**Table S1** – Frequency of missing values in the training dataset, internal validation dataset, and external validation datasets, presented as the count and percentage of cases with missing values for each variable.

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** | **Training** | **Internal Validation** | **External Validation** |
| Transfused | 0 (0%) | 0 (0%) | 0 (0%) |
| Age | 0 (0%) | 3 (< 0.01%) | 0 (0%) |
| Height | 59,919 (2.0%) | 23,367 (2.2%) | 177 (1.1%) |
| Weight | 34,516 (1.1%) | 13,835 (1.3%) | 89 (0.6%) |
| Gender | 3 (< 0.01%) | 119 (0.01%) | 0 (0%) |
| ASA Status | 7,477 (0.25%) | 2,458 (0.23%) | 18 (0.04%) |
| HTN | 0 (0%) | 0 (0%) | 1 (< 0.01%) |
| CHF | 0 (0%) | 0 (0%) | 1 (< 0.01%) |
| Smoking | 0 (0%) | 1 (< 0.01%) | 1 (< 0.01%) |
| COPD | 0 (0%) | 0 (0%) | 1 (< 0.01%) |
| Dialysis | 0 (0%) | 0 (0%) | 1 (< 0.01%) |
| Diabetes | 1 (< 0.01%) | 0 (0%) | 1 (< 0.01%) |
| Hematocrit | 489,076 (16%) | 172,675 (16%) | 3,639 (23%) |
| Platelet | 520,875 (17%) | 184,686 (17%) | 3,653 (23%) |
| INR | 1,844,389 (60%) | 672,110 (62%) | 8,656 (54%) |
| PTT | 2,139,041 (70%) | 775,483 (72%) | 9,137 (57%) |
| Creatinine | 581,318 (19%) | 213,555 (20%) | 3,022 (19%) |
| Sodium | 606,082 (20%) | 224,193 (21%) | 3,000 (19%) |
| Albumin | 1,480,598 (49%) | 529,041 (49%) | 7,158 (45%) |
| Bilirubin | 1,502,853 (49%) | 535,572 (50%) | 7,278 (45%) |
| Elective surgery | 4,706 (0.15%) | 1,350 (0.12%) | 0 (0%) |
| Procedure-specific transfusion rate | 0 (0%) | 14,546 (1.4%) | 0 (0%) |

**Table S2** – Values from the training dataset used for data preprocessing, i.e. median imputation, mean subtraction, and normalization to unit variance. These transformations were applied using the same values across the training, internal validation, and external validation datasets.

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** | **Median** | **Mean** | **Standard deviation** |
| Transfused | 0 | 0.044 | 0.206 |
| Age (yrs) | 58 | 56.7 | 16.9 |
| Height (in) | 66 | 66.9 | 4.1 |
| Weight (lbs) | 181 | 187.6 | 50.6 |
| Gender  (Female = 1) | 1 | 0.568 | 0.495 |
| ASA Status | 2 | 2.45 | 0.736 |
| HTN | 0 | 0.444 | 0.497 |
| CHF | 0 | 0.0085 | 0.0918 |
| Smoking | 0 | 0.169 | 0.375 |
| COPD | 0 | 0.042 | 0.202 |
| Dialysis | 0 | 0.013 | 0.113 |
| Diabetes | 0 | 0.213 | 0.531 |
| Hematocrit (%) | 40.1 | 39.64 | 5.32 |
| Platelet (1000 / uL) | 242 | 250.9 | 81.7 |
| INR | 1.00 | 1.078 | 0.288 |
| PTT (s) | 29.1 | 30.4 | 7.70 |
| Creatinine (mg/dL) | 0.85 | 1.00 | 0.856 |
| Sodium (mEq/L) | 139 | 139.2 | 3.25 |
| Albumin (g/dL) | 4.00 | 3.91 | 0.611 |
| Bilirubin (mg/dL) | 0.50 | 0.646 | 0.652 |
| Elective surgery  (Yes = 1) | 1 | 0.798 | 0.401 |
| Procedure-specific transfusion rate (%) | 0.29 | 2.40 | 5.74 |

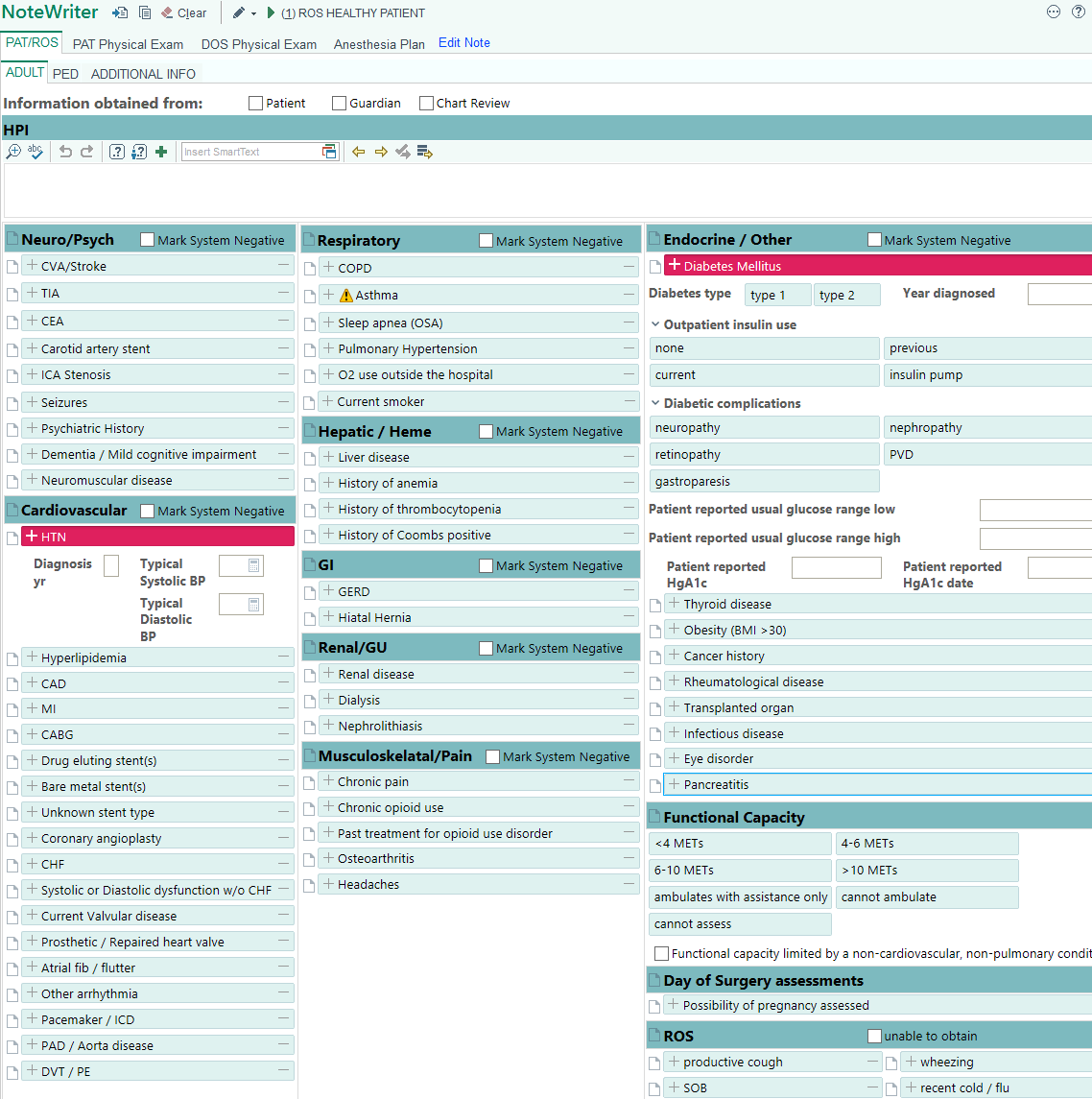
**Table S3** – Model performance on the external validation data for procedures with fewer than 50 occurrences in 2019.

Model performance depends on reliable procedure-specific transfusion data. Procedures with fewer than 50 historic observations have higher uncertainty in the estimate of historical procedure-specific transfusion rate. For example, historical transfusion data was only available for 2019 in the external validation dataset, and not for earlier time periods due to an electronic health record transition. Therefore, only 2019 data could be used to compute historical procedure-specific transfusion rates. If a procedure was only performed 10 times at the external validation institution in 2019, and 0 transfusions were observed, the naïve estimate of procedure-specific transfusion risk would be 0/10 = 0%. However, the true transfusion rate could be much higher and due to the small number of samples, 0% could have been observed by chance (95% CI 0 – 30.8%). If that procedure had occurred 50 times with 0 transfusions observed, the procedure-specific transfusion risk estimate is more precise (95% CI 0 – 7.7%).

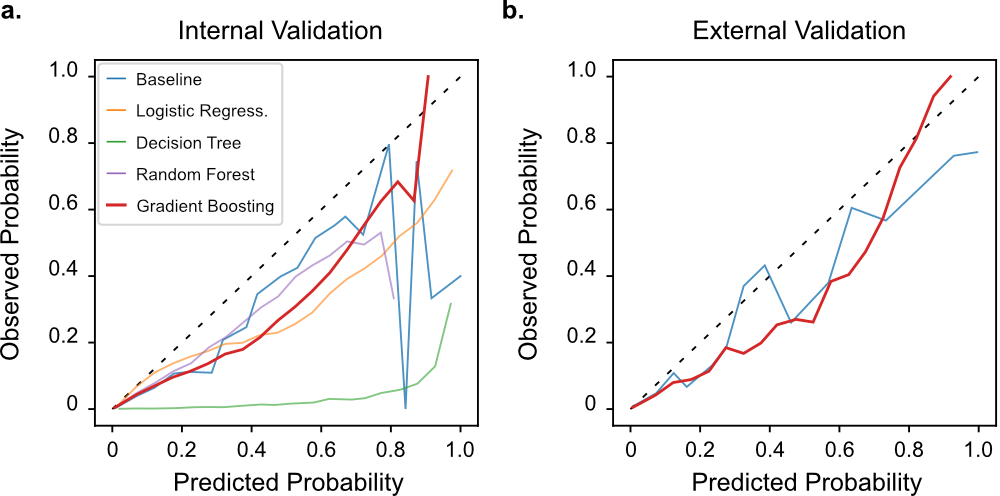
To evaluate model performance when imprecise estimates of procedure-specific transfusion rate are provided as input, we assessed model predictions only for surgical cases with fewer than 50 occurrences contributing to procedure-specific transfusion risk estimates in the external validation dataset. Overall discrimination as measured with c-statistic and average precision was poor in comparison to the results presented in the main manuscript (Table 3). In addition, we were unable to achieve 96% sensitivity to detect transfusion regardless of decision threshold without simply recommending type and screens for all patients. Therefore, we do not recommend using our model with imprecise procedure-specific transfusion rate estimates.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Model** | **c-statistic** | **Average Precision** | **Sensitivity** | **Specificity** | **Positive Predictive Value** | **% Positive** |
| Baseline | 0.816  (0.804-0.829) | 0.371  (0.349-0.396) | 0.840  (0.819-0.858) | 0.616  (0.609-0.622) | 0.150  (0.143-0.158) | 41.8%  (41.2-42.5) |
| Gradient Boosting Machine | 0.876  (0.868-0.886) | 0.460  (0.434-0.490) | 0.918  (0.904-0.932) | 0.582  (0.575-0.589) | 0.151  (0.144-0.158) | 45.5% (44.8-46.2) |

**Figure S1** – Illustration of pre-operative note template used at Barnes Jewish Hospital. Clinicians (nurse practitioners and resident physicians) at the pre-operative assessment clinic complete this note template for all pre-surgical patients based on the patient interview and chart review. Clinicians flag the patient as having discrete comorbidities based on synthesis of the available data by clicking the plus sign next to each listed comorbidity. Illustrated below, the clinician has flagged this patient as having hypertension and diabetes. The comorbidities used for this study were abstracted from the flagged comorbidities in these pre-operative note templates with mapping of the comorbidities to note template buttons as follows: hypertension = “HTN”, congestive heart failure = “CHF”, smoking = “Current smoker”, chronic obstructive lung disease = “COPD”, dialysis = “Dialysis”, and diabetes = “Diabetes Mellitus”.



**Figure S2** – Calibration plots for model performance on the internal validation (a) and external validation (b) data. Perfect calibration is indicated by the dashed black line. All models appear to over-estimate transfusion risk, with observed probability less than predicted probability.



**Figure S3** – Net benefit analysis using the external validation cohort.

Net benefit analysis is based on the idea that each possible decision threshold implies a value judgement on the tradeoff between true positives (ordering type and screens for patients who require transfusion) and false positives (ordering type and screens for patients who do not require transfusion)[31]; for example, a threshold of 0.05 implies that one would be willing to perform 20 additional type and screens to identify 1 additional patient who requires transfusion. For each possible decision threshold, its implied value judgement is used to compute the “value” or net benefit of each model based on its predictive performance.

In the figure, the baseline and gradient boosting models are compared to two other reference strategies for type and screen allocation: ordering type and screens for everyone, and ordering type and screens for no one. The gradient boosting model has the highest net benefit for all threshold probabilities > 0.0012. For threshold probabilities < 0.0012 (i.e., if one is willing to perform 833 or more type and screens to identify 1 patient who requires transfusion), the strategy of ordering type and screens for all patients has highest net benefit.

