Sentinel

Welcome to the Sentinel Innovation and Methods Seminar Series

The webinar will begin momentarily

Please visit www.sentinelinitiative.org for recordings of past sessions and details on upcoming webinars.

Note: closed-captioning for today's webinar will be available on the recording posted at the link above.

Sentinel

Machine learning for improving high-dimensional proxy confounder adjustment

An overview of the current literature

Richard Wyss, PhD, MSc

DoPE

Purpose of Manuscript and Presentation

- Wyss R, Yanover C, El-Hay Tal, Bennett D, Platt RW, et al. "*Machine learning for improving high-dimensional proxy confounder adjustment in healthcare database studies: An overview of the current literature*. Pharmacoepidemiology and Drug Safety. 2022;31(9):932-943
  - Review paper sponsored by ISPE
  - Purpose is to give a high-level overview of recent advancements and areas for future research.
  - Tailored to applied pharmacoepidemiologic audience.
    - Does not go in depth into any specific topic
    - Give a broad overview of papers where individuals can look to read more
- In this presentation, I'll give some highlights of challenges and trends we found in our review.

Background

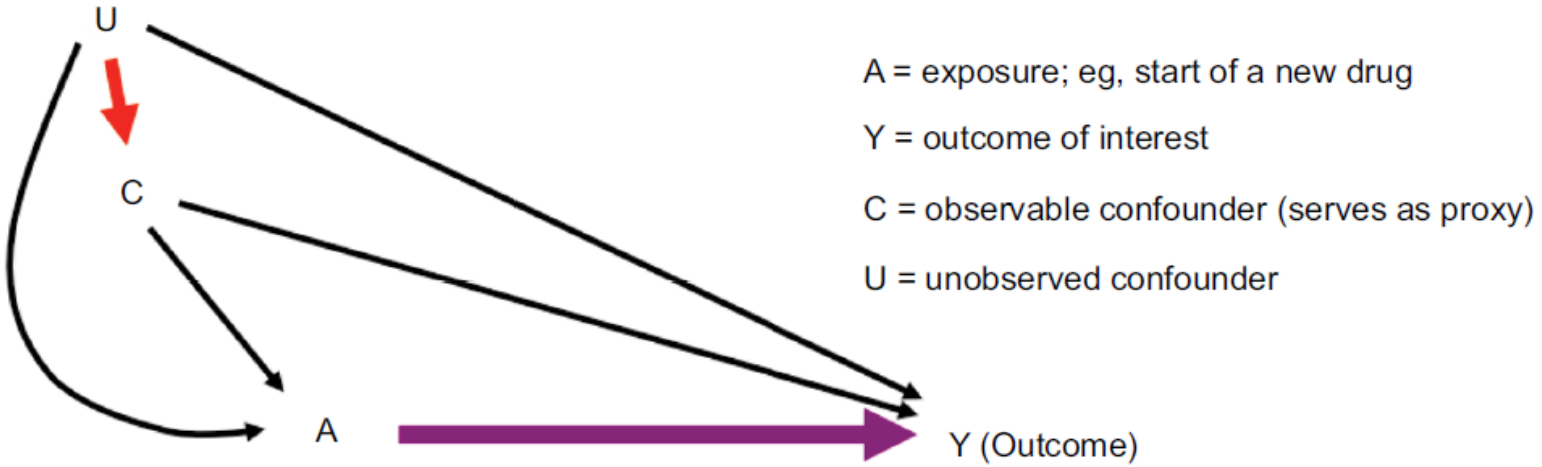
Background: Challenges for Confounding Control in RWE Studies

- Confounding arising from non-randomized treatment choices remains a fundamental challenge for extracting valid evidence to help guide treatment and regulatory decisions.

- Standard tools for confounding adjustment have typically relied on adjusting for a limited number of investigator specified variables.
  - Adjusting for investigator-specified variables alone is often inadequate
    - Some confounders are unknown at the time of drug approval
    - Many confounders are not directly measured in routine-care databases.

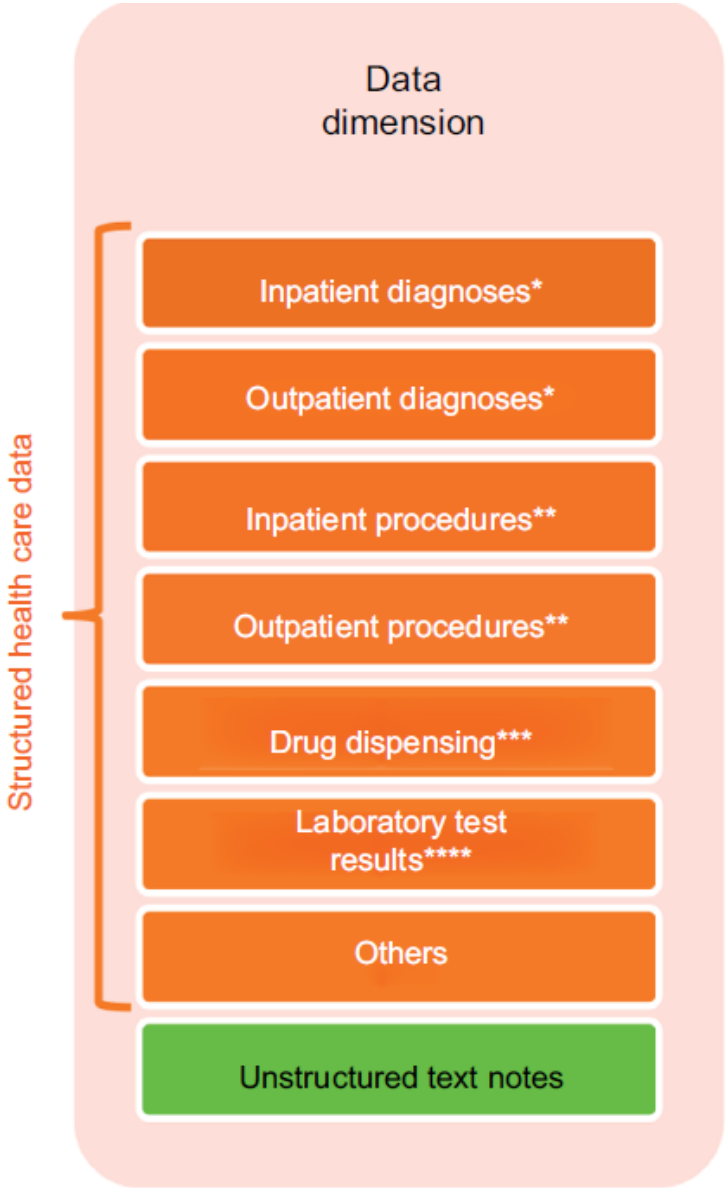
Background: Proxy Confounder Adjustment

- Healthcare databases may be understood and analyzed as a high-dimensional set of "proxy" factors that indirectly describe the health status of patients (Schneeweiss 2009, 2017).



Unobserved confounder	Observable proxy measurement	Coding examples
Very frail health	Use of oxygen canister	CPT-4
Sick but not critical	Code for hypertension during a hospital stay	ICD-9, ICD-10
Health-seeking behavior	Regular check-up visit; regular screening examinations	ICD-9, CPT-4, #PCP visits

Background: High-Dimensional Proxy Confounder Adjustment



How to generate/identify proxy variables?

- **Manual:** expert knowledge to identify claims codes that are thought to capture important confounder information.
  - **Limitations:**
    - difficult to know all confounders a priori
    - difficult to determine which combination of codes best capture information on those confounders
- **Kitchen sink approach:**
  - Balance all pre-treatment variables
    - Some dimension reduction is necessary in high-dimensional data
    - Harm efficiency and validity

Areas for high-dimensional proxy confounder adjustment

1. Feature generation
   - Transforming raw data into covariates (or features) to be used for proxy adjustment

2. Covariate prioritization, selection, and adjustment

3. Diagnostic assessment

Phase 1: Feature Generation

Feature Generation

- First challenge for proxy confounder adjustment is determining how to best leverage the full information content in healthcare databases
  - To generate features that best capture confounder information

- Range from very simple approaches that use binary indicators for each code, to approaches that first process information into a common data model format with common terminologies and coding schemes
  - Examples include OMOP Common Data Model maintained by OHDSI and PCORnet

- **Alternative strategy is to use data-driven approaches to generate features based on empirical associations and longitudinal coding patterns**
  - **More flexible since they can be independent of the coding system**

Feature Generation

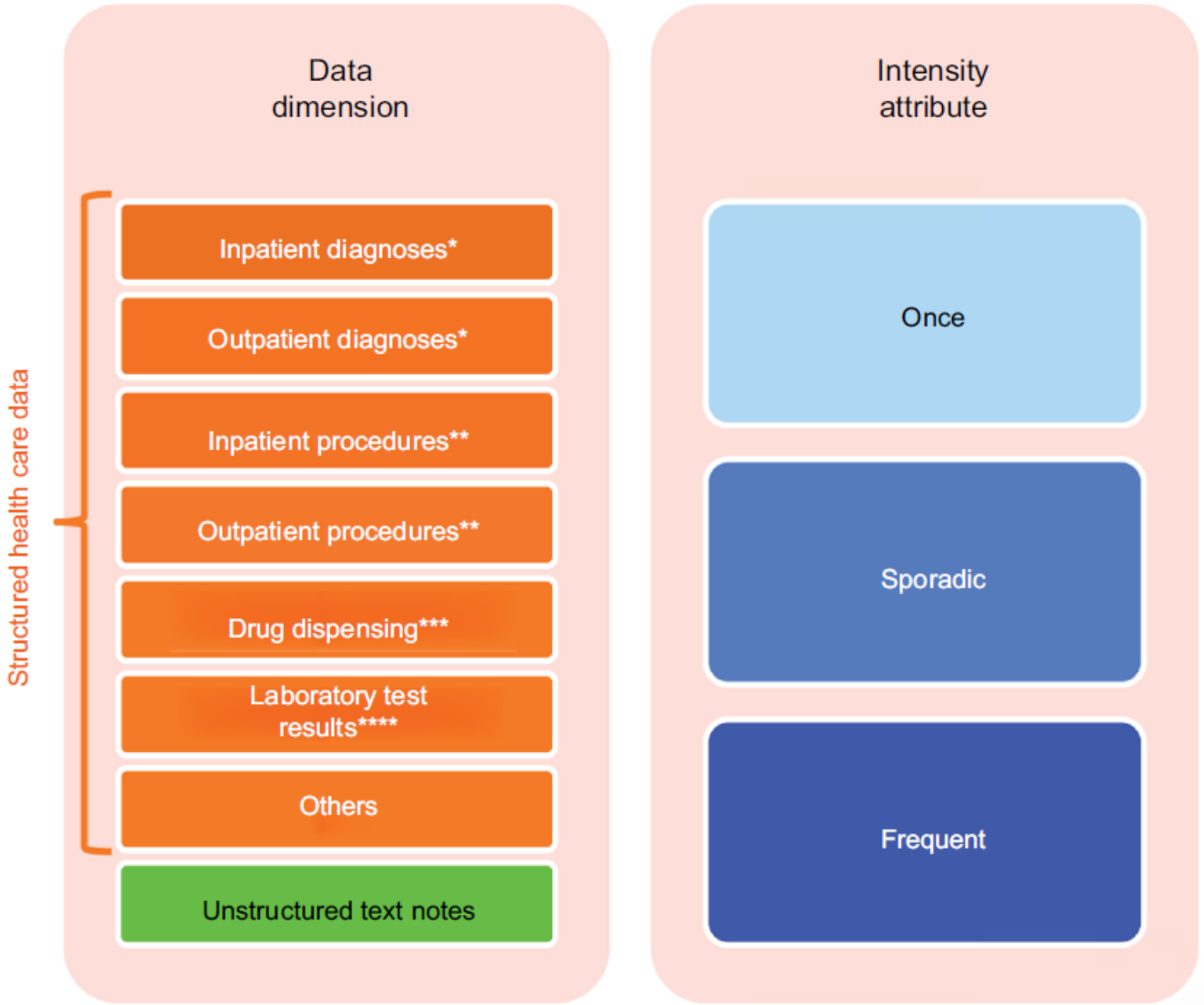
- Schneeweiss et al. (2009) first to propose automated proxy confounder adjustment in healthcare claims databases.

ORIGINAL ARTICLE

High-dimensional Propensity Score Adjustment in Studies of Treatment Effects Using Health Care Claims Data

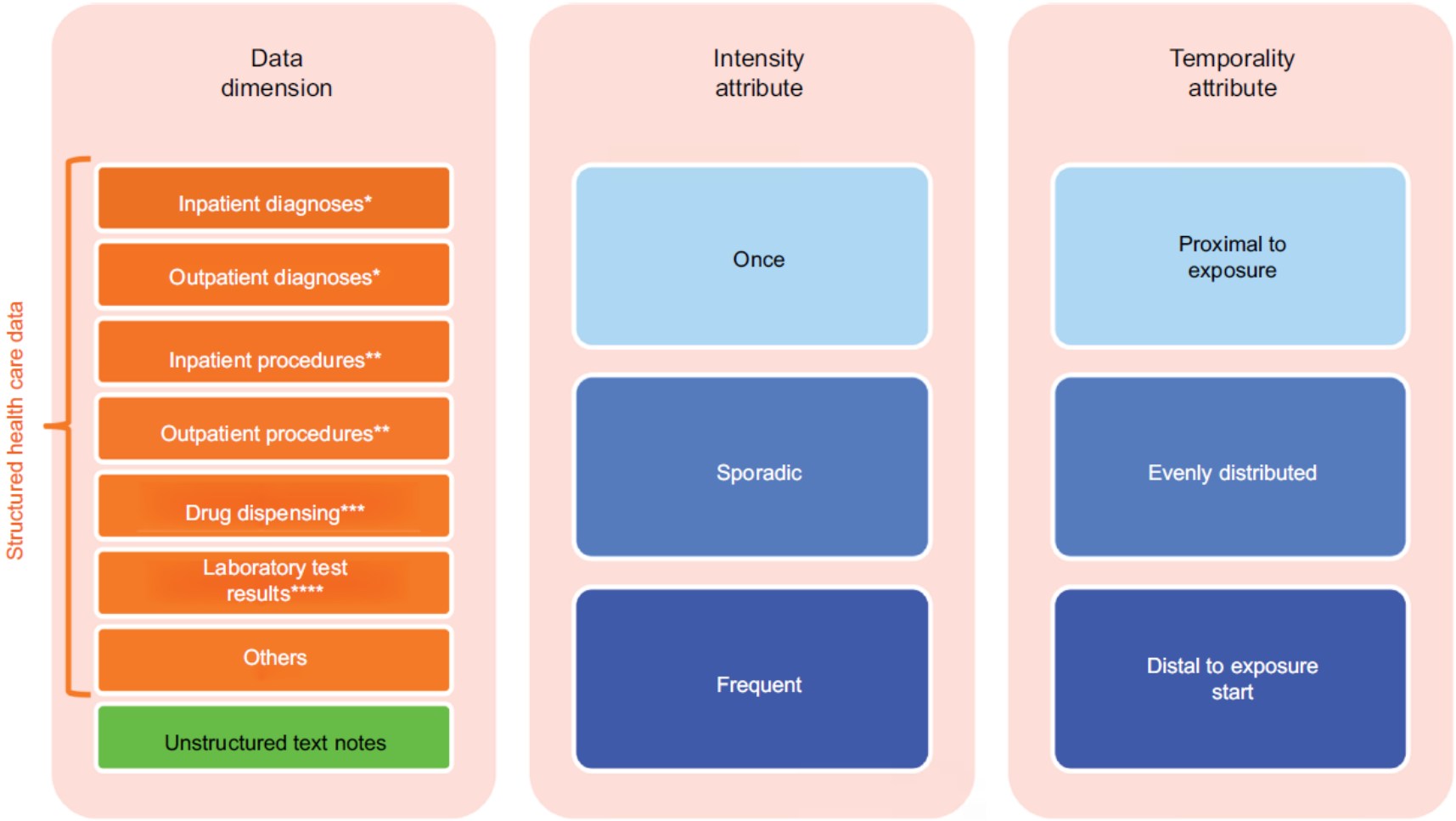
Sebastian Schneeweiss, Jeremy A. Rassen, Robert J. Glynn, Jerry Avorn, Helen Mogun, and M. Alan Brookhart

Feature Generation



- Evaluates thousands of diagnostic and procedural claims codes.
- For each code creates three binary variables based on the frequency of occurrence
- Evaluates potential confounding impact of generated variables by looking at strength of association with both the treatment and outcome

Feature Generation



- In theory, algorithms could consider more complex longitudinal coding patterns to capture additional information.
- For example, recent work has proposed using neural networks to model a patient's full course of care by taking into account temporality of code sequences.

Feature Generation

- There remain several limitations for large-scale feature generation for purposes of proxy confounder adjustment
  - Current approaches require data to be in a structured format.
    - How to best utilize NLP to generate features from unstructured electronic health records?

- Best practices for large-scale feature generation for proxy confounder adjustment remain unknown.
  - Still in early stages.
  - Relatively little has been written on this topic.

Phase 2: Covariate Prioritization, Selection, and Adjustment

Covariate Prioritization, Selection, and Adjustment

- HDPS has proven useful across a wide range of applications (>900 citations on google scholar).
  - **Limitations**:
    - How many variables to adjust for?
    - Evaluates empirical associations based only on univariate associations with treatment and outcome.

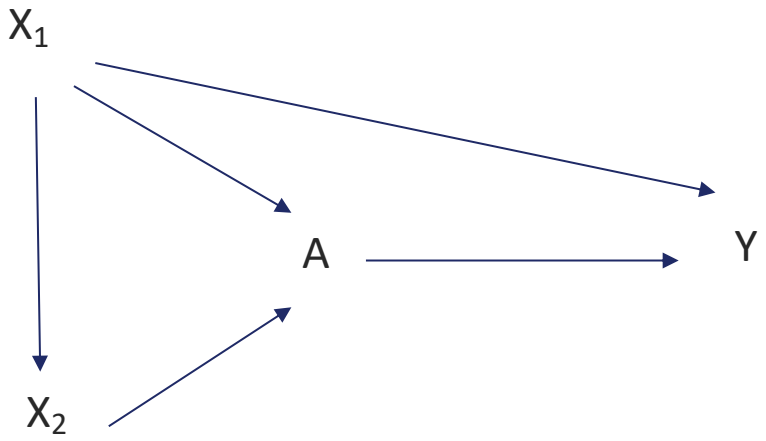


Figure. X_2 is independent of the outcome after conditioning on X_1

Covariate Prioritization, Selection, and Adjustment

- Recent developments in machine learning for causal inference have been developed that can potentially improve automated confounding control.
  - Consider joint correlation structure between variables

- Common theme is that these methods use ML prediction algorithms in a way to consider the joint associations of covariates with both the treatment and outcome simultaneously.
  - Targeted Learning (TMLE and CTMLE)
  - Many variations of regularized regression

Covariate Prioritization, Selection, and Adjustment

- Adjustment for proxy confounders: estimating nuisance functions
  - Recent work has shown that the flexibility of ML algorithms comes at a cost of slow convergence rates
    - Need big data
    - Without big data, can perform poorly when used within singly robust frameworks (i.e., propensity score methods or outcome regression)
  - **Doubly robust methods allow for slower convergence rates**
    - More work is needed on exploring the practical impact of slow convergence rates in settings relevant to real world evidence studies and high-dimensional confounding control.

Phase 3: Diagnostic Assessment

Diagnostic Assessment

- Prediction diagnostics
  - Out-of-sample prediction diagnostics most common way to evaluate ML model performance
    - e.g, cross validation
    - Discrimination (C-statistic)
    - Calibration plots

- **Previous work has shown that the use of prediction model diagnostics alone lead to suboptimal performance for causal inference.**

Diagnostic Assessment

- Diagnostics for Causal Inference:
  - Balancing
    - Challenging for high-dimensional confounding control (which variables to assess balance?)
    - Useful for PS analyses
    - Less useful for outcome-regression based approaches (G-computation, doubly robust analyses)
  - Estimand Diagnostics
    - Simulation-based approaches (simulate data structures reflecting complexity of RWD)
      - Plasmode simulation-based frameworks (use parametric bootstrap)
      - Wasserstein generative adversarial networks
      - Limitations
        - Can be difficult to closely mimic observed data (even on measured variables)
        - Can favor model choices made when simulating the data
    - Negative and Positive Controls
      - Gaining in popularity
      - Many different frameworks

Discussion

Discussion and conclusions

- There is a growing body of evidence showing that ML algorithms for high-dimensional proxy confounder adjustment can supplement background knowledge to improve covariate adjustment in healthcare database studies
- Three areas to consider for data-driven igh-dimensional proxy adjustment include:
  1. Feature generation
  2. Covariate prioritization, selection and adjustment
  3. Diagnostic assessment
- There is a large literature on methods for high-dimensional confounder prioritization/selection.
- Relatively little has been written on best practices for feature generation and diagnostic assessment.
  - Consequently, these areas have particular limitations and challenges when applying ML methods for high-dimensional proxy adjustment

Thanks to funding organization and collaborators

- Funding for review paper was provided by the *International Society for Pharmacoepidemiology* (ISPE) as part of the *Comparative Effectiveness Research Special Interest Group*.

- Coauthors:
  - Chen Yanover
  - Tal El-Hay
  - Dimitri Bennett
  - Robert W Platt
  - Andrew R Zullo
  - Grammati Sari
  - Xuerong Wen
  - Yizhou Ye
  - Hongbo Yuan
  - Mugdha Gokhale
  - Elisabetta Patorno
  - Kueiyu Joshua Lin

Thank you