Statistical methods for electronic health record data
International Conference on Health Policy Statistics

Jennifer F Bobb, R Yates Coley, Susan M Shortreed

Biostatistics Unit, Kaiser Permanente Washington Health Research Institute (KPWHRI)

jennifer.f.bobb@kp.org, rebecca.y.coley@kp.org, susan.m.shortreed@kp.org

Acknowledgements to entire Biostatistics Unit at KPWHRI

Course Outline

- Introduction to electronic health records (EHR) data for research purposes (20 minutes)
- Observational studies (30 minutes)
- Prediction studies (30 minutes)
- Pragmatic clinical trials (30 minutes)
- Questions (10 minutes)

Introduction to electronic health records for research purpose

What is EHR data?

- Electronic health record (EHR) data is data generated in the process of health care delivery
- EHRs initially developed for health care providers and insurers to manage healthcare billing (not to improve patient care)
- EHR adoption incentivized by 2009 HITECH Act for health care providers
- Two main sources of EHR data
    - Clinical records, also known as electronic medical record (EMR)
    - Health insurance claims data, also known as administrative claims

Health insurance, administrative claims data

- Generated when insurance billed for health-related services
- Includes information necessary for insurance claim
  - Diagnosis codes- International classification of disease (ICD) v9/10
  - Encounter of procedure codes- current procedural terminology (CPT)
  - Pharmacy fills for medications or supplies
- (Almost) all care observed, but not very much detail
- Unlikely to have information on severity of illness, lab results, vital signs, etc.
- Linked to particular health care coverage
  - Includes any paid services for health insurance enrollees while they are covered
- Some claims data sources available to researchers for purchase
  - Centers for Medicare and Medicaid Services (CMS)
  - Optum

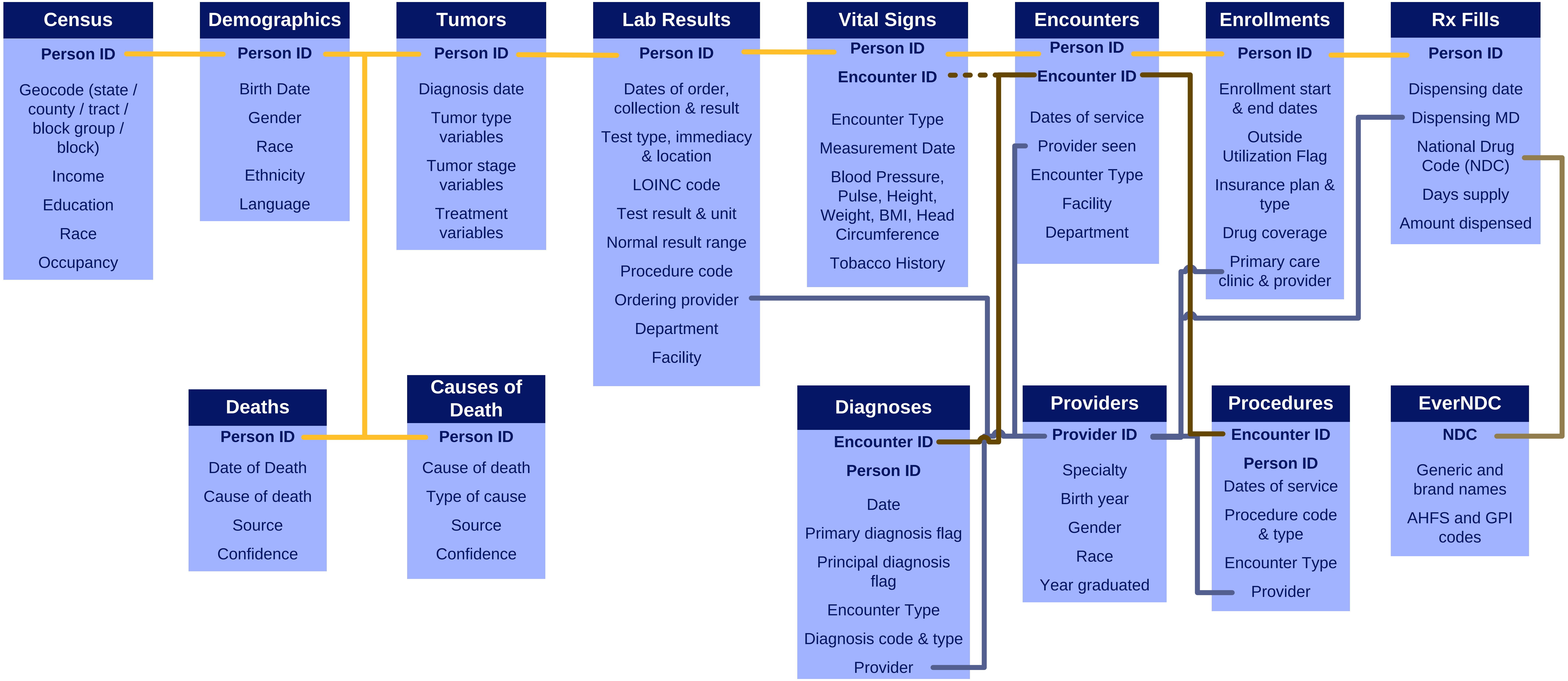
Clinical, EMR data

- Generated when patient seeks health care with a particular provider
- Rich source of information on all clinical care provided
  - Clinic visits/encounters, diagnoses, procedures (as in claims data)
  - Laboratory results, vital signs, imaging (not in claims data)
  - Pharmacy orders (vs. fills)
  - Structured text and embedded questionnaires, patient-reported outcomes (PROs)
  - Clinical notes, unstructured text (available for manual chart review or natural language processing [NLP], not covered in this course)
- Linked to a health system
  - Includes care provided in that health system (regardless of insurance coverage)
  - Care documented may be affected by health insurance reimbursement
  - Does not include care given outside of that health system

EHR data sources

Data Source	Strengths	Weaknesses
Claims data	• Easily defined patient population • Follow care for a patient across providers • Includes medication fills • Includes mortality status, (near) dates of death	• Doesn’t include details of care provided not associated with billing • Delay in data availability due to claims processing
Clinical data	• Data availability doesn’t depend on insurance coverage • Data available in (near) real time	• Doesn’t include care provided outside the health system • Challenging to define a patient population • Includes orders for medications, not fills • Only includes deaths that occur at health system facilities

EHR data are extensive and complex!



Census
Person ID
Geocode (state / county / tract / block group / block)
Income
Education
Race
Occupancy

Demographics
Person ID
Birth Date
Gender
Race
Ethnicity
Language

Tumors
Person ID
Diagnosis date
Tumor type variables
Tumor stage variables
Treatment variables

Lab Results
Person ID
Dates of order, collection & result
Test type, immediacy & location
LOINC code
Test result & unit
Normal result range
Procedure code
Ordering provider
Department
Facility

Vital Signs
Person ID
Encounter ID
Encounter Type
Measurement Date
Blood Pressure, Pulse, Height, Weight, BMI, Head Circumference
Tobacco History

Encounters
Person ID
Encounter ID
Dates of service
Provider seen
Encounter Type
Facility
Department

Enrollments
Person ID
Enrollment start & end dates
Outside Utilization Flag
Insurance plan & type
Drug coverage
Primary care clinic & provider

Rx Fills
Person ID
Dispensing date
Dispensing MD
National Drug Code (NDC)
Days supply
Amount dispensed

Deaths
Person ID
Date of Death
Cause of death
Source
Confidence

Causes of Death
Person ID
Cause of death
Type of cause
Source
Confidence

Diagnoses
Encounter ID
Person ID
Date
Primary diagnosis flag
Principal diagnosis flag
Encounter Type
Diagnosis code & type
Provider

Providers
Provider ID
Specialty
Birth year
Gender
Race
Year graduated

Procedures
Encounter ID
Person ID
Dates of service
Procedure code & type
Encounter Type
Provider

EverNDC
NDC
Generic and brand names
AHFS and GPI codes

Why do research with EHR data?

- Sample: Larger, more representative sample that one could easily (or affordably) obtain in a traditional study
    - Individual patient consent typically not required, but human subjects review still is!
    - More likely to include people of color, people who are not fluent in English, lower income patients
- More detailed comprehensive information on a patient’s healthcare than you could easily (or affordably) obtain in a study
    - With large health systems or claims data, information across specialties and providers
    - Patient history not susceptible to recall bias
- More complete, longer term capture of outcomes of interest (vs. relying on study visits)
- Structured data elements easily translated into quantitative inputs for analysis (e.g., presence of absence of exposure or event of interest)

Why do research with EHR data?

- EHR data can be used to supplement traditional research studies
- Example: NIA-funded longitudinal Adult Changes in Thought (ACT) cohort
    - In depth data collected biennially in cohort at in-person examinations and surveys
    - Linked to administrative data for rich studies on aging
    - For example, gold standard dementia assessment in 2022 can be associated with medication fills over prior decades
- EHR data can be used to assess generalizability of research findings

Challenges of research with EHR data

- Selection bias
- Measurement error, misclassification, and missing data
- Multilevel data
- Multi-site studies, Interoperability

Selection bias in defining study sample

- Insurance claims data limited to those covered by a particular insurance plan to define a population of patients for whom health utilization is completely observed
  - Eligibility criteria likely includes some minimum length of prior enrollment
  - Must balance selection bias with measurement error
- Clinical data studies in a health system must define population of patients for whom we expect to see most relevant care
  - Eligibility criteria typically based on prior utilization
- Health equity- patients who have the most barriers to care are less likely to have continuous insurance coverage, consistent care with the same providers; studies with EHR data will exclude these patients

Measurement error, misclassification, and missing data

- Will affect both covariates and outcomes
- Sources of measurement error/misclassification
  - Manually entered data may be incorrectly/imprecisely entered
    - Exacerbated by smart/predictive text or listing order for diagnosis codes
  - Lab values may have the wrong units indicated
  - Misdiagnosis
- Overlap between measurement error/misclassification and missing data:
  - If a “true” diagnosis doesn’t appear in the chart- provider may have incorrectly ruled it out or not assessed the patient for it (can’t distinguish between these)
  - Lack of prior enrollment or seeking care externally may cause missing, mismeasured data
- More straightforward to think of predictors from the EHR as, e.g., “diagnosis of X indicated in the medical record” instead of “patient has X condition”
- Health equity: EHR data reflects care **received**, not care **needed**. Populations with barriers to health care are less likely to have clinical needs accurately reflected in EHR data

Multilevel data

- Data are clustered within health system, clinic, provider, and patient
  - Clustering may be non-nested, e.g., if a patient sees multiple providers
- Cluster size may be informative
  - People who are more sick have more encounters with health system (exacerbated in EHR data vs. studies with scheduled visits)
  - Health equity: cluster size may also be associated with access to affordable, effective, and culturally competent care.
  - Marginal models for longitudinal data (including GEE) assume cluster size is non-informative, estimates average effect for average observation (not average effect for average observation in average cluster)

Multi-site studies, Interoperability

- Multi-site studies with EHR data can increase sample size, generalizability but pose many administrative, programming, and statistical challenges
- Administrative- Data use agreements, secure data transfer
- Programming- Interoperability
  - “If you’ve seen one EHR, you’ve seen one EHR”
  - Research networks with shared data models (e.g., PCORnet) facilitate data sharing
  - Health equity: Well-resourced health systems are more likely to have data infrastructure to support EHR studies, less likely to include safety net clinics (exception, OCHIN)
- Statistical- Comparability
  - Clinical documentation practices vary across site (ex, problem list vs. diagnosis code)
  - Measurement error and missing data patterns may also vary

Operationalizing clinical data for research

- Having clinical research partners is essential
    - Important to understand how data are coded in a particular health system
- Know underlying variables of interest before developing organizational definitions
    - Pharmacy data: EHR contains pharmacy orders and claims data contain pharmacy fills
        - Intervention to change physician behavior: use pharmacy orders
        - Identify individuals taking a particular medication: use pharmacy fills (picked up scripts)
- Know limitations of different data sources, for example: mortality
    - EHR only contains death happening in health care setting
        - Many people do not die in a health care setting
    - Health insurance companies can often determine death date but not cause of death
    - National death index covers the whole country but is quite delayed
        - State death records are generally for people who die in the state

Using clinical data to conduct observational studies

Outline

- Observational studies
  - Retrospective cohort
- Common “types” of bias and approaches to address bias in clinical database studies
  - Confounding
    - Treatment selection bias: differences between those observed to receive treatments
    - Unmeasured confounding: don’t have something you wish you had!
    - Indication bias: certain people have really low probability of treatment or outcome
  - Selection bias: who is in your sample and who is not
  - Measurement error: data you have may not accurately reflect the data you want
  - Informative observation times: Missing data is the complement of observed data
- Summary and conclusions

Observational studies

- No matter study type or data source, always several threats to validity in observational studies
- Can address (or minimize impact of) these threats by
  - **Design:** carefully select – study sample, comparison group, outcome and covariate definitions
  - **Analytic:** use math to address bias - e.g., weighted regression, outcome model adjustment
  - **Sensitivity analyses:** assess robustness of results to design and analytic decisions
    - How do things change if different decision had been made?
    - If an assumption is not valid, how different would things be?
  - **Use all three approaches!**
- There will always be limitations – be honest about what they are!
- **Retrospective cohort:** uses clinical data from visits, days, or years in the past
  - Just because data already collected doesn’t mean you should keep changing your study design
    - Think just as carefully about study design as you would for a prospective study

Designing a retrospective cohort

- Important to emulate a prospective study when designing a retrospective cohort study
    - Define eligibility at one point in time
        - Baseline covariates defined before or at this point
    - Start assessing exposure at time of eligibility
    - Assess outcome information going “forward” in time
        - Treat missing outcome information as missing data not as cohort exclusion criteria
            - Whenever possible… (example exception on the next slide)
- Approaches discussed in Hernan et al. valuable when designing retrospective cohort studies
    - Hernán MA, Robins JM. Using Big Data to Emulate a Target Trial When a Randomized Trial Is Not Available. *Am J Epidemiol*. 2016;183(8):758-764. doi:10.1093/aje/kwv254

Compare mom and baby outcomes of antihypertensive medications

- Retrospective cohort study of pregnant women with hypertension
  - Northern California, Southern California, and Washington regions of Kaiser Permanente
- Hypertension defined using diagnosis codes, medication fills, and blood pressures (BPs)
  - Prior studies used claims only; did not have BP values, pre-pregnancy weight, or expected due date
- Pregnancy is difficult to identify using EHR or claims data
  - Identified women who gave birth (live or stillbirth) and women who had termination after 20 weeks
    - Did not capture miscarriages or early terminations
- Compared labetalol (n=3017), methyldopa (1834), nifedipine (1105), and other beta-blockers (390)
  - Small for gestational age; preterm delivery; NICU admission; preeclampsia; stillbirth, maternal ICU
  - Inverse probability of treatment weights to account for confounding
- Results: adjusted prevalence for many outcomes similar across medication groups
  - Methyldopa had a lower rate of babies who were small for gestational age

Addressing treatment selection bias (i.e., confounding)

- Treatment selection bias (confounding): bias in treatment effect because of who was exposed to treatment compared to who was not
- **Design:** select comparator group that is similar to intervention group (except the exposure)
  - Definition of comparator groups affected by who is included in cohort
    - Collaborating with scientific partners essential
- **Analytic:** Use analytic methods to account for bias due to differences in who was observed to have the treatment and who was observed to be in the comparator group
  - Stratification, matching, adjustment for covariates
- **Sensitivity analyses:** how far off could we have been?
  - Design: Pick a primary comparison, redo analysis under different decisions, are they different?
    - Focus on point estimates not confidence intervals
  - Analytic: vary approach (set of covariates, form of adjustment)
- **Health equity:** given equal access to care, not everyone receives equal access to treatment
  - Examine characteristics of treated and untreated patients

Treatment selection bias

- Methods advancement in recent years for addressing treatment selection bias
  - **Propensity score methods** among the most popular, especially in EHR data
- EHR data can provide many, many covariates to adjust for in analyses
  - Propensity score (PS) is a summary measure: P(treatment | **X**)
    - Can help protect patient privacy, by sharing propensity score rather than all covariates
    - Popular ways to use propensity score to adjust for confounding: inverse probably weights, matching, outcome adjustment (must use flexible models), stratification
      - Everything but weighting, can be done directly with covariates (i.e., not propensity score)
    - Selecting which covariates to include in propensity score is important
      - Which variable included affects bias and statistical efficiency (i.e., big standard errors)
- We use both propensity score methods and non-propensity score methods
  - In rare outcome settings may not have degrees of freedom for outcome adjustment

Early pregnancy exposure to opioids and neural tube defects (NTDs)

- Health plan administrative data linked to birth certificate data, as part of the Medication Exposure in Pregnancy Risk Evaluation Program (MEPREP)
- Potential NTD outcomes identified via ICD codes and then validated by clinical experts
- Concern: women who use opioids long term too different from those not taking opioids?

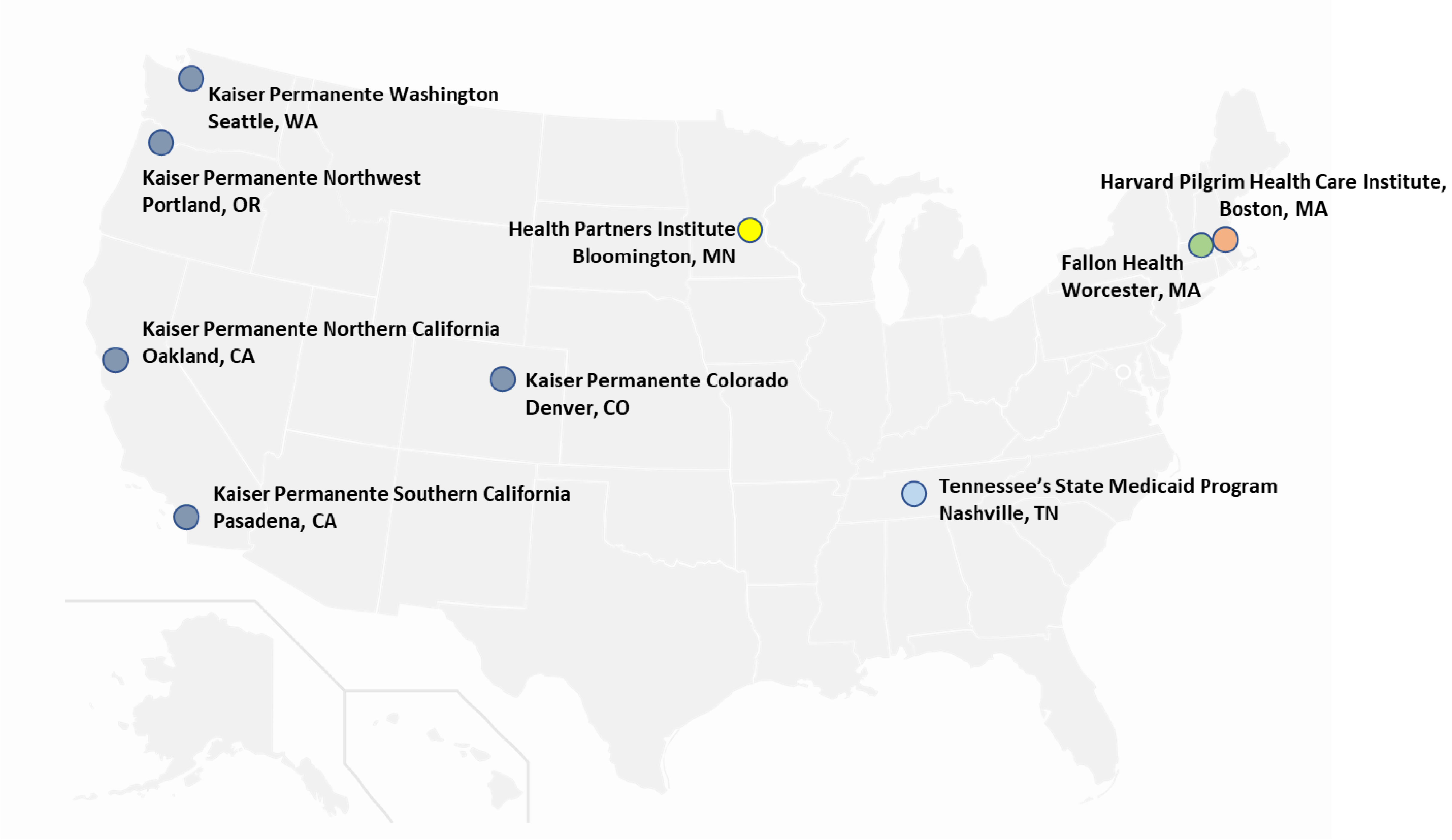
New user design:

925,533 eligible mom/baby pairs

19,072 pregnancies with "new" dispensing for opioid during critical exposure period (2.1%)

1:3 matched (maternal age, site)

57,219 unexposed pregnancies


Kaiser Permanente Washington
Seattle, WA
Kaiser Permanente Northwest
Portland, OR
Health Partners Institute
Bloomington, MN
Harvard Pilgrim Health Care Institute,
Boston, MA
Fallon Health
Worcester, MA
Kaiser Permanente Northern California
Oakland, CA
Kaiser Permanente Colorado
Denver, CO
Tennessee's State Medicaid Program
Nashville, TN
Kaiser Permanente Southern California
Pasadena, CA

Early pregnancy exposure to opioids and neural tube defects (NTDs)

Ultra rare outcome:

- Primary NTDs: just 11 cases in matched cohort
- Limited ability to adjust for potential confounders in outcome regression

Approach to address treatment selection bias:

- Estimated propensity score using logistic regression
- Inverse probability of treatment weighting (IPTW)
  - Used weights for "average treatment effect among the treated" (ATT)
- Weighted logistic regression to estimate odds ratios

Results of primary analysis non-statistically significant

- Odds ratio 3.0 (95% CI: 0.7, 12.7)

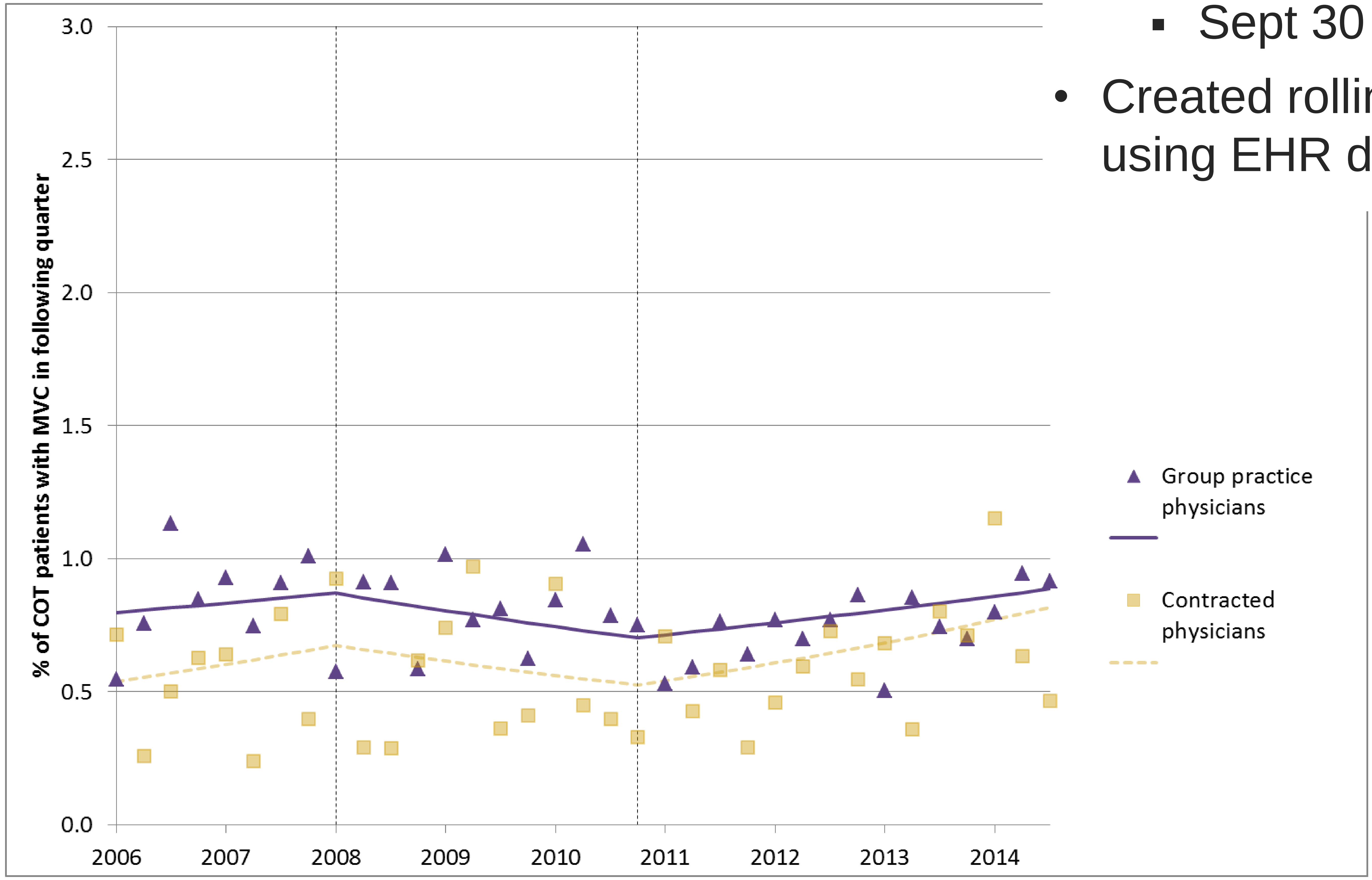
Covariates included in the propensity score	Categories of covariate
Site	each site (9 total)
Maternal age	15-19, 20-24, 25-29, 30-34, 35-39, 40-49
Calendar year of delivery	each year of study (14 total)
Maternal race and ethnicity	Hispanic, White, Asian, Black, other
Maternal education	≤ 12 years, > 12 years
Medicaid insurance	yes, no
Pre-gestational diabetes	yes, no
Parity	nulliparous, parous
Nitrosatable medication use	yes, no
Alcohol use disorder	yes, no

Interrupted time series – a type of historical control

- **Concurrent control** is most common comparison group
- **Historical controls** can be confounded by time
  - Can use both concurrent and historical control in same study!
    - Use concurrent control to account for changes over time unrelated to intervention
      - Think carefully in selecting concurrent control group
      - Adjust for potential confounders
- Use regression “discontinuity” approaches to model changes over time
  - Potential inflection points when interventions begin (or expected effect of)
  - Smooth or non-smooth changes
    - Health systems are usually giant ships and tend not to change on a dime
    - But EHR based interventions can be turned on with flip of a switch!

Did risk reduction initiative for patients on long-term opioid therapy (LTOT) reduce motor vehicle crashes?

- Regression discontinuity to assess impact of interventions
  - Jan 1 2008: intervention to alter prescribing behavior
  - Sept 30 2010: multifaceted risk reduction intervention
- Created rolling (quarterly) entry retrospective cohort study using EHR data, 2006-2014



- Identified people LTOT in each quarter
- Assessed if motor vehicle crash occurred within 90 days
  - Linked state motor vehicle crash registry to EHR data
- Concurrent control group: patients using LTOT seeing Group Health's contracted providers
  - Intervention only provided to Group Health clinicians

Unmeasured confounding

- **Design:**
  - Select study sample and comparator to reduce unobserved confounding!
    - Example: opioids and NTD study focused on new users of opioids and excluded prevalent users
  - Two-phase sampling collects more data on subset of observations (usually variables that are expensive to collect)
    - Can stratify sampling of subset for additional data collection (e.g., on exposure or outcome)
  - Can link to external data or conduct additional data collection to augment what is available in EHR

- **Analytic:**
  - Use analytic methods that aim to adjust for unobserved confounding
  - Instrumental variable (IV) analysis: IV only associated with outcome through exposure (can be hard to assess!)
    - Geographic proximity or provider preference are common IVs in clinical research

Unmeasured confounding

- **Sensitivity analysis:** Evaluate how sensitive effect estimates are to unobserved confounding
    - E-value: minimum association between unmeasured confounder and outcome needed to “explain away” estimated effect.
        - Larger E-value means effect estimate less likely due to unobserved confounding
    - Negative control: evaluate whether exposure has estimated impact on unrelated outcome
        - Can be helpful to assess if treatment appears protective because those with treatment are healthier
        - Alternatively, can evaluate whether exposure at unrelated time impacts outcome
            - Example: opioids and NTD study also examined association with 2nd and 3rd trimester exposure (after neural tubes already formed)

- **Health equity:** Clinical data rarely contains data on social determinants of health in a comprehensive way

Unmeasured confounding sensitivity analyses: bariatric surgery and macrovascular disease

- **Scientific question:** Does bariatric surgery reduce risk of macrovascular disease in patients with Type 2 diabetes and severe obesity compared to non-surgical treatment?
- **Challenge:** People who choose to have surgery are likely different than those who do not, and those differences may not be captured in the EHR
- Primary analysis found **lower risk of macrovascular disease for patients with bariatric surgery**, hazard ratio = 0.60 (95% CI: 0.42-0.86)
- E-value analysis done as a **sensitivity analysis**
  - Unobserved confounder would need to have a relative risk of 2.72 to overcome observed association (hazard ratio=0.60) between surgery and macrovascular disease
    - Conducted E-value calculations for upper bound of confidence interval
      - A relative risk of 1.6 would be needed to overcome an association of HR=0.86

Indication bias

- An **extreme case of treatment selection bias**
    - Happens when the clinical indication for selecting a treatment (severity of the illness) also affects the outcome
        - Patients with more severe illness more likely to receive treatment and therefore could appear to have poorer outcomes
- **Some approaches** to address indication bias
    - Smart comparator group definition (example on next slide)
    - Studying effects of medications: restrict to those with indication for the medication
        - Rarely can get clear indication from actual medication prescription, but can approximate in other ways
    - Self-control approach (historical comparison approach)
        - Need short-term outcomes (short follow-up time) for this study design to work

Antidepressant use in pregnancy and risk of gestational diabetes

- Prior studies show an increased risk, but may be affected by confounding by indication
  - Depression is associated with adverse health outcomes
  - Women taking antidepressants may be at higher risk than those not taking antidepressants
- **Cohort:** Kaiser Permanente Washington women with singleton birth between 2001 and 2004
  - Pregnancies associated with more than one baby at higher risk for gestational diabetes
- Restricted cohort to women taking antidepressant prior to pregnancy
  - More than one antidepressant fill at least 6 months prior to pregnancy
- Women with antidepressant fill during pregnancy referred to as continuers (n=1634)
  - Comparison group: those with no fills during pregnancy (n=1211, discontinuers)
  - Used inverse probability of treatment weights to control for confounding
- Relative risk comparing continuers to discontinuers: 1.10 95% CI: (0.84,1.44)

Selection bias

- **Designing the cohort:** who gets into your sample and how they get in is important
  - Patients included in the analysis could differ from the target population of interest
  - Retrospective studies can collect (minimal) information on people who are not in main cohort
    - Can assess how they are different
  - Keep in mind scientific question and power when defining the sample
    - A factor in defining sample for observational studies comes from requiring prior enrollment
      - Measurement error versus selection bias (see intro slides)
- **Analytic strategies** can be used to assess and account for selection bias (by measured variables)
  - E.g., can use weighting to weight sample back to target population of interest
  - Selection bias always exists in research studies; can use EHR data to assess selection bias in traditional studies too
- Can do **sensitivity analyses** on defining population, but it can get thorny quickly

Accounting for non-response in the Middle-Aged/Seniors Chronic Opioid Therapy (MASCOT) survey.

- Can **use EHR data to assess and account for non-response**
- Of 2489 people eligible and contacted, 1289 (51.8%) completed baseline interview
  - Assessed predictors of survey response
    - General patient characteristics: age, sex, race, ethnicity, Charlson
    - Opioid misuse risk factors: tobacco use, substance use disorders, mood/anxiety disorders, having excess opioids on hand due to early refills, receiving opioids from more than 3 prescribers, high opioid dose, co-prescribed sedative
  - **Older people were more likely to respond:** 51.7 % response rate for patients between 45 and 55 years old and 59.9% response rate for patients 75 years and older
    - Other than age **very few things related to survey non-response**!
- Increased confidence in study results
- (Similar analyses can also be used to assess generalizability of trial results)

Measurement error

- **Design**
  - Does EHR capture exposure, outcome, and relevant covariate information?
    - E.g., EHR data not a good source for over-the-counter medications
  - Validation studies to assess properties of measures (e.g., sensitivity, specificity) are key
    - Two-phase study designs can be used to efficiently combine error-prone data on full cohort, with more precise "gold standard" data on a subset
- **Analytic** approaches to address measurement error
  - Regression calibration
  - Can use multiple imputation or complex algorithms (references in course handout)
- **Sensitivity analyses**
  - Exposure misclassification example: require two or more fills of medication to be exposed
    - Addresses folks picking up meds but not taking them
- **Health equity:** less likely to have accurate information on people with short enrollment histories

Two-phase study examining association between elective induction of labor and pregnancy outcomes

- **Exposure groups:** Among women who were pregnant and had no indication of labor through week X of pregnancy (X = 38, 39, 40),
  - Electively induction (EI) of labor vs. “Expectant management” (i.e., pregnancy continues)
- **Measurement error:** Challenging to define a “phenotype” of EI using automated data – requires *absence* of indications; pregnancy outcomes also error-prone
- **Two phase study design:**

  Phase 1: cohort of eligible pregnancies identified via electronic/automated data

  Phase 2: stratified sample selected (oversampling those with apparent EI and outcomes)

  - Obtained “gold standard” measures of EI, mom and baby outcomes, and some confounders (e.g., smoking status) using chart review
- **Statistical analysis**: semiparametric maximum likelihood to efficiently use all data while accounting for outcome-dependent sampling
- **Outcomes differed by gestational age:** EI at 39 weeks was associated with cesarean delivery (prevalence 41% vs. 28%; OR 1.77 [1.14–2.81]).

Informative observation times

- How often someone accesses health care is tied to many things
    - Health care coverage complicated (e.g., co-pays, deductibles, visit limits, referrals, etc)
        - Barriers to health care beyond financial (e.g., geographical location, cultural competency, etc)
    - **“Sicker” patients tend to receive more care**
        - Have more visits and more chances for other diagnoses and information recorded
    - **Informative cluster size** may have an impact on study results
        - Must think about this in addition to accounting for correlation
- Informative observation times may impact studies addressing some scientific questions but not all
    - If modelling longitudinal trajectories, think carefully about when measurements occur
        - Especially if there are **differences in visit patterns between comparison groups**
- Health equity: health systems and culturally sensitive care can impact follow-up care

Does receipt of alcohol-related care differ for patients with HIV?

- Veterans Health Administration (VA) implemented annual alcohol screening in 2004
  - Evidence-based brief interventions for unhealthy alcohol in 2007
  - Alcohol use disorders identification test consumption (AUDIT-C)
    - Validated measure of alcohol use; can be used to identify unhealthy alcohol use
- **Goal:** compare rates of brief intervention in 12 months following AUDIT-C for patients living with HIV and patients not living with HIV
- Used data extracted from VA research data warehouse from 10/1/2009 through 5/30/2013
  - Identified all AUDIT-C with responses that indicated unhealthy alcohol use
    - Some individuals had one AUDIT-C a year; some had many AUDIT-Cs in a year
  - Primary analysis used multiple AUDIT-Cs per person (at least 9 months between AUDIT-C)
    - Adjusted relative risk=0.83 (0.80, 0.85)
  - Sensitivity analyses selected one AUDIT-C at random
    - Adjusted relative risk =0.86 (0.85, 0.88)

Summary and conclusions

- Retrospective cohort studies must be constructed with care just like prospective studies
  - Having access to lots of data can make it easy to say: let’s just to a sensitivity analyses
    - Keep track and check back before you finalize analyses
      - Might end up doing a lot of analyses, interpreting results can be troublesome
- Threats to validity in all studies
  - Rigorous studies use both design and analytic strategies to address bias
  - Targeted sensitivity analyses can strengthen study results

Predictive analytics with EHR data

Yates Coley, PhD

What are the uses of predictive analytics with EHR data?

- Identify patients with elevated risk of adverse outcome, intervene
- Target health system resources to those who would most benefit from an intervention
- Guide decision-making about likely benefits and harms of treatment options

- Used in many clinical settings
  - Inpatient hospitalization events for sepsis, rapid deterioration
  - Opportunities to intervene at outpatient visits for prevention, early detection
  - Additional outreach to patients by mail, phone, or online

Why do prediction *with* EHR data?

- Data in clinical records and claims data reflects information that would be available to guide clinical treatment outside of a study setting
- Using EHR data makes it easier to incorporate prediction models back “into” the EHR so that they can be integrated into clinical workflow and used in real-time
- While providers have access to information in the EHR, risk factors may be complex;
    - Prediction models can estimate the relationships between complex predictors and outcomes
    - Integration in the EHR enables quick calculation, summary of risk
    - Prediction models can also inform targeted outreach outside of in-person encounters
- For all of the same reasons you use EHR data for trials and observational studies (see introduction)

Prediction vs. Inference

- The goal of prediction studies is different than inferential studies, and this will have an impact throughout the process
- Goal of prediction study: predict risk of outcome with accuracy, correctly discriminate (order) risk
  - Not estimating association or causal effects of predictors (and you shouldn’t try to!)
  - Not estimating variability in estimated coefficients (for e.g., logistic regression)
- Goal of inference: estimate association (or effect) of a treatment or exposure with outcome.
  - Need to estimate variability to construct confidence intervals, perform hypothesis tests
- Must keep expected end use of prediction model in mind throughout modeling process

Prediction vs. Inference

- There is some use of inference in prediction studies
  - Assessing variability in prediction model <u>performance</u>
  - Comparing performance of candidate models (choose which one is better)
- Statistical power in prediction
  - Power to identify risk factors when estimating prediction model (analogous to Type II error). If a variable is associated with the outcome, you want that reflected in your prediction model.
  - Power to quantify performance of prediction model with precision

Case study: Predicting suicide risk

- Goal: Predict risk of suicide attempt or death following an outpatient mental health visit
- If we can identify people at elevated risk for suicide attempt at time of outpatient mental health visit, providers can take additional steps to evaluate and prevent suicide
  - Columbia Suicide Severity Rating Scale
  - Safety plan (e.g., securing firearms, removing means)
  - Interventions already recommended for patients reporting suicidal ideation on the 9-item patient health questionnaire (PHQ-9), but not all patients complete a PHQ-9 at every visit

- Simon, Johnson, … Shortreed (2018) “Predicting suicide attempts and suicide deaths following outpatient visits using electronic health records” *American Journal of Psychiatry*

Defining sampling framework, prediction instance

- Want data for estimating and validating prediction model to reflect decision-making (or other action) that prediction model will inform
- Unit for prediction: person- or encounter-level?
- Timeline- when is prediction being made, and what is length of follow-up?

- Case Study: 90-day suicide risk following outpatient mental health visits
  - Population is all outpatient mental health visits (people can have multiple visits)
    - 14 million visits by patients 13 and older at 7 health systems 2009-2017
    - Need to know when someone is at risk, not just who; risk changes over time
  - Prediction made at time of visit
  - 90-day outcome window (binary, with accommodations for censoring if needed)

Defining sampling framework, prediction instance

- Cohort sample – most straightforward and common for prediction
  - Representative of population of prediction instances
  - Predictions reflect population-level averages (e.g., absolute rates for binary outcome)
  - What we used for suicide risk prediction
- Case-control sample – sometimes required due to data collection or computing restraints
  - Selecting controls is tricky, must think about population represented
  - Need to adjust prediction to get population-level averages or rates
  - Some machine learning methods developed with “balanced” data in mind (e.g., “majority vote” for classification), but plenty of options for unbalanced data

Defining predictors- Timing of data availability

- Consider what data will be available to you at the time of prediction
  - Inference vs. prediction.
  - With inference, prioritize having complete baseline data, won’t use most recent data if they are not complete
  - With prediction, have to consider prospective use
- Health insurance claims data typically delayed 1-3 months
- Some providers won’t have access to claims data ever, so are limited to clinical data
  - Meanwhile, health insurance providers will not have clinical variables
- Patient-reported outcomes (PROs) may not be available in EHR in real time (if, for e.g., patients complete paper forms)

Defining predictors- Missing data and measurement error

- Both missing data and measurement error can be challenging to conceptualize in EHR data (see introduction slides)
- Prediction vs. inference
  - Inference- want to make accurate inference on parameter, important to model missing data and measurement error correctly
  - Prediction- want to make accurate predictions with data as they are available at the time of prediction
- Both missing data and measurement error will deteriorate prediction model performance, but you want to estimate prediction model with data as it will be observed in the future (when the model is being used)
  - Statistical methods for measurement error and missing data (e.g., multiple imputation) are not practical for generating new predictions in real time
  - Concerns about inference when you adjust for a “missing” category don’t apply to prediction

Defining outcomes- missingness and measurement error

- Outcomes may be missing or misclassified
- In contrast to predictors, we do want to adjust for missingness/measurement error (because you don’t need outcome information at the time of prediction, it doesn’t affect prospective use)
- Measurement error may be “differential”, that is, related to predictors. e.g., missed diagnoses may be more likely for patient with fewer risk factors
- Measure of how well prediction model estimates error-prone outcome /= how well model predicts true outcome
  - Can adjust for misclassification when estimating model performance (Wang et al 2016)
- Missingness/misclassification may degrade prediction model performance for “true” outcomes
  - Severity depends on whether misclassification is differential or non-differential

Case study: Predicting suicide risk, defining predictors

- 149 predictors include:
  - Demographics (age, sex, race, ethnicity, insurance type)
  - Comorbidities
  - Mental health diagnoses, medication fills, encounters with mental health diagnosis (inpatient, emergency department, or outpatient) in prior 5 years
  - Prior suicide attempts in 5 years prior to visit
  - PHQ-9 for patient reported depression symptoms, suicidal ideation, over prior year
- “Missing” categories used when information not available (e.g., race)
  - Not best practice for causal inference, but this is prediction! Don’t want to exclude visits with missing data
- Many correlated predictors (again, not great for inference, so shouldn’t try to interpret)
- Currently conducting analyses to assess impact of lag in health insurance claims data

Case study: Predicting suicide risk, defining outcomes

- Outcomes captured in the 90 days following the visit
  - Suicide death, state mortality data indicating death from self-inflicted injury or injury or poisoning with undetermined intent
  - Suicide attempt, diagnosis of injury or poisoning with ICD code indicating self-harm or undetermined intent or suicide death
- Event rate of 60 suicide attempts and 2 suicide deaths per 10,000 visits
- Suicide outcomes may be missing or misclassified
  - Person may be treated for self-harm in emergency department, but it is diagnosed as accidental
  - Person may not present for treatment for self-harm
  - Timing of self-harm could be mis-recorded (e.g., provider indicates diagnosis of self-harm in medical record when patient reports it later, not when it happened)
  - Patient may disenroll form health plan within 90 days, censoring follow-up
  - Ongoing analyses to evaluate impact of misclassification using misclassification rates obtained from manual chart review

Prediction model estimation

- Estimate the “best” prediction model you can!
- May require choosing between models (e.g., logistic regression vs. random forest)
  - Many, many options for prediction modeling
  - In addition to performance, consider the importance of interpretability and ease of implementation into the EHR. Logistic regression may preferred (and perform just as well as machine learning!)
  - Ensemble methods (e.g., Super Learner, van der Laan, Polley, and Hubbard 2007) combine predictions from multiple models- good statistical properties but also complex to implement
  - Recommend Hastie, Tibshirani, and Friedman *Elements of Statistical Learning*; James, Witten, Hastie, Tibshirani *Introduction to Statistical Learning*
- May require selecting tuning parameters, typically with cross-validation
- Suicide prediction case study:
  - Logistic regression with LASSO (Least absolute shrinkage selection operator, Tibshirani 1996)
  - Random forest, artificial neural networks have not shown meaningful improvement

Prediction model validation

- Evaluate the performance of the prediction model you've selected
  - **Validation** is used to establish prediction model performance for observations outside of the study sample (Altman and Royston 2000)
  - **Internal validation** aims to evaluate model performance in new observations from the same underlying population using the same data available for model validation
    - This is may also be used to inform choice between candidate models (above)
    - Split-sample or cross-validation
  - **External validation** estimates model performance in observations outside of the population represented in the study sample, a.k.a. generalizability
  - Important to select a validation method that reflects the setting in which you plan on using the prediction model

Measures of prediction model performance

- Evaluate model performance by comparing predictions to observed outcomes
- **Focus on prediction measures that reflect how model will be used**
- Measures of accuracy for a classification rule
  - For continuous predictors, define classification at threshold, e.g., 95th risk percentile
  - Sensitivity (a.k.a. true positive rate, recall)
  - Specificity (1- false positive rate)
  - Positive predictive value (PPV, a.k.a. precision)
  - Negative predictive value (NPV)
  - PPV and NPV depend on prevalence of outcome in population
  - F-score, function of sensitivity and PPV

Measures of prediction model performance

(continued…)

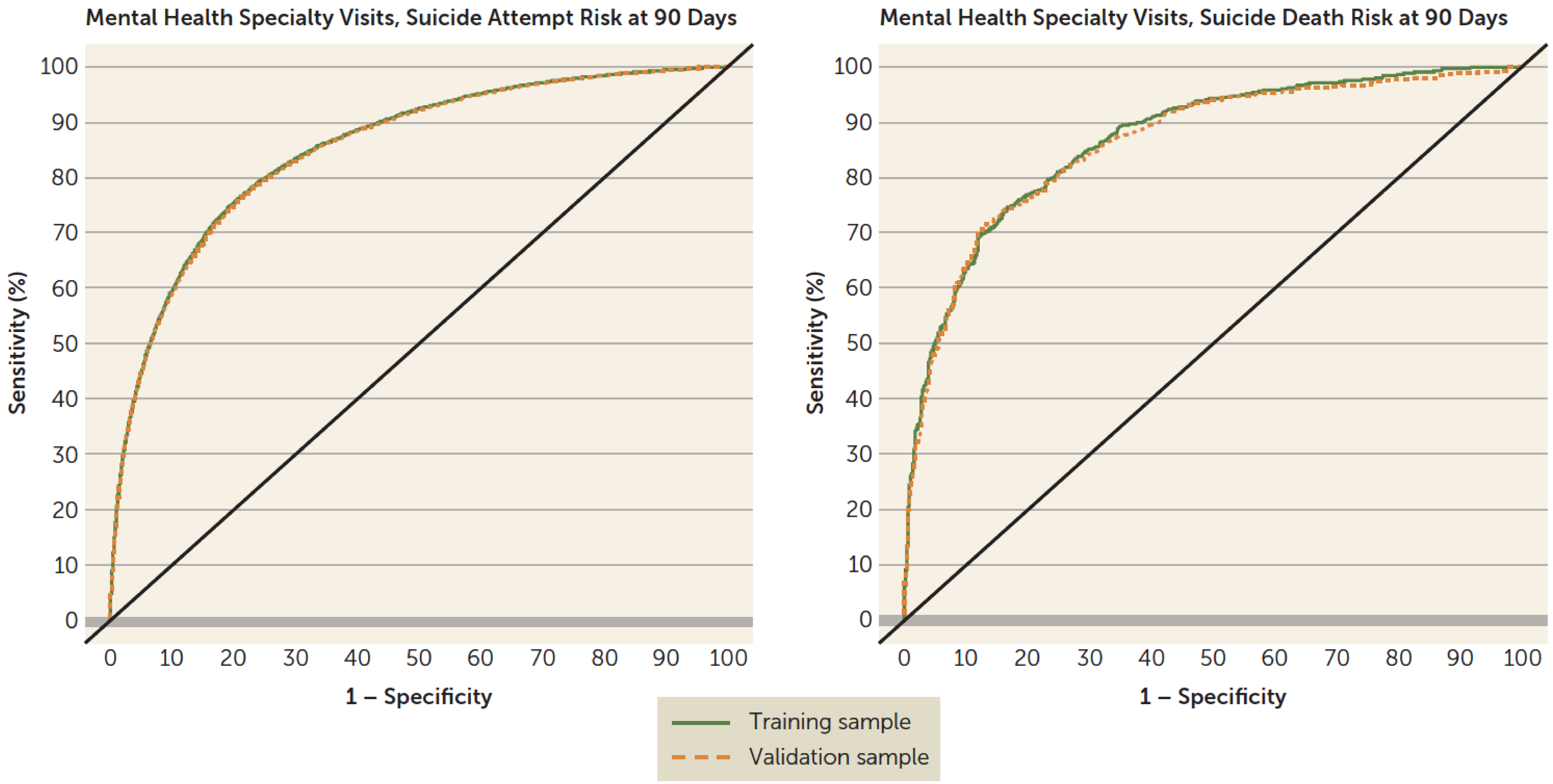
- AUC: Area under the receiver operating characteristic (ROC) curve (Hanley and McNeil 1983)
  - ROC curve plots Sensitivity vs. 1-Specificity for all unique thresholds of continuous risk score
  - Measures “discrimination”, how well risk ordered across entire sample
  - Influenced by distribution of predictors in population (more heterogeneity – higher AUC)
  - Sensitive to small changes in classification accuracy for small number of observed events
- Calibration (Steyerberg et al 2010)
  - Agreement between predicted and observed event rates
  - Calibration frequently doesn’t matter in classification problems, and prediction models can always be recalibrated when accuracy of continuous scores is important (e.g., propensity methods)
- Brier score (equivalent to mean squared error for unidimensional problems)
- Many others…

Case study: Predicting suicide risk, estimating and evaluating model

- First analysis presented in Simon et al (2018) for predicting suicide attempts and suicide death with logistic regression models
- Data randomly divided in to 65% training and 35% testing set
- 10-fold cross-validation used within training set to select shrinkage parameter for LASSO
  - Selected to optimize Bayesian information criterion (BIC)
  - 94 predictors selected for suicide attempt prediction model
  - 43 predictors selected for suicide death prediction model
- Performance measured in testing set
  - AUC= 0.851 (95% CI: 0.848, 0.853) for suicide attempt
  - AUC= 0.861 (95% CIL 0.848, 0.875) for suicide death
  - Also examined ROC curves, classification accuracy, calibration

Case study: Predicting suicide risk (logistic regression results)

FIGURE 1. Receiver Operating Characteristic Curves Illustrating Model Performance in the Validation Data Set for Prediction of Suicide Attempts and Suicide Deaths Within 90 Days of Visit in Seven Health Systems, 2009–2015[a]



Case study: Predicting suicide risk (logistic regression results)

FIGURE 1. Receiver Operating Characteristic Curves Illustrating Model Performance in the Validation Data Set for Prediction of Suicide Attempts and Suicide Deaths Within 90 Days of Visit in Seven Health Systems, 2009–2015[a]

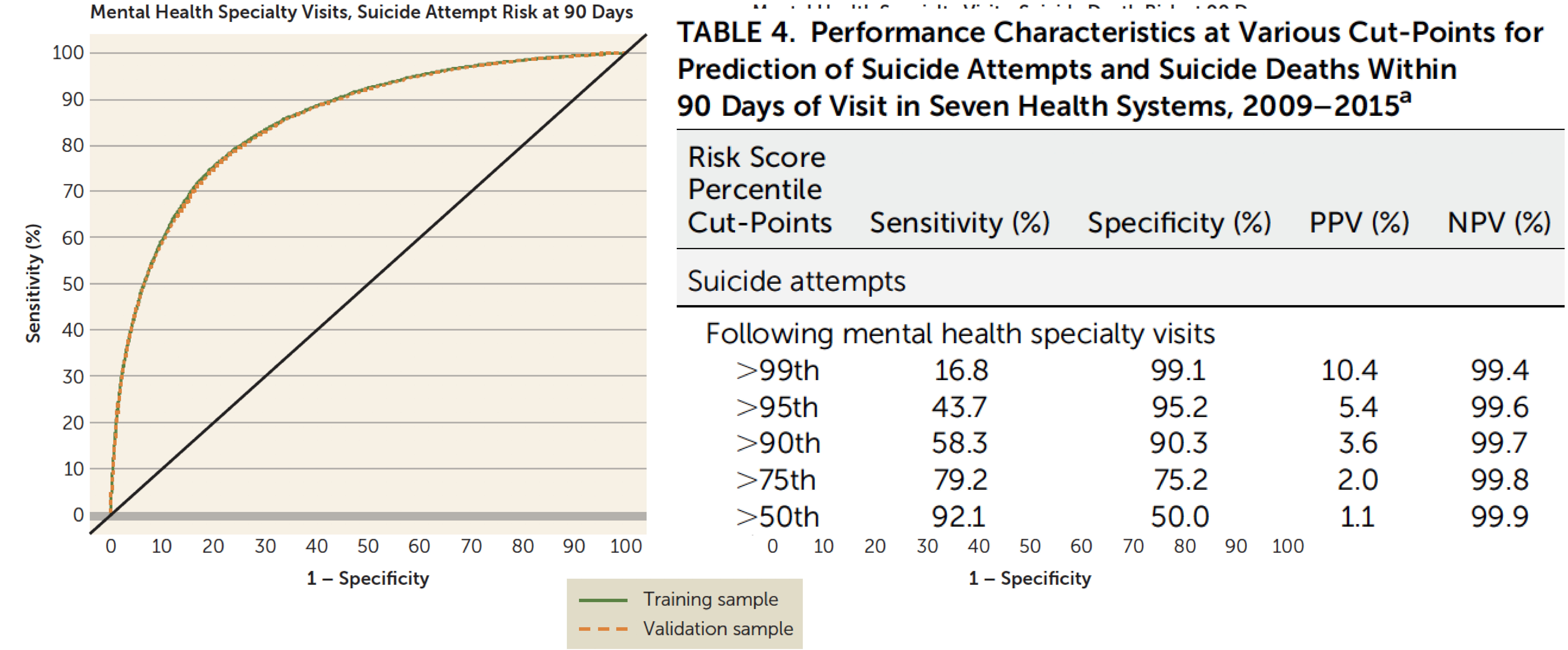
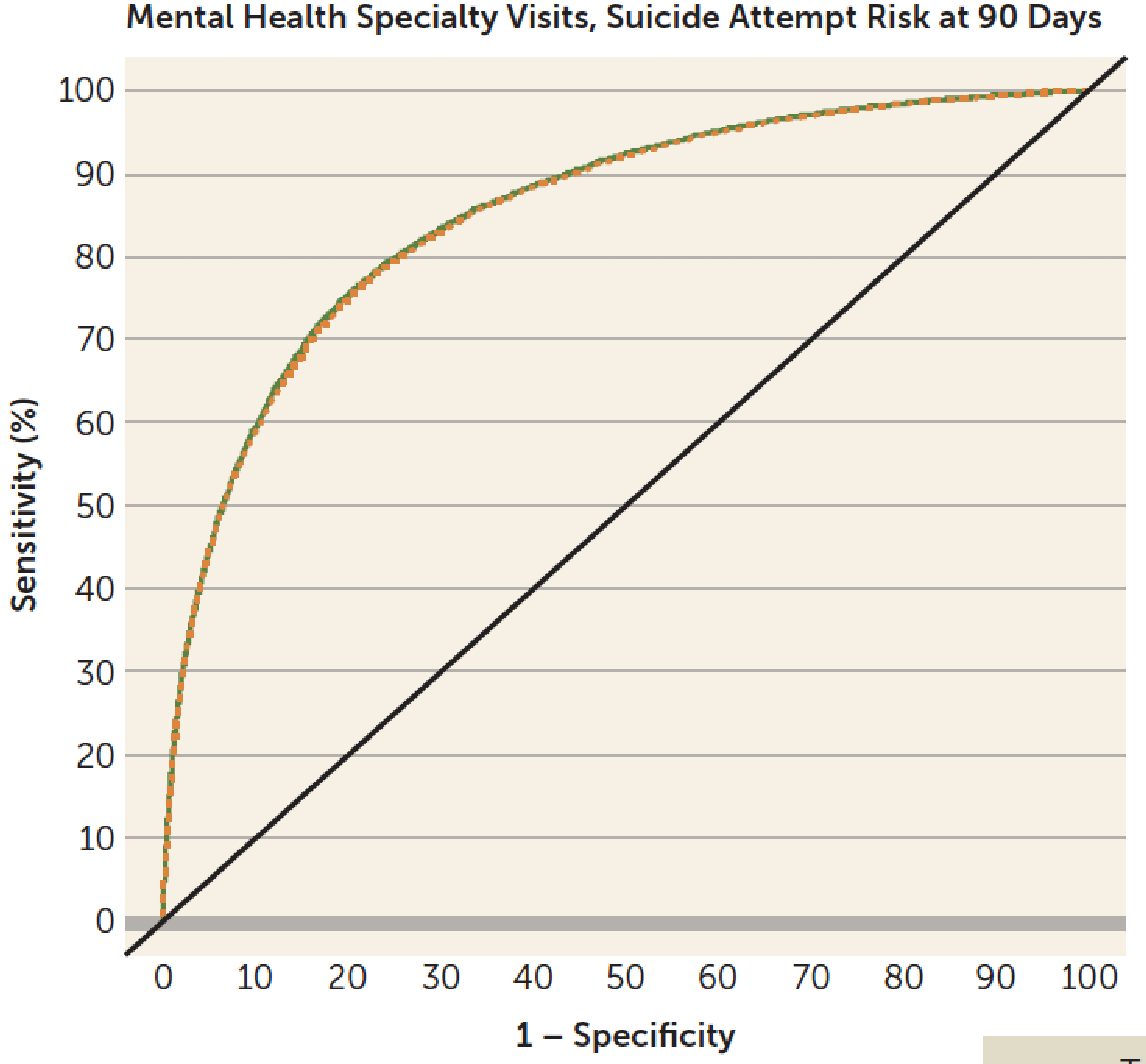


TABLE 4. Performance Characteristics at Various Cut-Points for Prediction of Suicide Attempts and Suicide Deaths Within 90 Days of Visit in Seven Health Systems, 2009–2015[a]

Risk Score Percentile Cut-Points	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)
Suicide attempts				
Following mental health specialty visits				
>99th	16.8	99.1	10.4	99.4
>95th	43.7	95.2	5.4	99.6
>90th	58.3	90.3	3.6	99.7
>75th	79.2	75.2	2.0	99.8
>50th	92.1	50.0	1.1	99.9

Case study: Predicting suicide risk (logistic regression results)

FIGURE 1. Receiver Operating Characteristic Curves Illustrating Model Performance in the Validation Data Set for Prediction of Suicide Attempts and Suicide Deaths Within 90 Days of Visit in Seven Health Systems, 2009–2015[a]



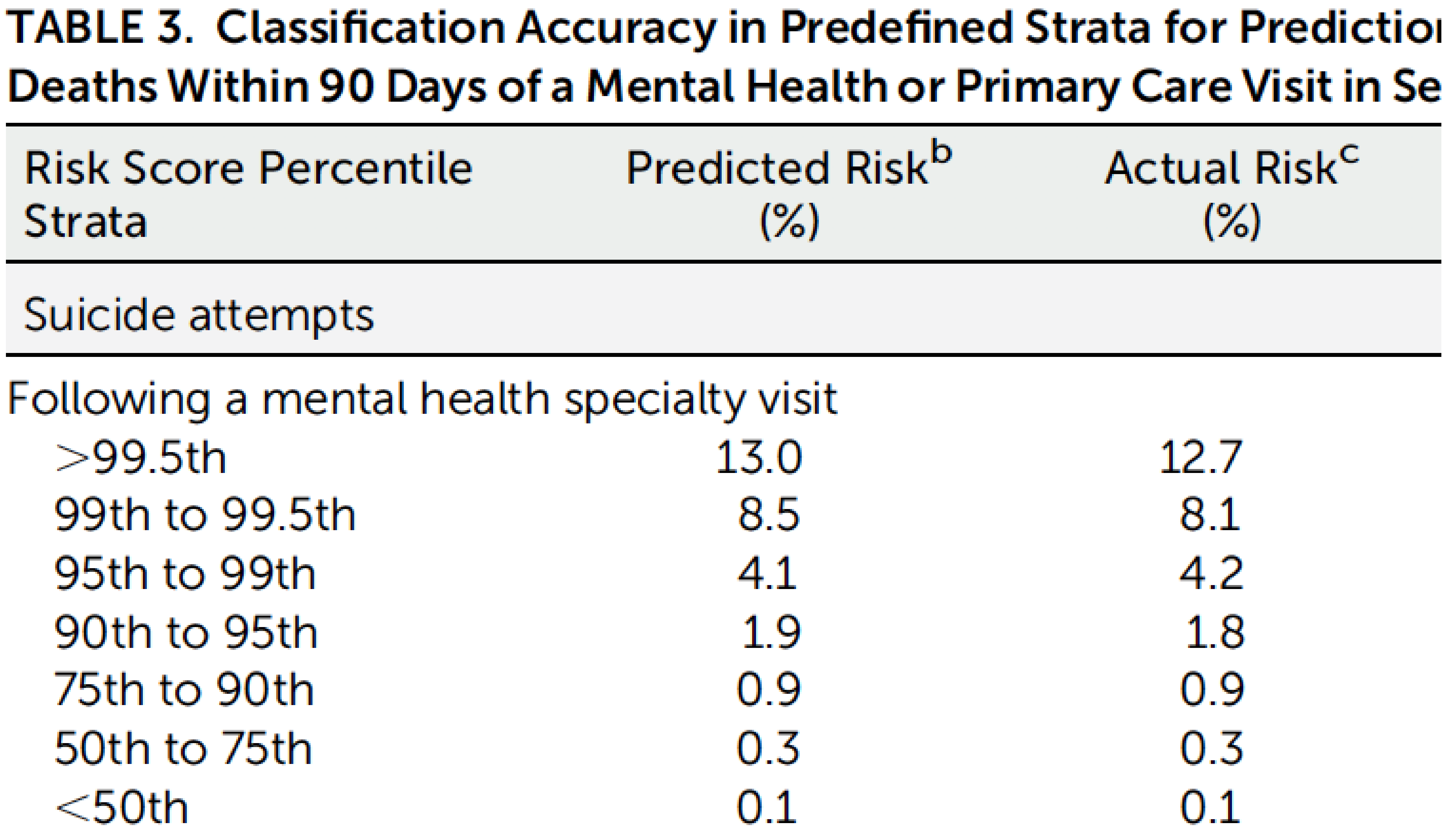


TABLE 3. Classification Accuracy in Predefined Strata for Predictio Deaths Within 90 Days of a Mental Health or Primary Care Visit in Se

Risk Score Percentile Strata	Predicted Risk[b] (%)	Actual Risk[c] (%)
Suicide attempts		
Following a mental health specialty visit		
>99.5th	13.0	12.7
99th to 99.5th	8.5	8.1
95th to 99th	4.1	4.2
90th to 95th	1.9	1.8
75th to 90th	0.9	0.9
50th to 75th	0.3	0.3
<50th	0.1	0.1

Correlation and clustering in prediction studies

- Correlation and clustering common in EHR data (See intro slides)
  - Suicide case study: median= 3 visits, IQR=1-8
- Cluster size is typically informative
  - Suicide case study: 1.7 attempts/1,000 visits for people with 1 visit vs 10 for people with 20+ visits
- Inference vs. prediction
  - In inference, correlated outcomes within a cluster affects variance estimates
  - In prediction, we don't care about interpreting predictor effects or estimate variance. But, clustering may have other impacts:
    - When dividing data for train/test split, prediction performance may be over-estimated if correlated outcomes appear both in training and testing data (similar impact on cross-validation)
    - If risk factors are different for people with many visits than those with few, prediction model may perform poorly in people with fewer visits
- Coley et al (2021) *Biometrical Journal*

Health equity in clinical prediction models

- Using prediction models estimated with observational clinical data to guide future medical care may exacerbate existing health disparities in access, quality, and outcome based on race, ethnicity, sexual orientation, gender identity, and other statuses.
  - Rajkomar et al (2018) *Ann Int Med,* Obermeyer et al (2019) *Science*
- Suicide case study:
  - Variability in suicide risk by race and ethnicity
  - Variability in patterns of mental health utilization and diagnosis by race and ethnicity
  - Black, Indigenous, and People of Color face barriers in accessing affordable, quality, culturally appropriate mental health care, and past discrimination shape current preferences
  - Suicide is a rare event, limiting the statistical power to identify race- and ethnicity-specific risk factors (i.e., may not be able to estimate interactions
  - Coley et al (2021) *JAMA Psychiatry*

Case study: Predicting suicide risk, health equity

Race/ethnicity	AUC	Sensitivity (at top 5% of risk)
White	0.83 (0.82, 0.84)	0.47 (0.44, 0.50)
Hispanic	0.86 (0.82, 0.89)	0.37 (0.30, 0.45)
Asian	0.83 (0.80, 0.87)	0.32 (0.22, 0.42)
Black	0.78 (0.69, 0.85)	0.07 (0, 0.17)
American Indian/Alaskan Native	0.60 (0.51, 0.69)	0.07 (0, 0.20)

Case study: Predicting suicide risk, health equity

- It is possible to set different thresholds for intervention for specific racial and ethnic groups to increase sensitivity (true positive rate)
  - But, there is a trade-off with specificity (false positive rate)
  - Using subgroup-specific thresholds to ensure equal sensitivity means there will be a higher false positive rate in subgroups with lower AUC—this may have negative consequences
- Equal prediction performance is necessary but not sufficient to ensure equitable use of prediction model
  - Populations with barriers to care less likely to have an encounter for which model is used
  - Benefits and harms of suicide prevention interventions may also vary by race/ethnicity (if they depend on access to safe, effective mental health care)
- Statisticians need to be mindful of how prediction models will be used and the impacts they might have throughout the modeling building and evaluation process

Other issues to consider for deployment of prediction models

- Monitoring prospective performance is important (want to make sure you are improving care), but may be challenging since implementation of prediction model will (should) change care
  - Pragmatic trials can be used to evaluate impact of prediction model use
- A clinical prediction model must be “actionable”, inform a decision
  - Should have an evidence-based intervention or action to inform
  - Benefits and harms/costs of an intervention must align with model accuracy
- Clinician and health system buy-in is crucial if a prediction model is going to be implemented
  - Providers must understand (and accept) limitations of prediction model
- Developing a prediction model is a very small part of a much larger project!
  - Building a model into the EHR, designing a workflow to show information to provider, and training providers to use prediction model and deliver intervention are all subsequent steps
  - Need to develop a plan to monitor model performance impact after implementation

Statistical methods for pragmatic trials

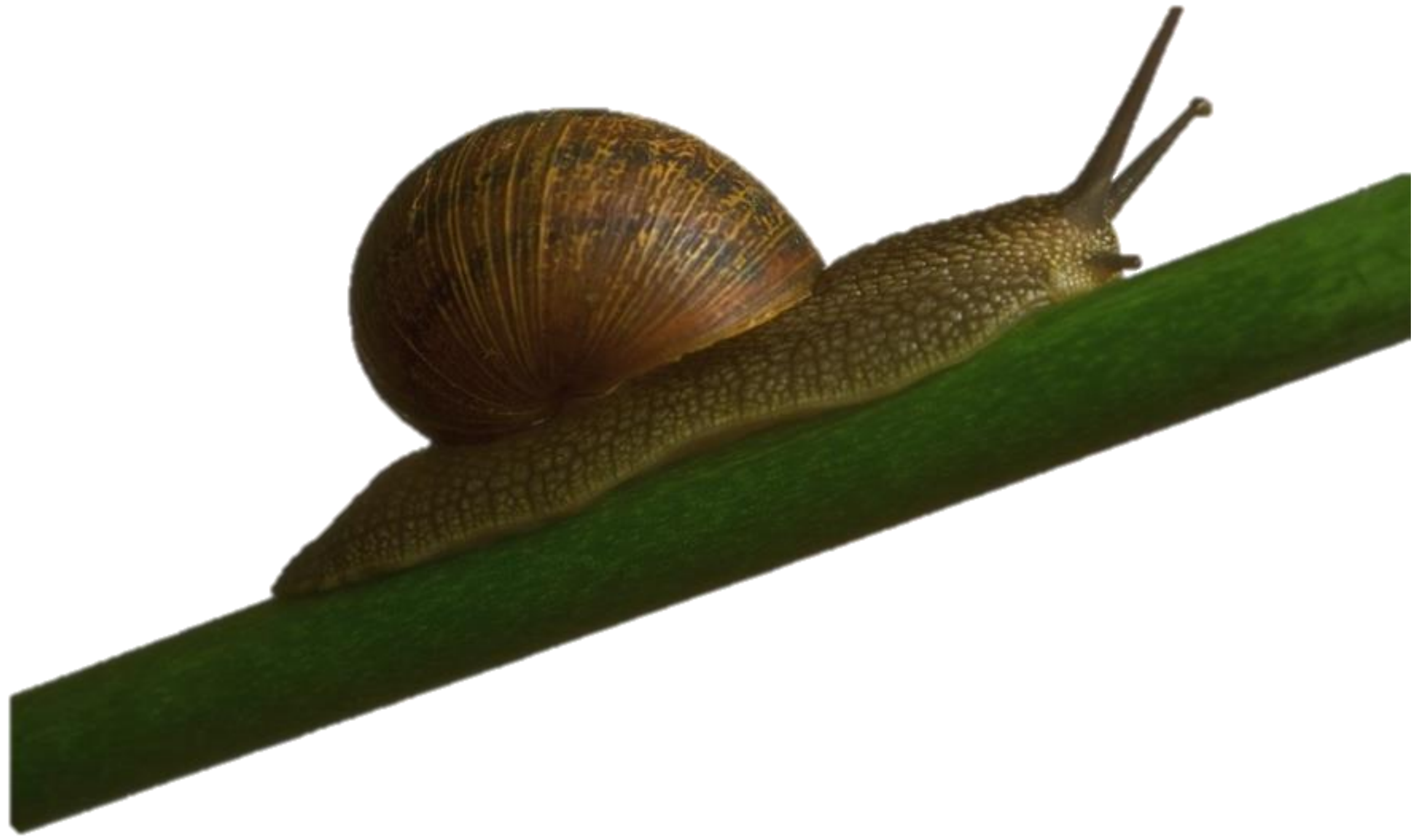
Outline

- Why pragmatic clinical trials (PCTs)?
- What are pragmatic trials?
- Design and analytic considerations
    - Common trial designs: cluster randomized, stepped wedge, Zelen
    - Challenges in using EHR data to define eligibility criteria and for outcome ascertainment
    - Health equity
- Interwoven case studies illustrating concepts
    - PROUD: cluster randomized trial
    - SPARC: stepped wedge trial
    - MICARE: Zelen trial

Clinical research is slow

- Traditional randomized controlled trials are slow and expensive—and don't always produce findings that are easily put into practice

- In fact, it takes an average of 17 years before research findings lead to widespread changes in care



NIH Collaboratory
Health Care Systems Research Collaboratory

Slides adapted from NIH Collaboratory

Clinical research is not always directly relevant to practice

- Traditional randomized trials study efficacy of treatments for carefully selected populations under ideal conditions

- Difficult to translate to real world
  - Patients rarely meet the exact (often strict) inclusion criteria
  - Certain populations have often been excluded
  - Real-life adherence is rarely the same as treatment protocols in trials

- When implemented into everyday clinical practice, often see a “voltage drop”— dramatic decrease in effectiveness

> “If we want more evidence-based practice, we need more practice-based evidence.”
>
> Green, LW. *American Journal of Public Health*, 2006.

NIH Collaboratory

Health Care Systems Research Collaboratory

Slides adapted from NIH Collaboratory

What are pragmatic clinical trials (PCTs)?

“Pragmatic clinical trials are performed in real-world clinical settings with highly generalizable populations to generate actionable clinical evidence at a fraction of the typical cost and time needed to conduct a traditional clinical trial.”

NIH COLLABORATORY
LIVING TEXTBOOK
of Pragmatic Clinical Trials

Advantages

- Large sample sizes
- Opportunity to study a diverse population including subgroups (e.g., youth, pregnant women) that are often excluded from explanatory trials
- Generalizability
- Cost (per patient) – but depends on existing infrastructure

Challenges

- Rely on big, often messy EHR and claims data not collected for research purposes
- Some design decisions are outside of the investigators’ control

Study design

Common study design: **cluster randomized trial**

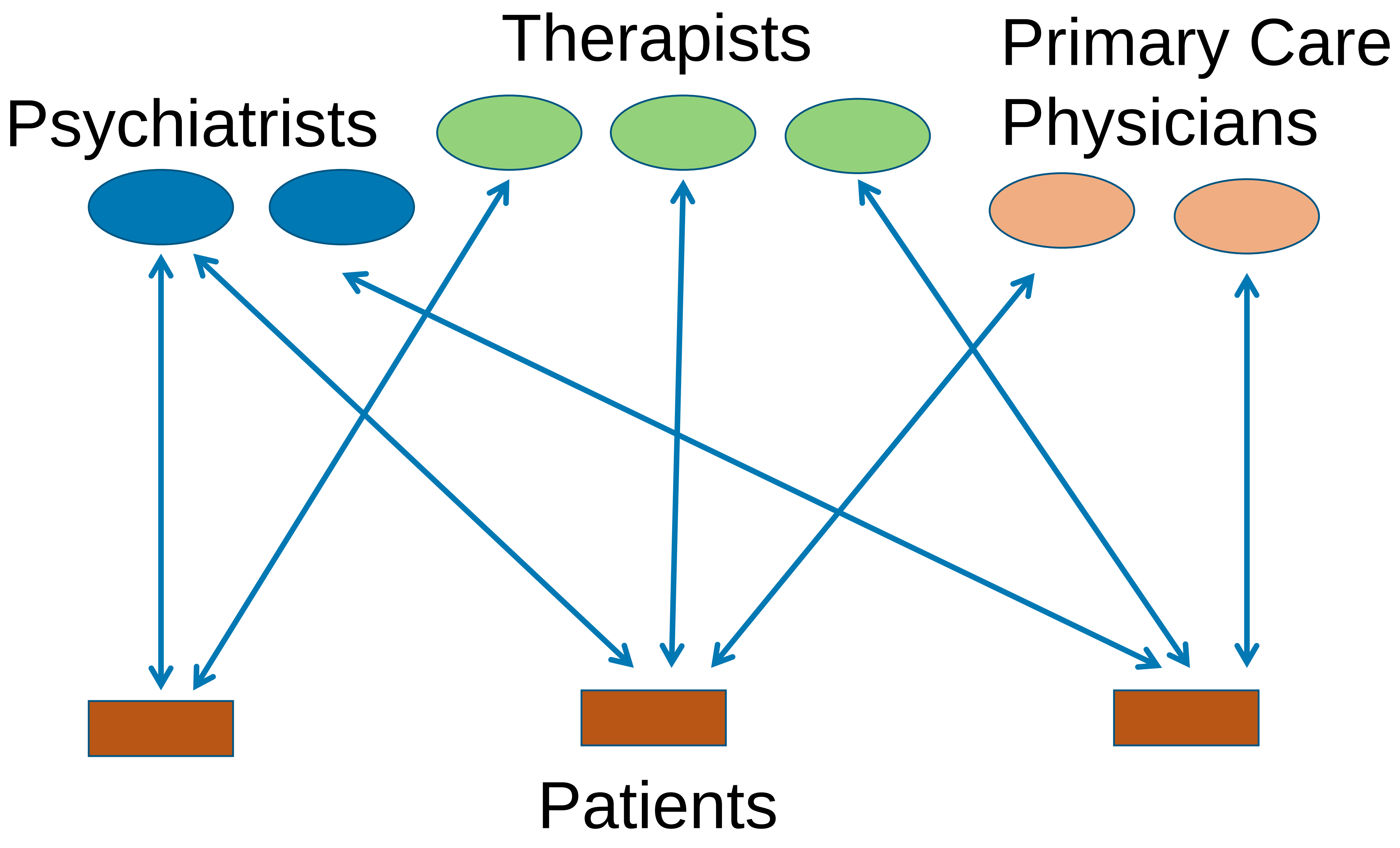
- Instead of randomizing individual patients, clusters of patients are randomized
- Intervention may be at the cluster-level (e.g., provider trainings)
- May be necessary to avoid contamination
  - For example, if providers are trained to provide the intervention to patients, and they have some patients randomized to the intervention and some randomized to control, it may be challenging to avoid contamination between the intervention arms
  - Contamination leads to diluted intervention effects
- Patients in the same cluster are correlated → needs to be accounted for in the analysis

Cluster randomized trials

Choice of randomization unit (cluster):

- Provider < Panel < Clinic < Region < Site
- Goal: smallest unit without (too much) contamination
  - For a given sample size (# patients), statistical power increases as the number of clusters increases
- More clusters are better if possible
- Smaller number of clusters leads to inference challenges
  - Chance imbalance in covariates
  - Elevated type 1 error rates for individual-level analyses (generalized estimating equations [GEE], generalized linear mixed-effect models [GLMM])

- May not be able to avoid contamination completely
  - Patients may go to multiple providers or clinics (especially if intervention period is long)
- Think about whether this is the best design



Randomization

- Simple randomization
- Stratified randomization
  - Stratify on a small set of covariates (note: covariate values needed prior to randomization)
- Covariate-constrained randomization
  - Balances a large number of characteristics
  - Requires that all clusters are recruited prior to randomization
  - Example of general approach:
    1. Enumerate all (or simulate many) cluster randomization assignments (A or B but not actual treatment)
    2. Across these possible randomization assignments assess characteristic “balance” using a pre-specified balance metric (several options available)
    3. Restrict to those assignments with balance (requires specifying a threshold)
    4. Randomly choose from the “constrained” pool a randomization scheme
    5. Randomly assign treatments to A or B

Cluster randomized trials

Analyses can be at the **cluster level** (cluster-level outcome) or at a **patient level**

- If at a patient level, power calculations/analyses must account for correlation of patients within a cluster
  - 50 clusters per condition, 100 patients per cluster, effect size = 0.1σ
    - ICC (0.01, 0.03, 0.05) → power (94%, 70%, 53%)
- Power can also be reduced with variable cluster sizes
  - 50 clusters per condition, 100 average cluster size, effect size = 0.1σ, ICC = 0.02
    - CV (0, 0.50, 1) → power (82%, 80%, 72%)

If possible, collect information on cluster's outcome rates at baseline (pre-randomization)

- Baseline value of outcome typically strongly prognostic
- Adjustment can increase power and control for imbalance at baseline

Stepped wedge cluster randomized trials

- All clusters receive the intervention, eventually
- Randomize timing of when the cluster is turned on to intervention
- Within each time point, clusters are exchangeable (i.e., expect balance in covariates)
- But intervention effect confounded with time
  - Analysis must account potential confounding due to secular trends

	Cluster	Baseline	Period 1	Period 2	Period 3	Period 4
Stepped Wedge	3	UC	INT	INT	INT	INT
	2	UC	UC	INT	INT	INT
	1	UC	UC	UC	INT	INT
	4	UC	UC	UC	UC	INT

Stepped wedge cluster randomized trials

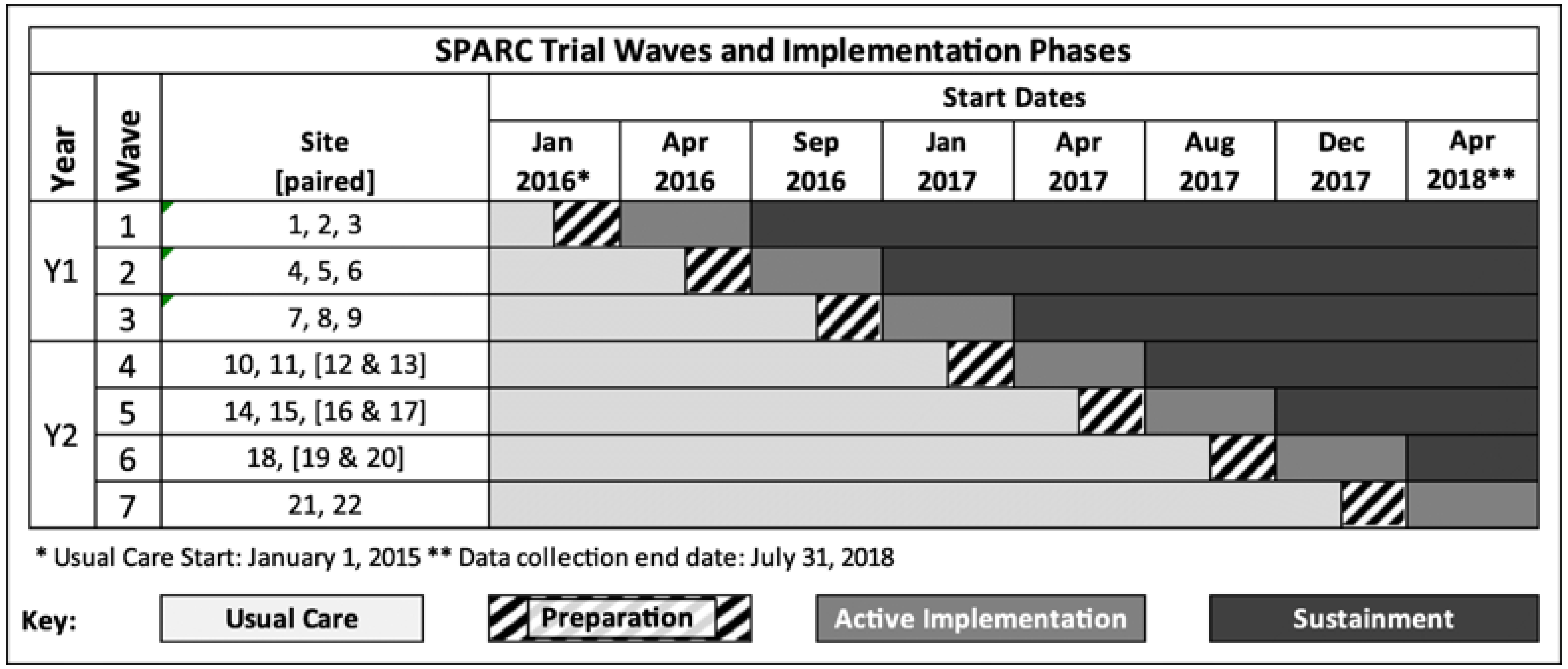
- Different types of stepped wedge trials
  - Cross-sectional: different patients measured at each time point
  - Closed cohort: same participants measured at each time point
  - Open cohort: patients enter (and leave) cohort at different time points
    - Common in EHR studies within health systems
- Analysis should use longitudinal data methods and account for correlation of
  - Patients from the same clinic (at the same or different time periods)
  - Repeated measures from same patient (closed or open cohort design)
- Pre-specify approach to account for time trends
- Power calculations more complex, though software exists
- Due to challenges in implementing and potential for confounding, conventional parallel group cluster trials generally preferred, if feasible

Case Study: The Sustained Patient-centered Alcohol Related Care (SPARC) Trial

Pragmatic, stepped-wedge implementation trial

Interventions:

- Integrated alcohol-related care
  - Screening for unhealthy alcohol use, brief alcohol counseling
  - Assessment, diagnosis, and treatment of alcohol use disorders (AUDs)
- Usual primary care



SPARC Trial Waves and Implementation Phases

Year	Wave	Site [paired]	Start Dates							
			Jan 2016*	Apr 2016	Sep 2016	Jan 2017	Apr 2017	Aug 2017	Dec 2017	Apr 2018**
Y1	1	1, 2, 3	░▨	▒	■	■	■	■	■	■
	2	4, 5, 6	░	░▨	▒	■	■	■	■	■
	3	7, 8, 9	░	░	░▨	▒	■	■	■	■
Y2	4	10, 11, [12 & 13]	░	░	░	░▨	▒	■	■	■
	5	14, 15, [16 & 17]	░	░	░	░	░▨	▒	■	■
	6	18, [19 & 20]	░	░	░	░	░	░▨	▒	■
	7	21, 22	░	░	░	░	░	░	░▨	▒

* Usual Care Start: January 1, 2015 ** Data collection end date: July 31, 2018

Key: ░ Usual Care ▨ Preparation ▒ Active Implementation ■ Sustainment

Sample: 19 primary care sites of Kaiser Permanente Washington

- 330,000 PC patients (2015-2018)

Pragmatic design feature: randomization scheme developed to meet health system constraints, including stratification (by implementation year)

Case Study: The Sustained Patient-centered Alcohol Related Care (SPARC) Trial

Statistical methods:

- GLMM with site- and person-specific random intercepts, adjusted for calendar time (4-month intervals) and stratum
- Computational challenges: big dataset, non-nested random effects, rare (binary) outcomes

Results:

- Increased brief alcohol counseling
  - though absolute magnitude of increase was small
- Did not increase AUD treatment engagement

Zelen design

- Zelen design / encouragement trial
  - Individually randomized
  - Instead of consent followed by randomization, randomize patients to be “offered” an intervention (a subset of whom may consent)
    - Primary intent-to-treat (ITT) analysis compare patients offered intervention to those who are not offered intervention (usual care)
    - Power calculations need to account for dilution of effect
      - Consent rate major driver of power
    - Patients randomized to usual care are never contacted
    - Need to be able to ascertain the outcome on everyone (including those who did not consent to the intervention and those who were randomized to usual care and never contacted)

Defining eligibility criteria of sample using EHR data

- In many pragmatic trials, no patient contact for research purposes
  - Patients may be contacted as part of the intervention but not for research
- EHR (and claims) data used to ascertain eligibility criteria of sample to be analyzed
- In **traditional cluster randomized trials**, if patients are recruited after randomization:
  - Intervention group assignment could affect who is recruited
  - Possibility for type of post-randomization selection bias called “recruitment bias”
    - Patients recruited in intervention clusters may be systematically different than those recruited in usual care clusters
- In **pragmatic cluster randomized trials** without patient contact:
  - If patients are identified for inclusion in trial analyses using post-randomization data, potential for similar type of bias (“identification bias”)

Approaches to address identification bias in cluster-randomized trials

Study design (preferred)

- Primary analysis: define sample using pre-randomization (pre-R) data
  - Randomization ensures comparability across intervention groups
- If sample defined using post-R data, some samples may be less likely to be affected by treatment assignment (e.g., entire clinic population)
  - But need to ensure adequately powered

Statistical analysis (if sample defined using post-randomization data)

- Can adjust for baseline characteristics
  - But recent work has highlighted that this will not fully remove bias in general
- Methods being developed to address identification bias analytically within the principal stratification framework for post-treatment selection bias

Case Study: The PRimary care Opioid Use Disorders treatment (PROUD) Trial

Pragmatic, cluster-randomized implementation trial

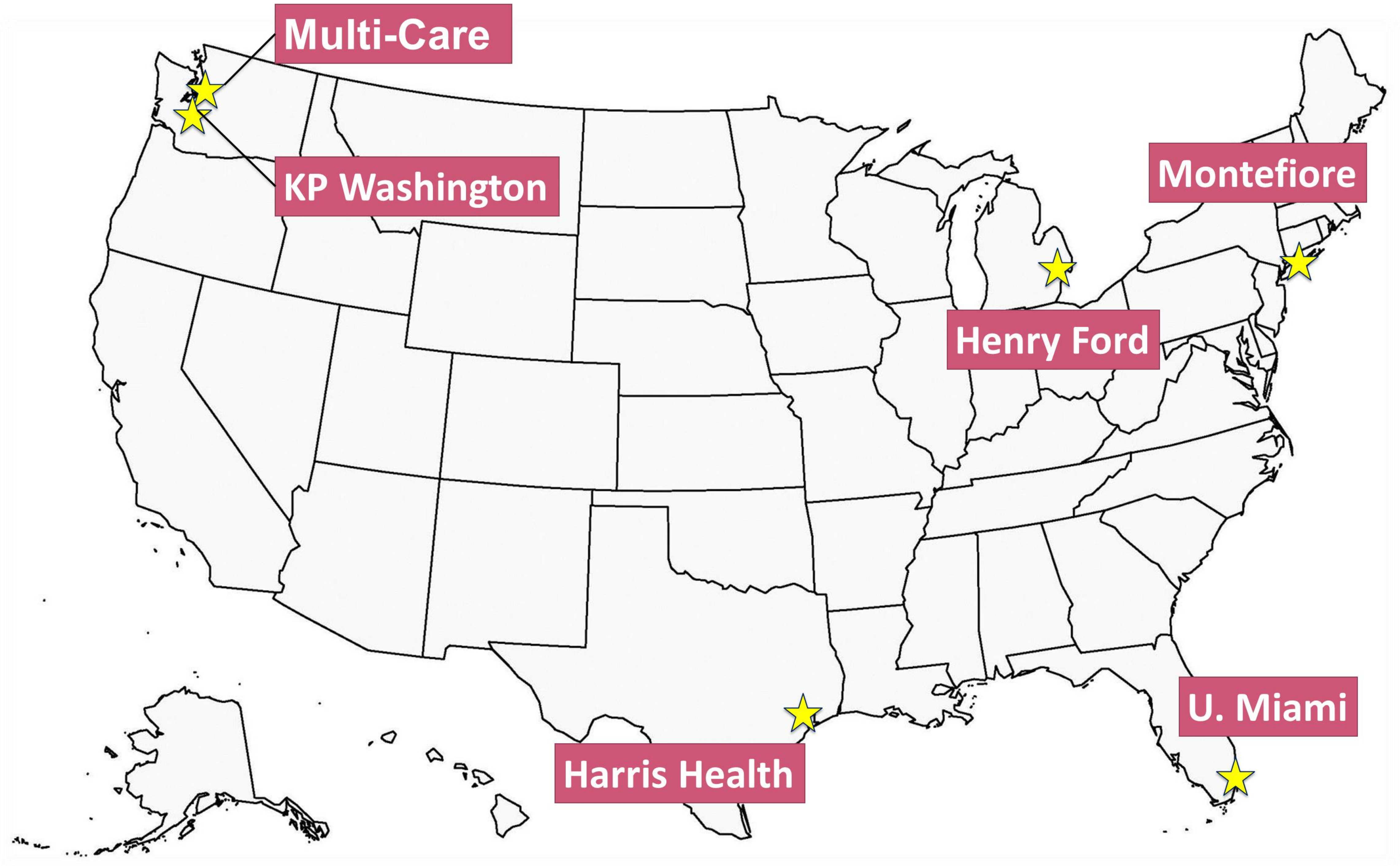
Interventions:

- Implementation of Massachusetts (MA) Model
  - nurse care management for opioid use disorder (OUD)
- Usual primary care

Sample: 12 primary care clinics within 6 diverse health care systems

- ~375,000 PC patients
- 3,335 OUD diagnosis

Randomization: stratified on the health system (1 MA Model, 1 usual primary care clinic)



Potential for identification bias

Effectiveness objective: Does MA Model reduce acute care utilization (days of emergency and hospital care) among patients with OUD?

Challenges in using EHR data to define eligibility criteria

- OUD is underdiagnosed
- MA Model is expected to increase diagnosis and attract patients to receive care

Potential for identification bias:

- Intervention affects who is diagnosed with OUD
- Patients diagnosed in the intervention arm are likely to be different (either sicker or healthier) than patients diagnosed in the control arm.
- Bias can be in either direction

Addressing identification bias

Design solution: only include individuals identified pre-randomization (pre-R)

- Primary sample: patients with an OUD diagnosis pre-randomization
  - Among 1,988 patients identified pre-R,
    - no intervention effect on acute care utilization
    - Also no effect on explanatory OUD treatment measures
  - Limitations: misses many patients who are potentially affected

Analytic solution: secondary sample additionally includes 1,347 patients with OUD newly recognized post-randomization (post-R Sample) and adjusts for additional covariates

- No intervention effect on acute care utilization
  - Explanatory OUD treatment measure: mean difference in days of OUD treatment
    - -11.7 (95% CI -38.6, 15.3) in pre-R sample
    - 32.2 (95% CI 4.7, 59.7) in post-R sample
  - Could reflect true differences or reflect residual unmeasured confounding

Using real world data for outcome ascertainment

- In efficacy trials, measures are typically collected at regularly spaced, pre-defined time points
    - With the same number of follow-up measures planned per person
- In pragmatic trials using real world data, this may not be possible
    - Measures collected as part of routine clinical care
    - Likely to be variability across patients and providers
    - May be external incentives in health system to obtain measures at certain time points (e.g., national performance metrics)
    - Need to consider potential for outcome ascertainment to differ across arms, potentially leading to bias

Case Study: More Individualized Care: Assessment and Recovery through Engagement (MICARE) Trial

Pragmatic, individually-randomized Zelen trial

Interventions:

- Telephone-based nurse care management for depression and OUD
- Usual primary care

Sample: Target of 800 patients within 2 health systems

Randomization: within each health system, 1:1 randomization to be offered the intervention vs. usual care (usual care patients not contacted)

Aims: Test whether offering the intervention:

- Increases OUD medication treatment (primary outcome)
- Reduces depressive symptoms on PHQ-9 (depression screen collected as part of routine care)

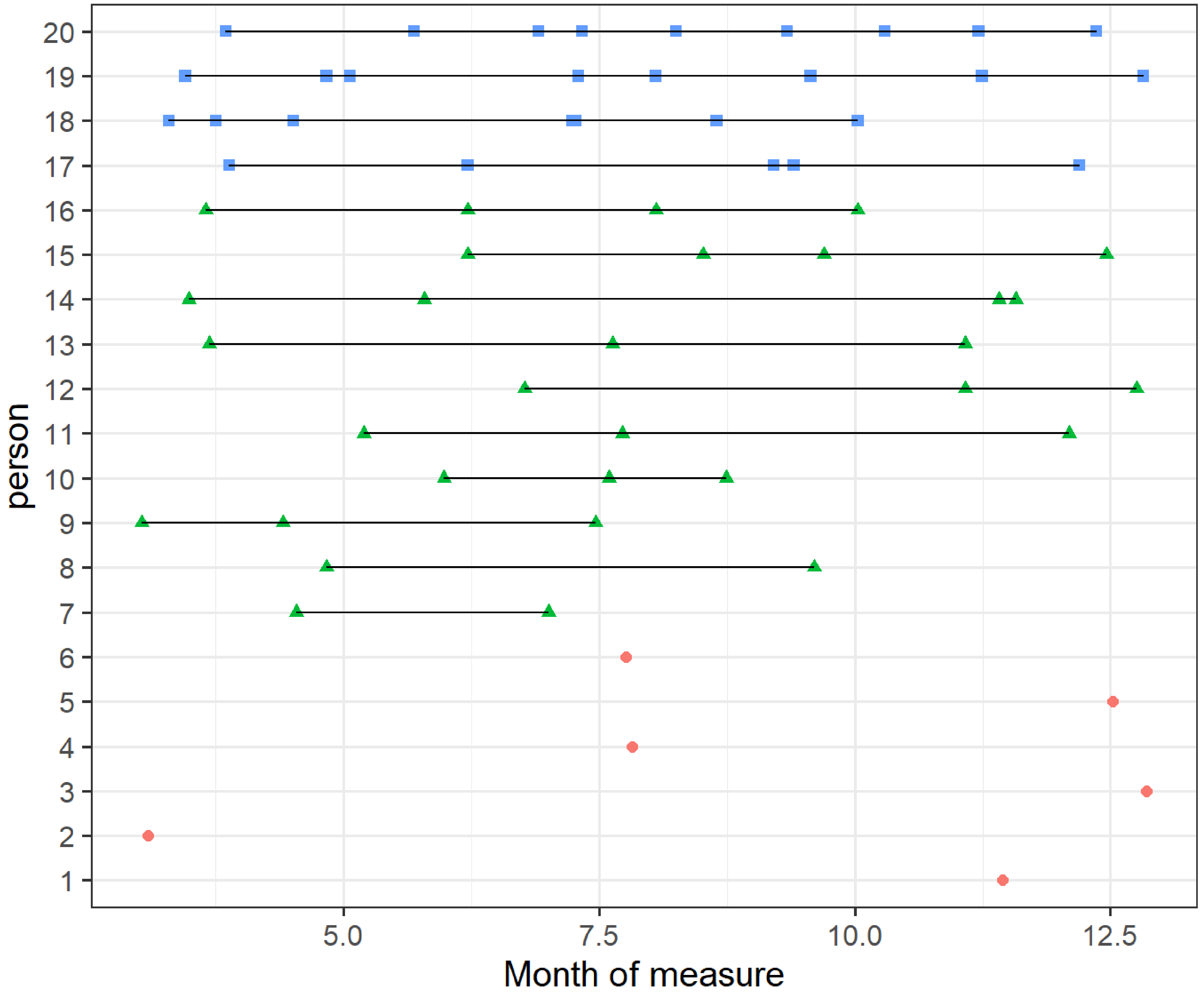
Using clinical EHR measures in MICARE study

For **depression outcomes**, we have

- Irregular timing and repeated measures
    - not controlled by study team
- Intervention expected to increase follow up

Need to **carefully consider** how to define outcome

- Using “best score” over follow-up could lead to high levels of bias
- Plan to use follow-up score closest to 12 months, adjusted for timing since baseline
    - Will conduct simulation study using **pre-trial retrospective cohort data** to assess potential of proposed approach for bias



Outcome ascertainment and missing data

- Consider potential for measurement error in outcome (see intro slides)
- May need to validate the outcomes you do observe
  - Depending on the outcome (PPV, sensitivity)
  - Depending on the cost (two-stage design?)
- Handling missing outcome data (e.g., if patient leaves the health system): range of different approaches (e.g., multiple imputation, weighting), depending on
  - Missingness mechanism: administrative missingness (MCAR), MAR, or MNAR?
  - Amount of missing data: how stable is your population being studied?

Health equity in PCTs

PCTs require administrative and data resources to conduct

- Low resource clinics and health systems don’t always have the resources to participate

Who is in your population?

- An advantage of PCTs can be ability to enroll more diverse patient sample
- Think about how using EHR to define eligibility criteria could impact your sample (see intro slides)

Heterogeneity of treatment effect analyses

- Need sufficient sample size to detect interactions
- Need careful interpretation of any subgroup effects
    - Consider social determinants of health and institutionalized barriers

Summary

- PCTs enable studying interventions in real world clinical settings and inform learning health systems
- PCTs add complication
  - Study design needs to be flexible to balance health system needs
- Many design challenges
  - Defining the eligibility criteria for the sample to be analyzed requires careful thought
  - Clinical measures have several challenges when used as trial outcomes
- Preliminary data (prior to randomization) can be used to examine assumptions and inform the study design
- Health equity: consider who is in your sample and whether the intervention effect may differ across populations
- Ongoing area of active research: many open design and statistical questions

Statistical methods for electronic health record data
International Conference on Health Policy Statistics

Jennifer F Bobb, R Yates Coley, Susan M Shortreed

Biostatistics Unit, Kaiser Permanente Washington Health Research Institute (KPWHRI)

jennifer.f.bobb@kp.org, rebecca.y.coley@kp.org, susan.m.shortreed@kp.org

Acknowledgements to entire Biostatistics Unit at KPWHRI