Scalable Phenotyping for Safety Outcomes Using Electronic Health Record Data

Brian D. Williamson, PhD
Kaiser Permanente Washington Health Research Institute
Disclaimer

The views expressed in this presentation represent those of the presenters and do not necessarily represent the official views of the U.S. FDA.
Acknowledgments

Sentinel Advanced Phenotyping Framework and Scalable Natural Language Processing (NLP) Teams

**FDA**
- Adebola Ajao
- Robert Ball
- Steven Bird
- Sara Karami
- Yong Ma
- Michael Nguyen
- Daniijela Stojanovic
- Sanrat Wittayanukorn
- Mingfeng Zhang
- Yueqin Zhao

**Harvard Pilgrim Health Care Institute**
- Adee Kennedy
- Judy Maro
- Elizabeth Messenger-Jones
- Kathleen Shattuck
- Mayura Shinde
- Darren Toh

**Kaiser Washington**
- Maralyssa Bann
- Will Bowers
- David Carrell
- David Cronkite
- James Floyd
- Monica Fujii
- Vina Graham
- Kara Haugen
- Eric Johnson
- Ron Johnson
- Ann Kelley
- Linda Kiel
- Jennifer Nelson
- Arvind Ramaprasan
- Mary Shea
- Brian Williamson
- Jing Zhou

**Univ. of Washington**
- Patrick Heagerty

**Kaiser Northwest**
- Andrew Felcher
- Brian Hazlehurst
- Denis Nyongesa
- Daniel Sapp
- Matthew Slaughter

**Putnam Data Sciences**
- Susan Gruber

**Vanderbilt University**
- Cosmin (Adi) Bejan
- Michael Matheny
- Joshua Osmanski
- Daniel Park
- Joshua Smith
- Dax Westerman
- Robert Winter

**Univ. of Michigan**
- Xu Shi

**HealthCore**
- Kevin Haynes

**Duke University**
- Keith Marsolo

**Brigham and Women’s Hospital, Harvard**
- Rishi Desai
- William Feldman
- Shamika More
- Shirley Wang

**Univ. of Pennsylvania**
- Kevin Johnson

**Indiana University**
- David Aronoff
Motivation

- Goal: improve safety surveillance using observational data
- Active Risk Identification and Analysis (ARIA) system:

Analytic Tools + Common Data Model = ARIA

- Electronic claims data
- Parameterized, re-usable tools and computable algorithms

Image courtesy of Michael Nguyen
Motivation

When is the ARIA Process Needed?

ARIÀ must be considered before a sponsor PMR can be issued
ARIA Sufficiency

• ARIA is sufficient when:
  • Outcome & exposure of interest, covariates can be identified from data
  • Methods can assess exposure-related risk with satisfactory precision
• 2016—2018: ARIA insufficient for 45 of 89 drug/outcome pairs
  • Inadequate identification of outcome: 38 pairs

Example ARIA sufficient* outcomes:
• GI bleeding
• Heart failure
• Lymphoma
• Major adverse cardiac events (MACE)
• Myocardial infarction
• Multiple sclerosis relapse
• Non-melanoma skin cancer
• Seizure
• Stroke

Example ARIA insufficient* outcomes:
• Acute pancreatitis
• Anaphylaxis
• Drug-induced liver injury
• Fatal MACE
• Malignancies (several)
• Nerve injury
• Suicide or suicidal ideation

*Sufficiency is highly dependent on the scientific question and regulatory context
Improving ARIA Sufficiency

- Our focus: outcome identification (phenotyping)
- Key considerations:
  - Gold-standard data creation
  - Feature engineering
  - Model development
  - Model evaluation
- Challenge: traditional chart review expensive (in time and resources)
- Approach: a general framework for scalable phenotyping algorithms
- Case studies: acute pancreatitis, anaphylaxis, severe COVID-19
Assessing Fitness for Purpose

Can a phenotyping effort succeed for the outcome of interest?

- Key considerations:
  - Downstream use of the predicted outcome
  - Ambiguity of the clinical condition (*clinical complexity*)
  - Ambiguity arising from healthcare data (*data complexity*)
Clinical and Data Complexity: Anaphylaxis

- Clinical complexity:
  - Diagnosis complex, relies on subjective assessment of signs and symptoms
  - 20% of charts at KPWA identified as “difficult” or discordant across two MD reviewers
  - Event often does not occur under direct observation
- Data complexity:
  - Relevant information captured in chart notes
Clinical and Data Complexity: Acute Pancreatitis

• Clinical complexity:
  • Established events criteria include pain, imaging results
• Data complexity:
  • Relevant information captured in ICD-10 diagnosis code and serum lipase laboratory value*

Gold-standard Data Creation

- Goal: identify true cases and controls for algorithm training
- Challenge: limited resources (time, personnel)
- Best practices:
  - Chart abstraction guidelines reflect clinical diagnostic criteria
  - Clinician oversight of chart abstractors
  - Dual review of samples to assess replicability
  - Use K-fold cross-validation
- Future work:
  - Can NLP-assisted methods reduce review time?
  - Can surrogate outcomes be incorporated in model training?
Feature Engineering

• Goal: identify useful features from the EHR
• Challenges:
  • Limited resources (time, personnel)
  • Local vocabulary reduces generalizability
• Best practices:
  • Incorporate clinical and domain knowledge
  • Engineer many features
  • Consider manual and automated approaches
• Future work:
  • Can automated approaches capture all relevant relationships?
  • Automated approaches with acute outcomes?
Model Development and Evaluation

• Goal: construct a **useful** prediction model

• Challenges:
  • Performance constrained by clinical and data complexity
  • Evaluation requires gold-standard outcomes

• Best practices:
  • Incorporate domain knowledge
  • Consider a **large, diverse** set of candidate prediction algorithms (including **machine learning**)
  • Evaluate performance using K-fold cross-validation
  • Consider many performance metrics
  • Final algorithm choice guided by downstream **performance, replicability, generalizability**

• Future work:
  • Under what conditions can models be **transported to new settings** without additional gold-standard evaluation?
Selected Results

Predicting Anaphylaxis

- Model a: Machine learning, structured & NLP (0.710)
- Model b: Traditional regression, structured & NLP (0.660)
- Model c: Machine learning, structured only (0.619)
- Model d: Traditional regression, structured only (0.584)

Predicting Symptomatic COVID-19

- Model Aggregate: AUC = 0.84
- Model Silver Label: AUC = 0.819
- Model Silver_NLP_1_COVID19_HTS: AUC = 0.816
- Model Silver_NLP_2_COVID19_CUI_Notes: AUC = 0.817
- Model Silver_NLP_2_COVID19_CUI_Notes: AUC = 0.825

Carrell et al., American Journal of Epidemiology (accepted)
Closing Thoughts

All aspects of phenotyping can be improved by

• Considering data and clinical complexity
• Incorporating domain knowledge
• Using a wide variety of tools (including machine learning), with proper evaluation

Our framework provides guidelines for **fully incorporating EHR data** into phenotyping analyses
Thank You

Brian Williamson
Assistant Investigator, Biostatistics Division
Kaiser Permanente Washington Health Research Institute
Contact: brian.d.williamson@kp.org