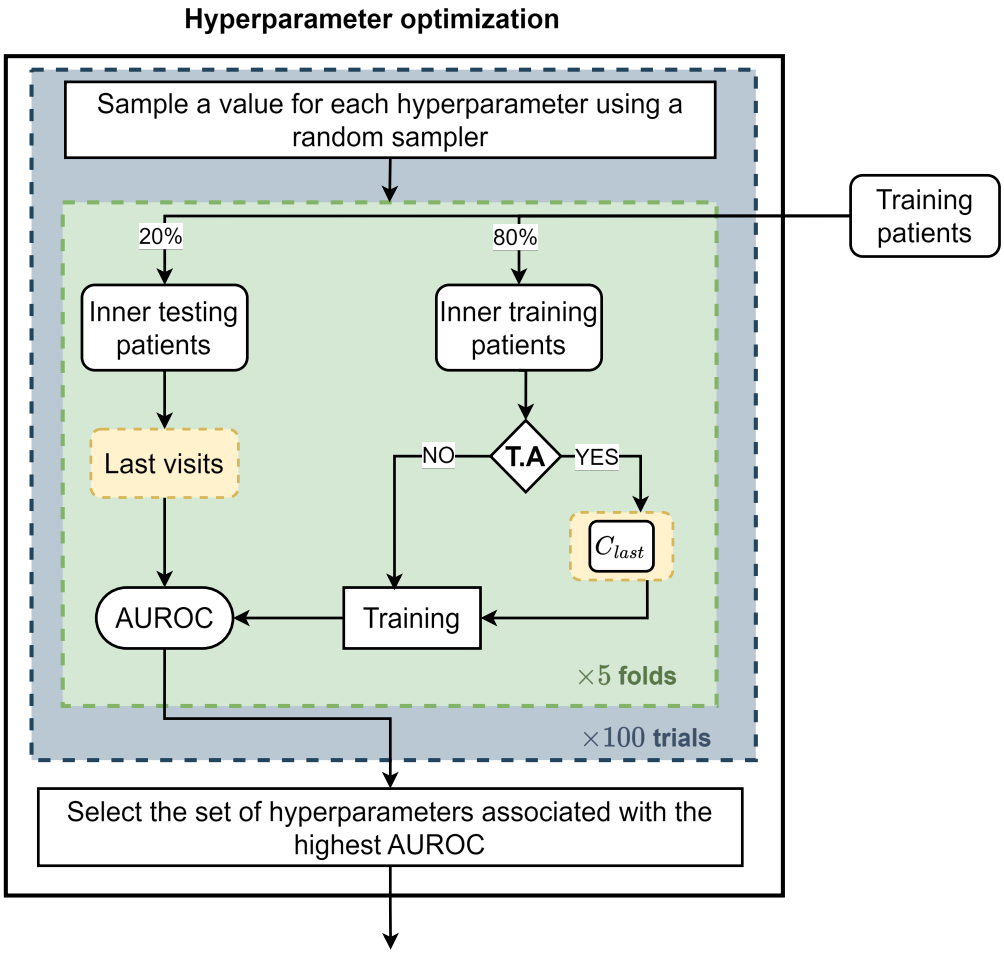
Supplementary materials



T.A: Temporal Analysis

Figure 1: Hyperparameter optimization for each model. The process is performed automatically using a random sampler from predefined search spaces for each hyperparameter, within the framework of the Optuna [1] Python library. The training patients are divided using a 5-fold cross-validation into inner training patients and inner testing patients. Temporal models are trained using the temporal cohort C_{last}. All models are tested on the last visits of patients. A total of 100 sets of hyperparameter values are sampled sequentially and evaluated on the same inner testing patients. The performance is measured using the mean of the AUROC on the 5 inner testing sets. The set of hyperparameters associated with the highest AUROC is used to train the model on the outer training patients.

Table 1: Random Forest' hyperparameters. The hyperparameters that are not mentioned were set as the default ones from the scikit-learn wrapper interface of version 0.8.0 of the *skranger* library[1]. The *weight* hyperparameter represents the weight of the positive class.

Hyperparameter	Search space
n_estimators	$\{128, 256, ..1024\}$
mtry	$\{10, 15, 20\}$
min_node_size	$\{10, 20, ..80\}$
weight	$[0.1, 0.9]$

Table 2: LSTMS' hyperparameters (BLSTM and LSTM_k). The *weight decay* refers to the coefficient multiplying the $\mathcal{L}_2$ penalty in the cross entropy loss. The *learning rate* refers to the initial learning rate given to the Adam optimizer [2] at the beginning of the training. The *hidden_size* refers to the number of neurons in the hidden layer. The *weight* hyperparameter represents the weight of the positive class.

Hyperparameter	Search space
weight decay	$\left[0, 10^{-4}\right]$
learning rate	$\left[10^{-5}, 10^{-3}\right]$
hidden_size	$\{16, 32, 48, 64\}$
weight	$[0.1, 0.9]$

[1] https://pypi.org/project/skranger/

Table 3: Descriptive analysis of the demographics and admission characteristics features along with four major comorbidities on the full dataset. We present the mode of each categorical feature along with its proportion in the dataset, and the mean of each continuous feature along with its standard deviation. The p-values are computed using the Welch's t-test [3] for continuous features (age, ambulance admissions count, ED visits count, weeks recently hospitalized) and the Pearson's chi-squared test [4] for categorical and binary features with the scipy [5] Python library.

Variable	All (n=250,812)	Survivors (n=214,095)	Deceased (n=36,717)	p-value
Demographics				
Age	61.12 ± 20.07	58.98 ± 20.18	73.62 ± 13.93	< 0.001
Sex	Female (54 %)	Female (55 %)	Male (54 %)	< 0.001
Admission characteristics				
Ambulance admission	0 (71 %)	0 (74 %)	1 (51 %)	< 0.001
Flu season	0 (75 %)	0 (75 %)	0 (74 %)	< 0.001
ICU admission	0 (97 %)	0 (97 %)	0 (95 %)	< 0.001
Urgent 30-d readmission	0 (90 %)	0 (92 %)	0 (79 %)	< 0.001
Ambulance admissions count	0.23 ± 0.75	0.17 ± 0.64	0.56 ± 1.19	< 0.001
ED visits count	0.8 ± 1.52	0.7 ± 1.42	1.38 ± 1.89	< 0.001
Weeks recently hospitalized	0.29 ± 0.99	0.21 ± 0.85	0.71 ± 1.5	< 0.001
Living status	Home (48 %)	Unknown (50 %)	Home (59 %)	< 0.001
Admission service	Cardiology (13 %)	Obstetrics (15 %)	I.M (14 %)	< 0.001
Admission type	Urgent (65 %)	Urgent (60 %)	Urgent (89 %)	< 0.001
Major comorbidities				
Dementia	0 (97 %)	0 (98 %)	0 (93 %)	< 0.001
Congestive heart failure	0 (94 %)	0 (95 %)	0 (86 %)	< 0.001
Metastatic solid cancer	0 (98 %)	0 (99 %)	0 (91 %)	< 0.001
Asthma	0 (97 %)	0 (97 %)	0 (96 %)	< 0.001

I.M: Internal Medicine.

Table 4: Descriptive analysis of the demographics and admission characteristics features along with four major comorbidities on the learning set. We present the mode of each categorical feature along with its proportion in the dataset, and the mean of each continuous feature along with its standard deviation. The p-values are computed using the Welch's t-test [3] for continuous features (age, ambulance admissions count, ED visits count, weeks recently hospitalized) and the Pearson's chi-squared test [4] for categorical and binary features with the scipy [5] Python library.

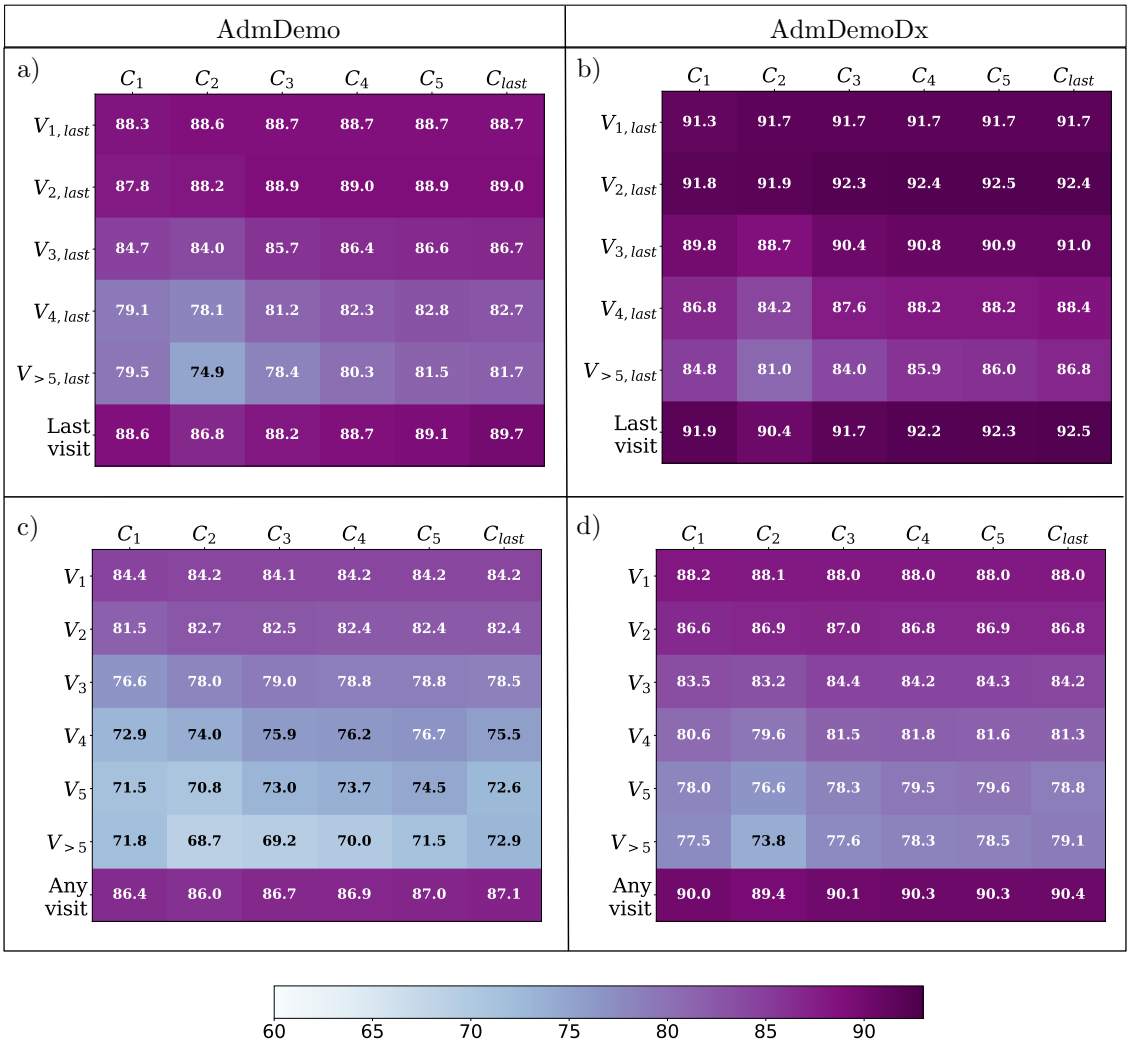
Variable	All (n=148,587)	Survivors (n=127,996)	Deceased (n=20,591)	p-value
Demographics				
Age	60.52 ± 20.27	58.45 ± 20.36	73.34 ± 13.99	< 0.001
Sex	Female (54 %)	Female (56 %)	Male (54 %)	< 0.001
Admission characteristics				
Ambulance admission	0 (70 %)	0 (74 %)	1 (52 %)	< 0.001
Flu season	0 (75 %)	0 (75 %)	0 (74 %)	0.003
ICU admission	0 (97 %)	0 (97 %)	0 (95 %)	< 0.001
Urgent 30-d readmission	0 (91 %)	0 (92 %)	0 (79 %)	< 0.001
Ambulance admissions count	0.24 ± 0.78	0.18 ± 0.66	0.59 ± 1.26	< 0.001
ED visits count	0.86 ± 1.57	0.75 ± 1.47	1.52 ± 1.99	< 0.001
Weeks recently hospitalized	0.29 ± 0.98	0.22 ± 0.86	0.73 ± 1.47	< 0.001
Living status	Unknown (49 %)	Unknown (52 %)	Home (61 %)	< 0.001
Admission service	Obstetrics (13 %)	Obstetrics (15 %)	F.M (16 %)	< 0.001
Admission type	Urgent (64 %)	Urgent (60 %)	Urgent (89 %)	< 0.001
Major comorbidities				
Dementia	0 (97 %)	0 (98 %)	0 (93 %)	< 0.001
Congestive heart failure	0 (94 %)	0 (95 %)	0 (85 %)	< 0.001
Metastatic solid cancer	0 (98 %)	0 (99 %)	0 (91 %)	< 0.001
Asthma	0 (97 %)	0 (97 %)	0 (96 %)	< 0.001

F.M: Family Medicine.

Table 5: Descriptive analysis of the demographics and admission characteristics features along with four major comorbidities on the holdout set. We present the mode of each categorical feature along with its proportion in the dataset, and the mean of each continuous feature along with its standard deviation. The p-values are computed using the Welch's t-test [3] for continuous features (age, ambulance admissions count, ED visits count, weeks recently hospitalized) and the Pearson's chi-squared test [4] for categorical and binary features with the scipy [5] Python library.

Variable	All (n=49,318)	Survivors (n=42,285)	Deceased (n=7,033)	p-value
Demographics				
Age	64.07 ± 16.54	62.85 ± 16.61	71.35 ± 14.02	< 0.001
Sex	Male (53 %)	Male (53 %)	Male (56 %)	< 0.001
Admission characteristics				
Ambulance admission	0 (73 %)	0 (75 %)	0 (61 %)	< 0.001
Flu season	0 (75 %)	0 (75 %)	0 (73 %)	0.002
ICU admission	0 (96 %)	0 (96 %)	0 (93 %)	< 0.001
Urgent 30-d readmission	0 (91 %)	0 (92 %)	0 (81 %)	< 0.001
Ambulance admissions count	0.11 ± 0.41	0.08 ± 0.36	0.24 ± 0.64	< 0.001
ED visits count	0.48 ± 1.07	0.42 ± 1.01	0.8 ± 1.35	< 0.001
Weeks recently hospitalized	0.22 ± 0.86	0.17 ± 0.75	0.56 ± 1.32	< 0.001
Living status	Unknown (72 %)	Unknown (75 %)	Unknown (59 %)	< 0.001
Admission service	Cardiology (15 %)	Cardiology (17 %)	H/O (17 %)	< 0.001
Admission type	Urgent (71 %)	Urgent (68 %)	Urgent (88 %)	< 0.001
Major comorbidities				
Dementia	0 (99 %)	0 (99 %)	0 (97 %)	< 0.001
Congestive heart failure	0 (98 %)	0 (99 %)	0 (96 %)	< 0.001
Metastatic solid cancer	0 (98 %)	0 (99 %)	0 (92 %)	< 0.001
Asthma	0 (99 %)	0 (99 %)	0 (98 %)	< 0.001

H/O: Hematology / Oncology



$V_{t,\text{last}}$: t^{th} visits of patients having exactly t visits; $V_{>t,\text{last}}$: last visits of patients having more than t visits; Last visit: last visits of all patients; V_t: t^{th} visits of patients having at least t visits; $V_{>t}$: random visit that occurred after the t^{th} visit for patients having more than t visits; Any visit: one visit per patient in the testing set selected randomly.

Figure 2: Performance of each LSTM_k trained with a cohort C_k on different groups of patients. (a) and (b) Performance on the last visits of patients. (c) and (d) Performance on the last and intermediary visits of patients. The rows represent the testing patients and the columns represent the training cohorts. The scores in the intersection of a row and a column correspond to the *mean* of the AUROC over the 5 folds of cross-validation of an LSTM trained with the corresponding cohort and tested on the corresponding patients.

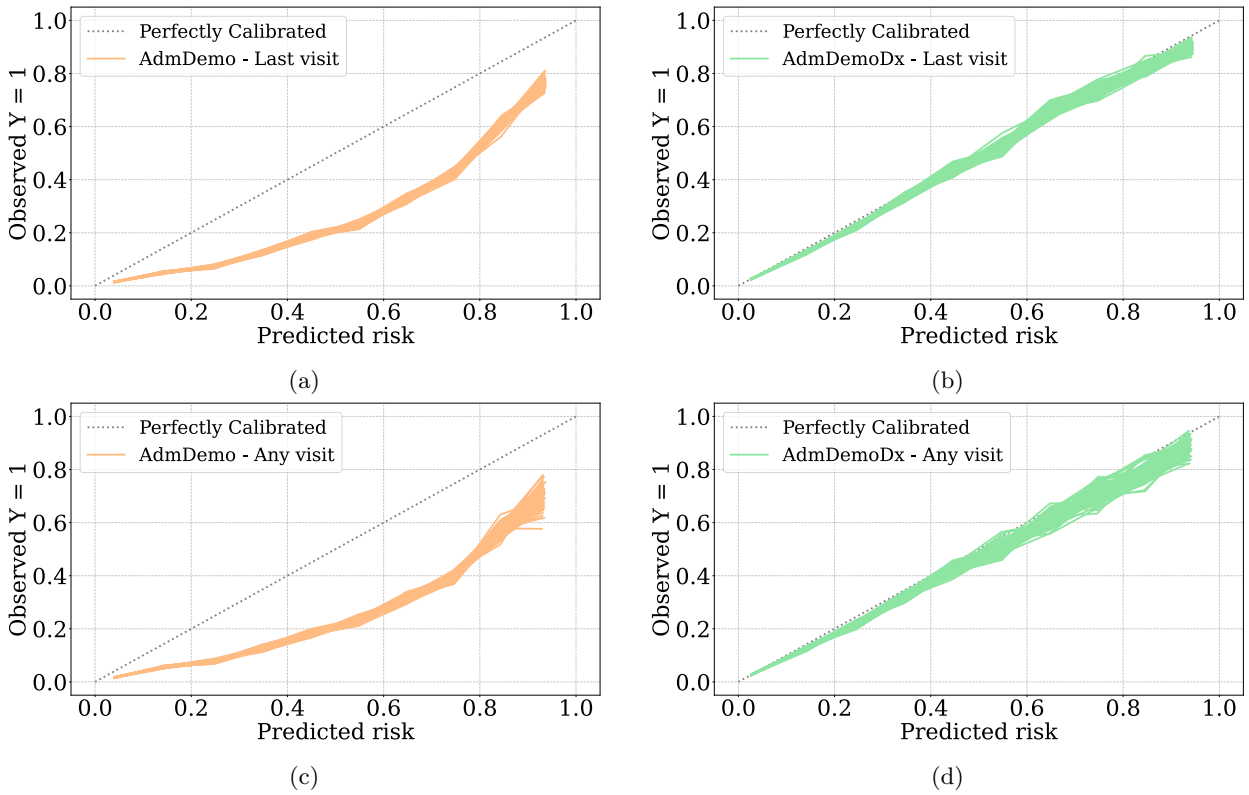


Figure 3: Calibration curves of the ELSTM with AdmDemo and AdmDemoDx predictors for each of the 100 bootstraps on the holdout set.

Table 6: Temporal validity of the ELSTM. The ELSTM is trained with patients admitted between july 1, 2011 and june 30, 2017 and tested on patients who are only admitted between july 1, 2017 and june 30, 2021, without excluding those who are not eligible for a GOC discussion. The goal is to evaluate the ELSTM when using the same rules for excluding visits for training and testing, but with data from different time periods. (a) Performance on the last visits of patients. (b) Performance on the last and intermediary visits of patients. The scores correspond to the *mean* ± *standard deviation* of the AUROC over 100 bootstraps.

(a)

Patients group	AdmDemo	AdmDemoDx
$V_{1,last}$	86.3 ± 0.4	89.6 ± 0.3
$V_{2,last}$	88.3 ± 0.5	91.6 ± 0.4
$V_{3,last}$	85.8 ± 0.9	88.7 ± 0.9
$V_{4,last}$	84.0 ± 1.4	88.9 ± 1.1
$V_{5,last}$	86.1 ± 1.8	90.4 ± 1.6
$V_{>5,last}$	81.5 ± 1.7	85.4 ± 1.5
Last visit	88.2 ± 0.2	91.1 ± 0.2

(b)

Patients group	AdmDemo	AdmDemoDx
V_1	84.0 ± 0.3	87.8 ± 0.3
V_2	84.4 ± 0.4	88.1 ± 0.4
V_3	81.5 ± 0.8	85.6 ± 0.7
V_4	79.0 ± 1.0	84.7 ± 1.0
V_5	78.8 ± 1.5	84.3 ± 1.4
$V_{>5}$	79.0 ± 1.8	83.4 ± 1.5
Any visit	86.3 ± 0.3	89.5 ± 0.2

$V_{t,\text{last}}$: t^{th} visits of patients having exactly t visits; $V_{>t,\text{last}}$: last visits of patients having more than t visits; Last visit: last visits of all patients; V_t: t^{th} visits of patients having at least t visits; $V_{>t}$: random visit that occurred after the t^{th} visit for patients having more than t visits; Any visit: one visit per patient in the testing set selected randomly.

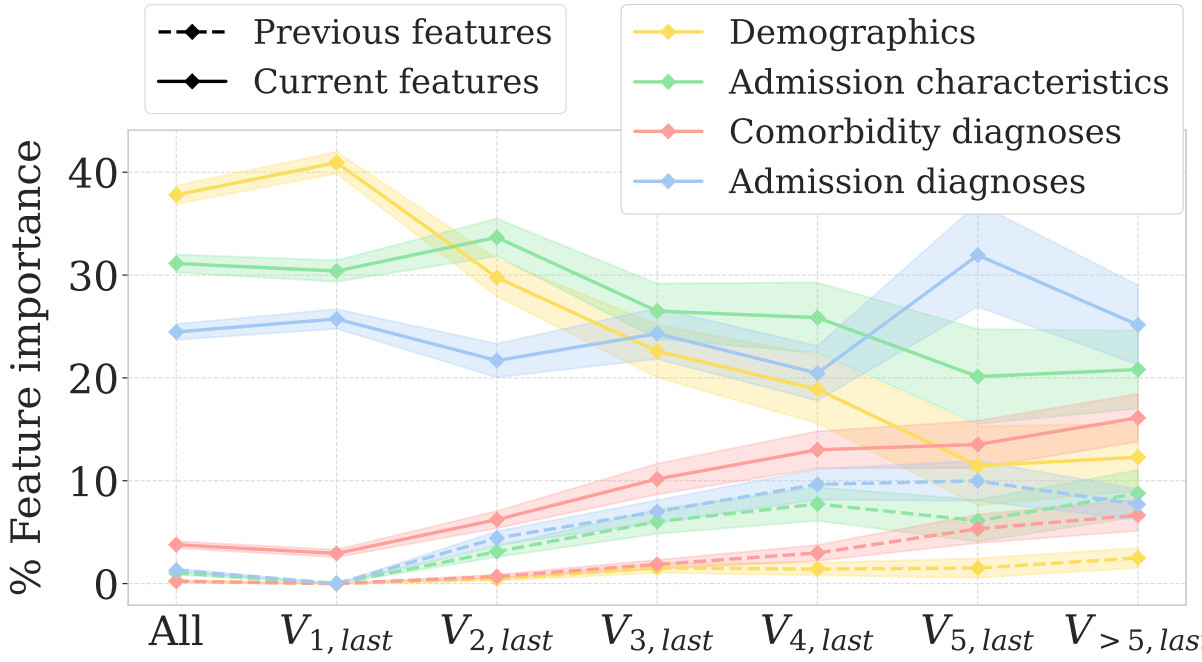


Figure 4: Post-hoc analyses of feature importance of the ELSTM trained with AdmDemoDx predictors when including the time gap between current and previous admissions as a predictor. Importance of each feature is computed using feature permutation [6] over 100 bootstraps. Shaded regions indicate variations within one standard deviation of the mean over 100 bootstraps. Importance of previous features increases as the size of patients' history gets longer.

Table 7: Performance of the ELSTM with AdmDemoDx predictors when including the time gap between current and previous admissions as a predictor.

	AUROC	Sensitivity	Specificity	Precision	NPV
Last visit	88.9 ± 0.3	79.5 ± 0.6	82.0 ± 0.2	38.4 ± 0.5	96.6 ± 0.1
Any visit	87.1 ± 0.3	75.3 ± 0.6	82.2 ± 0.2	34.1 ± 0.5	96.5 ± 0.1

Last visit: last visits of all patients; Any visit: one visit per patient in the holdout set selected randomly.

Table 8: Performance of the ELSTM using AdmDemoDx predictors on population subgroups of the holdout set. (a) Performance of the ELSTM on subpopulations of different age groups. (b) Performance of the ELSTM on subpopulations of males and females. The scores correspond to the *mean* $\pm$ *standard deviation* of the AUROC over 100 bootstraps.

(a)

	Age ≤ 50	$50 <$ Age < 65	Age ≥ 65
Last visit	91.5 ± 0.9	90.3 ± 0.6	84.9 ± 0.4
Any visit	89.8 ± 1.0	88.5 ± 0.7	82.8 ± 0.4

(b)

	Males	Females
Last visit	88.7 ± 0.4	89.4 ± 0.4
Any visit	87.0 ± 0.4	87.6 ± 0.4

Last visit: last visits of all patients; Any visit: one visit per patient in the testing set selected randomly.

References

[1] Takuya Akiba et al. "Optuna: A next-generation hyperparameter optimization framework". In: *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*. 2019, pp. 2623–2631.

[2] Diederik P Kingma and Jimmy Ba. "Adam: A method for stochastic optimization". In: *arXiv preprint arXiv:1412.6980* (2014).

[3] Bernard L Welch. "The generalization of 'STUDENT'S'problem when several different population varlances are involved". In: *Biometrika* 34.1-2 (1947), pp. 28–35.

[4] Karl Pearson. "X. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling". In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 50.302 (1900), pp. 157–175.

[5] Pauli Virtanen et al. "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python". In: *Nature Methods* 17 (2020), pp. 261–272. DOI: 10.1038/s41592-019-0686-2.

[6] Aaron Fisher, Cynthia Rudin, and Francesca Dominici. "All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously." In: *J. Mach. Learn. Res.* 20.177 (2019), pp. 1–81.