dHealth 2023
B. Pfeifer et al. (Eds.)

doi:10.3233/SHTI230006

Machine Learning Based Risk Prediction for Major Adverse Cardiovascular Events for ELGA-Authorized Clinics

Seda POLAT ERDENIZ[a,b,c,1], Diether KRAMER[a], Michael SCHREMPF [a,b], Peter P. RAINER[b], Alexander FELFERNIG[c], Trang N.T. TRAN[c], Tamim BURGSTALLER[c] and Sebastian LUBOS[c]
[a] *Steiermärkische Krankenanstaltengesellschaft m. b. H., Graz, Austria*
[b] *Medical University of Graz, Graz, Austria*
[c] *Graz University of Technology, Graz, Austria*

Abstract. Background: Artificial Intelligence (AI) has had an important impact on many industries as well as the field of medical diagnostics. In healthcare, AI techniques such as case-based reasoning and data driven machine learning (ML) algorithms have been used to support decision-making processes for complex tasks. This is used to assist medical professionals in making clinical decisions. A way of supporting clinicians is providing predicted prognoses of various ML models. Objectives: Training an ML model based on the data of a hospital and using it on another hospital have some challenges. Methods: In this research, we applied data analysis to discover required data filters on a hospital's EHR data for training a model for another hospital. Results: We applied experiments on real-world data of ELGA (Austrian health record system) and KAGes (a public healthcare provider of 20+ hospitals in Austria). In this scenario, we train the prediction model for ELGA-authorized health service providers using the KAGes data since we do not have access to the complete ELGA data. Conclusion: Finally, we observed that filtering the data with both feature and value selection increases the classification performance of the prediction model, which is trained for another system.

Keywords. Data Filtering, Model Training, Transfer Learning.

1. Introduction

ELGA GmbH [7] is the legal body of Austria responsible for the Austrian health record system. The Austria Federal Government, the Federal States and the Austrian Social Insurance and integration (so the data flow to ELGA), founded it in 2009 with Austrian hospitals starting mostly by 2015. The objective of the non-profit company is the provision of e-health services for the implementation of the national health record system (ELGA). The individual participants and the authorized health service providers (general practitioners, hospitals, laboratories, pharmacies etc.) can access respective health data.

However, ELGA still does not hold the complete patient data of Austrian hospitals. KAGes data after 2015 is being sent partly and the data before 2015 is completely missing at ELGA. From other Austrian hospitals, we do not know exactly which data is

[1] Corresponding Author: Seda Polat Erdeniz, Steiermärkische Krankenanstaltengesellschaft m.b.H., Graz, Austria, E-Mail: seda.polaterdeniz@kages.at

This study received approval from the Ethics Committee of the Medical University of Graz (approval no. 30-146 ex 17/18) and was supported by research grants of Gesundheitsfonds Steiermark and from ERA PerMed within the project "PreCareML".

not sent to ELGA since we do not have access to their EHR data. We only know that their data before 2015 is also missing at ELGA.

Therefore, if a machine-learning model has been developed based on the data of an Austrian hospital, this model cannot be directly applied at ELGA-authorized health service providers. The model should be able to work with the limited data of ELGA. The model should be retrained according to the data available at ELGA.

Our motivation is to make our MACE prognosis prediction model (which is trained on KAGes data) usable at ELGA-authorized health service providers. For this purpose, we did a research based on patient data and machine learning models at KAGes GmbH (Steiermärkische Krankenanstaltengesellschaft m.b.H.). KAGes, which is the public healthcare provider of 20+ hospitals in the federal state of Styria (Austria). KAGes has developed various machine learning based prognosis prediction models with its own patient data, which are running on KAGes hospitals (e.g., the prediction of delirium, dysphagia and MACE prognoses) [4,5]. In this research, we discuss challenges and propose solutions for running a machine learning based prognosis model on ELGA-authorized health service providers. We explain the technical approaches for adapting the MACE prognosis prediction model (trained with the data of KAGes) for ELGA-authorized health service providers.

In the following sections, we describe how we increased the classification performance of a prediction model, which is trained on a dataset (KAGes data) and tested on another dataset (ELGA data). We first applied descriptive analytics based on KAGes data and the data sent from KAGes to ELGA. Afterwards, we show how these findings lead us to apply a predictive analytics for MACE prognosis prediction based on ELGA EHR data. Finally, we compare the prediction model performances based on KAGes and ELGA data and conclude the paper with discussions on the results.

2. Methods

The target of this paper is to train a model (MACE prognosis prediction model) on a training dataset (KAGes data) which will provide predictions for a different test dataset (ELGA-authorized health service providers' data).

For improving the classification performance, we need to filter the KAGes training dataset according to ELGA data with "Data Filters". We know that the data flow to ELGA starts by 2016 (or end of 2015). However, even after 2016 there is still some part of the data, which does not flow to ELGA. When we find the filters of "not flowing data", we can also apply it to the past (the data between 2000-2016) data and build up an "ELGA-compatible" training dataset. Thus, "ELGA-compatible" dataset holds all KAGes data sent to ELGA after 2016 and also the data that would be sent to ELGA if the integration was done by 2000, which makes the trained model richer.

Therefore, we aimed to find such filters of "which data is not sent to ELGA now?" Afterwards, we could apply these rules on the complete training dataset of KAGes and obtain a subset of this dataset, which represents the "ELGA-compatible" training dataset. With this approach, we targeted to increase the relation between patient data of ELGA and the patient data in the training dataset. Because of this approach, we expected to obtain a higher classification performance for MACE prognosis prediction for ELGA-authorized health service providers.

2.1. Document types

The documents sent to ELGA from KAGes are laboratory (LAB), radiology (RBF), and doctoral letter (KHD) documents. For MACE prognosis prediction, the most important documents out of those are Laboratory and Doctoral Letter documents. With these two documents, we can obtain 3118 features for the training dataset. Demographic information (age, academic title, location, etc.) about the patient are already stored in all ELGA-authorized health service providers, so we assume these 30 features are also available.

Filter.1: Applied procedures (LEI, 3819 features) are missing at ELGA-authorized health service providers, so this feature group is filtered out of the training dataset.

Filter.2: Nursing assessments (PFASE, 188 features) in ELGA are observed in a very low amount; so this feature group is completely filtered out of the training dataset.

Therefore, in order to prepare the training dataset for a MACE prognosis prediction model for ELGA-authorized health service providers, we focused on analyzing two documents: Laboratory and Doctoral Letter documents.

2.1.1. Laboratories

Out of doctoral letter documents, we extract diagnoses as LAB features. First, we compared the number of laboratory documents in KAGes and ELGA. The data sharing with ELGA started by the end of 2015 as seen in Figure 1.

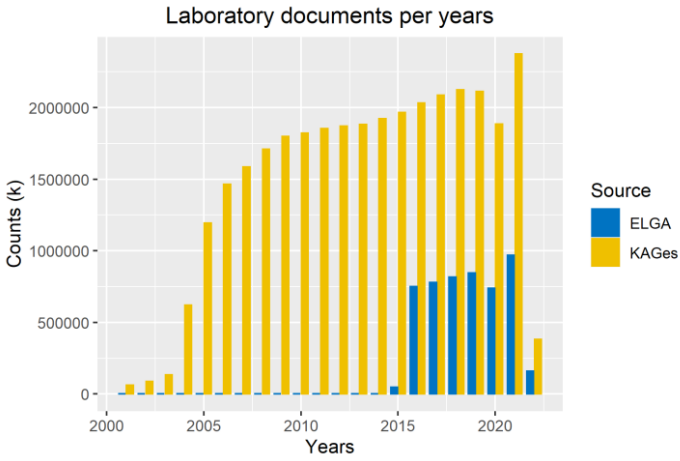

Figure 1. Laboratory documents sent to ELGA from KAGes vs. stored ones in KAGes until 02.03.2022.

Therefore, we take the start date as 2016 for our analyses to make them easier. However, still KAGes holds more laboratory documents than the ones sent to ELGA. Therefore, we made further analyses to figure out which laboratory documents have not been sent to ELGA by KAGes since 2016.

As observed in the ambulant vs. stationary laboratory documents analysis in Table 1, there is a high number of missing stationary laboratory documents in ELGA. There is

also a small gap in ambulant laboratory documents. In order to understand these gaps better, we tested several hypotheses.

Table 1. Ambulant and stationary laboratory documents between beginning of 2016 and end of 2021.

	KAGes	ELGA	Missing in ELGA
Ambulant	4.34M	4.28M	0.1M
Stationary	8.10M	0.42M	7.7M

The first hypothesis to explain the laboratory documents gap in ELGA after 2016 was "Only the latest version of laboratory documents are sent to ELGA from KAGes, so the number of documents of ELGA is much lower for stationary cases". This hypothesis could explain this status, since during a stationary case there can be multiple laboratory documents and only the latest one can be sent to ELGA. However, in ambulant cases, we do not expect multiple laboratory documents since patients do not stay in the hospital. Therefore, this hypothesis explains the gap pretty well. However, as we get the number of distinct documents without including version numbers, results did not change significantly. However, this was a wrong hypothesis.

Filter.3: Therefore, we decided to simply exclude stationary laboratory documents from the training dataset.

2.1.2. Diagnoses

Out of doctoral letter documents, we extract diagnoses as ICD features, medications as MEDI features and alcohol consumption, smoking and body mass index (BMI) as ASM features. As we compared the doctoral letters at KAGes and ELGA, we realized that the gap is very low compared to the laboratory documents.

Table 2. Ambulant and stationary diagnosis documents of 2016-2021.

	KAGes	ELGA	Missing in ELGA
Ambulant	0.072M	0.002M	0.07M
Stationary	1.80M	1.67M	0.13M

Moreover, the diagnosis codes extracted out of doctoral letters are also not completely same with the diagnosis codes in KAGes database. While creating a doctoral letter electronically using OpenMedocs[2] System (EHR system of KAGes), physicians are selecting some or all of the diagnoses (ICD codes) of this case recorded in the KAGes database. However, in some cases, they do not include all the ICD codes in doctoral letters for some reasons (adding only major diagnoses not the minor ones, excluding some diagnoses, which are not relevant for the case anymore, etc.). Therefore, KAGes database always has an equal or more amount of ICD codes per case compared to the doctoral letters sent to ELGA.

Table 3. ICD code occurrences comparison for 39k cases.

Comparison of doctoral letters and KAGes diagnoses database	Amount
Amount of ICD codes which occur the same times in both	36%
Amount of ICD codes which are missing in doctoral letters less than 15% of the cases	73%
Amount of ICD codes which are missing in doctoral letters greater than 50% of the cases	5%

Filter.4: After this comparison, from the training dataset we decided to remove the ICD features, which are missing in ELGA greater than 50% of the cases.

[2] https://www.landesrechnungshof.steiermark.at/cms/beitrag/12610583/136482471/

3. Result

We aimed to train a machine-learning model that can work at ELGA-authorised health care providers in high performance (in terms of predictive power). For this purpose, we trained two new random forest models: ELGA Model #1 and ELGA Model #2 based on the assumed available data at ELGA (according to the data analysis results).

Table 4. Provided set of features (before feature selection) for the HIS and ELGA MACE-prediction Models

Feature Type	Description	n (ELGA-Models)	n (HIS-Models)
socio demographic data	age, gender, area of residence, etc	30	30
approximate string matching strategy	smoking behaviour, alcohol consumption, obesity	3	3
Diagnosis codes	ICD-10 Codes, ICD-10 groups	1042	1069
Procedures Codes	diagnostic and curative procedures (e.g. CT)	0	3819
Medication	ATC-Codes	323	323
Nursing Protocols	body mass index, movement disorders, etc.	0	188
Labaratory	LOINC-Codes (e.g. thrombocytes, creatine)	1724	1724

We used two different training data sets for the two models. For ELGA Model #1, we used feature selection filters (Filter.1, Filter.2 and Filter.4); whereas for ELGA Model #2 we used both feature and value selection filters (Filter.1, Filter.2, Filter.3 and Filter.4) (see Table 4).

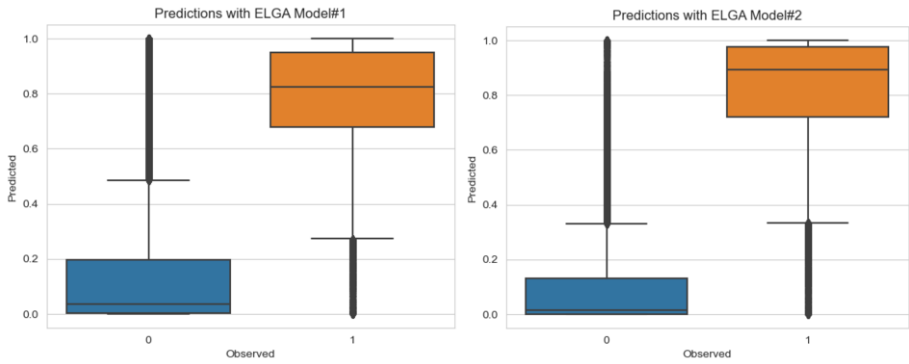

Figure 2. Classification results of ELGA Model #1 and ELGA Model #2 (0: NO-MACE, 1: MACE).

As it can be easily observed in Figure 2, ELGA Model #2 improves the discrimination power of the model since the probability range of no-MACE and MACE cases are far less overlapping than the ones in ELGA Model #1. In Figure 3, we compare the AUROC performance results of the models with their confidence intervals. As observed, ELGA Model #2 is improving classification performance (in terms of AUROC) significantly since ELGA Model #2 has a higher DeLong CI [6] (0.969-0.971) which is not overlapping Model #1's DeLong CI (0.934-0.937).

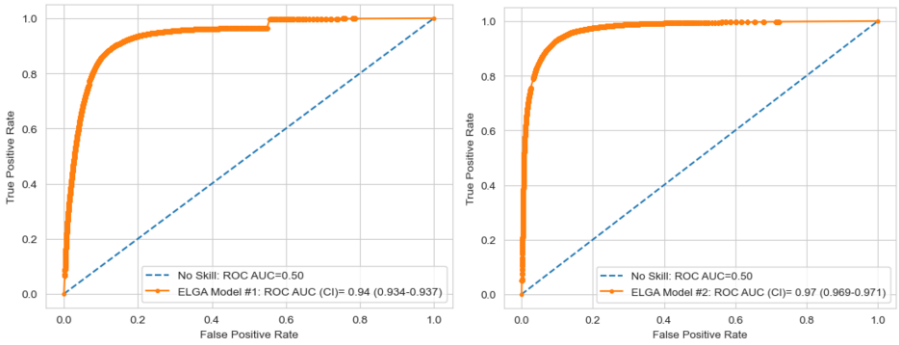

Figure 3. Performance results (AUC with CI) of ELGA Model #1 and ELGA Model #2.

4. Discussion

In this machine learning research in healthcare, we aimed to train a MACE prediction model for ELGA-authorized health care providers which is trained on EHR data at KAGes. For this purpose, we analyzed the differences in the data stored at KAGes and the data sent to ELGA from KAGes. Accordingly, we discovered four filters, which are applied on the training dataset of the KAGes MACE prediction model. We defined data filters as feature selection and value selection filters. Using this definition, we trained two models: ELGA Model #1 (only with feature selection filters) and ELGA Model #2 (both with feature and value selection filters). Based on the experiments, we observed that ELGA Model #2 works better than ELGA Model #1 on a simulated test data set. This showed us that when re-training a model for a subset of the training dataset, it is better to consider filtering the training data in terms of both feature and value selections, which increases the similarity with the test data and results with better classification performance of the prediction model.

References

[1] George W Beeler Jr, Patricia S Gibbons, and Christopher G Chute. 1992. Development of a clinical data architecture. In Proceedings of the Annual Symposium on Computer Application in Medical Care. American Medical Informatics Association, 244.

[2] Laxmaiah Manchikanti, Frank JE Falco, and Joshua A Hirsch. 2013. Ready or not! Here comes ICD-10. Journal of neurointerventional surgery 5, 1 (2013), 86–91.

[3] Clement J McDonald, Stanley M Huff, Jeffrey G Suico, Gilbert Hill, Dennis Leavelle, Raymond Aller, Arden Forrey, Kathy Mercer, Georges DeMoor, John Hook, et al. 2003. LOINC, a universal standard for identifying laboratory observations: a 5-year update. Clinical chemistry 49, 4 (2003), 624–633.

[4] Schrempf, Michael, et al. Development of an Architecture to Implement Machine Learning Based Risk Prediction in Clinical Routine: A Service-Oriented Approach. Studies in health technology and informatics 293 (2022): 262-269.

[5] Erdeniz, Seda Polat, et al. Explaining Machine Learning Predictions of Decision Support Systems in Healthcare. Current Directions in Biomedical Engineering 8.2 (2022): 117-120.

[6] DeLong ER, DeLong DM, Clarke-Pearson DL (1988) Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. Biometrics 44:837. https://doi.org/10.2307/2531595

[7] Herbek, Susanne, et al. "The Electronic Health Record in Austria: a strong network between health care and patients." European Surgery 44 (2012): 155-163.