

# “Complex models, marginal benefits—a multi-centre development and validation study of early warning scores across 2-16 million patient admissions addressing intercurrent medical interventions”

Alexandros Katsiferis<sup>a\*</sup>, Neil Scheidwasser<sup>a</sup>, Tri-Long Nguyen<sup>b</sup>, Theis Lange<sup>c</sup>, Mark P Khurana<sup>a</sup>, Pernille B Nielsen<sup>d</sup>, Kasper Karmark Iversen<sup>e,f,g</sup>, Christian S Meyhoff<sup>e,h</sup>, Eske Kvanner Aasvang<sup>i</sup>, Jesper Mølgaard<sup>i</sup>, Adrian G Zucco<sup>j</sup>, Tibor V Varga<sup>j</sup>, Samir Bhatt<sup>a,k</sup>

## Affiliations:

a Section of Health Data Science & AI, Department of Public Health, University of Copenhagen, Denmark

b Section of Epidemiology, Department of Public Health, University of Copenhagen, Denmark

c Department of Public Health, University of Copenhagen, Denmark

d Department Of Intensive Care, Center For Cancer And Organ Dysfunction, Copenhagen University Hospital - Rigshospitalet

e Department of Clinical Medicine, University of Copenhagen, Denmark

f Department of Emergency Medicine, Herlev Gentofte Hospital, Denmark

g Department of Cardiology, Herlev Gentofte Hospital, Denmark

h Department of Anesthesia and Intensive Care, Copenhagen University Hospital – Bispebjerg and Frederiksberg, Copenhagen, Denmark

i Department Of Anesthesia, Center For Cancer And Organ Dysfunction, Copenhagen University Hospital - Rigshospitalet

j Copenhagen Health Complexity Center, Department of Public Health, University of Copenhagen, Denmark

k MRC Centre for Global Infectious Disease Analysis, Imperial College London, London, United Kingdom

## Corresponding Author Details \*:

Alexandros Katsiferis, PhD

Section of Health Data Science & AI, Department of Public Health, University of Copenhagen, Copenhagen, Denmark

Address: Bartholinsgade 6Q

DK-1356 Copenhagen K

Denmark

Tel: +45 52724022

E-mail: [alexandros.katsiferis@sund.ku.dk](mailto:alexandros.katsiferis@sund.ku.dk)

## Author Details:

### Neil Scheidwasser

Work-Address: Bartholinsgade 6Q, DK-1356, Copenhagen K, Denmark

### Tri-Long Nguyen

Work-Address: Bartholinsgade 6Q, DK-1356, Copenhagen K, Denmark

### Theis Lange

Work-Address: Øster Farimagsgade, 1353 Copenhagen K, Denmark

### Mark P Khurana

Work-Address: Bartholinsgade 6Q, 1356 Copenhagen, Denmark

### Pernille B Nielsen

Work-Address: Blegdamsvej 9, 2100 Copenhagen, Denmark

### Kasper Karmark Iversen

Work-Address: Borgmester Ib Juulsvej 1, 2730 Herlev, Denmark

### Christian S Meyhoff

Work-Address: Bispebjerg Bakke 23, DK-2400 Copenhagen, Denmark

### Eske Kvanner Aasvang

Work-Address: Blegdamsvej 9, 2100 Copenhagen, Denmark

### Jesper Mølgaard

Work-Address: Blegdamsvej 9, 2100 Copenhagen, Denmark

### Adrian G Zucco

Work-Address: Bartholinsgade 6Q, DK-1356, Copenhagen K, Denmark

### Tibor V Varga

Work-Address: Bartholinsgade 2, building 22, DK-1356, Copenhagen K, Denmark

### Samir Bhatt

Work-Address: Bartholinsgade 6Q, DK-1356, Copenhagen K, Denmark

## Summary

**Background** The National Early Warning Score (NEWS) is a nationally recommended, clinically implemented system, used to prevent patient deterioration. While numerous studies have compared predictive models for clinical deterioration, large-scale evaluations of their potential clinical utility remain undetermined. Here, we compared NEWS's clinical net benefit against simplified scoring rules and modern machine learning to determine whether simpler approaches are sufficient or if complex models provide meaningful advantages in clinical practice.

**Methods** We included fifteen Danish hospitals with over 2·16 million patient admissions representing 829 610 unique patients over five years (2018 to 2023). We compared NEWS against both simpler and more complex approaches for predicting 24-hour mortality: NEWS-Light (NEWS without blood pressure and temperature), DEWS (NEWS-Light with age and sex; DEWS denotes the Demographic Early Warning Score), and a model based on eXtreme Gradient Boosting (XGB-EWS) incorporating vital signs, demographics, laboratory markers, plus medical history embeddings extracted using sentence transformers. We used propensity score weighting to mitigate intervention bias and evaluated performance using Area Under the Receiver Operating Characteristic Curve (AUC), calibration, and net benefit.

**Findings** XGB-EWS achieved the highest discrimination (AUC 0·932, 95% Confidence Interval [0·929-0·936]), followed by DEWS (0·908 [0·904-0·912]), NEWS (0·902, [0·898-0·906]), and NEWS-Light (0·879, [0·873-0·885]). Decision curve analysis showed maximum net benefit differences of 1·8 additional correct mortality identifications per 10 000 patients between XGB-EWS and NEWS, and 1·7 per 10 000 between NEWS and NEWS-Light, across the evaluated risk thresholds.

**Interpretation** Machine learning approaches provided marginal clinical utility improvements over traditional scoring systems, with NEWS-Light showing small performance decrements compared to full NEWS. The clinical significance of these differences must be weighed against workflow optimization benefits, suggesting healthcare systems should evaluate trade-offs between predictive performance and operational efficiency when selecting early warning approaches.

**Funding** Novo Nordisk Foundation

## Research in context

### Evidence before this study

Early warning score systems have evolved from single vital sign monitoring to standardized multivariable scores such as the National Early Warning Score (NEWS), to even more sophisticated machine learning and deep learning frameworks utilizing electronic health records data. All these early warning scores are primarily designed to guide clinical decision-making by helping identify patients at risk of clinical deterioration within hospitals. Despite these advances, validation studies predominantly focus on statistical metrics measuring discrimination performance rather than meaningful clinical utility. On Aug 12, 2025, we searched PubMed for English-language studies published in the last 10 years, using terms including "early warning score", "NEWS", "NEWS2", "machine learning", "artificial intelligence", "decision curve analysis", "net benefit", "clinical utility", and "24-hour mortality." While most of the published models, including the more sophisticated machine learning ones, have demonstrated better discrimination compared to traditional early warning scores, we found only one study that combined early warning score validation with clinical utility analysis for short-term clinical deterioration. Additionally, no studies have evaluated clinical utility across multiple patient subgroups while using causal inference methods to address intervention bias in a large-scale healthcare system validation.

### Added value of this study

The current study represents the largest multi-center early warning score validation to date, encompassing 2.16 million patient encounters across 15 Danish hospitals and nine distinct clinical specialties over a period of five years. We used the *predictimand* framework with causal inference methods to address intervention bias, a critical and novel methodological advance in the assessment of early warning scores. Unlike previous research that focused primarily on discrimination metrics, we evaluated clinical utility using decision curve analysis, providing evidence that sophisticated machine learning early-warning-score approaches delivered in overall only marginal clinical utility improvements (up to 1.7 additional correctly identified mortality cases per 10 000 patients) despite better discrimination. We find that a simplified version of NEWS (NEWS without blood pressure and temperature components) achieves a comparable clinical net benefit to full NEWS. Additionally, we provide the first quantitative assessment of the healthcare resources implications in this area of research, showing that simplified approaches could potentially redirect 98.1 full-time equivalent positions annually from routine vital sign collection to direct patient care, providing evidence for healthcare administrators to reallocate clinical resources toward patient interaction and care delivery.

### Implications of all the available evidence

We demonstrated that differences in the ability of the models to identify 24-hour mortality cases correctly were less than 2 per 10 000 patients. Given that often in clinical practice, a proportion of these identifications translates to lives being rescued, the real-world clinical advantage is likely even more modest. These modest gains of machine learning approaches, coupled with NEWS-Light displaying marginal performance decrements compared to the full NEWS, challenge the prevailing emphasis on algorithmic complexity over clinical value. NEWS-Light has the potential to enable workflow optimization, freeing approximately 3 minutes per patient encounter without compromising clinical effectiveness. Future research should conduct prospective cluster-randomized controlled trials comparing NEWS versus NEWS-Light implementation, while the broader prediction modeling research community should adopt utility-based evaluation frameworks to ensure that algorithmic advances translate to tangible improvements in patient care rather than statistical superiority alone.

## Introduction

Early warning systems are important components of modern healthcare services, providing standardized approaches to identify patients at risk of clinical deterioration.<sup>1</sup> Following its development by the Royal College of Physicians in 2012, the National Early Warning Score (NEWS) was implemented throughout National Health Service hospitals and has been adopted or validated in healthcare systems worldwide, including its implementation in Danish hospitals.<sup>1-8</sup> The score ranges from 0 to 20 and consists of seven vital sign parameters: respiratory rate, oxygen saturation, supplemental oxygen requirement, systolic blood pressure, pulse rate, level of consciousness, and temperature. Each parameter is assigned points based on deviation from normal values, with higher scores indicating greater risk of adverse outcomes within 24 hours. In parallel with scoring systems, such as NEWS, there has been interest in developing statistical and machine learning models that leverage broader electronic health records for predicting clinical deterioration. These approaches utilize the probabilistic output instead of a score for each patient and are based on either logistic regression or more modern machine learning architectures.<sup>1,4,7,9,10</sup> The latter typically incorporate laboratory results, medication and comorbidity histories, and procedural data to enhance prognostication beyond that provided by vital signs alone.

Retrospective studies of Early Warning Scores (EWSs) face methodological challenges that could undermine their clinical validity. One limitation is intervention bias: when effective early warning systems prompt interventions that rescue many high-risk patients, retrospective analysis can become problematic.<sup>11-13</sup> In practice, if high-risk patients receive effective care and survive, then their observed outcome (higher survival, that is, lower mortality) deviates from the initial baseline prediction, due to the rescuing intervention in itself. Thus, retrospective analysis may create misleading associations, potentially making high scores appear to overestimate mortality, with the latter being more prominent as healthcare systems become more responsive to EWSs.<sup>14</sup> This represents a core misalignment between how EWSs function clinically (as pre-intervention risk stratification tools) and how they are evaluated retrospectively (based on post-intervention outcomes). The pertinent clinical question – “What would the patient’s risk trajectory be without intervening on them?” – cannot be answered using approaches that measure observed outcomes after treatments have been administered. Beyond intervention bias, the literature has methodological gaps in performance measurement. While recent guidelines, such as TRIPOD+AI, now emphasize the importance of clinical utility metrics such as net benefit, the vast majority of validation research has focused on discrimination measures such as Area Under the Receiver Operator Characteristic Curve (hereinafter referred to as AUC) that are not designed to be interpretable in a clinical setting, cannot capture real-world clinical value or misclassification costs.<sup>1,7,15-18</sup> Moreover, existing subgroup evaluations remain limited to age- and disease-specific strata using discrimination metrics rather than clinical utility measures,<sup>4,19-21</sup> while evidence of practical implementation value, including cost-effectiveness and feasibility, remains scarce.<sup>22</sup>

To address these methodological limitations, we conducted – what we believe is – the largest multi-centre EWS development and validation study to date, comprising 2·16 million individual patient encounters across 15 Danish hospitals. We employed causal prediction approaches to mitigate intervention bias and estimate the natural EWS performance in the absence of clinical interventions. Our primary aim was to evaluate the discriminative performance, calibration, and potential clinical utility of NEWS for predicting 24-hour mortality, comparing it with simplified variants and machine learning models incorporating laboratory results, previous medical procedures, and diagnoses. Secondary aims included assessing performance heterogeneity across subgroups using clinical utility metrics and estimating the healthcare resource implications of implementing simplified approaches.

## Methods

### Study Design & Data

We conducted a multi-center retrospective cohort study, using data from 15 hospitals (10 hospital networks) across the Capital Region and Zealand Region of Denmark, between October 16, 2018, and October 16, 2023, including a diverse range of clinical settings, specialties, and patient populations. Since every Danish citizen has a unique identification number, we directly linked their demographic information (date of birth, sex, and date of death) from the Population Statistics Register with their clinical data available in the hospitals' records. All centers routinely collected NEWS data as part of standard clinical care during the study period. All patients were followed until death or completion of the 24-hour observation period after the initial NEWS score was recorded, whichever occurred first, with no patients lost to follow-up. The study adheres to the TRIPOD+AI guidelines.

### Participants

We included all adult patients ( $\geq 18$  years) admitted to the participating hospitals during the study period who had at least one NEWS assessment recorded and had hospital stays of more than 15 minutes. For individuals with multiple measurements taken during a single admission, we used the first one to establish a baseline risk before clinical interventions. Patients could contribute multiple separate hospital admissions to the study, with each admission treated as an independent encounter for analysis purposes.

### Outcome

The outcome used was all-cause mortality within 24 hours of the initial NEWS recording. We selected that time horizon as it represents the timeframe for which NEWS was originally designed, is the most frequently evaluated period in EWS validation studies, and is clinically relevant for the intended use of these systems, which aim to identify patients at imminent risk of deterioration.<sup>1,14</sup>

### Predictors

We extracted vital sign parameters routinely collected during clinical care: respiratory rate, oxygen saturation, supplemental oxygen, systolic and diastolic blood pressure, pulse rate, level of consciousness, and temperature. Laboratory parameters included hematological markers (hemoglobin, leukocytes, and platelets), biochemical indicators (creatinine, alanine aminotransferase, lactate dehydrogenase, albumin, and C-reactive protein), and specialized tests (arterial and venous lactate and troponin T). For patients with multiple measurements of the same marker, we calculated the median across all measurements of that specific test from the patient's medical history prior to and including the time of the initial NEWS recording. We used the Danish Healthcare Classification System (*Sundhedsvæsenets Klassifikationssystem* (SKS) codes) to acquire data on previous diagnoses and medical procedures from prior hospitalizations, also including Intensive Care Unit (ICU) or respiratory support episodes. Both laboratory data and clinical history variables were only available from 2018 onwards due to data system implementation. Additional predictors included demographic variables (age, sex), hospital of admission, number of previous admissions, and temporal variables related to the NEWS recording (day of week, time of day, month). We further captured SKS codes registered within 24 hours after the initial NEWS measurement to define exposure to major interventions for subsequent causal prediction analyses.

### Sample Size Considerations

We performed sample size calculations using the *pmsampsize* package to assess adequacy for prediction model development.<sup>23</sup> Based on an anticipated AUC of 0.8, 100 predictor parameters for the most complex model, and an observed 24-hour mortality prevalence of 0.22%, the minimum required sample size was 289 823 admissions with 638 mortality events to achieve an events-per-parameter ratio of 6.38.

### Missing Data

Median and mode imputation were respectively applied to continuous and categorical variables. Missing values in clinical history text fields (SKS descriptions) were preserved as missing indicators and handled implicitly through the text processing pipeline. For the eXtreme Gradient Boosting machine learning model (XGB-EWS), which was the only approach that incorporated laboratory parameters as predictors, the algorithm's built-in missing value treatment handled missing laboratory values implicitly. We provide a list with missingness percentages per variable in the final dataset (Appendix, p 2).

### Analytical Methods

We processed the clinical history text in Danish from previous diagnoses and medical procedures using text embeddings, which are high-dimensional numerical representations of textual data. Specifically, we used the “*minishlab/potion-multilingual-128M*” model, a static embedding sentence transformer distilled from a larger Transformer architecture.<sup>24,25</sup> The inputs to that model were the concatenated structured text strings of previous

diagnoses and medical procedures, with the output being 256-dimensional embeddings. Using standard Principal Component Analysis (PCA), we reduced this to 60 principal components, retaining 88.9% of the original variance. A composite embedding score was also created by averaging all principal components for use in the propensity score model, as mentioned below.

We defined 24-hour mortality risk in the absence of major interventions as the variable of interest to be predicted by the study's models. Major clinical interventions were defined as ICU admission, respiratory and ventilatory support, surgical operations, and anesthesia administration, within 24 hours following the initial NEWS recording. We defined the study's target population as the entire population of adult patients with a hospital stay > 15 minutes. Since some of those patients received major rescuing interventions due to a predicted high risk at baseline, we applied inverse probability weighting (IPW) in our EWS validation to correct for the intervention bias that would have distorted our results had we naively kept these patients in the evaluation (i.e. artifactual overestimation of mortality) or had we naively discarded those patients without weighting the remaining analytical sample (i.e. selection bias). IPW allowed us to weight the patients without major interventions such that they approximated the overall target population. In short, IPW allowed us to estimate the 24-hour mortality risk that would be observed without major interventions in the target population. This approach is referred to as "hypothetical strategy" in the novel predictimand framework.<sup>12</sup> To compute IPW, we estimated propensity scores for receiving major interventions using covariate balancing propensity scores, creating weights that made the analytical sample of patients without major interventions representative of the overall target population. The propensity score model included baseline characteristics measured at the first NEWS recording, number of previous hospitalizations, history of ICU/respiratory support, and the clinical history embedding composite score. We assessed covariate balance using standardized mean differences to evaluate whether patients without major interventions were representative of the overall hospitalized cohort. IPW validity relied on sufficient covariates for exchangeable predictive performance across intervention groups.

We developed three logistic regression models with different EWSs as predictors: NEWS, NEWS-Light (using a simplified score that excludes temperature and blood pressure parameters), and DEWS (Demographic Early Warning Score: age and sex in addition to simplified vital sign scoring), a modified version of the model proposed by Candel et al. (2023).<sup>4</sup> We further developed a machine learning tree-based model, XGB-EWS, with all available predictors, including NEWS's vital sign set plus diastolic blood pressure, laboratory values, history of ICU/respiratory support, clinical history embedding components, demographic variables (age, sex), and temporal features. All models incorporated the weights in both the model fitting and performance evaluation. A detailed description of the models used can be found in the Appendix (pp 3—4).

Model performance was assessed on the basis of discrimination, calibration, and clinical utility using grouped ten-fold cross-validation with hospitals as clusters to account for hospital-level effects. Discrimination was measured using the AUC, i.e., the probability that the model assigned higher risks to patients who died than to those who survived. Calibration was assessed using the Brier score, measuring the mean squared difference between predicted probabilities and observed outcomes, and through visual inspection of calibration plots. Clinical utility was evaluated using decision curve analysis, which compares the net benefit of using each model to guide clinical decisions against default strategies of treating all patients or treating no patients across different risk thresholds. Risk thresholds represent the probability cutpoints above which clinicians would initiate interventions or escalate care; for instance, a threshold of 0.01 (1%) means that patients with predicted mortality risk above 1% would receive intervention, while those below would not. We selected the risk thresholds as multiples of the weighted mortality prevalence (1×, 2×, 4×, 8×, 10×, 20×, and 30×), with subgroup analyses using subgroup-specific prevalences. The net benefit quantifies the clinical value of using a prediction model by calculating the proportion of true positives minus the proportion of false positive mortality cases weighted by the odds of the risk threshold. The latter reflects the relative importance of avoiding false positives versus identifying true positives at that specific threshold. An overview of the study's methodology is presented in Figure 1.

### **Role of the funding source**

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

## Results

During the study period from October 16, 2018, to October 16, 2023, we identified 18 877 954 EWS measurements from 2 355 905 hospital encounters across 899 100 unique patients in 10 hospital networks in Denmark. After excluding admissions with hospital stays of  $\leq 15$  minutes and restricting the analysis to the first NEWS assessment per hospitalization, the final cohort comprised 2 353 543 encounters of 898 662 individuals. We excluded those observations exposed to major 24-hour interventions (8% of the total), which resulted in a dataset of 2 161 689 hospital encounters of 829 610 individuals. We subsequently performed inverse probability weighting using the propensity score model to make this remaining sample comprising admissions without major interventions representative of the overall hospitalization cohort (Appendix, pp 5–6). The baseline characteristics of the final weighted cohort are presented in Table 1, stratified by mortality status. The weighted prevalence of 24-hour mortality was 0.22% (5 262 deaths), with the median NEWS score (Interquartile Range) being 1 (2.00), but higher (7.00 (6.00)) among patients who died within 24 hours. The weighted population comprised a majority (51.04%) of patients aged 18–65 years, with 50.36% female patients and 65.58% admitted to internal and acute medicine clinical specialties.

In terms of overall discrimination, XGB-EWS demonstrated the highest performance with an AUC of 0.932 (95% Confidence Interval 0.929–0.936), followed by DEWS at 0.908 (95% CI 0.904–0.912), NEWS at 0.902 (95% CI 0.898–0.906), and NEWS-Light at 0.879 (95% CI 0.873–0.885). The hospital-specific discriminative performance of the models (Figure 2) showed consistent patterns across the strata. We assessed model calibration using weighted deciles, comparing predicted versus observed 24-hour mortality frequencies across risk deciles. For the overall population, all models demonstrated reasonable calibration across the risk spectrum (Figure 3). When stratified by age group, a pattern of overestimation of risks was observed for the youngest group, which was balanced in those between 66 and 80 years of age, and reversed into underestimation in the oldest group for all models except the DEWS. The sex-stratified calibration analysis revealed similar calibration performance between male and female patients across all models. Hospital-specific Brier scores and detailed subgroup calibration analyses are provided in the Appendix (pp 7–9).

We conducted a decision curve analysis to evaluate the clinical utility of each model across a range of risk thresholds, the latter chosen as multiples of the weighted prevalence in the data. Table 2 shows the Net Benefit and the corresponding differences of XGB-EWS and the NEWS versions. For example, at a risk threshold of 0.0180 ( $8 \times$  weighted prevalence), XGB-EWS achieved a net benefit of 8.9 per 10 000 patients compared to NEWS's 7.1 per 10 000. The maximum net benefit difference across the overall sample between XGB-EWS and NEWS, as well as NEWS and NEWS-Light across the evaluated thresholds, was 1.8 and 1.7 per 10 000 patients, respectively. We provide a full decision curve for all models and the overall population in the Appendix (p 10). Across all evaluated subgroups, the net benefit differences between XGB-EWS and NEWS were the highest in patients admitted with a NEWS score of 4 or higher at admission (maximum 17.4 per 10 000 patients), those above 80 years of age (maximum 6.2 per 10 000), and patients with circulatory system diseases (maximum 5.3 per 10 000). NEWS versus NEWS-Light comparisons showed maximum differences of 14.0 per 10 000 patients with a NEWS score of 4 or higher at admission, 5.5 per 10 000 patients with blood-related diseases, and 5.0 per 10 000 patients with circulatory system diseases. The rest of the subgroup comparisons showed differences remaining below 4.5 per 10 000 patients for both comparisons (Appendix, pp 11–18).

We conducted a full-time equivalent (FTE) analysis to estimate the potential workforce impact of implementing NEWS-Light (Figure 4). Over the 5-year study period, the 10-hospital system recorded 18 877 954 EWS assessments (this includes all assessments per admission without exclusion), corresponding to 3 775 591 annual assessments. Assuming a conservative 3-minute time saving per assessment by removing blood pressure and temperature measurements, this represented 188 779 annual hours of potential workflow optimization. Based on the Danish healthcare system standard of 37 working hours per week (1924 hours annually per FTE), this could potentially correspond to 98.1 full-time equivalent positions ( $188\,779 \div 1924 = 98.1$ ). This must be considered further conservative as we assume all hours employed are effective work hours, which is not achievable in a clinical reality.

## Discussion

This large-scale validation study of 2·16 million hospital encounters represents the most comprehensive evaluation of EWSs to date, exceeding all previously reported validation studies. We found that our gradient boosting tree-based model achieved superior discriminative performance compared to traditional scoring systems, consistent with previous research.<sup>7,9,26,27</sup> However, when evaluating potential clinical utility through decision curve analysis, an assessment rarely performed in EWS studies<sup>1</sup>, the net benefit differences were small and likely of questionable clinical significance. While the decision curve analysis showed measurable differences between models, net benefit calculations represent additional case identifications rather than lives rescued. Thus, the detection of up to 1·8 additional correct mortality identification per 10 000 patients does not directly translate to mortality reductions, as clinical outcomes depend on how much an EWS score is used in clinical decisions, the subsequent care quality, treatment availability, patient factors beyond the model's reach, and, importantly, the severity of illness. The marginal clinical utility gains observed suggest that the additional complexity and implementation challenges of advanced models may not be solely justified when traditional scoring rules perform equivalently in terms of actionable clinical benefit across diverse hospital settings and patient populations.

The introduction of NEWS-Light, which eliminates blood pressure and temperature measurements while maintaining comparable predictive performance, is a robust and efficient alternative to improve workflow. By removing two of the most time-consuming and frequently impractical vital sign measurements, NEWS-Light may reduce workload during routine assessments. Blood pressure measurements, particularly in acutely ill patients, often require multiple attempts and can cause discomfort<sup>28</sup>, while temperature measurements may require patient repositioning and extended contact time, which could have been reallocated to other patients in need of care. The elimination of these measurements could potentially reduce each assessment episode by an estimated 3 minutes<sup>29</sup>, resulting in time savings that could be redirected toward patient interaction and care. From a patient-centered care perspective, this workflow optimization could enable healthcare providers to spend additional time listening to patient concerns, providing explanations about their condition, and addressing psychosocial needs that are often overlooked during busy clinical rounds.<sup>30</sup> The 98·1 annual full-time equivalent positions worth of time savings identified across our hospital system reflects potential capacity that could be reallocated from routine data collection to direct patient engagement. Additionally, reducing the frequency of impractical procedures may decrease patient anxiety and improve cooperation with care protocols, particularly important for elderly patients who comprised 19% of our cohort and often experience discomfort or sleep disruptions during repeated vital sign assessments.

Our study has several limitations. The retrospective observational design limits causal inference about the effectiveness of simplified EWS, and the Danish healthcare context may limit generalizability to other healthcare systems with different organizational structures or data availability. Despite our propensity score weighting approach incorporating vital signs, demographics, medical history embeddings from natural language processing, laboratory values, and hospital-specific factors, unmeasured variables from unobserved clinical factors such as clinical intuition, real-time resource constraints, or rapid clinical changes that influence both major intervention decisions and predictive performance of EWSs may persist, potentially biasing our hypothetical risk estimates. Additionally, our definition of major clinical interventions, while clinically informed, could be refined in future studies to explore sensitivity to different criteria. While our study evaluated 24-hour mortality prediction, future research could assess performance over other clinical deterioration endpoints and longer time horizons. Most importantly, while our study demonstrates comparable predictive performance and quantifies potential efficiency gains, clustered randomized controlled trials are needed to evaluate real-world implementation effectiveness. However, our study has several strengths that advance the field. Methodologically, our study is the first to apply the predictimand framework with causal inference methods to address intervention bias in EWS validation, providing a rigorous approach for estimating performance under counterfactual conditions that previous studies have not achieved. The scale of 2·16 million encounters provides the necessary robust statistical power for detecting clinically meaningful differences across diverse patient populations and enables detailed subgroup analyses in EWS research. The multi-hospital validation across 15 hospitals in 10 hospital networks provided thorough validation of model performance, demonstrating consistent results across varied institutional settings and patient populations, strengthening confidence in our findings. Uniquely, our study provides measurable estimates of potential healthcare resource implications, quantifying 98·1 full-time equivalent positions worth of time savings, offering healthcare administrators data for implementation decisions.

In conclusion, these findings indicate that healthcare systems should carefully weigh the trade-offs between predictive performance and workflow optimization when implementing EWSs. While simplified scoring

approaches showed small performance decrements, the potential for substantial time savings suggests that adoption decisions should balance these competing considerations based on local priorities and resources.

### **Author's contributions**

AK, NS, KKI, PBN, and TL conceptualized and designed the study. SB provided supervision for the project. AK and NS conducted all quantitative analyses. AK and SB wrote the first draft of the manuscript. All authors provided methodological insights on the manuscript, had access to the data, and were involved in revising the manuscript. All authors accept responsibility for submitting the manuscript for publication.

### **Declaration of interests**

CSM and EKA have founded a spin-out company, WARD24/7 ApS, with the aim of pursuing the regulatory and commercial activities of the WARD-project (Wireless Assessment of Respiratory and circulatory Distress, a project developing a clinical support system for continuous wireless monitoring of vital signs). WARD24/7 ApS has obtained license agreement for any WARD-project software and patents. One patent has been filed: "Wireless Assessment of Respiratory and circulatory Distress (WARD), EP 21184712.4 and EP 21205557.8". No other authors declared conflicts of interests.

### **Data sharing**

Access to the dataset supporting the conclusion of this article can only legally be accessed on acceptance from the Danish Data Protection Agency, the government body that regulates the use of Danish registry data. Interested parties can contact (PBN & KKI) for further information. We confirm that others, given a suitable project judged by the Danish Data Protection Agency, would be able to access these data in the same manner as the authors. We also confirm that the authors do not have any special access privileges that others do not have. The codebase for the analysis of the current study can be found online: [https://github.com/MLGlobalHealth/Early\\_Warning\\_Scores](https://github.com/MLGlobalHealth/Early_Warning_Scores)

### **Acknowledgments**

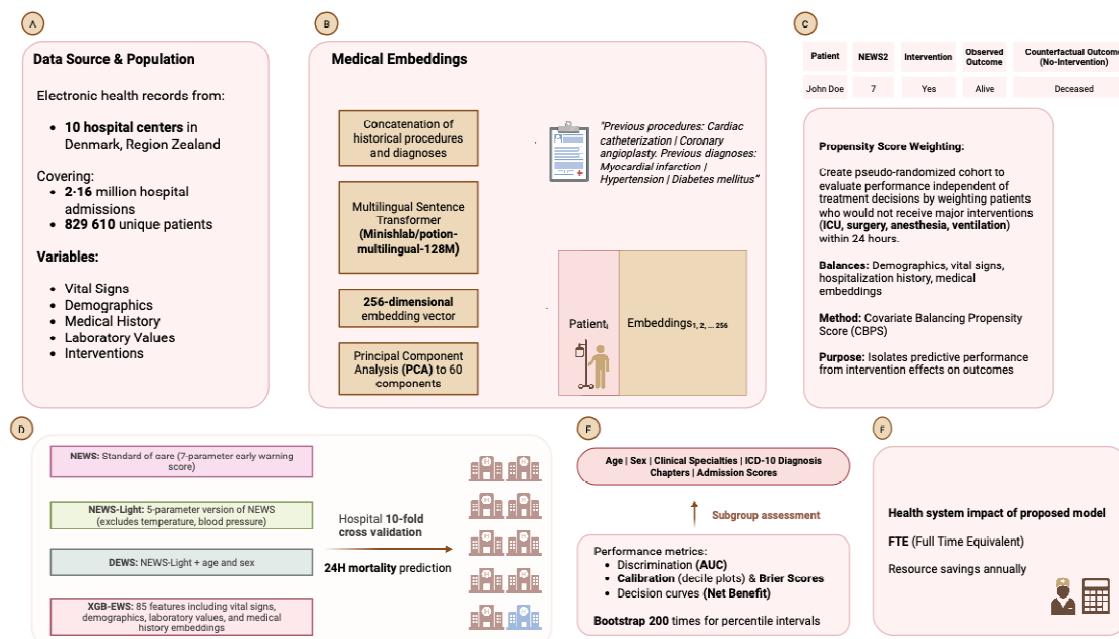
This work is funded from the Novo Nordisk Foundation via The Novo Nordisk Young Investigator Award (NNF20OC0059309). AK acknowledges funding from the Novo Nordisk Foundation via The Novo Nordisk Young Investigator Award (NNF20OC0059309). NS and MPK acknowledge funding from The Danish National Research Foundation via a chair grant (DNRF160). TVV is supported by the "Data Science Investigator - Emerging 2022" grant from the Novo Nordisk Foundation (NNF22OC0075284). AGZ and The Copenhagen Health Complexity Center is funded by TrygFonden. SB acknowledges funding from the MRC Centre for Global Infectious Disease Analysis (reference MR/X020258/1). SB acknowledges support from the Novo Nordisk Foundation via The Novo Nordisk Young Investigator Award (NNF20OC0059309). SB acknowledges the Danish National Research Foundation (DNRF160) through the chair grant. SB acknowledges support from Schmidt Sciences via the Schmidt Polymath Award (G-22-63345). We acknowledge the use of BioRender for the creation of Figure 1 (Katsiferis, A. (2025) <https://BioRender.com/f1b7o4s>).

## References

- 1 Gerry S, Bonnici T, Birks J, *et al.* Early warning scores for detecting deterioration in adult hospital patients: systematic review and critical appraisal of methodology. *BMJ* 2020; : m1501.
- 2 Smith GB, Prytherch DR, Meredith P, Schmidt PE, Featherstone PI. The ability of the National Early Warning Score (NEWS) to discriminate patients at risk of early cardiac arrest, unanticipated intensive care unit admission, and death. *Resuscitation* 2013; **84**: 465–70.
- 3 Mhd Yunin NS, Tan TL. National Early Warning Score 2 (NEWS) as a prognostic tool for adult patients in emergency department: A retrospective observational study. *PLoS One* 2025; **20**: e0326058.
- 4 Candel BGJ, Nissen SK, Nickel CH, *et al.* Development and External Validation of the International Early Warning Score for Improved Age- and Sex-Adjusted In-Hospital Mortality Prediction in the Emergency Department. *Crit Care Med* 2023; **51**: 881–91.
- 5 Bell D, Baker J, Williams C, Bassin L. A Trend-Based Early Warning Score Can Be Implemented in a Hospital Electronic Medical Record to Effectively Predict Inpatient Deterioration. *Crit Care Med* 2021; **49**: e961–7.
- 6 Nielsen PB, Langkjær CS, Schultz M, *et al.* Clinical assessment as a part of an early warning score—a Danish cluster-randomised, multicentre study of an individual early warning score. *Lancet Digit Health* 2022; **4**: e497–506.
- 7 Edelson DP, Churpek MM, Carey KA, *et al.* Early Warning Scores With and Without Artificial Intelligence. *JAMA Netw Open* 2024; **7**: e2438986.
- 8 Williams B. The National Early Warning Score: from concept to NHS implementation. *Clinical Medicine* 2022; **22**: 499–505.
- 9 Churpek MM, Carey KA, Snyder A, *et al.* Multicenter Development and Prospective Validation of eCARTv5: A Gradient-Boosted Machine-Learning Early Warning Score. *Crit Care Explor* 2025; **7**: e1232.
- 10 Rong R, Gu Z, Lai H, *et al.* A deep learning model for clinical outcome prediction using longitudinal inpatient electronic health records. *JAMIA Open* 2025; **8**. DOI:10.1093/jamiaopen/ooaf026.
- 11 Logan Ellis H, Palmer E, Teo JT, Whyte M, Rockwood K, Ibrahim Z. The early warning paradox. *NPJ Digit Med* 2025; **8**: 81.
- 12 van Geloven N, Swanson SA, Ramspek CL, *et al.* Prediction meets causal inference: the role of treatment in clinical prediction models. *Eur J Epidemiol* 2020; **35**: 619–30.
- 13 Balcarcel DR, Mehta SD, Dixon CG, *et al.* Feedback loops in intensive care unit prognostic models: an under-recognised threat to clinical validity. *Lancet Digit Health* 2025; : 100880.
- 14 Holland M, Kellett J. The United Kingdom's National Early Warning Score: should everyone use it? A narrative review. *Intern Emerg Med* 2023; **18**: 573–83.
- 15 Collins GS, Moons KGM, Dhiman P, *et al.* TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ* 2024; : e078378.
- 16 Holland M, Kellett J. A systematic review of the discrimination and absolute mortality predicted by the National Early Warning Scores according to different cut-off values and prediction windows. *Eur J Intern Med* 2022; **98**: 15–26.
- 17 Collins GS, Dhiman P, Ma J, *et al.* Evaluation of clinical prediction models (part 1): from development to external validation. *BMJ* 2024; : e074819.
- 18 Vickers AJ, Elkin EB. Decision curve analysis: A novel method for evaluating prediction models. *Medical Decision Making* 2006; **26**: 565–74.
- 19 Price C, Prytherch D, Kostakis I, Briggs J. Evaluating the performance of the National Early Warning Score in different diagnostic groups. *Resuscitation* 2023; **193**: 110032.
- 20 Alhמוד B, Bonnici T, Patel R, Melley D, Williams B, Banerjee A. Performance of universal early warning scores in different patient subgroups and clinical settings: a systematic review. *BMJ Open* 2021; **11**: e045849.

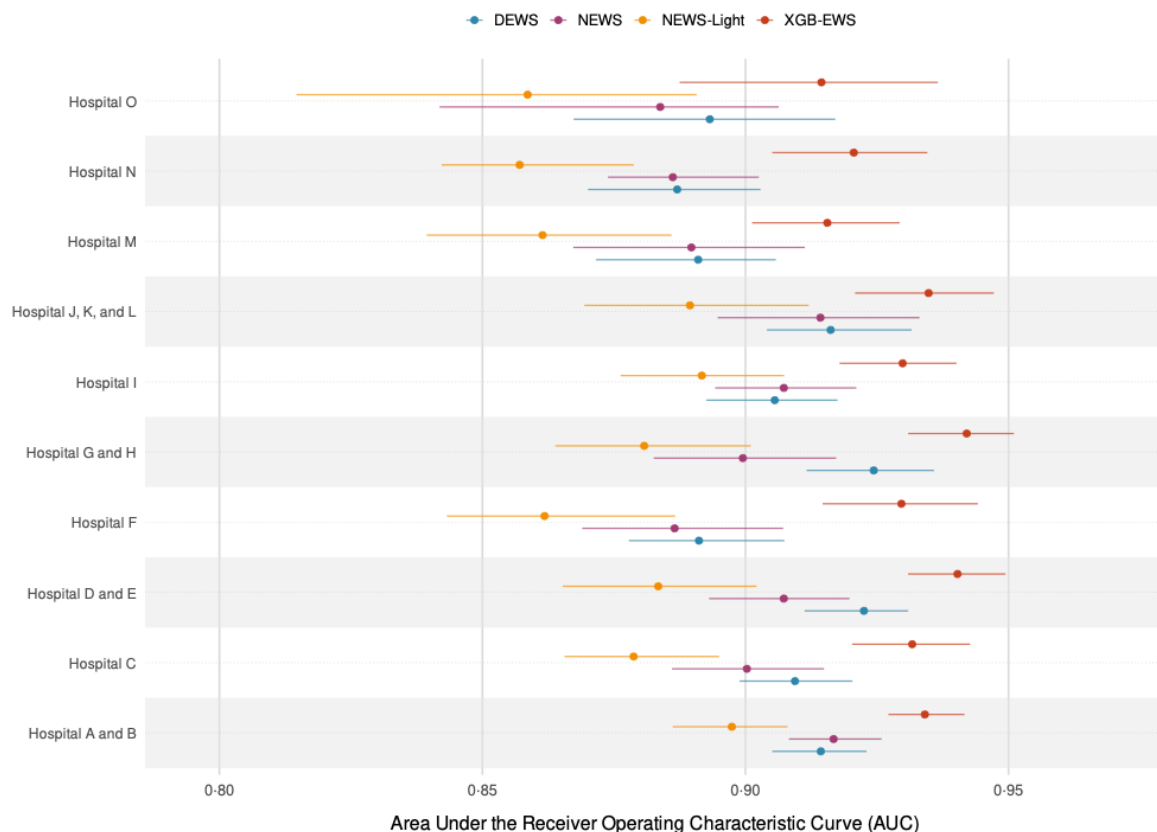
- 21 Bae SJ, Chung HS, Choi Y, *et al.* Comparison of early warning scores for predicting outcomes in adult and older patients in emergency department: Multicenter study. *Am J Emerg Med* 2025; **96**: 91–7.
- 22 Arshi B, Cowley LE, Rijnhart E, Reeve K, Smits LJ, Wynants L. External validation, impact assessment and clinical utilization of clinical prediction models: a prospective cohort study. *J Clin Epidemiol* 2025; **186**: 111902.
- 23 Riley RD, Snell KI, Ensor J, *et al.* Minimum sample size for developing a multivariable prediction model: PART II □ binary and time□to□event outcomes. *Stat Med* 2019; **38**: 1276–96.
- 24 Reimers N, Gurevych I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference* 2019; : 3982–92.
- 25 MinishLab/model2vec: Fast State-of-the-Art Static Embeddings.  
<https://github.com/MinishLab/model2vec> (accessed July 23, 2025).
- 26 Yuan S, Yang Z, Li J, Wu C, Liu S. AI-Powered early warning systems for clinical deterioration significantly improve patient outcomes: a meta-analysis. *BMC Med Inform Decis Mak* 2025; **25**: 203.
- 27 Veldhuis LI, Woittiez NJC, Nanayakkara PWB, Ludikhuizen J. Artificial Intelligence for the Prediction of In-Hospital Clinical Deterioration: A Systematic Review. *Crit Care Explor* 2022; **4**: e0744.
- 28 Tholl U, Forstner K, Anlauf M. Measuring blood pressure: pitfalls and recommendations. *Nephrology Dialysis Transplantation* 2004; **19**: 766–70.
- 29 Dall’Ora C, Griffiths P, Hope J, *et al.* How long do nursing staff take to measure and record patients’ vital signs observations in hospital? A time-and-motion study. *Int J Nurs Stud* 2021; **118**: 103921.
- 30 Kwame A, Petrucka PM. A literature-based study of patient-centered care and communication in nurse-patient interactions: barriers, facilitators, and the way forward. *BMC Nurs* 2021; **20**: 158.

## Figures



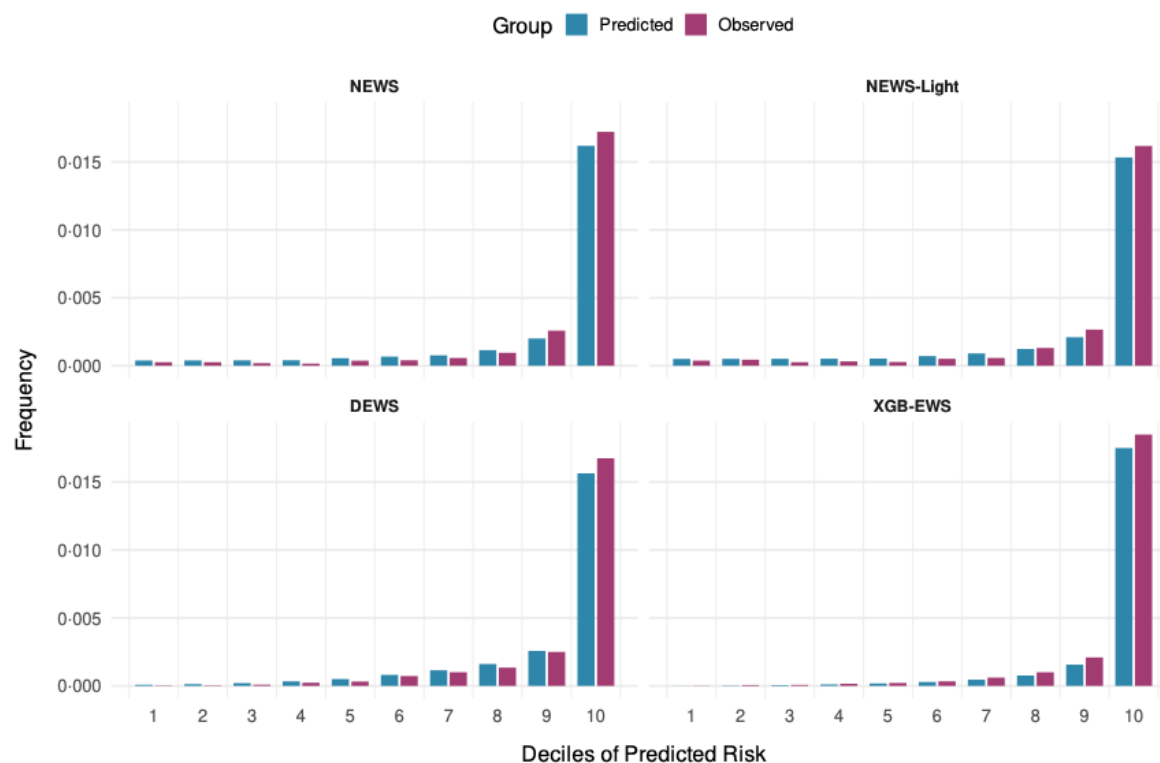
**Figure 1: Overview of the study's workflow**

(A) Electronic health records from Danish hospitals capturing patient data, including vital signs, demographics, medical history, laboratory values, and clinical interventions. (B) To incorporate historical medical information, previous procedures and diagnoses were concatenated into text strings, then transformed into numerical embeddings using multilingual sentence transformers and dimensionality reduction. (C) Propensity score weighting created a balanced cohort that allows evaluation of early warning score performance in the counterfactual scenario where major interventions had not occurred. (D) Four prediction models were compared using hospital-clustered cross-validation to predict 24-hour mortality, ranging from the standard NEWS to machine learning approaches incorporating all available data. (E) Model performance was assessed through multiple metrics and subgroup analyses across patient demographics and clinical contexts. (F) Potential healthcare resource savings were estimated for implementing simplified early warning approaches.



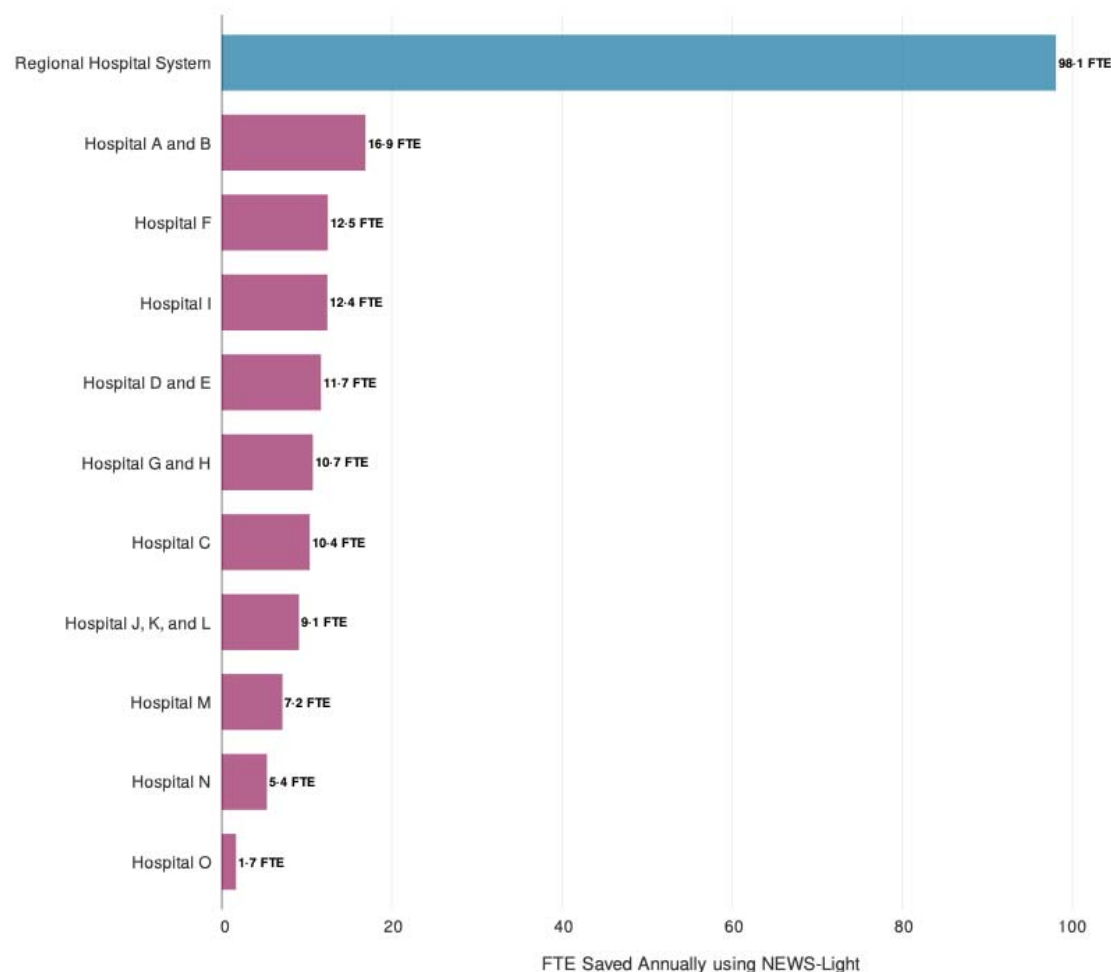
**Figure 2: Forest plot of hospital-specific discriminative performance of EWS algorithms**

The Area under the Receiver Operating Characteristic Curve (AUC) with 95% confidence intervals for predicting 24-hour mortality across the 10 hospital clusters is illustrated. Each point represents the discriminative performance of the four prediction models within individual hospital settings. Confidence intervals were calculated using 200 bootstrap resamples with the percentile method (2.5th and 97.5th percentiles). The four models were XGB-EWS, DEWS, NEWS, and NEWS-Light, and evaluated across all 15 participating hospitals in the study's hospital network.



**Figure 3: Model calibration across the overall population**

Calibration performance comparing predicted versus observed 24-hour mortality frequencies across deciles of predicted risk for all four models. Patients were divided into ten equal groups (deciles) based on predicted risk, with predicted frequencies (blue bars) compared against observed mortality frequencies (purple bars) within each decile. Perfect calibration is indicated when predicted and observed frequencies align closely. Frequencies were calculated using inverse probability weights to reflect the target population.



**Figure 4: Potential healthcare resource reallocation from NEWS-Light implementation**

Annual full-time equivalent (FTE) positions that could potentially be redirected to direct patient care by implementing NEWS-Light instead of standard NEWS across participating hospitals. Resource estimates are calculated based on eliminating blood pressure and temperature measurements, estimated at 3 minutes per assessment. FTE calculations assume 1924 annual working hours per position (Danish standard of 37 hours per week). The Regional Hospital System represents the combined annual reallocation across all individual hospitals. All estimates represent annual values derived from the 5-year study period assessment volumes.

Tables

Table 1: Baseline characteristics of untreated weighted patients by 24-Hour mortality status

| Characteristic                                                                       | Overall<br>N = 2342239 | Deceased<br>N = 5262 | Alive<br>N = 2336976 |
|--------------------------------------------------------------------------------------|------------------------|----------------------|----------------------|
| Age group                                                                            |                        |                      |                      |
| 18-65                                                                                | 1 195 536 (51·04%)     | 646 (12·28%)         | 1 194 890 (51·13%)   |
| 66-80                                                                                | 713 354 (30·46%)       | 1 948 (37·02%)       | 711 406 (30·44%)     |
| 80+                                                                                  | 433 349 (18·50%)       | 2 668 (50·70%)       | 430 680 (18·43%)     |
| Sex                                                                                  |                        |                      |                      |
| Female                                                                               | 1 179 468 (50·36%)     | 2 367 (44·98%)       | 1 177 101 (50·37%)   |
| Male                                                                                 | 1 162 771 (49·64%)     | 2 895 (55·02%)       | 1 159 876 (49·63%)   |
| Number of previous hospitalizations                                                  | 1 (3·00)               | 2 (4·00)             | 1 (3·00)             |
| Hospital                                                                             |                        |                      |                      |
| Hospital D and E                                                                     | 272 294 (11·63%)       | 555 (10·55%)         | 271 739 (11·63%)     |
| Hospital G and H                                                                     | 257 537 (11·00%)       | 423 (8·04%)          | 257 113 (11·00%)     |
| Hospital O                                                                           | 46 006 (1·96%)         | 143 (2·72%)          | 45 863 (1·96%)       |
| Hospital A and B                                                                     | 499 461 (21·32%)       | 1 143 (21·72%)       | 498 319 (21·32%)     |
| Hospital M                                                                           | 143 804 (6·14%)        | 364 (6·92%)          | 143 440 (6·14%)      |
| Hospital J, K, and L                                                                 | 177 265 (7·57%)        | 435 (8·27%)          | 176 830 (7·57%)      |
| Hospital C                                                                           | 280 667 (11·98%)       | 673 (12·79%)         | 279 994 (11·98%)     |
| Hospital N                                                                           | 124 838 (5·33%)        | 425 (8·08%)          | 124 413 (5·32%)      |
| Hospital F                                                                           | 280 724 (11·99%)       | 443 (8·42%)          | 280 281 (11·99%)     |
| Hospital I                                                                           | 259 643 (11·09%)       | 658 (12·50%)         | 258 985 (11·08%)     |
| Clinical specialty                                                                   |                        |                      |                      |
| Cardiopulmonary                                                                      | 169 784 (7·25%)        | 260 (5·00%)          | 169 518 (7·25%)      |
| Critical Care                                                                        | 75 717 (3·23%)         | 48 (0·91%)           | 75 669 (3·24%)       |
| Gastroenterology                                                                     | 24 072 (1·03%)         | 34 (0·65%)           | 24 038 (1·03%)       |
| Geriatrics & Palliative                                                              | 12 179 (0·52%)         | 20 (0·38%)           | 12 159 (0·52%)       |
| Internal & Acute Medicine                                                            | 1 536 061 (65·58%)     | 4 230 (80·40%)       | 1 531 831 (66·55%)   |
| Nephrology                                                                           | 17 621 (0·75%)         | 24 (0·46%)           | 17 596 (0·75%)       |
| Neurology                                                                            | 63 268 (2·70%)         | 121 (2·30%)          | 63 147 (2·70%)       |
| Oncology & Hematology                                                                | 55 616 (2·37%)         | 108 (2·05%)          | 55 508 (2·38%)       |
| Surgical                                                                             | 353 374 (15·09%)       | 388 (7·38%)          | 352 986 (15·10%)     |
| Other                                                                                | 34 547 (1·47%)         | 22 (0·42%)           | 34 525 (1·48%)       |
| Consciousness                                                                        |                        |                      |                      |
| A (Alert)                                                                            | 2 320 171 (99·06%)     | 3 912 (74·33%)       | 2 316 260 (99·11%)   |
| VPU (Verbal-Pain-Unresponsive)                                                       | 22 067 (0·94%)         | 1 351 (25·67%)       | 20 717 (0·89%)       |
| Supplemental oxygen                                                                  |                        |                      |                      |
| Air                                                                                  | 2 100 961 (89·70%)     | 1 956 (37·17%)       | 2 099 004 (89·82%)   |
| Oxygen                                                                               | 241 278 (10·30%)       | 3 306 (62·83%)       | 237 972 (10·18%)     |
| Respiratory rate (breaths/min)                                                       | 16 (2·00)              | 21 (10·00)           | 16 (2·00)            |
| Pulse (beats/min)                                                                    | 79 (21·00)             | 93 (33·00)           | 79 (22·00)           |
| Temperature (°C)                                                                     | 36·72 (0·72)           | 36·61 (1·00)         | 36·72 (0·72)         |
| Oxygen saturation (%)                                                                | 97 (3·00)              | 95 (5·00)            | 97 (3·00)            |
| Systolic blood pressure (mm Hg)                                                      | 133 (30·00)            | 116 (42·00)          | 133 (30·00)          |
| Diastolic blood pressure (mm Hg)                                                     | 76 (19·00)             | 68 (25·00)           | 76 (19·00)           |
| Previous ICU/respiratory support                                                     |                        |                      |                      |
| No history of ICU or respiratory support                                             | 2 304 589 (98·39%)     | 5 105 (97·02%)       | 2 299 484 (98·40%)   |
| Previous history of ICU or respiratory support                                       | 37 649 (1·61%)         | 157 (2·98%)          | 37 492 (1·60%)       |
| NEWS score                                                                           | 1·00 (2·00)            | 7·00 (6·00)          | 1·00 (2·00)          |
| Hemoglobin (g/L)                                                                     | 8·30 (1·45)            | 7·60 (1·70)          | 8·30 (1·45)          |
| Leukocytes (×10 <sup>9</sup> /L)                                                     | 7·60 (3·10)            | 8·30 (3·50)          | 7·60 (3·10)          |
| Platelets (×10 <sup>9</sup> /L)                                                      | 250 (95·00)            | 245 (114·00)         | 250 (95·00)          |
| Creatinine (μmol/L)                                                                  | 74 (27·00)             | 82 (45·00)           | 74 (27·00)           |
| Alanine aminotransferase (U/L)                                                       | 23 (14·00)             | 20 (13·00)           | 23 (14·00)           |
| Lactate dehydrogenase (U/L)                                                          | 192 (52·00)            | 210 (71·00)          | 192 (52·00)          |
| Albumin (g/L)                                                                        | 36 (7·00)              | 32 (7·00)            | 36·0 (7·00)          |
| C-reactive protein (mg/L)                                                            | 5 (20·00)              | 17 (40·00)           | 5 (20·00)            |
| Lactate, arterial blood (mmol/L)                                                     | 1·20 (0·85)            | 1·50 (1·40)          | 1·20 (0·85)          |
| Troponin T (ng/L)                                                                    | 20 (50·00)             | 47 (83·00)           | 20 (49·00)           |
| Lactate, venous blood (mmol/L)                                                       | 1·40 (1·00)            | 1·45 (1·10)          | 1·40 (1·00)          |
| Data are n (%) or Median (IQR). IQR = Interquartile Range (Quartile 3 – Quartile 3). |                        |                      |                      |

medRxiv preprint doi: <https://doi.org/10.1101/2025.10.14.25337794>; this version posted October 14, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license.

Table 2: Net Benefit Comparison of different strategies and models at specified risk thresholds

| Risk Threshold                                                                                                                                                                        | Net Benefit per 10 000 Patients |            |      |            |         | Net Benefit Difference per 10 000 Patients |                    |
|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------|------------|------|------------|---------|--------------------------------------------|--------------------|
|                                                                                                                                                                                       | Treat All                       | Treat None | NEWS | NEWS-Light | XGB-EWS | NEWS vs XGB-EWS                            | NEWS vs NEWS-Light |
| 0.0022                                                                                                                                                                                | 0.0                             | 0.0        | 15.0 | 14.4       | 15.8    | -0.8                                       | 0.6                |
| 0.0045                                                                                                                                                                                | -22.6                           | 0.0        | 12.6 | 11.7       | 13.8    | -1.2                                       | 0.9                |
| 0.0090                                                                                                                                                                                | -68.0                           | 0.0        | 10.1 | 8.4        | 11.4    | -1.3                                       | 1.7                |
| 0.0180                                                                                                                                                                                | -160.1                          | 0.0        | 7.1  | 5.9        | 8.9     | -1.8                                       | 1.2                |
| 0.0225                                                                                                                                                                                | -206.8                          | 0.0        | 6.4  | 5.2        | 8.1     | -1.7                                       | 1.2                |
| 0.0449                                                                                                                                                                                | -446.9                          | 0.0        | 4.1  | 2.8        | 5.9     | -1.8                                       | 1.3                |
| 0.0674                                                                                                                                                                                | -698.6                          | 0.0        | 2.8  | 1.9        | 4.6     | -1.8                                       | 0.9                |
| Positive values indicate NEWS has higher net benefit, whereas negative values indicate the comparator is better. Risk thresholds are chosen as multiples of the mortality prevalence. |                                 |            |      |            |         |                                            |                    |

A

Data Source & Population

Electronic health records from:

- 10 hospital centers in Denmark, Region Zealand

Covering:

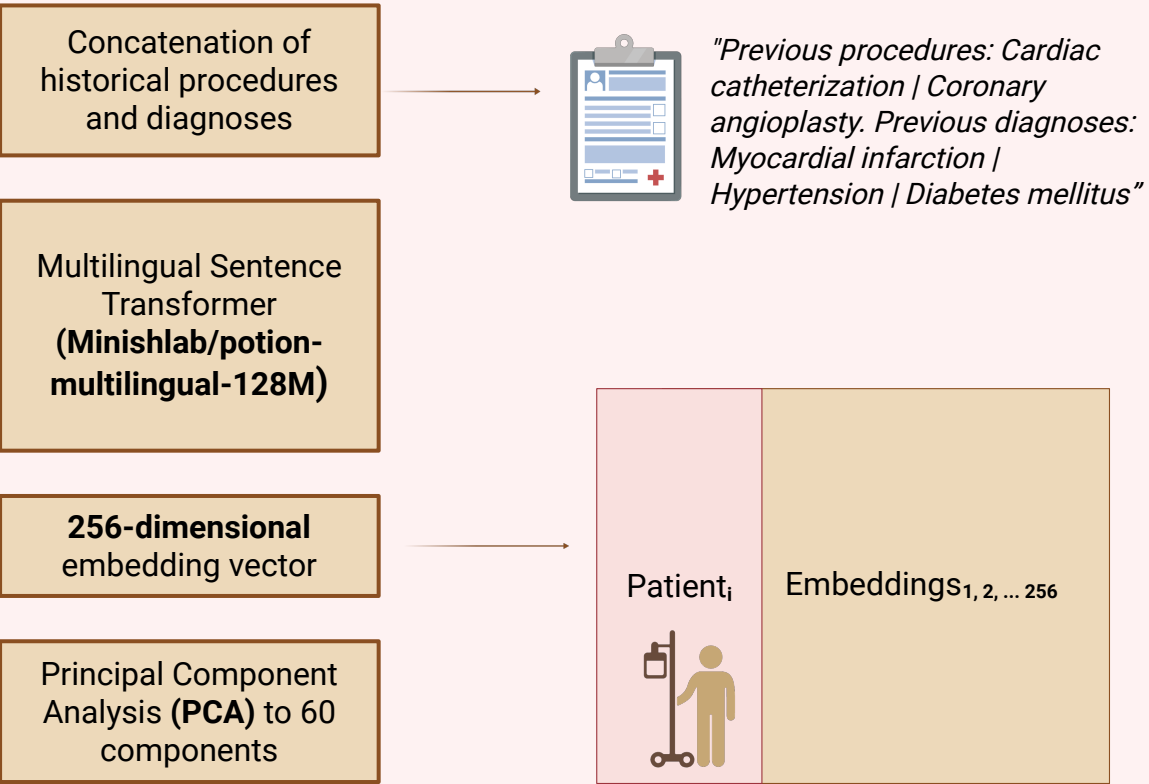
- 2.16 million hospital admissions
- 829 610 unique patients

Variables:

- Vital Signs
- Demographics
- Medical History
- Laboratory Values
- Interventions

B

Medical Embeddings



C

| Patient  | NEWS2 | Intervention | Observed Outcome | Counterfactual Outcome (No-Intervention) |
|----------|-------|--------------|------------------|------------------------------------------|
| John Doe | 7     | Yes          | Alive            | Deceased                                 |

Propensity Score Weighting:

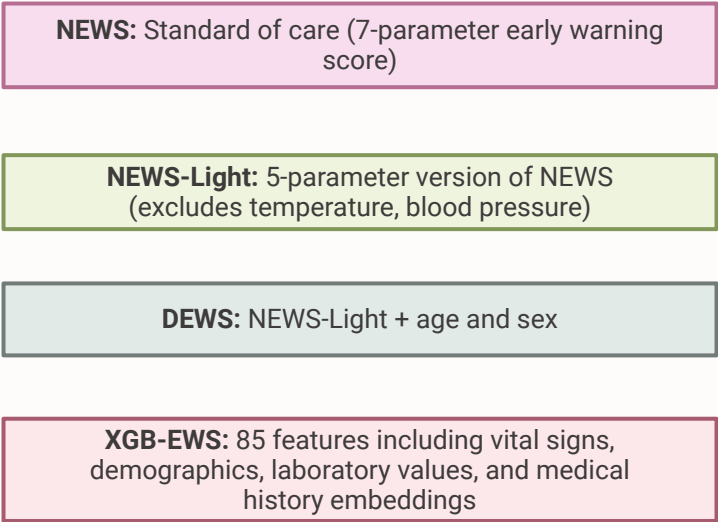
Create pseudo-randomized cohort to evaluate performance independent of treatment decisions by weighting patients who would not receive major interventions (**ICU, surgery, anesthesia, ventilation**) within 24 hours.

**Balances:** Demographics, vital signs, hospitalization history, medical embeddings

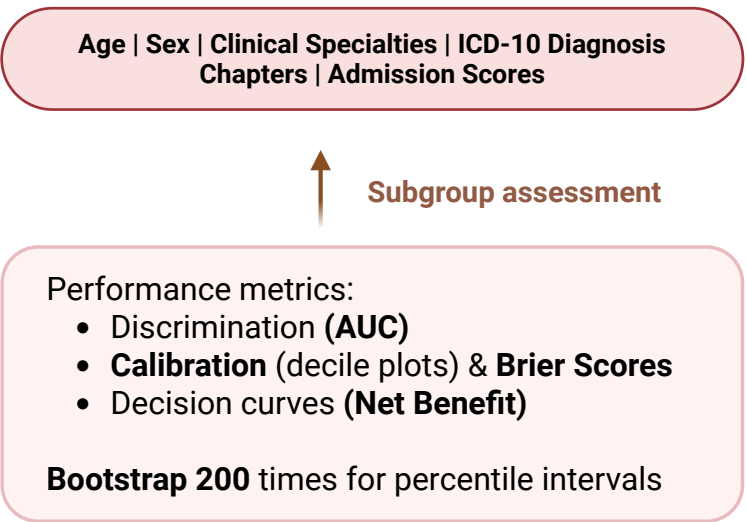
**Method:** Covariate Balancing Propensity Score (CBPS)

**Purpose:** Isolates predictive performance from intervention effects on outcomes

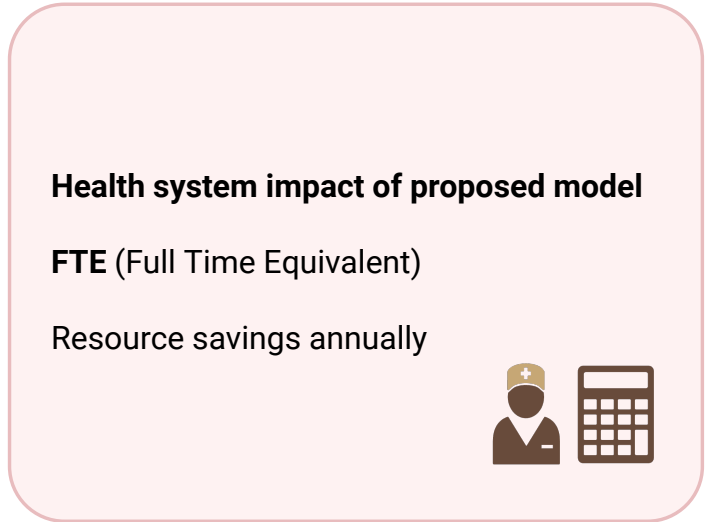
D

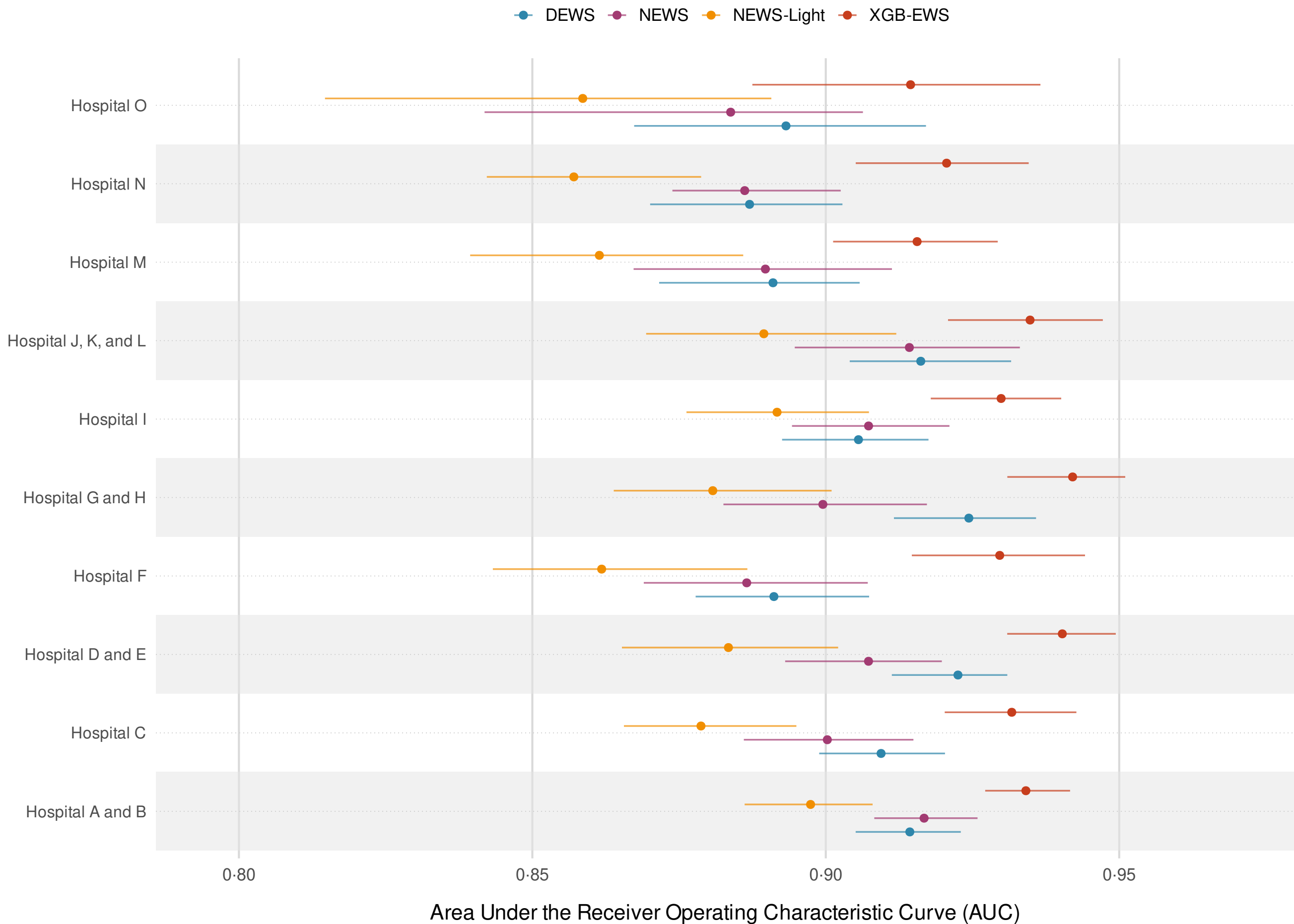


E



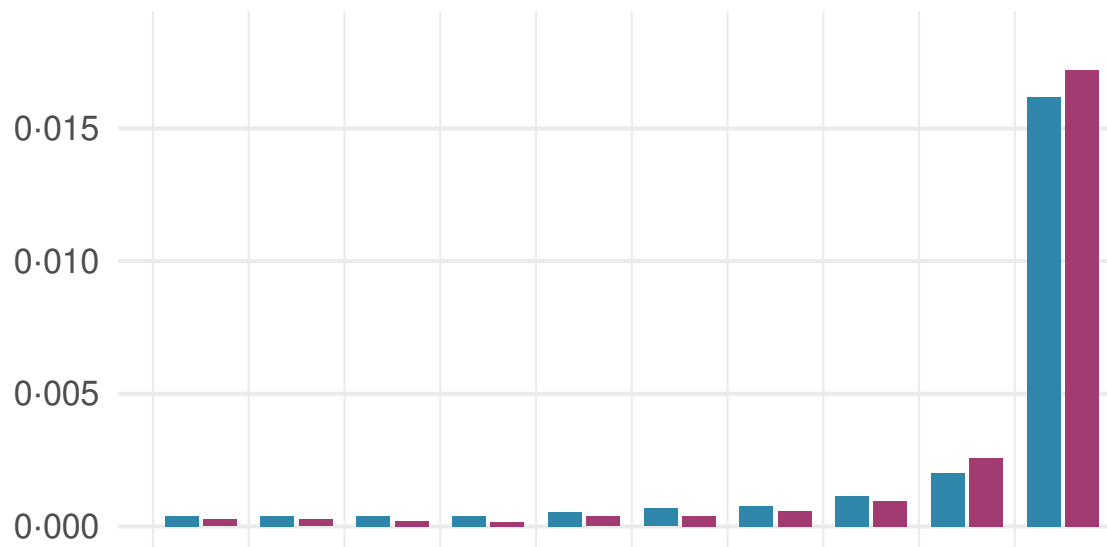
F



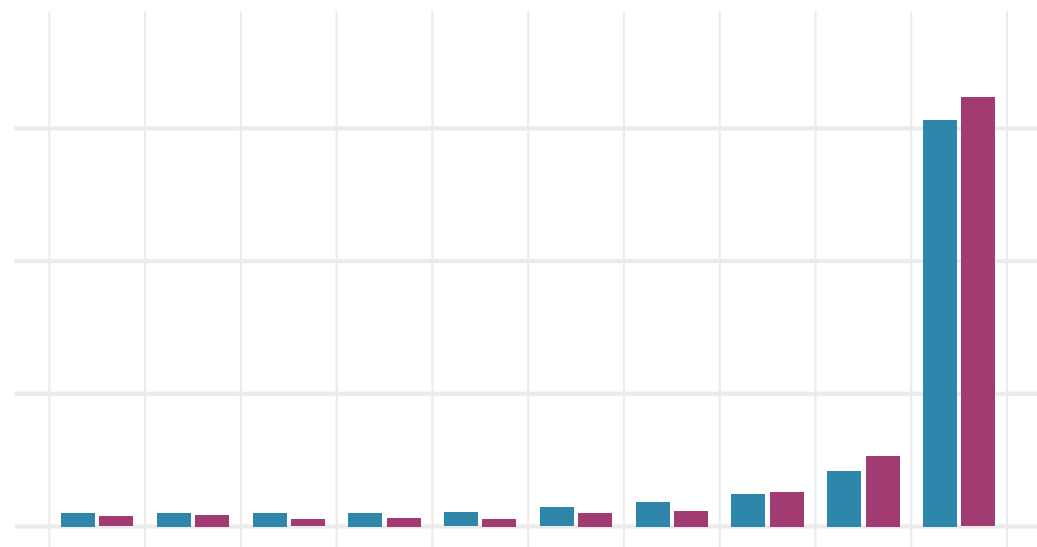


Group Predicted Observed

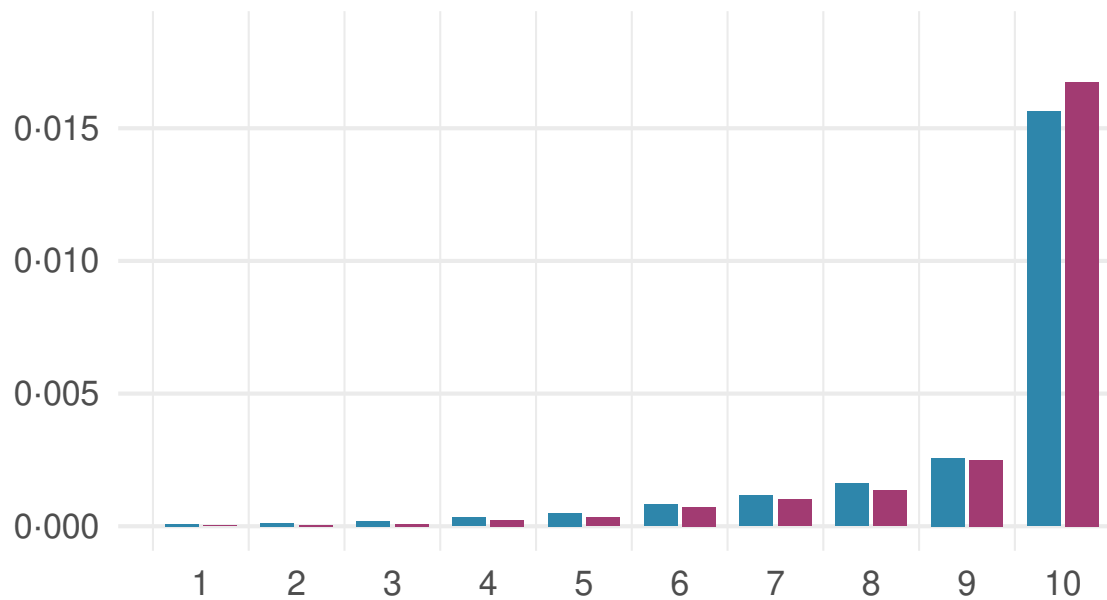
NEWS



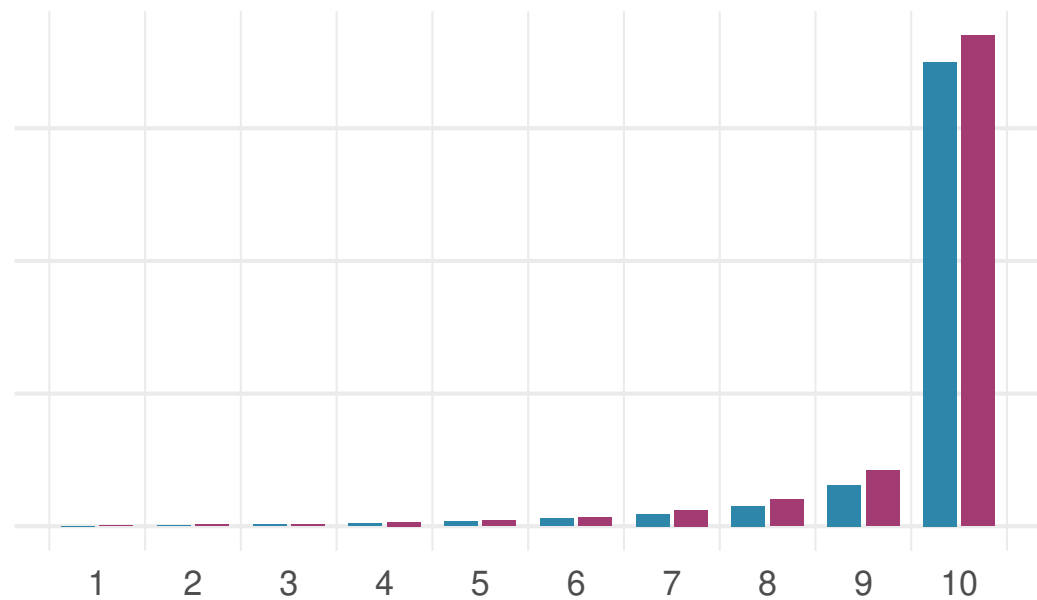
NEWS-Light



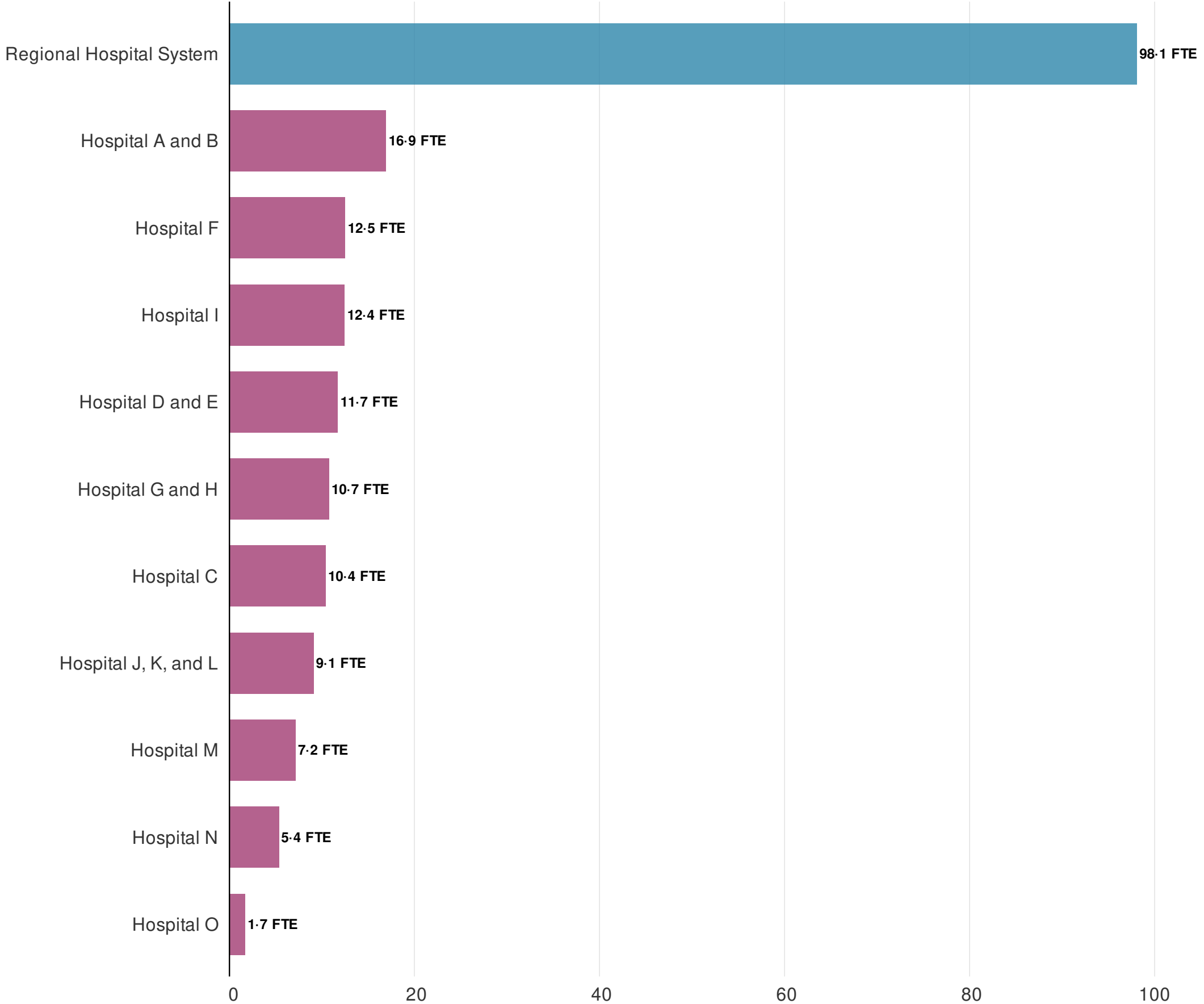
DEWS



XGB-EWS



Deciles of Predicted Risk



FTE Saved Annually using NEWS-Light