Leveraging patients' longitudinal data to improve the Hospital One-year Mortality Risk

Hakima Laribi, MSc[1], Nicolas Raymond, MSc[1], Ryeyan Taseen, MD, MSc[2], Dan Poenaru, MD, MHPE, MA, PhD[3,4], Martin Vallières, PhD[1,5,*]

[1] Department of Computer Science, Université de Sherbrooke, Sherbrooke, Canada
[2] Department of Medicine, Cambridge Memorial Hospital, Cambridge, Canada
[3] Department of Pediatric Surgery, McGill University Health Centre, Montreal, Canada
[4] Centre for Outcomes Research and Evaluation (CORE), Research Institute of the McGill University Health Centre, Montreal, Canada
[5] Centre de recherche du Centre hospitalier universitaire de Sherbrooke, Sherbrooke, Canada

* Corresponding author: Martin Vallières, PhD; Université de Sherbrooke, 2500, Boulevard de l'Université, Faculté des Sciences, Local D4-2005, Sherbrooke (Qc), Canada J1K 2R1; martin.vallieres@usherbrooke.ca; +1-819-821-8000 ext: 65116

Word count:
Abstract: 222
Main text: 3997

Key words:
Machine learning; Long Short-Term Memory neural networks; Longitudinal data, Mortality risk; Administrative data.

ABSTRACT

Objective

To develop and validate an Ensemble Long Short-term Memory neural network (ELSTM) that integrates patients' longitudinal data to predict the Hospital One-year Mortality Risk using patients' information collected routinely at admission. The aim is to identify patients at the end of life who may benefit from goals of care (GOC) discussions.

Materials and Methods

We evaluated our ELSTM (i) when including only predictors that can be reported upon admission (AdmDemo), and (ii) when adding also diagnoses available later during patients' stay (AdmDemoDx). We used records of 82,104 patients admitted between 2011 and 2017 to compare the temporal and non-temporal strategies. We also quantified the clinical utility of the best strategy on 33,898 patients eligible for GOC discussions admitted between 2017 and 2021.

Results

Our ELSTM used with AdmDemo and AdmDemoDx predictors demonstrated an increased performance with AUROCs between 0.73-0.90 and 0.79-0.93, respectively. The ELSTM-based decision-making increased prediction precision by up to 12.1% compared to the usual decision-making process, but it also reduced sensitivity by up to 3.8%.

Discussion

The integration of patients' longitudinal data provides better insights into the severity of illness and the overall condition of patients, especially when limited information is available during their hospitalization.

Conclusion

The proposed ELSTM is an automated and accurate model able to identify patients at high risk of one-year mortality, potentially usable in clinical decision support systems to improve end-of-life care.

1 BACKGROUND AND SIGNIFICANCE

Estimating the life expectancy of patients helps identifying high-risk individuals and improve the quality of care they receive in hospital settings.[1–3] Unlike patients with cancer who receive palliative care in their final months of life, patients with other less predictable conditions are only referred for these services in their final weeks or days, if at all.[4] In Canada, despite common individual preference for most individuals to die in community and other home-like settings,[5] 58% of those who died in 2015 were hospitalized more than once in their last year of life, and 61% died in hospital.[6] An early identification of these high-risk patients would allow important discussions with healthcare providers regarding end-of-life choices, to align their preferences with the care they receive.[7] Such discussions would enable goals-of-care (GOC) documentation, including Code Status Orders (CSOs) clarifying essential preferences for life-supporting therapy.[8,9] Early identification would also facilitate communication between clinicians and families regarding patients' life trajectories, ensuring informed shared decision-making[10] and potentially reduce depression and grief.[11] However, a clear and timely prognostication of high-risk patients in hospital settings is time-consuming and therefore challenging for workload-burdened clinicians.[12] An accurate automated tool not requiring human involvement could initially flag these patients, lightening the work burden of the clinical team.

Several studies have investigated the ability of data available in Electronic Health Records (EHRs) to predict the mortality risk of patients, potentially driving an automated clinical decision support system. *van Walraven et al*[13,14] introduced the Hospital One-year Mortality Risk (HOMR) score, representing the probability of death within one year of patient's admission. The original model consisted of a logistic regression using post-discharge administrative data routinely collected upon admission, evaluated using Area Under the Receiver Operating Characteristic curve (AUROC). Their goal was to flag high-risk individuals and initiate end-of-life discussions with them to decide in favor or against potentially aggressive and invasive interventions. To operate in real-time, subsequent versions modified the HOMR score according to the availability of data in each hospital, and included only variables available immediately when patients were admitted.[15,16] As a result, due to specific EHRs constraints, diagnostic codes were omitted from the predictors. More recently, *Taseen and Ethier*[9] explored the clinical utility of models predicting the HOMR score, in which they developed three random forest models based on variable sets available at different times during a patient's admission. The authors compared the discriminative power of such models with previously established linear regression models and evaluated their clinical utility within their hospital setting.

Nevertheless, these studies did not include valuable longitudinal information present in patients' records, as they focus on single visits and do not take into account the patient's history from previous hospital admissions. This approach diverges from the clinical reality, where clinicians consistently consider the entire patient history before making any prognostic prediction for any condition. Another approach has been to incorporate broader covariates (e.g., medical disease codes, clinicians' notes, social history) and aggregate patient information within and across admissions to predict their mortality risk in order to refer them for end-of-life care.[17,18] However, these studies did not explicitly quantify the impact of integrating patient history in developing more accurate solutions. Moreover, the proposed models are more challenging in terms of data acquisition and are therefore less likely to be deployed in a clinical decision support system — unlike HOMR-based models that have already been clinically deployed[16] or are in the process of deployment.[9]

2 OBJECTIVE

In this work, we have evaluated the benefits of integrating patients' longitudinal data to improve the accuracy of the HOMR score. We built on the work of *Taseen and Ethier*[9] by re-analyzing the same data routinely collected during patients' admissions, and also integrating additional recent visits. To assess the benefits of a temporal EHR analysis, we developed and compared a Long Short-Term Memory-based ensemble model (ELSTM) that leverages patients' longitudinal data, to baseline models that consider patients' visits independently, without including previous visits. Figure 1 shows an overview of our study. We further analyzed the predictive power of our model in two different scenarios with different requirements of data access: (i) including only demographics and admission characteristics available on patient's admission, and (ii) adding also admission diagnoses and comorbid diagnoses available during patient's hospitalization. In an effort to better inform about the clinical utility of such models, we quantified the gains and losses of our ELSTM in

terms of true and false positives as compared to standard human decision-making.

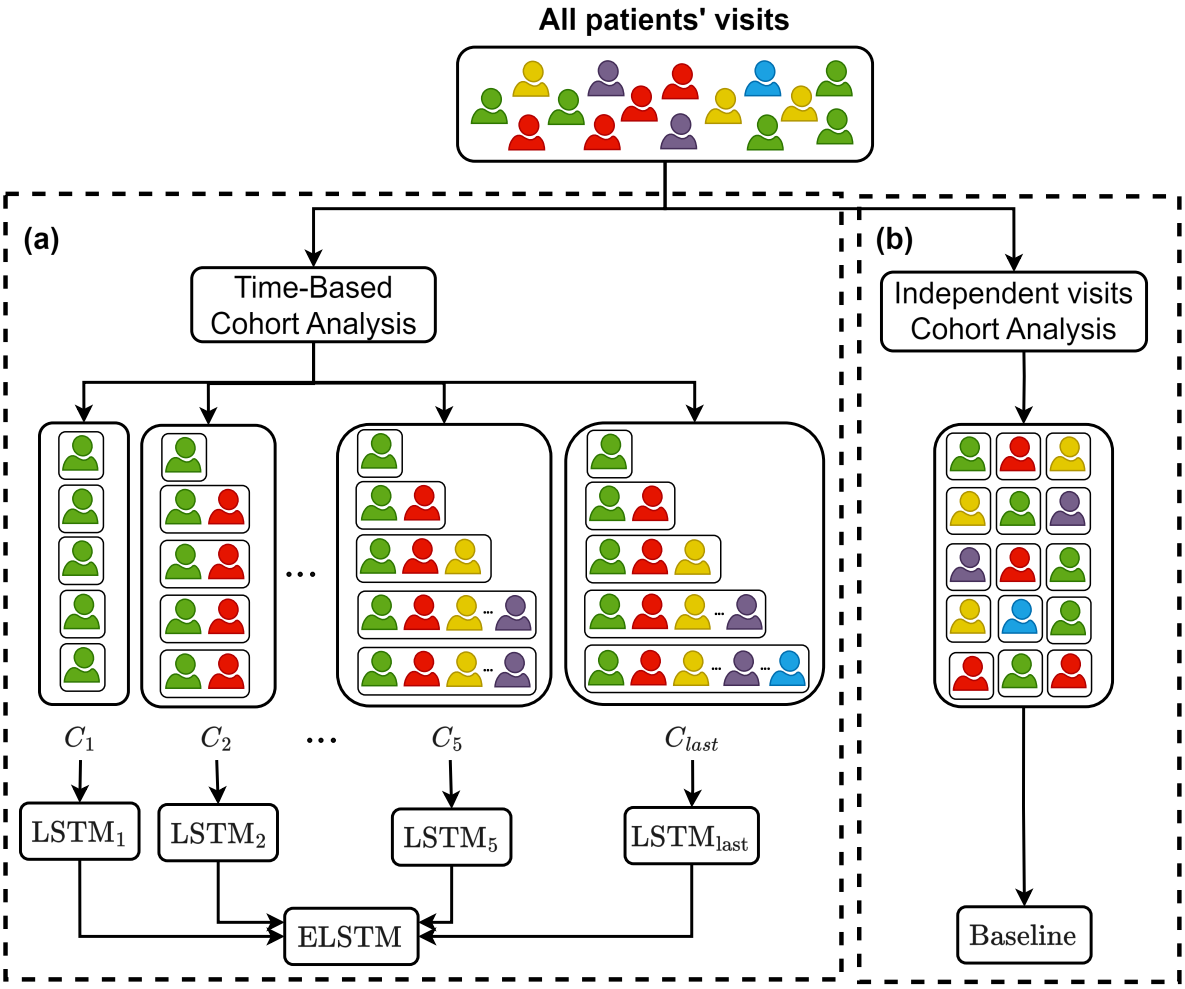


Figure 1: Study overview. (a) The ELSTM averages predictions of multiple LSTMs trained using different cohorts of the same patients. Each cohort includes the patient's history up to a specific visit. (b) Baseline models consider patients visits independently.

3 MATERIAL AND METHODS

3.1 Dataset

This retrospective study took place at an integrated university hospital network with 2 sites and 700 acute care beds in Sherbrooke, Quebec, Canada. Data were obtained from the institutional data warehouse, combining EHR and administrative information. The cohort included all adult patients admitted to a non-psychiatric service between July 1, 2011 and June 30, 2021, excluding admissions to infrequently admitting services (such as genetics) or admissions with a legal context (i.e. court-ordered). Mortality status was also extracted from the institutional data warehouse, which was sourced from the Quebec vital statistics registry. Institutional Review Board approval was obtained prior to data acquisition (Institutional Review Board of the CIUSSS de l'Estrie—CHUS Nagano #2022-4409). We followed the data extraction steps previously described by *Taseen and Ethier*[9] as we use the same source of data. Table 1 lists the predictors used for model comparisons. Comorbid diagnoses from prior visits became accessible in the information system 6 months following the respective visit, or only 2 weeks later for emergency department encounters.

Table 1: Covariates included in all predictive models as described in *Taseen and Ethier.*[9] AdmDemo predictors include only demographics and admission characteristics while AdmDemoDx predictors include demographics, admission characteristics and comorbid and admission diagnoses.

Group	Variable	Description
Demographics	Age	Age at admission in full years since birth.
	Sex	Sex at birth, female or male.
Admission characteristics	Ambulance admission	If the current admission is via ambulance.
	Flu season	If the current admission is in the month of December, January, or February.
	ICU admission	If the current admission is a direct admission to the ICU.
	Urgent 30-d readmission	If the current admission is an urgent readmission within 30 days of a previous discharge.
	Ambulance admissions count	Number of admissions to the hospital by ambulance in the year before admission.
	ED visits count	Number of visits to the emergency department in the year before admission.
	Weeks recently hospitalized	Number of full weeks hospitalized in the 90 days before admission.
	Admission service	Cardiac surgery, cardiology, critical care, endocrinology, family medicine, gastroenterology, general surgery, gynecology, hematology-oncology, internal medicine, maxillofacial surgery, nephrology, neurosurgery, neurology, obstetrics, ophthalmology, orthopedic surgery, otorhinolaryngology, palliative care, plastic surgery, respirology, rheumatology, thoracic surgery, trauma, urology, or vascular surgery.
	Admission type	Urgent, semi-urgent, elective, or obstetric.
	Living status	Living status at admission: chronic care hospital, nursing home, home, or unknown.
Comorbidity diagnoses	84 binary variables	ICD-10 codes from previous visits hospital discharge abstracts and emergency department information systems mapped to 84 binary variables.
	Visible comorbidities	If a previous hospitalization occurred between 5 years and 6 months before admission or if a previous visit to the emergency department occurred between 6 months and 2 weeks before admission. This binary variable flags the availability of comorbid diagnoses.
Admission diagnoses	147 binary variables	Free-text diagnosis on admission order form mapped to 147 binary variables using regular expressions.

Given the potential variations in data availability on admission across different hospital information systems, we explored the feasibility of early identification of high-risk patients in several scenarios. We evaluated two strategies with different data requirements: (i) "AdmDemo", including only demographics and admission characteristics and, (ii) "AdmDemoDx" including demographics, admission characteristics, comorbid diagnoses and admission diagnoses.

3.2 Ensemble Long Short-Term Memory neural network (ELSTM)

To evaluate the impact of incorporating a patient's longitudinal health record for improving the HOMR score, we introduce an Ensemble Long Short-Term Memory neural network (ELSTM) that leverages information

learned by multiple LSTMs trained at different stages of patients' admissions to hospital (Figure 1a1a) . We base our ensemble model on an LSTM architecture[19] since recurrent neural networks can handle sequences of different lengths without extra padding. This is particularly relevant in our case where patients can have varying numbers of previous visits.

More formally, we define C_k as the temporal cohort including the visits sequence of each patient up to their k^{th} visit; if a patient has less than k visits, C_k includes all their visits. C_{last} denotes the cohort including the visits sequence of each patient up to their last visit available in our dataset. The formal definition of C_k is given by:

$$C_k = \{\{V_j^i\}_{j=1}^{\min(k,M^i)}\}_{i=1}^{N} \tag{1}$$

where $N \in \mathbb{N}$ is the number of patients, $M^i \in \mathbb{N}$ the number of visits for the i^{th} patient and V_j^i the j^{th} visit of the i^{th} patient.

During the training phase, we train multiple LSTMs on temporal cohorts including patients with varying numbers of visits. The goal is to capture diverse information at different stages of patients' visit sequence. Each LSTM_k is trained using the temporal cohort C_k to aggregate a patient's visit sequence and estimate their mortality risk at their last visit available in C_k, with $k \in \{1, \ldots, \text{K}\} \cup \{\text{last}\}$. The ensemble model learns from multiple visits for each patient, while each LSTM_k is exclusively trained on a single visit sequence per patient. This setup guarantees that the training data for each LSTM_k are independent and identically distributed (iid). We set $K = 5$ given that only 5% of patients have more than 5 visits in our dataset. We chose not to restrict C_k to patients with only k visits in order to optimize each LSTM_k of the ensemble model on a larger set of data. Here, our assumption is that including patients with a full sequence of visits, even if the length was less than k, would make the distribution of training data more exhaustive and improve the model's predictive performance.

In the testing phase, the ELSTM averages the predictions of all LSTMs trained with patients having at least m visits to make a prediction at the m^{th} visit of a patient, as follows:

$$\text{ELSTM}(V_m^i) = \text{ELSTM}(\{V_j^i\}_{j=1}^{m}) = \frac{\sum_{k \in M} \text{LSTM}_k(\{V_j^i\}_{j=1}^{m})}{|M|} \tag{2}$$

with $M = \{k \in \{1, \ldots, \text{K}\} \mid k \geq m\} \cup \{\text{last}\}$.

3.3 Experimental setup

3.3.1 Baseline models

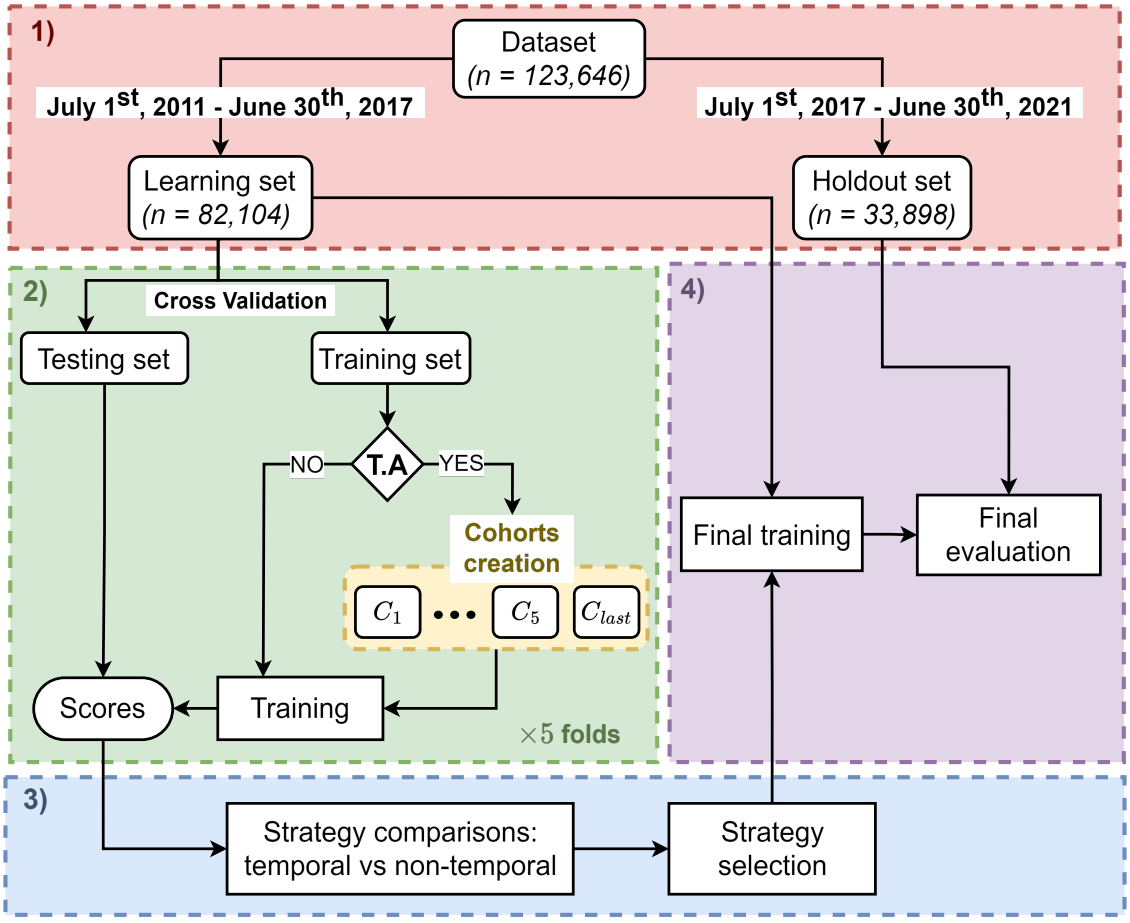
We conducted a comparative analysis of the ELSTM with two baseline models which do not use longitudinal data. The first model is the random forest (RF), as employed in prior work,[9] using the scikit-learn wrapper[20] from skranger library[1]. The second model is a basic LSTM (BLSTM) which does not consider previous information when making a prediction for a specific visit. Each LSTM-based model contains one single hidden layer followed by 2 fully connected layers and was implemented using the PyTorch library.[21] For a fair comparison, we added the visit count at each admission as a predictor to the baseline models.

3.3.2 Experimental design

We used the experimental setup illustrated in Figure 2 to evaluate the ELSTM and baseline models. The experiments are repeated for each group of predictors AdmDemo and AdmDemoDx. Following a similar approach to *Taseen and Ethier*,[9] we temporally split the dataset into a *learning set*, including admissions from July 1, 2011, to June 30, 2017, and a *holdout set*, including admissions from July 1, 2017, to June 30, 2021. We excluded patients admitted before June 30, 2017 from the holdout set to prevent data leakage. This design aimed to simulate the evaluation of a model trained on all available patients data and tested on subsequently admitted patients. As patients are exclusively in one set at a time, temporal models have only aggregated previous visits occurring within the last six years prior to the current admission. To evaluate the final model's clinical utility on the holdout set, we focused on the same population eligible for GOC discussions as in previous work.[9] Therefore, we excluded hospitalizations without an overnight stay from the

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

holdout set, since there would not be enough time for a GOC discussion to occur. Additionally, we omitted admissions to the obstetrics service, where such discussions are considered inappropriate, and admissions to the palliative care service, where GOC discussions have already occurred and are therefore unnecessary at this stage.



T.A: Temporal Analysis

Figure 2: Experimental setup for model comparisons and final evaluation. 1) Temporal division of the dataset into a *learning set* and a *holdout set*. 2) Evaluation of the predictive performance of each model on 5 *testing sets* using a 5-fold cross-validation over the patients of the learning set. The same data splits are used for all models. Baseline models (RF, BLSTM) are trained on all the visits, while each temporal model LSTM_k comprising the ELSTM is trained using a temporal cohort C_k. The scores are reported on specific patients of the *testing sets*. The training of each model includes the optimization of hyperparameters, except for the temporal models, for which we only optimize the hyperparameters of $\text{LSTM}_{\text{last}}$. Details on hyperparameter optimization are shown in Supplementary Figure 1. 3) Comparison of the temporal and non-temporal strategy and selection of the best strategy based on performance. The latter is measured using the mean and standard deviation of the scores on the 5 testing sets. 4) Final evaluation of the selected strategy on the holdout set. The final model predictions are then compared to usual care to quantify clinical utility.

3.3.3 Model selection procedure

In the model selection phase, we compare the performance of the ELSTM and baseline models to evaluate the benefits of incorporating the patients history in predicting their one-year mortality risk. To achieve this, we

used a nested 5-fold cross-validation scheme. We partitioned the learning set using a 5-fold cross-validation into distinct training and test sets. Each of the training sets was subsequently separated into distinct inner training and inner test sets with an inner 5-fold cross-validation. The inner sets were entirely dedicated to optimize the hyperparameters of the models for each outer training fold. The data splitting was based on patients rather than visits, ensuring that each patient exclusively belonged to one set at a time.

To train the LSTM-based models, we created an additional validation set (as well as an inner validation set) for each of the 5 cross-validation splits, enabling us to track model performance through training epochs and proceed to early stopping if necessary. Each (inner) validation set was created by randomly sampling 10% of patients from the corresponding (inner) training set.

At each (inner) training split, the baseline models are trained using all patients' visits, while each LSTM_k part of the ELSTM is trained using a temporal cohort C_k.

We assessed the benefits of the longitudinal data at each patient visit including their last visit available in our dataset, when we considered our patient's medical trajectory completed. We define V_t as all the t^{th} visits of patients having at least t visits, and $V_{t,last}$ as the last visits of patients having exactly t visits.

3.3.4 Final evaluation procedure

To evaluate the clinical utility of the best model selected in the previous phase, we compared its predictions to the usual care performed by clinicians on patients eligible for a GOC discussion. We aimed to quantify the gains and losses in terms of true positive and false positive alerts if this automated tool was used in a clinical decision support system to alert clinicians when a patient is identified as being at risk of one-year mortality. First, we extracted all CSOs of patients in the holdout set, and considered that a GOC occurred between a patient and a clinician (and that a patient at high risk of one-year mortality was identified by the clinical team) if a CSO was documented prior to the patient's discharge, whether during the current admission or a previous one. Similarly to *Taseen and Ethier*,[9] we defined:

- True Positives (TPs) as patients with a documented CSO who died within a year.
- False Positives (FPs) as patients with a documented CSO who survived beyond a year.
- False Negatives (FNs) as patients without a documented CSO who died within a year.
- True Negatives (TNs) as patients without a documented CSO who survived beyond a year.

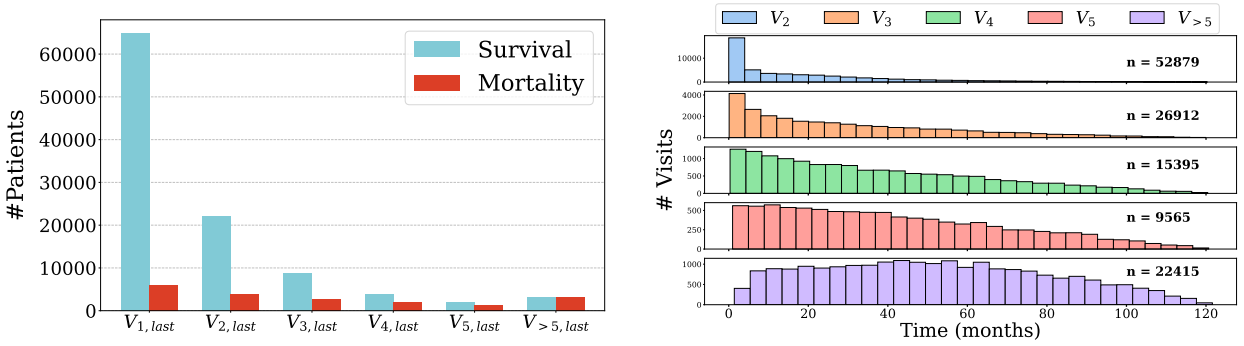
Next, we trained the previously selected model using the entire learning set and compared its predictions, which would represent the actions suggested by the automated tool, to the usual care the patients from the holdout set received.

3.3.5 Hyperparameters optimization

We optimized each model's hyperparameters to find the best set leading to the highest scores. We trained each LSTM-based model using the Adam optimizer[22] with parameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$, and a batch size of 100. We fixed the sizes of the fully connected layers to 2 and 1 respectively. Given that the ELSTM consists of multiple models, we chose to exclusively optimize the hyperparameters of LSTM_last and used the selected set to train each LSTM_k. This way, we ensured consistent probability scales within the models constituting the ensemble model. For each optimized model, we sampled 100 sets of hyperparameters values from predefined search spaces, using a random sampler from the Optuna Python library.[23] Each set of hyperparameters values was evaluated by training the model with the 5 inner training sets and then measuring the AUROC on their respective inner testing sets. Here, the inner test sets included only the last visit of each patient. The set associated with the highest AUROC was selected to train the model on the whole training set of the outer loop. Models' hyperparameters are provided in Supplementary Tables 1-2.

4 RESULTS

The overall cohort consisted of 123,646 patients and 250,812 hospitalizations, with 15% of patients experiencing mortality within one year of their last admission. The learning set included 82,104 patients and



(a) (b)

Figure 3: Distribution of visits across the entire dataset. (a) Proportions of survival and mortality per number of visits in the dataset. $V_{t,\text{last}}$ represents all the patients with exactly t visits in the dataset and $V_{>t,\text{last}}$ those with more than t visits. The number of patients decreases with the number of visits, in contrast to the mortality rate. (b) Distribution of visits over time after the first hospital discharge. V_t represents all the t^{th} visits in the dataset and $V_{>t}$ all the visits after the t^{th} visit.

148,587 hospitalizations, while the holdout set included 33,898 patients and 49,318 hospitalizations. Detailed descriptive analyses for each set can be found in Supplementary Tables 3-5. Figure 3a provides an overview of the proportion of mortality and survival per number of visits across the dataset. Patients who are frequently admitted to the hospital are generally fewer, but present a higher risk of one-year mortality. Figure 3b shows the distribution of visits over time after the first hospitalization discharge. The second and third visits occur mainly in the first months following the first hospital discharge, while subsequent visits are increasingly scattered across time.

4.1 Model selection on the learning set

In this part of our study, we explored the advantages of integrating patients' historical data to predict their HOMR score. We assessed the baselines and the ELSTM on the learning set at various stages of patients' hospital admissions, to understand the extent to which exploring patients' history proves beneficial. As described earlier, we considered two groups of predictors: AdmDemo and AdmDemoDx.

Table 2a presents the performance of the baseline models and ELSTM for the last visit of each patient. The ELSTM outperforms the non-temporal models with a higher AUROC for all patient groups with both sets of predictors. Statistical tests revealed a significant overall improvement, except for $V_{5,\text{last}}$ and $V_{>5,\text{last}}$, where we note a higher variance due to fewer patients ($\sim$ 300 and $\sim$ 500) that can diminish the statistical power of the test. Notably, even for patients without a historical record $V_{1,\text{last}}$, the temporal model was effective - emphasizing that absence of recurrent visits serves as valuable insight. Experiments in Table 2b show that the impact of longitudinal data is less pronounced on intermediate visits. We observe non-statistically significant increases or decreases, especially for AdmDemoDx predictors. Supplementary Figure 2 shows the performances of each individual LSTM_k within the ELSTM.

Next, the group of predictors with fewer variables (AdmDemo) achieved acceptable results for all patient sets with a predictably lower AUROC compared to AdmDemoDx (Tables 2a and 2b). The former seems to benefit more from the longitudinal data, as we observed a higher AUROC improvement across all patient sets compared to AdmDemoDx when using the ELSTM. This emphasizes the significance of incorporating longitudinal data in cases where premorbid variables are not available (e.g., comorbidity diagnoses), and their ability to provide a more comprehensive understanding of the patients through their history.

In addition, the ELSTM demonstrated an acceptable temporal validity across all patient groups when tested on patients admitted later in time (Supplementary Table 6). Overall, the ELSTM achieved the best performance for most patient groups, particularly on their last visits completing their medical trajectory. These results highlight the gains from integrating longitudinal patient data to predict the HOMR score.

Table 2: Performance of the baselines and the ELSTM on the testing sets of the learning set using both AdmDemo and AdmDemoDx predictors. (a) Performance on the last visits of patients. (b) Performance on the last and intermediate visits of patients. Each testing set from the 5-fold cross-validation was divided into different groups of patients according to their number of visits to evaluate when temporal modeling is beneficial. Each group of patients included a patient at most once. The scores correspond to the *mean* ± *standard deviation* of the AUROC over the 5 testing sets. For each group of patients, the highest AUROC is highlighted in bold. Significant difference was quantified using the one-sided Wilcoxon signed-rank test.[24] Each p-value corresponds to the significance of improvement of the ELSTM over the best baseline model for a specific group of patients.

(a)

	AdmDemo			AdmDemoDx		
Patients group	RF	BLSTM	ELSTM	RF	BLSTM	ELSTM
$V_{1,\text{last}}$	88.1 ± 0.4	88.2 ± 0.3	$\mathbf{88.7 \pm 0.3}$**	91.3 ± 0.2	91.4 ± 0.2	$\mathbf{91.8 \pm 0.3}$**
$V_{2,last}$	88.1 ± 0.5	88.4 ± 0.4	$\mathbf{89.1 \pm 0.5}$**	92.1 ± 0.6	92.0 ± 0.8	$\mathbf{92.6 \pm 0.7}$**
$V_{3,last}$	84.4 ± 0.5	85.2 ± 0.5	$\mathbf{86.8 \pm 0.9}$**	90.2 ± 0.8	90.4 ± 1.1	$\mathbf{91.1 \pm 1.0}$**
$V_{4,last}$	80.3 ± 1.1	80.6 ± 1.5	$\mathbf{83.3 \pm 0.8}$**	87.5 ± 0.9	87.7 ± 0.6	$\mathbf{88.7 \pm 0.9}$**
$V_{5,last}$	80.6 ± 4.1	81.7 ± 3.9	$\mathbf{82.5 \pm 4.4}$	86.1 ± 2.4	85.8 ± 2.3	$\mathbf{87.0 \pm 2.6}$*
$V_{>5,last}$	75.3 ± 1.6	75.0 ± 1.4	$\mathbf{75.6 \pm 1.4}$	81.7 ± 0.6	81.8 ± 0.5	$\mathbf{82.4 \pm 0.6}$*
Last visit	89.1 ± 0.2	89.2 ± 0.3	$\mathbf{89.8 \pm 0.2}$**	92.2 ± 0.2	92.3 ± 0.2	$\mathbf{92.6 \pm 0.3}$**

(b)

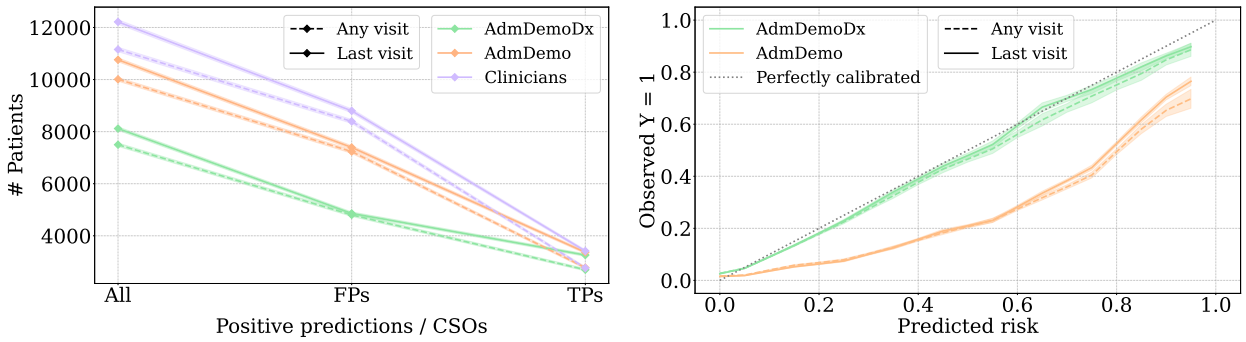
	AdmDemo			AdmDemoDx		
Patients group	RF	BLSTM	ELSTM	RF	BLSTM	ELSTM
V_1	84.2 ± 0.4	$\mathbf{84.5 \pm 0.3}$	84.3 ± 0.4	88.2 ± 0.2	$\mathbf{88.4 \pm 0.2}$	88.3 ± 0.3
V_2	82.3 ± 0.4	82.6 ± 0.2	$\mathbf{82.8 \pm 0.4}$**	$\mathbf{87.2 \pm 0.5}$	87.1 ± 0.4	$\mathbf{87.2 \pm 0.5}$
V_3	77.5 ± 0.5	78.0 ± 0.6	$\mathbf{79.3 \pm 0.7}$**	84.4 ± 0.7	84.6 ± 0.8	$\mathbf{84.7 \pm 0.7}$**
V_4	74.5 ± 0.5	75.0 ± 0.8	$\mathbf{76.9 \pm 0.6}$**	81.9 ± 0.5	81.9 ± 0.3	$\mathbf{82.1 \pm 0.6}$
V_5	73.4 ± 2.2	73.7 ± 2.1	$\mathbf{74.3 \pm 1.7}$	79.5 ± 1.1	79.4 ± 1.4	$\mathbf{79.9 \pm 1.3}$
$V_{>5}$	73.0 ± 1.1	$\mathbf{73.7 \pm 0.8}$	72.9 ± 0.7	$\mathbf{79.9 \pm 0.8}$	79.6 ± 0.7	79.1 ± 1.6
Any visit	86.8 ± 0.2	87.0 ± 0.2	$\mathbf{87.3 \pm 0.3}$**	90.3 ± 0.1	90.3 ± 0.2	$\mathbf{90.6 \pm 0.2}$**

$V_{t,\text{last}}$: 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; $V_{>t}$: one visit selected randomly 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.
* p-value < 0.1; ** p-value < 0.05

4.2 Final evaluation on patients eligible for a GOC discussion

In this section, we compared the ELSTM using AdmDemo or AdmDemoDx predictors with the usual care performed by clinicians for each patient in the holdout set. We optimized the decision threshold for considering a patient at risk of one-year mortality by maximizing the Youden's J index.[25] We set it at 0.34 for ELSTM-AdmDemo and 0.17 for ELSTM-AdmDemoDx.

Results in Table 3 revealed that the ELSTM with AdmDemo predictors constitutes an automated tool with similar predictions to the usual care performed by clinicians, with overall good precision and not too many inappropriate alerts relative to daily clinical practice.



(a) (b)

FPs: False Positives; TPs: True Positives.

Figure 4: Analyses of the final ELSTM tested on the holdout set with both AdmDemo and AdmDemoDx predictors. Shaded regions indicate variations within one standard deviation of the mean over 100 bootstraps. (a) Number of positive predictions by the ELSTM with AdmDemo and AdmDemoDx predictors, and CSOs documented by clinicians. The ELSTM shows a reduced number of false positives and a slight loss in the number of true positives. (b) ELSTM calibration curves with AdmDemo and AdmDemoDx predictors. We used interpolation to unify the predicted risk bins over the 100 bootstraps and generate a mean calibration curve with its variations. Supplementary Figure 3 shows the 100 calibration curves for each ELSTM in the bootstrap sampling. The ELSTM-AdmDemoDx is almost identical to a perfectly calibrated model, while the ELSTM-AdmDemo tends to overestimate the risk of mortality.

We also observe that, although the ELSTM with AdmDemoDx predictors achieved the highest AUROC, the model is less sensitive and detects slightly fewer patients who actually died within a year of their admission (Figure 4a). Nevertheless, this model considerably reduced the number of false positive notifications and increased the precision. The calibration curves in Figure 4b support this result, by showing a tendency of ELSTM-AdmDemo to overestimate the risk of death compared to ELSTM-AdmDemoDx.

Table 3: Comparisons of the final ELSTM with AdmDemo and AdmDemoDx predictors to the usual care performed by clinicians on patients of the holdout set. The scores correspond to the *mean* ± *standard deviation* of the metric over 100 bootstraps drawn with replacement. The highest value for each metric is highlighted in bold.

	Any visit			Last visit		
	Clinicians	AdmDemo	AdmDemoDx	Clinicians	AdmDemo	AdmDemoDx
AUROC	*N.A*	83.0 ± 0.3	$\mathbf{87.2 \pm 0.3}$	*N.A*	85.3 ± 0.3	$\mathbf{89.0 \pm 0.3}$
Sensitivity	74.9 ± 0.6	$\mathbf{75.2 \pm 0.7}$	73.1 ± 0.7	$\mathbf{81.7 \pm 0.5}$	80.5 ± 0.6	77.9 ± 0.6
Specificity	72.2 ± 0.2	76.0 ± 0.2	$\mathbf{84.1 \pm 0.2}$	70.4 ± 0.3	75.1 ± 0.2	$\mathbf{83.6 \pm 0.2}$
Precision	24.8 ± 0.4	27.7 ± 0.4	$\mathbf{36.0 \pm 0.5}$	28.0 ± 0.4	31.3 ± 0.4	$\mathbf{40.1 \pm 0.5}$
NPV	95.9 ± 0.1	$\mathbf{96.2 \pm 0.1}$	$\mathbf{96.2 \pm 0.1}$	$\mathbf{96.5 \pm 0.1}$	$\mathbf{96.5 \pm 0.1}$	96.4 ± 0.1

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

Finally, we analyzed the evolution of importance for each group of features along with the number of visits per patient in the ELSTM-AdmDemoDx. Post-hoc analyses of the importance assigned to each feature by a model provided important insights into their impact on the predicted scores. Figure 5 illustrates that in patients with fewer visits the model relied mostly on demographics to determine mortality risk. In contrast, patients with more visits required almost all their predictors, equally from both their current and previous

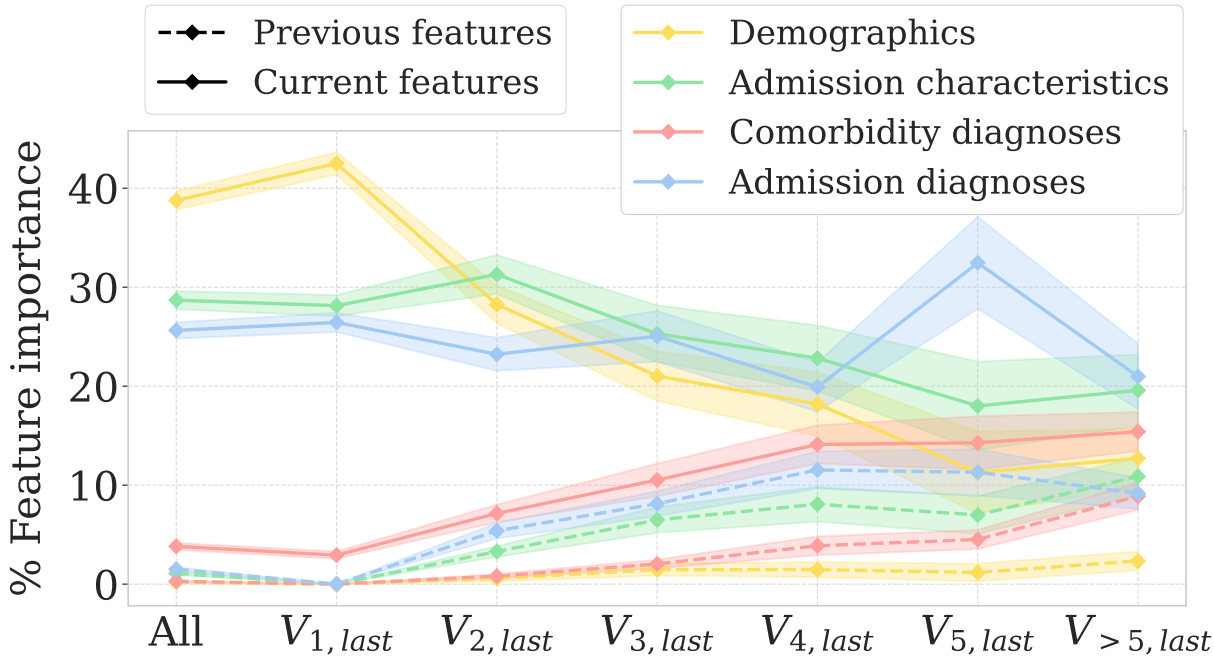


Figure 5: Post-hoc analyses of feature importance of the final ELSTM trained with AdmDemoDx predictors. Importance of each feature is computed using feature permutation[26] over 100 bootstraps. Shaded regions indicate variations within one standard deviation of the mean over 100 bootstraps. Importance of previous features increases as the size of patients' history gets longer.

visits. This highlights the importance of using longitudinal data for patients with a long medical history, and is consistent with clinical reality, where the frequently admitted patients' prognoses depend more on their overall health history than on their demographics. Feature importance and the overall performance of the ELSTM did not vary when we included the time gap between current and previous admissions (Supplementary Figure 4, Supplementary Table 7), demonstrating that the model was able to learn this information solely through the content of longitudinal records.

5 DISCUSSION

Recent years have seen efforts dedicated to developing automated models identifying patients at high risk of mortality, in order to improve end-of-life care and align patient preferences with the provided care. Recent works have explored the use of machine learning models to integrate patients' longitudinal data in several clinical contexts,[27–29] and presented interesting improvements over single-visit models. However, to date, these techniques have not been used for models predicting the HOMR score to enhance palliative care. This study introduces the Ensemble Long Short-Term Memory (ELSTM), a recurrent neural network-based ensemble model that integrates both admission and historical patient data to automatically identify individuals at an elevated risk of one-year mortality. The aim is to prompt the clinical team for end-of-life interventions, such as GOC discussions.

Firstly, we developed the ELSTM, an ensemble model built upon the LSTM neural network, that leverages information learned by different LSTMs at various stages of a patient's admission. We applied the ELSTM to patients with varying numbers of visits and estimated their HOMR score. We used patient self-reported predictors available upon admission (AdmDemo), as well as other comorbid diagnoses available in patients' EHR and admission diagnoses documented later during their stay (AdmDemoDx). A significant improvement in AUROC, the standard evaluation metric in the literature for measuring the discriminative power of the HOMR score,[13] was observed across the majority of patients groups using both sets of predictors. Within

the LSTM-based neural network, we believe the longitudinal data contributed to mortality prediction in two aspects. First, frequent visits to the hospital (i.e., more longitudinal data) likely indicate an increasing severity of illness, thus a higher risk of death. Second, the characteristics of each previous hospitalization (i.e., the content of longitudinal data) provide an overview of the patient's overall condition. Thus, the importance of previous data grows with the length of patient's history. These two aspects allow each LSTM to learn long- and short-term longitudinal patterns to accurately identify patients at high risk of one-year mortality.

Next, we compared the ELSTM to the usual care provided by clinicians to the population of interest eligible for end-of-life interventions. Both AdmDemo and AdmDemoDx strategies revealed considerable benefits as an automated alert prevalence tool of patients at high risk of one-year mortality in a clinical decision support system. More specifically, the ELSTM using AdmDemo predictors facilitates real-time data acquisition, as it requires fewer variables, all available immediately upon admission and can be self-reported. In addition, the model revealed similar results to human decision-making, and is hence useful in hospitals where diagnoses are encoded post-discharge as in previous studies.[15,16] On the other hand, even though the ELSTM using AdmDemoDx predictors is more challenging in terms of data acquisition, it can significantly reduce false positive notifications and therefore the risk of an alert fatigue, making it a suitable candidate for deployment in a clinical decision support system.

We have identified several limitations worth addressing in future studies. Firstly, although the model's overall performance seems satisfactory, an examination of population subgroups shows that the oldest patients, and potentially those with the most complicated medical conditions, are less accurately predicted (Supplementary Table 8). To better identify patients at high risk of mortality, models used on these patients should include not only administrative and diagnostic variables routinely collected on admission, but also admission-specific clinical variables such as vital signs, laboratory and imaging tests. Secondly, our evaluation of clinical utility assumes that a clinician would engage in a GOC discussion and document a CSO for all and only those patients suspected to be at high risk of death. However, this assumption has limitations. Not all high-risk patients may have the opportunity for a GOC discussion due to a lack of resources or time. Additionally, clinicians may document a patient's CSO not only based on their risk of mortality but also on the potential need for escalated care requiring intubation or ventilation. Thirdly, although the AUROC is the main metric in the literature to evaluate models predicting the HOMR score, model selection in clinical settings should primarily maximize clinical utility, which is extremely context-dependent (based on individual hospital services and resources, typical patient origin and profile, severity of admissions, length of stay, etc). Fourthly, although our model demonstrated an acceptable temporal validity, it was not validated using external datasets. We therefore have no evidence on how our model would translate to other institutions with different patient origins, characteristics and distributions. Finally, it is important to acknowledge that predicting a patient at high risk of mortality does not guarantee an effective GOC discussion. Subsequent research should therefore investigate the actual impact of early detection of these patients on the quality of their end-of-life care.

6 CONCLUSION

In this work, we developed an ELSTM, an ensemble recurrent neural network-based approach leveraging information available across different patient hospitalizations. We evaluated our model using data collected routinely during hospital admissions to predict the Hospital One-year Mortality Risk score and to identify individuals who might benefit from end-of-life discussions with healthcare providers. Our model outperformed existing approaches both when using only admission demographics and administrative variables as predictors (AdmDemo), and when integrating diagnoses as well (AdmDemoDx). Our study highlights the rich data potential available in patients' medical records, emphasizing their ability to generate predictive models for enhancing patient care, throughout the life spectrum and at the end of life.

STATEMENTS

All authors had full access to all the data in the study and accept responsibility to submit for publication.

Data availability

Software code allowing to run the experiments used to produce the results presented in this work is freely shared under the GNU General Public License v3.0 on the GitHub website at: https://github.com/MEDomics-UdeS/POYM. The hospitalization data analysed during the current study are not publicly available for confidentiality purposes overseen by the IRB (Institutional Review Board of the CIUSSS de l'Estrie—CHUS Nagano #2022-4409). However, a randomly generated dataset with the same format as used in our experiments is publicly shared in our GitHub repository to test the code implemented for this work.

Authors' contributions

Conceptualization: HL, MV
Data curation: HL, RT
Formal Analysis: HL
Funding acquisition: MV
Investigation: HL, MV, RT
Methodology: HL, MV
Project administration: HL, MV
Resources: MV
Software: HL, NR
Supervision: MV, DP
Validation: HL, DP
Visualization: HL
Writing – original draft: HL
Writing – review & editing: HL, MV, DP, RT, NR

Funding/Support

This study was supported by : (i) Canada CIFAR AI Chair, Mila; (ii) Natural Sciences and Engineering Research Council of Canada (NSERC), Discovery Grants Program (RGPIN-2021-03996); (iii) Fonds de recherche du Québec – Nature et technologies, programme relève professorale (312290).

Role of funding source

The funding sources had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

Competing interests

None.

Acknowledgements

We thank Jean-François Ethier, Associate Professor in the Department of Medicine at the Université de Sherbrooke, for data collection. We also thank Olivier Lefebvre, PhD student at Université de Sherbrooke, and Mahdi Ait Lhaj Loutfi, Master's student at Université de Sherbrooke, for helpful comments and suggestions throughout the project.

REFERENCES

1. Yourman LC, Lee SJ, Schonberg MA, *et al.* Prognostic indices for older adults: a systematic review. *JAMA*. 2012;307(2):182–192.

2. Clarke M, Kennedy K, MacDonagh R. Development of a clinical prediction model to calculate patient life expectancy: the measure of actuarial life expectancy (MALE). *Medical Decision Making.* 2009;29(2):239–246.

3. Kalra S, Basourakos S, Abouassi A, *et al.* The implications of ageing and life expectancy in prostate cancer treatment. *Nat Rev Urol.* 2016;13(5):289–295.

4. Seow H, O'Leary E, Perez R, *et al.* Access to palliative care by disease trajectory: a population-based cohort of Ontario decedents. *BMJ open.* 2018;8(4):e021147.

5. Gomes B, Calanzani N, Gysels M, *et al.* Heterogeneity and changes in preferences for dying at home: a systematic review. *BMC Palliat Care.* 2013;12(1):1–13.

6. Hsu AT, Garner RE. Associations between the receipt of inpatient palliative care and acute care outcomes: a retrospective study. *Health Reports.* 2020;31(10):3–13.

7. Brinkman-Stoppelenburg A, Rietjens JA, Heide A. The effects of advance care planning on end-of-life care: a systematic review. *Palliat Med.* 2014;28(8):1000–1025.

8. Huber MT, Highland JD, Krishnamoorthi VR, *et al.* Utilizing the electronic health record to improve advance care planning: a systematic review. *Am J Hosp Palliat Care.* 2018;35(3):532–541.

9. Taseen R, Ethier JF. Expected clinical utility of automatable prediction models for improving palliative and end-of-life care outcomes: Toward routine decision analysis before implementation. *JAMIA Open.* 2021;28(11):2366–2378.

10. Heyland DK, Allan DE, Rocker G, *et al.* Discussing prognosis with patients and their families near the end of life: impact on satisfaction with end-of-life care. *Open Medicine.* 2009;3(2):e101.

11. Yamaguchi T, Maeda I, Hatano Y, *et al.* Effects of end-of-life discussions on the mental health of bereaved family members and quality of patient death and care. *J Pain Symptom Manage.* 2017;54(1):17–26.

12. Lund S, Richardson A, May C. Barriers to advance care planning at the end of life: an explanatory systematic review of implementation studies. *PloS one.* 2015;10(2):e0116629.

13. Walraven C. The Hospital-patient One-year Mortality Risk score accurately predicted long-term death risk in hospitalized patients. *J Clin Epidemiol.* 2014;67(9):1025–1034.

14. Walraven C, McAlister FA, Bakal JA, *et al.* External validation of the Hospital-patient One-year Mortality Risk (HOMR) model for predicting death within 1 year after hospital admission. *Can Med Assoc J.* 2015;187(10):725–733.

15. Walraven C, Forster AJ. The HOMR-Now! model accurately predicts 1-year death risk for hospitalized patients on admission. *Am J Med Open.* 2017;130(8):991–e9.

16. Wegier P, Koo E, Ansari S, *et al.* mHOMR: a feasibility study of an automated system for identifying inpatients having an elevated risk of 1-year mortality. *BMJ Qual Saf.* 2019;28(12):971–979.

17. Guo A, Foraker R, White P, *et al.* Using electronic health records and claims data to identify high-risk patients likely to benefit from palliative care. *Am J Manag Care.* 2021;27(1).

18. Beeksma M, Verberne S, Bosch A, *et al.* Predicting life expectancy with a long short-term memory recurrent neural network using electronic medical records. *BMC Med Inform Decis Mak.* 2019;19(1):1–15.

19. Hochreiter S, Schmidhuber J. Long short-term memory. *Neural Comput.* 1997;9(8):1735–1780.

20. Pedregosa F, Varoquaux G, Gramfort A, *et al.* Scikit-learn: Machine learning in Python. *J Mach Learn Res.* 2011;12:2825–2830.

21. Paszke A, Gross S, Massa F, *et al.* Pytorch: An imperative style, high-performance deep learning library. *Adv Neural Inf Process Syst.* 2019;32.

22. Kingma DP, Ba J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980.* 2014.

23. Akiba T, Sano S, Yanase T, *et al.* Optuna: A next-generation hyperparameter optimization framework. in *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*:2623–2631 2019.

24. Wilcoxon F. Individual comparisons by ranking methods. in *Breakthroughs in Statistics: Methodology and Distribution*:196–202Springer 1992.

25. Youden WJ. Index for rating diagnostic tests. *Cancer.* 1950;3(1):32–35.

26. Fisher A, Rudin C, Dominici F. All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously. *J Mach Learn Res.* 2019;20(177):1–81.

27. Herman R, Vanderheyden M, Vavrik B, *et al.* Utilizing longitudinal data in assessing all-cause mortality in patients hospitalized with heart failure. *ESC Heart Fail.* 2022;9(5):3575–3584.

28. Nitski O, Azhie A, Qazi-Arisar FA, *et al.* Long-term mortality risk stratification of liver transplant recipients: real-time application of deep learning algorithms on longitudinal data. *Lancet Digit Health.* 2021;3(5):e295–e305.

29. Yang F, Zhang J, Chen W, *et al.* DeepMPM: a mortality risk prediction model using longitudinal EHR data. *BMC Bioinformatics.* 2022;23(1):423.