Opportunities and Challenges in Fall Risk Management using EHRs and Artificial Intelligence: A Systematic Review

Henrique D. P. dos Santos[1][a], Juliana O. Damasio[1][b], Ana Helena D. P. S. Ulbrich[2][c]
and Renata Vieira[3][d]

[1] *School of Technology, PUCRS, Porto Alegre, Brazil*
[2] *Nossa Senhora da Conceição Hospital, Porto Alegre, Brazil*
[3] *CIDEHUS, University of Évora, Portugal*

Keywords: Electronic Health Records, Artificial Intelligence, Fall Detection, Fall Risk Prediction.

Abstract: Electronic Health Records (EHRs) have led to valuable improvements to hospital practices by integrating patient information. In fact, this data can be used to develop clinical risk prediction tools. We performed a systematic literature review with the objective of analyzing current studies that use artificial intelligence techniques in EHRs data to identify in-hospital falls. We searched several digital libraries for articles that reported on the use of EHRs and artificial intelligence techniques to identify in-hospital falls. Articles were selected by three authors of this work. We compiled information on study design, use of EHR data types, and methods. We identified 21 articles, 11 about fall risk prediction and 10 covering fall detection. EHR data shows opportunities and challenges for fall risk prediction and in-hospital fall detection. There is room for improvement in developing such studies.

1 INTRODUCTION

Electronic Health Records (EHRs) have played an important role in hospital environments providing many benefits in terms of patient safety and health care quality (Buntin et al., 2011). EHRs are a rich source of information to build risk prediction models. Current machine learning and natural language processing (NLP) techniques can help in the efforts to prevent or identify several outcomes: readmission, fracture, diabetes, mortality, and length of stay, among others. A previous systematic review, presented in (Goldstein et al., 2017), described work on risk prediction models yet it did not consider falls. To the best of our knowledge, there are no previous studies addressing a systematic review for the use of machine learning and NLP methods over EHR data to identify in-hospital falls. Within hospitals and nursing homes, falls constitute the largest category of adverse event reports. Approximately 30% of in-patient falls result in injury, with 4% to 6% resulting in serious injury (Hitcho et al., 2004).

Traditional fall risk models (Morse et al., 1989) and fall detection tools (Resar et al., 2006) were developed for hospital environments without EHR systems. These tools are useful but time-consuming and do not consider cultural changes for a variety of hospitals and countries (De Souza Urbanetto et al., 2013). A previous review on the matter focused on non-automated models, listing articles about fall risk prediction models, predicting falls among inpatients and recording falls in the community (Walsh et al., 2016). Another fall-related review focused on sensor information, but used machine learning algorithms in wearable, ambience, and vision-based devices (Mubashir et al., 2013) not EHR.

Both fall detection and fall risk prediction play a crucial role for hospital risk management and prevention (Swift and Iliffe, 2014). Fall detection enables the nurse team to map the most common reasons for fall incidents and to create policies to avoid new incidents. Fall risk prediction is one of these strategies to prevent inpatient falls: it predicts the patient's risk and defines the level of care for this patient.

Thus, the main purpose of this systematic literature review (SLR) is to understand how EHRs data

[a] https://orcid.org/0000-0002-2410-3536
[b] https://orcid.org/0000-0001-8915-285X
[c] https://orcid.org/0000-0001-6910-8210
[d] https://orcid.org/0000-0003-2449-5477

Santos, H., Damasio, J., Ulbrich, A. and Vieira, R.
Opportunities and Challenges in Fall Risk Management using EHRs and Artificial Intelligence: A Systematic Review.
DOI: 10.5220/0010424306260633
In *Proceedings of the 23rd International Conference on Enterprise Information Systems (ICEIS 2021) - Volume 1*, pages 626-633
ISBN: 978-989-758-509-8

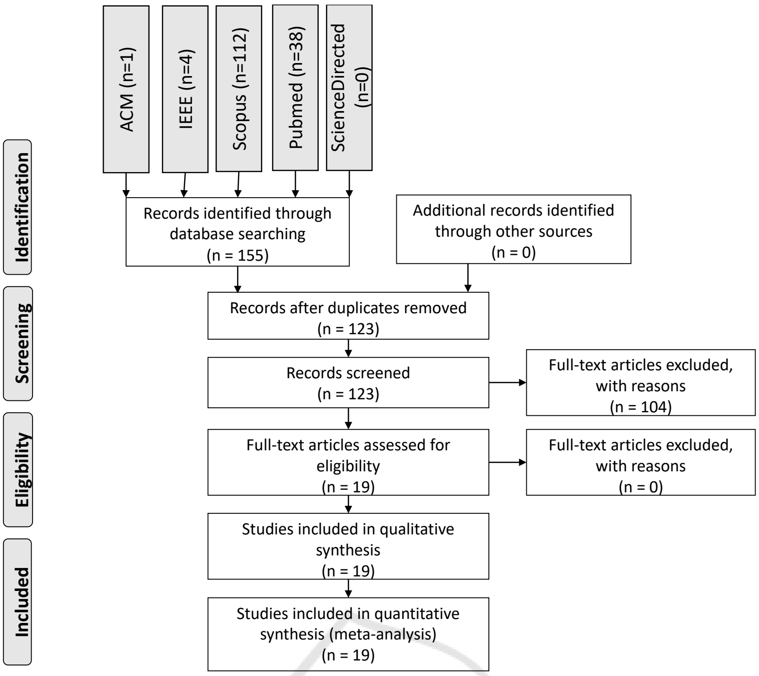
ACM (n=1)
IEEE (n=4)
Scopus (n=112)
Pubmed (n=38)
ScienceDirected (n=0)
Identification
Screening
Eligibility
Included
Records identified through database searching (n = 155)
Additional records identified through other sources (n = 0)
Records after duplicates removed (n = 123)
Records screened (n = 123)
Full-text articles excluded, with reasons (n = 104)
Full-text articles assessed for eligibility (n = 19)
Full-text articles excluded, with reasons (n = 0)
Studies included in qualitative synthesis (n = 19)
Studies included in quantitative synthesis (meta-analysis) (n = 19)

Figure 1: PRISMA flow diagram.

has been used to develop and validate automatically built models for identifying in-hospital fall risks. We focus on EHRs data because of a large amount of useful information is generated during the patient stay. We detailed the studies considering research design, data types, outcomes and evaluation, to summarize what is relevant and useful for these models. The goal is to analyse and discuss how the area has been developed and point to future promising directions.

The paper is structured as follows. Section 2 shows the protocol used in the SLR. Section 3.1 presents the results obtained with the application of the protocol. Section 4 presents a discussion about the results. Finally, in Section 5 the conclusions are presented.

2 MATERIALS AND METHODS

We used the SLR protocol proposed by (Kitchenham et al., 2009). The main goal of this study is to identify artificial intelligence techniques (machine learning, natural language processing, and neural networks) to analyze adverse events (falls) in electronic health records (EHRs).

2.1 Research Questions

Based on the previously defined objective, the following research questions were identified:
Research Design
RQ1: What is the purpose of the investigation?
RQ2: What types of data and features are used?
Model Development and Evaluation
RQ3: Which AI techniques are used?
RQ4: How are the algorithms evaluated?
RQ5: What is the performance of the algorithms?
Limitations
RQ6: What are the limitations indicated?
Final discussion
RQ7: What are the challenges and opportunities?

2.2 Search Process

We selected five relevant digital libraries in Computing Science and Health: ACM Digital Library[1], ScienceDirect[2], IEEExplore[3], Scopus[4], and Pubmed[5]. Scopus is a general database that indexes several other databases, covering approximately 19,500 titles from more than 5,000 international publishers,

[1] https://dl.acm.org/

[2] https://www.sciencedirect.com/

[3] https://ieeexplore.ieee.org/Xplore/home.jsp

[4] https://www.scopus.com/

[5] https://www.ncbi.nlm.nih.gov/pubmed/

including coverage of 16,500 peer-reviewed journals in the scientific, technical, and medical and social sciences. Afterwards, keywords related to the research topic were identified, such as “fall”, “electronic health records”, and “artificial intelligence techniques”. These terms were combined to create the search expressions. The search expressions were adapted according to the mechanism of each digital library, so as not to alter their logical sense. The searches were performed in the abstract, title, and keywords fields.

For “fall” concept were used the search expression `(fall OR falls) AND`; For “electronic health records” were used `(electronic health records OR EHR OR electronic medical records OR EMR OR narratives OR free-text records OR clinical notes)` AND; For “artificial intelligence techniques” were used `(machine learning OR data mining OR text mining OR neural networks OR natural language processing OR information extraction OR decision trees OR prediction)`.

2.3 Selection Criteria

We included all papers (including conference proceedings) published in English and regardless of year of publication. We followed some selection criteria for the inclusion and exclusion of publications:

Exclusion. In the case of similar or duplicate publications, only the most recent were considered; Results that did not use data from medical records or did not use computational methods or did not focus on fall adverse events; Books, PhD or Masters theses, and abstracts from conference presentations.

Inclusion. The results must bear some relation with the research topic of this work in the title, in the keywords, or in the abstract.

2.4 Quality Assessment, Data Collection, Data Analysis

We applied the search expression to each digital library in May 2019. We used the StArt[6] tool to help organize the SLR. The results extracted from each digital library were imported into StArt, with all relevant information about the publication, such as title, abstract, year, authors, and keywords. To check the quality of the publications we used the Kappa Method for Measurement of Interrater reliability (McHugh, 2012).

For that purpose, first one author applied the selection criteria in the publications; subsequently, two researchers individually reviewed the accepted and excluded publications. This was important to reduce bias. To better evaluate the articles, we consulted the TRIPOD (Moons et al., 2015) guidelines that state how to report risk prediction studies. Finally, we performed an analysis of the publications.

We first identified a total of 157 publications. After applying the selection criteria, 21 studies remained. Figure 1 shows the PRISMA flow diagram (Liberati et al., 2009) with the number of papers accepted each step of the research. Publications from the years 2009 to 2020 were found in our results.

These publications were related to two tasks: fall risk prediction and fall incident detection. Fall risk prediction is used to label a patient in her admission, to inform and aware the healthcare professionals about her condition. The fall detection is a Risk Management department task to identify adverse events in the hospital environment and create preventive risk protocols and safety policies.

We analysed the number of publications by country. The highest concentration of papers was the United States of America (13 papers), followed by Japan (n = 3), Korean (n = 2), and Brazil (n = 2). The number of papers per country was collected according to the country of the first author’s institutional affiliation. In regards to the distribution of publications by year, the year 2015 showed the largest amount of papers (n = 5), followed by 2019 (n = 4), 2016 (n = 3), 2018 (n = 3), 2012 (n = 2), and 2017 (n = 2).

We present a summary of the studies in Tables 1 and 2. The analysis of these works are presented in detail in the next section, answering questions presented in 2.1.

3 RESULTS

3.1 Design of EHR Prediction/Detection Studies

Referring to question **RQ1**, we see that the studies main purpose are divided in detection and prediction of falls. For papers on detection of in-hospital fall events (n = 8) we present Table 1. For papers related to the prediction of the patients’ risk of falling (n = 11) we present Table 2. Most studies used a cohort design from a single EHR system and a single hospital (n = 12) (Tremblay et al., 2009; Toyabe,

[6] http://lapes.dc.ufscar.br/tools/start_tool

Table 1: Characteristics of Fall Incident Detection Studies Included in the Review.

Author	Year	Patients	Data Points	Source	Algorithms	Eval	Ss	Sp	F1
Tremblay	2009	NR	2,157	notes	K-Means, LR	T/T	0.83	0.93	
Toyabe	2012	NR	4,821	notes, inc-rep	Syntactic Rules	DV	0.87	0.98	0.84
McCart	2013	2,241	26,010	notes	LR, SVM	T/T	0.93	0.94	0.85
Rochefort	2015	NR	NR	inc-rep	NR	NR	0.83	1.00	
Luther	2015	1,652	26,010	notes	SVM	T/T	0.94	0.83	0.90
Bates	2016	NR	8,288	radiology reports	SVM	CV	0.94		0.93
Shiner	2016	NR	2,730	notes	MaxEnt, CRF	T/T	0.97	0.44	
Topaz	2019	NR	750	notes	Random Forest	T/T	0.90		0.89
dos Santos	2019	367	1,078	notes	LSTM	CV	NR	NR	0.90
Santos	2020	NR	3,468	notes	BiLSTM-CRF	CV	NR	NR	0.81

NR = Not Reported; notes = Clinical Notes; inc-rep = Incident Reports; LR = Logistic Regression; Eval = Evaluation Method; T/T = Split dataset in train and test sets once; DV = Direct Validation; CV = Cross Validation; Ss = Sensitivity; Sp = Specificity; F1 = F-Measure; Every Fall Detection study used textual information as the source of detection.

Table 2: Characteristics of Fall Risk Prediction Studies Included in the Review.

Author	Year	Patients	Data Points	Source	Algorithms	Eval	Ss	Sp	AUC
Marier	2016	5,129	133,781	pat, adm, med	LR	T/T	NR		
Weed-Pfaff	2016	1,080	NR	adm, med	Syntactic Rules	DV	NR		
Lee	2016	8,390	NR	pat, adm	LR	T/T	0.95		
Yokota	2016	46,241	1,230,604	pat, adm	LR	T/T	0.71	0.66	
Yokota	2017	45,257	1,223,687	pat, adm	SVM	T/T	0.64	0.69	
Zhu	2017	114	1,558	notes	Query	DV	NR		
Bjarnadottir	2018	36,583	1,046,053	notes	Query	NR	NR		
Choi	2018	75,036	220,904	pat, adm	Syntactic Rules	DV	0.30		0.70
Lucero	2018	814	NR	pat, adm, med	LR	CV	0.73	0.85	0.88
Cho	2019	35,479	NR	adm, med	Bayes Network	CV			0.96
Oshiro	2019	57,678	NR	adm, pat, med	LR	T/T	0.73	0.70	0.72

NR = Not Reported; notes = Clinical Notes; adm = Administrative Data; med = Medication Orders; pat = Patient Data; LR = Logistic Regression; DV = Direct Validation; T/T = Split dataset in train and test sets once; CV = Cross Validation; Ss = Sensitivity; Sp = Specificity; AUC = Area Under the ROC Curve;

2012; Rochefort et al., 2015; Lee et al., 2016; Bates et al., 2016; Yokota and Ohe, 2016; Yokota et al., 2017; Zhu et al., 2017; Bjarnadottir and Lucero, 2018; Lucero et al., 2019; Topaz et al., 2019; Oshiro et al., 2019; dos Santos et al., 2019; Santos et al., 2020), the other studies used data from multiple hospitals (with the same EHR system)(n = 7) (McCart et al., 2013; Luther et al., 2015; Shiner et al., 2016; Marier et al., 2016; Weed-Pfaff et al., 2016; Choi et al., 2018; Cho et al., 2019).

Considering question **RQ2**, fall detection studies based their predictive models on textual information. Fall events reported in EHRs are usually presented as non-structured data (text). Clinical notes were the most common sources used (n = 6) (Tremblay et al., 2009; Toyabe, 2012; McCart et al., 2013; Luther et al., 2015; Shiner et al., 2016; Topaz et al., 2019; dos Santos et al., 2019; Santos et al., 2020), other sources included were radiology reports (n = 1) (Bates et al., 2016) and incident reports (n = 2) (Toyabe, 2012; Rochefort et al., 2015).

For fall risk detection models, most studies used structured data, as follows:

Administrative Data (n = 9) (Marier et al., 2016; Weed-Pfaff et al., 2016; Lee et al., 2016; Yokota and Ohe, 2016; Yokota et al., 2017; Choi et al., 2018; Lucero et al., 2019; Cho et al., 2019; Oshiro et al., 2019)**:** the average number of daily tests, blood transfusions, number of nurses, Charlson score, Morse Fall Scale score, other nursing scores, International Classification of Diseases (ICD), type of room, length of hospital stay, medical department, ward, unit, day of the week, nursing processes.

Patient Data (n = 7) (Marier et al., 2016; Lee et al., 2016; Yokota and Ohe, 2016; Yokota et al., 2017; Choi et al., 2018; Lucero et al., 2019; Oshiro et al., 2019)**:** gender, age, blood pressure, gait abnormality, other mental disorders, walking issues, Parkinson's disease, urinary incontinence, depression, maximum pulse rate, registration as a severely ill patient, activity and hyponatremia, osteoarthritis, osteoporosis, other symptoms.

Medication Data (n = 5) (Marier et al., 2016; Weed-Pfaff et al., 2016; Lucero et al., 2019; Cho et al., 2019; Oshiro et al., 2019)**:** medication period of nervous and circulation medicines, psychotropics, antipsychotic medication, anticonvulsant medications, in some cases all medications.

Each feature provides the model with a hypothesis for the possible risk of falls. Other studies did not use structured data, but only textual information from clinical notes (n = 2) (Zhu et al., 2017; Bjarnadottir and Lucero, 2018). In general, studies found a deficit in the Morse Fall Scale to represent all variables for risk factors. Some studies regarding fall risk models described the most relevant features: imbalance and gait are the main reason to determine a high fall risk.

3.2 Model Development and Evaluation

Regarding question **RQ3**, Logistic Regression models were the most common algorithms used (n = 7) (Tremblay et al., 2009; McCart et al., 2013; Marier et al., 2016; Lee et al., 2016; Yokota et al., 2017; Lucero et al., 2019; Oshiro et al., 2019) to develop the prediction model. Other approaches included syntactic rules or queries (n = 5) (Toyabe, 2012; Weed-Pfaff et al., 2016; Choi et al., 2018; Zhu et al., 2017; Bjarnadottir and Lucero, 2018), Support Vector Machine methods (n = 4) (McCart et al., 2013; Luther et al., 2015; Bates et al., 2016; Yokota et al., 2017), Recurrent Neural Network (n = 2) (dos Santos et al., 2019; Santos et al., 2020), Random Forests (n = 1) (Topaz et al., 2019), Bayes Network (n = 1) (Cho et al., 2019), and Conditional Random Field (n = 1) (Shiner et al., 2016). Most fall risk prediction studies incorporated some form of variable selection (n = 9) (Marier et al., 2016; Weed-Pfaff et al., 2016; Lee et al., 2016; Yokota and Ohe, 2016; Yokota et al., 2017; Choi et al., 2018; Lucero et al., 2019; Cho et al., 2019; Oshiro et al., 2019), most often via stepwise approaches.

Most of these works select from patient, administrative and medication data structured variables as features (n = 9) (Marier et al., 2016; Weed-Pfaff et al., 2016; Lee et al., 2016; Yokota and Ohe, 2016; Yokota et al., 2017; Choi et al., 2018; Lucero et al., 2019; Cho et al., 2019; Oshiro et al., 2019). On the other hand, only two papers used textual information (Zhu et al., 2017; Bjarnadottir and Lucero, 2018).

Considering fall incident detection, all studies used textual information. A variety of strategies were used to detect fall in texts as showed in Table 1. Most work used machine learning algorithms (n = 6) (Tremblay et al., 2009; McCart et al., 2013; Luther et al., 2015; Bates et al., 2016; Shiner et al., 2016; Topaz et al., 2019; dos Santos et al., 2019; Santos et al., 2020) and one used syntactic rules (Toyabe, 2012).

Answering question **RQ4**, all but two studies used some form of validation of the model. The most common form was split sample in train and test (n = 10) (McCart et al., 2013; Tremblay et al., 2009; Luther et al., 2015; Bates et al., 2016; Shiner et al., 2016; Topaz et al., 2019; Yokota et al., 2017; Yokota and Ohe, 2016; Lee et al., 2016; Marier et al., 2016), followed by cross-validation (n = 5) (Lucero et al., 2019; Cho et al., 2019; Bates et al., 2016; dos Santos et al., 2019; Santos et al., 2020) and direct validation (n = 4) (Toyabe, 2012; Zhu et al., 2017; Weed-Pfaff et al., 2016; Choi et al., 2018), with some studies using multiple forms of validation. Direct validation was used when the authors developed syntactic rules or queries. These rules are based on data information, author hypothesis, or/and statistical correlation with the outcome.

Splitting the sample in train and test is a common approach used in health science research groups. Computer science studies usually apply the cross-validation method. The split proportion varied in the studies in this review. The authors (Tremblay et al., 2009) and (Topaz et al., 2019) chose to split each dataset into a training dataset (80% of the data) and a validation dataset (20% of the data). In (McCart et al., 2013) and (Luther et al., 2015) models were trained with a stratified sample of 70% of documents from one location (dataset train) and then applied to the remaining unseen documents (dataset test). Oshiro et al. (Oshiro et al., 2019) used a similar split: 72% to train the model and 22% to test it. While (Marier et al., 2016) and (Yokota et al., 2017) performed a split by 50%. In (Shiner et al., 2016) and (Yokota and Ohe, 2016) did not explicit the split proportion they used.

For **RQ5**, we found that the most common metric to measure the algorithms' performance was Sensitivity (n = 14) (Tremblay et al., 2009; Toyabe, 2012; McCart et al., 2013; Rochefort et al., 2015; Luther et al., 2015; Bates et al., 2016; Shiner et al., 2016; Topaz et al., 2019; Lee et al., 2016; Yokota and Ohe, 2016; Yokota et al., 2017; Choi et al., 2018; Lucero et al., 2019; Oshiro et al., 2019), which measures the proportion of actual positives that are correctly identified. Specificity was also used in most studies (n

= 10) (Tremblay et al., 2009; Toyabe, 2012; McCart et al., 2013; Rochefort et al., 2015; Luther et al., 2015; Shiner et al., 2016; Yokota and Ohe, 2016; Yokota et al., 2017; Lucero et al., 2019; Oshiro et al., 2019); it measures the proportion of actual negatives that are correctly identified. Another common metric in machine learning studies was F-Measure (n = 5) (Toyabe, 2012; McCart et al., 2013; Luther et al., 2015; Bates et al., 2016; Topaz et al., 2019; dos Santos et al., 2019; Santos et al., 2020), the harmonic mean of precision and sensitivity. Precision is the fraction of correctly identified instances among all positives identified. Three other studies used the Area Under the Receiver Operating Characteristic (ROC) Curve (relation between sensitivity and specificity). Average Sensitivity for in-hospital fall detection was 0.90 (worst: 0.83, best: 0.97) and for fall risk prediction was 0.67 (worst: 0.30, best: 0.95). In Tables 1 and 2, we show the individual results for each study.

3.3 Limitations

Besides the contribution of the works, we also analysed what some authors listed as the limitations of their studies (**RQ6**). The most reported limitations were the data sample and data selection (n = 11) (Tremblay et al., 2009; McCart et al., 2013; Shiner et al., 2016; Yokota and Ohe, 2016; Weed-Pfaff et al., 2016; Marier et al., 2016; Zhu et al., 2017; Yokota et al., 2017; Choi et al., 2018; Oshiro et al., 2019; Topaz et al., 2019; Lucero et al., 2019; dos Santos et al., 2019). The subsequent related issue was the generalization problem of the model (n = 8) (Shiner et al., 2016; Marier et al., 2016; Zhu et al., 2017; Bjarnadottir and Lucero, 2018; Choi et al., 2018; Oshiro et al., 2019; Topaz et al., 2019; Lucero et al., 2019). Some studies did not discuss their limitations (n = 7) (Toyabe, 2012; Luther et al., 2015; Rochefort et al., 2015; Bates et al., 2016; Lee et al., 2016; Cho et al., 2019)

Regarding fall incident detection, Tremblay et al. (Tremblay et al., 2009) showed examples where the trained model misclassified fall-related adverse events. The examples featured words such as hip, pain, and knee, which are commonly found in fall incidents; however, that is not always the case. McCart et al. (McCart et al., 2013) discussed the pitfalls in the gathered dataset. A shortfall in the reported incidents added bias to the trained model.

The authors (Shiner et al., 2016) warned about their small and random sample to identify falls. Their study design reduced the possible variation in the way falls are described. They stated that further work should test multiple methods for fall identification, including incident reports, manual records reviews, and patient self-reports.

The reported limitations in fall risk prediction also referred to dataset issues. Marier et al. (Marier et al., 2016) pointed to limitations in the model generalization related to the selection procedure. Their sample was restricted to nursing homes that disproportionately represent higher-quality institutions. Furthermore, there was missing data on selected risk factors for some residents. Weed-Pfaff et al. (Weed-Pfaff et al., 2016) noted that errors of omission or inaccurately recorded data could have affected results. In (Yokota and Ohe, 2016) also mentioned human errors and estimated that fall incidents occurred 1.3 times more often than the reported falls. They stated that the process of choosing variables was a time-consuming task. In their following work (Yokota et al., 2017), Yakota and colleagues noted that the representation of a single day did not evaluate the patients' status changes over the course of the day. In this study, the constructed model was a black box. It did not specify which feature contributes more with the outcome.

Regarding dataset size and sample selection, Zhu et al. (2017) also noted the limited size of fall incidents. The authors (Bjarnadottir and Lucero, 2018) highlighted the lack of generalization when a model is trained over a single-site observation. Moreover, the lack of labeled data only allowed them to conclude that their model might contain risk factors that have been theoretically and empirically linked to fall risk. Choi et al. (2018) remembered the importance of evaluating the model generality again. They observed that their model was built upon discrete diagnosis information that are not usually present in EHR systems. Lucero et al. (2019) were also concerned about feature selection. Their findings presented limited generalization because the data derives from only one tertiary teaching hospital. Oshiro, Cho et al.(2019) debated that further models should improve the identification of injurious falls and the detection of falls overall.

The drawbacks listed above provide inputs on opportunities and challenges. The next section discusses the possible room for improvement in the development of fall incident detection and fall risk prediction models.

4 DISCUSSION

Over the last decade, many studies on the development of models for fall detection and risk prediction using EHR data have been published. In this section, we discuss the challenges and opportunities for artificial intelligence in fall risk management, answering question **RQ7**.

The EHR data has the advantage of providing a large number of patients for cohort studies. On top of that, it is also able to provide a large number of features: potential predictors. However, we verified that many studies did not fully use every piece of information about the patients available in their medical records as predictor variables.

The validation process adopted by some studies did not assure the applicability of the proposed model in other scenarios. A multivariable model for fall risk prediction should be validated with an independent sample and should evaluate its impact in real scenarios before being used as a clinical decision-support system. Only four studies used multiple sites, but none validated the model across these sites. This shows a lost opportunity to validate the prediction algorithm in external data. Even if the scores perform worse externally, it is important to answer how well the models will fit another site.

Four insights for improvements in development of fall incident detection and fall risk prediction: *Dataset.* Combining all EHR data: laboratory, medication, patient data, and clinical notes; *Algorithms.* Evaluating Machine and Deep Learning approaches over all features; *Model.* Training site-specific models since each hospital has different patient profiles and specific environments; *Validation.* Consider multicenter studies and evaluations on real scenarios;

5 CONCLUSIONS

Our results indicate that many models for fall detection (to map the occurrence of incidents) and fall risk prediction (to avoid new incidents) have been developed, in the last ten years, based on information from Electronic Health Records. Both tasks play a crucial role in fall prevention for hospital risk management. Most studies in risk prediction used structured data related to administrative, medication, and patient information. The most commonly used algorithms were Generalized linear models. Thus, the most common metric to measure the algorithms' performance was Sensitivity. We believe that one of the challenges for the next few years is to use all available EHR data to build both predictive and detection models. In addition, further work should focus on improving and validating existing models, considering the TRIPOD guidelines to provide quality reporting. Also, it is important to create automated methods that focus on patient safety avoiding spending the time of health care professionals with questionnaires, annotations, protocol assessments and notifications.

ACKNOWLEDGMENTS

This work was partially funded by CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior) Foundation (Brazil), UFRGS (Federal University of Rio Grande do Sul), Google Latin America Research Awards, and the Foundation for Science and Technology of Portugal, under the projects UIDB/00057/2020, CEECIND/01997/2017 and by CAPES - Finance Code 001.

REFERENCES

Bates, J., Fodeh, S., Brandt, C., and Womack, J. (2016). Classification of radiology reports for falls in an hiv study cohort. *Journal of the American Medical Informatics Association*, 23(e1):e113–e117.

Bjarnadottir, R. I. and Lucero, R. J. (2018). What can we learn about fall risk factors from ehr nursing notes? a text mining study. *eGEMs*, 6(1):1–8.

Buntin, M. B., Burke, M. F., Hoaglin, M. C., and Blumenthal, D. (2011). The benefits of health information technology: a review of the recent literature shows predominantly positive results. *Health affairs*, 30(3):464–471.

Cho, I., Boo, E.-H., Chung, E., Bates, D., and Dykes, P. (2019). Novel approach to inpatient fall risk prediction and its cross-site validation using time-variant data. *Journal of medical Internet research*, 21(2):e11505.

Choi, Y., Staley, B., Henriksen, C., Xu, D., Lipori, G., Brumback, B., and Winterstein, A. (2018). A dynamic risk model for inpatient falls. *American Journal of Health-System Pharmacy*, 75(17):1293–1303.

De Souza Urbanetto, J., Creutzberg, M., Franz, F., Ojeda, B., da Silva Gustavo, A., Bittencourt, H., Steinmetz, Q., and Farina, V. (2013). Morse fall scale: Translation and transcultural adaptation for the portuguese language [morse fall scale: Tradução e adaptação transcultural para a língua portuguesa]. *Revista da Escola de Enfermagem*, 47(3):569–575.

dos Santos, H. D. P., Silva, A. P., Maciel, M. C. O., Burin, H. M. V., Urbanetto, J. S., and Vieira, R. (2019). Fall detection in ehr using word embeddings and deep learning. In *2019 IEEE 19th International Conference on Bioinformatics and Bioengineering (BIBE)*, pages 265–268.

Goldstein, B., Navar, A., Pencina, M., and Ioannidis, J. (2017). Opportunities and challenges in developing risk prediction models with electronic health records data: A systematic review. *Journal of the American Medical Informatics Association*, 24(1):198–208.

Hitcho, E. B., Krauss, M. J., Birge, S., Claiborne Dunagan, W., Fischer, I., Johnson, S., Nast, P. A., Costantinou, E., and Fraser, V. J. (2004). Characteristics and circumstances of falls in a hospital setting: a prospective analysis. *Journal of general internal medicine*, 19(7):732–739.

Kitchenham, B., Brereton, O. P., Budgen, D., Turner, M., Bailey, J., and Linkman, S. (2009). Systematic literature reviews in software engineering–a systematic literature review. *Information and software technology*, 51(1):7–15.

Lee, J., Jin, Y., Piao, J., and Lee, S.-M. (2016). Development and evaluation of an automated fall risk assessment system. *International Journal for Quality in Health Care*, 28(2):175–182.

Liberati, A., Altman, D. G., Tetzlaff, J., Mulrow, C., Gøtzsche, P. C., Ioannidis, J. P., Clarke, M., Devereaux, P. J., Kleijnen, J., and Moher, D. (2009). The prisma statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. *PLoS medicine*, 6(7):e1–e34.

Lucero, R., Lindberg, D., Fehlberg, E., Bjarnadottir, R., Li, Y., Cimiotti, J., Crane, M., and Prosperi, M. (2019). A data-driven and practice-based approach to identify risk factors associated with hospital-acquired falls: Applying manual and semi- and fully-automated methods. *International Journal of Medical Informatics*, 122:63–69.

Luther, S., McCart, J., Berndt, D., Hahm, B., Finch, D., Jarman, J., Foulis, P., Lapcevic, W., Campbell, R., Shorr, R., Valencia, K., and Powell-Cope, G. (2015). Improving identification of fall-related injuries in ambulatory care using statistical text mining. *American Journal of Public Health*, 105(6):1168–1173.

Marier, A., Olsho, L., Rhodes, W., and Spector, W. (2016). Improving prediction of fall risk among nursing home residents using electronic medical records. *Journal of the American Medical Informatics Association*, 23(2):276–282.

McCart, J., Berndt, D., Jarman, J., Finch, D., and Luther, S. (2013). Finding falls in ambulatory care clinical documents using statistical text mining. *Journal of the American Medical Informatics Association*, 20(5):906–914.

McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia medica: Biochemia medica*, 22(3):276–282.

Moons, K. G., Altman, D. G., Reitsma, J. B., Ioannidis, J. P., Macaskill, P., Steyerberg, E. W., Vickers, A. J., Ransohoff, D. F., and Collins, G. S. (2015). Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (tripod): explanation and elaboration. *Annals of internal medicine*, 162(1):W1–W73.

Morse, J., Morse, R., and Tylko, S. (1989). Development of a scale to identify the fall-prone patient. *Canadian Journal on Aging / La Revue canadienne du vieillissement*, 8(4):366–377.

Mubashir, M., Shao, L., and Seed, L. (2013). A survey on fall detection: Principles and approaches. *Neurocomputing*, 100:144 – 152. Special issue: Behaviours in video.

Oshiro, C., Frankland, T., Rosales, A., Perrin, N., Bell, C., Lo, S., and Trinacty, C. (2019). Fall ascertainment and development of a risk prediction model using electronic medical records. *Journal of the American Geriatrics Society*.

Resar, R., Rozich, J., Simmonds, T., and Haraden, C. (2006). A trigger tool to identify adverse events in the intensive care unit. *Joint Commission Journal on Quality and Patient Safety*, 32(10):585–590.

Rochefort, C., Buckeridge, D., and Abrahamowicz, M. (2015). Improving patient safety by optimizing the use of nursing human resources. *Implementation Science*, 10(1).

Santos, J., dos Santos, H. D. P., and Vieira, R. (2020). Fall detection in clinical notes using language models and token classifier. In *2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS)*, pages 283–288.

Shiner, B., Neily, J., Mills, P., and Watts, B. (2016). Identification of inpatient falls using automated review of text-based medical records. *Journal of Patient Safety*.

Swift, C. G. and Iliffe, S. (2014). Assessment and prevention of falls in older people–concise guidance. *Clinical medicine*, 14(6):658.

Topaz, M., Murga, L., Gaddis, K., McDonald, M., Bar-Bachar, O., Goldberg, Y., and Bowles, K. (2019). Mining fall-related information in clinical notes: Comparison of rule-based and novel word embedding-based machine learning approaches. *Journal of Biomedical Informatics*, 90.

Toyabe, S.-I. (2012). Detecting inpatient falls by using natural language processing of electronic medical records. *BMC Health Services Research*, 12(1).

Tremblay, M., Berndt, D., Luther, S., Foulis, P., and French, D. (2009). Identifying fall-related injuries: Text mining the electronic medical record. *Information Technology and Management*, 10(4):253–265.

Walsh, M., Frances Horgan, N., Walsh, C., and Galvin, R. (2016). Systematic review of risk prediction models for falls after stroke. *Journal of Epidemiology and Community Health*, 70(5):513–519.

Weed-Pfaff, S., Nutter, B., Bena, J., Forney, J., Field, R., Szoka, L., Karius, D., Akins, P., Colvin, C., and Albert, N. (2016). Validation of predictors of fall events in hospitalized patients with cancer. *Clinical Journal of Oncology Nursing*, 20(5):E126–E131.

Yokota, S., Endo, M., and Ohe, K. (2017). Establishing a classification system for high fall-risk among inpatients using support vector machines. *CIN - Computers Informatics Nursing*, 35(8):408–416.

Yokota, S. and Ohe, K. (2016). Construction and evaluation of find, a fall risk prediction model of inpatients from nursing data. *Japan Journal of Nursing Science*, 13(2):247–255.

Zhu, V. J., Walker, T. D., Warren, R. W., Jenny, P. B., Meystre, S., and Lenert, L. A. (2017). Identifying falls risk screenings not documented with administrative codes using natural language processing. In *AMIA annual symposium proceedings*, volume 2017, page 1923. American Medical Informatics Association.