PHILIPS

eICU Program

White paper

Introduction to the Philips Critical Care Outcomes Prediction models – Mortality prediction

The purpose of this white paper is:

- To describe the methods for developing the Philips Critical Care Outcomes Prediction models
- To provide a comprehensive review of the performance statistics in comparison with APACHE IVa and IVb
- To provide a basis for obtaining feedback from eICU partners regarding the model design and evaluation

Executive summary

Background: Since 2005, Philips has been providing benchmarking reporting for customers using Cerner's release of Acute Physiology and Chronic Health Evaluation (APACHE) IVa prediction models (Cerner Corporation, Kansas City, MO) to obtain severity-adjusted predictions for ICU outcomes.[1-4] Evidence has pointed to limitations in this model, including performance decay over time and sub-optimal calibration of APACHE models in the eICU customer install base.[5]

Philips has developed a new set of benchmarking models and this work focuses on mortality prediction in the hospital and ICU using a recent cohort of eRI (eICU Research Institute data). The objectives for this initiative were to:

- Reduce the documentation burden to obtain mortality risk predictions
- Reduce the bias introduced through variations in documentation practice
- Achieve comparable accuracy performance as APACHE IVa and IVb
- Improve calibration performance in comparison with APACHE IVa and IVb
- Develop a system for routine recalibration to eliminate issues with model decay over time

Methods

The Philips Critical Care Outcomes ICU and hospital mortality models were developed using 2017-2018 eRI data. A manuscript detailing the development of the models is in process. This white paper further describes the secondary validation of the models among the entire eICU benchmarking cohort. We applied the models to the entire eICU customer install base with APACHE licenses for all years available (2004-2019). The IVb analysis was restricted to the years 2014-2019, as Cerner Corporation (Kansas City, MO) restricted APACHE IVb to years from 2014 forward. We examined the behavior and performance of the three models (APACHE IVa/IVb, and Philips Critical Care Outcomes Prediction models) for mortality and length of stay.

Results

The Philips Critical Care Outcomes Prediction models displayed the following traits when compared with APACHE IVa/IVb for mortality prediction after IVa/IVb:

- The Philips Critical Care Outcomes Prediction models showed significantly higher model discriminative performance over APACHE IVa/IVb (AUROCs and 95% confidence intervals)
  - Restricted years 2014-2019 (IVa vs. IVb vs. the Philips Critical Care Outcomes Prediction models)
    - ICU mortality AUROC: 0.884 [0.884,0.887] vs. 0.886 [0.886,0.888] vs. 0.926 [0.926,0.928]
    - Hospital mortality AUROC: 0.864 [0.864,0.865] vs. 0.864 [0.864,0.866] vs. 0.905 [0.905,0.906]
  - All years 2004-2019 (IVa vs. the Philips Critical Care Outcomes Prediction models)
    - ICU mortality AUROC: 0.881 [0.882,0.883] vs. 0.922 [0.922,0.924]
    - Hospital mortality AUROC: 0.860 [0.860,0.862] vs. 0.901 [0.900,0.902]
- The Philips Critical Care Outcomes Prediction models showed improved calibration against the actual mortalities than APACHE IVa/IVb
  - Restricted years 2014-2019 (IVa vs. IVb vs. the Philips Critical Care Outcomes Prediction models)
    - ICU mortality A:P ratios: 0.756 vs. 0.874 vs. 1.020
    - Hospital mortality A:P ratios: 0.732 vs. 0.944 vs. 1.017
  - All years 2004 -2019 (IVa vs. the Philips Critical Care Outcomes Prediction models)
    - ICU mortality A:P ratios: 0.749 vs. 0.975
    - Hospital mortality A:P ratios: 0.740 vs. 1.021
- The Philips Critical Care Outcomes Prediction models showed consistently improved performance across a variety of subgroups including ICU types, admission diagnoses, admission sources and years (see Table 2, 3 ,4 /Figure 3.1, 3.2, 3.3, 4.1, 4.2, and Appendix Figure 2 for details)
- The Philips Critical Care Outcomes Prediction models were able to mitigate abrupt changes in Glasgow Coma Scale documentation practice in two health systems compared to APACHE IVa that produced dramatically altered predictions (Figure 5.1, 5.2).

In summary, the Philips Critical Care Outcomes Prediction models had very good accuracy and calibration[6] when applied to all available eICU data between 2004 and 2019 with the following statistics:

	AUROC*	A:P Ratio**
ICU Mortality	0.92	0.98
Hospital Mortality	0.90	1.02

* AUROC = area under the receiver operating characteristic curve.
**A:P Ratio = actual deaths divided by the predicted number of deaths.

Table 1. Showing ICU and hospital mortality for the Philips Critical Care Outcome Prediction Models.

The above table shows the results for the Philips models. In comparison, the Apache IVa values are: AUROC: 0.881 for ICU mortality and AUROC: 0.860 for Hospital mortality. The Philips models provided an improvement for both values.

Introduction

How Philips uses APACHE models

Philips Healthcare has been providing clinical outcomes reports using the APACHE risk models for quarterly reporting since 2005.[1] Additionally, APACHE algorithms have been incorporated into eSearch since eSearch v3.0 was released in 2009. The model predictions are all based on first-day patient characteristics and represent the expected outcomes according to the APACHE national equations methodology.

Changes in medical practice and overall population health warrant review and recalibration of predictive models over time, as models may become less accurate. Although APACHE has recalibrated its predictive algorithms, the release of APACHE model IVa was developed in a cohort of patients assembled in 2006-2008,[1] and the IVb model was trained in a cohort of patients in 2014-2015.[2] Historically, mortality rates are traditionally lower in the eICU population compared with the APACHE cohorts.[5]

This white paper introduces the process of developing the Philips Critical Care Outcomes Prediction models using Philips eICU program customer data, reveals the new model's performance relative to APACHE IVa/IVb, and aims to engage a broadened customer base for feedback to improve the model continually.

The Philips Critical Care Outcomes Prediction model development

In order to develop the Philips Critical Care Outcomes Prediction models to meet the need for benchmarking, we have:

- Defined baseline patient features to draw information from both structured and unstructured data sources, with the aim to accurately reflect the actual patient status at the time of ICU admission.
- Used machine learning modeling techniques to harness the predictive values of the patient baseline characteristics in the eICU patient cohort and, subsequently, serve as an accurate risk adjustment for benchmarking.
- Designed the Philips Critical Care Outcomes Prediction models to be updated frequently, so they reflect the current performance of ICUs in the eICU install base.
- Designed a novel ICU stay definition to capture independent and clinically relevant ICU stays and improve handling of data anomalies.
- Defined a simple time window definition of the patient admission baseline.
- Selected model features by considering clinical domain knowledge, complexity of data collection and a data-driven approach. High priority was given to objective measurements not requiring manual data entry. Features such as urinary output, active treatments, and chronic conditions are not required.
- Engineered variables with a high risk for misclassification into features designed to reduce the potential for bias, such as admission diagnosis groups and GCS score.
- Used an advanced modeling technique allowing vital sign patterns to be associated with different risks based on the patient's admission diagnosis.

The patient cohort for model development

We used the eICU Research Institute (eRI) database that houses all the historical data collected from participating customers for the model development and validation. After excluding ICUs that did not use eCareManager and have not maintained reliable data flow, we included all patient unit stays discharged from the hospital in two years between 1/1/2017 and 12/31/2018 within the eRI database.

We examined the documentation pattern of the unit stays in the eCareManager system and implemented criteria to define independent unit stays more reflective of the real patient unit stays, with the following rules applied:

- Excluded patient unit stays not classified as "ICU" stays
- Excluded patient unit stays with irrational admission/discharge time stamps
- Combined back-to-back or overlapping stays
- Allowed patients to be briefly moved out of ICU for surgical operation and immediately readmitted to the ICU; treat the two adjacent ICU stays as one continuous ICU stay
- Updated conflicting admission/discharge information using all data available

Based on the reconciled ICU patient unit stays, we excluded stays with a length of stay <4 hours or >365 days, or patients <16 years of age.

The baseline time window

We established the first 24 hours of ICU admission as the time window to represent patients' baseline risk; we also included data points up to six hours before admission when no data was available in the first 24 hours of ICU admission.

Model features

The list of features included vital signs, critical laboratory measures, and essential user-documented data of admission diagnosis and GCS scores.

We extracted information from all possible sources from the eRI, including structured and unstructured data, reconciled the conflicts and errors, and generated summarized statistics of individual inputs as model features. The only exception, is that we did not use data charted in the eCareManager Patient Registry, which was not part of the de-identified eRI database at the time of this study.

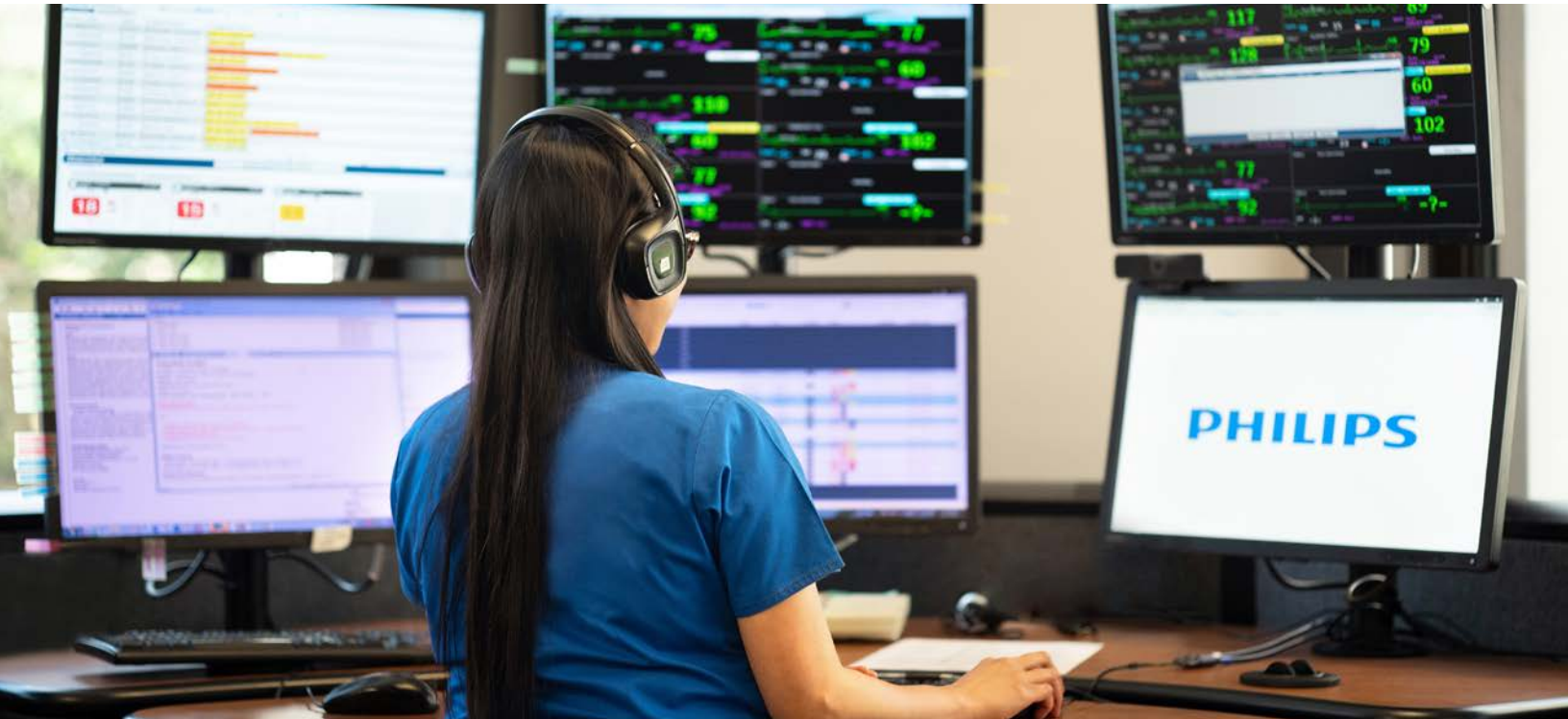
We have required a list of variables that are commonly measured at ICU admission to be present to be eligible for prediction, while allowing less frequently measured variables to be missing.

We have made specific adaptations of the feature definition to accommodate known issues of documentation in the eCareManager system, for example:

- The GCS score – We confirmed in the data the existence of the potentially-biased documentation of GCS score, mainly when sites vary in their approach to documenting patients under sedation or mechanical ventilation as 'not responding' (lowest GCS score) compared to 'unable to score GCS due to medication/ other.' We made accommodations in our GCS feature design to mitigate the impact of these practice variations.
- The admission diagnosis – We also have examined the unique admission diagnosis strings as documented in our system and regrouped them to reflect novel patient subgroups, which are clinically similar within the group and remain relatively stable over time.

Modeling techniques

We developed the Philips Critical Care Outcomes Prediction models using the generalized additive model (GAM) framework, which is an extension of a standard linear model by allowing non-linear functions of continuous predictors while maintaining the additivity of multivariate linear regression. We also included a nested random effect for diagnosis groups, along with other essential interactions between features.



Model validation

The ICU and hospital mortality models have been validated in the eRI cohort. A manuscript is in preparation for peer review describing, in detail, the development and validation within the eRI cohort.

To further validate the model developed from eRI data, we applied the final models to the eCareManager archived databases used for quarterly benchmarking. The Philips Critical Care Outcomes Prediction model was available for all health systems regardless of their APACHE licensing status.

To make a side-by-side comparison of the Philips Critical Care Outcomes Prediction model to APACHE, we identified a cohort of patient unit stays that could be directly linked between APACHE stays and the new combined stays. The majority of the linked stays reflect the same patient unit stays. As the new rules were implemented to define clinically-relevant ICU stays better, they may slightly differ from the APACHE rules.

We limited the primary analysis among the subgroup for which APACHE IVa, IVb, and the Philips Critical Care Outcomes Prediction model predictions all produced valid predictions (2014-2019) because APACHE IVb is not available before 2014. To directly compare IVa and the Philips Critical Care Outcomes Prediction model, we expanded the year limit to 2004-2019, including patient unit stays with both IVa and the Philips Critical Care Outcomes Prediction model producing valid prediction.

We assessed the model discrimination using the area under the receiver operating characteristic curve (AUROC). We evaluated the model calibration by the actual/predicted ratio. We repeated the analysis in essential patient subgroups (e.g., unit type, admission source, admission diagnosis) and by each hospital year/quarter active in the eCareManager archive database. Each eICU will receive a custom report comparing model performance in its population.

To illustrate the models' robustness against changes in GCS documentation practice, we have approached health systems that have confirmed significant changes in the GCS documentation practice. With their consent, we analyzed and presented the model calibration performance by the quarterly average, median, and inter-quartile range (IQR) of predicted hospital mortality, and the discriminative performance by AUROCs according to APACHE IVa and the Philips Critical Care Outcomes Prediction model before and year after the change in documentation practice. We could not select IVb in this analysis because APACHE IVb was not available before 2014. Given that APACHE IVb was a simple recalibration of IVa, we expect IVb to behave similarly to IVa when challenged by a shift in a significant feature documentation pattern.

Results

ICU stay cohorts

From the eCareManager archived database, we identified 5,289,859 patient unit stays from unit years in which there was at least one valid APACHE IVa prediction, representing:

- 46 health systems
- 420 hospitals
- 732 ICUs
- 16 years (2004-2019, IVb only available from 2014-2019)

From the 5,289,859 patient unit stays:

- APACHE identified 4,412,156 APACHE unit stays
- The Philips Critical Care Outcomes Prediction model identified 4,195,994 independent unit stays, reflecting the reconciled definition of an independent ICU stay.

Missing predictions in different ICU stay cohorts

Predictions cannot be generated for any of the models if a required data element is missing.

Out of the 4,412,156 APACHE stays (2004-2019):

- 3,988,109 (90.3%) received IVa ICU mortality prediction
- 3,768,239 (85.4%) received IVa hospital mortality prediction

Out of the 2,831,455 APACHE stays in the years IVb was available (2014-2019):

- 2,619,117 (92.5%) received IVb ICU mortality prediction
- 2,470,692 (87.3%) received IVb hospital mortality prediction

Out of the 4,195,994 new combined patient stays (2004-2019):

- 3,597,283 (85.7%) received valid ICU and hospital predictions

We have identified the following causes for discrepancies in the percentage of the unit stays scored for APACHE IVa and the Philips Critical Care Outcomes Prediction model:

- The difference of patient unit stay filtering and recombination rules between APACHE and the Philips Critical Care Outcomes Prediction model
- The Philips Critical Care Outcomes Prediction model requirement for admission height, weight, and commonly measured laboratory studies for mortality prediction while APACHE IVa predicts mortality despite all laboratory values “missing”

Detailed requirements for the Philips Critical Care Outcomes Prediction models inputs are in Appendix Table 1.1 and 1.2.

The cohort for a side-by-side comparison of APACHE IVa/IVb and the Philips Critical Care Outcomes Prediction models

To directly compare APACHE IVa and the Philips Critical Care Outcomes Prediction models, we identified a group of patient stays for which we could directly match the APACHE stays with the new combined stay (defined by and used in the Philips Critical Care Outcomes Prediction models) by the unique patient unit stay ID.

Out of 4,091,932 patient unit stays with a one-to-one match between APACHE stays and new combined stays:

- 3,144,009 stays matched for APACHE IVa and the Philips Critical Care Outcomes Prediction models with valid predictions (ICU and hospital mortality predictions, all years from 2004 to 2019)
- 2,081,163 stays matched for APACHE IVa, IVb and the Philips Critical Care Outcomes Prediction models with valid predictions (ICU and hospital mortality predictions, restricted years from 2014 to 2019)

The primary analysis comparing APACHE IVa, IVb, and the Philips Critical Care Outcomes Prediction models was done in the cohort of 2,081,163 one-to-one matched stays with all predictions available.

The impact of GCS on model performance was made in the cohort of one-to-one matched stays with both IVa and Philips Critical Care Outcomes model prediction available. We did not directly compare the performance of IVb and the Philips Critical Care Outcomes Prediction models given that some of the GCS documentation pattern change may have happened before the year 2014 when APACHE IVb was not available.

The specific patient cohorts used by this analysis are in Appendix Figure 1. In addition to the summary results presented in this white paper, each health system will receive an eICU-specific set of data, in which we also examined the model performance among hospitals of each health system.

Model performance

The AUROCs and the calibration measured by actual: predicted ratios for the Philips Critical Care Outcomes Prediction models were higher, or better than that of APACHE IVa, and IVb in the final cohort of 2,081,163 patient unit stays with all predictions available (2014-2019).

Side by side comparison of model performance: APACHE IVa, IVb and the Philips Critical Care Outcomes models

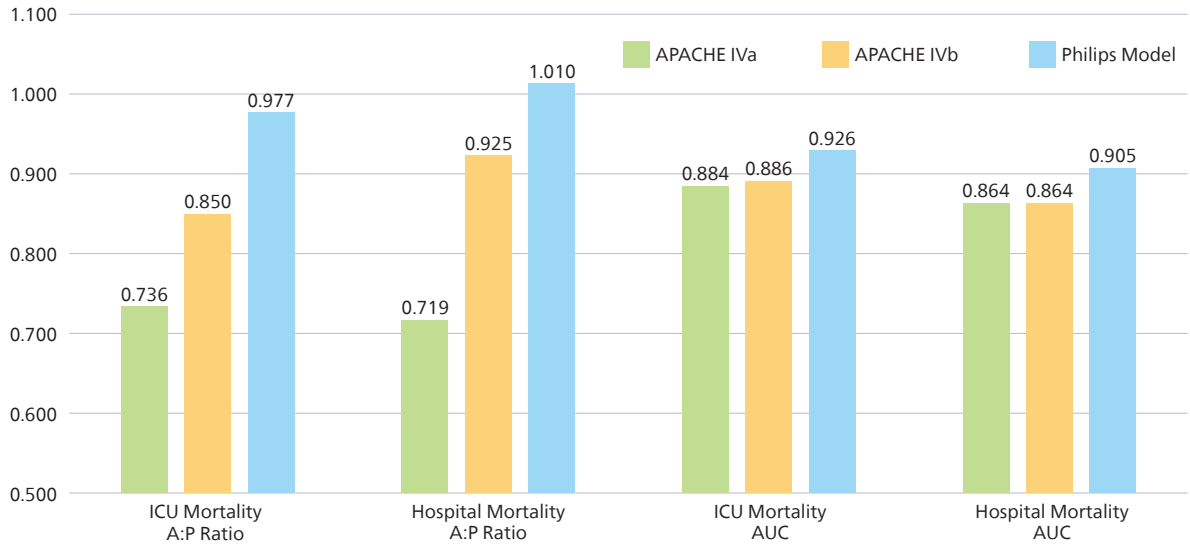


Figure 1. Model performance of APACHE IVa, IVb, and the Philips Critical Care Outcomes Prediction models (2014-2019).

Similar findings were shown below when comparing APACHE IVa and the Philips Critical Care Outcomes Prediction models among the expanded cohort of 3,144,009 patient unit stays with both IVa and the Philips Critical Care Outcomes Prediction models prediction available (2004-2019).

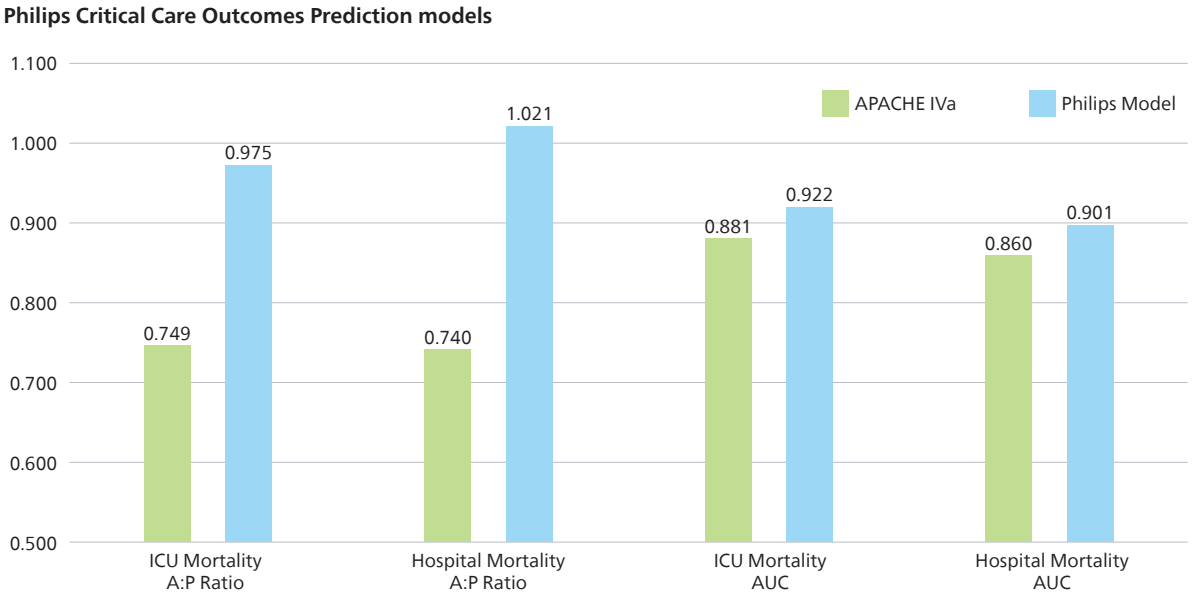


Figure 2. Model performance of APACHE IVa and the Philips Critical Care Outcomes Prediction models (2004-2019).

Among the 2,081,163 one-to-one linked stays that had all predictions available (APACHE IVa, IVb, and Philips Critical Care Outcomes Prediction models, ICU/Hospital), we examined the model performance in the following subgroups:

- By admission diagnosis strings (Figure 3.1-3.2)
- By admission diagnosis groups (Figure 3.3, appendix figure 2)
- By patient unit stay type (Table 2)
- By ICU admission source (Table 3)
- By hospital discharge year/quarter (Table 4 and Figure 4.1-4.2)

We have observed that the Philips Critical Care Outcomes Prediction model's discriminative and calibration performance was higher than that of APACHE IVa and IVb, consistently in the majority of subgroups, as defined above (detailed tabulation of the model performance by subgroups are in the files going to each health system).

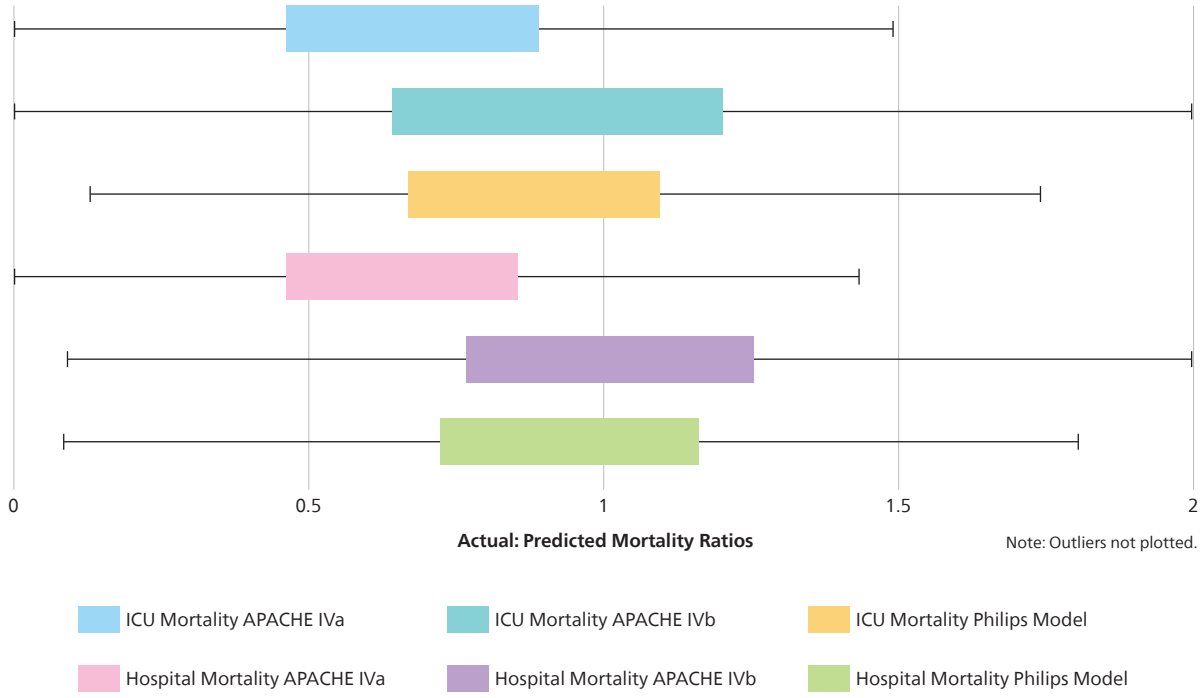


Figure 3.1 Model calibration performance (Actual/Predicted ratios) by diagnosis strings.*

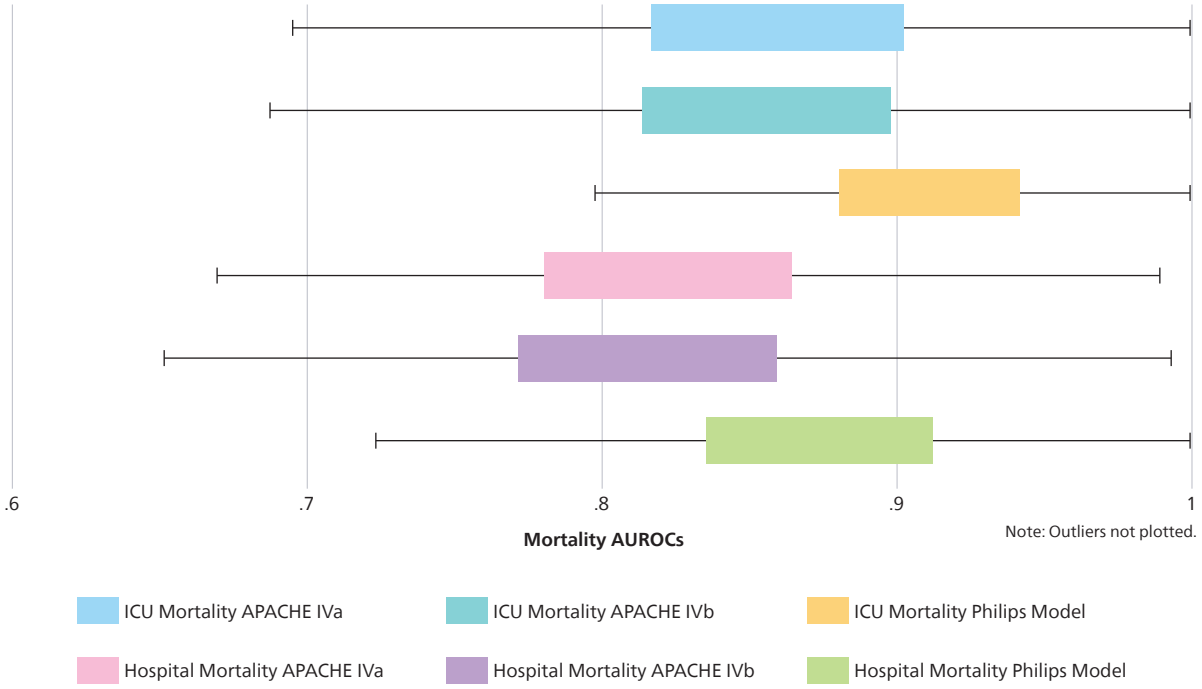


Figure 3.2 Model discriminative performance (AUROCs) by diagnosis strings.*

*Among 2,081,163 patient stays with predictions available for each model.

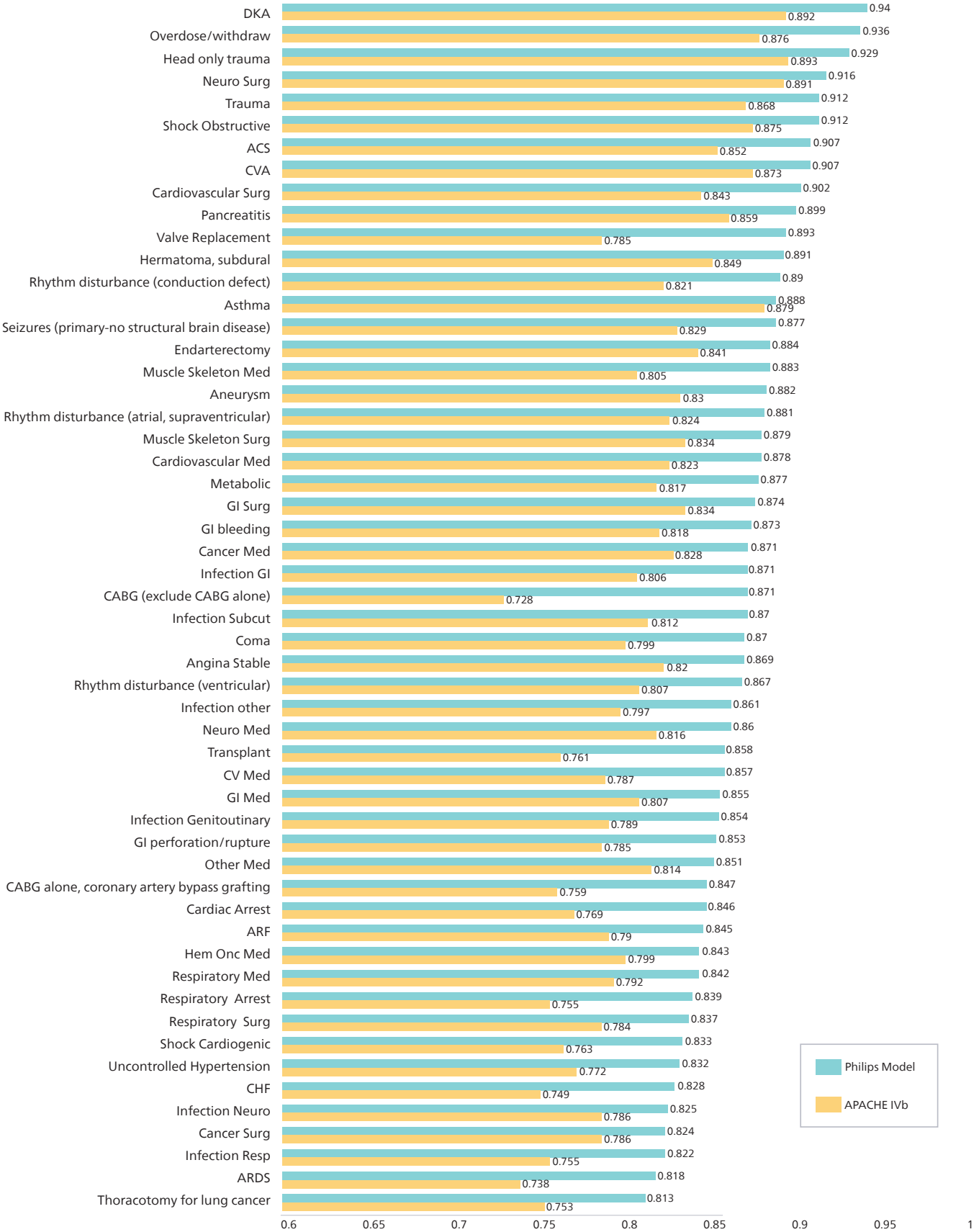


Figure 3.3 Model performance (Hospital mortality AUROCs) by admission diagnosis groups: APACHE IVb vs. the Philips Critical Care Outcomes Prediction model (2014-2019).

		ICU mortality						Hospital mortality					
		AUROC			A:P ratio			AUROC			A:P ratio		
Unit types	N	IVa	IVb	Philips	IVa	IVb	Philips	IVa	IVb	Philips	IVa	IVb	Philips
Burn-Trauma ICU	382	0.867	0.871	0.950	1.119	1.424	1.741	0.876	0.859	0.925	1.123	1.422	1.524
CCU-CTICU	95075	0.890	0.891	0.931	0.776	0.867	0.929	0.874	0.873	0.914	0.769	0.964	1.000
CSICU	51718	0.885	0.886	0.945	0.800	0.930	1.081	0.866	0.865	0.919	0.684	0.881	0.981
CTICU	53493	0.880	0.879	0.928	0.830	1.000	1.000	0.866	0.862	0.907	0.767	1.037	1.000
Cardiac ICU	133408	0.892	0.892	0.931	0.808	0.900	1.050	0.872	0.872	0.911	0.772	0.969	1.056
MICU	252062	0.869	0.870	0.912	0.750	0.828	0.960	0.851	0.851	0.893	0.783	0.960	1.082
Med-Surg ICU	1147915	0.883	0.884	0.926	0.733	0.846	0.965	0.860	0.860	0.902	0.711	0.905	0.977
Neuro ICU	139019	0.898	0.902	0.932	0.680	0.864	1.159	0.877	0.879	0.915	0.661	0.944	1.151
SICU	181067	0.891	0.891	0.929	0.672	0.818	0.978	0.867	0.865	0.907	0.667	0.889	1.014
Trauma ICU	25722	0.899	0.896	0.935	0.712	0.940	1.205	0.888	0.883	0.920	0.690	0.945	1.113
Vascular ICU	1302	0.923	0.918	0.937	0.696	0.750	1.000	0.888	0.890	0.898	0.694	0.855	0.952

Footnote: model performance matrices (AUROCs) were colored formatted separately for ICU and hospital mortality, with minimum value colored red, maximum value colored green, and median value colored yellow. N= number of patient stays.

Table 2. Model performance (AUROCs and Actual/Predicted ratios) by unit types (2014-2019).

		ICU mortality						Hospital mortality					
		AUROC			A:P ratio			AUROC			A:P ratio		
Admission source	N	IVa	IVb	Philips	IVa	IVb	Philips	IVa	IVb	Philips	IVa	IVb	Philips
Acute Care/Floor	145614	0.834	0.835	0.891	0.858	0.919	0.958	0.812	0.814	0.864	0.787	1.000	0.986
Chest Pain Center	3411	0.915	0.909	0.942	0.667	0.667	0.900	0.904	0.901	0.926	0.591	0.722	0.929
Direct Admit	141638	0.888	0.888	0.923	0.716	0.840	1.033	0.861	0.861	0.899	0.720	0.941	1.044
Emergency Department	1138573	0.891	0.892	0.931	0.679	0.791	0.964	0.868	0.869	0.910	0.689	0.866	1.012
Floor	127700	0.833	0.834	0.886	0.858	0.919	0.968	0.814	0.816	0.862	0.789	1.014	1.014
ICU	196	0.778	0.793	0.869	1.079	1.148	1.116	0.795	0.811	0.886	1.121	1.340	1.209
Observation	556	0.917	0.940	0.933	0.600	0.679	1.029	0.895	0.887	0.915	0.462	0.636	0.845
Operating Room	320739	0.878	0.873	0.932	0.676	1.000	0.926	0.853	0.851	0.903	0.650	1.000	0.975
Other	711	0.917	0.916	0.929	0.447	0.507	0.618	0.909	0.913	0.920	0.444	0.566	0.636
Other Hospital	63892	0.854	0.855	0.887	0.895	0.927	1.052	0.826	0.825	0.861	0.841	0.987	1.057
Other ICU	303	0.843	0.854	0.903	0.728	0.937	0.881	0.827	0.827	0.882	0.730	1.055	0.975
PACU	51160	0.895	0.885	0.934	0.696	1.067	0.889	0.853	0.852	0.894	0.652	1.154	0.909
Recovery Room	45150	0.898	0.889	0.930	0.720	1.059	0.947	0.846	0.853	0.887	0.694	1.172	1.000
Step-Down Unit (SDU)	41517	0.816	0.818	0.880	0.939	1.000	1.070	0.803	0.804	0.854	0.865	1.092	1.071

Footnote: model performance matrices (AUROCs) were colored formatted separately for ICU and hospital mortality, with minimum value colored red, maximum value colored green, and median value colored yellow. N= number of patient stays.

Table 3. Model performance (AUROCs and Actual/Predicted ratios) by ICU admission source.

		ICU mortality						Hospital mortality					
		AUROC			A:P ratio			AUROC			A:P ratio		
Year quarter	N	IVa	IVb	Philips	IVa	IVb	Philips	IVa	IVb	Philips	IVa	IVb	Philips
2014Q1	62769	0.879	0.881	0.919	0.808	0.926	1.016	0.862	0.861	0.900	0.776	0.990	1.043
2014Q2	69493	0.879	0.880	0.922	0.699	0.823	0.911	0.860	0.861	0.902	0.686	0.890	0.942
2014Q3	71890	0.883	0.885	0.927	0.714	0.833	0.909	0.859	0.861	0.903	0.711	0.931	0.976
2014Q4	75716	0.883	0.884	0.925	0.736	0.855	0.930	0.860	0.861	0.902	0.712	0.923	0.966
2015Q1	81844	0.881	0.883	0.919	0.756	0.868	0.967	0.854	0.853	0.894	0.740	0.949	1.000
2015Q2	78146	0.886	0.887	0.926	0.712	0.825	0.945	0.862	0.862	0.903	0.706	0.913	1.000
2015Q3	77248	0.889	0.890	0.931	0.712	0.839	0.945	0.867	0.868	0.909	0.709	0.912	1.000
2015Q4	82095	0.885	0.885	0.926	0.747	0.875	0.982	0.865	0.866	0.905	0.719	0.935	1.012
2016Q1	87033	0.881	0.881	0.923	0.759	0.870	1.000	0.859	0.859	0.902	0.754	0.960	1.056
2016Q2	87435	0.890	0.891	0.928	0.707	0.815	0.964	0.869	0.868	0.905	0.700	0.903	1.000
2016Q3	88665	0.890	0.891	0.930	0.726	0.841	0.981	0.868	0.868	0.908	0.709	0.912	1.012
2016Q4	91817	0.887	0.886	0.926	0.750	0.864	1.000	0.865	0.865	0.906	0.740	0.948	1.046
2017Q1	97808	0.879	0.881	0.921	0.737	0.843	1.000	0.856	0.857	0.898	0.734	0.931	1.044
2017Q2	96297	0.887	0.888	0.928	0.733	0.846	1.000	0.866	0.865	0.904	0.711	0.915	1.024
2017Q3	95621	0.891	0.892	0.932	0.730	0.844	1.000	0.873	0.873	0.913	0.712	0.913	1.012
2017Q4	98394	0.884	0.884	0.928	0.740	0.851	1.000	0.864	0.864	0.905	0.732	0.938	1.023
2018Q1	100491	0.877	0.880	0.926	0.744	0.859	1.000	0.859	0.859	0.903	0.746	0.942	1.043
2018Q2	101775	0.886	0.887	0.928	0.720	0.831	0.964	0.867	0.868	0.910	0.694	0.894	0.988
2018Q3	101518	0.889	0.890	0.931	0.708	0.823	0.962	0.869	0.870	0.911	0.707	0.911	1.012
2018Q4	104593	0.882	0.882	0.926	0.730	0.844	0.982	0.866	0.865	0.907	0.731	0.946	1.036
2019Q1	103646	0.876	0.877	0.922	0.750	0.864	0.966	0.856	0.857	0.902	0.732	0.938	1.011
2019Q2	95119	0.884	0.885	0.927	0.726	0.841	0.964	0.861	0.860	0.903	0.695	0.891	0.976
2019Q3	94553	0.889	0.891	0.930	0.750	0.871	1.019	0.868	0.869	0.909	0.698	0.900	0.988
2019Q4	37197	0.894	0.895	0.933	0.733	0.833	1.019	0.874	0.874	0.913	0.672	0.863	0.965

Footnote: model performance matrices (AUROCs) were colored formatted separately for ICU and hospital mortality, with minimum value colored red, maximum value colored green, and median value colored yellow. N= number of patient stays.

Table 4. Model performance (AUROCs and Actual/Predicted ratios) by hospital discharge year-quarter.

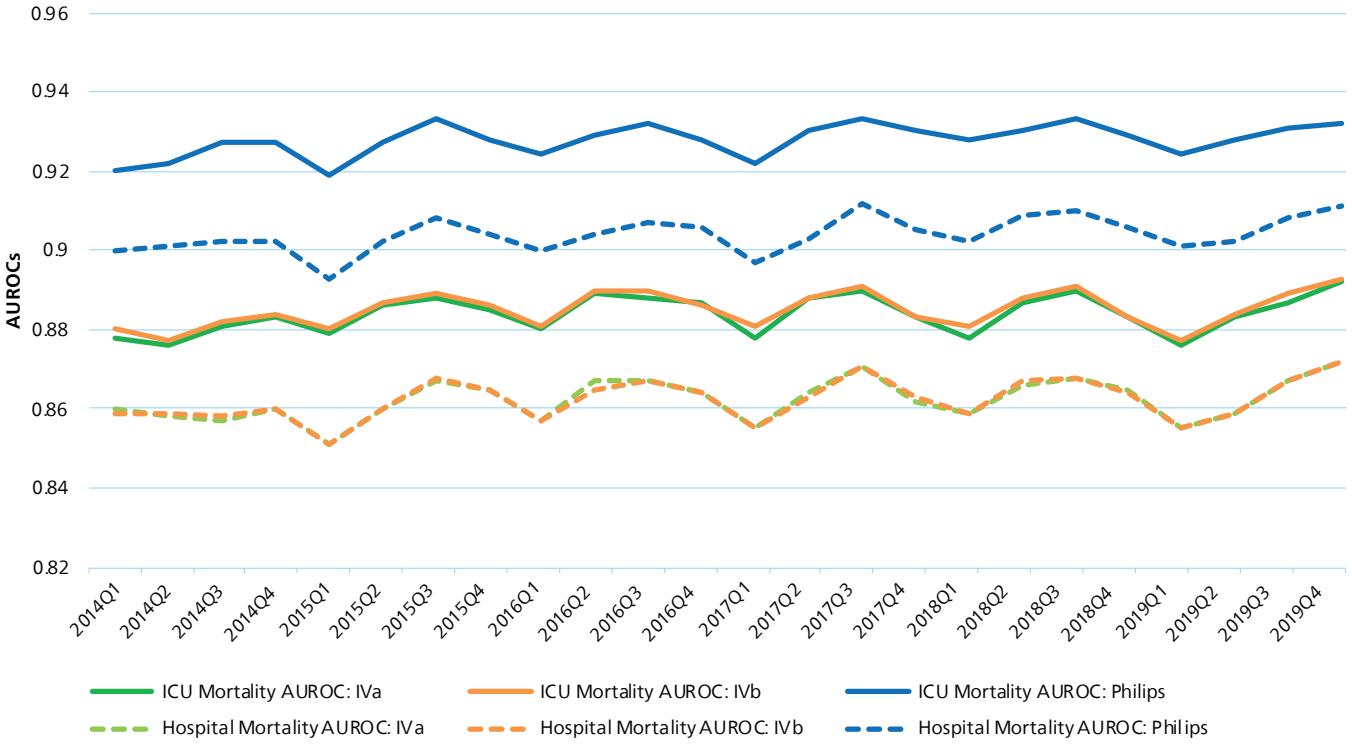


Figure 4.1 Model performance (AUROCs) by hospital discharge year-quarter.

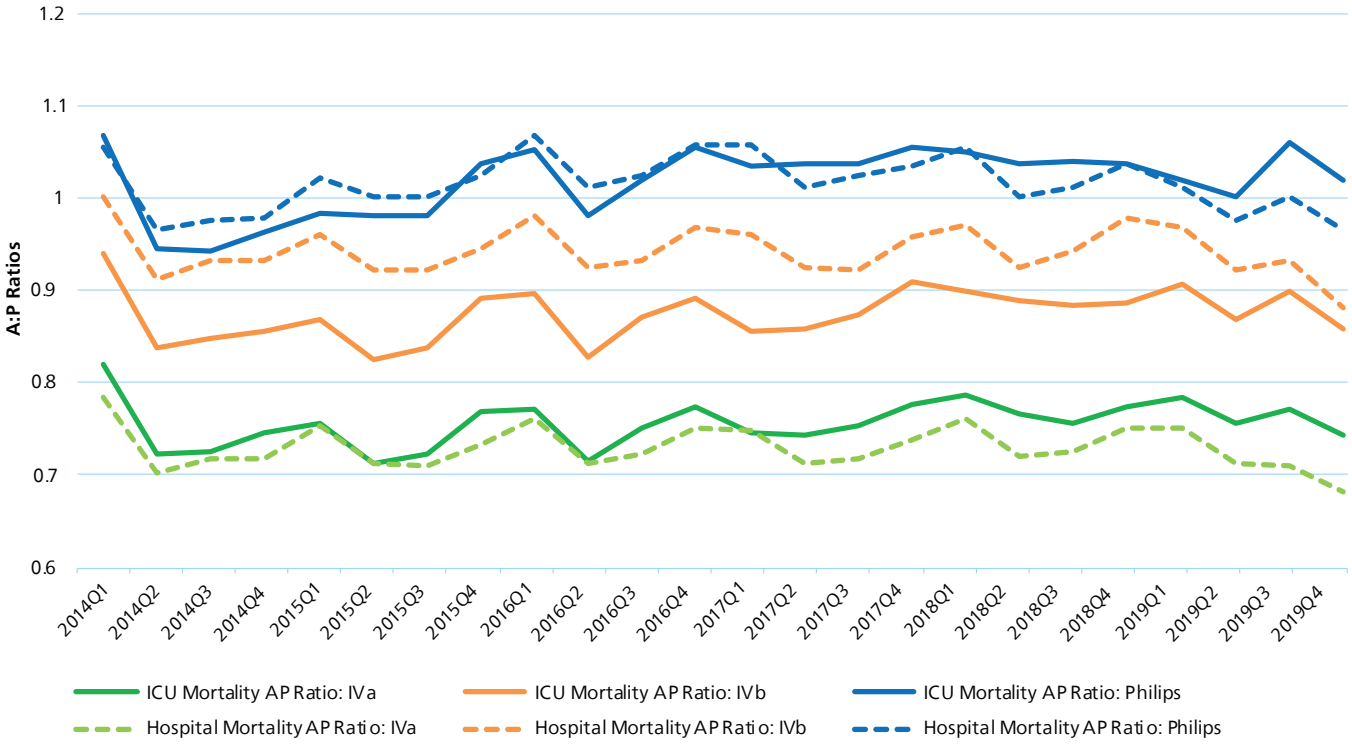


Figure 4.2 Model performance (Actual/Predicted Ratios) by hospital discharge year-quarter.

Model's performance against significant documentation pattern change

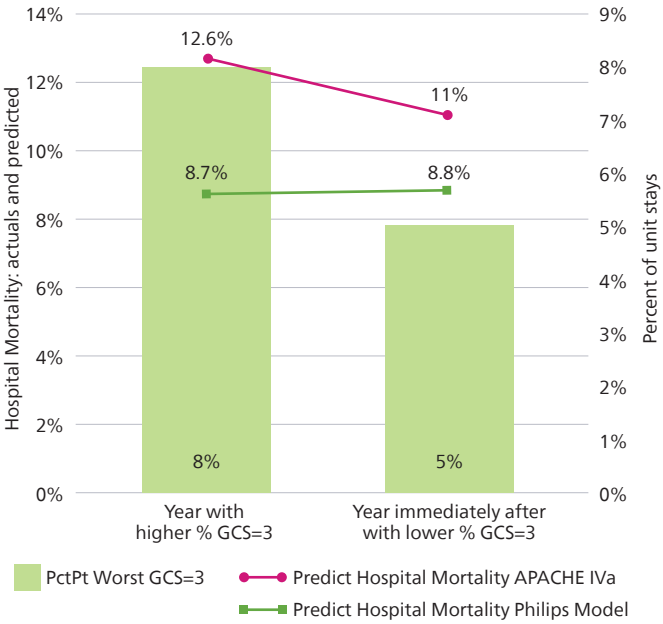
One of the most prominent features used by ICU risk models is the GCS score.[1-4] Critical as it is for patient severity assessment, GCS is heavily dependent on the reliable evaluation of the patient's neurologic status, and we have noticed a varying pattern of documentation in our customer base.

With customer consent, we identified two health systems that have gone through significant alterations in their practice of GCS assessment and confirmed the change of GCS scores on the health system level before and after the transition.

- In Health System A, the change was an unintended consequence of an updated EMR interface. Health System B made a deliberate change in their GCS documentation practice. Both changes resulted in a significantly decreased percentage of patients receiving the lowest GCS score of 3.
- We identified one year before the transition as the 'before' and one year after as the 'after', given that the health systems changed the practice rather quickly.

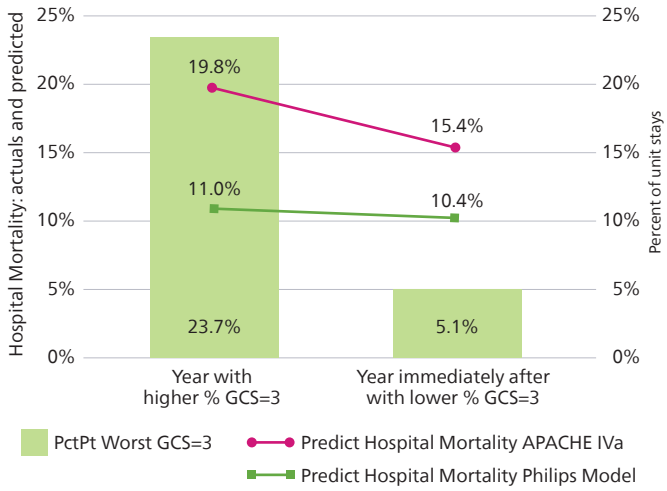
Through this analysis, we have confirmed that:

- The Philips Critical Care Outcomes Prediction model had better model discriminative performance in terms of AUROC, before and after the GCS change.
  - Health system A: AUROCs = 0.91 vs. 0.84 for the Philips Critical Care Outcomes Prediction model vs. IVa, respectively.
  - Health system B: AUROCs = 0.90 vs. 0.81 for the Philips Critical Care Outcomes Prediction model vs. IVa, respectively.
- More importantly, we have observed that the Philips Critical Care Outcomes Prediction model appeared to mitigate the abrupt changes in GCS scores assessment with better calibration (Figure 5.1-5.2).
  - APACHE IVa hospital mortality predictions fluctuated significantly along with GCS scores.
  - In contrast, differences in the Philips Critical Care Outcomes Prediction model predictions before and after the GCS change were muted.



Footnote: Health system A inadvertently changed their process for capturing GCS to align more closely with APACHE methodology resulting in a decrease in the proportion of patients with a total GCS score of 3 from 8% to 5% of the population. AUROCs = 0.91 vs. 0.84 for the Philips Critical Care Outcomes Prediction model vs. IVa, respectively.

Figure 5.1 Change in predicted mortality after an inadvertent change in GCS documentation practice in two consecutive years.



Footnote: Health system B deliberately changed its process for capturing GCS to align more closely with APACHE methodology, resulting in a decrease in the proportion of patients with a total GCS score of 3 from 23.7% to 5.1% of the population. AUROCs = 0.90 vs. 0.81 for the Philips Critical Care Outcomes Prediction model vs. IVa, respectively.

Figure 5.2 Change in predicted mortality after a deliberate change in GCS documentation practice in two consecutive years.

In addition, we have prepared PowerPoint slides for each individual health system. From the slides, customers can track APACHE IVa and the Philips Critical Care Outcomes Prediction model performance for each hospital, and track the impact on their program by hospital discharge year/quarter, especially around the time of any known changes in GCS documentation to evaluate the impact at their institution. Upon request, we will also provide a detailed view of the model impact for patient subgroups within each health system.

Discussion

Risk-adjustment is essential to provide fair and meaningful comparisons between units, hospitals, or health systems in benchmarking analysis. A well-calibrated risk prediction model is the centerpiece of the risk adjustment. It is a valid concern that risk prediction models developed in different patient samples and at different times, may not reflect the real risk for current patient populations. As predictive models tend to lose discrimination over time, recalibration becomes necessary. It is crucial to understand the value and the risk of using any predictive model before implementing it in a risk-adjustment analysis.[7]

Philips' eSearch tool has been providing APACHE IV, IVa, and IVb for customers for several years, and our benchmarking tools and reports currently use APACHE IVa for risk-adjustment when ranking programs and facilities. Although APACHE has released several versions of risk prediction models, the latest update as of 2020, was APACHE IVb which was based on 2014-2015 data.[2]

The transition from manually collected data to automated data collection from EMR, has led to an increased risk of other forms of bias, introduced by variations in documentation patterns that may exist across hospitals and across time. The most prominent issue raised by the customer is around the GCS documentation. Customers have noticed substantial changes in their APACHE scores and predictions after their known changes in how they document GCS scores (Figures 5.1 and 5.2).

We built a systematic approach to automate data collection, monitor the quality of data over the entire pipeline, and develop new models to mitigate biases introduced by changes in documentation significantly, and properly service and maintain risk models over time and across different patient cohorts.

This analysis demonstrated that a Philips Critical Care Outcomes Prediction model trained using the eICU data, with specific considerations given to the local data structure, reliability, and known data documentation behavior, outperformed APACHE IVa and IVb in many aspects.

We believe our ability to accomplish the above efforts was in part due to several advantages we have:

- Better visibility into the customer install base, improving awareness of potential biases in the data sources, and ways of documentation. This enhanced our ability to design customized model features that are more resistant to documentation bias
- A large sample size of the heterogeneous population over multiple institutions
- Abilities to monitor the behaviors of models over the years
- Availabilities of advanced machine learning approach accessible to the analytic platform
- High-resolution electronic data collection automated from the EMR

However, we did identify that because the Philips Critical Care Outcomes Prediction model requires admission BMI and some other commonly-measured laboratory values to make a prediction, the Philips Critical Care Outcomes Prediction model scored slightly fewer patients than the APACHE model, though very similar to APACHE's hospital mortality models. It is important to recognize that the overall documentation burden has been reduced with the Philips Critical Care Outcomes Prediction model in comparison with APACHE, as comorbidities, active treatments, and urinary output are no longer required.

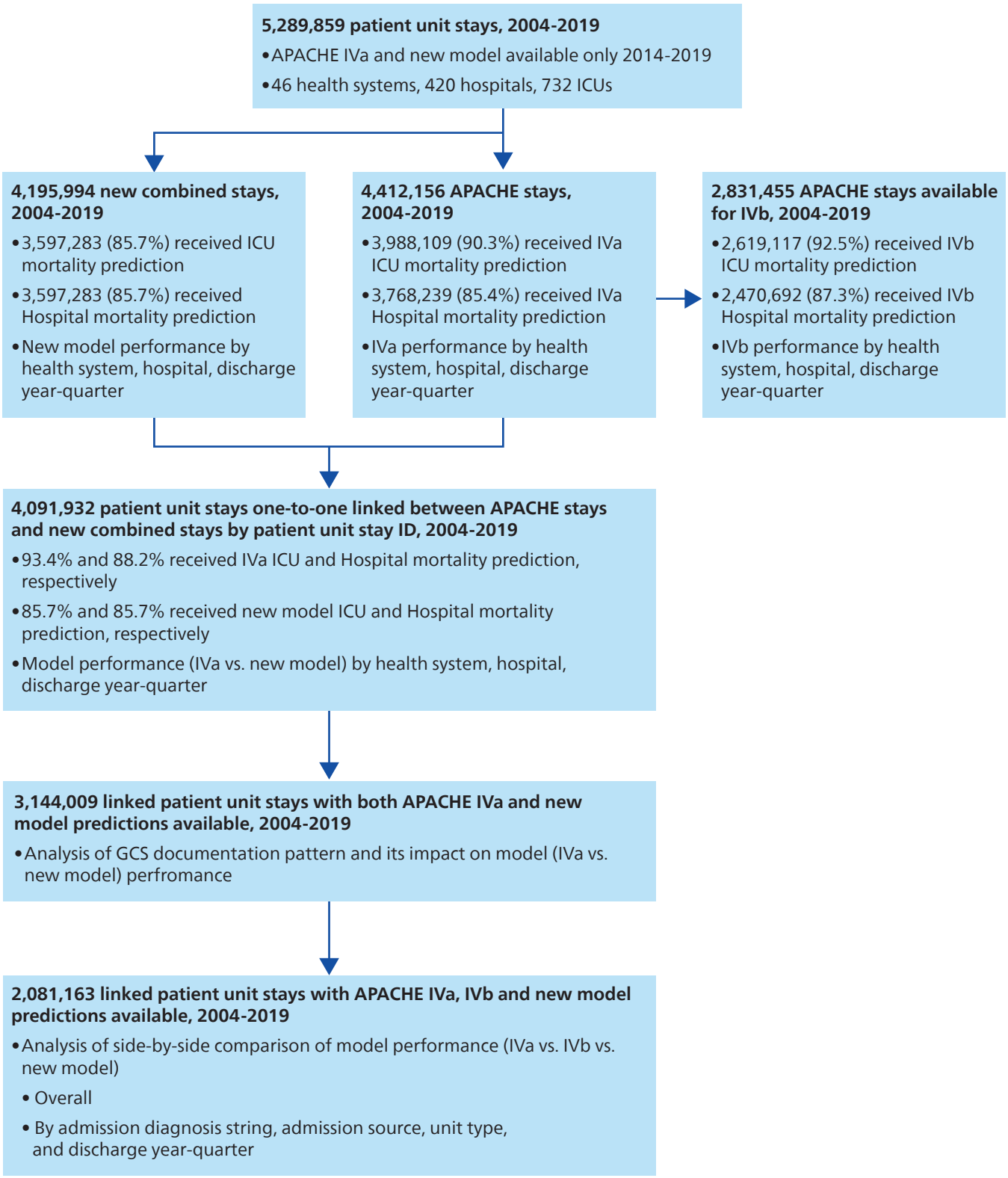
Conclusions and future steps

In this analysis, we have demonstrated an improved performance of Philips' new mortality benchmarking models over APACHE IVa and IVb. The Philips Critical Care Outcomes Prediction model not only outperformed both versions of APACHE in the overall model calibration and discriminative performances, but also retained more stability over time and was able to mitigate the impact of significant biases typically introduced by the varying documentation pattern of user inputs.

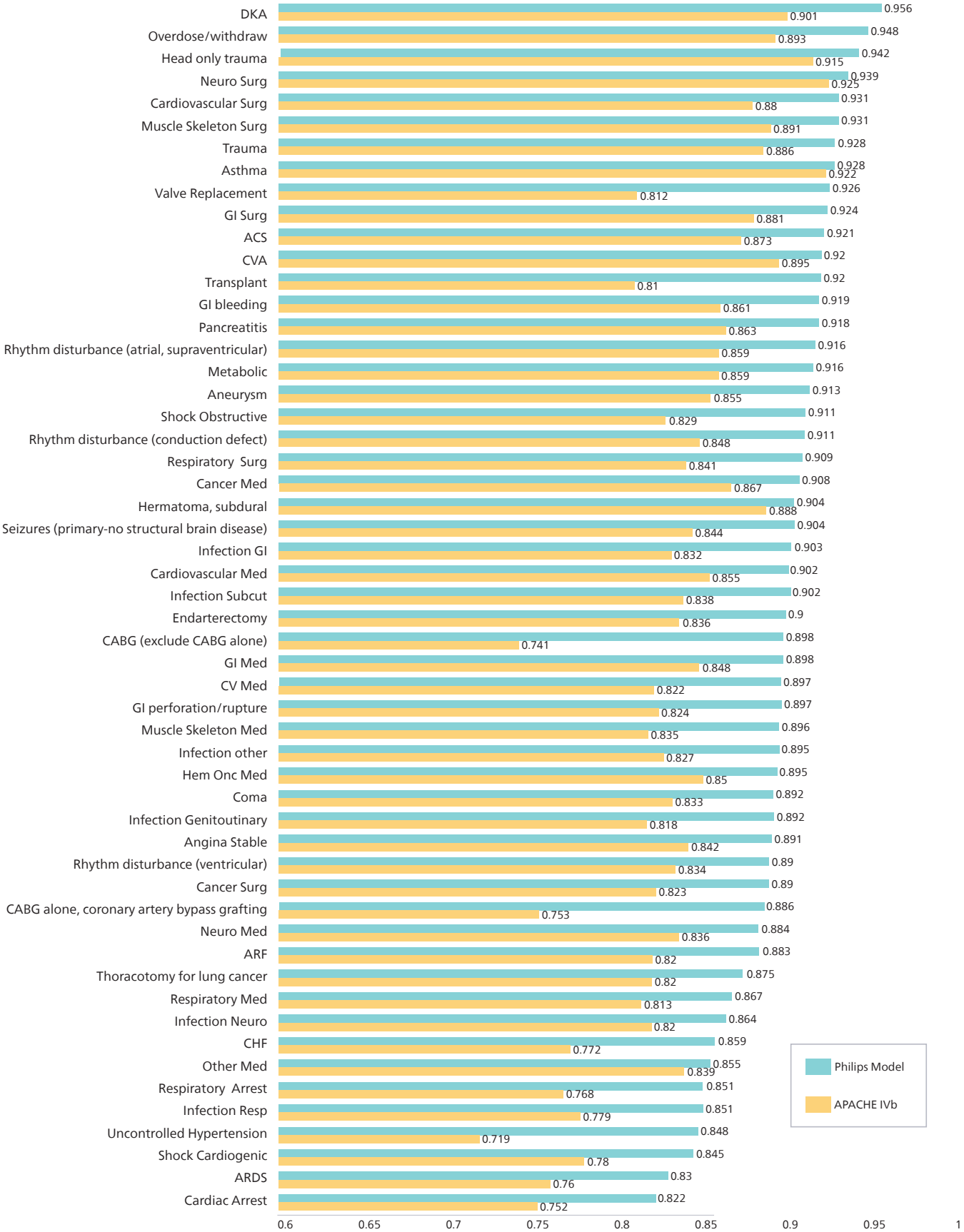
Based on the findings described in this paper, we plan to:

- Implement the new mortality prediction model for our customer install base and offer it for customers to select for benchmarking ICUs, hospitals, and eICU programs within the Benchmarking & Data Analytics platform.
- Obtain feedback from customers around missing data approaches and features required for a patient to be scored (e.g., BMI).
- Continue work on the other critical benchmarking outcomes, including the length of stay and ventilation days.
- Develop and implement a regular update of the risk prediction models.
- Proactively identify potential biases introduced over time and make changes to the predictive models and benchmarking.

Appendix



Appendix Figure 1. Patient cohort chart and type of analysis done in each cohort.



Appendix Figure 2. Model performance (ICU mortality AUROCs) by admission diagnosis groups: APACHE IVb vs. the Philips Critical Care Outcomes Prediction model (2014-2019).

Data input category	Data input	Detailed definition
Basic characteristics	BMI*	Kg/m2
Basic characteristics	Age*	Years
Basic characteristics	Gender	Female, non-Female (or NA)
Basic characteristics	Pre-ICU admission lead time*	Hours in the hospital before ICU
Basic characteristics	ICU admission source	i.e., floor, ER, unspecified
Basic characteristics	Ventilation status*	Yes vs. No, at hour 24 of ICU admission
Basic characteristics	Admitted with elective surgery status*	Yes vs. No
Vital signs	Mean blood pressure*	mmHg, mean, variability
Vital signs	Systolic blood pressure*	mmHg, mean
Vital signs	Diastolic blood pressure*	mmHg, mean
Vital signs	Heart rate*	Rate per minute, mean, variability
Vital signs	Respiratory rate*	Rate per minute, mean, variability
Vital signs	Oxygen saturation, SaO2*	%, Mean
Labs	Blood glucose*	mg/dl, mean
Labs	Blood white blood cell*	Count per ml, mean
Labs	Blood sodium*	mEq/L, mean
Labs	Blood potassium*	mEq/L, mean
Labs	Blood creatinine*	mmol/L, mean
Labs	Blood hemoglobin*	g/dl, mean
Labs	Blood albumin	g/dl, mean, with missing
Labs	Blood lactate	mmol/L, mean, with missing
Labs	Arterial blood gas, PH	Mean, with missing
Labs	Arterial blood gas, PaCO2	mmHg, mean, with missing
Provider assessment	Admission diagnosis	Categories allowing un-specified
Provider assessment	Total Glasgow coma scale score	GCS scores (3-15) with unable to score due to medication, NA; last entry at 24 hours of ICU admission

Footnote: * required data for model development; Abbreviations: NA: Not available.

Appendix Table 1.1 Data inputs used in new benchmark mortality model.

Variable	% Missing among 463,221 linked stays received valid APACHE IVa ICU mortality prediction, but no Philips Critical Care Outcomes Prediction model prediction
avgWBC	46.3%
avgHGB	42.3%
BMI	40.4%
avgCreatinine	35.7%
avgSodium	35.2%
avgPotassium	33.2%
avgGlucose	21.7%
avgDias	1.8%
avgMap	1.8%
varMap	1.8%
avgSys	1.8%
avgSaO2	1.8%
varSaO2	1.8%
avgRR	1.2%
avgHeartRate	0.6%
varHeartRate	0.6%
calcAge	0.0%
catGender	0.0%
catUnitAdm~e	0.0%
preAdmissionLeadTime	0.0%
electiveSurgery	0.0%
dxGroup	0.0%
catAvgLactate	0.0%
catAvgpH	0.0%
catAvgpaCO2	0.0%
catAvgAlbumin	0.0%
catVentilation	0.0%
gcsTotalLast	0.0%

Footnote: among 3,731,478 patient unit stays one-to-one linked between APACHE stays and new combined stays by patient unit stay ID, there were 463,221 stays received valid APACHE IVa ICU mortality prediction, but no Philips Critical Care Outcomes Prediction model prediction.

Appendix table 1.2 missing data pattern of the Philips Critical Care Outcomes Prediction model, in comparison to APACHE IVa ICU prediction.

References

1. Cerner. The APACHE IV equations: benchmarks for mortality and Resource Use (white paper). In. Kansas City: Cerner Corporation; 2005.
2. Cerner. APACHE IVb White Paper Report. Cerner Corporation;2016.
3. Zimmerman JE, Kramer AA, McNair DS, Malila FM. Acute Physiology and Chronic Health Evaluation (APACHE) IV: hospital mortality assessment for today's critically ill patients. *Crit Care Med*. 2006;34(5):1297-1310.
4. Zimmerman JE, Kramer AA, McNair DS, Malila FM, Shaffer VL. Intensive care unit length of stay: Benchmarking based on Acute Physiology and Chronic Health Evaluation (APACHE) IV. *Crit Care Med*. 2006;34(10):2517-2529.
5. Lilly CM, Swami S, Liu X, Riker RR, Badawi O. Five-Year Trends of Critical Care Practice and Outcomes. *Chest*. 2017;152(4):723-735.
6. Hosmer DW, Lemeshow S. Assessing the fit of the model. In: Applied Logistic Regression. 2nd ed. New York, NY: Q Wiley-Interscience Publication; 2000:143-202.
7. Glance, Laurent G. et al. Benchmarking in Critical Care. *CHEST*, Volume 121, Issue 2, 326 - 328.

4522 991 70891 * AUG 2021

www.philips.com