Quantifying medical histories with the Kaplan-Meier (KM) estimator for feature extraction of electronic health records in machine learning

Herdiantri Sufriyana

(1) Graduate Institute of Biomedical Informatics, College of Medical Science and Technology, Taipei Medical University, 250 Wu-Xing Street, Taipei 11031, Taiwan. (2) Department of Medical Physiology, Faculty of Medicine, Universitas Nahdlatul Ulama Surabaya, 57 Raya Jemursari Road, Surabaya 60237, Indonesia. https://orcid.org/0000-0001-9178-0222

Yu Wei Wu

(1) Graduate Institute of Biomedical Informatics, College of Medical Science and Technology, Taipei Medical University, 250 Wu-Xing Street, Taipei 11031, Taiwan. (2) Clinical Big Data Research Center, Taipei Medical University Hospital, 250 Wu-Xing Street, Taipei 11031, Taiwan. https://orcid.org/0000-0002-5603-1194

Emily Chia-Yu Su (emilysu@tmu.edu.tw)

(1) Graduate Institute of Biomedical Informatics, College of Medical Science and Technology, Taipei Medical University, 250 Wu-Xing Street, Taipei 11031, Taiwan. (2) Clinical Big Data Research Center, Taipei Medical University Hospital, 250 Wu-Xing Street, Taipei 11031, Taiwan. (3) Research Center for Artificial Intelligence in Medicine, Taipei Medical University, 250 Wu-Xing Street, Taipei 11031, Taiwan. https://orcid.org/0000-0003-4801-5159

Method Article

Keywords: medical history, electronic health record, Kaplan-Meier estimator, feature extraction, machine learning

Posted Date: October 13th, 2021

DOI: https://doi.org/10.21203/rs.3.pex-1635/v1

Abstract

This protocol aimed to describe data transformation procedure of medical histories from electronic health records (EHRs) to historical rates by Kaplan-Meier (KM) estimation. The applicability is to extract features from real-world, time-varying data of EHRs, for developing but not limited to a machine learning prediction model. By this extraction technique, machine can learn medical history of a condition in each healthcare provider, as a differential quantity through time in term of affecting a future health state, without a need to access EHRs of other healthcare providers. However, this protocol needs a sufficient amount of longitudinal data from the most subjects in EHRs. The key stages consisted of time interval computation, historical rate derivation, and data transformation into historical rates.

Introduction

Implementation of electronic health records (EHRs) is a high-resource investment which increases healthcare costs and burden in clinical workflow. Meanwhile, efficiency without exposing harm to patient is mandated due to emerging adoption of insurance-based healthcare worldwide. This adoption also warrants more preventive measures in current practice; thus, clinical prediction models have been developed for several health outcomes, including those using either EHRs or machine learning.[1-4] The most of well-known successes were clinical prediction models by machine learning, that used data from medical devices stored in EHRs.[5-7] Medical histories of diagnoses and procedures in EHRs are abundant but not optimally explored compared to other data types, e.g. laboratory tests, medical imaging, and genomic data.[8]

Recorded as a diagnosis/procedure code, a condition affects a health state in the future relative to the time interval between that condition and the health state at the time of the prediction. The effect may be improving, devastating, or undifferentiated, but there should be a quantity that differentiates through time; thus, a computer is able to associate such a trend of a condition with health states among individuals. That quantity should also differ among conditions given the same time interval to the time of prediction, and be inferred from the population as a generalized quantity for each individual given the time interval. This is the intuition behind application of the Kaplan-Meier (KM) estimator in this protocol. It also allows comparable medical history across healthcare providers, since this type of data is often isolated within each healthcare provider. By population-level historical rate, it is possible to utilize the isolated data across healthcare providers without accessing the respective databases.

This protocol was already applied in a project consisting multiple studies that compared a human-machine learning algorithm with those applying human learning with statistical methods and those applying other machine learning algorithms. Ethical clearance was waived by the Taipei Medical University Joint Institutional Review Board (TMU-JIRB number: N202106025). We aimed to describe data transformation procedure of medical histories from electronic health records (EHRs) to historical rates by Kaplan-Meier (KM) estimation for developing but not limited to a machine learning prediction model.

Reagents

Equipment

To conduct this protocol, we used R 4.0.2 programming language (R Foundation, Vienna, Austria). The integrated development environment software was RStudio 1.3.959 (RStudio PBC, Boston, MA, USA). We used Bioconductor 3.11 to ensure reproducibility;[9] thus, versions of the included R packages were all in sync according to versions in this Bioconductor version. Since the main context of this protocol was machine learning predictive modeling, we used an R package of caret 6.0.86, particularly for data partition, and we also created an R package medhist 0.1.0 which facilitated most steps of this protocol. Our package is available for download from this repository https://github.com/herdiantrisufriyana. Details on other R package versions and all of the source codes (vignette) for the data analysis are available in https://github.com/herdiantrisufriyana/hist_rate.

To reproduce our work, a set of hardware requirements may be needed. We used a single machine. It was equipped by 8 logical processors for the 3.40 GHz central processing unit (CPU) (Core (TM) i7-4770, Intel®, Santa Clara, CA, USA), and 16 GB RAM. But, one can use a machine with only 4 logical processors and 4 GB RAM, if the sample size is smaller than that of dataset we used in this protocol.

Procedure

1. Compute time intervals between code or variable encounters and each visit of any subject

This step was to quantify medical histories into time intervals. In addition to a single-code variable, we also used variables that assigned multiple codes of diagnosis and procedure. Details of the codes are available in the source codes. We computed the number of days for a code or variable in the latest encounter before the current visit (the time of prediction or d_0).

2. Split a dataset randomly for a derivation and validation set

We only used the derivation set to infer KM estimates at the population level. Later, we also used it for training set of a prediction model; thus, our models were blinded to the distribution of KM estimates of any external validation sets.

3. Calculate KM estimate for each code or variable

The KM estimate (equation in Figure 1),[10] as we denote it as an estimator of the historical function *H(d)*, is a probability or a fraction of visits that a condition is longer than d days before d_0. This consists of d_i as a day when at least an encounter occurred for a code or variable that refers to that condition, e_i is the

number of encounters for the code or variable on day d_i, and v_j is a visit recorded that did not encounter the code or variable (it was censored) up to day d_i.

4. Apply interpolation between time points of historical rates

Because there might be a day on which no code or variable is encountered in the population, we applied a linear interpolation between time points at which a KM estimate was able to be calculated. Other methods of interpolation are also available in medhist 0.1.0, an R package, that allows future investigators to implement this historical rate.

5. Compute provider-wise time intervals

Medical histories for derivation set were nationwide, while those for training set were provider-wise by estimation. This means our prediction models only used medical histories recorded by a healthcare provider, which was blinded to those recorded by others. This reflects most real-world situations in which a healthcare provider does not have access to medical records of other providers.

6. Transform a provider-wise time interval into a nationwide KM estimate.

A KM estimate of a code for an individual was then determined given the number of days from a code or variable encounter to the current visit. The number of days was matched with those in historical rates of each code. We already have nationwide historical (KM) rates for each code or variable, derived from the training set only. All medical histories in days of a subject in each healthcare provider were transformed into historical rates. This technique allowed generalization of individual data based on nationwide, population-level data, without the need to access data from other providers.

Troubleshooting

Step 1 and 5

Problem

Premature stop of computation

Possible reason

Dataset consists of a large sample size or number of variables.

Solution

Consider to use computer with larger memory size or RAM. Alternatively, split table into ones consisting unique sets of variables.

Step 2

Problem

Unexpected error

Possible reason

Proportion of a partition may be too small.

Solution

Consider to increase the proportion.

Step 3 to 4

Problem

Unable to calculate KM estimate

Possible reason

A variable is only available in a visit.

Solution

Remove this kind of variable.

Step 6

Problem

Unexpected error

Possible reason

There is no expected variable in the source of historical rates.

Solution

Check if the variable exists or the variable name is correct.

Time Taken

All steps

Approximate time: 10 to 20 minutes (pre-computed)

Step 1 and 5

Approximate time: 5 to 10 minutes (pre-computed)

Step 2

Approximate time: 1 to 2 minutes

Step 3 to 4

Approximate time: 1 to 2 minutes

Step 6

Approximate time: 1 to 2 minutes

Anticipated Results

A historical rate ranges from 0 to 1. If interpolation is not used, then one may find multiple instances with the same historical rate although the time intervals are different. This was why we need interpolation to estimate the rate in a period of two time points, in which there was no available data for a code or variable.

References

1. Sufriyana, H., et al. Comparison of Multivariable Logistic Regression and Other Machine Learning Algorithms for Prognostic Prediction Studies in Pregnancy Care: Systematic Review and Meta-Analysis. JMIR Med Inform 8, e16503 (2020).

2. Fleuren, L.M., et al. Machine learning for the prediction of sepsis: a systematic review and meta-analysis of diagnostic test accuracy. Intensive Care Med 46, 383-400 (2020).

3. Lee, Y., et al. Applications of machine learning algorithms to predict therapeutic outcomes in depression: A meta-analysis and systematic review. J Affect Disord 241, 519-532 (2018).

4. Gonem, S., Janssens, W., Das, N. & Topalovic, M. Applications of artificial intelligence and machine learning in respiratory medicine. Thorax 75, 695-701 (2020).

5. Bien, N., et al. Deep-learning-assisted diagnosis for knee magnetic resonance imaging: Development and retrospective validation of MRNet. PLoS Med 15, e1002699 (2018).

6. Hannun, A.Y., et al. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. Nat Med 25, 65-69 (2019).

7. Rajpurkar, P., et al. Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists. PLoS Med 15, e1002686 (2018).

8. Scott, I., Carter, S. & Coiera, E. Clinician checklist for assessing suitability of machine learning applications in healthcare. BMJ Health Care Inform 28(2021).

9. Huber, W., et al. Orchestrating high-throughput genomic analysis with Bioconductor. Nat Methods 12, 115-121 (2015).

10. Stalpers, L.J.A. & Kaplan, E.L. Edward L. Kaplan and the Kaplan-Meier Survival Curve. BSHM Bulletin: Journal of the British Society for the History of Mathematics 33, 109-135 (2018).

Acknowledgements

The social security administrator for health or *badan penyelenggara jaminan sosial (BPJS) kesehatan* in Indonesia gave permission to access the sample dataset in this protocol (dataset request approval number: 5064/I.2/0421). This protocol was funded by the Ministry of Science and Technology (MOST) in Taiwan (grant number MOST109-2221-E-038-018 and MOST110-2628-E-038-001) and the Higher Education Sprout Project from the Ministry of Education (MOE) in Taiwan (grant number DP2-110-21121-01-A-13) to Emily Chia-Yu Su.

Figures

$$\hat{H}(d) = \prod_{i=1}^{d} \left(1 - \frac{e_i}{v_i}\right)$$

Figure 1

Equation of historical rate

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- suppprotocolhistrate.pdf