A Framework for Understanding Bias and Real-World Clinical Electronic Health Record Data

Haeun Lee[1], MS, Harold Lehmann[1], MD, PhD, Christelle Xiong[1], BS, Derek Baughman[1], MD, MS, Benjamin Martin[1], PhD, Jiayi Tong[2], PhD, Paul Nagy[1], PhD

[1] Biomedical Informatics and Data Science, Johns Hopkins School of Medicine, Johns Hopkins University, Baltimore, MD, USA

[2] Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA

INTRODUCTION

- Real-world data (RWD) from electronic health records (EHRs), combined with AI/ML advances, enables the analysis of healthcare outcomes across large and diverse patient groups in the real-world clinical practice, complementing traditional clinical trials and informing clinical research, drug development, and patient care.
- However, biases persist at the healthcare system level and research level, threatening the validity and generalizability of real-world evidence (RWE) findings.
- Current frameworks lack systematic mapping of where and how biases originate and propagate in large-scale observational healthcare research.

OBJECTIVES

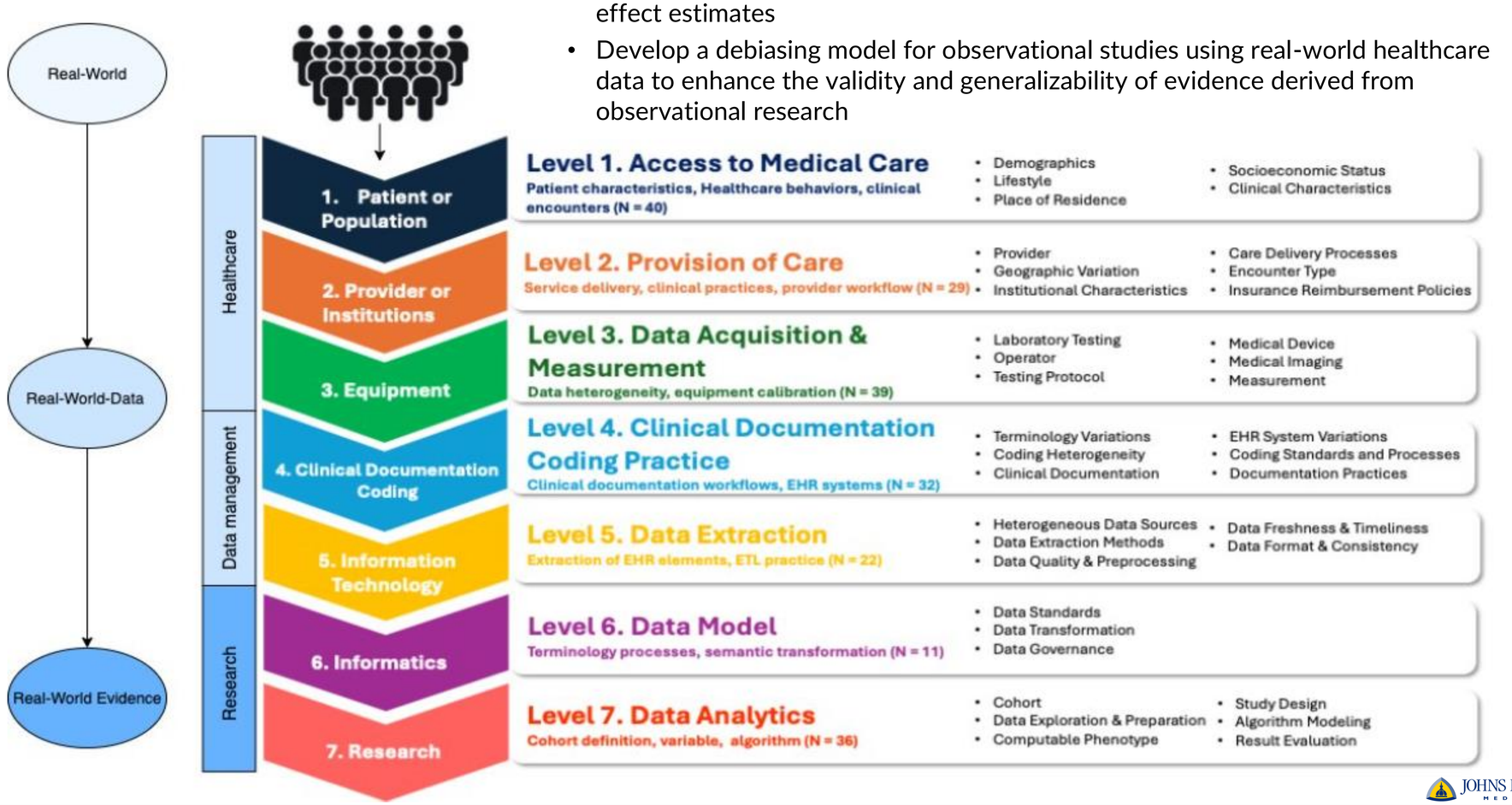
- Develop a conceptual framework mapping the pathway from bedside clinical encounters to research evidence generation, identifying where biases can emerge at each stage
- Systematically categorize documented bias sources using a hierarchical schema

METHODS

- We developed a bias framework through synthesis of established RWE generation pathways and iterative review by a multidisciplinary expert panel of clinicians, informaticians, and biostatisticians.
- To examine bias in RWD, we conducted a scoping review following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews (PRISMA-ScR) framework, searching PubMed and Web of Science for English publications (2016-2025) using “electronic health records” AND “bias” combined with related terms including “patient”, “healthcare”, “diagnostic”, “coding”, “data”, “extraction”, “standards”, “phenotype”, “CDM”, “ETL”, “transformation”, and “mapping”.

RESULTS AND HIGHLIGHTS

- Analysis of 215 papers identified 209 distinct bias sources across three levels (Healthcare, Data, Management, Research) and seven domains, showing how biases cascade through the clinical-to-research pipeline
- The majority of bias sources originated from access to care (Level 1, 19.1%), followed by data acquisition and measurement (Level 3, 18.7%), and data analytics (Level 7, 17.2%)
- Healthcare-level biases (N=108) were most frequently mentioned in the literature



CONCLUSION

- This study developed a multi-level framework spanning healthcare, data management, and research domains, identifying and categorizing 209 bias sources in clinical evidence generation.
- Biases showed interconnected properties that span multiple levels and amplify throughout the data lifecycle, indicating the need for comprehensive bias evaluation across the entire data journey.
- This framework can server as a trigger to help methodologists make sure, they have considered important sources of bias.

NEXT STEPS

- Develop a hierarchical 7-level taxonomy mapping bias sources to specific bias types and their impact on research validity
- Create OMOP-based quantification metrics to operationalize and measure bias indicators across healthcare sites, assessing inter-site variability and generating site-specific bias profiles
- Quantify real-world impact on research findings through clinical scenarios, comparing biased versus corrected analyses to demonstrate bias magnitude on effect estimates
- Develop a debiasing model for observational studies using real-world healthcare data to enhance the validity and generalizability of evidence derived from observational research

www.PosterPresentations.com