0.630 (0.614-0.646) 0.630 (0.616-0.645)			
Hightower et al 2010 USA 32	To compare the risk predictive value of preoperative physiological capacity (PC: defined by gas exchange measured during cardiopulmonary exercise testing) with the ASA physical status classification in the same patients (n=32) undergoing major abdominal cancer surgery.	Patients (18+) undergoing elective major abdominal cancer surgery(Gastrectomy, Pancreatectomy, Radical cystectomy, Radical nephrectomy, Radical transabdominal tumour debulking, Pelvic exenteration, Low anterior	AUC	Postoperative morbidity	0.688 (0.52315 - 0.85185)			Three newly identified PC measures and the ASA rank were significantly associated with postoperative morbidity; none showed a statistically greater association compared with the others.

REF Country Sample size	Aim and design	Population, surgery and (surgical specialty category)	Validation measures	Outcomes	C-statistic	O/E ratio	Brier score	Comments
	Prospective	resection, Retroperitoneal lymph node dissection). (General surgery)						
Lim et al (2017) USA 6,148	This study aimed to evaluate the predictive value of ASA classification system on 30- day morbidity following single-level elective anterior cervical discectomy and fusion. Retrospective	Patients who underwent Single-level elective anterior cervical discectomy and fusion (SLE-ACDF) between 2011 and 2013 were selected from the NSQIP database. (Orthopaedic surgery)	C-statistic	Any complication Catastrophic outcome Airway complication Pneumonia Unplanned intubation Ventilator >48hours	0.706 (0.48–1.41) 0.793 (0.39–2.71) 0.719 (0.67–7.29) 0.700 (0.33–4.56) 0.678 (0.41–5.36) 0.825 (0.69–50.96)			C-statistics used in the regression models demonstrate sufficient discriminability of these models.
McConaghy et al (2021) USA 202,488 (THA) 230,823 (TKA)	The purpose of this study was to review the (1) mCCI; (2) ASA physical status classification system; (3) ECM; and (4) mFI-5 to determine their ability to accurately predict 30-day mortality, 30-day rate of major and minor complications, discharge disposition, and extended length of stay in the hospital following Total hip arthroplasty or Total knee arthroplasty Retrospective	Patients who underwent elective Total hip arthroplasty or Total knee arthroplasty from January 2011 through December 2019 from the NSQIP database (Orthopaedic surgery)	C-statistic	Any major complications (30 days) Any minor complications (30 days) Discharge to not home Length of stay >1 day	THA 0.615 (0.609-0.621) TKA 0.615 (0.607-0.617) THA 0.608 (0.603-0.613) TKA 0.608 (0.603-0.611) THA 0.640 (0.636-0.642) TKA 0.639 (0.632-0.636) THA 0.580 (0.572-0.582) TKA 0.561 (0.559-0.563)			ASA was not predictive of major and minor complications.

REF Country Sample size	Aim and design	Population, surgery and (surgical specialty category)	Validation measures	Outcomes	C-statistic	O/E ratio	Brier score	Comments
Meng et al (2018) USA 1,516	Primary study aim was to compare the discriminative ability of the ASA, CCI, and mFI with the occurrence of postoperative adverse outcomes in patients undergoing Radical Cystectomy. Retrospective	Patients undergoing elective Radical cystectomy (RC) for bladder cancer were extracted from the 2005 to 2011 NSQIP database (Urology surgery)	AUC	Any adverse event Minor adverse event Serious adverse event Infectious adverse events Length of stay Discharge to a higher level of care	0.510 (0.485-0.534) 0.514 (0.486-0.542) 0.519 (0.491-0.547) 0.517 (0.490-0.543) 0.536 (0.510-0.561) 0.590 (0.556-0.624)			
Ondeck et al (2018) USA 16,495	This study aimed to compare the discriminative ability of ASA, mCCI, and mFI, as well as demographic factors including age, body mass index, and gender for perioperative adverse outcomes following Posterior lumbar fusion. Retrospective	Patients undergoing elective posterior lumbar fusion (PLF) with or without interbody fusion were extracted from the 2011–2014 NSQIP database. (Orthopaedic surgery)	AUC	Any adverse event Severe adverse event Minor adverse event Infectious adverse event Extended length of stay Discharge to higher level of care	0.567 (0.555–0.579) 0.571 (0.553–0.588) 0.571 (0.557–0.585) 0.580 (0.560–0.599) 0.564 (0.556–0.572) 0.631 (0.621–0.641)			The combination of the best overall performing comorbidity index across the board (ASA) and demographic factor across the board (age) for the occurrence of any adverse event following PLF led to an AUC of 0.603 (95% CI: 0.590–0.616), which was statistically better than ASA alone and age alone.
Sankar et al (2014) Canada 10,864	The primary objective was to evaluate the inter-rater agreement of ASA-PS scores assigned at outpatient preoperative assessment clinics vs operating theatres. Retrospective	Adult patients (≥18 yr) who underwent elective non-cardiac surgery (ENT surgery, General surgery, Gynaecology, Neurosurgery, Opthalmology, Orthopaedic surgery, Plastic surgery, Thoracic	AUC	Myocardial injury (Ratings in the clinic) Myocardial injury (Ratings in the operating theatre)	0.700 (0.65 – 0.75) 0.750 (0.71–0.79)			

REF Country Sample size	Aim and design	Population, surgery and (surgical specialty category)	Validation measures	Outcomes	C-statistic	O/E ratio	Brier score	Comments
		surgery, Urology, Vascular surgery). (Mixed surgery)						
Wolters et al 2006 Germany 107	To prospectively apply various scoring systems in order to estimate outcome in patients undergoing aortobifemoral surgery due to arterial occlusive disease at the aorto-iliac level Prospective	Patients who received an aortobifemoral or aorto-iliac graft due to arterial occlusive disease between 1996 to 2000. (Vascular surgery)	AUC	Morbidity	0.518			In this study there was no significant correlation between the ASA classification and the postoperative mortality or morbidity.

Table 16. Data Extraction Tables: RCRI

REF Country Sample Size	Aim and design	Population, surgery and (surgical specialty category)	Validation measures	outcomes	C-statistic	O/E ratio	Brier score	Comments
Alphonsus et al (2021) South Africa 3,045	Using RCRI to quantify perioperative risk for complications in these patients. Retrospective	Patients aged ≥45 years presenting for surgery in the combined SASOS and ASOS data set. Non-cardiac surgery (Orthopaedic, Breast, Obstetrics, Gynaecology, Upper gastrointestinal, Lower gastrointestinal, Hepatobiliary, Urology, Vascular, Head and Neck, Plastics, Thoracic, Neurosurgery, other) (Mixed surgery)	AUC	Cardiac complications	**0.680 (0.57 -0.79)**	-	-	The discrimination of the RCRI for cardiac complications was described using the receiver operating characteristic (ROC) and area under the curve (AUC). The calibration between the original derivation cohort and the ASOS cohort was compared using the RR for cardiac complications in the African cohort for each RCRI risk factor.

REF Country Sample Size	Aim and design	Population, surgery and (surgical specialty category)	Validation measures	outcomes	C-statistic	O/E ratio	Brier score	Comments
								The likelihood ratios for low- (0 RCRI risk factors), intermediate- (1 - 2 RCRI risk factors) and high-risk (≥3 RCRI risk factors) categories were calculated.
Alrezk et al (2017) USA 172,905	To investigate the performance of the RCRI and Gupta MICA perioperative cardiac risk models in a geriatric population. Retrospective	NSQIP geriatric cohort >65 non-cardiac surgery (Hernia, Anorectal, Aortic, Bariatric, Brain, Breast, ENT, Foregut/ hepatopancreatobiliary, GBAAS/intestinal, Neck, Obstetric/gynecologic, Orthopedic, Other abdomen, Peripheral vascular, Skin, Spine, Thoracic, Vein, Urology) (Mixed surgery)	AUC	Cardiac risk	0.680 (0.67-0.69)			Although the RCRI underestimates risk for low-risk patients, it is well calibrated for the highest-risk group (2.4% to 5.4% in RCRI). The RCRI was not derived to predict the cardiac risk within 30 days of surgery but is aimed solely at predicting the risk during a hospital stay. In the modern world, using equations developed so long ago and on unique populations (i.e., RCRI) is of questionable value, particularly in an era when curated data sources such as the NSQIP and other large data sets are readily available.
Andersson et al (2015)	To assess the relationship of RCRI with	Patients >25 undergoing major elective, non-cardiac surgery (Ear-	C-Statistic	Major adverse cardiovascular events	0.761	-	-	Modelling RCRI classes as a continuous variable, C

REF Country Sample Size	Aim and design	Population, surgery and (surgical specialty category)	Validation measures	outcomes	C-statistic	O/E ratio	Brier score	Comments
Denmark 447,352	major adverse cardiovascular events in an unselected cohort of patients undergoing elective, non-cardiac surgery. Retrospective	nose-throat, Major orthopedic, Minor orthopedic, Abdominal (bowel) Abdominal (nonbowel), Breast, Plastic, Endocrine, Eye, Female reproductive, Male reproductive, Intercranial, Neuro, Nonarterial vessels, Thoracic, Urologic, Vascular) **(Mixed surgery)**						statistic was highest among age group 56 to 65 years (0.772) and lowest for those aged >85 years (0.683). In a nationwide unselected cohort, the performance of the RCRI was similar to that of the original cohort. Having ≥1 risk factor was of moderate sensitivity, but high negative predictive value for all ages.
Archan et al (2010) Austria 225	To determine if the Revised Cardiac Risk Index (Lee) is useful for stratification of patients by risk of both perioperative cardiac morbidity and long-term all-cause mortality Retrospective	Patients with abdominal aortic aneurysms admitted to the authors' institution from 1999 to 2006. All patients underwent endovascular aortic aneurysm repair (Vascular surgery)	AUC	Postoperative cardiac events	0.730			The AUCs for the prediction of postoperative cardiac events by age and GAS (Glasgow Aneurysm Score) were 0.60 and 0.67, respectively. In the present study, the authors found a highly significant association between a Lee index ≥3 and postoperative cardiac events (p 0.004). These results indicate that the Lee index, even if its overall discriminatory ability is fair at best, is a useful tool for the identification of high-

REF Country Sample Size	Aim and design	Population, surgery and (surgical specialty category)	Validation measures	outcomes	C-statistic	O/E ratio	Brier score	Comments
								risk patients from both a short- and long-term perspective. The Lee index was developed for the prediction of prospectively detected major cardiac complications.
Bronheim et al 2018 USA 52066	To determine the ability of RCRI to predict non-cardiac adverse events after Posterior Lumbar Decompression (PLD) Retrospective	Adult (≥18 years), undergoing Posterior Lumbar Decompression (PLD) between 2006 and 2014 (Orthopaedic surgery)	AUC	Unplanned intubation Pulmonary embolism Ventilated >48 hours Acute renal failure Cerebrovascular accident/stroke with neurologic deficit Coma >24 hours Sepsis Septic shock Reoperation Superficial Surgical site infection (SSI) Deep incisional SSI Organ space SSI Wound dehiscence Pneumonia Progressive renal insufficiency Urinary tract infection Peripheral nerve injury Bleeding transfusions Deep vein thrombosis/thrombophlebitis Readmission	0.837 (0.833-0.842) 0.409 (0.403 -0.415) 0.845 (0.841 -0.850) 0.881 (0.876 -0.887) 0.747 (0.742 -0.753) 0.900 (0.868 -0.933) 0.826 (0.821 -0.830) 0.845 (0.840 -0.850) 0.850 (0.845 -0.855) 0.717 (0.711 -0.722) 0.879 (0.875 -0.883) 0.876 (0.872 -0.879) 0.717 (0.712 -0.723) 0.739 (0.733 -0.744) 0.845 (0.839 -0.851) 0.738 (0.733 -0.744) 0.073 (0.071 -0.076) 0.712 (0.706 -0.717) 0.707 (0.701 -0.713) 0.835 (0.829 -0.840)			RCRI had a poor discriminative ability (AUC = 0.623) for a composite of any non-cardiac complication.
Bronheim et al 2019 USA 52066	To determine the ability of Revised Cardiac Risk Index (RCRI) to predict adverse	Adult (≥18 years), undergoing Posterior Lumbar Decompression (PLD) between 2006 and 2015	AUC	Cardiac arrest requiring CPR Myocardial Infarction	0.855 (0.851-0.860) 0.876 (0.872-0.880)			RCRI had a good discriminative ability to predict both MI [area under the curve (AUC) = 0.876] and cardiac

REF Country Sample Size	Aim and design	Population, surgery and (surgical specialty category)	Validation measures	outcomes	C-statistic	O/E ratio	Brier score	Comments
	cardiac events following posterior lumbar decompression Retrospective	(Orthopaedic surgery)						arrest requiring CPR (AUC = 0.855).
Carabini et al (2014) Lebanon 547	To determine the accuracy of the RCRI in predicting MACE in patients undergoing spine fusion surgery of 3 or more bony levels. Retrospective	Patients undergoing spine fusion surgery of 3 levels or more (Orthopaedic surgery)	AUC	Major adverse cardiac events	0.540 (0.47-0.61)			The RCRI did not predict cardiac morbidity in our patients undergoing major spine fusion surgery, despite being extensively validated in low-risk non-cardiac surgical patients.
Chrisant et al 2024 Tanzania 184	To assess the accuracy of RCRI compared to the ASA-PS classification system in preoperative risk assessment for pulmonary and cardiac problems among non-cardiothoracic surgery patients Prospective	Patients ≥18 admitted for elective non-cardiothoracic surgery (Mixed surgery)	AUC	Pulmonary complications Cardiac complications	0.710 (0.59 - 0.83) 0.730 (0.56 – 0.91)			There was no significant difference in the predictive ability of the two tools. Both RCRI and ASA-PS classification systems were equally effective in predicting these complications.

REF Country Sample Size	Aim and design	Population, surgery and (surgical specialty category)	Validation measures	outcomes	C-statistic	O/E ratio	Brier score	Comments
Cohn and Ros 2018 USA 663	This study compared these risk calculators and a reconstructed Revised Cardiac Risk Index in predicting postoperative cardiac complications, both during hospitalization and 30 days after operation, in a patient cohort who underwent select surgical procedures in various risk categories. Retrospective	A patient cohort who underwent select surgical procedures in various risk categories Non cardiac surgery (15.7% high risk; 62.9% Intermediate risk; 21.4% low risk) (Mixed surgery)	ROC	All cardiac complications (in hospital) All cardiac complications (-30 days) Major cardiac complications	0.850 0.780 0.550			In this study, the C-statistics (95% confidence interval) for RCRI and R-RCRI showed good discrimination for cardiac complications.
Davis et al (2013) Canada 9,519	To re-examine the validity of the inclusion of the two predictors, diabetes and chronic renal failure, in the RCRI using a large modern prospectively collected data set.	Patients over the age of 50 undergoing elective non-cardiac surgery with an expected length of stay > two days (Abdominal, Orthopaedic, Thoracic, Vascular, genitourinary, neurosurgery, and ear nose and throat surgeries (Mixed surgery)	AUC	Major cardiac complications	0.790 (0.76 - 0.83)			A simplified 5-Factor model using a high-risk type of surgery, a history of ischemic heart disease, congestive heart failure, cerebrovascular disease, and a preoperative GFR \30 mLmin-1 results in superior prediction of major cardiac complications following

REF Country Sample Size	Aim and design	Population, surgery and (surgical specialty category)	Validation measures	outcomes	C-statistic	O/E ratio	Brier score	Comments
	Prospective							elective non-cardiac surgery compared to RCRI.
Dunn et al (2019) USA 503	To compare the utility of the three different CV risk calculators using data from our own transplant center, to assess which of the models most accurately predicts long-term and short term Major Adverse Cardiac Events (MACE) Retrospective	Transplant database, all adult patients who underwent kidney transplant from 2005 to 2010 (Urology / Transplant)	AUC	30 day major adverse cardiac event 1 year major adverse cardiac event	0.633 0.661			This study focussed on which of the three models performed best: Of the three calculators, PORT performed best when the sensitivity was set at a clinically relevant level.
Moses et al 2019 USA 856	To compare and further validate online calculators with actual adverse cardiac outcomes at a single institution Retrospective	All patients from January 2011 through December 2015 on the vascular surgical service. Vascular operations: carotid endarterectomy (CEA), infrainguinal lower extremity bypass, open abdominal aortic aneurysm (AAA) repair, and endovascular aneurysm repair (EVAR) (Vascular surgery)	O/E	Adverse cardiac events	-	2.231	-	For open AAA repair RCRI strongly underpredicted the adverse events (P ≤ 0.0001). Pooled data showed adverse cardiac events (ACEs) were underpredicted by RCRI (P ≤ .001).

REF Country Sample Size	Aim and design	Population, surgery and (surgical specialty category)	Validation measures	outcomes	C-statistic	O/E ratio	Brier score	Comments
Pandey et al (2015) USA 1,568	To evaluate the association between the presence of stable anginal symptoms and risk for adverse postoperative cardiovascular outcomes in patients with a history of MI scheduled to undergo a major non-cardiac surgery. Retrospective	Patients who underwent 1 of the 15 elective non-cardiac procedures who had a concomitant documented history of a recent MI from the NSQIP database (2005-2011). (Carotid Endarterectomy, Lower extremity bypass, Abdominal aortic aneurysm repair, Total hip replacement, Total knee replacement, Bariatric surgery, Esophagectomy, Gastrectomy, Pancreatectomy, Colectomy, Nephrectomy, Cystectomy, Prostatectomy, Hysterectomy, Pneumonectomy (Mixed surgery)	AUC	Adverse cardiac event	0.590	-	-	We quantified the predictive accuracy of the RCRI and the modified RCRI (treating preoperative angina and MI as separate inputs in the score) in predicting an adverse cardiac outcome in the present patient population using the area under the curve (AUC) analysis. All input variables were given a score of 0 or 1 (0 = no, 1 = yes), in accordance with the original RCRI described. The Mantel-Haenszel test was used to assess the statistical significance of the difference in the discriminatory ability between the 2 models.
Peterson et al (2016) USA 1,096	To determine whether the NSQIP MICA risk calculator could accurately discriminate perioperative MICA in patients undergoing elective hip and	Adult patients at Penn State Milton S. Hershey Medical Center from January 1, 2013, to December 31, 2014, Elective total hip replacement and total knee replacement surgeries (Orthopaedic surgery)	C-Statistic	Adverse cardiac events	0.90 (0.75-1)			The NSQIP and RCRI calculators perform comparably in discriminating adverse cardiac events in patients undergoing elective hip and knee surgery.

REF Country Sample Size	Aim and design	Population, surgery and (surgical specialty category)	Validation measures	outcomes	C-statistic	O/E ratio	Brier score	Comments
	knee replacement surgery and compare with RCRI. Retrospective							
Schmidt et al (2024) Germany 199	To compare the predictive value of the RCRI, AUB-HAS2, and Andersson scores with preoperative NT-proBNP for the postoperative 30-day morbidity in an observational non-cardiac and non-vascular surgery cohort Retrospective	Patients aged>65 years scheduled for elective non-cardiac non-vascular surgery in general anaesthesia with intermediate or high surgical risk (elective intracranial, thoracic, head and neck, trauma and orthopaedic, or abdominal surgery, including visceral, urological, and gynaecological operations) (Mixed surgery)	AUC/ROC	Morbidity Rehospitalisation Acute kidney injury Infection Acute decompensated heart failure	0.560 0.457 0.687 0.573 0.590	-	-	AUB-HAS2, but not RCRI or Andersson score, significantly predicted the CME (AUB-HAS2: AUCROC 0.646, p<0.001; RCRI: AUCROC 0.560, p=0.126; Andersson: AUCROC 0.487, p=0.760). In our analysis, the modified RCRI, including the weighted preoperative NT-proBNP cut-of and original RCRI components, showed improved predictively compared to the original RCRI.
Yap et al (2018) Philippines 3,045	To validate the ACS-NSQIP calculator in a Filipino population and compare its predictive ability with the Revised	All patients aged 19 years and older admitted from January 2016 to March 2017, referred for preoperative evaluation and cardiopulmonary risk stratification before non-cardiac surgery. Open, laparoscopic and	AUC/ Brier	Major adverse cardiac event	0.930			Primary outcome: MACE—cardiac arrest, acute myocardial infarction (ST elevation or non-elevation acute coronary syndrome), heart failure. Secondary outcomes: 2. Morbidity—any

REF Country Sample Size	Aim and design	Population, surgery and (surgical specialty category)	Validation measures	outcomes	C-statistic	O/E ratio	Brier score	Comments
	Cardiac Risk Index. Prospective	percutaneous abdominal surgeries, anorectal surgeries, breast surgeries, thyroid surgeries, head and neck surgeries, orthopaedic surgeries, urologic surgeries, excision and incision biopsies of superficial masses, wound debridement, vascular surgeries, and neurosurgical procedures **(Mixed surgery)**						postoperative complication

Table 17. Data Extraction Tables: P-POSSUM

REF Country Sample Size	Aim and design	Population, surgery and (surgical specialty category)	Validation measures	Outcomes	C-statistic	O/E ratio	Brier score	Comments
Angelucci et al 2024 Italy 567	To investigate the accuracy of ACS-NSQIP and P-POSSUM risk calculators in predicting postoperative outcomes for patients undergoing retroperitoneal sarcoma surgery	Adult patients (≥18), undergoing elective comprehensive resection for primary or persistent adult-type retroperitoneal sarcoma with ASA score I–IV (General surgery)	AUC, Brier score	Any complication Severe complication	0.660 (0.61–0.70) 0.630 (0.58–0.68)		0.229 0.205	

	Retrospective							
Bodea et al (2018) Romania 113	To evaluate the capacity of P-POSSUM risk scores concerning the morbidity and mortality following standard ncreaticod -uodenectomy and pancreaticod -uodenectomy associated with portal vein Retrospective	Consecutive pancreaticoduodenectomy performed for periampullary malignant tumors during July 2013-December 2015. (General surgery)	ROC	All complications Pancreaticoduodenectomy specific complications	0.610 0.640	1.660 0.890		The predictive power of P-POSSUM score was higher in PD specific complications and subsequent mortality compared to the whole morbidity and mortality rates.
He et al 2023 China 349	To assess the suitability of POSSUM and its modified versions, E-PASS and its modified score, SRS, and SORT scores for predicting postoperative complications and mortality in patients undergoing laparoscopic radical gastrectomy for gastric cancer Retrospective	Patients who were histopathologically diagnosed with gastric adenocarcinoma and underwent R0 resection (Laparoscopic radical gastrectomy) (General surgery)	AUC	Early complications after laparoscopic radical gastrectomy	0.781			P-POSSUM performed moderately for the prediction of postoperative complication morbidity in laparoscopic surgery. P-POSSUM underestimated post-operative mortality.
Jones et al (2022) Ireland 67	To determine whether the risk estimates for the Portsmouth scoring system could be used in major head and neck	Patients undergoing resection for a temporal bone malignancy in a single head and neck centre (ENT surgery)	AUC	Complications Morbidity	0.766	1.013		The optimal cut-off for the scoring system was then calculated using the Youden index from the receiver operating

	reconstructive surgery. Retrospective							characteristic curve, which was 40.5 % in this case. Although it was a valid predictor of morbidity overall, we found that the Portsmouth Physiological and Operative Severity Score for the Enumeration of Mortality and Morbidity was most useful in higher-risk patients, with an optimal cut-off score of 40.5 per cent.
Karabulut et al Turkey 2024 300	To investigate whether these two risk calculators (ACS surgical risk calculator and P-POSSUM) accurately reflect actual mortality and morbidity outcomes when retrospectively assessed based on preoperative risk scores, to compare their predictive superiority against each other, and to assess their applicability to major hepatobiliary surgery Retrospective	Patients undergoing major hepatobiliary surgeries between August 2016 and December 2021 Major hepatopancreaticobiliary surgery, pancreaticoduodenectomy, pylorus-preserving pancreaticoduodenectomy, total pancreatectomy, distal pancreatectomy, hepatectomy (right, left, partial, trisegmentectomy), and hepaticojejunostomy surgeries (General surgery)	C-statistic, O/E, Brier	Morbidity	0.672 (0.611-0.732)	1.480	0.257	
Sahiner et al	The sensitivity and specificity of the	Patients who were admitted to the general	AUC	Morbidity	0.734 (0.635-0.833)			The AUC result of e CRP/ALB was found

(2020) Turkey 119	CRP/ALB ratio in predicting postoperative morbidity and mortality in patients undergoing colorectal surgery were investigated by comparing it with P-POSSUM, Cr-POSSUM, and ACPGBICRC scores. Retrospective	surgery outpatient clinic for colorectal malignancy, operated and followed up in the intensive care unit between January 2015 and November 2018 Surgery for colorectal cancer (Right hemicolectomy, Left hemicolectomy, Transverse colectomy, Anterior resection, Low anterior resection, Total colectomy). (General surgery)						to be higher (0.817) than the value of 0.734 calculated for P-POSSUM. P-POSSUM appears to be more valuable than CRP/ALB in predicting morbidity
Sevinyan et al UK 153	To evaluate the accuracy and reliability of the P-POSSUM and ACS-NSQIP surgical risk calculators in predicting postoperative complications in gynaecological–oncological robotic surgery Retrospective	Patients who had undergone Da Vinci-assisted Robotic Surgery for suspected or confirmed gynaecological malignancies (Gynaecology surgery)	AUC, Brier score	Morbidity	0.551	-	0.183	Morbidity was much better predicted by ACS-NSQIP than by P-POSSUM (AUC-0.608 vs. AUC-0.551) with the same result in mortality prediction (Brier score 0.0000). Moreover, a statistically significant overestimation of morbidity has been shown by the P-POSSUM calculator (p = 0.018).

7.3 Information available on request

The protocol for this review, and a clinical summary are available on request.

.

8. ADDITIONAL INFORMATION

8.1 Conflicts of interest

The authors declare they have no conflicts of interest to report.

8.2 Acknowledgements

The Public Health Wales team would like to thank Dr Claire Dunstan, Dr Meredith Graham, Dr Linda Warnock, Meredith Graham, the Clinical Implementation Network, and Praveena Pemmasani for their contributions during stakeholder meetings in guiding the focus of the review and interpretation of findings.

9. APPENDIX

9.1 Appendix 1 Data extraction table of studies that include mortality

Table 18. Data extraction table of studies that include mortality

Authors	Country	No. of Participants/age group	Surgery/Category	Outcomes of interest	Accuracy Scores
P-POSSUM					
Parmar et al. (2010)	UK	344/mean (SD) age 70 (9.9) years.	Consecutive reconstructive vascular surgical procedures/Vascular category.	Major adverse cardiac event (MACE) within 30 days of surgery, defined as 1 or more of the following: myocardial infarction, coronary revascularization, sudden death, and left ventricular failure.	AUC 0.82 (95% CI: 0.73 to 0.91). Good level of accuracy in predicting MACE after vascular surgery.
ACS-NSQIP Surgical Risk Calculator tool					
Baker et al. (2018)	USA	283/ above 18 years old.	An abdominal operation for oncology indications. These included open and laparoscopic operations in the following subcategories: 78 colorectal (27%), 46 pancreatic (16%), 35 gastro esophageal (12%), 30 liver (11%), 29 small bowel (10%), 29 retroperitoneal (10%), and 36 other (13%). General Category.	Composite Grade 1 complications: minor (bedside procedures such as an nasogastric tube insertion) Grade 2 complications: moderate (blood transfusions or need for total parenteral nutrition), Grade 3 complications: severe (interventions without general anesthesia). Grade 4 complications: severe (requiring interventions with general anesthesia). Grade 5 complications: severe (involve organ system failure). Grade 6 complications (post-operative death).	ACSNSQIP's AUC is reported for overall post-operative morbidity. AUC 0.606 Poor level of accuracy in predicting complications across a wide variety of abdominal surgical procedures.
Suresh et al.(2019)	USA	264/ average age 50.8 ± 12.6 years	Panniculectomy/ Plastic and Reconstructive	Postoperative complications were defined as major (return to the operating room within 30 days or death) or minor (surgical site infection, deep vein thrombosis/pulmonary embolism, wound dehiscence, use of postoperative antibiotics, postoperative blood transfusion, prolonged hospital length of stay, or seroma/hematoma formation).	AUC for major complication 0.514 Very poor level of accuracy in predicting major complication after Panniculectomy surgery. AUC for any outcome 0.619.

					Poor level of accuracy in predicting any complication after Panniculectomy surgery.
ASA					
Parmar et al. (2010)	UK	344/mean (SD) age 70 (9.9) years.	Consecutive reconstructive vascular surgical procedures/ Vascular Category.	MACE within 30 days of surgery, defined as 1 or more of the following: myocardial infarction, coronary revascularization, sudden death, and left ventricular failure.	AUC 0.67 (95% CI: 0.56 to 0.78). Poor level of accuracy in predicting MACE after vascular surgery.
Huisman et al.(2014)	UK	263/ aged≥70 years	Laparotomies (Colorectal cancer, gastric cancer and pancreatic cancer) and breast surgery/ General surgery.	Endpoint was morbidity during the first 30 days after surgery. Morbidity was registered using the Clavien-Dindo classification, a scale ranking severity of complications from 'any deviation from the normal post-operative course without the need for pharmacological treatment or surgical, endoscopic and radiological interventions' (grade one) to 'death of a patient' (grade five). Morbidity was dichotomized into minor (Clavien-Dindo grade one and two) and major complications (Clavien-Dindo grade three to five).	AUC is only for major complication. AUC 0.58, (95%CI:0.49 to 0.67). Very poor level of accuracy in predicting post-operative complications among onco-geriatric patients.
RCRI					
Bryce et al. (2012)	UK	106/66 to 77 years old.	Patients undergoing open elective abdominal aortic aneurysm (AAA)/ Vascular surgery.	MACE (major adverse cardiac event) defined as non-fatal myocardial infarction and cardiac death. Major adverse cardiac event included myocardial infarction, cerebrovascular disease, congestive cardiac failure, chronic renal failure, chronic obstructive pulmonary disease, white cell count, prothrombin time.	AUC 0.525 (95% CI: 0.364 to 0.687) Very poor predictive accuracy for MACE in those patients receiving open elective abdominal aortic aneurysm surgery.
Che et al. (2017)	China	1202/aged 69.5±5.3 years	Patients who underwent noncardiac surgery.	Postoperative major cardiac event (PoMCE) within 30 days. PoMCEs defined as cardiac death, nonfatal myocardial infarction, nonfatal cardiac arrest, and heart failure.	AUC 0.53 (95% CI: 0.45 to 0.61). Very poor predictive ability in identifying patients' cardiac risk among Chinese patients with CAD undergoing noncardiac surgery.

Cho et al. (2021)	South Korea	256/68.7±12.1 years	Patients who underwent both a nuclear stress test and noncardiac surgery.	Postoperative major adverse cardiac event (MACE) as a composite of cardiac death, nonfatal myocardial infarction, and pulmonary edema.	AUC 0.612 (95% CI 0.549 to 0.672). Poor predictive accuracy for MACE in those patients receiving non cardiac surgery.
Fronczek et al. (2019)	Poland	870/aged≥45 years old	Patients undergoing in-hospital noncardiac surgery	Major cardiac complications defined as a composite of a nonfatal MI, a nonfatal cardiac arrest, or a cardiac death within 30 days after surgery.	C-statistic 0.60 (95% CI 0.54 to 0.65). Poor predictive ability for major cardiac complications within 30 days of noncardiac surgery.
Palamuthusingam et al.(2024)	Australia and New Zealand	5094/aged≥18 years old.	Patients receiving Chronic Kidney Replacement Therapy (KRT) who underwent elective abdominal surgery, /Urology.	Major adverse cardiovascular event (MACE), defined as nonfatal myocardial infarction, nonfatal stroke, non-fatal cardiac arrest and cardiovascular mortality at 30 days.	AUC 0.67 (95% CI 0.63 to 0.70). Poor predictive ability for MACE at 30 days among the patients on chronic KRT undergoing elective surgery.
Parmar et al.(2010)	UK	344/mean (SD) age 70 (9.9) years.	Consecutive reconstructive vascular surgical procedures/ Vascular surgery category.	MACE within 30 days of surgery, defined as 1 or more of the following: myocardial infarction, coronary revascularization, sudden death, and left ventricular failure.	AUC 0.68 (95% CI 0.57 to 0.83). Poor ability in predicting MACE after vascular surgery.
Payne et al.(2013)	UK	252/ aged≥70 years old.	Patients undergoing major vascular surgery (aortic surgery, infra-inguinal bypass surgery, amputation)/ Vascular surgery category.	Major adverse cardiac event (MACE), defined as non-fatal myocardial infarction and cardiac mortality.	AUC 0.538. Very poor predictive accuracy for MACE.
Roshanov et al.(2021)	14 Countries, name not reported.	35,815/ aged≥45 years old.	Major elective noncardiac surgery.	Major cardiac complication including death up to 30 days after surgery. Primary outcome as a composite of Myocardial injury after noncardiac surgery (MINS), myocardial infarction, nonfatal cardiac arrest, or cardiac death. secondary outcome as a composite of myocardial infarction, nonfatal cardiac arrest, or cardiac death.	Primary Outcome: C-statistic 0.65 (95% CI 0.62 to 0.68). Secondary Outcome: C-statistics 0.69 (95% CI 0.65 to 0.72). Poor ability to predict both primary and secondary outcomes for noncardiac surgery.
Welten et al. (2007)	Netherlands	2642/above 18 years. ≤55 years (396), 56 to 65 years (650), 66 to 75 years (1058)	Patients underwent open non-cardiac vascular surgical procedures/Vascular surgery.	Major adverse cardiac events (MACE) within 30 days after surgery. MACE within 30-days after surgery, was defined as cardiac death (which was defined	AUC 0.65. Poor predictive accuracy for MACE.

		and above 75 years (538).		as any death with a cardiac complication as the primary or secondary cause, including deaths following myocardial infarction, cardiac arrhythmia and heart failure), myocardial infarction or coronary revascularization (PCI or CABG).	
Wotton et al. (2013)	UK	703/median age 68 years (range 14 to 89)	Patients underwent thoracotomy and lung resection/Cardiothoracic surgery.	Post-operative cardiac complication including death within 30 days. Post-operative major cardiac complications included pulmonary oedema, myocardial infarction (MI), ventricular fibrillation (VF) arrest, supraventricular arrhythmia, atrial fibrillation (AF) and mortality occurring within 30 days.	AUC 0.59 (95% CI 0.51 to 0.67). Very Poor discriminative ability for predicting post-operative cardiac complications.
Yao et al. (2021)	Australia	1366/ aged≥45 years old.	Patients undergoing elective non-cardiac surgery.	Major adverse cardiac events (MACE) within 30 days of surgery, including myocardial infarction, pulmonary oedema, complete heart block or cardiac death.	AUC 0.73 (95% CI 0.60 to 0.86). Fair level of accuracy in predicting MACE.
Zou et al. (2024)	China	453/aged over 60 years.	Elderly patients who underwent endovascular aortic aneurysm repair (EVAR)/ Vascular Surgery.	Major adverse cardiac and cerebrovascular events (MACCE). MACCE was defined as a composition of death, myocardial infarction, stroke, chronic cardiac failure and repeat revascularization.	AUC 0.588. Very poor level of accuracy in predicting MACCE.

9.2 Appendix 2. Modified versions of surgical risk prediction tools

Below are references identified through our screening process that satisfied our eligibility criteria, but that used modified versions of included tools.

ARISCAT score (for Postoperative Pulmonary Complications)
ARISCAT plus predicted postoperative forced expiratory volume (ppoFEV1)

Zorrilla-Vaca, A., et al. (2023) Performance Comparison of Pulmonary Risk Scoring Systems in Lung Resection. *Journal of Cardiothoracic and Vascular Anesthesia* - Volume 37, Issue 9, pp.1734 EP – 1743.

ASA Physical Status/ ASA Classification

No modified tools identified meeting out eligibility criteria

Carlisle Risk/Carlisle Calculator

No modified tools identified meeting out eligibility criteria

CFS (Clinical Frailty Scale, also known as Rockwood)

No modified tools identified meeting out eligibility criteria

CPET (Cardiopulmonary exercise testing)

No modified tools identified meeting out eligibility criteria

DASI (Duke Activity Status Index)
Modified DASI

Li, M. H. G., Rosser, M., Blitz, J. (2024). A Retrospective Cohort Study Examining the Validation of the Modified Duke Activity Status Index in the Non-cardiac Surgical Population. Journal of PeriAnesthesia Nursing. S1089-9472(24)00361-7

NELA PRS (National Emergency Laparotomy Audit (Parsimonious Risk Score)

No modified tools identified meeting out eligibility criteria

NRS-2002 (Nutrition Risk Screening 2002)

No modified tools identified meeting out eligibility criteria

NSQIP (National Surgical Quality Improvement Program) universal surgical risk calculator)
Adapted model

Karabulut, A., Umman, V., Oral, G., et al. (2024) Assessing the effectiveness of ACS surgical risk calculator versus P-POSSUM in predicting mortality and morbidity for major hepatobiliary surgery: An observational study. Medicine - Volume 103, Issue 28, pp. 1-6.

P-POSSUM score (Physiological and Operative Severity Score for the enumeration of Mortality and Morbidity)
Modified/ improved POSSUM

Nie, Y., Li, Z., Su, T., et al. (2019) Application of Improved POSSUM Score Combined with Clavien-Dindo Classification in Predicting the Incidence of Severe Complications After Thoracoscopic Lung Surgery. *Indian Journal of Surgery* - Volume 82, Issue 6, pp. 1031-1037.

Adapted model

Karabulut, A., Umman, V., Oral, G., et al. (2024) Assessing the effectiveness of ACS surgical risk calculator versus P-POSSUM in predicting mortality and morbidity for major hepatobiliary surgery: An observational study. Medicine - Volume 103, Issue 28, pp. 1-6.

Cr-POSSUM

Şahiner, Y., Yıldırım, M. B. (2020). Can the c-reactive protein-to-plasma albumin ratio be an alternative scoring to show mortality and morbidity in patients with colorectal cancer? Ulusal Travma ve Acil Cerrahi Dergisi - Volume 26, Issue 4, pp. 580-585.

Prabakaran, V., Thangaraju, T., Mathew, A. C., et al. (2019). *Indian Journal of Surgical Oncology* - Volume 10, Issue 1, pp. 174 EP – 179.

E-POSSUM

Kim, S. Y., Kim, J. H., Chin, H., et al. (2020). Prediction of postoperative mortality and morbidity in octogenarians with gastric cancer - Comparison of P-POSSUM, O-POSSUM, and E-POSSUM: A retrospective single-center cohort study. International Journal of Surgery - Volume 77, Issue 0, pp. 64 EP – 68.

O-POSSUM

He, H., Liu, Y., Liu, X., et al. (2023). Evaluation of different scoring systems in the prediction of complications, morbidity, and mortality after laparoscopic radical gastrectomy. World Journal of Surgical Oncology - Volume 21, Issue 1.

Kim, S. Y., Kim, J. H., Chin, H., et al. (2020). Prediction of postoperative mortality and morbidity in octogenarians with gastric cancer - Comparison of P-POSSUM, O-POSSUM, and E-POSSUM: A retrospective single-center cohort study. International Journal of Surgery - Volume 77, Issue 0, pp. 64 EP – 68.

PD-POSSUM

Zhang, Z. L., Chen, L., Peng, L., et al. (2020). A newly improved POSSUM scoring system for prediction of morbidity in patients with pancreaticoduodenectomy. *Translational Cancer Research* - Volume 9, Issue 9, pp. 5517 EP – 5527.

PONV (Apfel Score for Postoperative Nausea and Vomiting)

No modified tools identified meeting out eligibility criteria

RCRI (Revised Cardiac Risk Index for Pre-Operative Risk)

Modified RCRI

Ackland, G. L., Harris, S., Ziabari, Y., et al. (2020). Revised cardiac risk index and postoperative morbidity after elective orthopaedic surgery: a prospective cohort study. British journal of anaesthesia - Volume 105, Issue 6, pp. 744-52 - published 2010-01-01.

Pandey, A., Sood, A., Sammon, J. D., et al. (2015). Effect of preoperative angina pectoris on cardiac outcomes in patients with previous myocardial infarction undergoing major noncardiac surgery (data from ACS-NSQIP). *American Journal of Cardiology* - Volume 115, Issue 8, pp. 1080-1084.

Schmidt, G., Frieling, N., Schneck, E., et al. (2024). Comparison of preoperative NT-proBNP and simple cardiac risk scores for predicting postoperative morbidity after non-cardiac surgery with intermediate or high surgical risk. *Perioperative Medicine* - Volume 13, Issue 1, pp. 44.

Reconstructed RCRI

Cohn, S. L., Fernandez Ros, N. (2018). Comparison of 4 Cardiac Risk Calculators in Predicting Postoperative Cardiac Complications After Noncardiac Operations. American Journal of Cardiology - Volume 121, Issue 1, pp. 125 EP – 130.

Davis, C., Tait, G., Carroll, J., et al. (2013). The Revised Cardiac Risk Index in the new millennium: a single-centre prospective cohort re-evaluation of the original variables in

9,519 consecutive elective surgical patients. Canadian journal of anaesthesia - Volume 60, Issue 9, pp. 855-63.

ThRCRI

No modified tools identified meeting out eligibility criteria

SORT (Surgical Outcome Risk Tool)

No modified tools identified meeting out eligibility criteria

9.3 Appendix 3. Breakdown of all outcomes reported across studies by tool and surgical specialty.

Below is a breakdown of the different outcomes reported for ACS NSQIP, P-POSSUM, the RCRI and the ASA classification system. Where more than one study was reported for a particular surgical specialty, the findings have also been tabulated below. Please note some of the composite complications may include mortality as the composites were not always clearly defined within the studies.

Table 19 – ACS NSQIP (40 studies) all studies

Studies included: Alzahrani et al (2020), Angelucci et al (2024), Blair et al (2018), Botejue et al (2023), Boyd et al (2020), Campagnaro et al (2023), Choi et al (2018), Chudgar et al (2021), Chudgar et al (2022), Cohn and Ros (2018), Donadon et al (2020), Fruscione et al (2018), Golan et al (2017), Gray et al (2023), Houdek et al (2020), Hsiao et al (2022), Im et al (2024), Karabulut et al (2024), Labbot et al (2021a), Labbot et al (2021b), Lone et al (2019), Ma et al (2019), Manhabusqui et al (2023), Mannas et al (2020), McCarthy et al (2019), Moses et al (2019), Narain et al (2021), Pierce et al (2021), Ravindran et al (2020), Rivard et al (2016), Ryvlin et al (2023), Sevinyan et al (2024), Tierney et al (2019), van der Hulst et al (2022), Vos et al (2020), Wang et al (2017), Wherley et al (2020), Willoughby et al (2023), Yap et al (2018), Yung et al (2022)		**Studies[1]**	**Predictive accuracy score**				
			Discrimination (C-statistic)		**Calibration (O/E ratio)**		**Accuracy (Brier score)**
			Range	**Median**	**Range**	**Median**	**Range**
COMPLICATION CATEGORIES	**Any complication***	31	0.780 to 0.880	0.607	0.838 to 3.412	1.359	0.00000196 to 0.688
	Systematic complications	1	0.716		-		-
	Severe complication	27	0.515 to 0.756	0.607	0.440 to 3.667	1.086	0.00001444 to 0.722
	Minor complication (Clavien–Dindo grade 1–5)	1	0.610		-		0.361
	Cardiac complication	24	0.585 to 0.930	0.728	0.455 to 11.833	0.553	0.00003136 to 0.08

	Renal failure/complications	23	0.480 to 0.958	0.783	0.833 to 20.000	2.834	0.000 to 0.100
	Sepsis	7	0.560 to 0.795	0.6335	0.417 to 1.462	1.000	0.00002704 to 0.306
	Surgical site infection	22	0.427 to 0.845	0.592	0.241 to 10.818	1.328	0.00000576 to 0.349
	Anastomotic leakage	1	0.570		1.850		-
	Superficial skin infection	2	0.732				0.186 to 2.250
	Pneumonia	27	0.490 to 2.301	0.661	0.757 to 5.500	1.333	0.00013924 to 0.966
	Urinary tract infection	23	0.515 to 0.841	0.619	0.473 to 35.211	1.356	0.000 to 0.230
	Venous thromboembolism	24	0.204 to 0.846	0.600	0.336 to 4.805	1.761	0.00005625 to 0.967
		Studies[1]	**Predictive accuracy score**				
			Discrimination (C-statistic)		**Calibration (O/E ratio)**		**Accuracy (Brier score)**
			Range	**Median**	**Range**	**Median**	**Range**
HEALTHCARE UTILISATION AND RECOVERY CATEGORIES	Readmission	24	0.440 to 0.843	0.590	0.042 to 3.130	1.176	0.00008281 to 0.934
	Return to Operating room	25	0.311 to 0.780	0.578	0.659 to 3.485	1.196	0.000 to 0.220
	Length of stay	7	0.530 to 0.590	0.560	0.917 to 3.630	1.122	0.235 to 0.431
	Discharge to a facility other than home	17	0.585 to 0.900	0.700	0.129 to 3.630	0.938	0.003 to 0.084
	Adverse Discharge	1	0.710		1.607		

*includes composites such as ‘all complications’, ‘morbidity’ etc.
- No study provided a score
[1] The number of studies reflects the number of studies reporting any of the predictive accuracy scores
Where a single study reported the outcome no range or median is available
Outcomes in bold denote the composite outcomes
Healthcare utilisation outcomes have been grouped e.g. discharge to a facility other than home could incorporate discharge to rehabilitation centre, or discharge to a nursing home etc.

Table 20 – ACS NSQIP (14 studies) General Surgery

Studies included: Alzahrani et al. (2020), Angelucci et al (2024), Botejue et al (2023), Campagnaro et al (2023), Choi et al (2018), Donadon et al (2020), Fruscione et al (2018), Gray et al (2023), Hsiaso et al (2022), Karabulut et al (2024), Ravindran et al (2020), Rivard et al (2016), Van der Hulst et al (2022), Vos et al (2020)		**Studies[1]**	**Predictive accuracy score**				
			Discrimination (C-statistic)		**Calibration (O/E ratio)**		**Accuracy (Brier score)**
			Range	**Median**	**Range**	**Median**	**Range**
COMPLICATION CATEGORIES	**Any complication***	10	0.530 to 0.725	0.610	1.360 to 2.060	1.552	0.015 to 0.688
	Systematic Complications	1	0.716		-		-
	Severe complication	9	0.554 to 0.756	0.610	0.737 to 2.658	1.659	0.025 to 0.722
	Cardiac complication	8	0.667 to 0.830	0.717	0.455 to 11.833	4.125	0.0004 to 0.040
		Studies[1]	**Predictive accuracy score**				
			Discrimination (C-statistic)		**Calibration (O/E ratio)**		**Accuracy (Brier score)**
			Range	**Median**	**Range**	**Median**	**Range**
HEALTHCARE UTILISATION AND RECOVERY CATEGORIES	Readmission	9	0.440 to 0.767	0.603	0.042 to 3.130	1.366	0.026 to 0.934
	Reoperation	9	0.533 to 0.610	0.582	0.659 to 3.485	2.400	0.000 to 0.431
	Length of stay	4	0.560		0.917 to 1.878	1.106	0.235 to 0.431
	Discharge not to home	5	0.608 to 0.900	0.728	0.181 to 1.442	0.697	0.005 to 0.031

*includes composites such as ‘all complications’, ‘morbidity’ etc.
- No study provided a score
[1] The number of studies reflects the number of studies reporting any of the predictive accuracy scores
Where a single study reported the outcome no range or median is available

Outcomes in bold denote the composite outcomes
Healthcare utilisation outcomes have been grouped e.g. discharge to a facility other than home could incorporate discharge to rehabilitation centre, or discharge to a nursing home etc.

Table 21 – ACS NSQIP (6 studies) Orthopaedic Surgery

Studies included: Labbot et al (2021a), Labbot et al (2021b), Manhabusqui et al (2023), McCarthy et al (2019), Narain et al (2021), Willoughby et al (2023)		**Studies[1]**	**Predictive accuracy score**				
			Discrimination (C-statistic)		**Calibration (O/E ratio)**		**Accuracy (Brier score)**
			Range	**Median**	**Range**	**Median**	**Range**
COMPLICATION CATEGORIES	**Any complication***	6	0.450 to 0.719	0.489	0.854 to 3.412	1.303	0.110 to 0.485
	Severe complication	4	0.549 to 0.707	0.637	0.572 to 3.667	1.333	0.100
	Cardiac complication	1	0.585				

		Studies[1]	**Predictive accuracy scores**			
			Discrimination (C-statistic)		**Calibration (O/E ratio)**	**Accuracy (Brier score)**
			Range	**Median**	**Range**	**Range**
HEALTHCARE UTILISATION AND RECOVERY CATEGORIES	Readmission	3	0.571 to 0.701	0.580	0.084	-
	Reoperation	2	0.530 to 0.578	0.554	-	-
	Extended length of stay	1	0.530		-	-
	Discharge not to home	1	0.673		-	-
	Adverse discharge	1	0.710		1.607	

*includes composites such as ‘all complications’, ‘morbidity’ etc.
- No study provided a score
[1] The number of studies reflects the number of studies reporting any of the predictive accuracy scores

Where a single study reported the outcome no range or median is available
Outcomes in bold denote the composite outcomes
Healthcare utilisation outcomes have been grouped e.g. discharge to a facility other than home could incorporate discharge to rehabilitation centre, or discharge to a nursing home etc.

Table 22 – ACS NSQIP (4 studies) Neurosurgery

Studies included: Houdek et al (2020), Im et al (2024), Ryvlin et al (2023), Wang et al (2017)		**Studies[1]**	**Predictive accuracy score**				
			Discrimination (C-statistic)		**Calibration (O/E ratio)**		**Accuracy (Brier score)**
			Range	**Median**	**Range**	**Median**	**Range**
COMPLICATION CATEGORIES	**Any complication***	3	0.653 to 0.683	0.657	1.753		0.230 to 0.485
	Severe complication	3	0.555 to 0.666	0.595	0.950 to 1.636	1.079	0.241 to 0.230
	Cardiac complication	1	0.648		-		-
		Studies[1]	**Predictive accuracy score**				
			Discrimination (C-statistic)		**Calibration (O/E ratio)**		**Accuracy (Brier score)**
			Range	Median	Range	Median	Range
HEALTHCARE UTILISATION AND RECOVERY CATEGORIES	Readmission	2	0.630 to 0.843	0.736	1.533		0.120
	Reoperation	2	0.311 to 0.600	0.455	2.204		0.110
	Discharge not to home	1	0.590		1.202		0.260
	Length of stay	1	-		0.922		-

*includes composites such as ‘all complications’, ‘morbidity’ etc.
- No study provided a score
[1] The number of studies reflects the number of studies reporting any of the predictive accuracy scores
Where a single study reported the outcome no range or median is available
Outcomes in bold denote the composite outcomes

Healthcare utilisation outcomes have been grouped e.g. discharge to a facility other than home could incorporate discharge to rehabilitation centre, or discharge to a nursing home etc.

Table 23 – ACS NSQIP (4 studies) Urology

Studies included: Blair et al (2018), Golan et al (2017), Lone et al (2019), Mannas et al (2020)		**Studies[1]**	**Predictive accuracy score**				
			Discrimination (C statistic)		**Calibration (O/E ratio)**		**Accuracy (Brier score)**
			Range	**Median**	**Range**	**Median**	**Range**
COMPLICATION CATEGORIES	**Any complication***	4	0.500 to 0.610	0.590	1.108 to 1.833	1.501	0.189 to 0.361
	Severe complication	4	0.530 to 06.00	0.565	0.791to 1.365	0.938	0.121 to 0.240
	Cardiac complication	4	0.590 to 0.800	0.690	0.955 to 2.00	1.381	0.020 to 0.38
		Studies[1]	**Predictive accuracy score**				
			Discrimination (C statistic)		**Calibration (O/E ratio)**		**Accuracy (Brier score)**
			Range	**Median**	**Range**	**Median**	**Range**
HEALTHCARE UTILISATION AND RECOVERY CATEGORIES	Readmission	3	0.520 to 0.600	0.550	0.667 to 1.296	1.191	0.111 to 0.180
	Reoperation	4	0.520 to 0.590	0.580	0.841 to 1.400	1.260	0.034 to 0.67
	Discharge not to home	3	0.690 to 0.750	0.720	0.938 to 3.630	0.990	0.034 to 0.090
	Length of stay	2	0.590		1.148 to 3.630	2.389	0.419

*includes composites such as 'all complications', 'morbidity' etc.
- No study provided a score
[1] The number of studies reflects the number of studies reporting any of the predictive accuracy scores
Where a single study reported the outcome no range or median is available
Outcomes in bold denote the composite outcomes
Healthcare utilisation outcomes have been grouped e.g. discharge to a facility other than home could incorporate discharge to rehabilitation centre, or discharge to a nursing home etc.

Table 24 – ACS NSQIP (3 studies) mixed Surgery

Studies included: Boyd et al (2020), Cohn and Ros (2018), Yap et al (2018)		**Studies[1]**	**Predictive accuracy score**				
			Discrimination (C statistic)		**Calibration (O/E ratio)**		**Accuracy (Brier score)**
			Range	**Median**	**Range**	**Median**	**Range**
COMPLICATION CATEGORIES	**Any complication***	2	0.547 to 0.880	0.713	-		0.160
	Severe complication	1	0.529		-		-
	Cardiac complication	3	0.650 to 0.930	0.890	-		0.080
		Studies[1]	**Predictive accuracy score**				
			Discrimination (C statistic)		**Calibration (O/E ratio)**		**Accuracy (Brier score)**
			Range	**Median**	**Range**	**Median**	**Range**
HEALTHCARE UTILISATION AND RECOVERY CATEGORIES	Readmission	1	0.512				
	Reoperation	2	0.519 to 0.780	0.649	-		0.220
	Discharge not to home	1	0.848		-		-

*includes composites such as 'all complications', 'morbidity' etc.
- No study provided a score
[1] The number of studies reflects the number of studies reporting any of the predictive accuracy scores
Where a single study reported the outcome no range or median is available
Outcomes in bold denote the composite outcomes
Healthcare utilisation outcomes have been grouped e.g. discharge to a facility other than home could incorporate discharge to rehabilitation centre, or discharge to a nursing home etc.

Table 25 – ACS NSQIP (2 studies) Plastic

Studies included: Tierney et al (2019), Yung et al (2022)		**Studies[1]**	**Predictive accuracy score**					
			Discrimination (C statistic)		**Calibration (O/E ratio)**		**Accuracy (Brier score)**	
			Range	**Median**	**Range**	**Median**	**Range**	
COMPLICATION CATEGORIES	**Any complication***	1	0.699		1.027		0.087	
	Severe complication	1	0.539		0.901		0.273	
	Cardiac complication	1	0.750		2.571		0.026	
		Studies[1]	**Predictive accuracy score**					
			Discrimination (C statistic)		**Calibration (O/E ratio)**		**Accuracy (Brier score)**	
			Range	**Median**	**Range**	**Median**	**Range**	
HEALTHCARE UTILISATION AND RECOVERY CATEGORIES	Readmission	1	0.591		0.698		0.048	
	Reoperation	2	0.493 to 0.500	0.496	0.801 to 1.187	0.994	0.128 to 0.129	0.128
	Discharge not to home	2	0.585 to 0.700	0.643	0.705 to 1.841	1.273	0.084 to 0.187	0.135

*includes composites such as 'all complications', 'morbidity' etc.
- No study provided a score
[1] The number of studies reflects the number of studies reporting any of the predictive accuracy scores
Where a single study reported the outcome no range or median is available
Outcomes in bold denote the composite outcomes
Healthcare utilisation outcomes have been grouped e.g. discharge to a facility other than home could incorporate discharge to rehabilitation centre, or discharge to a nursing home etc.

Table 26 – ACS NSQIP (2 studies) Vascular

Studies included: Ma et al (2018), Moses et al (2019)	**Studies[1]**	**Predictive accuracy score**		
		Discrimination (C statistic)	**Calibration (O/E ratio)**	**Accuracy (Brier score)**

			Range	Median	Range	Median	Range
COMPLICATION CATEGORIES	**Any complication***	1	0.599		1.552		-
	Severe complication	1	0.588		1.618		-
	Cardiac complication	2	0.601		1.605 to 1.795	1.700	-
		Studies[1]	**Predictive accuracy score**				
			Discrimination (C statistic)		**Calibration (O/E ratio)**		**Accuracy (Brier score)**
			Range	**Median**	**Range**	**Median**	**Range**
HEALTHCARE UTILISATION AND RECOVERY CATEGORIES	Reoperation	1	0.535		1.884		-
	Discharge not to home	1	0.693		1.058		-

*includes composites such as ‘all complications’, ‘morbidity’ etc.
- No study provided a score
[1] The number of studies reflects the number of studies reporting any of the predictive accuracy scores
Where a single study reported the outcome no range or median is available
Outcomes in bold denote the composite outcomes
Healthcare utilisation outcomes have been grouped e.g. discharge to a facility other than home could incorporate discharge to rehabilitation centre, or discharge to a nursing home etc.

P-POSSUM

Table 27 – P-POSSUM (7 studies) all studies

Studies included: Angelucci et al (2024); Bodea et al (2018); He et al (2023); Jones et al (2022); Karabulut et al (2024); Sahiner et al (2020); Sevinyan et al (2024)		**Studies[1]**	**Predictive accuracy scores**				
			Discrimination (C statistic)		**Calibration (O/E ratio)**		**Accuracy (Brier score)**
			Range	**Median**	**Range**	**Median**	**Range**
COMPLICATION CATEGORIES	**Any complication***	7	0.551 to 0.781	0.567	1.48 to 1.66	1.57	0.186 to 0.257
	Severe complication	1	0.63		-		0.205
	Pancreaticoduodenectomy specific complications	1	0.64		0.89		-

*includes composites such as 'all complications', 'morbidity' etc.

- No study provided a score

[1] The number of studies reflects the number of studies reporting any of the predictive accuracy scores
Where a single study reported the outcome no range or median is available

Outcomes in bold denote the composite outcomes

Table 28 – P-POSSUM (5 studies) General Surgery

Studies included: Angelucci et al (2024); Bodea et al (2018); He et al (2023); Karabulut et al (2024); Sahiner et al (2020)		**Studies[1]**	**Predictive accuracy score**				
			Discrimination (C statistic)		**Calibration (O/E ratio)**		**Accuracy (Brier score)**
			Range	**Median**	**Range**	**Median**	**Range**
COMPLICATION CATEGORIES	**Any complication***	5	0.610 to 0.781	0.672	1.660 to 1.48	1.57	0.229 to 0.257
	Severe complication	1	0.63		-		0.205
	Pancreaticoduodenectomy specific complications	1	0.64		0.89		-

*includes composites such as 'all complications', 'morbidity' etc.

- No study provided a score

[1] The number of studies reflects the number of studies reporting any of the predictive accuracy scores
Where a single study reported the outcome no range or median is available

Outcomes in bold denote the composite outcomes

RCRI

Table 29 – RCRI (16 studies) all studies							
Studies included: Alphonsus et al (2021), Alrezk et al (2017), Andersson et al (2015), Archan et al (2010), Bronheim (2018), Bronheim (2019), Carabini et al (2014), Chrisant et al (2024), Cohn and Ros (2018), Davis et al (2013), Dunn et al (2019), Moses (2019), Pandey et al (2015), Peterson et al (2016), Schmidt et al (2024), Yap et al (2018)		**Studies[1]**	**Predictive accuracy score**				
			Discrimination (C statistic)		**Calibration (O/E ratio)**		**Accuracy (Brier score)**
			Range	Median	Range	Median	Range
COMPLICATION CATEGORIES	**Any complication***	2	0.560 to 0.850	0.623	-		-
	Pulmonary complication	1	0.760		-		-
	Cardiac complication	11	0.550 to 0.930	0.730	2.23		-
	Cardiac risk	1	0.680				
	Unplanned intubation	1	0.837		-		-
	Pulmonary embolism	1	0.409		-		-
	Ventilated >48 hours	1	0.845		-		-
	Acute renal failure	2	0.687 to 0.881	0.784	-		-
	Cerebrovascular accident/ stroke with neurologic deficit	1	0.747		-		-
	Coma >24 hours	1	0.900		-		-
	Sepsis	1	0.826		-		-
	Septic shock	1	0.845		-		-
	Superficial Surgical site infection (SSI)	1	0.717		-		-
	Deep incisional SSI	1	0.879		-		-
	Organ space SSI	1	0.876		-		-
	Wound dehiscence	1	0.717		-		-
	Pneumonia	1	0.739		-		-

	Progressive renal insufficiency	1	0.845		-	-
	Urinary tract infection	1	0.738		-	-
	Peripheral nerve injury	1	0.073		-	-
	Bleeding transfusions	1	0.712		-	-
	Deep vein thrombosis/ thrombophlebitis	1	0.707		-	-
	Cardiac arrest requiring CPR	1	0.855		-	-
	Myocardial Infarction/injury	1	0.876		-	-
	Acute decompensated heart failure	1	0.590			
	Infectious adverse event	1	0.573		-	-
		Studies[1]	**Predictive accuracy score**			
			Discrimination (C statistic)		**Calibration (O/E ratio)**		**Accuracy (Brier score)**
			Range	**Median**	**Range**	**Median**	**Range**
HEALTHCARE UTILISATION AND RECOVERY CATEGORIES	Readmission	2	0.457 to 0.835	0.646	-		-
	Reoperation	1	0.850				

*includes composites such as ‘all complications’, ‘morbidity’ etc.
- No study provided a score
[1] The number of studies reflects the number of studies reporting any of the predictive accuracy scores
Where a single study reported the outcome no range or median is available
Outcomes in bold denote the composite outcomes

Table 30 – RCRI (9 studies) Mixed Surgery

Studies included: Alphonsus et al. (2021), Alrezk et al. (2017), Andersson et al. (2015), Christant et al (2024), Cohn and Ros (2018), Davis et al. (2013), Pandey et al. (2015), Schmidt et al. (2024), Yap et al. (2018)		**Studies[1]**	**Predictive accuracy score**				
			Discrimination (C statistic)		**Calibration (O/E ratio)**		**Accuracy (Brier score)**
			Range	**Median**	**Range**	**Median**	**Range**
COMPLICATION CATEGORIES	**Any complication***	1	0.560		-		-
	Cardiac risk	1	0.680		-		-
	Pulmonary complication	1	0.710		-		-
	Cardiac complication	7	**0.550 to 0.930**	**0.761**	-		-
		Studies[1]	**Predictive accuracy score**				
			Discrimination (C statistic)		**Calibration (O/E ratio)**		**Accuracy (Brier score)**
			Range	**Median**	**Range**	**Median**	**Range**
HEALTHCARE UTILISATION AND RECOVERY CATEGORIES	Readmission	1	0.457		-		-

*includes composites such as 'all complications', 'morbidity' etc.
- No study provided a score
[1] The number of studies reflects the number of studies reporting any of the predictive accuracy scores
Where a single study reported the outcome no range or median is available
Outcomes in bold denote the composite outcomes

Table 31 – RCRI (4 studies) Orthopaedic Surgery

Studies included: Bronheim et al. (2018), Bronheim et al. (2019), Carabini et al. (2014), Peterson et al. (2016)	**Studies[1]**	**Predictive accuracy score**				
		Discrimination (C statistic)		**Calibration (O/E ratio)**		**Accuracy (Brier score)**
		Range	**Median**	**Range**	**Median**	**Range**

COMPLICATION CATEGORIES	**Any complication***	1	0.623		-		-
	Cardiac complication	2	**0.540 to 0.900**	**0.72**	-		-
		Studies[1]	**Predictive accuracy score**				
			Discrimination (C statistic)		**Calibration (O/E ratio)**		**Accuracy (Brier score)**
			Range	**Median**	**Range**	**Median**	**Range**
HEALTHCARE UTILISATION AND RECOVERY CATEGORIES	Readmission	1	0.835		-		-

*includes composites such as 'all complications', 'morbidity' etc.
- No study provided a score
[1] The number of studies reflects the number of studies reporting any of the predictive accuracy scores
Where a single study reported the outcome no range or median is available
Outcomes in bold denote the composite outcomes

Table 32 – RCRI (2 studies) Vascular surgery

Studies included: Archan et al. (2010), Moses et al. (2019)		**Studies[1]**	**Predictive accuracy score**				
			Discrimination (C statistic)		**Calibration (O/E ratio)**		**Accuracy (Brier score)**
			Range	**Median**	**Range**	**Median**	**Range**
COMPLICATION CATEGORIES	**Cardiac complication**	2	0.730		2.231		-

*includes composites such as 'all complications', 'morbidity' etc.
- No study provided a score
[1] The number of studies reflects the number of studies reporting any of the predictive accuracy scores
Where a single study reported the outcome no range or median is available
Outcomes in bold denote the composite outcomes

ASA classification system

Table 33 – ASA Classification system (13 studies) All studies							
Studies included: Aceto et al (2021), Bronheim et al (2018), Bronheim et al. (2019), Chrisant et al (2024), Hightower et al (2010), Feghali et al (2022), Fu et al (2018), Lim et al (2017), McConaghy et al (2023), Meng et al (2018), Ondeck et al (2018), Sankar et al (2014), Wolters et al (2006)		**Studies[1]**	**Predictive accuracy score**				
			Discrimination (C statistic)		**Calibration (O/E ratio)**		**Accuracy (Brier score)**
			Range	**Median**	**Range**	**Median**	**Range**
COMPLICATION CATEGORIES	**Any complication***	5	0.510 to 0.706	0.567	-		-
	Catastrophic outcome	1	0.793		-		-
	Severe complication	4	0.519 to 0.615	0.589	-		-
	Minor complication	3	0.514 to 0.608	0.589	-		-
	Pulmonary complication	2	0.69 to 0.76	0.725	-		-
	Cardiac complication	1	0.75		-		-
	Airway complication	1	0.719		-		-
	Unplanned intubation	2	0.678 to 0.741	0.710	-		-
	Pulmonary embolism	1	0.812		-		-
	Ventilated >48 hours	2	0.742 to 0.825	0.784	-		-
	Acute renal failure	1	0.789		-		-
	Cerebrovascular accident/ stroke with neurologic deficit	1	0.838		-		-
	Coma >24 hours	1	0.653		-		-
	Sepsis	1	0.905		-		-
	Septic shock	1	0.759		-		-
	Superficial Surgical site infection (SSI)	1	0.842		-		-
	Deep incisional SSI	1	0.950		-		-
	Organ space SSI	1	0.776		-		-

	Wound dehiscence	1	0.792		-		-
	Pneumonia	2	0.700 to 0.824	0.762	-		-
	Progressive renal insufficiency	1	0.814		-		-
	Urinary tract infection	1	0.825		-		-
	Peripheral nerve injury	1	0.513		-		-
	Bleeding transfusions	1	0.800		-		-
	Deep vein thrombosis/ thrombophlebitis	1	0.782		-		-
	Cardiac arrest requiring CPR	1	0.674		-		-
	Myocardial Infarction/injury	2	0.700 to 0.799	0.750	-		-
	Infectious adverse event	2	0.517 to 0.580	0.549	-		-
		Studies[1]	**Predictive accuracy score**				
			Discrimination (C statistic)		**Calibration (O/E ratio)**		**Accuracy (Brier score)**
			Range	**Median**	**Range**	**Median**	**Range**
HEALTHCARE UTILISATION AND RECOVERY CATEGORIES	Readmission	1	0.906		-		-
	Reoperation	1	0.866				
	Extended length of stay	3	0.561 to 0.630	0.572	-		-
	Length of stay	1	0.536		-		-
	Discharge not to home	4	0.590 to 0.640	0.631	-		-

*includes composites such as ‘all complications’, ‘morbidity’ etc.
- No study provided a score
[1] The number of studies reflects the number of studies reporting any of the predictive accuracy scores
Where a single study reported the outcome no range or median is available
Outcomes in bold denote the composite outcomes
Healthcare utilisation outcomes have been grouped e.g. discharge to a facility other than home could incorporate discharge to rehabilitation centre, or discharge to a nursing home etc.

Table 34 – ASA Classification system (6 studies) Orthopaedic Surgery

Studies included: Bromnheim et al (2018), Bronheim et al (2019), Fu et al (2017), Lim et al (2017), McConaghy et al (2023), Ondeck et al (2018)		**Studies[1]**	**Predictive accuracy score**				
			Discrimination (C statistic)		**Calibration (O/E ratio)**		**Accuracy (Brier score)**
			Range	**Median**	**Range**	**Median**	**Range**
COMPLICATION CATEGORIES	**Any complication***	3	0.567 to 0.706	0.607	-		-
	Catastrophic outcome	1	0.793		-		-
	Severe complication	3	0.571 to 0.615	0.5975	-		-
	Airway complication	1	0.719				
	Minor complication	1	0.608		-		-
	Pulmonary complication	1	0.760		-		-
	Cardiac complication	1	0.750		-		-
		Studies[1]	**Predictive accuracy score**				
			Discrimination (C statistic)		**Calibration (O/E ratio)**		**Accuracy (Brier score)**
			Range	**Median**	**Range**	**Median**	**Range**
HEALTHCARE UTILISATION AND RECOVERY CATEGORIES	Readmission	1	0.906		-		-
	Reoperation	1	0.866				
	Extended length of stay	3	0.561 to 0.630	0.572	-		-
	Length of stay	1	0.536		-		-
	Discharge not to home	3	0.590 to 0.640	0.631	-		

*includes composites such as ‘all complications’, ‘morbidity’ etc.
- No study provided a score
[1] The number of studies reflects the number of studies reporting any of the predictive accuracy scores
Where a single study reported the outcome no range or median is available
Outcomes in bold denote the composite outcomes
Healthcare utilisation outcomes have been grouped e.g. discharge to a facility other than home could incorporate discharge to rehabilitation centre, or discharge to a nursing home etc.

Table 35 – ASA Classification system (3 studies) mixed surgery

Studies included: Aceto et al (2021), Chrisant et al (2024), Sanker et al (2014)		Studies[1]	Predictive accuracy score			
			Discrimination (C statistic)		Calibration (O/E ratio)	Accuracy (Brier score)
			Range	Median		
COMPLICATION CATEGORIES	**Pulmonary complication**	2	0.690 to 0.760		-	-
	Cardiac complication	1	0.750		-	-

*includes composites such as 'all complications', 'morbidity' etc.
- No study provided a score
[1] The number of studies reflects the number of studies reporting any of the predictive accuracy scores
Where a single study reported the outcome no range or median is available
Outcomes in bold denote the composite outcomes

Table 36 – ASA Classification system (2 studies) Vascular Surgery

Studies included: Wolters et al (2006), Feghali et al (2022)		Studies[1]	Predictive accuracy score			
			Discrimination (C statistic)		Calibration (O/E ratio)	Accuracy (Brier score)
			Range	Median		
COMPLICATION CATEGORIES	**Any complication***	1	0.518		-	-
	Severe complication	1	Range 0.541 to 0.606 Median 0.5735		-	-

*includes composites such as 'all complications', 'morbidity' etc.
- No study provided a score
[1] The number of studies reflects the number of studies reporting any of the predictive accuracy scores
Where a single study reported the outcome no range or median is available
Outcomes in bold denote the composite outcomes

9.4 Appendix 4. Medline search strategies

1. ((NSQIP or "National Surgical Quality Improvement Program") adj4 ("Surgical Risk Calculator" or "Universal Surgical Risk Calculator")).ti,ab. 166
2. ((POSSUM and ((preop* or "pre-op*" or "postop*" or "post-op*" or risk or predict* or tool* or scor*) and (surg* or repair* or operat*))) or "P-POSSUM" or "Portsmouth-POSSUM" or "Physiological and Operative Severity Score").ti,ab. 680
3. ("Surgical Outcome Risk Tool" or (SORT adj3 ((preop* or "pre-op*" or risk or predict* or tool* or scor*) and surg*))).ti,ab. 137
4. or/1-3 972
5. (case reports or editorial or guideline or letter or meta analysis or patient education handout or practice guideline or "review" or "systematic review" or comment).pt. 7995825
6. 4 not 5 874
7. limit 6 to (english language and humans) 686
8. exp animals/ not humans.sh. 5288190
9. 7 not 8 686
10. limit 9 to yr="2019 -Current" 198

1. (ARISCAT or ("Postoperative Pulmonary Complications" adj2 scor*)).ti,ab. 69
2. ("American Society of Anesthesiologists Classification System" or "ASA Classification System" or "American Society of Anesthesiologists Physical Status Classification System" or "ASA Physical Status Classification System").ti,ab. 232
3. "Carlisle adj2 Calculator".ti,ab. or (Carlisle.au,ax. and calculator.ti,ab.) 1
4. (("Clinical Frailty Scale" or CFS or "Clinical Frailty Scor*" or (Rockwood adj2 scor*) or (Rockwood adj2 scale*) or (Rockwood adj1 frailty)) and (preoperat* or "pre-operat*" or ((before or prior or advance or pre or prepar*) adj3 (surg* or operat* or anesthes* or anaesthes* or sedat*)))).ti,ab. 239
5. (("Cardiopulmonary Exercise Test*" or CPET) and (preoperat* or "pre-operat*" or ((before or prior or advance or pre or prepar*) adj3 (surg* or operat* or anesthes* or anaesthes* or sedat*)))).ti,ab. 459
6. (DASI or "Duke Activity Status Index").ti,ab. 406
7. ("NELA PRS" or "NELA risk scor*" or "NELA Parsimonious Risk Scor*" or (Parsimonious adj2 "Risk Scor*") or "National Emergency Laparotomy Audit Parsimonious Risk Score").ti,ab. 13
8. ("Nutrition Risk Screening" or ("NRS-2002" and (preoperat* or "pre-operat*" or ((before or prior or advance or pre or prepar*) adj3 (surg* or operat* or anesthes* or anaesthes* or sedat*))))).ti,ab. 364
9. ("Apfel Score for Postoperative Nausea and Vomiting" or (("Postoperative Nausea and Vomiting" or PONV or Apfel) adj2 (scor* or tool* or scale*))).ti,ab. 361
10. (("Revised Cardiac" adj2 Index) or RCRI).ti,ab. 402
11. or/1-10 2519
12. ("Children's Fear Scale" or "Chronic Fatigue Syndrome" or "Corneal Fluorescein Staining" or "Diagnostic Autism Spectrum Interview*" or "Numeric Rating Scale*" or "Palin PRS" or "Personal Response System*" or "Pervasive Refusal Syndrome" or "Pierre Robin Sequence" or "Polygenic Risk Score*" or "R-CODOX-M/R-IVAC" or collision* or "rheumatoid factor cross reactive idiotype").ti,ab. 69494
13. 11 not 12 2460
14. (case reports or editorial or guideline or letter or meta analysis or patient education handout or practice guideline or "review" or "systematic review" or comment).pt. 7995825
15. 13 not 14 2188
16. limit 15 to (english language and humans) 1639

Health and Care
Research Wales
Evidence Centre
Canolfan Dystiolaeth
Ymchwil Iechyd a
Gofal Cymru

The Health and Care Research Wales Evidence Centre

Our dedicated team works together with Welsh Government, the NHS, social care, research institutions and the public to deliver vital research to tackle health and social care challenges facing Wales.

Funded by Welsh Government, through Health and Care Research Wales, the Evidence Centre answers key questions to improve health and social care policy and provision across Wales.

Along with our collaborating partners, we conduct reviews of existing evidence and new research, to inform policy and practice needs, with a focus on ensuring real-world impact and public benefit that reaches everyone.

Director: Professor Adrian Edwards

Associate Directors: Dr Alison Cooper, Dr Natalie Joseph-Williams, Dr Ruth Lewis

@EvidenceWales @tystiolaethcym

healthandcareevidence@cardiff.ac.uk

www.researchwalesevidencecentre.co.uk