Supplementary File for '**Using TeleTriage to Model the Risk of Hospital Admission at the Time of Registration in an Emergency Department**'

Figure A1. Word cloud for top 10 chief complaints for attendances at the JR hospital during the training period.

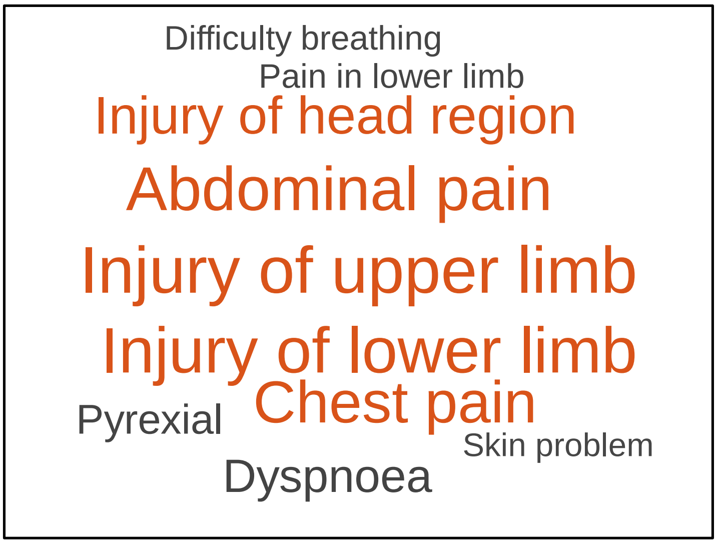
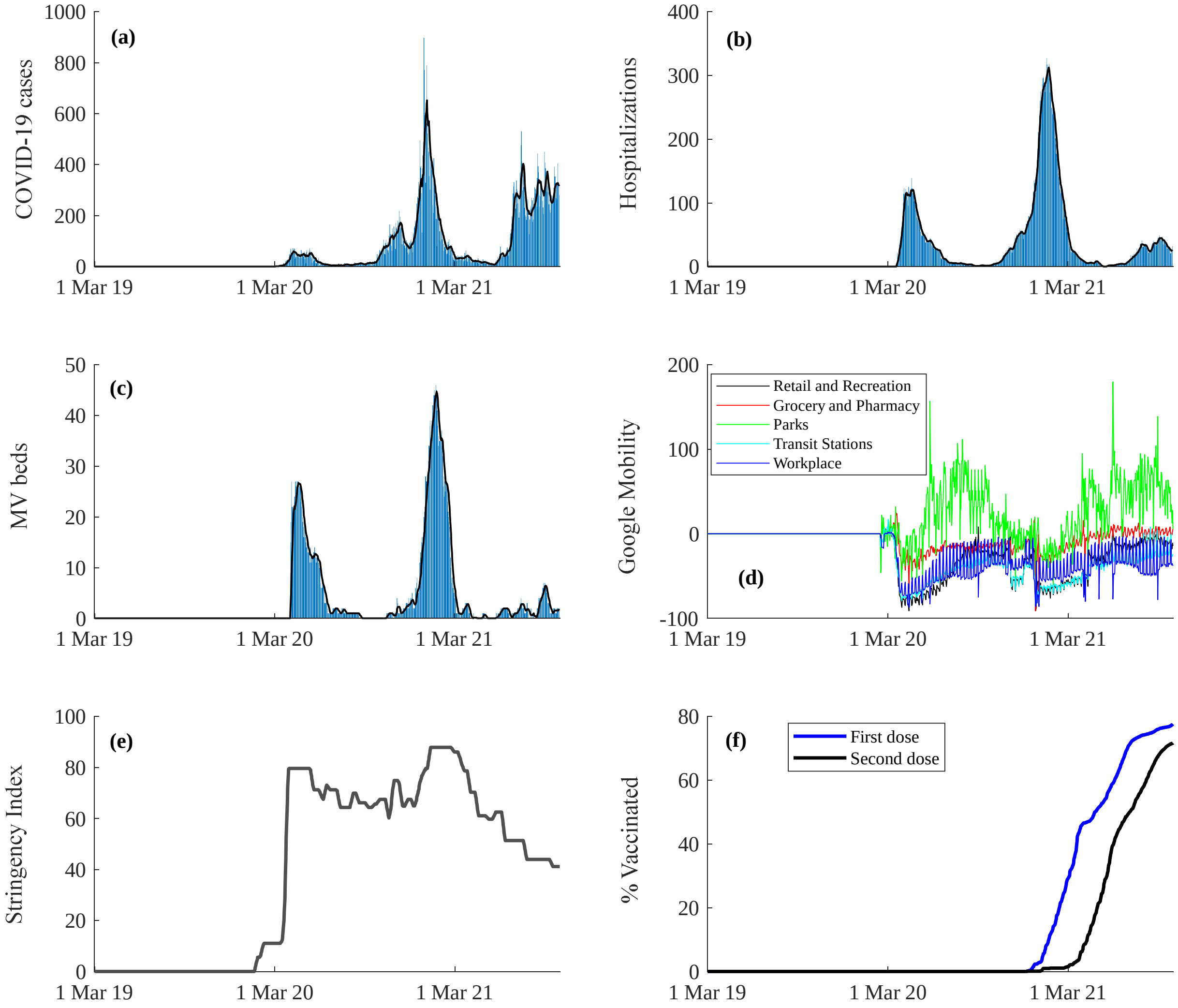


Figure A2. Plots of population-level features for Oxfordshire recorded daily. Panel **1a**: COVID-19 cases, **1b**: COVID-19 hospitalizations at the Oxford University Hospitals (OUH), **1c**: number of mechanical ventilator beds in use at the OUH, **1d**: mobility data from Google for (i) retail and recreation, (ii) grocery and pharmacy, (iii) parks, (iv) transit stations, and (v) workplace, **1e**: Stringency index, and **1f**: Percentage vaccinated. The data is presented for the period 1 March 19-30 Sept 21, inclusive.



Note: Google mobility data was not available for the pre-pandemic period and was thus treated as zero.

Random Forests (RFs)

In addition to RF with priors, we investigated two other approaches to deal with class imbalance, namely RUSBoost and RF and ADASYN, which are discussed below:

RUSBoost – This algorithm of Seiffert et al. (2010) is a hybrid approach based on the concept of random undersampling (RUS) and boosting. It comprises two steps: (1) Using only the training data, we randomly undersample observations from the majority class (without replacement) until both classes had the same number of observations. (2) Boosting is used to train the classification trees iteratively using the resampled (balanced) dataset (Freund and Schapire 1996). Initially, all training observations are assigned equal weight. With each iteration, misclassified observations are given a higher weight. With subsequent iterations, the model focuses more on observations that are harder to classify correctly. After the iterations, a weighted voting scheme is used across the classification trees to issue a final prediction. This algorithm required the estimation of the number of trees, the minimum size of the leaf node, the number of features to use for split-point selection, and a learning rate for shrinkage.

RF with ADASYN – A drawback of under-sampling approaches is a loss of information. Oversampling algorithms mitigate this by replicating observations from the minority class. We employed the adaptive synthetic (ADASYN) oversampling approach that generates synthetic data by performing slight perturbations on the minority class observations (He et al. 2008). An attractive feature of ADASYN is that it generates more synthetic data for those minority class observations that are harder to learn by adaptively shifting the decision boundary (i.e., the surface that corresponds to the demarcation of observations from the two classes). ADASYN calculates the number of synthetic observations to be generated using a *k*-nearest neighbour approach to achieve a predefined balance, which we set as 50:50. The synthetic observations are generated by linear combining two minority observations that belong to the same neighbourhood. ADASYN required the estimation of *k* (number of nearest neighbours), the number of trees, the minimum size of the leaf node, and the number of features for split-point selection.

Table A1. Out-of-sample mean AUC (and 95% confidence intervals) for predicting the risk of hospital admission from the ED at the JR using the methods RUSBoost, RF with ADASYN, and RF with priors, and feature matrices X_1, X_1 U X_2, and X_1 U X_2 U X_3.

Internal Validation: The JR Hospital

Method	X_1	X_1 U X_2	X_1 U X_2 U X_3
RUSBoost	0.605 (0.601-0.608)	0.838 (0.836-0.840)	0.836 (0.834-0.838)
RF with ADASYN	0.721 (0.718-0.724)	0.871 (0.869-0.872)	0.877 (0.875-0.879)
RF with priors	**0.728 (0.725-0.731)**	**0.881 (0.879-0.883)**	**0.882 (0.880-0.884)**

Note: Higher values of the AUC are better. Bold indicates the best result in each column.

Table A2. Internal validation: Out-of-sample AUC values (95% CI) for forecasting risk of patient admission at the JR hospital using: RUSBoost, RF with ADASYN, and RF with priors, whereby for each classifier, we employ: only X_1, only $X_1 \text{U} X_2$, and $X_1 \text{U} X_2 \text{U} X_3$. AUC for a perfect classifier is 1.

JR Hospital	**Performance Score (AUC) –Top 5 reasons for admissions**		
Chief Complaint	$\text{AUC}_1 : X_1$	$\text{AUC}_2 : X_1 \text{U} X_2$	$\text{AUC}_3 : X_1 \text{U} X_2 \text{U} X_3$
1. Abdominal pain			
RUSBoost	0.547-0.569	0.633-0.655	0.635-0.654
RF with ADASYN	0.608-0.629	0.681-0.702	0.683-0.703
RF with priors	**0.612-0.632**	**0.686-0.708**	**0.688-0.709**
2. Dyspnoea			
RUSBoost	0.579-0.605	0.701-0.727	0.735-0.760
RF with ADASYN	0.725-0.750	0.824-0.845	0.837-0.855
RF with priors	**0.731-0.756**	**0.841-0.860**	**0.842-0.860**
3. Chest pain			
RUSBoost	0.535-0.557	0.577-0.599	0.568-0.589
RF with ADASYN	0.623-0.644	0.685-0.708	0.699-0.718
RF with priors	**0.633-0.653**	**0.705-0.724**	**0.713-0.733**
4. Injury of lower limb			
RUSBoost	0.672-0.701	0.948-0.956	0.945-0.953
RF with ADASYN	0.869-0.888	0.956-0.963	0.957-0.964
RF with priors	**0.867-0.887**	**0.958-0.964**	**0.957-0.965**
5. Instability of gait			
RUSBoost	0.526-0.574	0.613-0.660	0.608-0.654
RF with ADASYN	0.601-0.648	0.653-0.697	0.656-0.696
RF with priors	**0.617-0.661**	**0.653-0.696**	**0.661-0.705**

Note: Highest AUC values using only X_1, only $X_1 \text{U} X_2$, and $X_1 \text{U} X_2 \text{U} X_3$ are highlighted in **bold**.

Table A3. Internal and external validation: out-of-sample Brier score, calibration-in-the-large, and calibration slope values, for forecasting the risk of patient admission at the JR and HG hospital using: RF with ADASYN and RF with priors, whereby for each classifier, we employ: only X_1, only $X_1 \text{U} X_2$, and $X_1 \text{U} X_2 \text{U} X_3$.

A.	**Internal Validation: The JR Hospital**		
	Brier score		
Method	X_1	$X_1 \text{U} X_2$	$X_1 \text{U} X_2 \text{U} X_3$
RF with ADASYN	0.196	0.137	0.132
RF with priors	0.188	0.130	0.130
	Calibration intercept		
Method	X_1	$X_1 \text{U} X_2$	$X_1 \text{U} X_2 \text{U} X_3$
RF with ADASYN	-0.365	-0.377	-0.140
RF with priors	-0.050	-0.002	0.048
	Calibration slope		
Method	$\boldsymbol{X_1}$	$\boldsymbol{X_1 \text{U} X_2}$	$\boldsymbol{X_1 \text{U} X_2 \text{U} X_3}$
RF with ADASYN	1.038	1.050	1.274
RF with priors	0.958	1.070	1.207

B.	**External Validation: The HG Hospital**		
	Brier Score		
Method	X_1	$X_1 \text{U} X_2$	$X_1 \text{U} X_2 \text{U} X_3$
RF with ADASYN	0.159	0.153	0.127
RF with priors	0.148	0.111	0.113
	Calibration intercept		
Method	X_1	$X_1 \text{U} X_2$	$X_1 \text{U} X_2 \text{U} X_3$
RF with ADASYN	-0.858	-1.366	-1.064
RF with priors	-0.667	-0.545	-0.607
	Calibration slope		
Method	X_1	$X_1 \text{U} X_2$	$X_1 \text{U} X_2 \text{U} X_3$
RF with ADASYN	1.459	1.097	1.559
RF with priors	1.358	1.340	1.458

Figure A3. Calibration plots for out-of-sample probability forecasts of hospital admission from the ED at the JR for RF with priors, using feature matrices X_1 and X_1 U X_2.

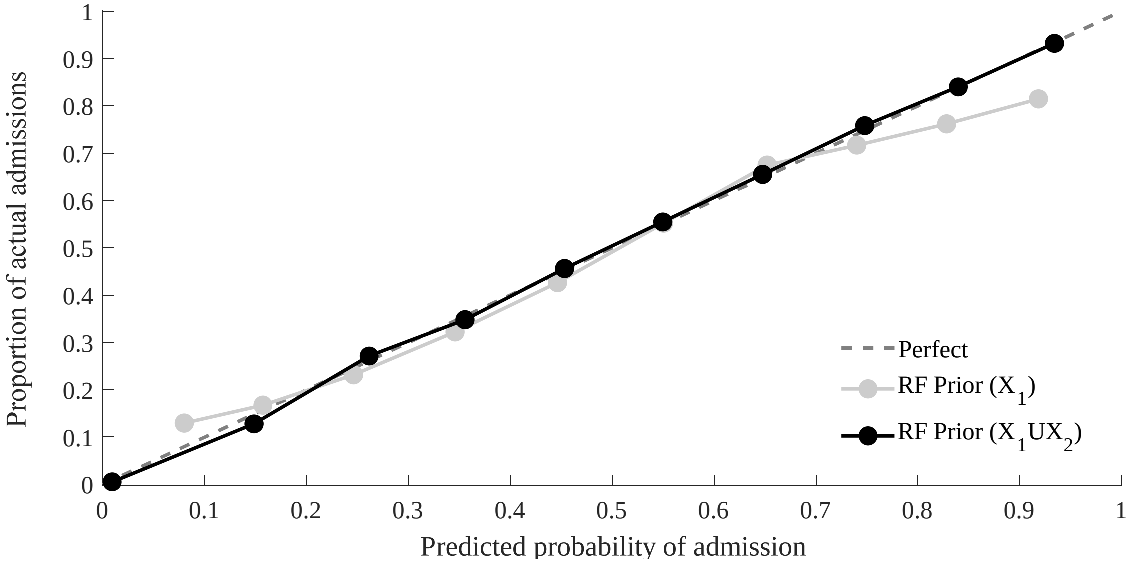


Figure A4. Calibration plots and evaluation of costs associated with decision-making for the out-of-sample probability forecasts of hospital admission from the ED at the JR and HG for RF with priors using X_1UX_2.

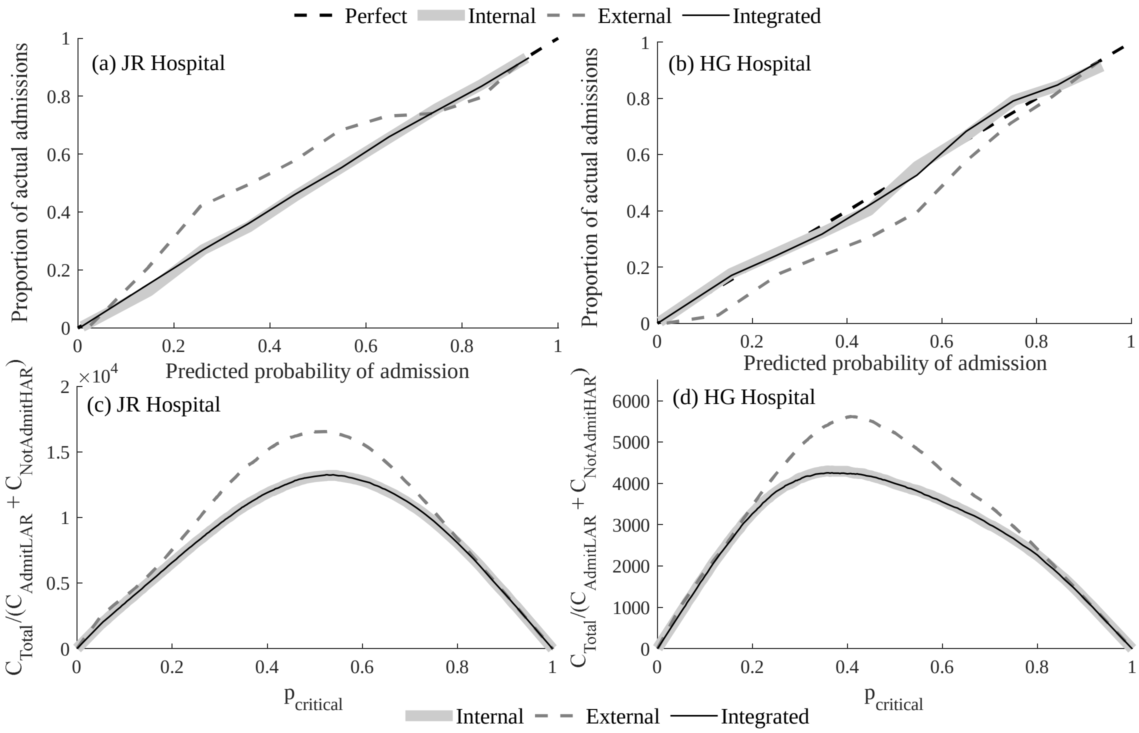


Table A4. Confusion matrix for predicting the triage category.

		Prediction		
		Major injury	Minor injury	Urgent care
Actual	Major injury	80.0%	18.8%	11.2%
	Minor injury	21.8%	77.1%	1.1%
	Urgent care	41.3%	53.3%	5.4%

Note: higher values along the diagonal are better.

Confusion matrix to assess the out-of-sample accuracy for predicting the triage category. The triage category, as assigned by the clinical staff, was assumed to be the actual "gold standard" (or the label). The model could predict patients with 'major injury', 'minor injury', and 'urgent care' with an accuracy of 80%, 77.1%, and 5.4%, respectively. Note that Patients triaged as 'urgent care' comprised only 3.4% of visits in the out-of-sample period.

To predict the triage category, we implemented a single classification tree with default hyperparameters (a minimum leaf size of one). For random forests, artificial neural networks (ANNs), and logistic regression, we estimated model hyperparameters using the last month of the training data as the cross-validation hold-out sample. For random forests, the selected hyperparameters were as follows: number of trees = 10, minimum leaf size = 5, number of features to use for split point selection = square root of the total number of features considered (rounded to the nearest integer). For ANNs, the selected architecture was as follows: number of hidden layers = 1, and number of nodes in a hidden layer = 5. The out-of-sample model performance for this multi-class classification problem was quantified using *accuracy*, which is calculated by dividing the correct predictions (true positive + true negative) by the total number of observations (true positive + true negative + false positive + false negative). The out-of-sample accuracy for the four modelling approaches was as follows (in brackets): classification tree (0.76), random forests (0.72), ANNs (0.67), and logistic regression (0.68). We acknowledge that finer tuning of the hyperparameter selection process and trying different activation functions for ANNs may have improved the performance of the sophisticated approaches. Nonetheless, we used a single classification tree due to its ease of implementation and interpretation.

Table A5. Patient group number, brief description, and the corresponding attendances (as a percentage of the total attendances for the JR hospital, calculated using only the training data).

Patient Group Number: Brief description and corresponding attendance (as %)	
10: road traffic accident	0.8%
20: assault	1.6%
30: deliberate self-harm	2.3%
40: sports injury	0.6%
50: fireworks injury	< 0.001%
60: other accident	94.3%
70: brought in dead	0%
80: other than above	0.3%

Figure A5. Relative frequencies of the total length of stay for the original dataset and model simulation generated under risk scenarios.

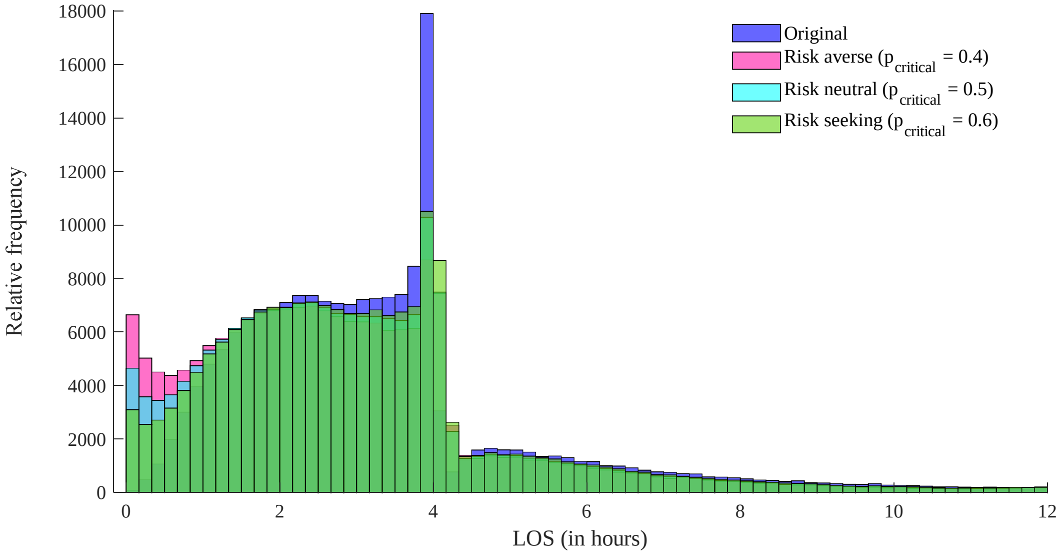
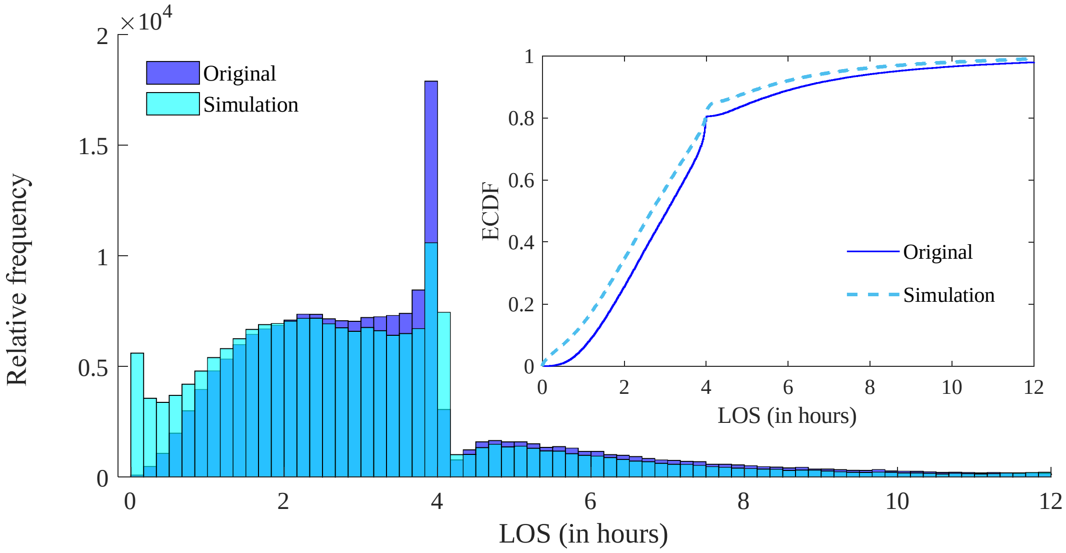


Figure A5 presents the original LOS with corresponding LOS obtained from model simulations for different values of threshold ($p_{critical}$) used to decide if a patient would need admission. Note that a patient is admitted if the predicted probability is higher than the pre-defined threshold ($p > p_{critical}$). In the above figure, we keep penalty (δ) fixed at 2.5 minutes and vary $p_{critical}$. The figure shows that lower values of $p_{critical}$ are associated with incentivizing an early intervention to admit a patient, which results in early decision-making (at the time of patient registration), while a higher value of $p_{critical}$ is associated with a slightly delayed LOS. For all values of $p_{critical}$ considered in this study, the impact of the 4-hour waiting time target on the LOS distribution was less prominent compared to the actual dataset.

Figure A6. Relative frequencies of the total length of stay (along with corresponding ECDFs in the inset) for the original dataset and model simulation. Simulation results are presented for a risk-neutral strategy ($p_{critical}$ = 0.5) and a penalty (δ) of 1 minute.



In Figure A6, compared to the original LOS distribution, the simulated LOS is lower, and the distribution is relatively smoother, with a higher proportion of patients being admitted around the time of registration. The inset in Figure A6 presents the empirical cumulative distribution function (ECDF) of the original and simulated LOS, showing that the stratification of patients using admission-risk at registration would be the preferred option based on comparison of the two ECDFs using a probabilistic approach (i.e., stochastic dominance). However, for a large penalty associated with decision-making (δ of 10 minutes), the simulated LOS is slightly higher, as shown in Figure A7. This result suggests that for higher penalty (δ), there is little value in using model prediction to decide if a patient needs admission at the time of registration (compared to admitting a patient based on clinical assessment).

Figure A7. Relative frequencies of the total length of stay (along with corresponding ECDFs in the inset) for the original dataset and model simulation. Simulation results are presented for a risk-neutral strategy ($p_{critical}$ = 0.5) and a penalty (δ) of 10 minutes.

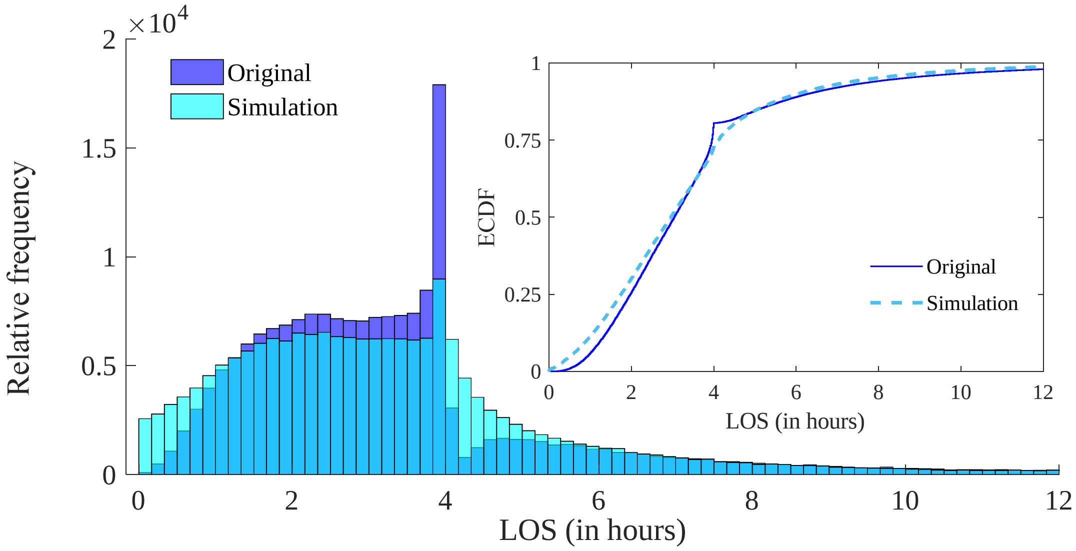
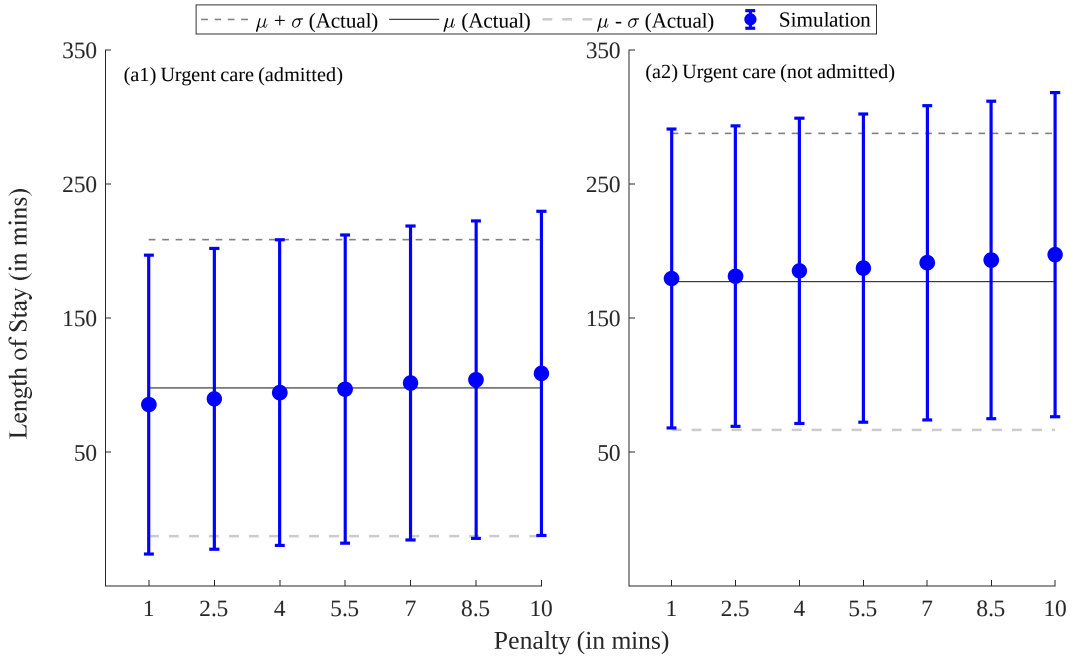
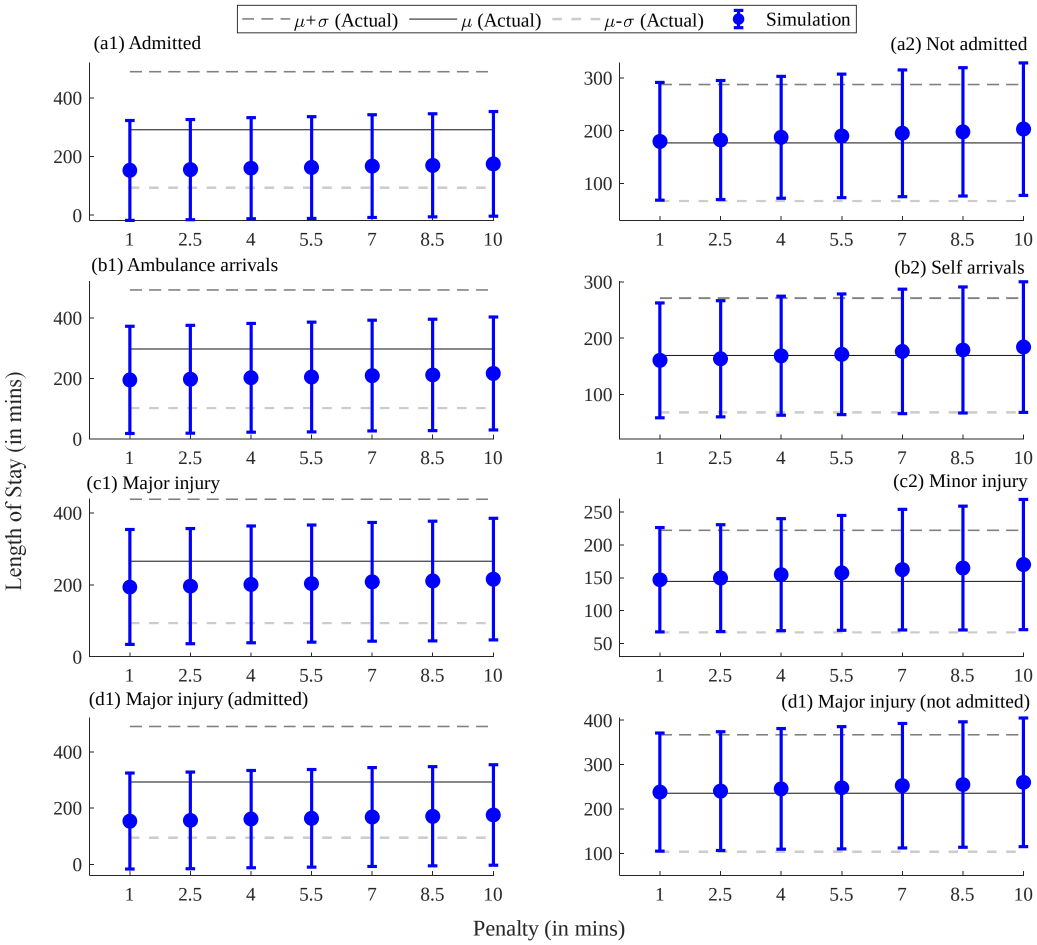


Figure A8. Mean and standard deviation of LOS (in minutes) resulting from the simulation for patients triaged as ‘urgent care’ in the out-of-sample period along with corresponding actual LOS. Different values considered (on the x-axes) for the penalty δ (in minutes) associated with a false positive and a false negative. Predictions were generated using the feature matrix, X_1 U X_2, using the integrated dataset from the two ED sites. The simulated LOS is presented using blue vertical lines, where the blue circles in the centre denote the mean and the horizontal bars at the end denote one standard deviation from the mean.



In Figure A8, we see that the model can help reduce waiting times for patients triaged as ‘urgent care’ that need admission (panel a1) if the penalty imposed for a wrong decision to admit/not admit a patient is less than around 5.5 minutes (using $\delta_{FP} = \delta_{FN}$). This improvement, however, comes at a cost of an increase in waiting times for patients (with the same triage category) that do not need admission (panel a2), for all values of the penalties considered in this study. These findings are overall consistent with patients who were triaged as ‘major injury’, as discussed in Section 5.2.

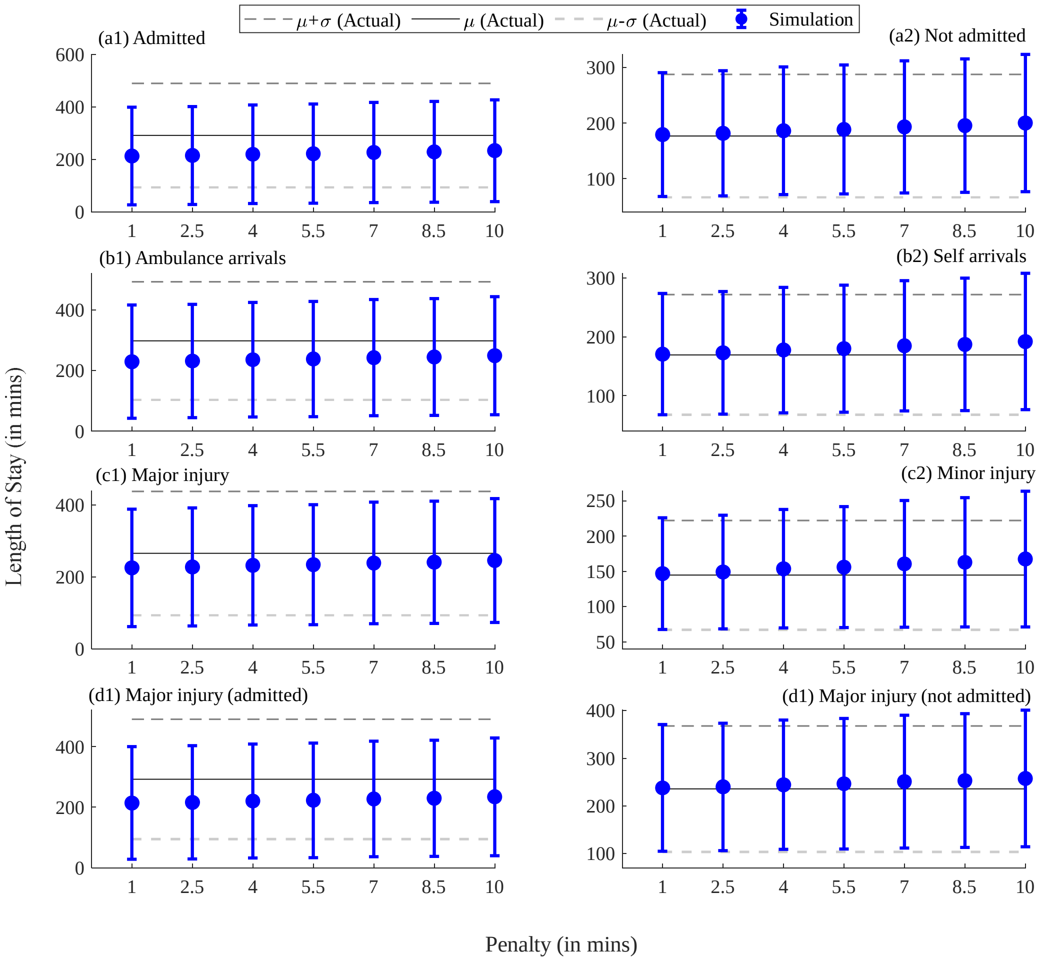
Figure A9. Length of stay (LoS, in minutes) for the JR and HG hospital sites for a range of penalty values (in minutes) associated with a false positive and a false negative. Predictions were generated using RF with priors (with X_1 and X_2) using the joint dataset from the two ED sites. LoS (mean and standard deviation) is presented for patients in the out-of-sample period. The decision to admit a patient were based on a risk-averse strategy ($p_{crit} = 0.4$).



Note: lower LOS values are better. Black and grey lines denote the mean and one standard deviation from the mean of the original LOS, respectively. Blue error bars indicate LOS from simulations for different penalties.

In Figure A9, we present the original and simulated LOS for different penalties associated with a wrong decision to admit/not admit a patient. We use a lower threshold for admitting a patient (risk-averse strategy, $p_{critical}$ = 0.4). Note that lower values of $p_{critical}$ incentivizes admission, which, as expected, translates into a further improvement in the LOS for high admission risk patients. However, this comes at the cost of a higher increase in LOS for low admission risk patients. Similarly, using a higher threshold to admit a patient (less risk-averse strategy, $p_{critical}$ = 0.6) results in a smaller improvement in LOS for high admission risk patients as a higher $p_{critical}$ disincentivizes admission at the time of registration, as presented in Figure A10.

Figure A10. Length of stay (LoS, in minutes) for the JR and HG hospital sites for a range of penalty values (in minutes) associated with a false positive and a false negative. Predictions were generated using RF with priors (with X_1 and X_2) using the joint dataset from the two ED sites. LoS (mean and standard deviation) is presented for patients in the out-of-sample period. The decision to admit a patient were based on a less risk-averse strategy ($p_{crit} = 0.6$).



Note: lower LOS values are better. Black and grey lines denote the mean and one standard deviation from the mean of the original LOS, respectively. Blue error bars indicate LOS from simulations for different penalties.

Figure A11. Mean and standard deviation of LOS (in minutes) resulting from the simulation for patients in the out-of-sample period along with corresponding actual LOS. Three different scenarios are considered (on the x-axes): (1) $\delta_{FP} = \delta_{FN}$, (2) $\delta_{FP} = 0.5\times\ \delta_{FN}$, and (3) δ_{FP} =1.5× δ_{FN}, for the penalty δ_{FN} chosen as ten minutes. Predictions were generated using the feature matrix, X_1 U X_2, using the integrated dataset from the two ED sites, for a risk-neutral strategy ($p_{critical}$ = 0.5) and the highest penalty value, $\delta_{FN} = 10$ minutes. The simulated LOS is presented using blue vertical lines, where the blue circles in the centre denote the mean and the horizontal bars at the end denote one standard deviation from the mean.

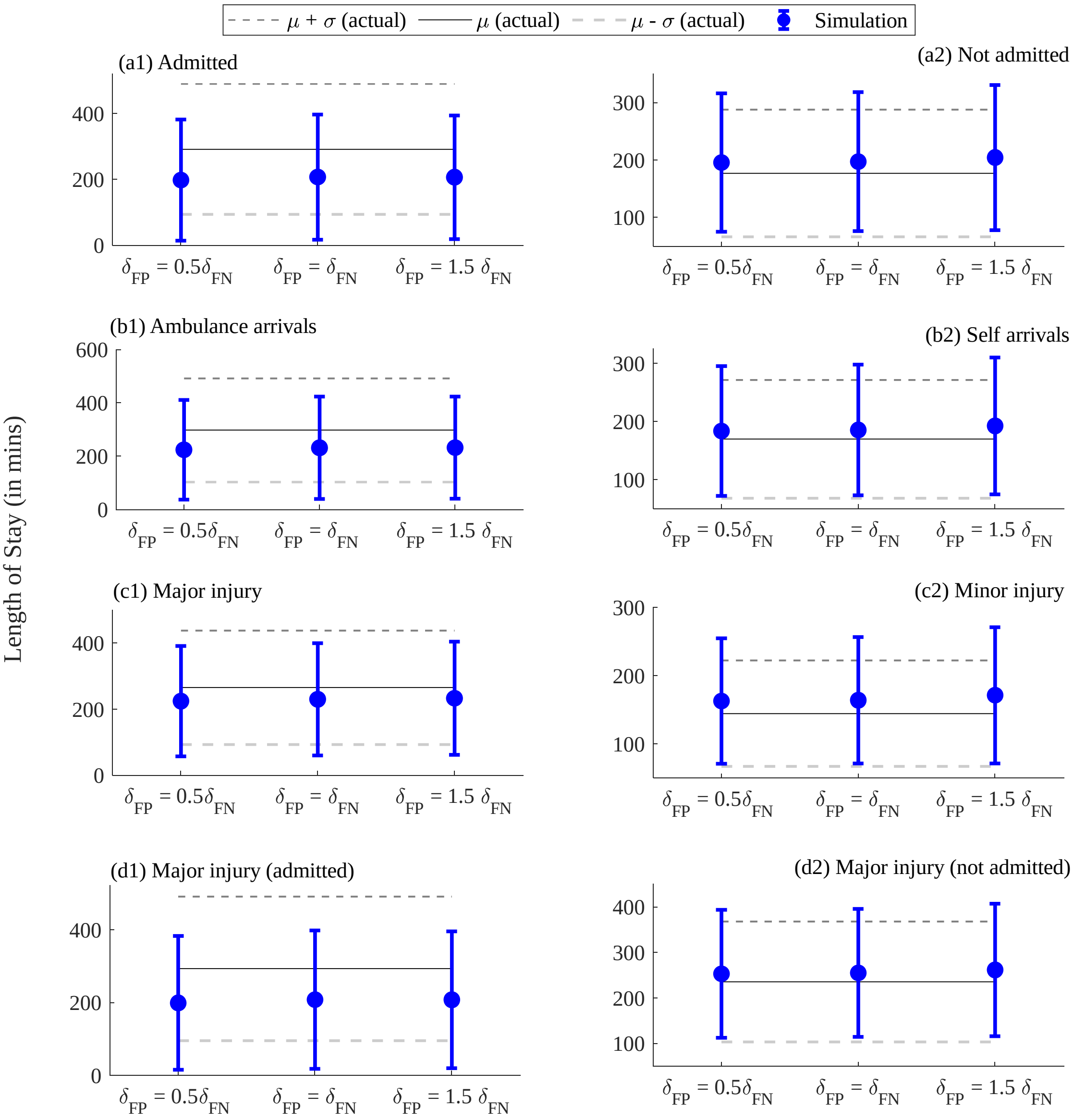


Table A6. Out-of-sample MAE and RMSE for estimating the number of bed requests from the EDs to the hospital at the JR and HG using triage category that is clinician-administered and digitally assessed, calculated using a 2-hour moving window.

Triage Category	Clinician-administered		Digitally assessed (TeleTriage)	
	MAE	RMSE	MAE	RMSE
Integrated dataset	1.84	2.41	2.31	3.08
JR dataset	1.69	2.22	1.93	2.57
HG dataset	0.93	1.21	1.01	1.33

Table A7. Out-of-sample MAE and RMSE for estimating the number of bed requests from the EDs to the hospital at the JR and HG using triage category that is clinician-administered and digitally assessed, calculated using a 3-hour moving window.

Triage Category	Clinician-administered		Digitally assessed (TeleTriage)	
	MAE	RMSE	MAE	RMSE
Integrated dataset	2.34	3.04	3.14	4.15
JR dataset	2.24	2.90	1.93	2.57
HG dataset	1.16	1.52	1.30	1.72

Table A8. TRIPOD checklist

Section/Topic	Item		Checklist Item	Page
Title and abstract				
Title	1	D;V	Identify the study as developing and/or validating a multivariable prediction model, the target population, and the outcome to be predicted.	1
Abstract	2	D;V	Provide a summary of objectives, study design, setting, participants, sample size, predictors, outcome, statistical analysis, results, and conclusions.	1
Introduction				
Background and objectives	3a	D;V	Explain the medical context (including whether diagnostic or prognostic) and rationale for developing or validating the multivariable prediction model, including references to existing models.	2-5
	3b	D;V	Specify the objectives, including whether the study describes the development or validation of the model or both.	2-5
Methods				
Source of data	4a	D;V	Describe the study design or source of data (e.g., randomized trial, cohort, or registry data), separately for the development and validation data sets, if applicable.	5-8
	4b	D;V	Specify the key study dates, including start of accrual; end of accrual; and, if applicable, end of follow-up.	9
Participants	5a	D;V	Specify key elements of the study setting (e.g., primary care, secondary care, general population) including number and location of centres.	5-8
	5b	D;V	Describe eligibility criteria for participants.	9
	5c	D;V	Give details of treatments received, if relevant.	9-10
Outcome	6a	D;V	Clearly define the outcome that is predicted by the prediction model, including how and when assessed.	3-5
	6b	D;V	Report any actions to blind assessment of the outcome to be predicted.	-
Predictors	7a	D;V	Clearly define all predictors used in developing or validating the multivariable prediction model, including how and when they were measured.	11-12
	7b	D;V	Report any actions to blind assessment of predictors for the outcome and other predictors.	-
Sample size	8	D;V	Explain how the study size was arrived at.	9-10
Missing data	9	D;V	Describe how missing data were handled (e.g., complete-case analysis, single imputation, multiple imputation) with details of any imputation method.	9-10
Statistical analysis methods	10a	D	Describe how predictors were handled in the analyses.	11-12
	10b	D	Specify type of model, all model-building procedures (including any predictor selection), and method for internal validation.	13
	10c	V	For validation, describe how the predictions were calculated.	14-16
	10d	D;V	Specify all measures used to assess model performance and, if relevant, to compare multiple models.	14-19
	10e	V	Describe any model updating (e.g., recalibration) arising from the validation, if done.	13
Risk groups	11	D;V	Provide details on how risk groups were created, if done.	9
Development vs. validation	12	V	For validation, identify any differences from the development data in setting, eligibility criteria, outcome, and predictors.	14-19
Results				
Participants	13a	D;V	Describe the flow of participants through the study, including the number of participants with and without the outcome and, if applicable, a summary of the follow-up time. A diagram may be helpful.	5-8
	13b	D;V	Describe the characteristics of the participants (basic demographics, clinical features, available predictors), including the number of participants with missing data for predictors and outcome.	9-10
	13c	V	For validation, show a comparison with the development data of the distribution of important variables (demographics, predictors and outcome).	-
Model development	14a	D	Specify the number of participants and outcome events in each analysis.	9-10
	14b	D	If done, report the unadjusted association between each candidate predictor and outcome.	-
Model specification	15a	D	Present the full prediction model to allow predictions for individuals (i.e., all regression coefficients, and model intercept or baseline survival at a given time point).	-
	15b	D	Explain how to the use the prediction model.	13
Model performance	16	D;V	Report performance measures (with CIs) for the prediction model.	16-17, 25
Model-updating	17	V	If done, report the results from any model updating (i.e., model specification, model performance).	14-25
Discussion				
Limitations	18	D;V	Discuss any limitations of the study (such as nonrepresentative sample, few events per predictor, missing data).	31, 33-34
Interpretation	19a	V	For validation, discuss the results with reference to performance in the development data, and any other validation data.	16-25
	19b	D;V	Give an overall interpretation of the results, considering objectives, limitations, results from similar studies, and other relevant evidence.	16-25
Implications	20	D;V	Discuss the potential clinical use of the model and implications for future research.	20-32
Other information				
Supplementary information	21	D;V	Provide information about the availability of supplementary resources, such as study protocol, Web calculator, and data sets.	S1
Funding	22	D;V	Give the source of funding and the role of the funders for the present study.	34

References

Seiffert, C., Khoshgoftaar, T. M., Van Hulse, J., and Napolitano, A. 2010. "Rusboost: A Hybrid Approach to Alleviating Class Imbalance," *IEEE Transactions on Systems Man and Cybernetics Part A-Systems and Humans* (40:1), 185-197.

Freund, Y., and Schapire, R. E. 1996. "Experiments with a New Boosting Algorithm," *Proceedings of the Thirteenth International Conference on International Conference on Machine Learning*, Bari, Italy: Morgan Kaufmann Publishers Inc., 148–156.

He, H. B., Bai, Y., Garcia, E. A., and Li, S. T. 2008. "Adasyn: Adaptive Synthetic Sampling Approach for Imbalanced Learning," *2008 IEEE International Joint Conference on Neural Networks, Vols 1-8*, 1322-1328.