Mining Electronic Health Records for Real-World Evidence

Chengxi Zang
chz4001@med.cornell.edu
Weill Cornell Medical College,
Cornell University.
New York, NY, USA

Weishen Pan
wep4001@med.cornell.edu
Weill Cornell Medical College,
Cornell University.
New York, NY, USA

Fei Wang
few2001@med.cornell.edu
Weill Cornell Medical College,
Cornell University.
New York, NY, USA

ABSTRACT

The rapid accumulation of large-scale Electronic Health Records (EHR) presents considerable opportunities to generate real-world evidence to inform clinical decision-making and accelerate drug development. However, the complexity of EHR has turned them into a formidable testing ground for cutting-edge AI algorithms. Furthermore, a significant gap still exists between algorithm development in the computer science community and clinical translation within the healthcare community. This tutorial aims to bridge this divide by fostering mutual understanding between the two communities by discussing using advanced machine learning and data mining technologies tailored to tackle real-world healthcare challenges, including 1) using EHR and trial emulation for understanding Long Covid and drug repurposing for Alzheimer's disease, and 2) risk prediction and associated fairness, interpretability, generalizability, etc., issues. We will conclude this tutorial by delving into potential opportunities for future research and unveiling the prospects of a career as a health data scientist.

CCS CONCEPTS

• **Applied computing** → **Health informatics**; *Health care information systems*; • **Computing methodologies** → **Causal reasoning and diagnostics**.

KEYWORDS

Real-World Data, Real-World Evidence, Electronic Health Records, Healthcare, Trial Emulation, Predictive modeling, Causal Inference

ACM Reference Format:
Chengxi Zang, Weishen Pan, and Fei Wang. 2023. Mining Electronic Health Records for Real-World Evidence. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '23), August 6–10, 2023, Long Beach, CA, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3580305.3599566

1 INTRODUCTION

Real-world data (RWD) are usually referred to as patients' data collected during the delivery of health care. Common real-world data sources include electronic health records (EHRs), administrative claims, etc. Taking EHRs as an example, they can have a variety of data from structured domains (e.g., diagnoses, prescriptions, procedures, laboratory tests, vital signs, etc.) to unstructured domains (e.g., clinical notes, medical images, etc.). Real-World Evidence (RWE) is defined by the FDA as "clinical evidence about the usage and potential benefits or risks of a medical product derived from analysis of RWD" [2], or can be extended to clinical evidence generated from observational noninterventional study [3]. In this tutorial, we aim to introduce how to use EHRs and machine learning methods to solve real-world healthcare challenges. We will introduce machine-learning-driven trial emulation methods and how to use them to improve our understanding of and ability to predict, treat, and prevent the post-acute sequelae of SARS-CoV-2 (or Long COVID) [6, 9, 16, 19, 20, 22], and to do comparative effectiveness analysis [12] and drug repurposing [18] for Alzheimer's disease.

On the other hand, we will introduce machine learning and EHR-based risk prediction [8, 11, 13–15, 17] and highlight critical issues associated including fairness, interpretability, and generalizability. We first explore methods to measure and address algorithmic disparities (potential discrimination against certain disadvantaged subpopulations) in risk prediction models[4, 5]. Then, we will discuss the need for interpretability [1, 10] and introduce the methods to explain risk prediction models [7]. Lastly, we will introduce how to train a risk prediction model with better generalizability when applied to different populations or datasets [21]. Tutorial materials are available at www.calvinzang.com/ehr4rwe_kdd2023.html and the outline is summarized below:

(1) Introduction (30 min)
(2) Trial Emulation for Generating Real-world Evidence (60 min)
    - Randomized Controlled Trial, Trial Emulation, and Machine Learning-driven Trial Emulation for Causal Inference
    - Using EHR and Trial Emulation to understand Long COVID
    - Using EHR and Trial Emulation for Alzheimer's disease drug repurposing
(3) Advancements in Risk Prediction for Healthcare (60 min)
    - Machine Learning for Risk Prediction in Health Care
    - Quantifying and Addressing Algorithmic Disparity
    - Explaining Models by Causal Path Decomposition
    - Improving Model Generalizability across Multiple Sites
(4) Conclusion: Discussion and Future Direction (20 min)

Societal Impacts and Audience. This tutorial tries to bridge the gap between methodology (CS community) and clinical translation (medical community). Machine learning tools (e.g., causal inference, predictive modeling) for mining EHRs tailored to specific healthcare applications will be introduced. This tutorial will be highly accessible to all data mining researchers, students, and practitioners who are interested in health data science. The tutorial will be self-contained and no special prerequisite knowledge is required.

KDD '23, August 6–10, 2023, Long Beach, CA, USA

ACM ISBN 979-8-4007-0103-0/23/08.
https://doi.org/10.1145/3580305.3599566

2 PRESENTER INFORMATION

Chengxi Zang is currently an Instructor in the Department of Population Health Sciences, at Weill Medical College of Cornell University. He got his Ph.D. from Tsinghua University with an Excellent Ph.D. Dissertation Award in the Computer Science Department. His research focus is using AI, Machine Learning, and large-scale Real-World Health Data to solve challenging healthcare problems, including drug repurposing for Alzheimer's Disease, suicide prevention, understanding Long Covid, etc. His research has been published in the top venues of related areas such as Nature Medicine, Nature Communications, Journal of General Internal Medicine, Scientific Reports, Cell Patterns, TKDE, KDD, AAAI, ICDM, and his papers have won ICDM'18 Best Paper Candidate and the Best Paper Award at AAAI'20 Workshop on Deep Learning on Graphs.

Weishen Pan is currently a postdoctoral research associate in the Department of Population Health Sciences, Weill Cornell Medicine, Cornell University. He got his Ph.D. from Tsinghua University. His primary research interest is machine learning algorithms development in computational medicine, particularly on model fairness and interpretability. He has published on top machine learning and data mining conferences including KDD and NeurIPS. His research on explaining the algorithmic disparity by causal pathway decomposition was highlighted in AMIA 2021 Year-in-Review Session. He won the data challenge on PTHrP results prediction organized by AACC as the core team member.

Fei Wang is currently an Associate Professor of Health Informatics in the Department of Population Health Sciences, Weill Cornell Medicine, Cornell University. His major research interest is data mining and its applications in health data science. His papers have received over 25,000 citations so far with an H-index 77. His papers have won 7 best paper awards at top international conferences on data mining and medical informatics. His team won the championship of the NIPS/Kaggle Challenge on Classification of Clinically Actionable Genetic Mutations in 2017 and Parkinson's Progression Markers' Initiative data challenge organized by Michael J. Fox Foundation in 2016. Dr. Wang is the recipient of the NSF CAREER Award in 2018, the inaugural research leadership award in IEEE International Conference on Health Informatics (ICHI) 2019, Amazon AWS Machine Learning for Research Award in 2017 and 2019, as well as Google Faculty Research Award. Dr. Wang's Research has been supported by NSF, NIH, ONR, PCORI, MJFF, AHA, etc. Dr. Wang is the chair of the Knowledge Discovery and Data Mining working group in the American Medical Informatics Association (AMIA).

ACKNOWLEDGMENTS

This work was supported by NSF 1750326.

REFERENCES

[1] Bing Bai, Jian Liang, Guanhua Zhang, Hao Li, Kun Bai, and Fei Wang. 2021. Why attentions may not be interpretable?. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 25–34.

[2] Office of the Commissioner. 2023. Real-World Evidence. https://www.fda.gov/science-research/science-and-research-special-topics/real-world-evidence Publisher: FDA.

[3] John Concato and Jacqueline Corrigan-Curay. 2022. Real-world evidence-where are we now? *The New England journal of medicine* 386, 18 (2022), 1680–1682.

[4] Sen Cui, Weishen Pan, Jian Liang, Changshui Zhang, and Fei Wang. 2021. Addressing algorithmic disparity and performance inconsistency in federated learning. *Advances in Neural Information Processing Systems* 34 (2021), 26091–26102.

[5] Sen Cui, Weishen Pan, Changshui Zhang, and Fei Wang. 2021. Towards model-agnostic post-hoc adjustment for balancing ranking fairness and algorithm utility. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 207–217.

[6] Dhruv Khullar, Yongkang Zhang, Chengxi Zang, Zhenxing Xu, Fei Wang, Mark G Weiner, Thomas W Carton, Russell L Rothman, Jason P Block, and Rainu Kaushal. 2023. Racial/Ethnic Disparities in Post-acute Sequelae of SARS-CoV-2 Infection in New York: an EHR-Based Cohort Study from the RECOVER Program. *Journal of General Internal Medicine* 38, 5 (2023), 1127–1136.

[7] Weishen Pan, Sen Cui, Jiang Bian, Changshui Zhang, and Fei Wang. 2021. Explaining algorithmic fairness through fairness-aware causal path decomposition. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 1287–1297.

[8] Chang Su, Robert Aseltine, Riddhi Doshi, Kun Chen, Steven C Rogers, and Fei Wang. 2020. Machine learning for suicide risk prediction in children and adolescents with electronic health records. *Translational psychiatry* (2020), 413.

[9] Jay K Varma, Chengxi Zang, Thomas W Carton, Jason P Block, Dhruv J Khullar, Yongkang Zhang, Mark G Weiner, Russell L Rothman, Edward J Schenck, Zhenxing Xu, et al. 2023. Excess burden of respiratory and abdominal conditions following COVID-19 infections during the ancestral and Delta variant periods in the United States: An EHR-based cohort study from the RECOVER Program. *medRxiv* (2023), 2023–02.

[10] Fei Wang, Rainu Kaushal, and Dhruv Khullar. 2020. Should health care demand interpretable artificial intelligence or accept "black box" medicine? , 59–60 pages.

[11] Tingyi Wanyan, Hossein Honarvar, Suraj K Jaladanki, Chengxi Zang, Nidhi Naik, Sulaiman Somani, Jessica K De Freitas, Ishan Paranjpe, Akhil Vaid, Jing Zhang, et al. 2021. Contrastive learning improves critical event prediction in COVID-19 patients. *Patterns* 2, 12 (2021), 100389.

[12] Jie Xu, Fei Wang, Chengxi Zang, Hao Zhang, Kellyann Niotis, Ava L Liberman, Cynthia M Stonnington, Makoto Ishii, Prakash Adekkanattu, Yuan Luo, et al. 2023. Comparing the effects of four common drug classes on the progression of mild cognitive impairment to dementia using electronic health records. *Scientific Reports* 13, 1 (2023), 8102.

[13] He S Yang, Yu Hou, Ljiljana V Vasovic, Peter AD Steel, Amy Chadburn, Sabrina E Racine-Brzostek, Priya Velu, Melissa M Cushing, Massimo Loda, Rainu Kaushal, et al. 2020. Routine laboratory blood tests predict SARS-CoV-2 infection using machine learning. *Clinical chemistry* 66, 11 (2020), 1396–1404.

[14] He S Yang, Daniel D Rhoads, Jorge Sepulveda, Chengxi Zang, Amy Chadburn, and Fei Wang. 2022. Building the ModelChallenges and Considerations of Developing and Implementing Machine Learning Tools for Clinical Laboratory Medicine Practice. *Archives of Pathology & Laboratory Medicine* (2022).

[15] Chengxi Zang, Marianne Goodman, Zheng Zhu, Lulu Yang, Ziwei Yin, Zsuzsanna Tamas, Vikas Mohan Sharma, Fei Wang, and Nan Shao. 2022. Development of a screening algorithm for borderline personality disorder using electronic health records. *Scientific Reports* 12, 1 (2022), 1–12.

[16] Chengxi Zang, Yu Hou, Edward Schenck, Zhenxing Xu, Yongkang Zhang, Jie Xu, Jiang Bian, Dmitry Morozyuk, Dhruv Khullar, Anna Nordvig, et al. 2023. Risk Factors and Predictive Modeling for Post-Acute Sequelae of SARS-CoV-2 Infection: Findings from EHR Cohorts of the RECOVER Initiative. *Research Square* (2023), rs–3.

[17] Chengxi Zang and Fei Wang. 2021. SCEHR: Supervised Contrastive Learning for Clinical Risk Prediction using Electronic Health Records. In *2021 IEEE International Conference on Data Mining (ICDM)*. IEEE, 857–866.

[18] Chengxi Zang, Hao Zhang, Jie Xu, Hansi Zhang, Sajjad Fouladvand, Shreyas Havaldar, Feixiong Cheng, Kun Chen, Yong Chen, Benjamin S Glicksberg, et al. 2022. High-throughput clinical trial emulation with real world data and machine learning: a case study of drug repurposing for Alzheimer's disease. *medRxiv* (2022), 2022–01.

[19] Chengxi Zang, Yongkang Zhang, Jie Xu, Jiang Bian, Dmitry Morozyuk, Edward J Schenck, Dhruv Khullar, Anna S Nordvig, Elizabeth A Shenkman, Russell L Rothman, et al. 2023. Data-driven analysis to understand long COVID using electronic health records from the RECOVER initiative. *Nature Communications* 14, 1 (2023), 1948.

[20] Hao Zhang, Chengxi Zang, Zhenxing Xu, Yongkang Zhang, Jie Xu, Jiang Bian, Dmitry Morozyuk, Dhruv Khullar, Yiye Zhang, Anna S Nordvig, et al. 2023. Data-driven identification of post-acute SARS-CoV-2 infection subphenotypes. *Nature Medicine* 29, 1 (2023), 226–235.

[21] Xi Sheryl Zhang, Fengyi Tang, Hiroko H Dodge, Jiayu Zhou, and Fei Wang. 2019. Metapred: Meta-learning for clinical risk prediction with limited patient electronic health records. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*. 2487–2495.

[22] Yongkang Zhang, Hui Hu, Vasilios Fokaidis, Jie Xu, Chengxi Zang, Zhenxing Xu, Fei Wang, Michael Koropsak, Jiang Bian, Jaclyn Hall, et al. 2023. Identifying environmental risk factors for post-acute sequelae of SARS-CoV-2 infection: An EHR-based cohort study from the recover program. *Environmental Advances* 11 (2023), 100352.