Prediction of postoperative ICU requirements: closing the translational gap with a real-world clinical benchmark for artificial intelligence approaches

Alexander Althammer, Felix Berger, Oliver Spring, Philipp Simon, Felix Girrbach, Maximilian Dieing, Jens O. Brunner, Sergey Shmygalev, Christina C. Bartenschlager, Axel R. Heller

Angaben zur Veröffentlichung / Publication details:

Althammer, Alexander, Felix Berger, Oliver Spring, Philipp Simon, Felix Girrbach, Maximilian Dieing, Jens O. Brunner, Sergey Shmygalev, Christina C. Bartenschlager, and Axel R. Heller. 2025. "Prediction of postoperative ICU requirements: closing the translational gap with a real-world clinical benchmark for artificial intelligence approaches." *Information* 16 (10): 888. https://doi.org/10.3390/info16100888.

information

MDPI

Article

Prediction of Postoperative ICU Requirements: Closing the Translational Gap with a Real-World Clinical Benchmark for Artificial Intelligence Approaches

Alexander Althammer [1,*], Felix Berger [1], Oliver Spring [1], Philipp Simon [1], Felix Girrbach [1], Maximilian Dieing [2], Jens O. Brunner [2,3,4], Sergey Shmygalev [1], Christina C. Bartenschlager [1,5,†] and Axel R. Heller [1,†]

[1] Anaesthesiology and Operative Intensive Care, Faculty of Medicine, University of Augsburg, 86156 Augsburg, Germany; felix.berger@uk-augsburg.de (F.B.); oliver.spring@uk-augsburg.de (O.S.); philipp.simon3@uk-augsburg.de (P.S.); felix.girrbach@uk-augsburg.de (F.G.); sergey.shmygalev@uk-augsburg.de (S.S.); christina.bartenschlager@th-nuernberg.de (C.C.B.); axel.heller@uk-augsburg.de (A.R.H.)
[2] Faculty of Business and Economics, University of Augsburg, Universitätsstraße 2, 86159 Augsburg, Germany; maximilian.dieing@wiwi.uni-augsburg.de (M.D.); jotbr@dtu.dk (J.O.B.)
[3] Department of Technology, Management, and Economics, Technical University of Denmark, Anker Engelunds Vej 1, Bygning 101A, 2800 Kongens Lyngby, Denmark
[4] Center for Excellence in Healthcare Operations Planning, Next Generation Technology, Region Zealand, Ærtekildevej 1, 4100 Ringsted, Denmark
[5] Applied Data Science in Health Care, Nürnberg School of Health, Ohm University of Applied Sciences Nuremberg, 90489 Nuremberg, Germany
* Correspondence: alexander.althammer@uk-augsburg.de
† These authors contributed equally to this work.

Abstract

Background: Accurate prediction of postoperative care requirements is critical for patient safety and resource allocation. Although numerous approaches involving artificial intelligence (AI) and machine learning (ML) have been proposed to support such predictions, their implementation in practice has so far been insufficiently successful. One reason for this is that the performance of the algorithms is difficult to assess in practical use, as the accuracy of clinical decisions has not yet been systematically quantified. As a result, models are often assessed purely from a technical perspective, neglecting the socio-technical context. Methods: We conducted a retrospective, single-center observational study at the University Hospital Augsburg, including 35,488 elective surgical cases documented between August 2023 and January 2025. For each case, preoperative care-level predictions by surgical and anesthesiology teams were compared with the actual postoperative care provided. Predictive performance was evaluated using accuracy and sensitivity. Since this is a highly imbalanced dataset, in addition to sensitivity and specificity, the balanced accuracy and the F_β-score were also calculated. The results were contrasted with published Machine-Learning (ML)-based approaches. Results: Overall prediction accuracy was high (surgery: 91.2%; anesthesiology: 87.1%). However, sensitivity for identifying patients requiring postoperative intensive care was markedly lower than reported for ML models in the literature, with the largest discrepancies observed in patients ultimately admitted to the ICU (surgery: 38.05%; anesthesiology: 56.84%; ML: 70%). Nevertheless, clinical judgment demonstrated a superior F_1-score, indicating a more balanced performance between sensitivity and precision (surgery: 0.527; anesthesiology: 0.551; ML: 0.28). Conclusions: This study provides the first real-world benchmark of clinical expertise in postoperative care prediction and shows a way in which modern ML approaches must be evaluated in a specific sociotechnical context. By quantifying the predictive performance of surgeons and

Academic Editors: Yutong Xie and Mohamed Bennasar

Received: 31 July 2025
Revised: 8 October 2025
Accepted: 9 October 2025
Published: 13 October 2025

Citation: Althammer, A.; Berger, F.; Spring, O.; Simon, P.; Girrbach, F.; Dieing, M.; Brunner, J.O.; Shmygalev, S.; Bartenschlager, C.C.; Heller, A.R. Prediction of Postoperative ICU Requirements: Closing the Translational Gap with a Real-World Clinical Benchmark for Artificial Intelligence Approaches. *Information* **2025**, *16*, 888. https://doi.org/10.3390/info16100888

anesthesiologists, it enables an evaluation of existing ML approaches. Thus the strength of our work is the provision of a real-world benchmark against which all ML methods for preoperative prediction of ICU demand can be systematically evaluated. This enables, for the first time, a comparison of different approaches on a common, practice-oriented basis and thus significantly facilitates translation into clinical practice, thereby closing the translational gap. Furthermore it offers a data-driven framework to support the integration of ML into preoperative decision-making.

Keywords: elective surgery; postoperative care requirements; prediction; artificial intelligence; machine learning; translation gap

1. Introduction

The accurate prediction of postoperative care requirements following elective surgical procedures plays a central role, both from a medical and an economical perspective [1]. For instance, if a patient is mistakenly assumed to require only PACU (Post-Anesthesia Care Unit) followed by timely transfer to a general ward, this can lead to last-minute scheduling changes with significant consequences for operating room and intensive care resources [2,3], ultimately compromising patient safety [4–6]. On the other hand, planning for intensive care resources that are not actually needed can also have negative medical and economic consequences [7], for example delay of surgeries that by local protocol require postoperative transfer to an ICU. Numerous risk factors and scoring systems have already been described for general [8–18] and specific postoperative complications [19], such as respiratory complications [20] or acute kidney injury [21], which may necessitate postoperative admission to a unit with advanced monitoring and therapeutic capabilities. A predictive model titled "SURPAS" [22] identified several risk factors for the necessity of advanced postoperative monitoring such as the ASA classification, preoperative functional status, and surgical specialty. The machine learning (ML)-based model for predicting ICU admission presented by Chiew et al. demonstrated a specificity of 98%, a sensitivity of 50%, and an AUROC of 0.96. Despite the existence of such models in the literature, there are currently no clearly defined recommendations for clinical practice. No standardized guidelines currently exist that call for routine ICU admission based on preoperative variables or planned surgical procedures—neither from surgical, anesthesiologic, nor intensive care perspectives [23,24]. In practice, the preoperative assessment of the required level of care—whether ICU, Intermediate Care (IMC) [25], prolonged PACU, or PACU—is still primarily based on internal hospital protocols and the individual judgment of physicians. Despite promising results on technical benchmarks, the clinical adoption of AI in medicine has been limited. As Sokol et al. recently emphasized, focusing on 'superhuman' performance in artificial settings does not necessarily translate into meaningful clinical impact. Instead, AI must be evaluated within a sociotechnical framework. In this framework, its value is measured by its ability to support real-world clinical reasoning and decision-making. While several studies have compared clinician performance with machine learning (ML) models in other fields—such as dermatology [26,27], radiology [28] and pathology [29]—a systematic, practice-oriented quantification of real-world clinical decision-making in the perioperative setting has not yet been established and, to our knowledge, no such benchmark currently exists. In the perioperative domain, MySurgeryRisk [30] demonstrated that ML-based prediction of postoperative complications could complement physician judgment, and Chiew et al. proposed an ML model for ICU prediction that explicitly called for direct comparison with clinical decision-making [4]. However, these approaches did not establish a structured benchmark

that translates subjective physician judgment into reproducible, imbalance-aware performance metrics. This Problem is described by Sokol et al. as the "translation gap" [4,31]. Our study addresses this gap by introducing a practice-oriented reference framework against which future AI- and ML-based approaches can be systematically evaluated. The absence of widespread adoption of promising data-driven and ML-based approaches can also be attributed to the lack of direct performance comparisons with subjective clinical decision-making [4,31–35]. The aim of this study is to evaluate a real-world dataset to generate robust insights into the performance of subjective physician decision-making, thereby enabling practical recommendations for the integration of modern ML approaches.

This article addresses the question: Can the real-world performance of clinicians serve as a pragmatic benchmark for evaluating AI- and ML-based models in predicting postoperative ICU requirements, thereby representing a crucial step towards their future integration into clinical practice?

The data-driven concept in this work aims to improve the prediction of postoperative care requirements and thus optimize the management of elective surgical procedures.

2. Materials and Methods

This retrospective analysis was conducted using an anonymized dataset from the University Hospital Augsburg and was previously reviewed by the Ethics Committee of Ludwig Maximilian University of Munich (Nr 25-0377-KB).

To classify the postoperative care requirements of elective surgery patients, four levels were defined:

- A patient assigned to Level 0 requires standard postoperative care in the recovery room followed by timely transfer to a general ward (=standard care) when transfer criteria are met. This corresponds to the international standard of a PACU. This level was set as the system default.
- Level 1 patients are expected to stay in the PACU for an extended period of at least 4 h, e.g., including overnight monitoring. Level 2 patients are monitored at the IMC during the postoperative phase.
- For Level 3 patients, postoperative care in an ICU is assumed to be necessary.

2.1. Clinical Practice at the University Hospital Augsburg

In daily clinical practice at University Hospital Augsburg, the initial prediction of the required postoperative level of care was made by the responsible surgical team during the scheduling of the operation. This was later reassessed during the anesthesiologic consultation and preoperative evaluation by anesthesiology physicians (see Figure 1). Typically, the anesthesiologist was aware of the surgical team's initial level-of-care prediction prior to making their own assessment. In selected cases, an additional interdisciplinary case conference is conducted shortly before the start of surgery in the morning to determine the appropriate prognosis for the respective patient.

During the observation period, 46,830 patient cases were recorded. Of these, 11,342 emergency cases were excluded from the study, resulting in 35,488 elective procedures. The performance of the decision-making process was retrospectively evaluated based on this cohort of 35,488 interdisciplinary elective surgical procedures conducted between 1 August 2023, and 31 January 2025. For each case, both the surgical and anesthesiologic preoperative predictions, as well as the actual postoperative level of care provided, were taken into account.

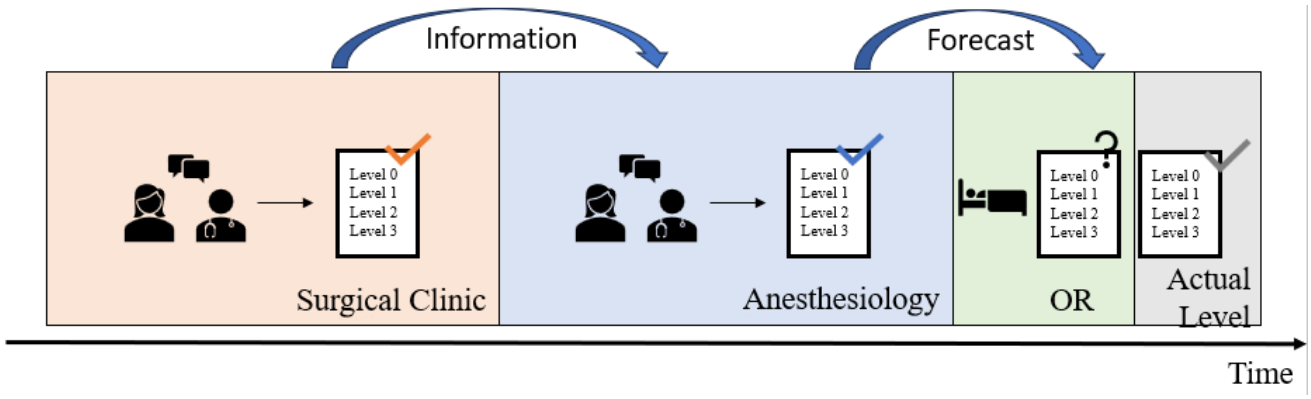

Figure 1. Current process for elective procedures and level designation: OR: Operating room; Level 0: Recovery room stay followed by prompt transfer to the general ward (PACU); Level 1: Recovery room stay followed by delayed transfer to the general ward (prolonged PACU); Level 2: Postoperative transfer to the IMC; Level 3: Postoperative transfer to the ICU.

2.2. Statistics

To assess the predictive accuracy of the physicians' classifications, both disciplines—surgery and anesthesiology—were analyzed individually and comparatively. Evaluation metrics included accuracy, sensitivity, specificity, balanced accuracy, and precision. To evaluate the quality of the level-of-care predictions, several performance metrics were employed. Accuracy represents the overall proportion of correctly classified patients across all categories. Sensitivity measures the proportion of true positives correctly identified for each care level, whereas specificity reflects the model's ability to correctly exclude patients who did not require a particular level of care. Given the imbalanced distribution of care levels in the dataset, balanced accuracy was additionally calculated, as it incorporates both sensitivity and specificity and thus provides a more robust measure. Precision, defined as the proportion of correctly predicted positive cases among all positive predictions, indicates the likelihood that a predicted care level matches the level actually required. The F_β-score was further applied as a generalized measure that combines precision and recall into a single metric, allowing for flexible weighting of the two. When $\beta = 1$, precision and recall are equally weighted, yielding the classical F_1-score. The F_1-score, a special case of the F_β-score with $\beta = 1$, represents the harmonic mean of precision and recall and therefore balances both metrics equally. It is widely used as a single summary measure of classification performance, especially in imbalanced datasets. Values of $\beta > 1$ increase the weight of recall, which emphasizes minimizing false negatives, whereas values of $\beta < 1$ increase the weight of precision, emphasizing the reduction in false positives. This flexibility makes the F_β-score particularly suitable for imbalanced datasets, where the relative importance of recall versus precision depends on the clinical context.

F_β was evaluated for $\beta = 0.5$, 1 and 2 to reflect precision- vs. recall-oriented priorities.

This approach allows for a nuanced evaluation of subjective clinical decision-making and its potential alignment—or divergence—from actual postoperative needs.

Cohen's Kappa was used to quantify the interrater agreement between the two medical assessments, correcting for agreement expected by chance. To test whether the performance of the assessments differed significantly, a Chi-square test was conducted. Statistical analyses were performed using R (version 02.04.24) and Python (version 03.12.11).

3. Results

In the following sections, the performance of the surgical and anesthesiologic assessments is analyzed separately. This is followed by a comparative evaluation of the predictions made by both disciplines.

3.1. Surgery-Based Predictions

Overall, the surgeons show an accuracy of 91.17%, meaning that in approximately 9% of cases, the surgical assessment did not match the level of care actually provided postoperatively. For 2008 patients, a postoperative care level higher than Level 0 was predicted. Among patients who ultimately required Level 3 care, 656 (38.05%) were correctly identified preoperatively. Table 1 compares the surgical prediction with the actual postoperative observation.

Table 1. Surgery-based prediction and actual level (Kappa = 0.0263, p-value < 0.001).

Actual Postoperative Level of Care	Surgery-Based Prediction				
	Level 0	**Level 1**	**Level 2**	**Level 3**	**Total**
Level 0	31,480 (98.21%)	296 (0.92%)	125 (0.39%)	152 (0.47%)	32,053
Level 1	736 (86.59%)	84 (9.88%)	21 (2.47%)	9 (1.06%)	850
Level 2	492 (57.14%)	41 (4.76%)	324 (37.63%)	4 (0.46%)	861
Level 3	772 (44.78%)	28 (1.62%)	268 (15.55%)	656 (38.05%)	1724
total	33,480	449	738	821	35,488

Patients requiring Level 0 care were identified with a sensitivity of 98.21%. The sensitivities for Level 1, Level 2, and Level 3 patients were 9.88%, 37.63%, and 38.05%, respectively. The specificity for the Level 0 group was 41.78%, while specificities for Levels 1 through 3 were 98.95%, 98.80%, and 99.51%, respectively. For patients ultimately requiring Level 3 care, the surgical assessment achieved a balanced accuracy of 68.78%. When a surgeon predicted Level 3 care, this was correct in 79.9% of cases. The precision for Level 0 predictions was 94.03%.

3.2. Anesthesiology-Based Predictions

Overall, the anesthesiologic prediction showed an accuracy of 87.12%, meaning that in approximately 13% of cases, the anesthesiologic assessment did not match the actual postoperative level of care. The overall performance is influenced by the high number of patients requiring only Level 0 care. For 5227 patients, a higher level of postoperative care than standard care (Level 0) was predicted. Among these, 1717 (32.85%) cases were overestimated, and 1049 (20.07%) cases were underestimated in terms of care level. Within the group of patients who ultimately required Level 3 care, 980 (56.84%) were correctly classified. Table 2 compares the anesthesiologic care level predictions with the levels that were actually realized.

With a sensitivity of 91.14%, the Level 0 patient group demonstrated the highest sensitivity. In contrast, sensitivities for Level 2 and Level 3 patients were 57.84% and 56.84%, respectively, while the identification of Level 1 patients showed the lowest sensitivity at 26.59%. The relatively low specificity for Level 0 (69.46%) indicates that patients are often incorrectly classified as requiring only standard care. Specificity increases for higher levels of care.

The balanced accuracy of the anesthesiologic assessment for Level 3 patients was 77.12%, although Level 0 remained the most reliably predicted category. Overall, the

positive predictive values for Level 1 through 3 (11.82%, 34.11%, and 52.83%, respectively) indicate limited precision in predicting higher care levels.

Table 2. Anesthesiology-based prediction and actual level in 35,488 cases (Kappa = 0.0408, *p*-value < 0.001).

Actual Postoperative Level of Care	Anesthesiology-Based Prediction				
	Level 0	**Level 1**	**Level 2**	**Level 3**	**Total**
Level 0	29,212 (91.14%)	1544 (4.82%)	532 (1.66%)	765 (2.39%)	32,053
Level 1	517 (60.82%)	226 (26.59%)	85 (10.00%)	22 (2.59%)	850
Level 2	197 (22.88%)	78 (9.06%)	498 (57.84%)	88 (10.22%)	861
Level 3	335 (19.43%)	64 (3.71%)	345 (20.01%)	980 (56.84%)	1724
total	30,261	1912	1460	1855	35,488

3.3. Comparison of Predictions

A direct comparison between the two disciplines—anesthesiology and surgery—reveals that the surgical assessments demonstrated slightly higher overall accuracy (87.12% vs. 91.17%, $\chi^2 = 2252.32$, $p < 0.001$).

Figure 2 provides detailed information on the different predictions, stratified by the actual postoperative level of care. Level 0 was the most frequently predicted category—even in cases where the actual postoperative care requirement corresponded to Level 1 through 3. In 28,976 instances, both specialties predicted Level 0 care, which was subsequently confirmed by the observed outcome. However, among these presumed low-risk patients, a relevant number ultimately required a higher level of care (Level 1: $n = 485$; Level 2: $n = 114$; Level 3: $n = 161$), indicating potential underestimation of postoperative needs or the occurrence of unexpected intraoperative complications. Discrepancies between surgical and anesthesiologic assessments were also evident. For example, 405 cases were classified as Level 0 by surgery but as Level 3 by anesthesiology, whereas 127 cases showed the inverse outcome. Although the total number of patients jointly predicted to require intensive postoperative care (Level 3) by both disciplines was comparatively low ($n = 502$), this group showed a high concordance with actual outcomes.

Figure 3 compares the sensitivities and specificities. While surgical clinicians achieved higher sensitivity for Level 0 patients, anesthesiology physicians demonstrated significantly higher sensitivities across all other care levels. The conclusions regarding specificity are vice versa.

Table 3 summarizes the diagnostic performance metrics and presents them in a comparative manner. In the comparison of predictive performance across risk levels, substantial effects of the highly imbalanced dataset became evident. As Level 0 constituted the overwhelming majority of cases (90.3%), both anesthesiologists and surgeons achieved high sensitivity for this class (0.911 vs. 0.982). However, specificity differed considerably (0.695 vs. 0.418), reflecting that anesthesiologists were more cautious and thus less likely to incorrectly classify higher-risk patients as Level 0. The dominance of this class also explained the uniformly high positive predictive values (>0.94), which were largely driven by prevalence rather than discriminative ability. Negative predictive values for Level 0 remained low, highlighting the limited reliability of a "non-Level 0" classification in ruling out higher risk. For the intermediate Level 1 category, predictive performance was consistently poor. With very low sensitivity (0.266 vs. 0.099) and near-zero positive predictive values (0.118 vs. 0.187), this category proved almost indistinguishable in real-world practice. The scarcity of cases in this group likely exacerbated the inability of both specialties to consistently recognize Level 1. In Level 2, anesthesiologists achieved somewhat higher sensitivity than

surgeons (0.578 vs. 0.376), though both groups retained high specificity (>0.97). Positive predictive values remained moderate (0.341 vs. 0.439), again reflecting the small prevalence of this class. The clinically most relevant Level 3 (ICU mandatory) category demonstrated complementary strengths of the two specialties. Anesthesiologists detected a greater proportion of ICU cases (sensitivity 0.568 vs. 0.381), whereas surgeons demonstrated higher precision (PPV 0.799 vs. 0.528).

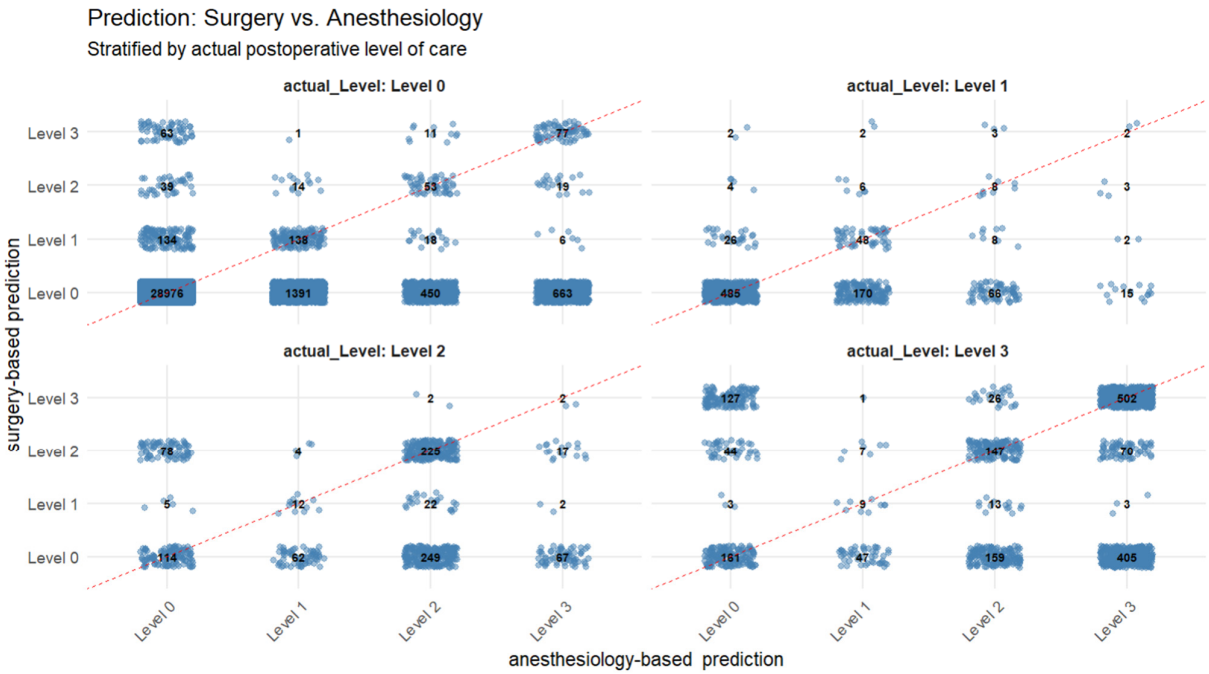

Figure 2. Comparison of surgery- and anesthesiology-based predictions, grouped by the actual level of care.

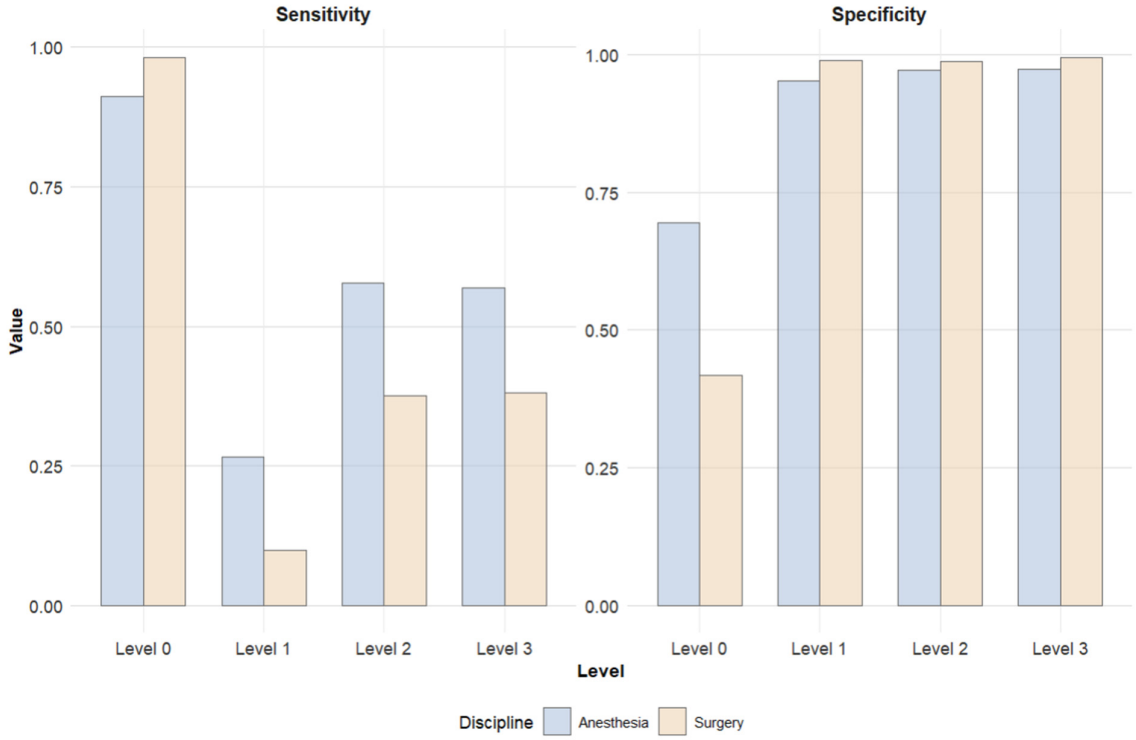

Figure 3. Comparison of specialties with regard to sensitivity (**left**) and specificity (**right**).

Table 3. Test performance metrics: surgery-based prediction (SUR) and anesthesiology-based prediction (AIN).

Actual Postoperative Level of Care	Sensitivity AIN	Sensitivity SUR	Specificity AIN	Specificity SUR	PPV AIN	PPV SUR	NPV AIN	NPV SUR	Prevalence AIN	Prevalence SUR
Level 0	0.911	0.982	0.695	0.418	0.965	0.940	0.456	0.715	0.903	0.903
Level 1	0.266	0.099	0.951	0.989	0.118	0.187	0.981	0.978	0.024	0.024
Level 2	0.578	0.376	0.972	0.988	0.341	0.439	0.989	0.985	0.024	0.024
Level 3	0.568	0.381	0.974	0.995	0.528	0.799	0.978	0.969	0.049	0.049

Figure 4 presents an aggregated comparison: all levels of care exceeding standard recovery (Level 1–3) were combined and contrasted with PACU (Level 0). This visualization allows for assessment of how well patients requiring enhanced postoperative care were identified. The comparison highlights that anesthesiologic assessments resulted in fewer patients being incorrectly classified as Level 0, thereby ensuring that critical resources were more appropriately utilized. However, this approach also led to a higher proportion of false positive classifications.

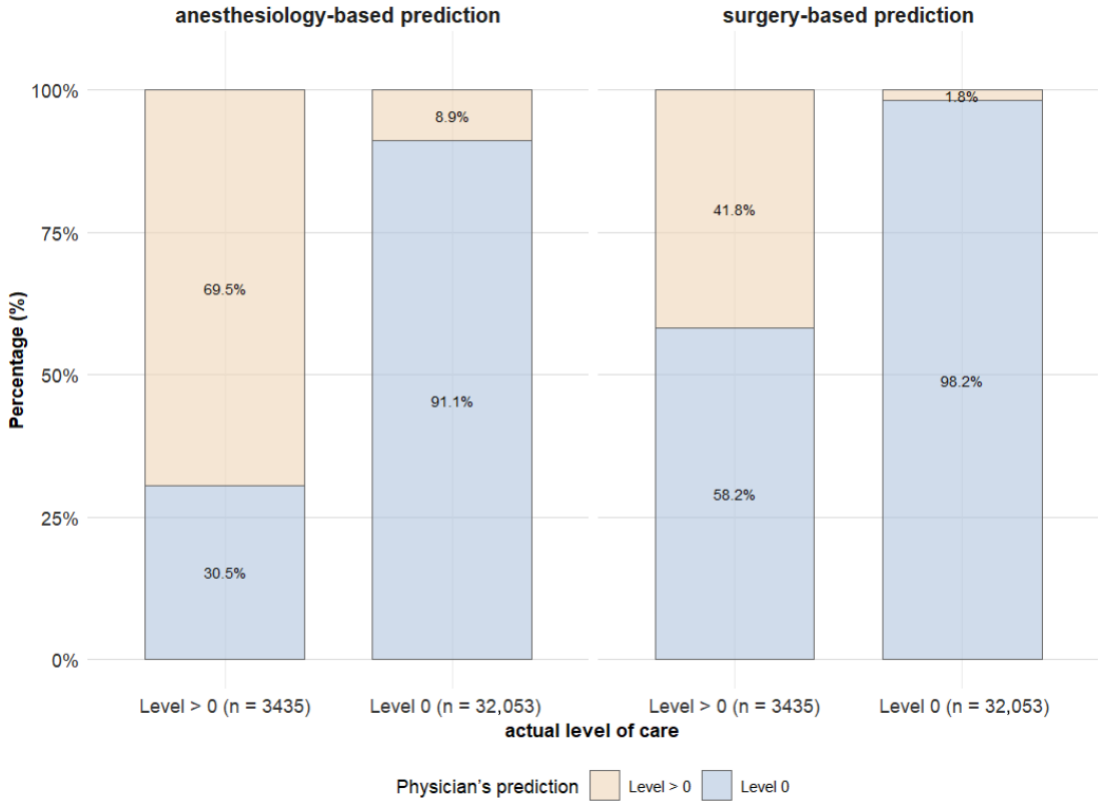

Figure 4. Comparison of the percentage distribution of anesthesiology-based (**left**) and surgery-based predictions (**right**) in cases with aggregated elevated postoperative care requirements.

As can also be seen in Figure 4, the actual distribution of care requirements is, as previously noted, heavily skewed toward the Level 0 group (n = 3435 vs. n = 32,053). To adequately account for this imbalance in performance evaluation, we calculated the F_β-score [36,37]. Moreover, the F_β-score can be weighted and adapted to clinical priorities—such as placing greater emphasis on avoiding false negatives, i.e., underestimating the need for ICU admission—by adjusting the beta coefficient. The F_β-score provides an appropriate metric for evaluating imbalanced data. While, in an ideal setting, a perfectly

accurate prediction would be desirable, in the clinical context it is reasonable to assume that patients incorrectly classified as Level 0 carry a more serious consequence than those incorrectly classified as Level > 0. This asymmetry can be captured using the weighted F_β-score, as illustrated in Figure 5. In our setting, false negative classifications correspond to patients who were incorrectly predicted as Level 0, despite actually requiring a higher level of postoperative care. To account for the greater clinical risk associated with such misclassifications, a higher β value was applied in the F_β-score. By increasing β, recall is weighted more strongly than precision, thereby penalizing false negatives more severely than false positives. This adjustment reflects the clinical priority of minimizing the underestimation of postoperative care needs. When applying higher β values in the F_β-score, misclassifications of patients as Level 0—despite requiring higher levels of care—were penalized more strongly. This weighting reflects the increased clinical relevance of false negative predictions. Accordingly, anesthesia showed higher F_β-scores under recall-oriented weighting (β > 1), whereas surgery performed better under precision-oriented weighting (β < 1).

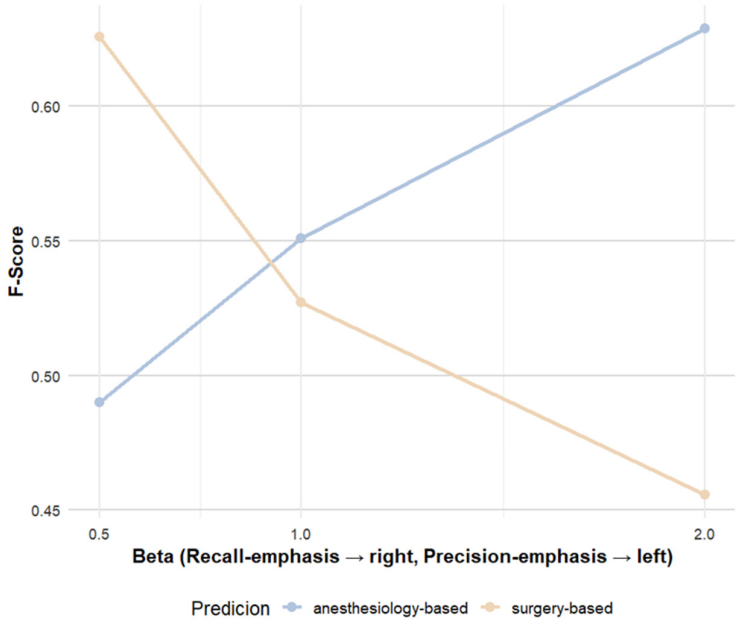

Figure 5. F_β-Score of the anesthesiology-based and surgery-based prediction for varying β weights.

4. Discussion

4.1. Summary of the Results and Their Significance for the Establishment of Future AI Models

Our study shows that both anesthesiologic and surgical predictions demonstrate a high, though not perfect, concordance with the care that was ultimately provided. In approximately 11% of cases, an incorrect prediction was made. Only about half of the patients who ultimately required postoperative intensive care were correctly identified as such during the preoperative assessment by either the surgical or the anesthesiology team, resulting in a large number of patients requiring an unplanned ICU admission. Since ICUs are commonly required to have a high bed utilization due to economic reasons, multiple unplanned ICU admissions of elective cases can pose major challenges.

A variety of AI and ML methods have been proposed for forecasting postoperative ICU demand, as highlighted in the introduction. In their systematic review and meta-analysis, Arina et al. summarized relevant AI models, highlighting three major limitations: the overall modest performance of existing approaches, the challenges of imbalanced datasets, and the lack of a pragmatic clinical benchmark [32]. For preoperative ICU prediction, the

authors identified the study by Chiew et al. as particularly impactful, as it demonstrated a procedure-independent approach relying solely on preoperatively available information [4]. The work of Chiew et al. is based exclusively on preoperative parameters and was applied independently of the surgical procedure. Notably, Chiew et al. themselves explicitly emphasized the need for direct comparisons between model predictions and clinical decision-making. Building on this rationale, the present study uses the work of Chiew et al. as a reference to contextualize and benchmark the performance of subjective physician decision-making [4].

Using ten preoperative parameters, Chiew et al. trained multiple ML models that demonstrated superior sensitivity and, to a lesser extent, specificity compared to the physician-based predictions reported in the present study [4].

Clinical benchmarking yielded F1-scores by anesthesiologists (0.551) and surgeons (0.527) that were comparable to each other and to the ML prediction reported by Chiew et al. (0.28), although the latter performed slightly worse. This indicates that clinical judgment and ML approaches achieve results in a similar range, with minor differences in predictive balance [4]. For implementation, patient safety must remain the primary goal of preoperative risk stratification; thus, false positives are a tolerable trade-off, whereas underestimation poses a direct risk to patients [7].

It is important to emphasize that this comparison is intended as an outlook rather than a direct evaluation. Given the substantial differences in datasets, case mixes, and class distributions, these results are not directly comparable and should be interpreted as illustrative rather than conclusive.

This observation is consistent with the findings of the meta-analysis by Arina et al. and underscores the importance of establishing a benchmark against clinical expertise [32]. Only after such a benchmark has been defined, can it be meaningfully assessed whether, and in what form, an ML tool should be implemented in clinical practice. In the present case, an upstream model appears most appropriate, leveraging the high sensitivity of the ML approach to optimally support clinical decision-makers in their daily work. Taking these requirements into account, Figure 6 outlines a data-driven concept developed from the findings of this study, aimed at guiding practical implementation in real-world clinical settings. The proposed approach could thus improve overall sensitivity without negatively affecting specificity. In practical terms, this would allow for more accurate preoperative planning of ICU bed allocation—at least for elective procedures. Over the approximately one-year observation period, this concept could lead to a maximum increase of 1309 correctly predicted ICU admissions. This analysis provides insights into the quality of preoperative clinical assessments regarding postoperative care requirements from a practical point of view. The results highlight the challenges associated with preoperative assessment of postoperative care requirements in the real world. According to the findings of this study, integrating an appropriate upstream ML-model into the clinical decision-making process would particularly enhance the sensitivity for the most critical patient group-those requiring Level 3 postoperative care.

The novelty of our work lies not only in directly benchmarking surgical and anesthesiologic decision-making but also in formalizing this process into a structured, practice-oriented framework. Previous Head-to-Head comparisons of clinician and AI performance—such as in dermatology [26,27], radiology [28], and pathology [29]—have provided valuable insights, but they remained confined to image-based or narrowly defined diagnostic tasks. In the perioperative domain, only isolated evaluations of predictive models exist [38], yet none have conducted a direct Head-to-Head comparison with real-world physician judgment. To date, no study has, to our knowledge, systematically transformed perioperative decision-making into a reproducible benchmark that accounts for class im-

balance and the socio-technical realities of care. By embedding clinical performance into such a benchmark, our study advances the field conceptually and methodologically: it provides actionable, imbalance-aware performance targets for future AI models, enables fair Head-to-Head evaluations, and thereby supports the translational pathway of AI into routine practice.

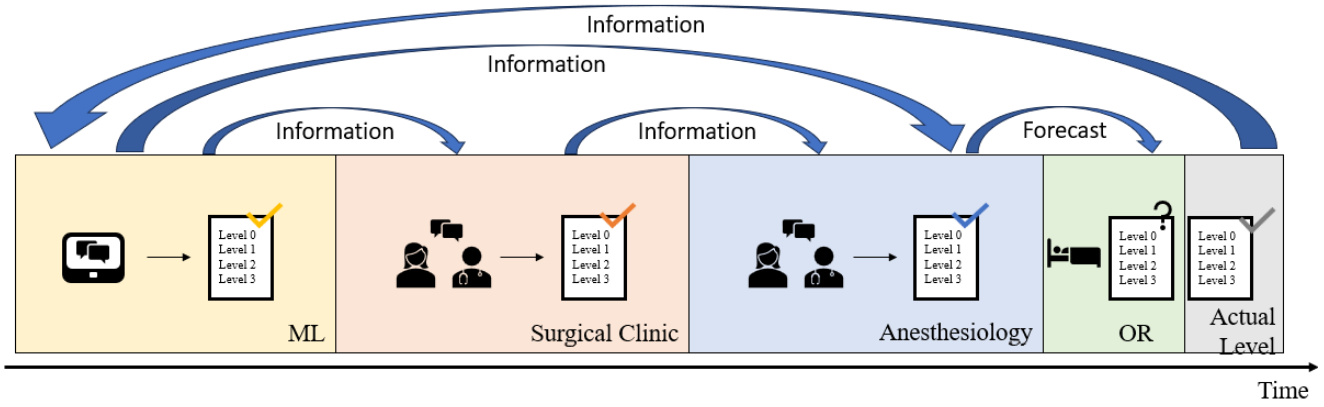

Figure 6. Data-Driven Process for Elective Procedures and Level Classification: ML: Machine-learning; OR: Operating Room; Level 0: Recovery room stay followed by prompt transfer to the general ward (PACU); Level 1: Recovery room stay followed by delayed transfer to the general ward (prolonged PACU); Level 2: Postoperative transfer IMC; Level 3: Postoperative transfer to ICU.

4.2. Limitations

This study is subject to some limitations: the overall performance of physician-based predictions, particularly accuracy, is influenced by the high proportion of patients classified as Level 0. This effect is further amplified by the system's default setting to Level 0. The training level of the participating physicians was not systematically recorded. Therefore, differences in the proportion of residents, board-certified specialists, and senior physicians across both disciplines may have biased the results. Furthermore, the physicians' reasons for their predictions -for example, comorbidities, procedure severity, the interplay of different variables, or the overall clinical impression—were not systematically recorded and analyzed. Our study therefore reflects only the performance (i.e., the outcome) of the clinical decision. Investigating the specific reasons underlying physicians' decisions represents an interesting topic for future research. A key limitation of this study lies in its single-center design, which inherently restricts the ability to assess potential institution-specific biases or regulatory frameworks and limits the generalizability of the findings. Additionally, the need for a higher level of postoperative observation is influenced by center-specific protocols, where patients undergoing certain procedures are always transferred to an ICU postoperatively, regardless of their assumed risk.

The reasons for the actual level of postoperative care realized were not available. It remains unclear whether the observed care levels were based solely on medical necessity or also influenced by organizational factors such as limited bed availability in IMC/ICU units. Consequently, no judgment could be made regarding the medical appropriateness of the care allocation observed in this dataset.

Furthermore, current ML models only predict the need for ICU admission-corresponding to Level 3 in our methodology [4]. Looking ahead, the development and prospective validation of ML models capable of predicting care needs across all levels—including Levels 1 and 2—would be a desirable advancement.

We did not develop or train any machine-learning (ML) model because it is not required to answer the study's primary question—namely, to quantify real-world clinician

performance and establish a clinical benchmark. ML appears only via published results from the literature, which we use to situate our benchmark. This design choice is intentional to help close the translational gap as a first step: by first defining the exact meaning of "better than clinical practice", our benchmark provides actionable, imbalance-aware targets (e.g., minimum recall for Level 3 patients at an acceptable precision) that future models should meet before real-world adoption. The comparison with published ML results must be interpreted with caution. The authors explicitly state that this is meant as an outlook, not as a direct evaluation. Differences in datasets, case structures, and prevalence patterns preclude a one-to-one comparison; the results should therefore be understood as illustrative only. The resulting clinical reference enables fair, prospective, and multicenter head-to-head evaluations of candidate ML systems against clinicians, directly aligning model claims with clinically meaningful goals.

4.3. Strengths

A key strength of this study lies in the use of a large, real-world dataset comprising over 35,000 elective surgical procedures from a tertiary care center. This allows for robust, practice-oriented insights into current clinical decision-making processes. By incorporating both surgical and anesthesiologic assessments and comparing them to the actual postoperative level of care, the study offers a comprehensive evaluation of current practice patterns. The dual-perspective design adds depth and realism, reflecting interdisciplinary workflows in perioperative planning.

Moreover, the classification system employed-differentiating between four distinct postoperative care levels (PACU, prolonged PACU, IMC, ICU)-closely mirrors commonly applied structures in many European hospitals. Despite the inherent heterogeneity in how observation units are defined and utilized across institutions, the applied model is broadly representative and thus enhances the generalizability of findings.

Finally, the study provides a valuable reference benchmark for evaluating future ML-based prediction models. By quantifying the performance of human clinical judgment, it enables a meaningful comparison and helps guide the development and implementation of data-driven decision support systems.

5. Conclusions

This study offers real-world insights into the preoperative prediction of postoperative care requirements following elective surgical interventions.

By establishing a pragmatic benchmark through the performance analysis of clinicians, this work provides a reference framework against which future AI- and ML-based models can be measured. Such a benchmark may represent a crucial step towards overcoming one of the key obstacles that have thus far hindered the implementation of these models in everyday clinical practice. The data-driven concept introduced here shows potential to enhance existing clinical decision-making processes. As demonstrated, early findings from the literature indicate that ML-based approaches can provide meaningful support in this context [4]. Through the present analysis, this potential has, for the first time, been objectively quantified in a real-world clinical setting.

Looking ahead, future interdisciplinary research should prioritize the development of data-driven decision support systems that effectively bridge the interface between human expertise and algorithmic assistance. The goal must be to create actionable, user-centered tools that deliver measurable value for both patients and clinicians in anesthesiology and surgery. In the future, AI methods should be evaluated against the clinical benchmark as established in this study. Particular attention should be paid to handling imbalanced data and to the clinical weighting of misclassifications. Our work provides a solid best-practice

approach and highlights suitable parameters for this purpose. For the adoption of an AI method, it is essential to address issues of missing data and limited generalizability. The framework developed in this study represents a promising foundation for such efforts. It illustrates a pragmatic approach to optimizing resource allocation in perioperative care. The key challenge for future research lies in translating the predictive performance of ML-based models into clinical practice—while navigating the regulatory, ethical, and organizational constraints inherent to healthcare systems. In the end, the impact of AI tools in perioperative medicine will not be determined by technical performance alone, but by how well they align with the clinical benchmark and integrate into the sociotechnical environment of everyday practice.

Author Contributions: Conceptualization, A.R.H., P.S., and O.S.; methodology, A.A. and C.C.B.; software, A.A.; validation, A.A. and C.C.B.; formal analysis, A.A. and C.C.B.; investigation, A.A., C.C.B. and F.B.; resources, A.R.H. and S.S.; data curation, C.C.B. and A.A.; writing—original draft preparation, A.A.; writing—review and editing, P.S., F.G., F.B., and S.S.; visualization, A.A.; supervision, A.R.H. and P.S.; project administration, A.R.H., C.C.B., and J.O.B.; funding acquisition, C.C.B. and J.O.B.; M.D.: review and editing + visualization. All authors have read and agreed to the published version of the manuscript.

Funding: This work was conducted as part of the KISIK project, which is funded by the German Federal Ministry of Education and Research (BMBF). (Funding reference number: 16SV9030).

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki, and approved by the Ethics Committee of Ludwig Maximilian University of Munich (Nr 25-0377-KB, 15 April 2025).

Informed Consent Statement: This study analyzed routinely collected clinical data in pseudonymized form. According to the regulations of the Ludwig Maximilian University of Munich explicit informed consent was not required, as the use of such data for scientific research purposes is permitted under Article 9(2)(j) of the General Data Protection Regulation (GDPR) and §27 of the German Federal Data Protection Act (BDSG). The study protocol was approved by the responsible ethics committee of Ludwig Maximilian University of Munich (Nr 25-0377-KB). The Ethics Committee has confirmed that no formal consultation or full review was required for this study.

Data Availability Statement: For data access inquiries, please contact the corresponding author.

Acknowledgments: We thank the physicians of the surgical and anesthesiologic departments at the University Hospital Augsburg for providing the level predictions. During the preparation of this work the authors used ChatGPT-5 in order to improve language use and readability. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

ICU	Intensive Care Unit
ML	Machine Learning
PACU	Post-Anesthesia Care Unit
IMC	Intermediate Care
SVM	Support Vector Machine
OR	Operating Room
AI	Artificial Intelligence

References

1. van Klei, W.A.; Moons, K.G.M.; Rutten, C.L.G.; Schuurhuis, A.; Knape, J.T.A.; Kalkman, C.J.; Grobbee, D.E. The effect of outpatient preoperative evaluation of hospital inpatients on cancellation of surgery and length of hospital stay. *Anesth. Analg.* **2002**, *94*, 644–649. [CrossRef] [PubMed]
2. Turunen, E.; Miettinen, M.; Setälä, L.; Vehviläinen-Julkunen, K. Financial cost of elective day of surgery cancellations. *J. Hosp. Adm.* **2018**, *7*, 30. [CrossRef]
3. Thevathasan, T.; Copeland, C.C.; Long, D.R.; Patrocínio, M.D.; Friedrich, S.; Grabitz, S.D.; Kasotakis, G.; Benjamin, J.; Ladha, K.; Sarge, T.; et al. The Impact of Postoperative Intensive Care Unit Admission on Postoperative Hospital Length of Stay and Costs: A Prespecified Propensity-Matched Cohort Study. *Anesth. Analg.* **2019**, *129*, 753–761. [CrossRef]
4. Chiew, C.J.; Liu, N.; Wong, T.H.; Sim, Y.E.; Abdullah, H.R. Utilizing Machine Learning Methods for Preoperative Prediction of Postsurgical Mortality and Intensive Care Unit Admission. *Ann. Surg.* **2020**, *272*, 1133–1139. [CrossRef] [PubMed]
5. Grass, F.; Behm, K.T.; Duchalais, E.; Crippa, J.; Spears, G.M.; Harmsen, W.S.; Hübner, M.; Mathis, K.L.; Kelley, S.R.; Pemberton, J.H.; et al. Impact of delay to surgery on survival in stage I-III colon cancer. *Eur. J. Surg. Oncol.* **2020**, *46*, 455–461. [CrossRef]
6. Kompelli, A.R.; Li, H.; Neskey, D.M. Impact of Delay in Treatment Initiation on Overall Survival in Laryngeal Cancers. *Otolaryngol. Head Neck Surg.* **2019**, *160*, 651–657. [CrossRef]
7. Nothofer, S.; Geipel, J.; Aehling, K.; Sommer, B.; Heller, A.R.; Shiban, E.; Simon, P. Postoperative Surveillance in the Postoperative vs. Intensive Care Unit for Patients Undergoing Elective Supratentorial Brain Tumor Removal: A Retrospective Observational Study. *J. Clin. Med.* **2025**, *14*, 2632. [CrossRef]
8. Silva, J.M.; Rocha, H.M.C.; Katayama, H.T.; Dias, L.F.; de Paula, M.B.; Andraus, L.M.R.; Silva, J.M.C.; Malbouisson, L.M.S. SAPS 3 score as a predictive factor for postoperative referral to intensive care unit. *Ann. Intensive Care* **2016**, *6*, 42. [CrossRef]
9. Moonesinghe, S.R.; Mythen, M.G.; Das, P.; Rowan, K.M.; Grocott, M.P.W. Risk stratification tools for predicting morbidity and mortality in adult patients undergoing major surgery: Qualitative systematic review. *Anesthesiology* **2013**, *119*, 959–981. [CrossRef]
10. Devereaux, P.J.; Bradley, D.; Chan, M.T.V.; Walsh, M.; Villar, J.C.; Polanczyk, C.A.; Seligman, B.G.S.; Guyatt, G.H.; Alonso-Coello, P.; Berwanger, O.; et al. An international prospective cohort study evaluating major vascular complications among patients undergoing noncardiac surgery: The VISION Pilot Study. *Open Med.* **2011**, *5*, e193–e200.
11. Cabrera, A.; Bouterse, A.; Nelson, M.; Razzouk, J.; Ramos, O.; Chung, D.; Cheng, W.; Danisa, O. Use of random forest machine learning algorithm to predict short term outcomes following posterior cervical decompression with instrumented fusion. *J. Clin. Neurosci.* **2023**, *107*, 167–171. [CrossRef] [PubMed]
12. van de Sande, D.; van Genderen, M.E.; Verhoef, C.; van Bommel, J.; Gommers, D.; van Unen, E.; Huiskens, J.; Grünhagen, D.J. Predicting need for hospital-specific interventional care after surgery using electronic health record data. *Surgery* **2021**, *170*, 790–796. [CrossRef] [PubMed]
13. Wijnberge, M.; Geerts, B.F.; Hol, L.; Lemmers, N.; Mulder, M.P.; Berge, P.; Schenk, J.; Terwindt, L.E.; Hollmann, M.W.; Vlaar, A.P.; et al. Effect of a Machine Learning-Derived Early Warning System for Intraoperative Hypotension vs Standard Care on Depth and Duration of Intraoperative Hypotension During Elective Noncardiac Surgery: The HYPE Randomized Clinical Trial. *JAMA* **2020**, *323*, 1052–1060. [CrossRef] [PubMed]
14. Li, L.; He, H.; Xiang, L.; Wang, Y. A warning model for predicting patient admissions to the intensive care unit (ICU) following surgery. *Perioper. Med.* **2025**, *14*, 60. [CrossRef]
15. Xu, Z.; Yao, S.; Jiang, Z.; Hu, L.; Huang, Z.; Zeng, Q.; Liu, X. Development and validation of a prediction model for postoperative intensive care unit admission in patients with non-cardiac surgery. *Heart Lung* **2023**, *62*, 207–214. [CrossRef]
16. Turan, E.I.; Baydemir, A.E.; Şahin, A.S.; Özcan, F.G. Effectiveness of ChatGPT-4 in predicting the human decision to send patients to the postoperative intensive care unit: A prospective multicentric study. *Minerva Anestesiol.* **2025**, *91*, 259–267. [CrossRef]
17. Stieger, A.; Schober, P.; Venetz, P.; Andereggen, L.; Bello, C.; Filipovic, M.G.; Luedi, M.M.; Huber, M. Predicting admission to and length of stay in intensive care units after general anesthesia: Time-dependent role of pre- and intraoperative data for clinical decision-making. *J. Clin. Anesth.* **2025**, *103*, 111810. [CrossRef]
18. Özçıbık Işık, G.; Kılıç, B.; Erşen, E.; Kaynak, M.K.; Turna, A.; Özçıbık, O.S.; Yıldırım, T.; Kara, H.V. Prediction of postoperative intensive care unit admission with artificial intelligence models in non-small cell lung carcinoma. *Eur. J. Med. Res.* **2025**, *30*, 293. [CrossRef]
19. Hariharan, S.; Zbar, A. Risk scoring in perioperative and surgical intensive care patients: A review. *Curr. Surg.* **2006**, *63*, 226–236. [CrossRef]
20. Brueckmann, B.; Villa-Uribe, J.L.; Bateman, B.T.; Grosse-Sundrup, M.; Hess, D.R.; Schlett, C.L.; Eikermann, M. Development and validation of a score for prediction of postoperative respiratory complications. *Anesthesiology* **2013**, *118*, 1276–1285. [CrossRef]
21. Kheterpal, S.; Tremper, K.K.; Heung, M.; Rosenberg, A.L.; Englesbe, M.; Shanks, A.M.; Campbell, D.A. Development and validation of an acute kidney injury risk index for patients undergoing general surgery: Results from a national data set. *Anesthesiology* **2009**, *110*, 505–515. [CrossRef]

22. Rozeboom, P.D.; Henderson, W.G.; Dyas, A.R.; Bronsert, M.R.; Colborn, K.L.; Lambert-Kerzner, A.; Hammermeister, K.E.; McIntyre, R.C.; Meguid, R.A. Development and Validation of a Multivariable Prediction Model for Postoperative Intensive Care Unit Stay in a Broad Surgical Population. *JAMA Surg.* **2022**, *157*, 344–352. [CrossRef] [PubMed]
23. Cashmore, R.M.J.; Fowler, A.J.; Pearse, R.M. Post-operative intensive care: Is it really necessary? *Intensive Care Med.* **2019**, *45*, 1799–1801. [CrossRef] [PubMed]
24. Park, C.-M.; Suh, G.Y. Who benefits from postoperative ICU admissions?-more research is needed. *J. Thorac. Dis.* **2018**, *10*, S2055–S2056. [CrossRef] [PubMed]
25. Waydhas, E.; Herting, S.; Kluge, A.; Markewitz, G.; Marx, E.; Muhl, T.; Nicolai, K.; Notz, V. *Intermediate Care Station Empfehlungen zur Ausstattung und Struktur*; Deutsche Interdisziplinäre Vereinigung für Intensiv-und Notfallmedizin eV (DIVI): Berlin, Germany, 2017.
26. Esteva, A.; Kuprel, B.; Novoa, R.A.; Ko, J.; Swetter, S.M.; Blau, H.M.; Thrun, S. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **2017**, *542*, 115–118. [CrossRef]
27. Brinker, T.J.; Hekler, A.; Enk, A.H.; Berking, C.; Haferkamp, S.; Hauschild, A.; Weichenthal, M.; Klode, J.; Schadendorf, D.; Holland-Letz, T.; et al. Deep neural networks are superior to dermatologists in melanoma image classification. *Eur. J. Cancer* **2019**, *119*, 11–17. [CrossRef]
28. Rajpurkar, P.; Irvin, J.; Ball, R.L.; Zhu, K.; Yang, B.; Mehta, H.; Duan, T.; Ding, D.; Bagul, A.; Langlotz, C.P.; et al. Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists. *PLoS Med.* **2018**, *15*, e1002686. [CrossRef]
29. Steiner, D.F.; MacDonald, R.; Liu, Y.; Truszkowski, P.; Hipp, J.D.; Gammage, C.; Thng, F.; Peng, L.; Stumpe, M.C. Impact of Deep Learning Assistance on the Histopathologic Review of Lymph Nodes for Metastatic Breast Cancer. *Am. J. Surg. Pathol.* **2018**, *42*, 1636–1646. [CrossRef]
30. Loftus, T.J.; Tighe, P.J.; Filiberto, A.C.; Efron, P.A.; Brakenridge, S.C.; Mohr, A.M.; Rashidi, P.; Upchurch, G.R.; Bihorac, A. Artificial Intelligence and Surgical Decision-making. *JAMA Surg.* **2020**, *155*, 148–158. [CrossRef]
31. Sokol, K.; Fackler, J.; Vogt, J.E. Artificial intelligence should genuinely support clinical reasoning and decision making to bridge the translational gap. *npj Digit. Med.* **2025**, *8*, 345. [CrossRef] [PubMed]
32. Arina, P.; Kaczorek, M.R.; Hofmaenner, D.A.; Pisciotta, W.; Refinetti, P.; Singer, M.; Mazomenos, E.B.; Whittle, J. Prediction of Complications and Prognostication in Perioperative Medicine: A Systematic Review and PROBAST Assessment of Machine Learning Tools. *Anesthesiology* **2024**, *140*, 85–101. [CrossRef]
33. Ive, J.; Olukoya, O.; Funnell, J.P.; Booker, J.; Lam, S.H.M.; Reddy, U.; Noor, K.; Dobson, R.J.; Luoma, A.M.V.; Marcus, H.J. AI assisted prediction of unplanned intensive care admissions using natural language processing in elective neurosurgery. *npj Digit. Med.* **2025**, *8*, 549. [CrossRef] [PubMed]
34. Cao, Y.; Wang, Y.; Liu, H.; Wu, L. Artificial intelligence revolutionizing anesthesia management: Advances and prospects in intelligent anesthesia technology. *Front. Med.* **2025**, *12*, 1571725. [CrossRef] [PubMed]
35. Montomoli, J.; Hilty, M.P.; Ince, C. Artificial intelligence in intensive care: Moving towards clinical decision support systems. *Minerva Anestesiol.* **2022**, *88*, 1066–1072. [CrossRef] [PubMed]
36. Sokolova, M.; Japkowicz, N.; Szpakowicz, S. Beyond Accuracy, F-Score and ROC: A Family of Discriminant Measures for Performance Evaluation. In Proceedings of the AI 2006: Advances in Artificial Intelligence: 19th Australian Joint Conference on Artificial Intelligence, Hobart, Australia, 4–8 December 2006; Springer: Berlin, Heidelberg, 2006; pp. 1015–1021, ISBN 978-3-540-49787-5.
37. Hicks, S.A.; Strümke, I.; Thambawita, V.; Hammou, M.; Riegler, M.A.; Halvorsen, P.; Parasa, S. On evaluation metrics for medical applications of artificial intelligence. *Sci. Rep.* **2022**, *12*, 5979. [CrossRef]
38. Bihorac, A.; Ozrazgat-Baslanti, T.; Ebadi, A.; Motaei, A.; Madkour, M.; Pardalos, P.M.; Lipori, G.; Hogan, W.R.; Efron, P.A.; Moore, F.; et al. MySurgeryRisk: Development and Validation of a Machine-learning Risk Algorithm for Major Complications and Death After Surgery. *Ann. Surg.* **2019**, *269*, 652–662. [CrossRef]