Sentinel

Natural Language Processing in Pharmacoepidemiology: Lessons from the Multi-source Observational Safety study for Advanced Information Classification using NLP (MOSAIC-NLP)

2024 ISPE Annual Meeting

August 2024

Contents

Disclosures

Dena Jaffe and **Elise Berliner** are full-time employees of Oracle Health and Oracle Life Sciences. Dena and Elise declare no relevant or material financial interests that relate to the research described in this. Sentinel has contracted with Oracle to complete this work.

Dr. Desai reports serving as Principal Investigator on investigator-initiated grants to the Brigham and Women's Hospital from Novartis, Vertex, and Bayer on unrelated projects.

Sarah Dutcher has no conflicts of interest to disclose.

This project was supported by **Task Order 75F40119F19002** under **Master Agreement 75F40119D10037** from the US Food and Drug Administration (FDA). The views expressed in this presentation are those of the authors and do not necessarily represent the official views of the U.S. FDA.

Presenters

Dena Jaffe
Lead RWD Strategist
Oracle Health

Elise Berliner
Global Senior Principal Real World Evidence
Oracle Life Sciences

Rishi Desai
Associate Professor
Brigham and Women's Hospital, Harvard Medical School

Sarah K. Dutcher
Lead Epidemiologist
US Food and Drug Administration

Leveraging NLP in Pharmacoepidemiology

Rishi J Desai, MS, PhD

Brigham & Women's Hospital/Harvard Medical School

Public Law 110–85
110th Congress

An Act

To amend the Federal Food, Drug, and Cosmetic Act to revise and extend the user-fee programs for prescription drugs and for medical devices, to enhance the postmarket authorities of the Food and Drug Administration with respect to the safety of drugs, and for other purposes.

Sept. 27, 2007
[H.R. 3580]

Be it enacted by the Senate and House of Representatives of the United States of America in Congress assembled,

SECTION 1. SHORT TITLE.

Food and Drug Administration Amendments Act of 2007.
21 USC 301 note.

This Act may be cited as the "Food and Drug Administration Amendments Act of 2007".

SEC. 905. ACTIVE POSTMARKET RISK IDENTIFICATION AND ANALYSIS.

(a) IN GENERAL.—Subsection (k) of section 505 of the Federal Food, Drug, and Cosmetic Act (21 U.S.C. 355) is amended by adding at the end the following:

"(3) ACTIVE POSTMARKET RISK IDENTIFICATION.—

"(A) DEFINITION.—In this paragraph, the term 'data' refers to information with respect to a drug approved under this section or under section 351 of the Public Health Service Act, including claims data, patient survey data, standardized analytic files that allow for the pooling and analysis of data from disparate data environments, and any other data deemed appropriate by the Secretary.

"(B) DEVELOPMENT OF POSTMARKET RISK IDENTIFICATION AND ANALYSIS METHODS.—The Secretary shall, not later than 2 years after the date of the enactment of the Food and Drug Administration Amendments Act of 2007, in collaboration with public, academic, and private entities—

"(i) develop methods to obtain access to disparate data sources including the data sources specified in subparagraph (C);

"(ii) develop validated methods for the establishment of a postmarket risk identification and analysis system to link and analyze safety data from multiple sources, with the goals of including, in aggregate—

"(I) at least 25,000,000 patients by July 1, 2010; and

"(II) at least 100,000,000 patients by July 1, 2012; and

"(iii) convene a committee of experts, including individuals who are recognized in the field of protecting data privacy and security, to make recommendations to the Secretary on the development of tools and methods for the ethical and scientific uses for, and communication of, postmarketing data specified under subparagraph (C), including recommendations on the development of effective research methods for the study of drug safety questions.

"(C) ESTABLISHMENT OF THE POSTMARKET RISK IDENTIFICATION AND ANALYSIS SYSTEM.—

"(i) IN GENERAL.—The Secretary shall, not later than 1 year after the development of the risk identification and analysis methods under subparagraph (B), establish and maintain procedures—

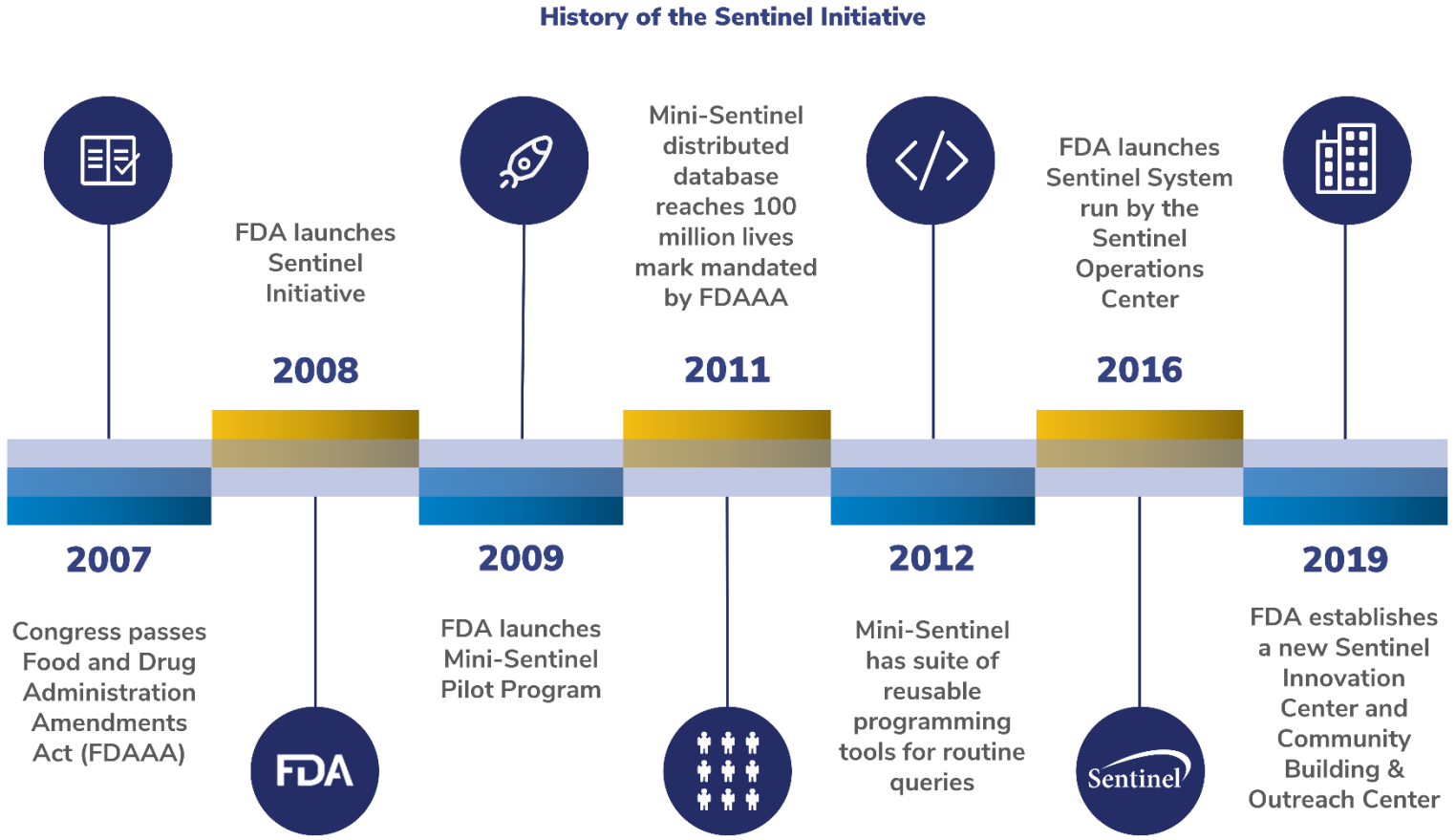
https://www.govinfo.gov/content/pkg/PLAW-110publ85/pdf/PLAW-110publ85.pdf

Establishment of a
postmarket risk identification and analysis system

https://www.govinfo.gov/content/pkg/PLAW-110publ85/pdf/PLAW-110publ85.pdf

FDA's Sentinel System

- 2007 FDA Amendments Act mandates FDA to establish ***active surveillance system*** for monitoring drugs using electronic healthcare data
- Through the Sentinel Initiative, FDA aims to assess the post-marketing safety of approved medical products



Sentinel Distributed Database (SDD)

1. Aetna, a CVS Health company
2. Carelon Research/Elevance Health
3. Duke University School of Medicine: Department of Population Health Sciences (Medicare Fee-for-Service and Medicaid data)
4. HealthPartners Institute
5. Humana, Inc.
6. Kaiser Permanente Colorado Institute for Health Research
7. Kaiser Permanente Hawai'i, Center for Integrated Health Care Research
8. Kaiser Foundation Health Plan of the Mid-Atlantic States, Inc.
9. Kaiser Permanente Northwest Center for Health Research
10. Kaiser Permanente Washington Health Research Institute
11. Marshfield Clinic Research Institute
12. Optum
13. Vanderbilt University Medical Center, Department of Health Policy (Tennessee Medicaid data)

Recognizing the Need to Harness Alternative Data Sources and Methods

FDA Budget Matters: A Cross-Cutting Data Enterprise for Real World Evidence

Share | Tweet | Linkedin | Email | Print

June 10, 2018

By: Scott Gottlieb, M.D.

Over time, as our experience with new medical products expands, our knowledge about how best to maximize their benefits and minimize any potential risks, sharpens with each data point we gather. Every clinical use of a product produces data that can help better inform us about its safety and efficacy.

The FDA is committed to developing new tools to help us access and use data collected from all sources. This includes ways to expand our methodological repertoire to build on our understanding of medical products throughout their lifecycle, in the post market. We don't limit our knowledge to pre-market information, traditional de novo post-market studies, and passive reporting. Newer methodologies enable us to collect data

FDA Commissioner Scott Gottlieb, MD

npj | Digital Medicine

www.nature.com/npjdigitalmed

PERSPECTIVE OPEN

Check for updates

Broadening the reach of the FDA Sentinel system: A roadmap for integrating electronic health record data in a causal analysis framework

Rishi J. Desai[1✉], Michael E. Matheny[2], Kevin Johnson[2], Keith Marsolo[3], Lesley H. Curtis[3], Jennifer C. Nelson[4], Patrick J. Heagerty[5], Judith Maro[6], Jeffery Brown[6], Sengwee Toh[6], Michael Nguyen[7], Robert Ball[7], Gerald Dal Pan[7], Shirley V. Wang[1], Joshua J. Gagne[1,8] and Sebastian Schneeweiss[1]

The Sentinel System is a major component of the United States Food and Drug Administration's (FDA) approach to active medical product safety surveillance. While Sentinel has historically relied on large quantities of health insurance claims data, leveraging longitudinal electronic health records (EHRs) that contain more detailed clinical information, as structured and unstructured features, may address some of the current gaps in capabilities. We identify key challenges when using EHR data to investigate medical product safety in a scalable and accelerated way, outline potential solutions, and describe the Sentinel Innovation Center's initiatives to put solutions into practice by expanding and strengthening the existing system with a query-ready, large-scale data infrastructure of linked EHR and claims data. We describe our initiatives in four strategic priority areas: (1) data infrastructure, (2) feature engineering, (3) causal inference, and (4) detection analytics, with the goal of incorporating emerging data science innovations to maximize the utility of EHR data for medical product safety surveillance.

npj Digital Medicine (2021)4:170 ; https://doi.org/10.1038/s41746-021-00542-0

Methodological Focal Points to Advance Causal Inference in Sentinel

Design Layer

Achieve causal study design

Considering:
- Study question
- Design choice
- Bias reduction

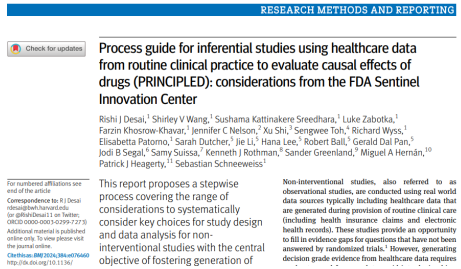
PHARMACOEPIDEMIOLOGY AND DRUG SAFETY 2012; **21**(S1): 32–40
Published online in Wiley Online Library (wileyonlinelibrary.com) **DOI**: 10.1002/pds.2316

ORIGINAL REPORT

Design considerations in an active medical product safety monitoring system

Joshua J. Gagne[1]*, Bruce Fireman[2], Patrick B. Ryan[3], Malcolm Maclure[4,5], Tobias Gerhard[6], Sengwee Toh[7], Jeremy A. Rassen[1], Jennifer C. Nelson[8,9] and Sebastian Schneeweiss[1]

[1]*Division of Pharmacoepidemiology and Pharmacoeconomics, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA*
[2]*Division of Research, Kaiser Permanente Northern California, Oakland, CA, USA*
[3]*Johnson & Johnson, Titusville, NJ, USA*
[4]*Department of Anesthesiology, Pharmacology and Therapeutics, University of British Columbia, Vancouver, BC, Canada*
[5]*Pharmaceutical Services Division, BC Ministry of Health Services, Victoria, BC, Canada*
[6]*Ernest Mario School of Pharmacy, Rutgers, The State University of New Jersey, New Brunswick, NJ, USA*
[7]*Department of Population Medicine, Harvard Medical School and Harvard Pilgrim Healthcare Institute, Boston, MA, USA*
[8]*Biostatistics Unit, Group Health Research Institute, Seattle, WA, USA*
[9]*Department of Biostatistics, University of Washington, Seattle, WA, USA*



RESEARCH METHODS AND REPORTING

Process guide for inferential studies using healthcare data from routine clinical practice to evaluate causal effects of drugs (PRINCIPLED): considerations from the FDA Sentinel Innovation Center

Rishi J Desai,[1] Shirley V Wang,[1] Sushama Kattinakere Sreedhara,[1] Luke Zabotka,[1] Farzin Khosrow-Khavar,[1] Jennifer C Nelson,[2] Xu Shi,[3] Sengwee Toh,[4] Richard Wyss,[1] Elisabetta Patorno,[1] Sarah Dutcher,[3] Jie Li,[3] Hana Lee,[5] Robert Ball,[5] Gerald Dal Pan,[5] Jodi B Segal,[6] Samy Suissa,[7] Kenneth J Rothman,[8] Sander Greenland,[9] Miguel A Hernán,[10] Patrick J Heagerty,[11] Sebastian Schneeweiss[1]

This report proposes a stepwise process covering the range of considerations to systematically consider key choices for study design and data analysis for non-interventional studies with the central objective of fostering generation of

Non-interventional studies, also referred to as observational studies, are conducted using real world data sources typically including healthcare data that are generated during provision of routine clinical care (including health insurance claims and electronic health records). These studies provide an opportunity to fill in evidence gaps for questions that have not been answered by randomized trials.[1] However, generating decision grade evidence from healthcare data requires

Measures Layer

Achieve fit-for-purpose measurement

Considering:
- sensitivity
- specificity,
- completeness
- mean sqr diff

Filling Rx
Prescribing Rx, self-report, infusers, pill caps, UDI from OR notes

EXPOSURE

Dx, Px codes
Labs, imaging, digital health dev, physician notes, patient reports

OUTCOME

Dx, Px, Rx codes
Labs, stage, imaging reports, BMI, genomics, physician notes

CONFOUNDERS

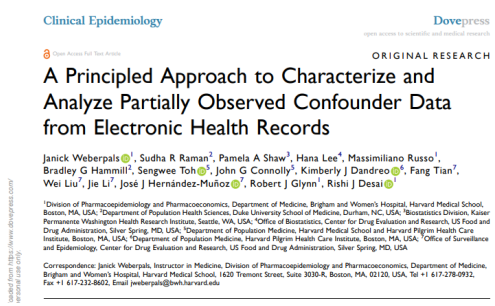
Dx, Px, Rx codes
Monitors, physician notes, biomarker, omics, behavior, socio-econ

TARGET POP^N

Analytics Layer

Achieve causal analysis

Considering:
- Confounding adjustment
- Missing data
- Robustness evaluations



Clinical Epidemiology — Dovepress

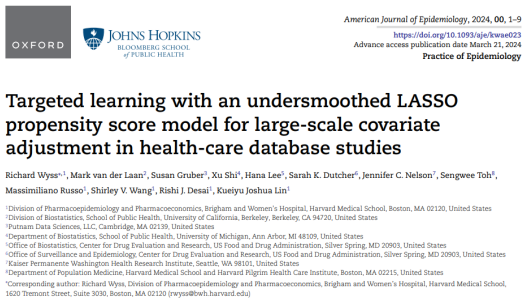
ORIGINAL RESEARCH

A Principled Approach to Characterize and Analyze Partially Observed Confounder Data from Electronic Health Records

Janick Weberpals[1], Sudha R Raman[2], Pamela A Shaw[3], Hana Lee[4], Massimiliano Russo[1], Bradley G Hammill[4], Sengwee Toh[5], John G Connolly[3], Kimberly J Dandreo[6], Fang Tian[7], Wei Liu[7], Jie Li[7], José J Hernández-Muñoz[7], Robert J Glynn[1], Rishi J Desai[1]

[1]Division of Pharmacoepidemiology and Pharmacoeconomics, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA; [2]Department of Population Health Sciences, Duke University School of Medicine, Durham, NC, USA; [3]Biostatistics Division, Kaiser Permanente Washington Health Research Institute, Seattle, WA, USA; [4]Office of Biostatistics, Center for Drug Evaluation and Research, US Food and Drug Administration, Silver Spring, MD, USA; [5]Department of Population Medicine, Harvard Medical School and Harvard Pilgrim Health Care Institute, Boston, MA, USA; [6]Department of Population Medicine, Harvard Pilgrim Health Care Institute, Boston, MA, USA; [7]Office of Surveillance and Epidemiology, Center for Drug Evaluation and Research, US Food and Drug Administration, Silver Spring, MD, USA

Correspondence: Janick Weberpals, Instructor in Medicine, Division of Pharmacoepidemiology and Pharmacoeconomics, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, 1620 Tremont Street, Suite 3030-R, Boston, MA, 02120, USA, Tel +1 617-278-0932, Fax +1 617-232-8602, Email jweberpals@bwh.harvard.edu



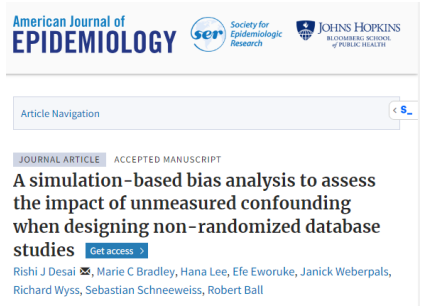
OXFORD — JOHNS HOPKINS BLOOMBERG SCHOOL of PUBLIC HEALTH

American Journal of Epidemiology, 2024, 00, 1–9
https://doi.org/10.1093/aje/kwae023
Advance access publication date March 21, 2024
Practice of Epidemiology

Targeted learning with an undersmoothed LASSO propensity score model for large-scale covariate adjustment in health-care database studies

Richard Wyss[a,1], Mark van der Laan[2], Susan Gruber[3], Xu Shi[4], Hana Lee[5], Sarah K. Dutcher[6], Jennifer C. Nelson[7], Sengwee Toh[8], Massimiliano Russo[1], Shirley V. Wang[1], Rishi J. Desai[1], Kueiyu Joshua Lin[1]

[1]Division of Pharmacoepidemiology and Pharmacoeconomics, Brigham and Women's Hospital, Harvard Medical School, Boston, MA 02120, United States
[2]Division of Biostatistics, School of Public Health, University of California, Berkeley, Berkeley, CA 94720, United States
[3]Putnam Data Sciences, LLC, Cambridge, MA 02139, United States
[4]Department of Biostatistics, School of Public Health, University of Michigan, Ann Arbor, MI 48109, United States
[5]Office of Biostatistics, Center for Drug Evaluation and Research, US Food and Drug Administration, Silver Spring, MD 20903, United States
[6]Office of Surveillance and Epidemiology, Center for Drug Evaluation and Research, US Food and Drug Administration, Silver Spring, MD 20903, United States
[7]Kaiser Permanente Washington Health Research Institute, Seattle, WA 98101, United States
[8]Department of Population Medicine, Harvard Medical School and Harvard Pilgrim Health Care Institute, Boston, MA 02215, United States
[a]Corresponding author: Richard Wyss, Division of Pharmacoepidemiology and Pharmacoeconomics, Brigham and Women's Hospital, Harvard Medical School, 1620 Tremont Street, Suite 3030, Boston, MA 02120 (rwyss@bwh.harvard.edu)



American Journal of EPIDEMIOLOGY — Society for Epidemiologic Research — JOHNS HOPKINS BLOOMBERG SCHOOL of PUBLIC HEALTH

Article Navigation

JOURNAL ARTICLE ACCEPTED MANUSCRIPT

A simulation-based bias analysis to assess the impact of unmeasured confounding when designing non-randomized database studies Get access

Rishi J Desai, Marie C Bradley, Hana Lee, Efe Eworuke, Janick Weberpals, Richard Wyss, Sebastian Schneeweiss, Robert Ball

Methodological Focal Points to Advance Causal Inference in Sentinel

Design Layer

Achieve causal study design

Considering:
- Study question
- Design choice
- Bias reduction

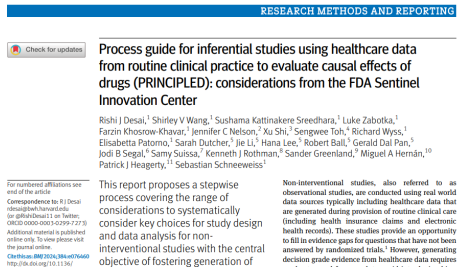
PHARMACOEPIDEMIOLOGY AND DRUG SAFETY 2012; **21**(S1): 32–40
Published online in Wiley Online Library (wileyonlinelibrary.com) **DOI**: 10.1002/pds.2316

ORIGINAL REPORT

Design considerations in an active medical product safety monitoring system

Joshua J. Gagne[1]*, Bruce Fireman[2], Patrick B. Ryan[3], Malcolm Maclure[4,5], Tobias Gerhard[6], Sengwee Toh[7], Jeremy A. Rassen[1], Jennifer C. Nelson[8,9] and Sebastian Schneeweiss[1]

[1]*Division of Pharmacoepidemiology and Pharmacoeconomics, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA*
[2]*Division of Research, Kaiser Permanente Northern California, Oakland, CA, USA*
[3]*Johnson & Johnson, Titusville, NJ, USA*
[4]*Department of Anesthesiology, Pharmacology and Therapeutics, University of British Columbia, Vancouver, BC, Canada*
[5]*Pharmaceutical Services Division, BC Ministry of Health Services, Victoria, BC, Canada*
[6]*Ernest Mario School of Pharmacy, Rutgers, The State University of New Jersey, New Brunswick, NJ, USA*
[7]*Department of Population Medicine, Harvard Medical School and Harvard Pilgrim Healthcare Institute, Boston, MA, USA*
[8]*Biostatistics Unit, Group Health Research Institute, Seattle, WA, USA*
[9]*Department of Biostatistics, University of Washington, Seattle, WA, USA*



RESEARCH METHODS AND REPORTING

Process guide for inferential studies using healthcare data from routine clinical practice to evaluate causal effects of drugs (PRINCIPLED): considerations from the FDA Sentinel Innovation Center

Rishi J Desai,[1] Shirley V Wang,[1] Sushama Kattinakere Sreedhara,[1] Luke Zabotka,[1] Farzin Khosrow-Khavar,[1] Jennifer C Nelson,[2] Xu Shi,[3] Sengwee Toh,[4] Richard Wyss,[1] Elisabetta Patorno,[1] Sarah Dutcher,[5] Jie Li,[5] Hana Lee,[5] Robert Ball,[5] Gerald Dal Pan,[5] Jodi B Segal,[6] Samy Suissa,[7] Kenneth J Rothman,[8] Sander Greenland,[9] Miguel A Hernán,[10] Patrick J Heagerty,[11] Sebastian Schneeweiss[1]

This report proposes a stepwise process covering the range of considerations to systematically consider key choices for study design and data analysis for non-interventional studies with the central objective of fostering generation of

Non-interventional studies, also referred to as observational studies, are conducted using real world data sources typically including healthcare data that are generated during provision of routine clinical care (including health insurance claims and electronic health records). These studies provide an opportunity to fill in evidence gaps for questions that have not been answered by randomized trials.[1] However, generating decision grade evidence from healthcare data requires

Measures Layer

Achieve fit-for-purpose measurement

Considering:
- sensitivity
- specificity,
- completeness
- mean sqr diff

Filling Rx
Prescribing Rx, self-report, infusers, pill caps, UDI from OR notes

Dx, Px codes
Labs, imaging, digital health dev, physician notes, patient reports

Dx, Px, Rx codes
Labs, stage, imaging reports, BMI, genomics, physician notes

Dx, Px, Rx codes
Monitors, physician notes, biomarker, omics, behavior, socio-econ

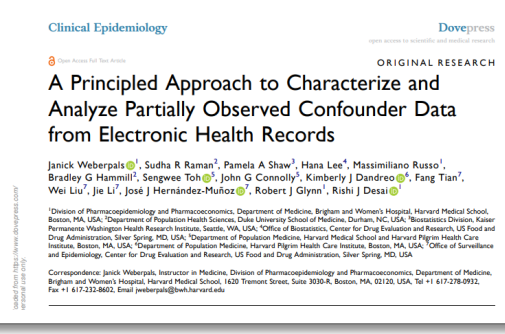
Natural Language Processing (NLP)

EXPOSURE *OUTCOME* *CONFOUNDERS* *TARGET POP^N*

Analytics Layer

Achieve causal analysis

Considering:
- Confounding adjustment
- Missing data
- Robustness evaluations



Clinical Epidemiology — Dovepress

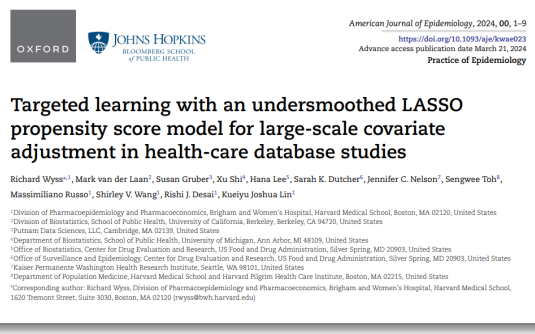
ORIGINAL RESEARCH

A Principled Approach to Characterize and Analyze Partially Observed Confounder Data from Electronic Health Records

Janick Weberpals[1], Sudha R Raman[2], Pamela A Shaw[3], Hana Lee[4], Massimiliano Russo[1], Bradley G Hammill[2], Sengwee Toh[5], John G Connolly[2], Kimberly J Dandreo[6], Fang Tian[7], Wei Liu[7], Jie Li[7], José J Hernández-Muñoz[7], Robert J Glynn[1], Rishi J Desai[1]

[1]Division of Pharmacoepidemiology and Pharmacoeconomics, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA; [2]Department of Population Health Sciences, Duke University School of Medicine, Durham, NC, USA; [3]Biostatistics Division, Kaiser Permanente Washington Health Research Institute, Seattle, WA, USA; [4]Office of Biostatistics, Center for Drug Evaluation and Research, US Food and Drug Administration, Silver Spring, MD, USA; [5]Department of Population Medicine, Harvard Medical School and Harvard Pilgrim Health Care Institute, Boston, MA, USA; [6]Department of Population Medicine, Harvard Pilgrim Health Care Institute, Boston, MA, USA; [7]Office of Surveillance and Epidemiology, Center for Drug Evaluation and Research, US Food and Drug Administration, Silver Spring, MD, USA

Correspondence: Janick Weberpals, Instructor in Medicine, Division of Pharmacoepidemiology and Pharmacoeconomics, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, 1620 Tremont Street, Suite 3030-R, Boston, MA, 02120, USA, Tel +1 617-278-0932, Fax +1 617-232-8602, Email jweberpals@bwh.harvard.edu



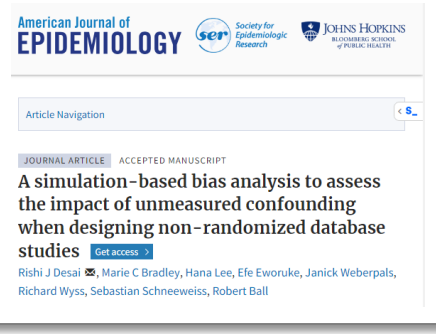
OXFORD — JOHNS HOPKINS BLOOMBERG SCHOOL of PUBLIC HEALTH

American Journal of Epidemiology, 2024, 00, 1–9
https://doi.org/10.1093/aje/kwae023
Advance access publication date March 21, 2024
Practice of Epidemiology

Targeted learning with an undersmoothed LASSO propensity score model for large-scale covariate adjustment in health-care database studies

Richard Wyss[*,1], Mark van der Laan[2], Susan Gruber[3], Xu Shi[4], Hana Lee[5], Sarah K. Dutcher[6], Jennifer C. Nelson[7], Sengwee Toh[8], Massimiliano Russo[1], Shirley V. Wang[1], Rishi J. Desai[1], Kueiyu Joshua Lin[1]

[1]Division of Pharmacoepidemiology and Pharmacoeconomics, Brigham and Women's Hospital, Harvard Medical School, Boston, MA 02120, United States
[2]Division of Biostatistics, School of Public Health, University of California, Berkeley, Berkeley, CA 94720, United States
[3]Putnam Data Sciences, LLC, Cambridge, MA 02139, United States
[4]Department of Biostatistics, School of Public Health, University of Michigan, Ann Arbor, MI 48109, United States
[5]Office of Biostatistics, Center for Drug Evaluation and Research, US Food and Drug Administration, Silver Spring, MD 20903, United States
[6]Office of Surveillance and Epidemiology, Center for Drug Evaluation and Research, US Food and Drug Administration, Silver Spring, MD 20903, United States
[7]Kaiser Permanente Washington Health Research Institute, Seattle, WA 98101, United States
[8]Department of Population Medicine, Harvard Medical School and Harvard Pilgrim Health Care Institute, Boston, MA 02215, United States
*Corresponding author: Richard Wyss, Division of Pharmacoepidemiology and Pharmacoeconomics, Brigham and Women's Hospital, Harvard Medical School, 1620 Tremont Street, Suite 3030, Boston, MA 02120 (rwyss@bwh.harvard.edu)



American Journal of EPIDEMIOLOGY — Society for Epidemiologic Research — JOHNS HOPKINS BLOOMBERG SCHOOL of PUBLIC HEALTH

Article Navigation

JOURNAL ARTICLE ACCEPTED MANUSCRIPT

A simulation-based bias analysis to assess the impact of unmeasured confounding when designing non-randomized database studies Get access

Rishi J Desai, Marie C Bradley, Hana Lee, Efe Eworuke, Janick Weberpals, Richard Wyss, Sebastian Schneeweiss, Robert Ball

Leveraging NLP in Sentinel

American Journal of Epidemiology
© The Author(s) 2022. Published by Oxford University Press on behalf of the Johns Hopkins Bloomberg School of Public Health. This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (https://creativecommons.org/licenses/by-nc/4.0), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journalpermissions@oup.com.

Vol. 192, No. 2
https://doi.org/10.1093/aje/kwac182
Advance Access publication: November 4, 2022

Practice of Epidemiology

Improving Methods of Identifying Anaphylaxis for Medical Product Safety Surveillance Using Natural Language Processing and Machine Learning

David S. Carrell*, Susan Gruber, James S. Floyd, Maralyssa A. Bann, Kara L. Cushing-Haugen, Ron L. Johnson, Vina Graham, David J. Cronkite, Brian L. Hazlehurst, Andrew H. Felcher, Cosmin A. Bejan, Adee Kennedy, Mayura U. Shinde, Sara Karami, Yong Ma, Danijela Stojanovic, Yueqin Zhao, Robert Ball, and Jennifer C. Nelson

* Correspondence to Dr. David Carrell, Kaiser Permanente Washington Health Research Institute, 1730 Minor Avenue, Suite 1600, Seattle, WA 98101 (e-mail: david.s.carrell@kp.org).

Initially submitted August 11, 2021; accepted for publication October 11, 2022.

Scalable Incident Detection via Natural Language Processing and Probabilistic Language Models

Colin G. Walsh, Drew Wilimitis, Qingxia Chen, Aileen Wright, Jhansi Kolli, Katelyn Robinson, Michael A. Ripperger, Kevin B. Johnson, David Carrell, Rishi J. Desai, Andrew Mosholder, Sai Dharmarajan, Sruthi Adimadhyam, Daniel Fabbri, Danijela Stojanovic, Michael E. Matheny, Cosmin A. Bejan

doi: https://doi.org/10.1101/2023.11.30.23299249

[Submitted on 17 May 2024]

High-dimensional multiple imputation (HDMI) for partially observed confounders including natural language processing-derived auxiliary covariates

Janick Weberpals, Pamela A. Shaw, Kueiyu Joshua Lin, Richard Wyss, Joseph M Plasek, Li Zhou, Kerry Ngan, Thomas DeRamus, Sudha R. Raman, Bradley G. Hammill, Hana Lee, Sengwee Toh, John G. Connolly, Kimberly J. Dandreo, Fang Tian, Wei Liu, Jie Li, José J. Hernández-Muñoz, Sebastian Schneeweiss, Rishi J. Desai

Journal of the American Medical Informatics Association, 2023, 1–9
https://doi.org/10.1093/jamia/ocad241
Research and Applications

AMIA INFORMATICS PROFESSIONALS. LEADING THE WAY.
OXFORD

Research and Applications

Data-driven automated classification algorithms for acute health conditions: applying PheNorm to COVID-19 disease

Joshua C. Smith, PhD[1],*, Brian D. Williamson, PhD[2], David J. Cronkite, MS[2], Daniel Park, BS[1], Jill M. Whitaker, MSN[1], Michael F. McLemore, BSN[1], Joshua T. Osmanski, MS[1], Robert Winter, BA[1], Arvind Ramaprasan, MS[2], Ann Kelley, MHA[2], Mary Shea, MA[2], Saranrat Wittayanukorn, PhD[3], Danijela Stojanovic, PharmD, PhD[3], Yueqin Zhao, PhD[3], Sengwee Toh, ScD[4], Kevin B. Johnson, MD, MS[5], David M. Aronoff, MD[6], David S. Carrell, PhD[2]

[1]Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN 37203, United States, [2]Kaiser Permanente Washington Health Research Institute, Seattle, WA 98101, United States, [3]Center for Drug Evaluation and Research, US Food and Drug Administration, Silver Spring, MD 20903, United States, [4]Harvard Pilgrim Health Care Institute, Boston, MA 02215, United States, [5]Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania, Philadelphia, PA 19104, United States, [6]Department of Medicine, Indiana University School of Medicine, Indianapolis, IN 46202, United States

*Corresponding author: Joshua C. Smith, PhD, Department of Biomedical Informatics, Vanderbilt University Medical Center, 2525 West End Avenue, Suite No. 1400, Nashville, TN 37203 (joshua.smith@vumc.org)

Methodological Considerations for NLP in Pharmacoepidemiology Studies

Elise Berliner, PhD

Oracle Life Sciences

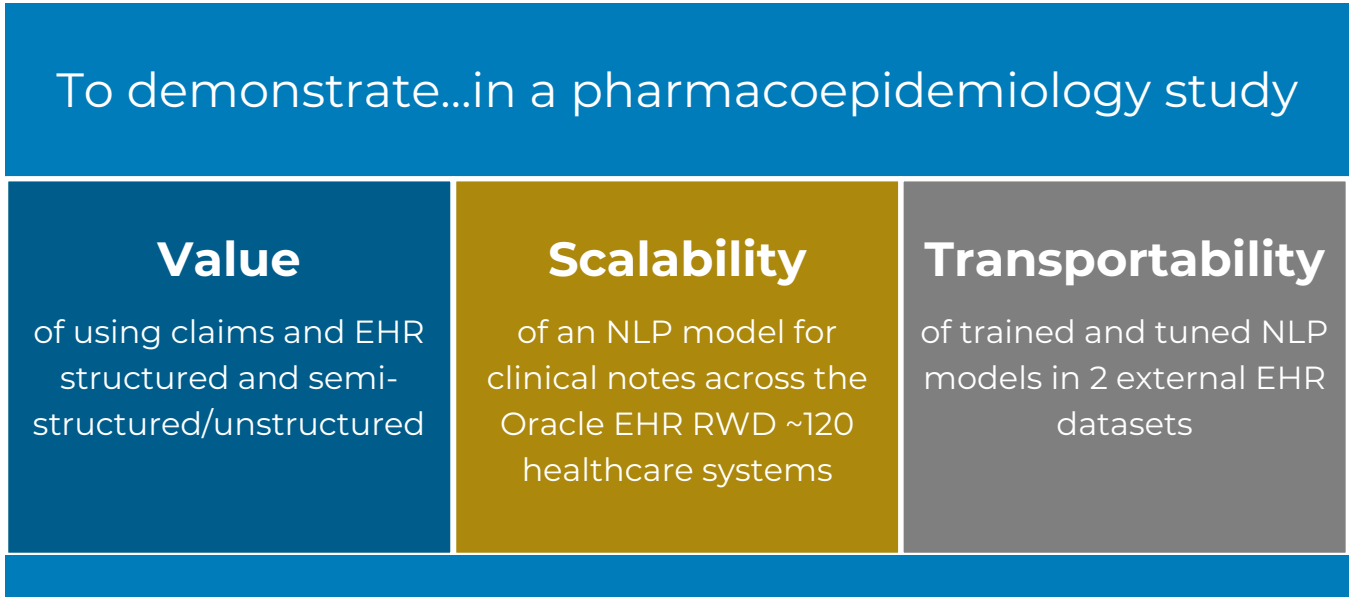
Content

1. **Introduce MOSAIC-NLP**
2. MOSAIC-NLP Study Design Considerations
3. Entity Extraction

MOSAIC-NLP

Multi-source Observational Safety study for Advanced Information Classification using NLP

Use of Natural Language Processing in a Pharmacoepidemiology Study: The Examination of Neuropsychiatric Events and Incident Use of Montelukast Among Patients with Asthma



Project Team

FDA

- **Sarah Dutcher**, Epidemiologist
- **Jummai Apata,** Epidemiologist
- **Robert Lim,** Medical Officer
- **Jie (Jenni) Li,** Epidemiologist
- **Jamal Jones,** Epidemiologist
- **Yong Ma,** Biostatistician
- **Tiffany Austin**, Project Manager

Sentinel Operations Center/Harvard

- **Meighan Driscoll**, Program Manager
- **Kimberly Gegear**, Project Manager
- **Sam McGown**, Research Assistant
- **Darren Toh**, Co-investigator, Pharmacoepidemiologist
- **Jenna Wong**, Pharmacoepidemiologist

Mass General Brigham

- **Richard Wyss,** Co-Investigator, Epidemiologist
- **Jie Yang**, Principal Investigator
- **Rishi Desai**, Operations Chief
- **Josh Lin**, Epidemiologist

Oracle Life Sciences/ Oracle Health

- **Elise Berliner**, Principal Investigator
- **Dena Jaffe**, Principal Investigator, Epidemiologist
- **Jenny Cai**, Project Manager
- **Sonam Lama**, Project Manager
- **Nathan Vavroch**, Data Strategist
- **Mike Jones**, Data Strategist
- **Vineela Kommuri**, Senior Data Engineer
- **Sravan Kumar Burla**, Software Engineer
- **Bridget Balkaran**, Lead Biostatistician
- **Austin Yue**, Biostatistician
- **Kyla Finlayson,** Biostatistician
- **Stacey Purinton**, Data Manager
- **Rob Taylor**, Data Manager
- **Eliza Celenti**, Medical Writer

National Jewish Health

- **Michael Wechsler,** Pulmonologist
- **David Beuther,** Pulmonologist
- **Lior Seluk**, Pulmonologist
- **Pearlanne Zelarney,** Research Informatics
- **Alicia Mitchell,** Developer
- **Sarah Rhoads**, Pulmonologist

Children's Hospital of Orange County

- **Louis Ehwerhemuepha,** Clinical Data Scientist
- **Hoang Nguyen,** Psychiatrist
- **Michael Chu,** Psychiatrist
- **Heather Huszti,** Psychologist
- **Olga Guijon,** Pediatrician and Asthma specialist

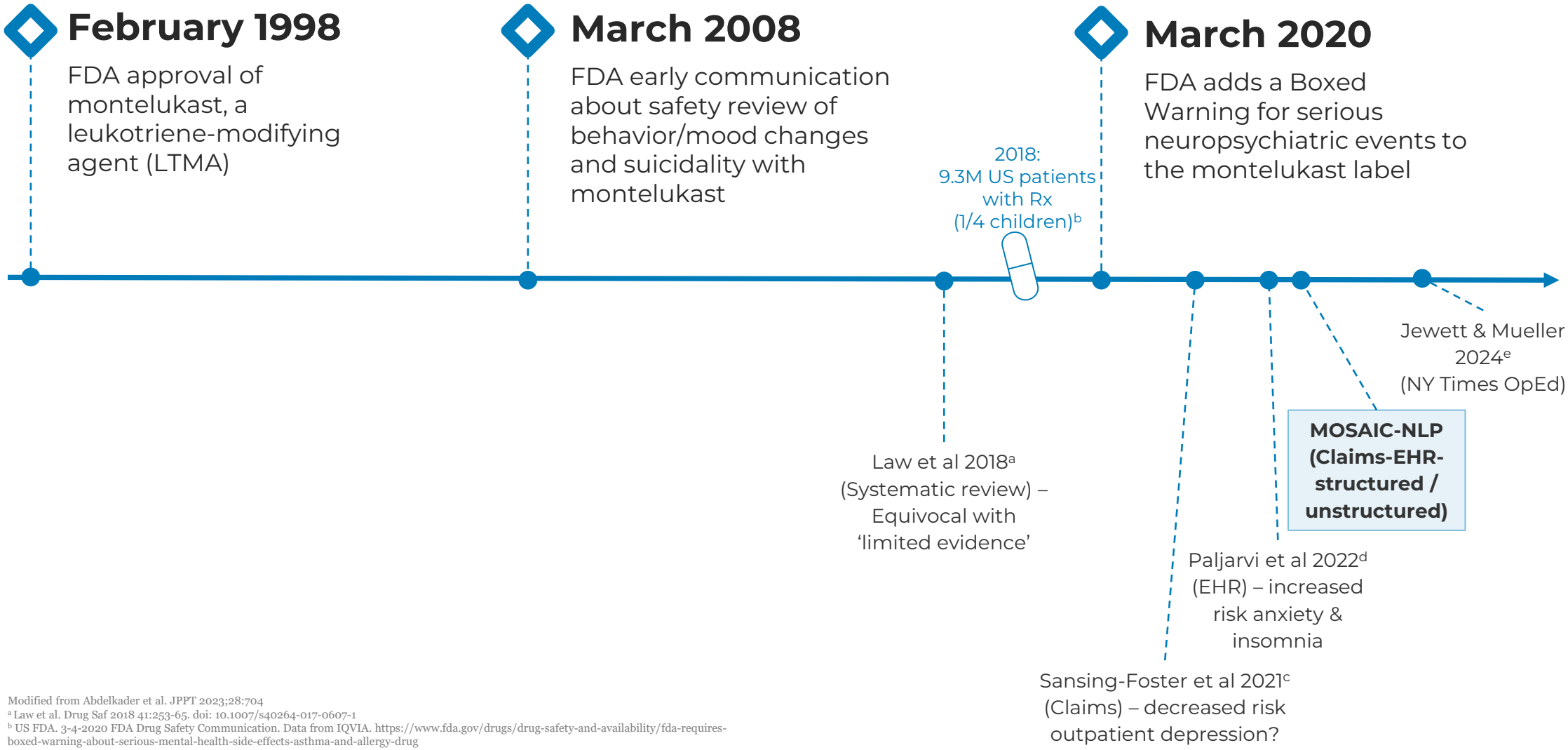
John Snow Labs

- **David Talby**, CTO
- **Ace Vo**, Project Manager
- **Hasham UI Haq**, Lead Senior NLP Data Scientist
- **Veysel Kocaman**, Data Scientist
- **Gursev Pirge**, Data Scientist
- **Ahmet Emin Tek**, Data Scientist
- **Andrei Marian Feier**, Clinical Annotation Lead
- **Denisa Popa**, Data Annotator
- **Aleksei Zhakarov**, Annotator
- **Jay Gil**, Annotator
- **Zhenya Nargizyan**, Annotator
- **Jiri Dobles**, Project Manager

Kaiser Permanente Washington Health Research Institute

- **David Carrell,** NLP Expert Consultant

Use Case - Montelukast



Modified from Abdelkader et al. JPPT 2023;28:704

[a] Law et al. Drug Saf 2018 41:253-65. doi: 10.1007/s40264-017-0607-1

[b] US FDA. 3-4-2020 FDA Drug Safety Communication. Data from IQVIA. https://www.fda.gov/drugs/drug-safety-and-availability/fda-requires-boxed-warning-about-serious-mental-health-side-effects-asthma-and-allergy-drug

[c] Sansing-Foster et al. J Allergy Clin Immunol Pract 2021;9:385-93 doi: 10.1016/j.jaip.2020.07.052

[d] Paljarvi et al. JAMA Netw Open 2022;5:e2213643. doi: 10.1001/jamanetworkopen.2022.13643

[e] Jewett and Mueller. NY Times Jan 9, 2024 https://www.nytimes.com/2024/01/09/health/fda-singulair-asthma-drug-warning.html

Further Research is Required

"All epidemiology is an exercise in the rational use of limited resources" (Matthew Fox, 2023)

- Incomplete outcome detection
  - Structured data typically report more severe outcomes and not mild symptoms
  - Identify only outcomes used for billing (claims)

- Timing of study
  - Coding changes for self harm or suicidal behavior from ICD-9 to ICD-10
  - FDA communications regarding this risk beginning in 2008 may have resulted in higher-risk patients avoiding montelukast altogether or stopping montelukast upon experiencing minor neuropsychiatric symptoms, thereby reducing occurrence of serious AEs
- Incomplete confounder control:
  - Socio-economic status (higher SES -> seek care, early intervention or increased diagnosis)
  - Psychiatric history

MOSAIC-NLP

Value

of using claims and EHR structured and semi-structured/unstructured

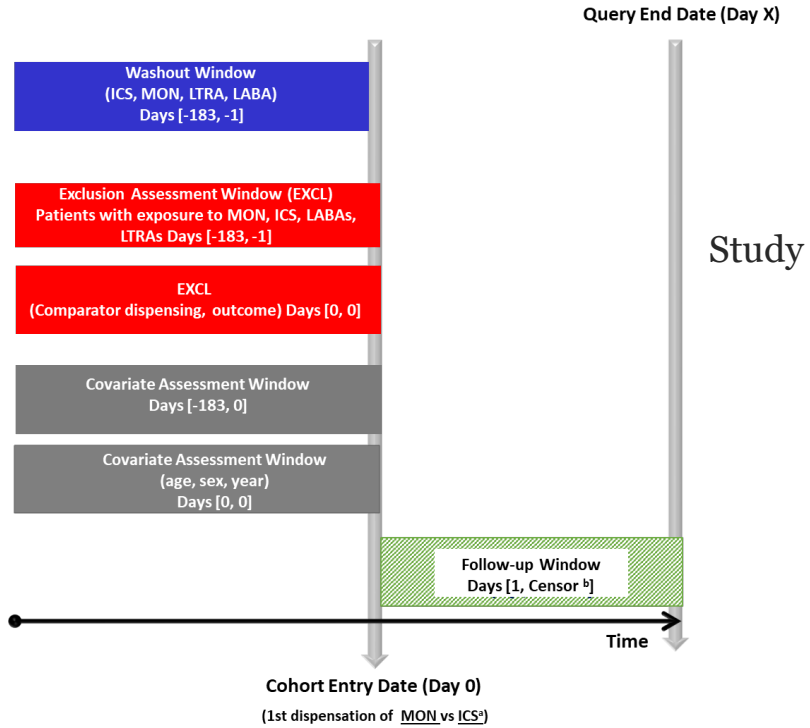
- ***Study Design:*** Retrospective cohort study
- ***Study Data:*** EHR-claims linked structured and unstructured data (2015-2022)
- ***Study Cohort:*** Patients with asthma newly initiating montelukast or inhaled corticosteroids monotherapy (comparator)
- ***Study Outcomes:*** Neuropsychiatric events

MOSAIC-NLP

Value

of using claims and EHR structured and semi-structured/unstructured

- ***Study Design:*** Retrospective cohort study
- ***Study Data:*** EHR-claims linked structured and unstructured data (2015-2022)
- ***Study Cohort:*** Patients with asthma newly initiating montelukast or inhaled corticosteroids monotherapy (comparator)
- ***Study Outcomes:*** Neuropsychiatric events



Study Design – Based on Sansing-Foster et al 2021

MOSAIC-NLP

Scalability

of an NLP model for clinical notes across the Oracle EHR RWD 120+ healthcare systems

- ***Study Cohort:*** 109,076 patients with asthma
- ***Healthcare Systems:*** 119
- ***Clinical Notes:*** 17+ million

MOSAIC-NLP

Scalability

of an NLP model for clinical notes across the Oracle EHR RWD 120+ healthcare systems

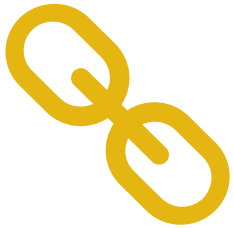
- ***Study cohort:*** 109,076 patients
- ***Healthcare systems:*** 119
- ***Clinical notes:*** 17+ million

	Oracle EHR RWD	Claims
Patients	105+ Million	200+ Million
National representation	✓	✓
Care and coverage	Pediatric hospitals Critical care hospitals Acute care hospitals Physician groups IDN	Commercial Medicare Advantage Medicaid Managed Care
Encounters and claims	125M emergency encounters 56M inpatient encounters 972M outpatient encounters	Closed medical claims Closed pharmacy claims

Content

1. Introduce MOSAIC-NLP
2. **MOSAIC-NLP Study Design Considerations**
3. Entity Extraction

Study Design Considerations for Claims-EHR Linked Study



Limitations of Linking Data

Linkable data (tokenization)

Claims limits to insured only

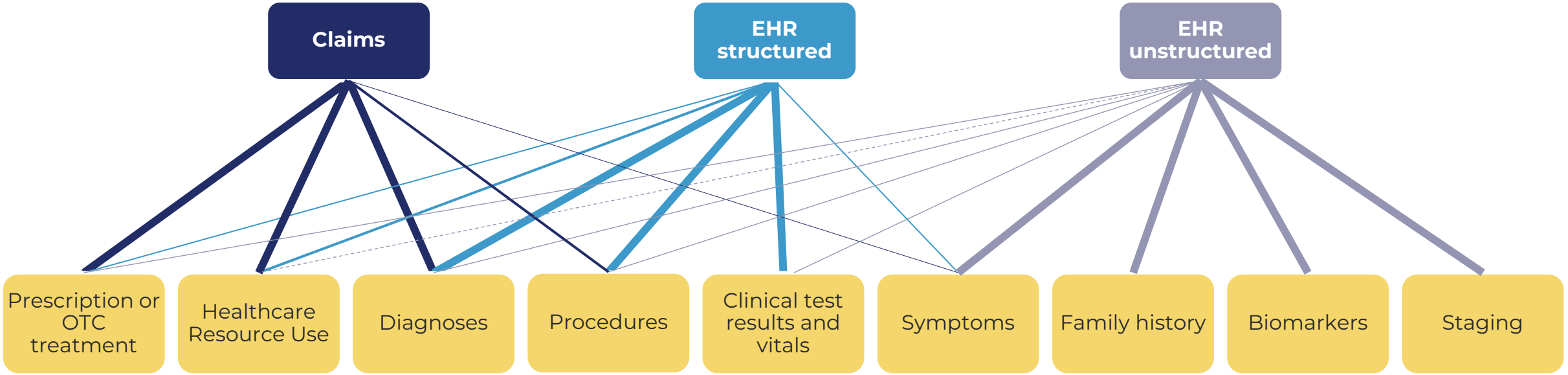
Bounded by dates of datasets

Measures

Defining patient eligibility

Disease and treatment history vs. problem list

Source of Measures



MOSAIC-NLP Study Design

Cohort	Claims + EHR structured + EHR unstructured

MOSAIC-NLP Study Design

Cohort	Claims + EHR structured + EHR unstructured

Inclusion

- Eligibility:
  - Present in both claims (medical and Rx) and EHR-structured (ever)
  - ≥1 diagnosis of asthma (claims and EHR-structured) (ever)
  - Valid text note (non-null content, non-scannded) (EHR-unstructured)
- Medication use (claims)
- Asthma diagnosis during study period (claims or EHR-structured)
- Age (claims)

Exclusion

- Prior medication use (claims)
- Concomitant medication use (claims)

MOSAIC-NLP Design of Measures

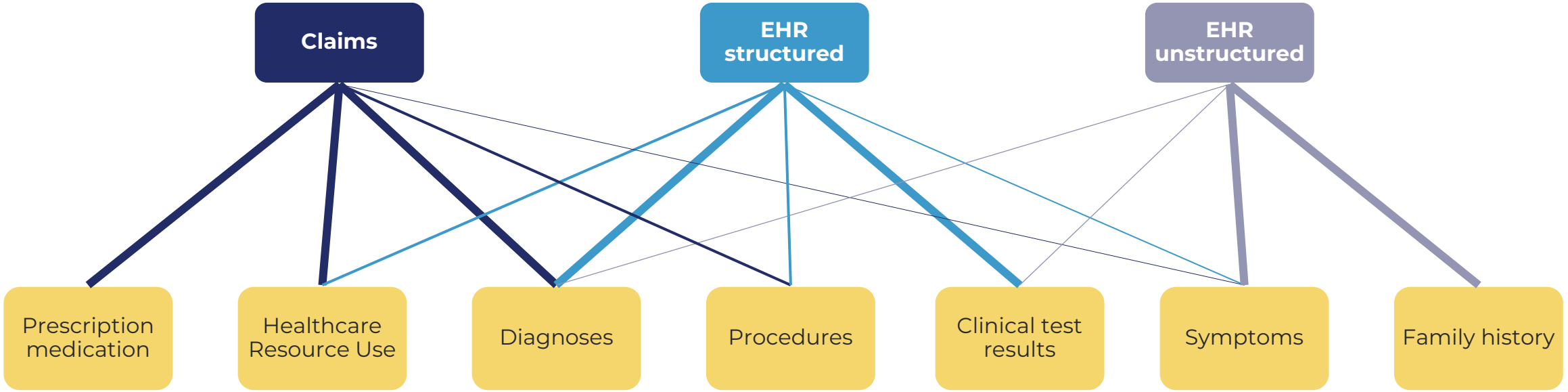
Cohort	Claims + EHR structured + EHR unstructured

Covariates	Analysis 1	Claims
	Analysis 2	Claims + EHR structured
	Analysis 3	Claims + EHR structured + EHR unstructured

MOSAIC-NLP Design of Measures

Cohort		Claims + EHR structured + EHR unstructured
Covariates	Analysis 1	Claims
	Analysis 2	Claims + EHR structured
	Analysis 3	Claims + EHR structured + EHR unstructured
Outcomes	Analysis 1	Claims
	Analysis 2	Claims + EHR structured
	Analysis 3	Claims + EHR structured + EHR unstructured

Source of Measures – MOSAIC-NLP



Outcomes: Neuropsychiatric Events

FDA's Boxed Warning

- Agitation, including aggressive behavior or hostility
- Attention problems
- Bad or vivid dreams
- Depression
- Disorientation or confusion
- Feeling anxious
- Hallucinations
- Irritability
- Memory problems
- Obsessive-compulsive symptoms
- Restlessness
- Sleepwalking
- Stuttering
- Suicidal thoughts and actions
- Tremor or shakiness
- Trouble sleeping
- Uncontrolled muscle movements

Claims/EHR Structured Data

Hospitalization/ER

OR

Diagnosis *AND/OR* Treatment of

- ❑ Depression
- ❑ Self harm
- ❑ Psychotic disorder
- ❑ Mood disorder
- ❑ Anxiety disorder
- ❑ OCD
- ❑ Manic or bipolar disorder
- ❑ Personality disorder
- ❑ Hyperactivity or aggressive behavior or harm

Diagnosis OR Treatment of Sleep Disorder

- ❑ Insomnia
- ❑ Hypersomnia
- ❑ Circadian rhythm disorder
- ❑ Parasomnia
- ❑ Movement disorder
- ❑ Other undefined sleep disorder

Outcomes: Neuropsychiatric Events

FDA's Boxed Warning

- Agitation, including aggressive behavior or hostility
- Attention problems
- Bad or vivid dreams
- Depression
- Disorientation or confusion
- Feeling anxious
- Hallucinations
- Irritability
- Memory problems
- Obsessive-compulsive symptoms
- Restlessness
- Sleepwalking
- Stuttering
- Suicidal thoughts and actions
- Tremor or shakiness
- Trouble sleeping
- Uncontrolled muscle movements

Claims/EHR Structured Data

Hospitalization/ER

OR

Diagnosis *AND/OR* Treatment of

- ❑ Depression
- ❑ Self harm
- ❑ Psychotic disorder
- ❑ Mood disorder
- ❑ Anxiety disorder
- ❑ OCD
- ❑ Manic or bipolar disorder
- ❑ Personality disorder
- ❑ Hyperactivity or aggressive behavior or harm

Diagnosis OR Treatment of Sleep Disorder

- ❑ Insomnia
- ❑ Hypersomnia
- ❑ Circadian rhythm disorder
- ❑ Parasomnia
- ❑ Movement disorder
- ❑ Other undefined sleep disorder

Unstructured Data

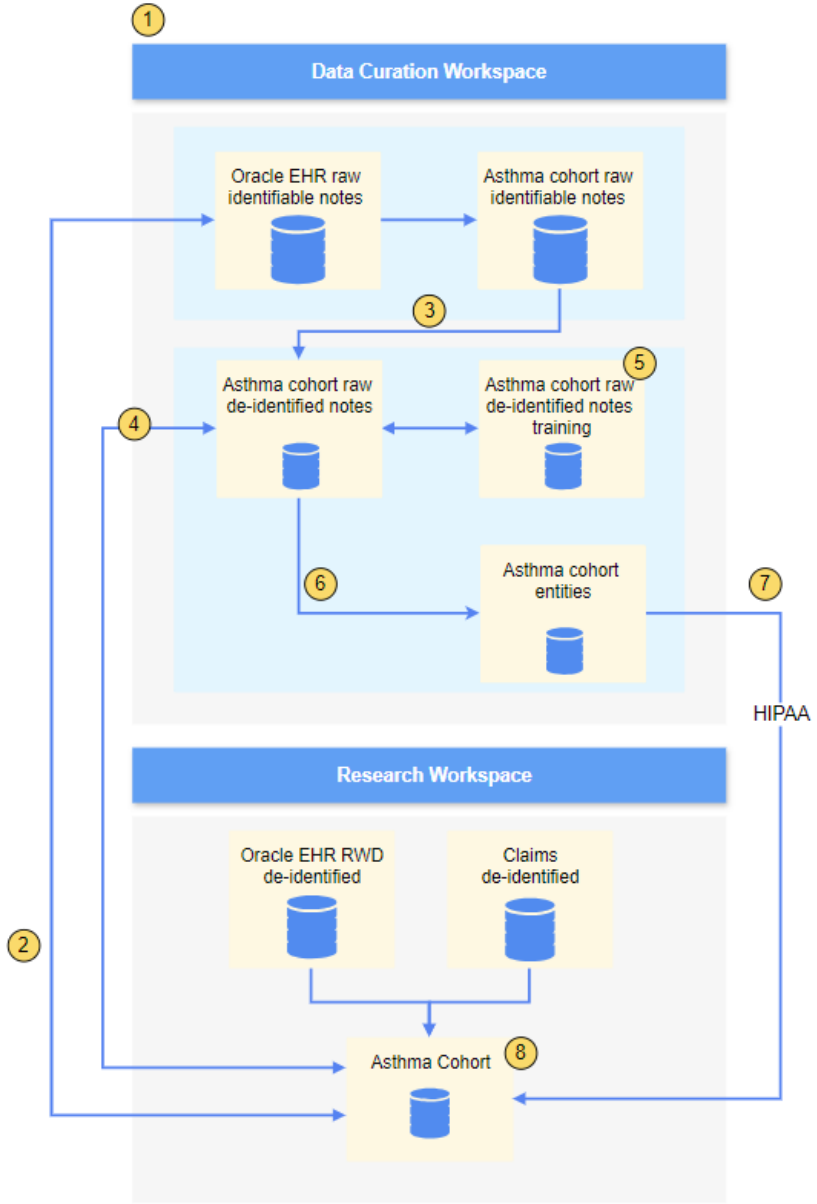
- ❑ Aggressive behavior or hostility
- ❑ Agitation
- ❑ Attention problems
- ❑ Bad or vivid dreams
- ❑ Depression
- ❑ Disorientation or confusion
- ❑ Dream abnormalities
- ❑ Feeling anxious
- ❑ Hallucinations
- ❑ Irritability
- ❑ Memory problems
- ❑ Obsessive-compulsive symptoms
- ❑ Restlessness
- ❑ Sleepwalking
- ❑ Stuttering
- ❑ Suicidal thoughts and actions
- ❑ Tremor or shakiness
- ❑ Trouble sleeping
- ❑ Uncontrolled muscle movements

Content

1. Introduce MOSAIC-NLP
2. MOSAIC-NLP Study Design Considerations
3. **Entity Extraction**

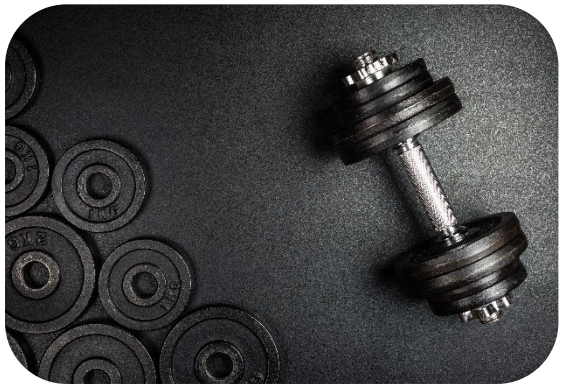
Entity Extraction Environment and Process

1. Secure environment
   - Separate workspace for data curation and research
   - Secure access
2. Use de-ID structured data + identifiable notes to create cohort AND study note dataset
3. De-ID notes within an acceptable level
4. Sampling frame for training dataset
   - Use de-ID structured data for note meta data
   - Identify important areas of variability: Healthcare system, age group, note/encounter type (mental health notes!)
5. Train and tune model
   - Review and revise entities
   - Qualitative and quantitative assessment
6. Run NLP and extract entities
7. HIPAA approval of new data fields
8. Data management



Entities and NLP Models

Annotation Guidelines



Model Training

Model Tuning

Entity Identification

- Aggressive behavior or hostility
- Agitation
- **Attention problems**
- Bad or vivid dreams
- Depression
- Disorientation or confusion
- Dream abnormalities
- Feeling anxious
- Hallucinations
- Irritability
- Memory problems
- Obsessive-compulsive symptoms
- Restlessness
- Sleepwalking
- Stuttering
- Suicidal thoughts and actions
- Tremor or shakiness
- Trouble sleeping
- Uncontrolled muscle movements

Annotation Guidelines

Attention problems

In NLP Lab: Attention_Problems

Definition: this entity contains mentions of clinical findings related to attention problems.

Extraction rules: extract only symptoms related to attention problems, but not ADHD since it is extracted under a different entity.

Examples:

1. Trouble concentrating **AttentionProblems Absent**: not at all
2. Trouble concentrating **AttentionProblems**: more than half the days
3. Her mother refers she struggles with attention **AttentionProblems** in school.
4. He complains of brain fog and attention difficulties **AttentionProblems** as side effects.
5. Watch for these signs: feeling restless, agitated, or hopeless, having trouble concentrating **AttentionProblems Hypothetical** or making decisions, having unexplained physical complaints, feeling irritable, angry, or aggressive.
6. Attention: Easily distracted **AttentionProblems**., Thought Content: Appropriate.

Assertions: Past, Absent, Family History, Someone else, Possible, Hypothetical, Present

Annotation Guidelines

Attention problems

In NLP Lab: Attention_Problems

Definition: this entity contains mentions of clinical findings related to attention problems.

Extraction rules: extract only symptoms related to attention problems, but not ADHD since it is extracted under a different entity.

Examples:

1. Trouble concentrating **AttentionProblems Absent**: not at all
2. Trouble concentrating **AttentionProblems**: more than half the days
3. Her mother refers she struggles with attention **AttentionProblems** in school.
4. He complains of brain fog and attention difficulties **AttentionProblems** as side effects.
5. Watch for these signs: feeling restless, agitated, or hopeless, having trouble concentrating **AttentionProblems Hypothetical** or making decisions, having unexplained physical complaints, feeling irritable, angry, or aggressive.
6. Attention: Easily distracted **AttentionProblems**., Thought Content: Appropriate.

Assertions: Past, Absent, Family History, Someone else, Possible, Hypothetical, Present

Decision to include

- Include attention problems for outcomes and control of psychiatric history
- Pediatric psychiatrists recommended adding ADHD to entities, as children are often diagnosed with ADHD as attention problems.

Entity Model Training

Clinical text: Persistent Depressive Disorder = Depression?

Clinicians "No"
Two different clinical entities

- As an **outcome** persistent depressive disorder would not be 'caused' by the exposure
- As a **covariate** persistent depressive disorder could be a moderator

Entity Model Training

Clinical text: Persistent Depressive Disorder = Depression?

Clinicians "No"
Two different clinical entities

- As an **outcome** persistent depressive disorder would not be 'caused' by the exposure
- As a **covariate** persistent depressive disorder could be a moderator

Annotation Guideline Options

- Persistent depressive disorder new entity
- Persistent depressive disorder = depression

Entity Model Training

Clinical text: Persistent Depressive Disorder = Depression?

Clinicians "No"
Two different clinical entities

- As an **outcome** persistent depressive disorder would not be 'caused' by the exposure
- As a **covariate** persistent depressive disorder could be a moderator

Annotation Guideline Options

- Persistent depressive disorder new entity
- Persistent depressive disorder = depression

DECISION

- ✓ Rare
- ✓ Clinicians noted the variability of using 'official' DSM diagnoses by physician type
- ✓ Important moderator/covariate

Entity Metrics

	Entity exists	Entity does not exist
Prediction an entity exists	True positive predict positive, and is positive → Correct / True prediction	False positive predict positive, but is negative → Incorrect / False positive Type I error
Prediction an entity does not exist	False negative predict negative, but is positive → Incorrect / False prediction Type II error	~~True negative~~ ~~predict negative, and is negative~~ ~~→ Correct / True prediction~~ (there is no true negative in NLP)

	Text	True Label	Predicted Label	
Entity	Trouble Concentrating	Attention problems	Attention problems	TP
Entity	Trouble Concentrating	Attention problems	0	FN
Entity	School	0	Attention problems	FP

Classification Metrics

Name	Description	Interpretation
Precision	$\frac{true\ positives}{true\ positives + false\ positives}$	When the prediction is positive, how often is it correct?
Recall	$\frac{true\ positives}{true\ positives + false\ negatives}$	When prediction is positive, how often does it predict yes?
F-1 score	A measure that balances precision and recall. $F-1\ score = 2 * \frac{precision\ *\ recall}{precision\ +\ recall}$	Balance score between precision and recall → use F-1 score

As a rule of thumb, ~80 in f-1 is considered a good model

Entity Model Tuning

Entity	TP	FP	FN	Total labels	Precision	Recall (Sensitivity)	F1
Self-harm	950	34	37	1021	0.965447	0.962513	0.963978
ADHD	135	6	7	148	0.957447	0.950704	0.954064
Suicide Attempt	89	9	10	108	0.908163	0.898990	0.903553
Aggressive / Hostility	229	33	43	305	0.874046	0.841912	0.857678
Agitation	58	12	11	81	0.828571	0.840580	0.834532
Suicidal Ideation	61	4	25	90	0.938462	0.709302	0.807947
Attention Problems	53	5	22	80	0.913793	0.706667	0.796993
Completed Suicide	0	0	9	9	--	0	0
Stuttering	0	0	2	2	--	0	0

F1 is the weighted average of precision (PPV; TP/(TP+FP)) and recall (sensitivity; TP/(TP+FN)) metrics

Entity Model Tuning

Entity	TP	FP	FN	Total labels	F1
Self-harm	950	34	37	1021	0.963978
ADHD	135	6	7	148	0.954064
Suicide Attempt	89	9	10	108	0.903553
Aggressive / Hostility	229	33	43	305	0.857678
Agitation	58	12	11	81	0.834532
Suicidal Ideation	61	4	25	90	0.807947
Attention Problems	53	5	22	80	0.796993
Completed Suicide	0	0	9	9	0
Stuttering	0	0	2	2	0

Decision to remove

Rare event/difficult to identify in model

Since for HIPAA conformance this cannot be used for patient, the added value is low for family history or ‘someone else’

Decision to remove

Rare event/difficult to identify in model

High cost time/effort

Implications of Adding Unstructured Data to Pharmacoepidemiology Studies

Dena Jaffe, MSc, PhD

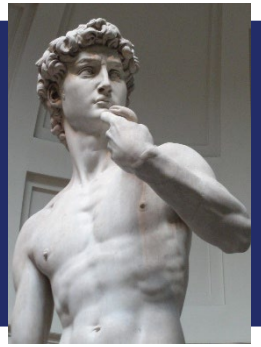
Oracle Health

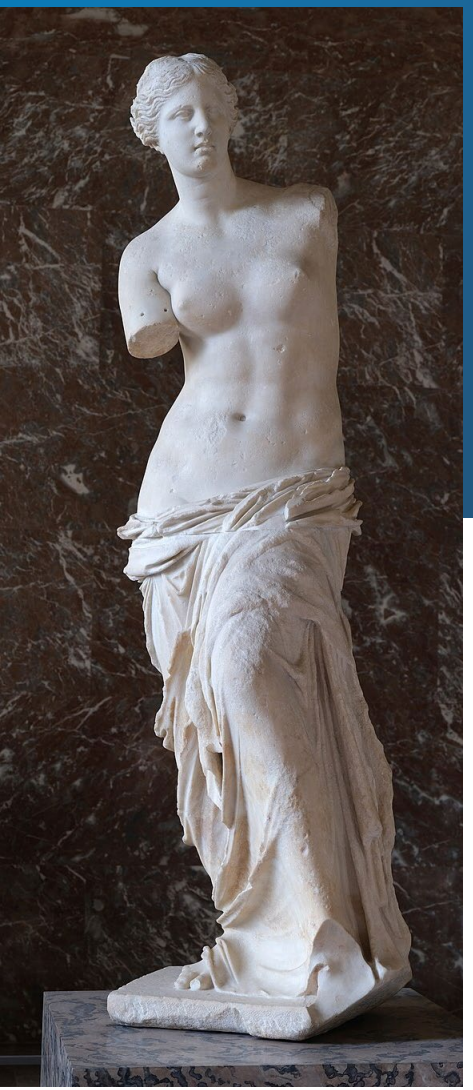
Extracted Data to a Full Study-Ready Dataset

Data from Clinical notes extracted using NLP models

?

Additional study data





Sculpting from Marble – Unstructured Data Management

Things to consider:

- Confidence score cutoffs
- Current variables and what is going to be added and how
  - Unstructured data sources (e.g., types of entities, types of assertions, types of modifiers).
    - scope and scale
- Temporality
- Input from healthcare providers

Confidence Score Cutoffs

Data presented with confidence scores sorted by delusion, greatest to least

documentId	delusion	substance_abuse	exercise	obsessive_compulsive	aggressive_behv_hostility
10008228609	0.999	0.9631	0.0	0.0	0.0
10024091330	0.9985	0.0	0.0	0.0	0.0
10024706670	0.8582	0.0	0.0	0.0	0.0
1002789805	0.9985	0.6772	0.0	0.0	0.9877
10038824818	0.9986	0.0	0.0	0.0	0.0
10066168172	0.9103	0.0	0.0	0.0	0.0
10066710469	0.9984	0.0	0.0	0.0	0.0
10069957827	0.9982	0.6732	0.0	0.0	0.0
10074486383	0.9981	0.0	0.0	0.0	0.0
10075733850	0.9981	0.0	0.0	0.0	0.0
10080306729	0.9989	0.4907	0.0	0.0	0.0

Note: JSL confidence cutoff >0.5

Confidence Score Cutoffs

Data presented with confidence scores sorted by delusion, greatest to least

docum...	delusion	substance_abuse	exercise	obsessive_compulsive	aggressive_behv_hostility
10008228609	1	1	0	0	0
10024091330	1	0	0	0	0
10024706670	1	0	0	0	0
1002789805	1	1	0	0	1
10038824818	1	0	0	0	0
10066168172	1	0	0	0	0
10066710469	1	0	0	0	0
10069957827	1	1	0	0	0
10074486383	1	0	0	0	0
10075733850	1	0	0	0	0
10080306729	1	1	0	0	0

Consider implications of using a different cutoff score for identifying binary data

Data Management: What is going to be added and how

		Assertions				
	Variable	Description Absent	Description Present	Description Possible	Description family history	Description Past
Modifiers	anxiety	current anxiety absent	present anxiety	possible anxiety	family history of current anxiety	past anxiety
	anxiety-worsened	worsened anxiety absent	present worsened anxiety	possible worsened anxiety	family history of worsened anxiety	past worsened anxiety
	anxiety-mild	mild anxiety absent	present mild anxiety	possible mild anxiety	family history of mild anxiety	past mild anxiety
	anxiety-moderate	moderate anxiety absent	present moderate anxiety	possible moderate anxiety	family history of moderate anxiety	past moderate anxiety
	anxiety-severe	severe anxiety absent	present severe anxiety	possible severe anxiety	family history of severe anxiety	past severe anxiety
	anxiety-alleviated	alleviated anxiety absent	present alleviated anxiety	possible alleviated anxiety	family history of alleviated anxiety	past alleviated anxiety
	anxiety-other_modifier	other modifier anxiety absent	present other modifier anxiety	possible other modifier anxiety	family history of other modifier anxiety	past other modifier anxiety
	anxiety-intermittent	intermittent anxiety absent	present intermittent anxiety	possible intermittent anxiety	family history of intermittent anxiety	past intermittent anxiety

- Include anxiety from structured data as:
  - Covariate (pre-index)
  - Outcome (post-index)
- How to distill all this extracted NLP data into pre-index data and post-index data?
- How to add to structured data variables?

Temporality Example: Anxiety

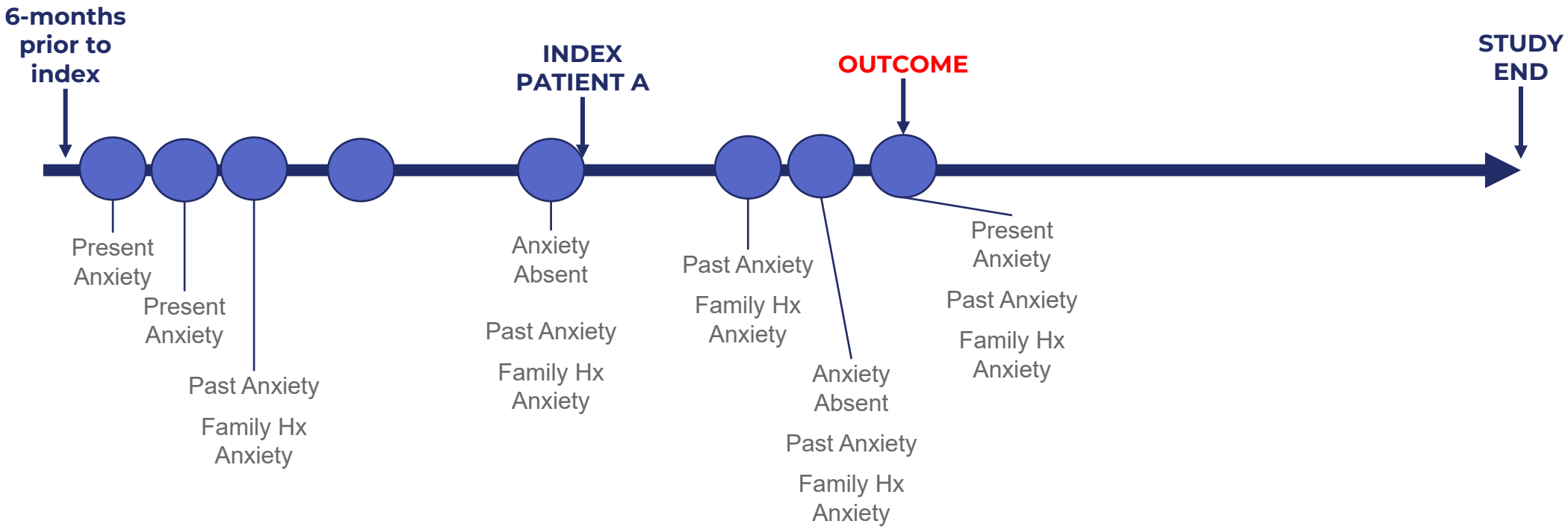
Extraction rules:

Present: No assertion label has to be added, entities are considered to have the present assertion by default if they do not have other assertion label.

Past entities are found in phrases in past tense or that include words such as *past, before, remote, previously, former, etc.*

Absent or negated entities are found in phrases that include words such as no, without, lack, etc.

Family history: Only consanguinity relations are to be asserted as family history.



Temporality Example: Possible

Possible
In NLP Lab: Possible

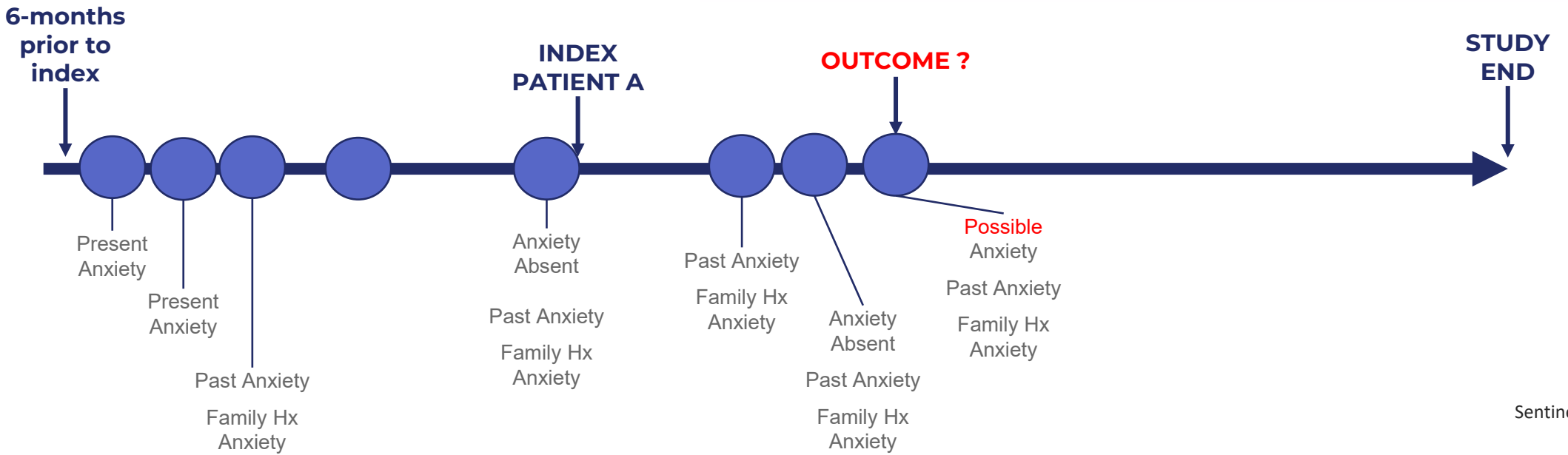
Definition: This label is assigned to entities that are possible but not confirmed.

Extraction rules: Possible entities are found in phrases that include words such as *might*, *maybe*, *perhaps*, *could*, *likely*, *unlikely*, *to rule out*, *suspects* etc.

Use Hypothetical assertion instead of Possible when a general description of a disease is provided, like in patient information forms, or when literature quotations are made.

Examples:

1. I will also order other tests, including tests to rule out other conditions, such as COPD COPD Possible
2. Additional tests will be done to rule out other possible causes of your symptoms such as GAD Anxiety Disorder Possible
3. It has not been confirmed yet, but the staff suspects it was a suicide attempt Suicide attempt Possible.
4. Based on the symptoms, circadian rhythm disorder circadian rhythm disorder Possible is suspected.

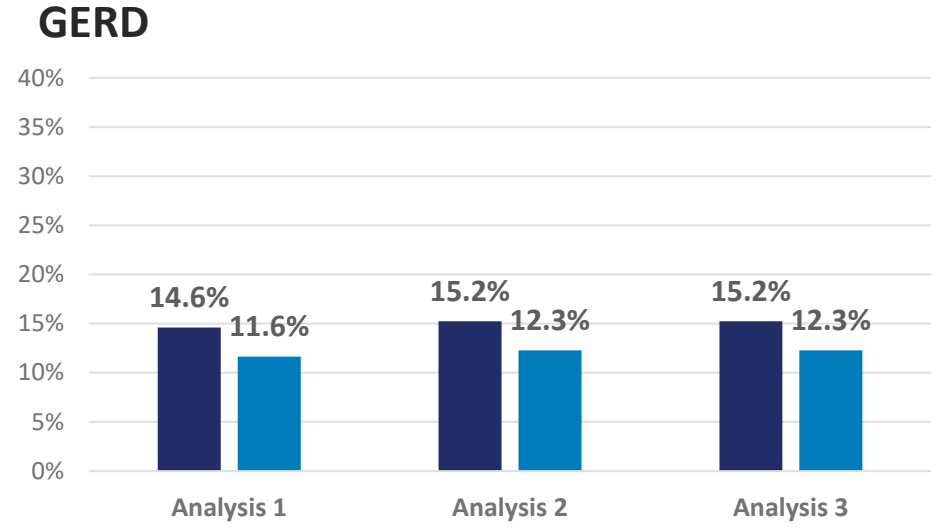


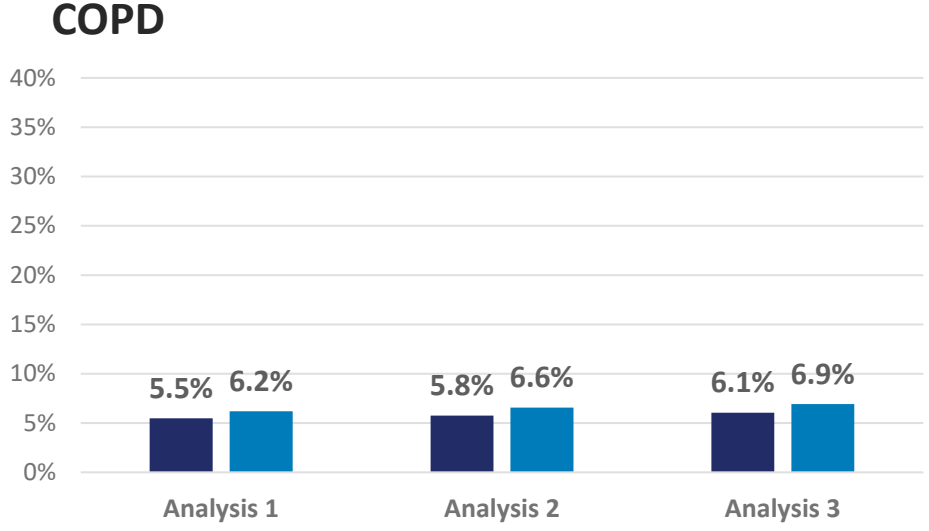
Temporality & Covariates: Added Value of Structured and Unstructured EHR

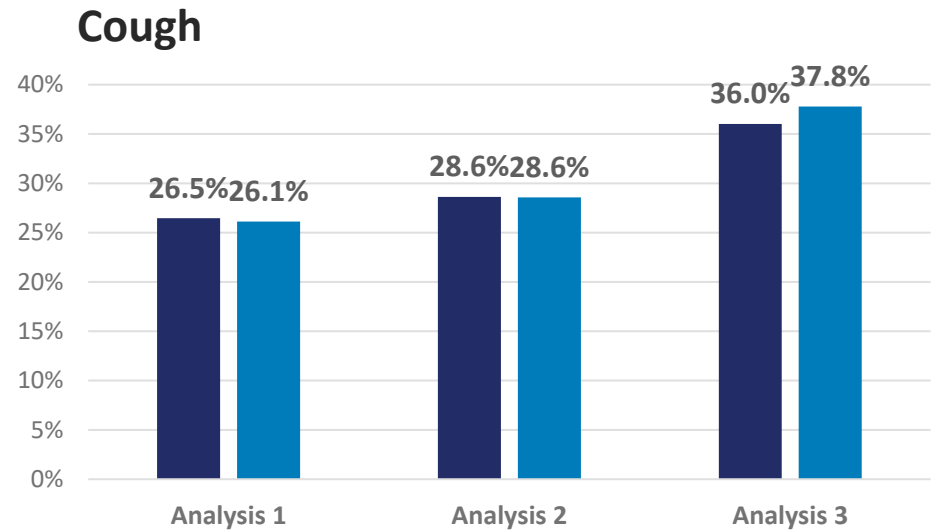
(6m prior to index)

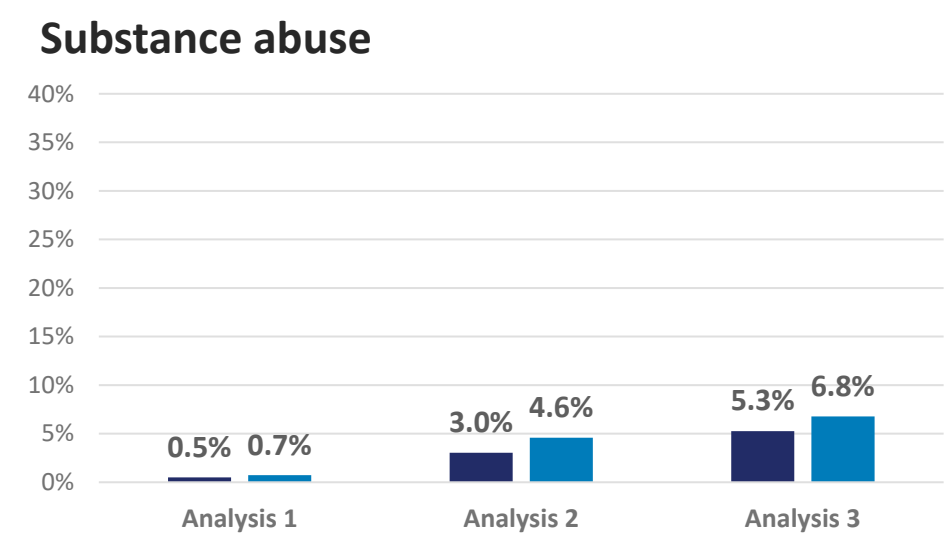
Montelukast N=39,665

ICS N=69,411



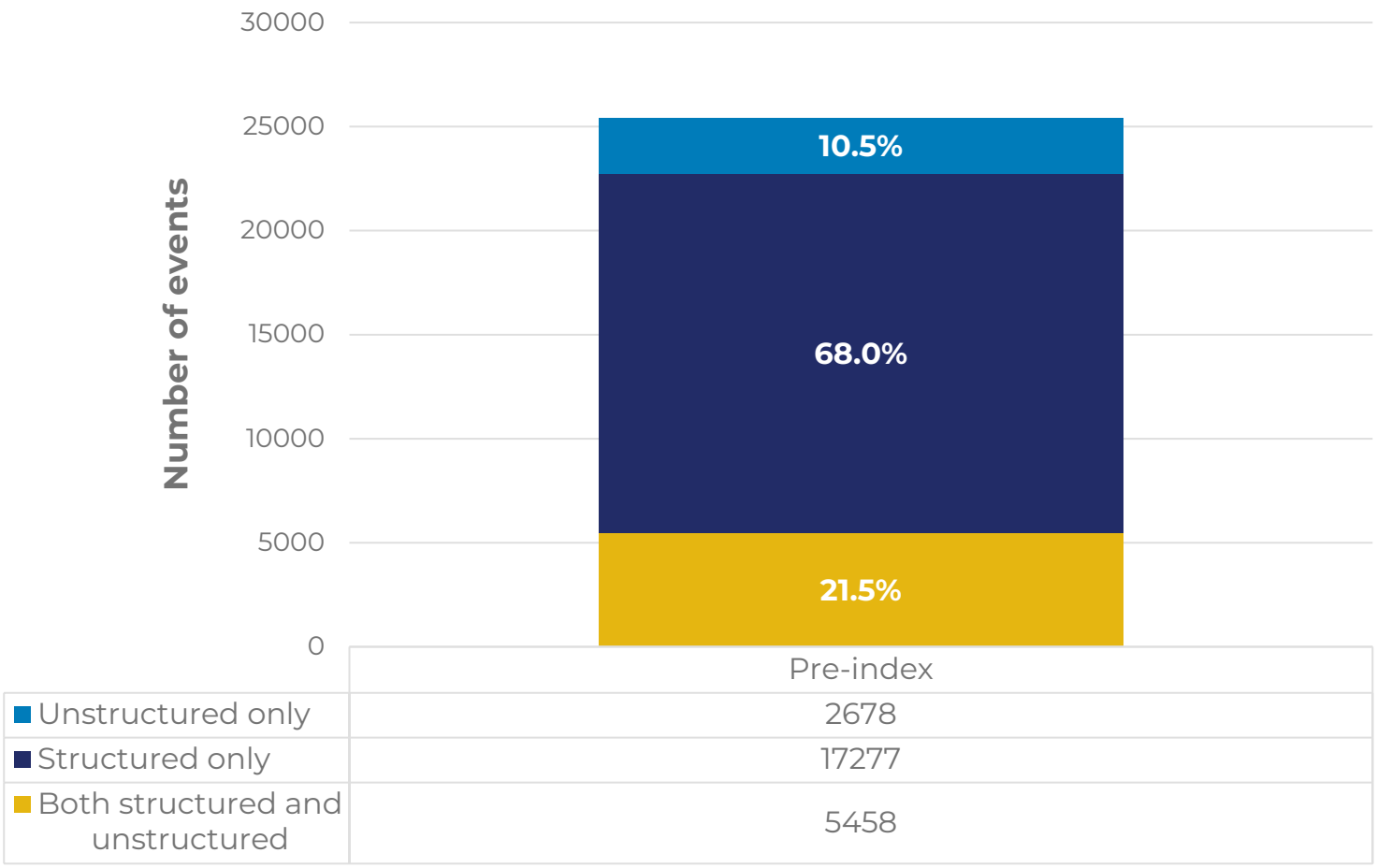






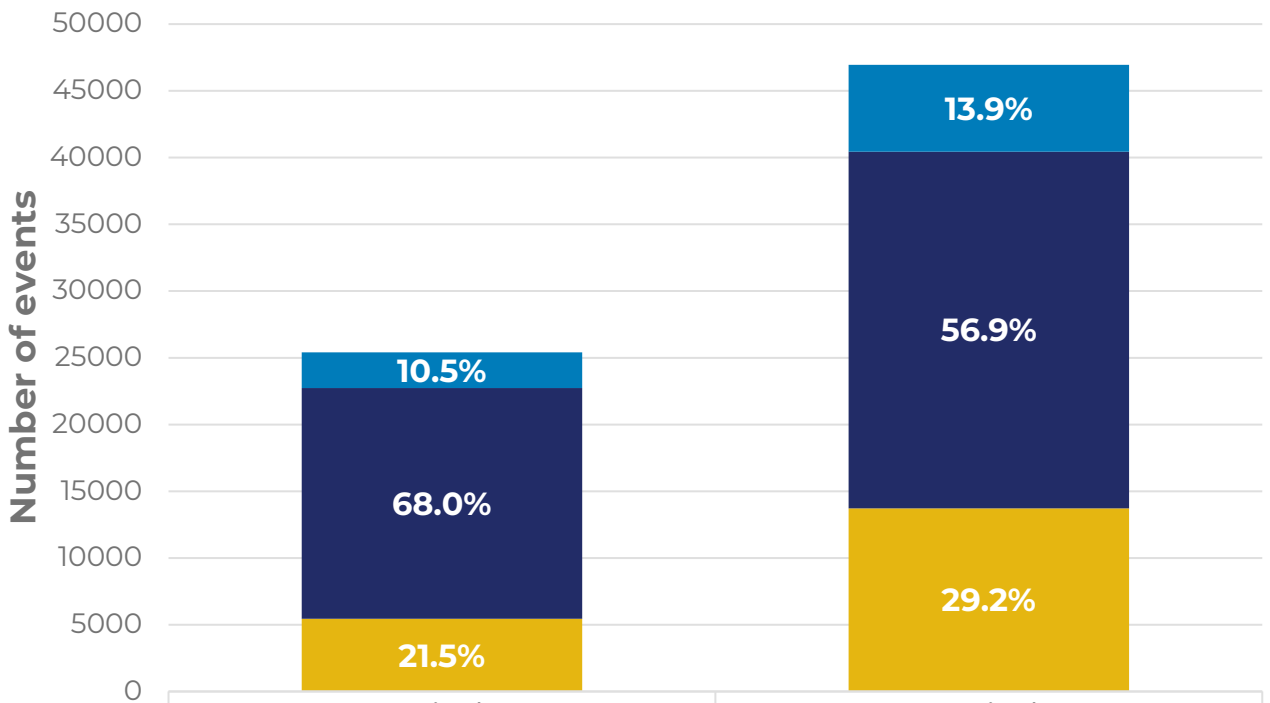
Analysis 1 = claims only data; Analysis 2 = claims + EHR structured data; Analysis 3 = claims + EHR structured data + EHR unstructured data

Anxiety Data Sources in MOSAIC-NLP



Pre-index (unmatched covariates) – Anxiety 6-month prior to index (n = 25,413)

Anxiety Data Sources in MOSAIC-NLP



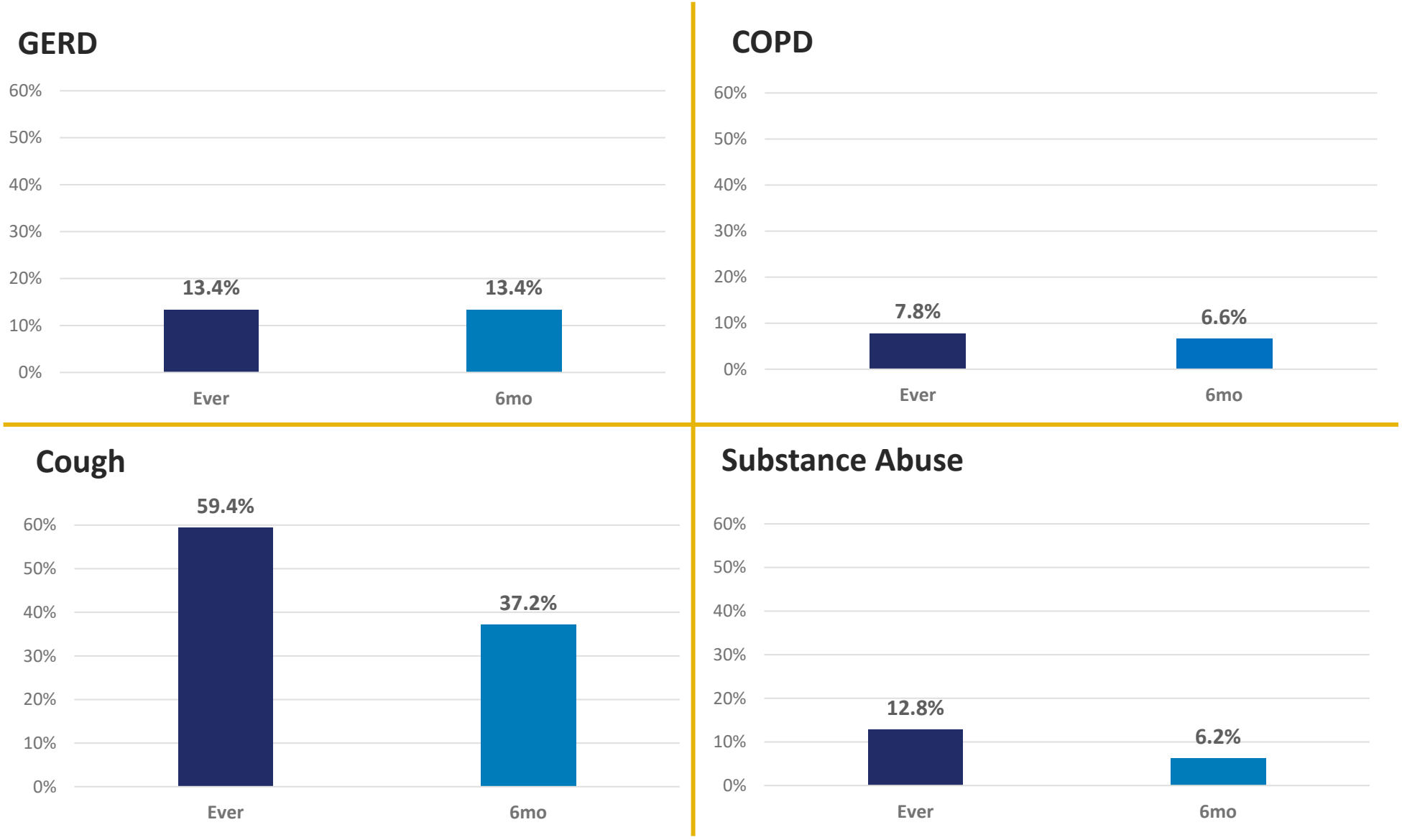
	Pre-index	Post-index
Unstructured only	2678	6502
Structured only	17277	26716
Both structured and unstructured	5458	13725

Pre-index (unmatched covariates) – Anxiety 6-months prior to index (n = 25,413)
Post-index (unmatched outcomes) – Anxiety post index (n = 46,943)

Temporality & Covariates: Look back window (Analysis 3: Ever vs 6 months)

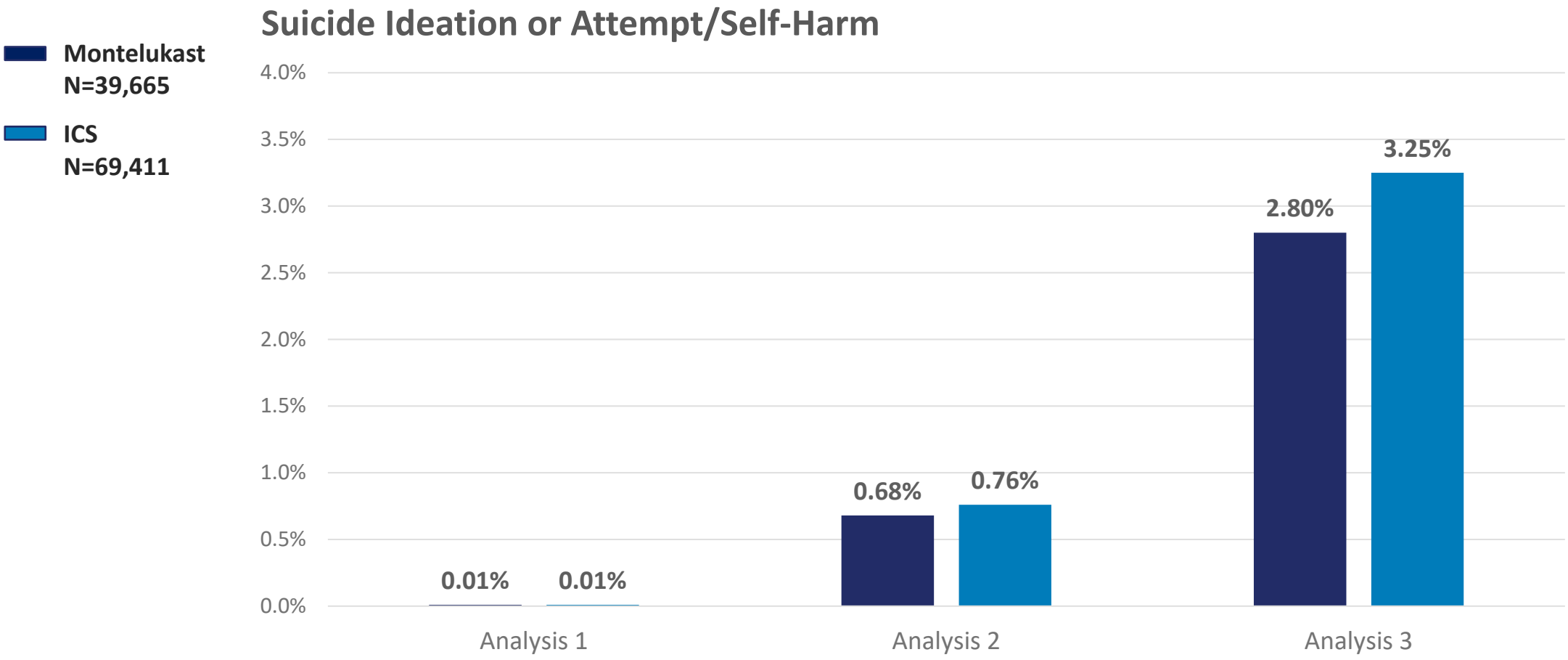
Montelukast N=39,665

ICS N=69,411



Analysis 3 = claims + EHR structured data + EHR unstructured data

Healthcare Providers & Covariates: Added Value of Structured and Unstructured EHR
(6m prior to index)



Study 1 = claims only data; Study 2 = claims + EHR structured data; Study 3 = claims + EHR structured data + EHR unstructured data

Stay tuned for results

Huge thank you to my fellow sculptors (statisticians and data scientists)

Bridget Balkaran

Kyla Finlayson

Rob J. Taylor

Austin Yue

NLP and Drug Safety at the US FDA

Sarah K Dutcher, PhD, MS
US Food and Drug Administration

FDA and Artificial Intelligence

Figure 1. Four areas of focus regarding the development and use of AI across the medical product lifecycle.

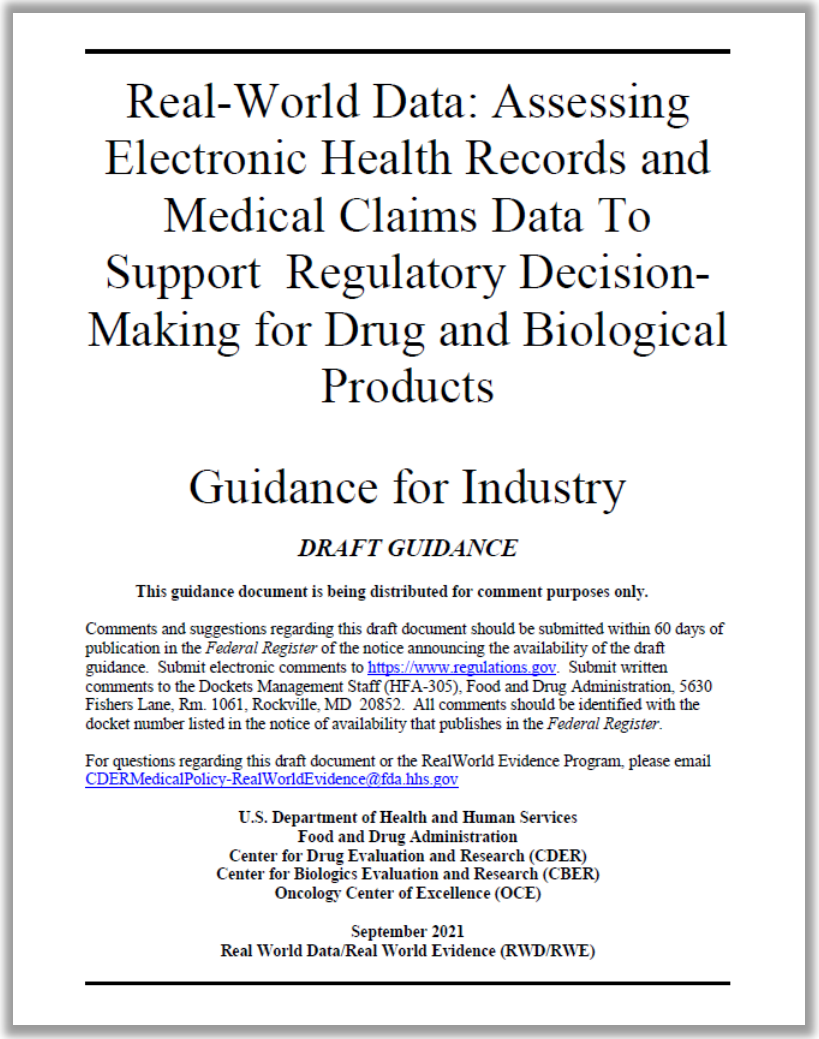
What role is AI/ML playing in drug development?

FDA recognizes the increased use of AI/ML throughout the drug development life cycle and across a range of therapeutic areas. In fact, FDA has seen a significant increase in the number of drug and biologic application submissions using AI/ML components over the past few years, with more than 100 submissions reported in 2021. These submissions traverse the landscape of drug development — from drug discovery and clinical research to postmarket safety surveillance and advanced pharmaceutical manufacturing.

Additionally, AI/ML is increasingly integrated in areas where FDA is actively engaged, including Digital Health Technologies (DHTs), and Real-World Data (RWD) analytics.

https://www.fda.gov/science-research/science-and-research-special-topics/artificial-intelligence-and-machine-learning-aiml-drug-development

FDA Guidance on Real World Data



Real-World Data: Assessing Electronic Health Records and Medical Claims Data To Support Regulatory Decision-Making for Drug and Biological Products

Guidance for Industry

DRAFT GUIDANCE

This guidance document is being distributed for comment purposes only.

Comments and suggestions regarding this draft document should be submitted within 60 days of publication in the *Federal Register* of the notice announcing the availability of the draft guidance. Submit electronic comments to https://www.regulations.gov. Submit written comments to the Dockets Management Staff (HFA-305), Food and Drug Administration, 5630 Fishers Lane, Rm. 1061, Rockville, MD 20852. All comments should be identified with the docket number listed in the notice of availability that publishes in the *Federal Register*.

For questions regarding this draft document or the RealWorld Evidence Program, please email CDERMedicalPolicy-RealWorldEvidence@fda.hhs.gov

U.S. Department of Health and Human Services
Food and Drug Administration
Center for Drug Evaluation and Research (CDER)
Center for Biologics Evaluation and Research (CBER)
Oncology Center of Excellence (OCE)

September 2021
Real World Data/Real World Evidence (RWD/RWE)

Technological advances in AI (NLP, ML) permit more rapid processing of unstructured EHR data to:

1. Extract data elements from structured fields and unstructured text in EHRs
2. Develop computer algorithms to identify outcomes
3. Evaluate images or laboratory results

These computer-assisted methods currently require significant human-aided curation and decision-making, injecting an additional level of data variability and quality considerations into the final analytic dataset.

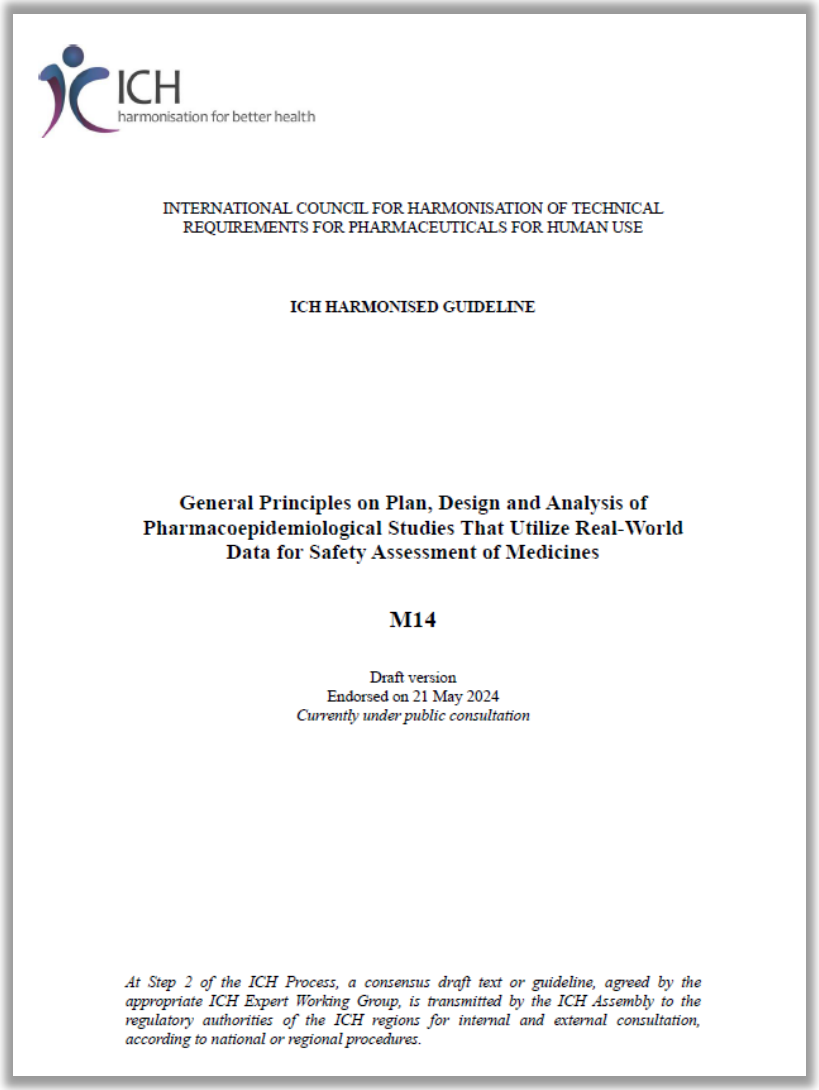
Study protocols should specify:

- Assumptions and parameters of the computer algorithms used
- The data source from which the information was used to build the algorithm
- Whether the algorithm was supervised (using expert input and review) or unsupervised
- Metrics associated with validation of the methods
- Relevant impacts on data quality

https://www.fda.gov/regulatory-information/search-fda-guidance-documents/real-world-data-assessing-electronic-health-records-and-medical-claims-data-support-regulatory

ICH M14 Guideline

- International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH) M14 draft guideline focuses on non-interventional pharmacoepidemiological studies and includes basic principles that may apply to these studies when RWD elements are included
  - Endorsed by ICH 21 May 2024
  - Released for public consultation
- "Key clinical information are often unstructured data within EHRs, either as free text fields (such as healthcare practitioner notes)…To enhance the efficiency of data abstraction, a range of approaches, including both existing and emerging technologies (e.g., natural language processing…) are increasingly being used to convert unstructured data into a computable, structured data format."



ICH
harmonisation for better health

INTERNATIONAL COUNCIL FOR HARMONISATION OF TECHNICAL REQUIREMENTS FOR PHARMACEUTICALS FOR HUMAN USE

ICH HARMONISED GUIDELINE

General Principles on Plan, Design and Analysis of Pharmacoepidemiological Studies That Utilize Real-World Data for Safety Assessment of Medicines

M14

Draft version
Endorsed on 21 May 2024
Currently under public consultation

At Step 2 of the ICH Process, a consensus draft text or guideline, agreed by the appropriate ICH Expert Working Group, is transmitted by the ICH Assembly to the regulatory authorities of the ICH regions for internal and external consultation, according to national or regional procedures.

https://www.ich.org/page/multidisciplinary-guidelines

Use of NLP for Pharmacovigilance: FAERS

- Use of the FDA Adverse Event Reporting System (FAERS) for pharmacovigilance complements pharmacoepidemiology studies in FDA's overall approach to monitor and promote drug safety
  - FAERS receives over 2 million reports every year
- The Center for Drug Evaluation and Research's (CDER) Office of Surveillance and Epidemiology (OSE) implemented the **Information Visualization Platform (InfoViP)** in 2022 to support safety reviewer's examination of individual case safety reports in FAERS
- InfoViP incorporates NLP capabilities, machine learning (ML), and advanced data visualizations in a tool to support postmarket safety surveillance
  1. InfoViP uses NLP to scan case report narratives to find and visually display relevant clinical information in a timeline
  2. InfoViP uses NLP to scan, extract, and compare numerous data points among a large group of ICSRs to detect duplicates automatically
  3. InfoViP uses ML to classify case reports based on their level of information quality, which safety reviewers can use to triage high-quality reports for priority review to detect safety concerns more rapidly

https://www.fda.gov/drugs/cder-conversations/information-visualization-platform-infovip-cders-new-artificial-intelligence-safety-surveillance

Use of NLP for Pharmacovigilance: FAERS

Pharmaceutical Medicine (2021) 35:307–316
https://doi.org/10.1007/s40290-021-00398-5

ORIGINAL RESEARCH ARTICLE

Leveraging Case Narratives to Enhance Patient Age Ascertainment from Adverse Event Reports

Phuong Pham[1,2] · Carmen Cheng[2] · Eileen Wu[2] · Ivone Kim[2] · Rongmei Zhang[3] · Yong Ma[3] · Cindy M. Kortepeter[2] · Monica A. Muñoz[1,2]

Accepted: 5 August 2021 / Published online: 2 September 2021
This is a U.S. government work and not under copyright protection in the U.S.; foreign copyright protection may apply 2021

Abstract
Introduction Missing age presents a significant challenge when evaluating individual case safety reports (ICSRs) in the FDA Adverse Event Reporting System (FAERS). When age is missing in an ICSR's structured field, it may be in the report's free-text narra
Objectives Th
processing (NL

frontiers | Frontiers in Drug Safety and Regulation

TYPE Brief Research Report
PUBLISHED 14 November 2022
DOI 10.3389/fdsfr.2022.1020943

Check for updates

OPEN ACCESS

EDITED BY
G. Niklas Norén,
Uppsala Monitoring Centre, Sweden

REVIEWED BY
Luis Pinheiro,
European Medicines Agency, Netherlands
Xiaofei Ye,
Second Military Medical University, China

*CORRESPONDENCE
Vivian Dang,
Vivian.Dang@fda.hhs.gov

SPECIALTY SECTION
This article was submitted to Advanced Methods in Pharmacovigilance and

Evaluation of a natural language processing tool for extracting gender, weight, ethnicity, and race in the US food and drug administration adverse event reporting system

Vivian Dang [1]*, Eileen Wu[1], Cindy M. Kortepeter [1], Michael Phan[1], Rongmei Zhang[2], Yong Ma[2] and Monica A. Muñoz [1]

- FDA evaluated an NLP tool's ability to extract age, gender, weight, ethnicity, race from FAERS case report narratives, when missing in structured fields
- NLP tool implementation provided meaningful improvements in the availability of age and gender information to support pharmacovigilance activities conducted with FAERS data
- NLP tools had minimal impact on the extraction of weight, ethnicity, or race from free-text fields largely because the information was infrequently provided by the reporter

Dang V, Wu E, Kortepeter CM, et al. Front Drug Saf Regul. 2022;2:1020943.
Pham P, Cheng C, Wu E, et al. Pharmaceut Med. 2021 Sep;35(5):307-316.

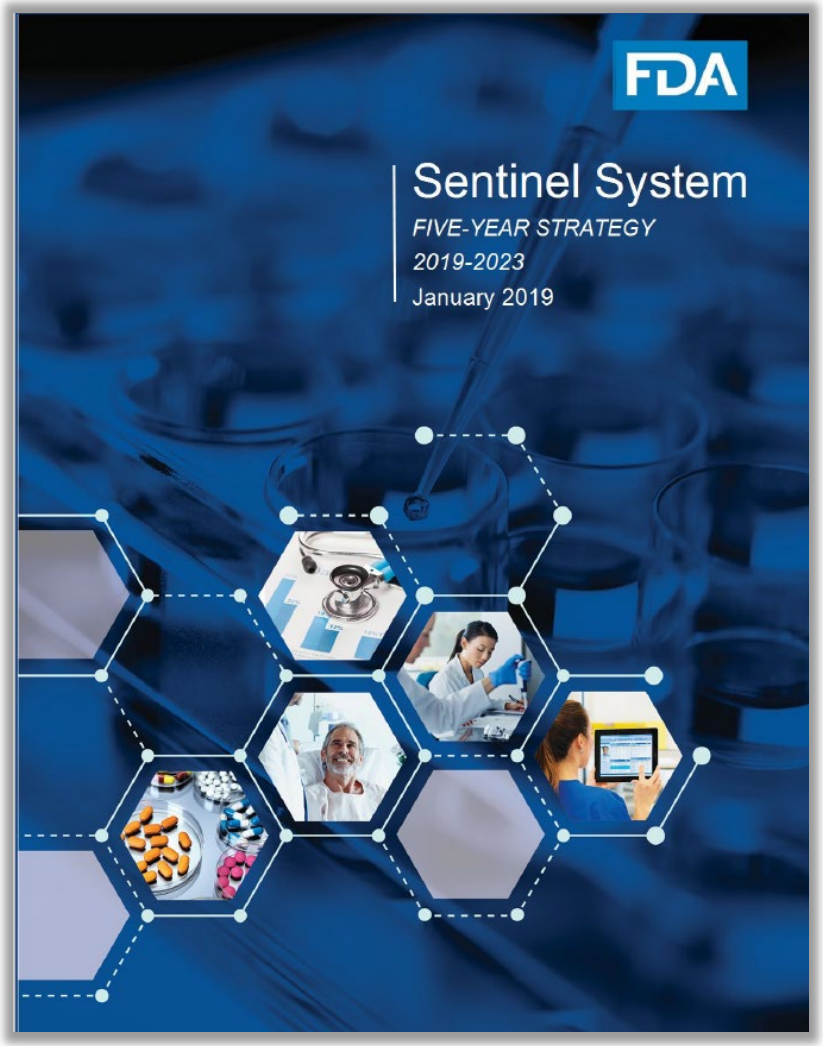
Sentinel's Strategic Plan

"FDA will focus its investment on innovations emerging from new data science disciplines, such as natural language processing and machine learning, and seek to expand its access to and use of electronic health records (EHRs)"

Goal: Use Sentinel projects using NLP of unstructured data to establish standards and inform best practices for regulatory use

Thinking evolved during implementation of the strategic plan:

- Original emphasis on use of NLP-extracted data to identify previously undetected complex health outcomes that require multiple data elements
- Using NLP to extract data and implementing into a study is very complex
  - Can be used to improve capture for other data elements, not just outcomes
  - NLP algorithm transportability across healthcare systems cannot be assumed
  - Efficiency gains, but still requires substantial manual input



https://www.fda.gov/media/120333/download

Summary

- Benefits of NLP
  - Improves identification or extraction of multiple data elements needed for drug safety assessments (exposures, outcomes, covariates)
  - Increases efficiency: speed (time saving), scale (can process a larger volume of unstructured data for the same manual effort)
  - Automated approach reduces potential for manual error or disagreement
- Considerations for using NLP in pharmacoepidemiology studies
  - NLP algorithm accuracy is dependent on multiple factors: source data system, selection of notes for training
  - Users need to understand how study results may be impacted by the performance of NLP algorithms and how NLP-extracted data are operationalized for the study
  - Requires a team with a variety of expertise

Applying NLP to semi- and unstructured EHR data can capture valuable clinical and patient information that enriches structured data available via claims data, thus enhancing our abilities to assess medical product safety

Thank You

Rishi J. Desai
rdesai@bwh.harvard.edu

Dena Jaffe
dena.jaffe@oracle.com

Elise Berliner
elise.berliner@oracle.com

Sarah K. Dutcher
Sarah.dutcher@fda.hhs.gov

Questions and Answers

Discussion Questions

- Look back window
- Weight given to entities in notes
- Events are therapeutic area dependent
- Credibility Assessment Framework – outline considerations on new methods/technologies of acceptability