The Benefits and Challenges of Leveraging Existing and Secondary Data for Pragmatic Research

- David M. Vock, Ph.D.

COPRH Con

Colorado Pragmatic Research in Health Conference

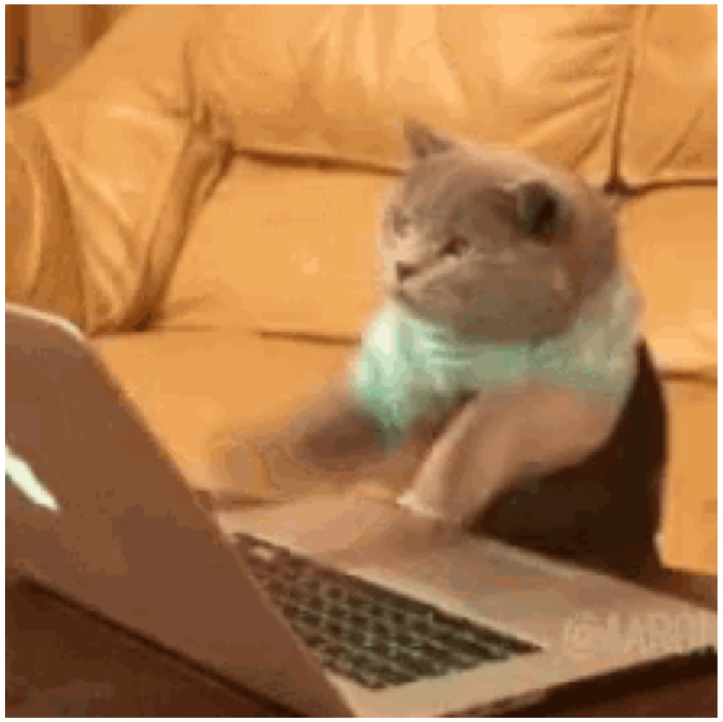
ACCORDS
UNIVERSITY OF COLORADO
CHILDREN'S HOSPITAL COLORADO

Colorado Clinical and Translational Sciences Institute (CCTSI)
UNIVERSITY OF COLORADO DENVER | ANSCHUTZ MEDICAL CAMPUS

Please Reach Out During This Talk!

Chat Box!

@docvock



COPRH Con
Colorado Pragmatic Research in Health Conference

ACCORDS

Meeting of Two Worlds

Using only existing/secondary data

- Health services research
- Health economics
- Outcomes research

Using only prospective (new) data

- Traditional RCT
- Epidemiologic Cohort Study

- Is there some middle ground in these research approach?
- In particular, what can practitioners of prospective data collection leverage from existing/secondary data

COPRH Con
Colorado Pragmatic Research in Health Conference

ACCORDS

CCTSI

Sources of Existing/Secondary Data

Medical Record

Source: Providers of Medical Care

Examples:

- HealthPartners
- Kaiser Permanente
- VA
- Allina Health
- Fairview/M Health
- Mayo Clinic

Insurance Claims

Source: Payers of Medical Care

Examples:

- HealthPartners
- Kaiser Permanente
- VA
- Medica
- Medicare

Registries/Completed Studies

Source: Government Agencies

Examples:

- State Death Certificates
- Social Security Death Master File
- National Death Index
- Scientific Registry for Transplant Recipients

Existing/Secondary Data

- Strengths
  1. Contemporaneous
  2. Captures information on those who seek care (diverse, large)
  3. Contains information available and relevant to clinicians and patients
  4. Large sample and inexpensive
- Weaknesses
  1. Irregular and inconsistently collected data
  2. Confounding by indication (challenging to compare interventions)
  3. Patients may seek care outside the system of interest (e.g., no heart attack recorded does not mean one did not occur)
  4. Fewer data quality checks, more measurement error

Key Points

- Existing/secondary data can and should be used at many points in the lifecycle of pragmatic research
- Any source of data should be interrogated for the not only what it includes but also what it does not capture.
- Using existing and secondary data requires a data integration and security plan.
- The limitations of existing/secondary data should be ameliorated in the design and analysis plan.
- Many perceived weaknesses of existing and secondary data should be reframed as strengths for pragmatic research.

Case Studies

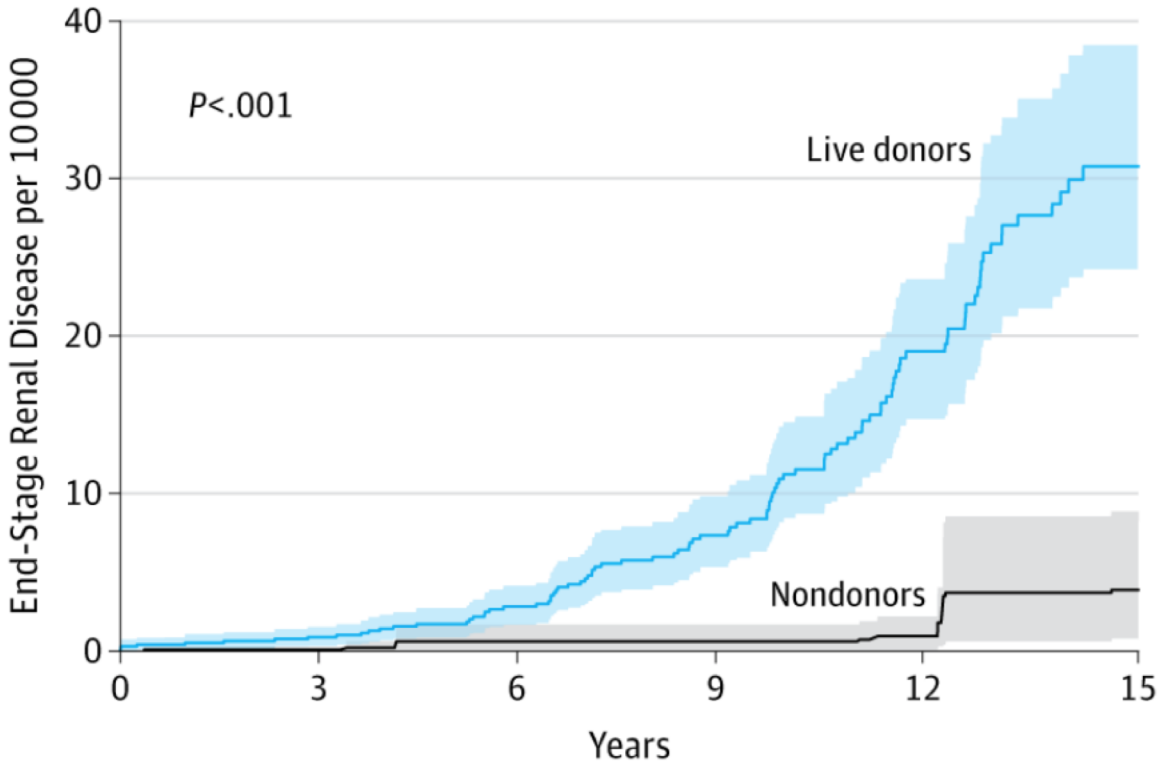
- Assessing the Long-Terms Health Consequences of Living Kidney Donation
- Evaluating the Effect of the Heart of New Ulm – a 10-year population-based intervention – On Cardiovascular Risk Factors and Outcomes
- Developing Cardiovascular Risk Prediction Algorithms Using EHD

Living Kidney Donation

- Definitive treatment for patients with end-stage kidney disease
- Outcomes with living donor are better than with deceased donor and LD recipients often require shorter time on dialysis
- BUT kidney donation is not without risk including peri-operative mortality of 3 per 10,000 and major peri-operative complications of 3-6%
- Increased focus on long-term outcomes of kidney nephrectomy
- In general, while donors initially lose half of kidney function, they can expect to regain ~70% of pre-donation function within 1 year

Muzaale et al. 2014 JAMA

A Cumulative incidence of end-stage renal disease



No. at risk						
Live donors	96 217	77 587	58 979	39 231	21 573	8781
Nondonors	96 217	95 930	95 422	94 734	94 199	50 124

Mjoen et al. 2014 Kidney International

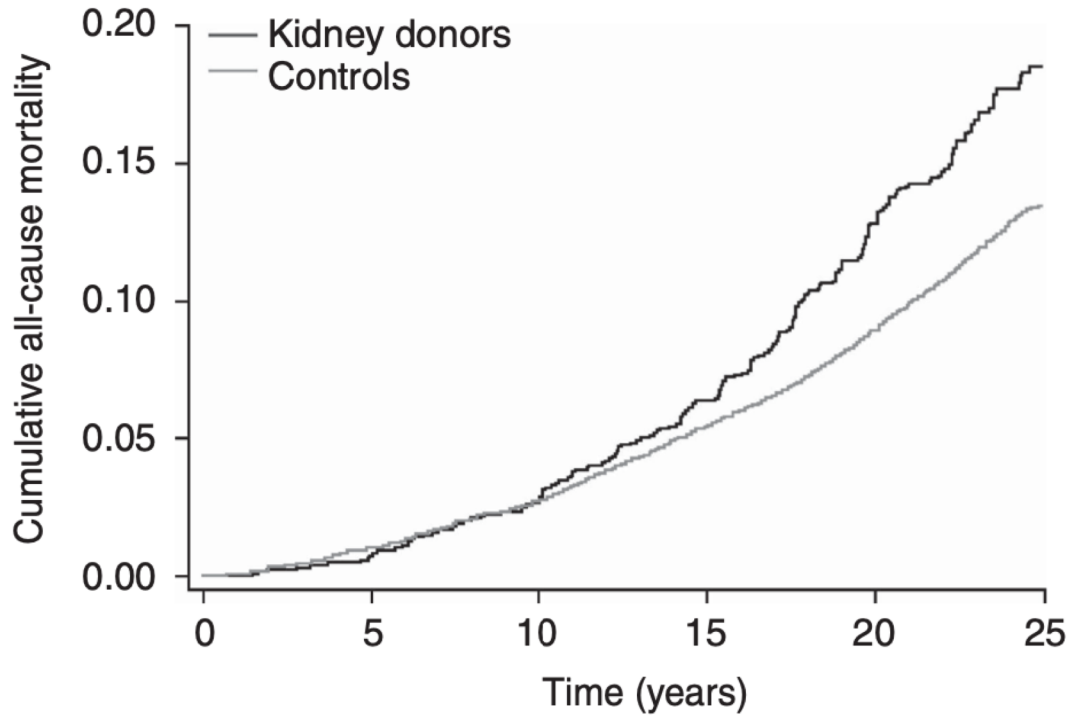
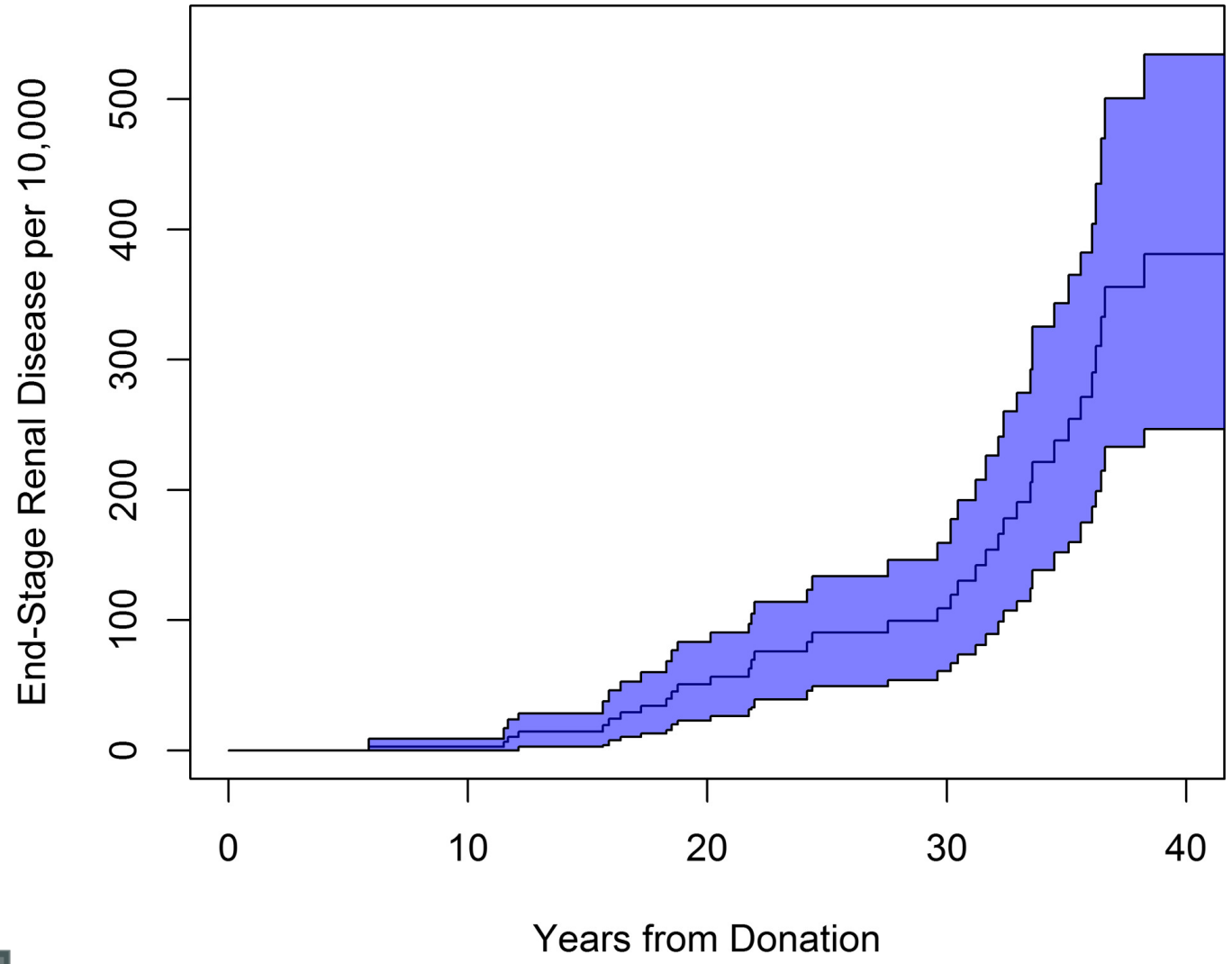


Figure 2 | Cumulative mortality risk in kidney donors and controls, adjusted for year of donation. Controls are matched to donors for age, sex, systolic blood pressure, body mass index, and smoking status.

Long-Term Trajectory – Matas et al. 2018 AJT

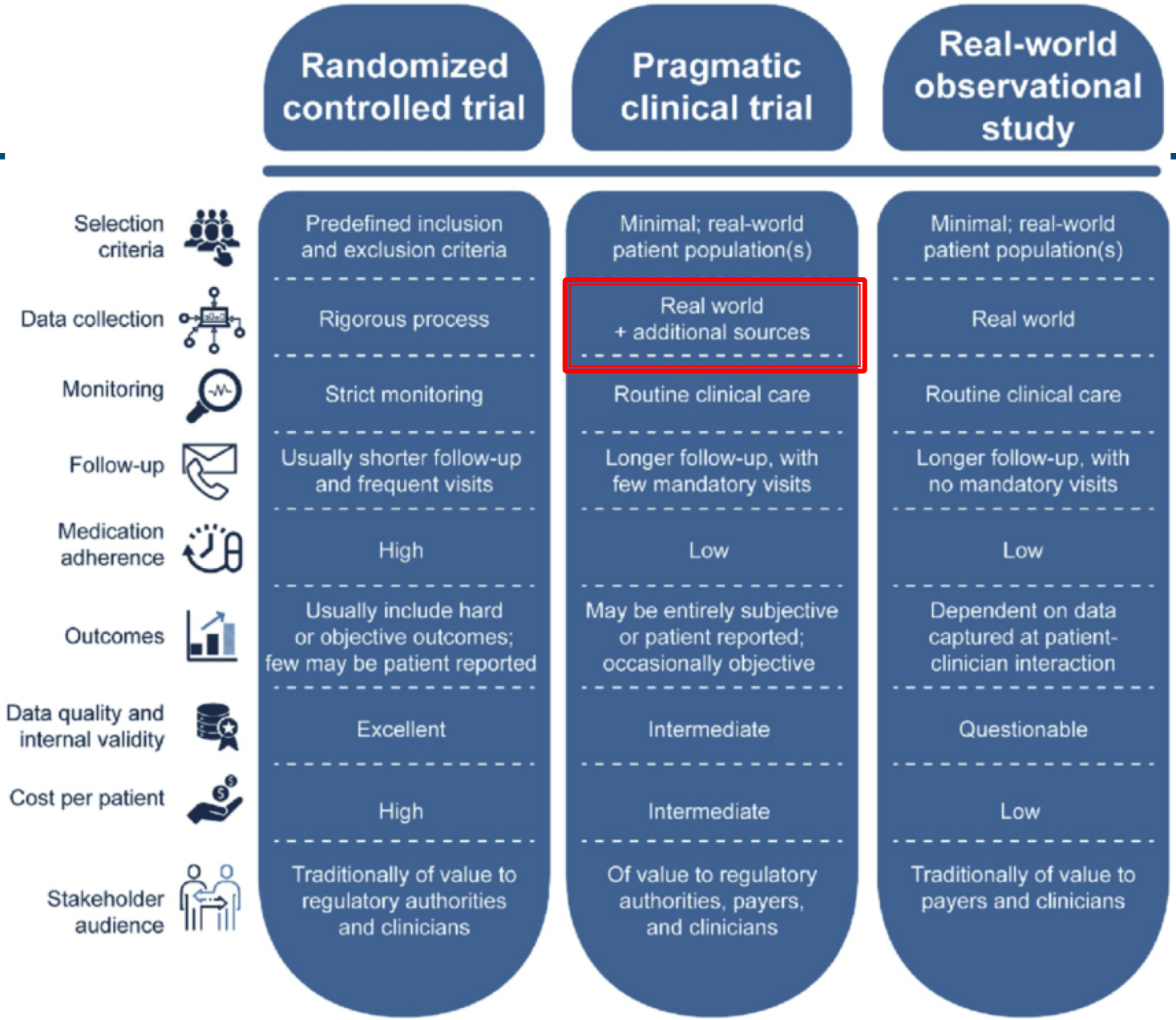


Key Question Necessitates Existing/Secondary Data

- Want to understand the long-term risks (e.g., >20 years) of kidney donation on development of ESKD and other intermediate endpoints (CKD, HTN) and secondary outcomes (all-cause mortality, CVD, etc.).
- National registries of living donors relatively recent development; no significant registries of those approved for donation but never donating or even generally healthy people over last 50 years

Key Points

- Existing/secondary data can and should be used at many points in the lifecycle of pragmatic research
- Any source of data should be interrogated for the not only what it includes but also what it does not capture.
- Using existing and secondary data requires a data integration and security plan.
- The limitations of existing/secondary data should be ameliorated in the design and analysis plan.
- Many perceived weaknesses of existing and secondary data should be reframed as strengths for pragmatic research.



Anzueto and Kaplan, 2020

Major Components of Any Study

- Design and Planning Phase
- Study Execution, Data Collection, & Monitoring Phase
- Analysis and Dissemination Phase

How can existing/secondary be used in each of these three phases of study design?

What changes to these phases must be done to accommodate use of existing/secondary data?

This the Warm-Up/Reminder

- Design and Planning Phase
- Study Execution, Data Collection, & Monitoring Phase
- Analysis and Dissemination Phase

This the Warm-Up/Reminder – Monday Talks

- Design and Planni

Track 2: Data Science and Biostatistics

Plenary - Recent Developments in Statistical Methods for Pragmatic, Stepped Wedge Cluster Randomized Trials
Fan Li, PhD

- Study Execution, Data Collection, & Monitoring Phase

- Analysis and Disseminati

Analyzing Correlated Data: Basics of the Linear Mixed Effects Model
John Rice, PhD

Clinical Prediction Models
Krithika Suresh, PhD; Katie Colborn, PhD

Interrupted Time Series with Individual Level Data
Elizabeth Juarez-Colunga, PhD; Angela Moss, MS

Causal Inference via Trial Emulation
Nandita Mitra, PhD

COPRH Con
Colorado Pragmatic Research in Health Conference

ACCORDS

CCTSI

This the Warm-Up/Reminder – Today's Talks

- Design and Planning Phase
- Study Execution, Data C
- Analysis and Dissemina

Building the Tower of Babel – Tricks and Traps in Harmonizing EHR Data
Lisa Schilling, MD, MSPH; Patrick Hosokawa, MS

Using Population-Based Data in Secondary Analysis
Allison Kempe, MD; Art Davidson, MD

Opportunities for Using Healthcare Claims Data for Pragmatic Sustainability Assessments
Mark Gritz, PhD

Digital Health Data Access, Management, and Use
Susan L. Moore, PhD, MSPH

Data Quality Assessment Issues and Methods for Secondary Data Use
Michael Kahn, MD, PhD

Methods for Linking Records Across Disparate Data Sources
Toan Ong, PhD; Jenna Reno, PhD

Watson: Attics, Guesswork and Clay. Sleuthing Your Way into Biomedical Natural Language Processing
Seth Russell, MS

Mining and Analyzing Data from Social Media Data Sources
Bethany Kwan, PhD; Jenna Reno, PhD

COPRH Con
Colorado Pragmatic Research in Health Conference

ACCORDS

CCTSI

This the Warm-Up/Reminder – Wednesday Talk

- Design and Planning Phase

Plenary - Implementing Pragmatic Trials via Electronic Platforms: Practical and Ethical Considerations for Consent, Participation, and Analysis
Andrea Troxel, ScD

- Study Execution, Data Collection, & Monitoring Phase

- Analysis and Dissemination Phase

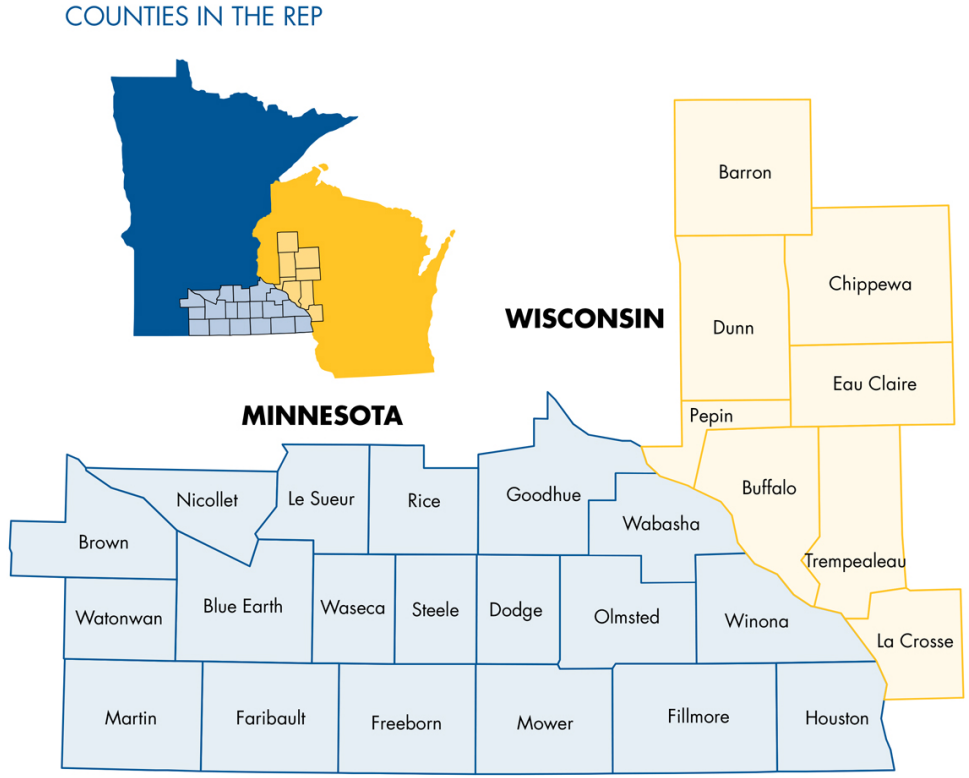
COPRH Con
Colorado Pragmatic Research in Health Conference
ACCORDS
CCTSI

UMN Donor Surveillance

- >4000 living kidney donors at University of Minnesota since program inception in early 1960s
- Since 2000, donors are contacted and complete survey of health history every 3 years
- Donors are asked to have medical records forwarded or consent to contact their physicians as part of this survey
- Some follow-up of donors within UMN/Fairview system but donors come from all over and live all over
- Excellent source of potential donors

Rochester Epidemiology Project (REP)

- Collaboration of clinics, hospitals, and other medical facilities in Minnesota and Wisconsin
- Involves community members who have agreed to share their medical records for research
- Two main providers in Olmstead Country – Mayo Clinic and Olmstead Medical Center
- Diagnosis and procedure codes date to 1966
- Laboratory values (electronic) date to the 1990s but paper charts are available
- Source of identifying potential controls to prospectively collect data



ROCHESTER EPIDEMIOLOGY PROJECT
EST. 1966

A Minnesota and Wisconsin Collaboration

COPRH Con
Colorado Pragmatic Research in Health Conference

ACCORDS

Source of Controls

- Using the REP to identify potential controls → design phase of study
- Match UMN donors to 4 participants in the REP (“potential controls”)
  - Exact match on race and gender, must be within 5 years of age
  - Must have one visit before and after index date
  - At index date, cannot have any diagnosis code which is an ”always exclude” condition
- The charts of these potential controls are then reviewed to ensure that there is not additional information which would exclude the participant
- Process ensures that the controls are contemporaneous to donors and typically from same geographic

Key Points

- Existing/secondary data can and should be used at many points in the lifecycle of pragmatic research
- Any source of data should be interrogated for the not only what it includes but also what it does not capture.
- Using existing and secondary data requires a data integration and security plan.
- The limitations of existing/secondary data should be ameliorated in the design and analysis plan.
- Many perceived weaknesses of existing and secondary data should be reframed as strengths for pragmatic research.

Interrogate Data for Limitations

- Only have data from Mayo System and OMC. Patients who do not seek medical care are not eligible
- Patients may seek care outside the system of interest (e.g., no heart attack recorded does not mean one did not occur)
- Common lab measurements (e.g., BP, height, weight) were not routinely recorded in older era. Other lab measurements are not always common among young and healthy (e.g., creatinine). Those that do have measurements may be less “healthy”
- Even so-called healthy controls might have different distribution of covariates between compared to donors
- For follow-up data, how the data are collected differs between the two sources (survey with medical record follow-up and medical record abstraction)

Key Points

- Existing/secondary data can and should be used at many points in the lifecycle of pragmatic research
- Any source of data should be interrogated for the not only what it includes but also what it does not capture.
- Using existing and secondary data requires a data integration and security plan.
- The limitations of existing/secondary data should be ameliorated in the design and analysis plan.
- Many perceived weaknesses of existing and secondary data should be reframed as strengths for pragmatic research.

Ameliorate the Limitations

- As part of the study, we will survey all potential controls and donors about their health history before and after the index date (potentially > 20K surveys)
- Initially mailed survey, follow-up by phone
- Learn some exclusionary conditions prior to index date
- A source of data ascertained in the same way between the donors and the potential controls
- Note: this is not cheap

Ameliorate the Limitations

- As part of the study, we will also query the USRDS and NDI to ascertain date of ESKD and death dates
- Another source of secondary/existing data!
- A source of data ascertained in the same way between the donors and the potential controls
- Sources ascertained in the same way can help calibrate outcome data available from REP and UMN database which has been differentially assessed

Key Points

- Existing/secondary data can and should be used at many points in the lifecycle of pragmatic research
- Any source of data should be interrogated for the not only what it includes but also what it does not capture.
- Using existing and secondary data requires a data integration and security plan.
- The limitations of existing/secondary data should be ameliorated in the design and analysis plan.
- Many perceived weaknesses of existing and secondary data should be reframed as strengths for pragmatic research.

Data Integration and Security

- Multiple sources of information including REP, UMN database, survey results, USRDS, NDI
- Each needs to be housed consistent with their data use agreements
- Each institution (Mayo and UMN) must query USRDS and NDI separately
- Creating a central database at one institution requires careful understanding of what (de-identified) information can be shared

Key Points

- Existing/secondary data can and should be used at many points in the lifecycle of pragmatic research
- Any source of data should be interrogated for the not only what it includes but also what it does not capture.
- Using existing and secondary data requires a data integration and security plan.
- The limitations of existing/secondary data should be ameliorated in the design and analysis plan.
- Many perceived weaknesses of existing and secondary data should be reframed as strengths for pragmatic research.

Ameliorate the Limitations

- Not just data integration but also statistical integration of data from multiple sources: existing data (UMN database and REP), survey responses, national registries
- Some may give conflicting outcomes (presence/absence of the event, date of first occurrence, etc.)
- Follow-up will be different (e.g., time to last visit in REP may be much older than survey which has health information until the present)
- Some sources will be missing
- Active research to determine optimal statistical methods

Ameliorate the Limitations - Confounding

- Once we have identified subjects in REP as being healthy enough donation plan to re-match and/or adjust for common comorbidities and risk factors of CVD and ESKD (e.g., BMI, BP, etc.)
- Many controls will not have lab values available particularly from an earlier era. Consider several sensitivity analyses where we assume that these values are normal or exclude if those values not assessed

Key Points

- Existing/secondary data can and should be used at many points in the lifecycle of pragmatic research
- Any source of data should be interrogated for the not only what it includes but also what it does not capture.
- Using existing and secondary data requires a data integration and security plan.
- The limitations of existing/secondary data should be ameliorated in the design and analysis plan.
- Many perceived weaknesses of existing and secondary data should be reframed as strengths for pragmatic research.

Reformulate Limitations as Strengths

- Irregular and inconsistently collected data → data on outcomes that are clinically meaningful to patients (i.e., sought care for condition)
- Fewer data quality checks, more measurement error → data available to patients and physicians which they are using to make decisions
- Other limitations can be viewed as an exchange → in exchange for the challenge of integrating multiple data sources have a more complete understanding of health outcomes in a cohort without selection bias

Thought Questions

- What are some key barriers to using existing and secondary data in your research? How can they be overcome?
- How can the limitations of existing and secondary data be rephrased as relative strengths of the sources?
- What can methodologists do to improve the suite of available methods to make using existing and secondary data more palatable?

Conclusion

- Existing/secondary data can and should be used at many points in the lifecycle of pragmatic research
- Any source of data should be interrogated for the not only what it includes but also what it does not capture.
- Using existing and secondary data requires a data integration and security plan.
- The limitations of existing/secondary data should be ameliorated in the design and analysis plan.
- Many perceived weaknesses of existing and secondary data should be reframed as strengths for pragmatic research.

Thank you!

vock@umn.edu

@docvock

directory.sph.umn.edu/bio/sph-a-z/david-vock

SCHOOL OF PUBLIC HEALTH
UNIVERSITY OF MINNESOTA

The Heart of New Ulm (MN) Project (NOHU)

- Launched by Allina Health in 2009 supported by Minneapolis Heart Institute Foundation
- New Ulm: Rural community, single hospital and clinic (Allina)
- Population-based project
- Focus on cardiovascular health
- Triple Aim: Improve health care quality, population health & reduce costs

Interventions in HONU

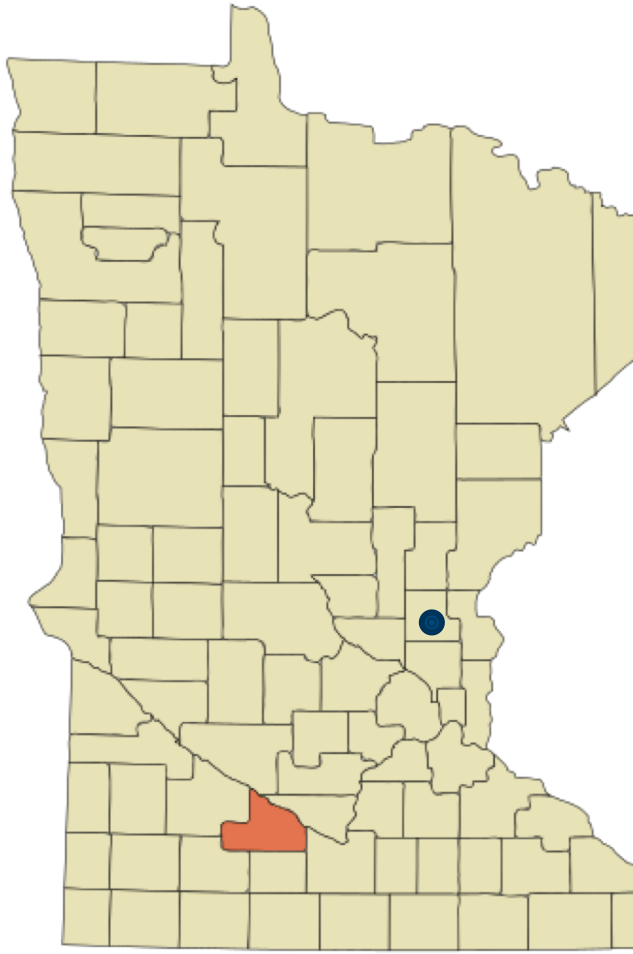
- **Community Interventions**: Heart Health Screenings, Community Health Summits, Formal Run/Walk events, Community Health Challenges, General Education, Small Community Events, Food Environment Improvements, Social Marketing Campaign
- **Health Care Interventions**: HeartBeat Connections, Heart & Vascular Prevention Clinic, Weight Management Phone Coaching, Grand Rounds
- **Worksite Interventions**: Worksite Assessments, Heart Health Screenings conducted at worksites, Worksite behavioral change programs, Business leader engagement and education

Surveillance of Effectiveness of Intervention

- Population of interest: All adults 40-79 years of age living in New Ulm zip code (~7855 residents)
- Active follow-up to assess risk factors on the target population or a sample would have been prohibitively expensive
- Allina operates only clinic and hospital in NU → routine care on target population captured in EHR
- EMR used to monitor changes in CVD risk factors

Comparison Study

- No perfect control group
- Cambridge, MN: rural community with single hospital and clinic
- Challenge: Cambridge population may be different than the New Ulm population in terms of health



COPRH Con
Colorado Pragmatic Research in Health Conference

ACCORDS

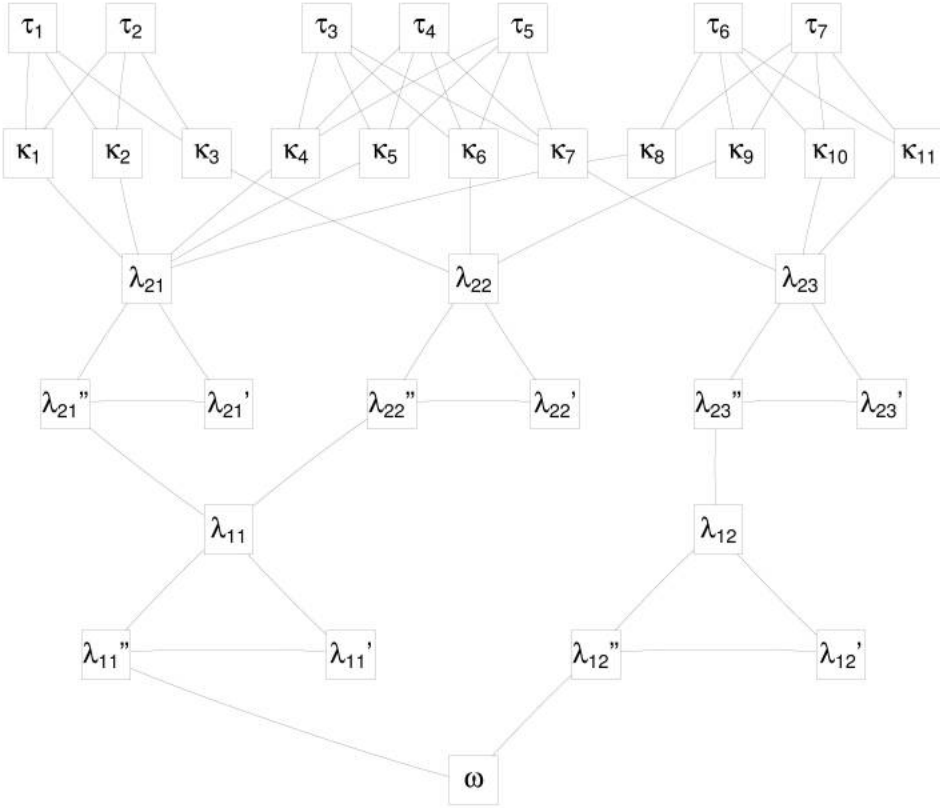
CCTSI

Matched Pairs Design

- One solution would be for every person in New Ulm find someone “identical” to them in Cambridge in 2008/09
- “Matched pairs” form the analysis cohort
- Idea is that the “matched pairs” are identical at baseline so any differences subsequently are due to intervention

Key Questions & Answers

- Key Questions/Answers:
  1. What characteristics should we consider when trying to match two people?: Age, gender, systolic blood pressure, low density lipoprotein cholesterol, glucose, BMI, clinic visits
  2. What metric should we use to define how similar two people are?: Mixture of exact matching on a few covariates (age range and gender), near fine balance on categorization of other covariates, and then Mahalanobis distance
  3. What algorithm should we use to find match pairs? Sparse optimal matching using network flow optimization algorithms



COPRH Con
Colorado Pragmatic Research in Health Conference

ACCORDS

CCTSI

Analytical Approach and Challenges

- Information on CV risk factors extracted from EMR in two-year periods – 08/09 (baseline), 10/11, 12/13, 14/15
- Fit a longitudinal model of risk factor trajectory with factors for time period, city, and city/time periods interaction using mixed models
  1. Accounts for correlation between repeated measures
  2. Allows for missing outcomes in some periods

HONU Comparator Study

CVD Risk Factor	Location	2008/09	2010/11	2012/13	2014/15	p-value
BP at Goal (<140/90 mm/Hg)	NU	79.2%	81.4%	83.5%	85.4%	<0.001
	Camb	80.1%	80.7%	81.4%	82.0%	
LDL at Goal (< 130 mg/dL)	NU	76.8%	76.2%	75.5%	74.9%	0.002
	Camb	77.9%	76.0%	73.9%	71.8%	
Glucose at Goal (<100 mg/dL)	NU	54.9%	50.5%	46.2%	41.9%	0.490
	Camb	56.7%	52.0%	47.2%	42.4%	
Major Adverse Cardiac Events	NU	1.2%	1.2%	1.3%	1.4%	0.088
	Camb	1.0%	1.2%	1.5%	1.8%	

COPRH Con Colorado Pragmatic Research in Health Conference

ACCORDS CCTSI

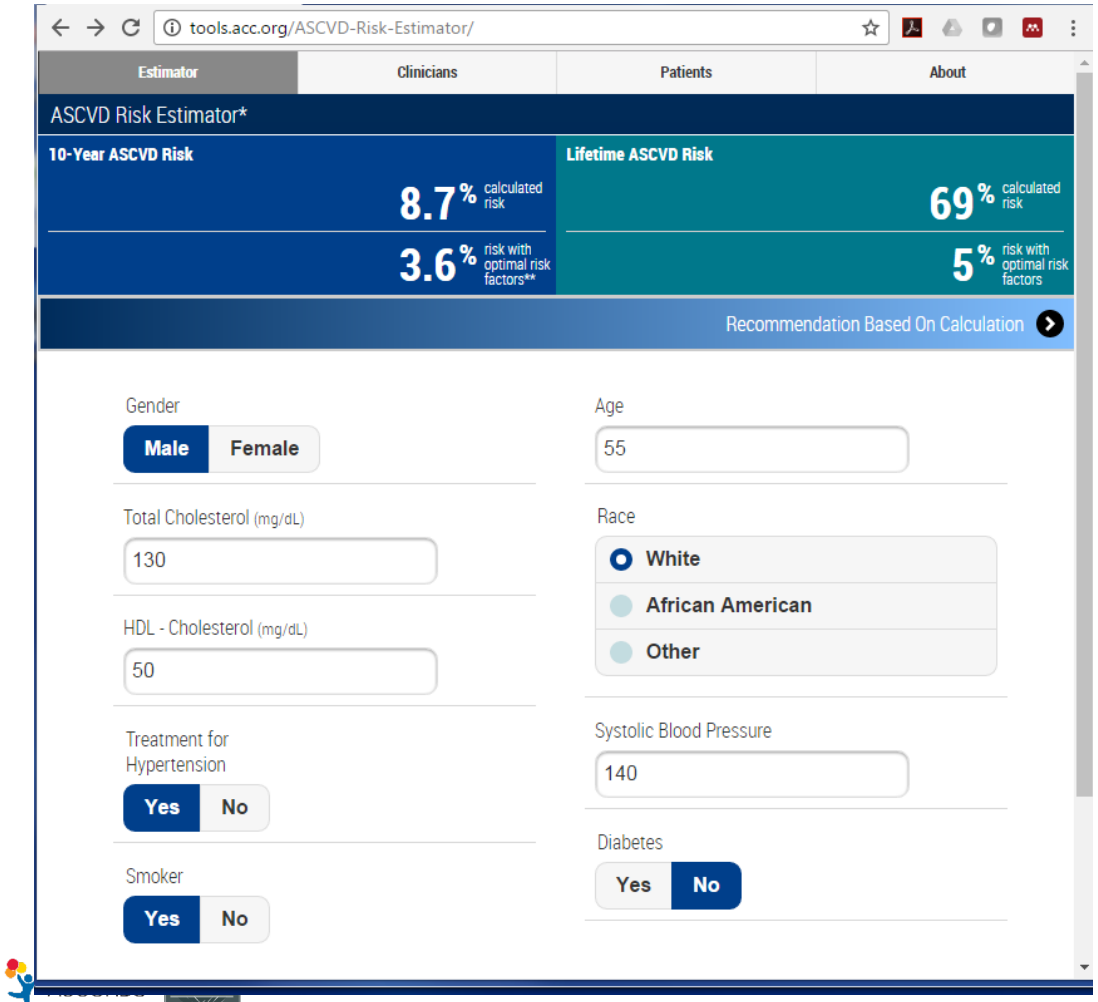
Sidebottom AC, **Vock DM**, Sillah A, Pereira R, Benson G et al. (2016) In Press

Cardiovascular Risk Prediction

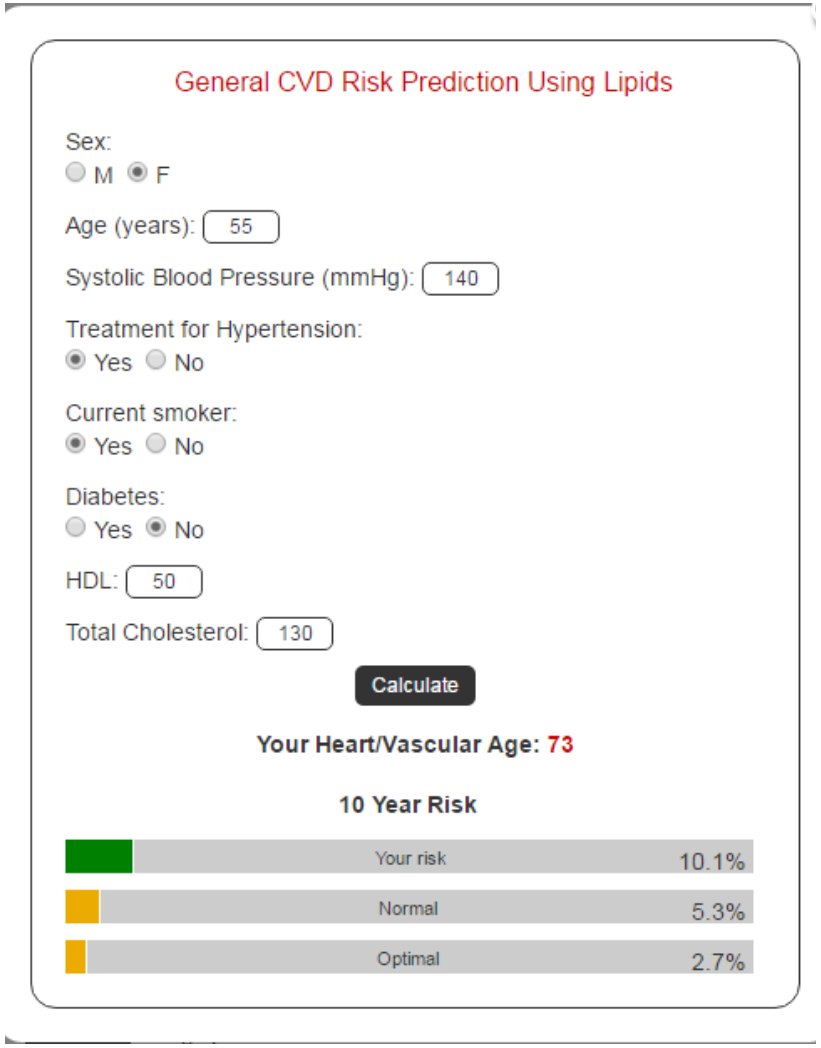
- Clinical risk prediction: Given some information about a patient (e.g., gender, blood pressure) gives the probability of an outcome (e.g., heart attack) over a specific time period (5 years)
- Systematic reviews found that there are over 100 risk models produced between 1999 and 2009 (Cooney et al. 2009, 2010; Matheny et al. 2011)
- Common ones include Framingham, SCORE, ASSIGN-SCORE, QRISK1, QRISK2, PROCAM, WHO/ISH, Reynolds Risk Score, AHA/ACC Pooled Cohort Equations

Cardiovascular Risk Prediction

AHA/ACC Pooled Cohort Equations



Framingham Risk Score



Importance of Cardiovascular Risk Prediction

- Raise awareness of the substantial burden of cardiovascular disease (CVD) and risk factors associated with CVD
- Help clinicians prioritize care and motivate patients to remain adherent to any interventions
- Recent AHA/ACC guidelines for statin therapy are based 10-year risk prediction of cardiovascular events
- Risk prediction will be routinely integrated in primary care as part of Decision Support Systems

Evaluating a Risk Prediction Model

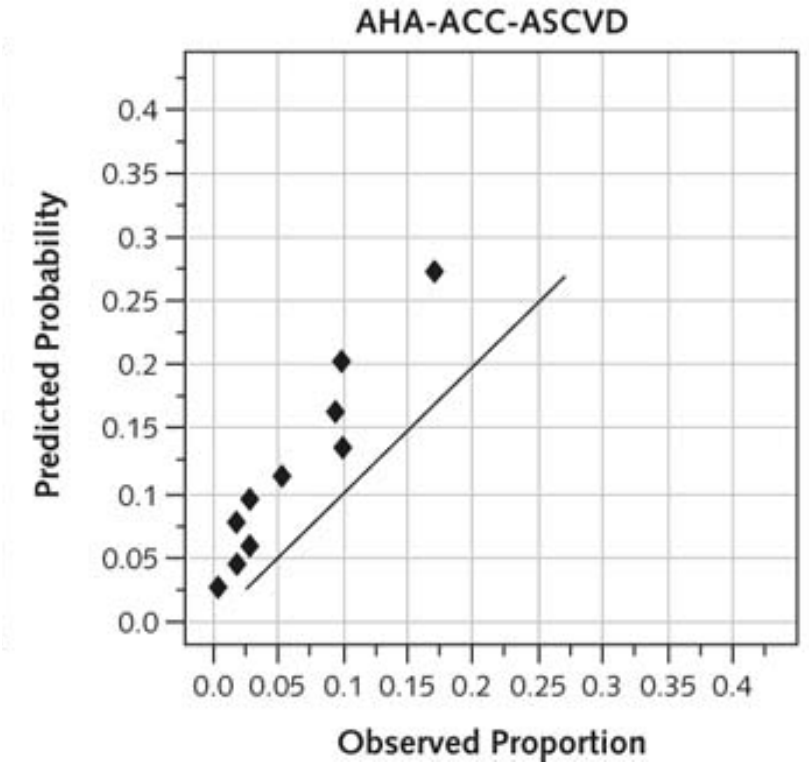
- **Calibration**: Are the predictions accurate? Do 20% of people with 20% risk experience CVD?
- **Discrimination:** Can we separate the high risk from the low risk patients?

Controversy Over AHA/ACC Pooled Cohort Equations

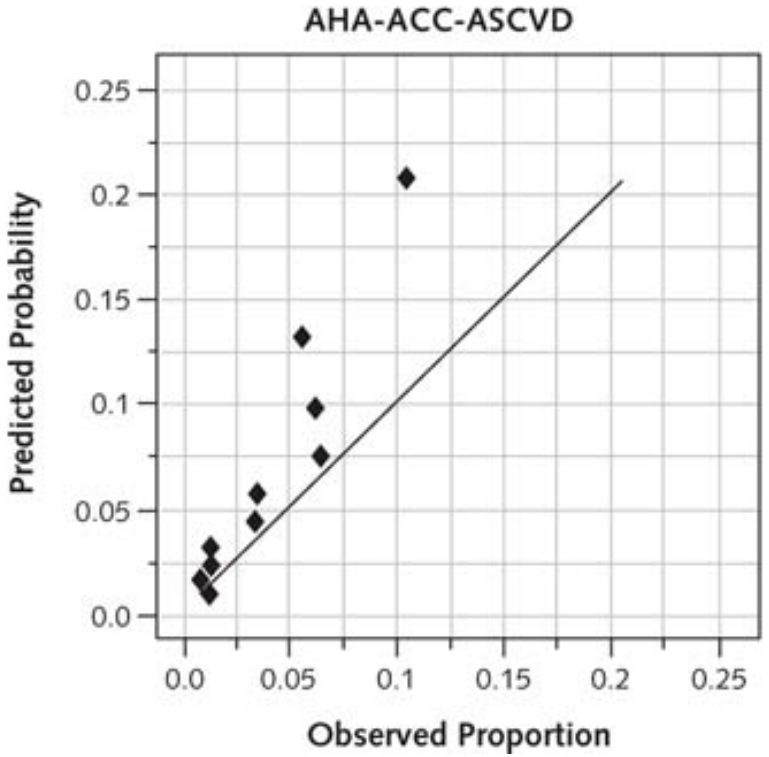
- AHA/ACC PCE were developed in late 2013 using data from several epidemiological cohort studies
- In addition, an expert panel recommended that statin (e.g., Lipitor) medication for high cholesterol **OR** 10-year risk > 7.5%
- Several studies have questioned whether the PCE are well-calibrated

Calibration in MESA Cohort (*DeFilippis et al. Annals Intern Med.)*

- Graph #1

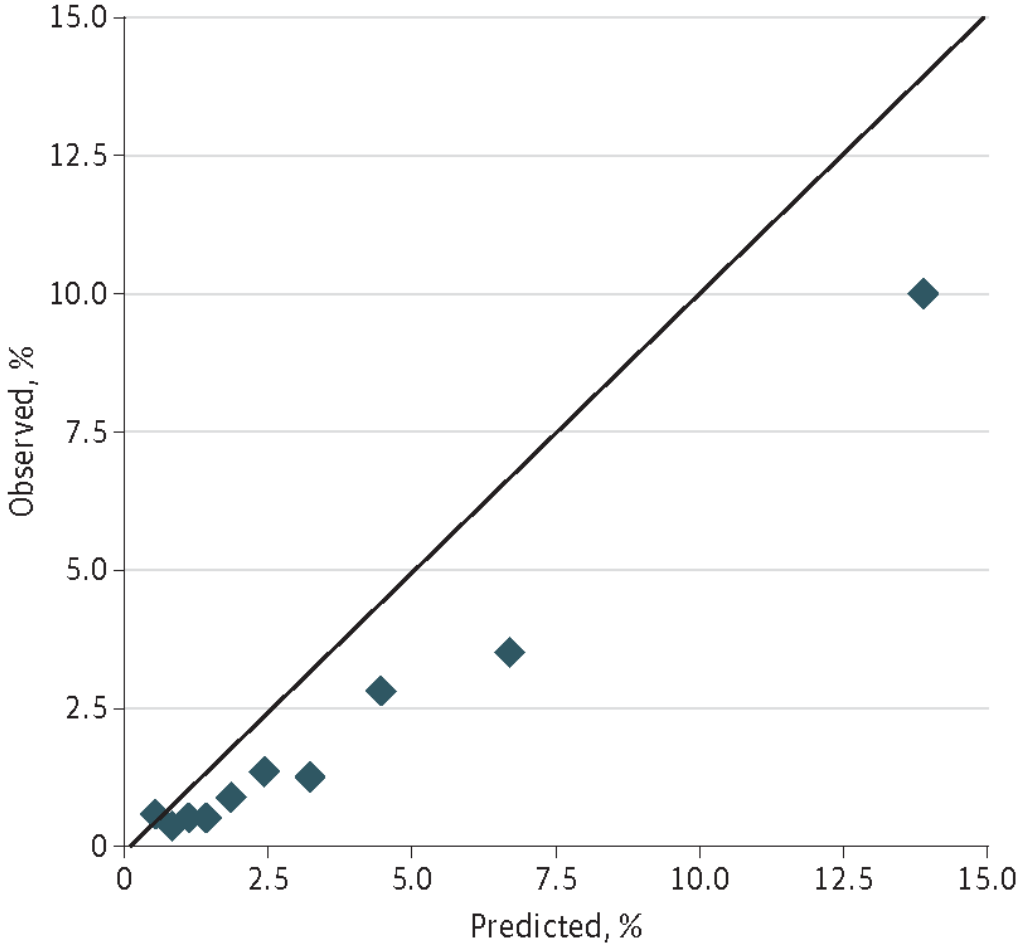


Men

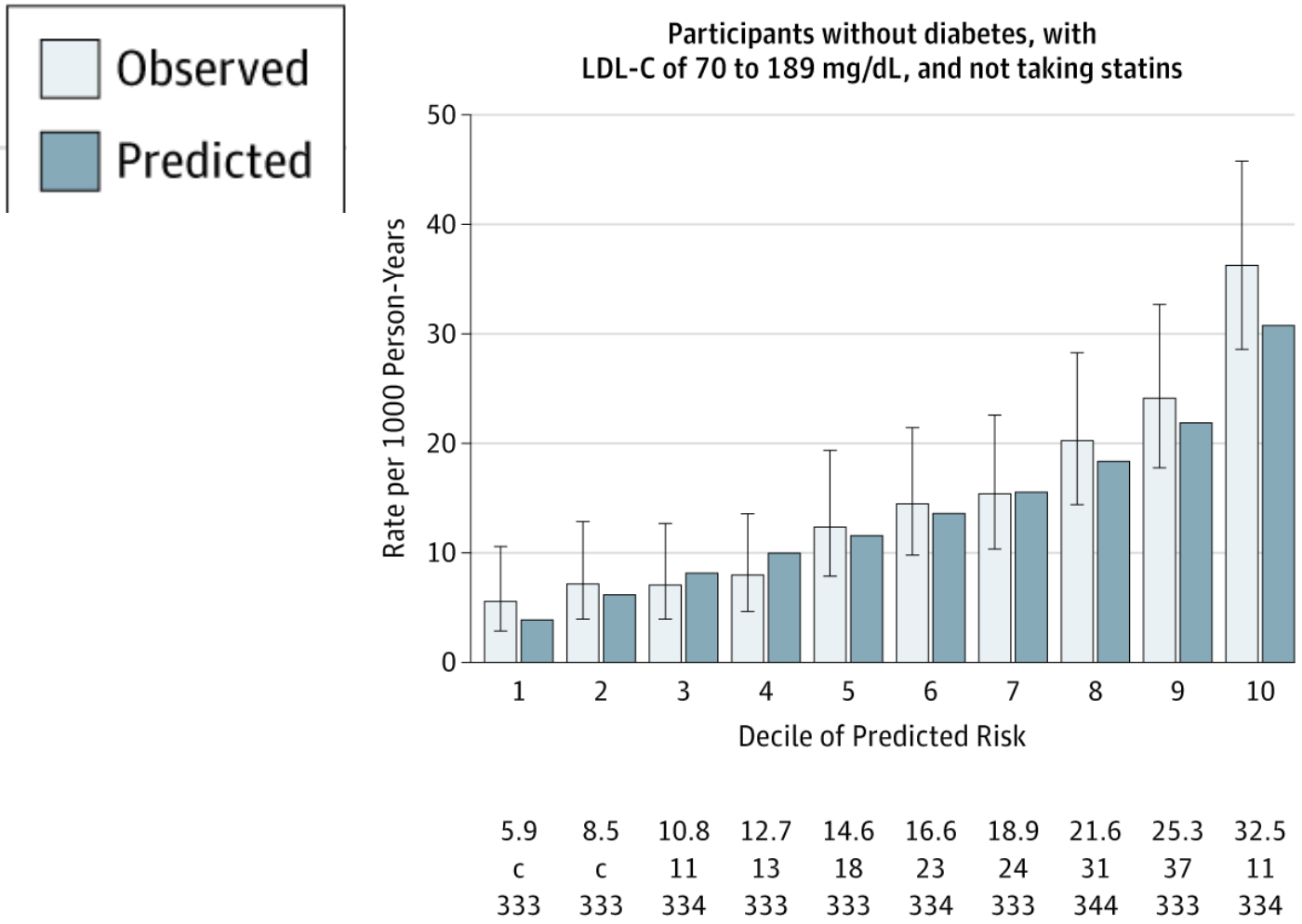


Women

Calibration in Women's Health Study (Cook and Ridker, *JAMA Intern Med)*



Calibration in REGARDS Cohort (Muntner et al. *JAMA*)

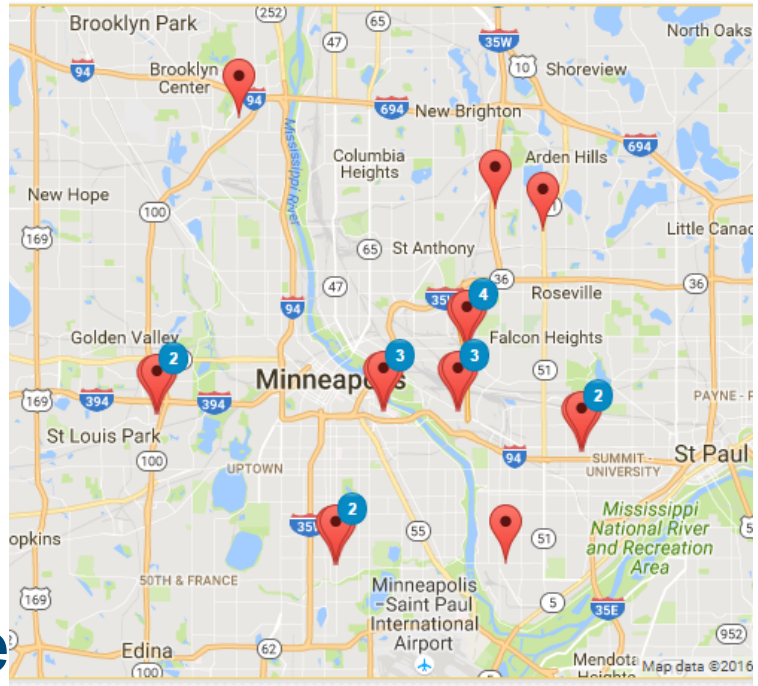


Limitation of Previous Validation Studies

- Validation studies performed in epidemiological cohort studies
- Procedures and schedule for obtaining data different in routine clinical practice
- Many cohorts are relatively homogeneous (e.g., racial/ethnic, comorbid conditions, limited age ranges)
- Cohorts include subjects from over 40 years ago (diet, interventions, etc. have changed)
- Likely cohort selection effects

CVD Risk Prediction Using EMR

- HealthPartners: Twin Cities-based healthcare delivery organization
- Operate a network of clinics and hospitals AND insurance plan
- Open and partially overlapping system
- Predict CVD risk and incorporate into Clinical Decision Support



Wolfson J, **Vock DM**, Bandyopadhyay S, Vazquez-Benitez G, Johnson PE, Adomavicius G, et al.(2016) J Am Heart Assoc (In Press)

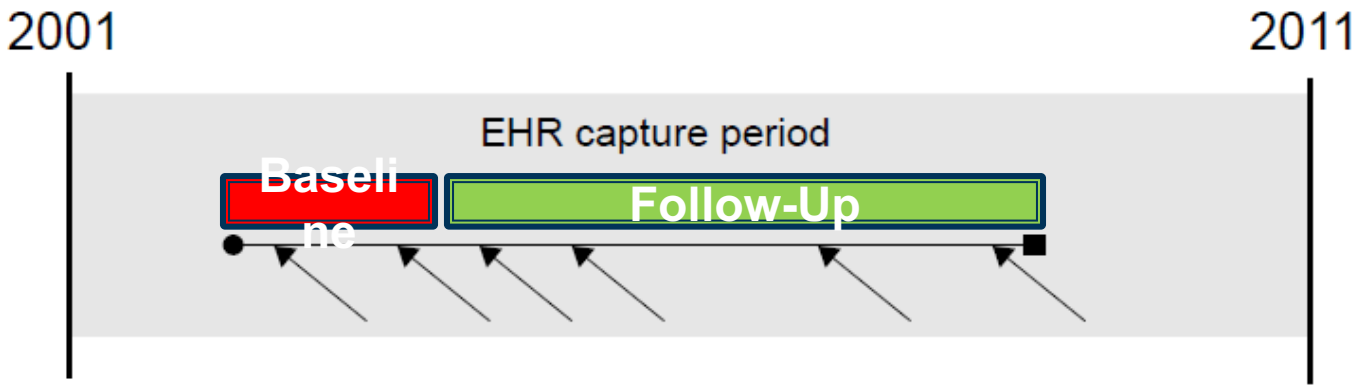
Defining the Cohort of Interest

- Goal is to use risk prediction equations in primary care clinic
- Patients with two medical encounters in the in-network ambulatory clinics (non-urgent care) with blood pressure information at least 30 days but at most 1.5 years apart
- Insurance and drug coverage for at least one year
- Age 40-79 years
- ~86,000 patients

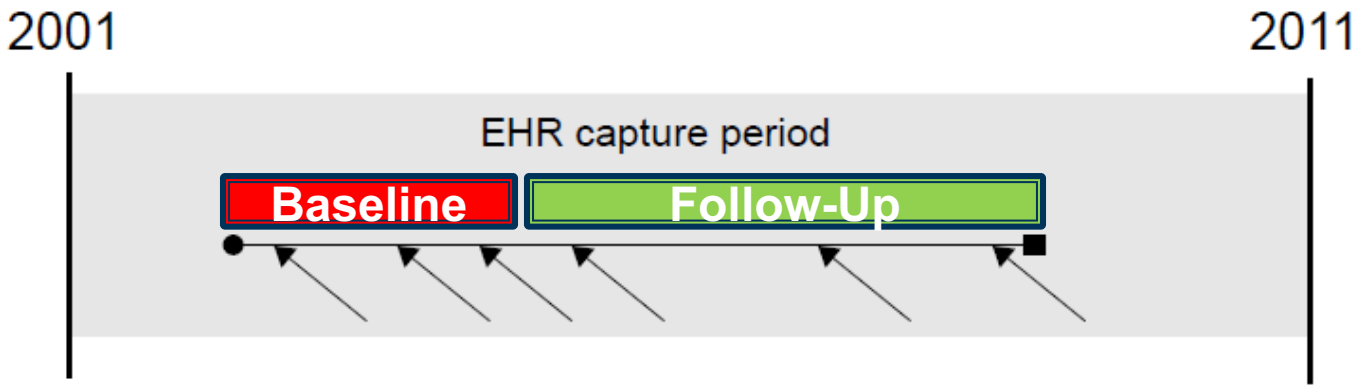
Baseline Versus Follow-Up

- No predefined baseline visit where risk factor ascertainment occurs and follow-up begins
- Not all risk factors may be collected during a single encounter → want to have a baseline period where we observe risk factors
- Longer baseline period: more risk factors collected but leads to shorter follow-up and clinical status may not be constant over that time period
- Visits are irregular so we have some leeway on how when baseline begins

Baseline Versus Follow-Up



Baseline Versus Follow-Up



Baseline Versus Follow-Up

- Baseline period consisted of the time between the first blood pressure reading during the enrollment period and the date of the final blood pressure reading at most 1.5 years from the first measurement.
- Follow-up period for a patient begins at the end of the baseline period and continues until
  1. Patient experiences a CV event
  2. Patient disenrolls from the insurance plan for more than 90 days
  3. Data capture period ends (in 2011), whichever comes first.
- Covariates were generally averaged over the baseline window

Outcome of Interest

- Obtained from diagnosis and procedure codes recorded by physicians as part of the medical record or insurance claims
- Can typically infer from claims information any major medical events even if care is not sought in-network
- Non-adjudicated outcomes

CVD Risk Prediction Using Data from the EMR is Accurate

Framingham Risk Score

5-year predicted risk group	Total N	Predicted Events, N (Rate)	Observed Events*, N (Rate)
0-2.5%	16,574	247 (0.015)	247 (0.015)
2.5-5%	11,885	427 (0.036)	449 (0.038)
5-7.5%	5,694	348 (0.061)	371 (0.065)
7.5-10%	3,050	263 (0.086)	317 (0.104)
>10%	4,855	782 (0.161)	757 (0.156)

Pooled Cohort Equations

5-year predicted risk group	Total N	Predicted Events, N (Rate)	Observed Events*, N (Rate)
0-2.5%	28,921	257 (0.009)	373 (0.013)
2.5-5%	6,650	234 (0.035)	273 (0.041)
5-7.5%	2,842	173 (0.061)	158 (0.056)
7.5-10%	1,471	127 (0.086)	108 (0.074)
>10%	2,174	321 (0.148)	254 (0.117)

Wolfso s G, et al.(201

Conclusions and Limitations

- Standard CVD risk prediction equations using data collected from primary care are (relatively) well-calibrated
- Events inferred from claims data → likely over-identifies CV events
- REGARDS study used Medicare claims to supplement CV event identification