Using population-based data for secondary analysis

Breakout session for the Colorado Pragmatic Research in Health (COPRH) Conference:

Implementation & Conduct of Pragmatic Research: Ensuring Rigor and Relevance in Practice

May 26, 2021

- ***Arthur Davidson, MD, MSPH***
- ***Allison Kempe, MD, MPH***

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

Learning Objectives

- Describe and categorize an array of population-based data resources for use in secondary analyses
- Describe the advantages and disadvantages of secondary datasets and how to access population-based data
- Describe examples of important publicly available datasets

Breakout Session Outline

- Orientation – the value of an Analytic Plan
- Brief introduction to secondary data
- Some publicly available databases
- Example of using secondary data – immunizations
- Interactive exercise

Orientation – the value of an Analytic Plan

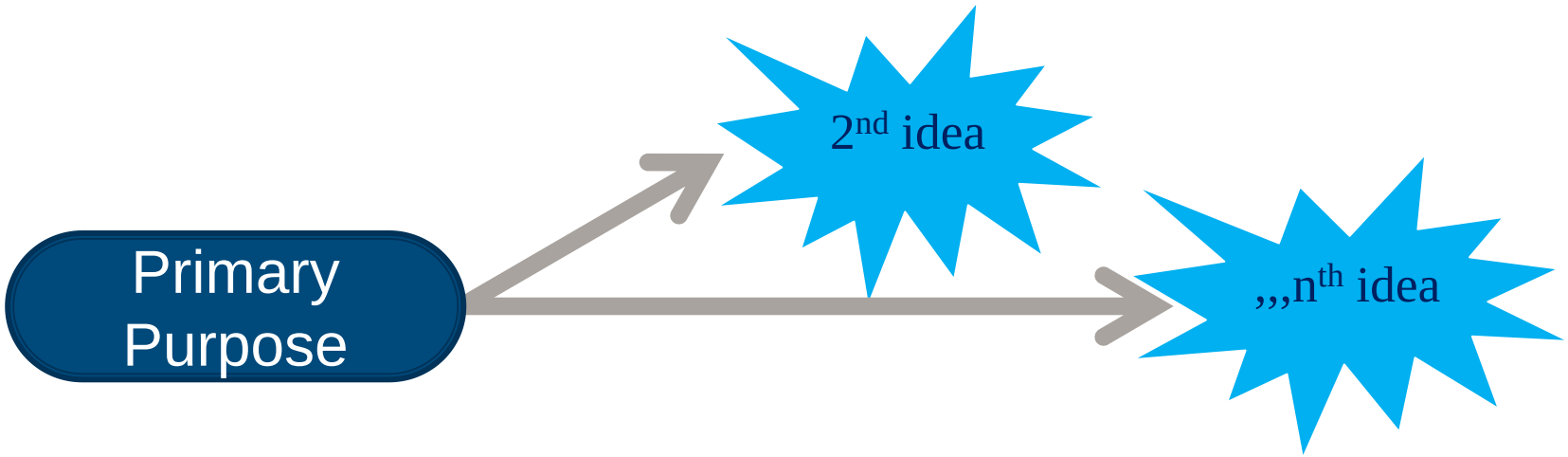
Goal: Think of a research question of interest to you or within your current focus area

- Listen to this short presentation in the context of that research question
- Look at the analytic plan template (on website); sharpen and elaborate the research question (*after the session*)
- Consider available secondary dataset(s) to assist in answering that question and any associated issues (e.g., unit of measurement, data required, target population, issues of privacy)
- Use the immunization example to think critically of potential challenges/problems related to your question and potential dataset(s) under consideration

What are “Secondary” Data?

Context: computer-readable, data libraries and/or archives

- **Secondary data** implies data were collected for a primary purpose, then made available for research conducted by other individuals/groups for a purpose other than that for which it was collected.



Examples of Secondary Data

- Primary Data: treatment/research/surveillance data collected by clinicians, investigators or organizations used for other purposes.
- Administrative Data: by-product of administering health services, health insurance member enrollment, health services reimbursing
    - Utilization claims data (All Payer Claims Database), Medicare, Medicaid
- Official Statistics: data collected/processed by government agencies; often available serially; uniform across regions

  State:
    - vital statistics – birth/death records
    - immunization information systems (IIS)
    - communicable diseases, cancer registry

  Federal:
    - US Census (decennial), American Community Survey (annual)
    - Nation Center for Health Statistics (CDC), Agency for Healthcare Research and Quality
- Routine Statistics: derived from analysis of large data sets in health and social services - large volume, available early/late?
    - hospital discharge data: Health Care Utilization Project, State In-patient Database, National Inpatient Sample, Colorado Hospital Association

Value of Secondary Data Sets

- Large
- National (and state) samples
- Population-based
- Decreased bias
  - *Systematic/Systemic or Selection bias: inherent tendency of a process to favor particular outcomes or with scientific observations to result in systemic error.*
- Less costly than primary data collection

Secondary Data Sets: Advantages ….. Caveats

1. Low cost …
   - as long as you don't need to wait for it, merge it, correct it, or impute missing values
2. Large N…
   - may still be inadequate size for relevant subgroups
   - statistical > clinical significance
3. Representative…
   - as long as you know who gets into data base, and how long they remain
4. Timely…
   - given access and analytic expertise

Secondary Data Sets: Disadvantages ….. *Cautions*

- Data not collected for research purposes; less attention to data accuracy and consistency than for primary research data. Unless you collect the data, you don't understand its "warts"
- Data not collected at specified time points; need to consider time-dependency of data (e.g., intermittent enrollment)
- Relevant events (e.g., hospitalization) may not be included in the data set
- Missing data, in non-random patterns
- Lack clinical detail or precision (e.g., ICD-9/10 codes only)
- Difficult to control for bias and effect modification (e.g., comorbidity)
- Issues of access: privacy, data sharing agreements
- Risk of drowning: need to stick to clearly formulated hypotheses
- You are a prisoner of the data: may provide right answer to wrong question

What help might be required?

If you attempt research with these datasets, you will rapidly discover:

- few datasets come prepared for immediate access by your favorite application,
- getting your data in an appropriate form for analysis is complicated, and
- organizing is often more complicated than the most complex statistical analyses you will employ. ***(use the analytic plan template to organize)***

KEY LESSION: ***Earlier and greater specificity of research question diminishes time to organize***

Using Secondary Data

OBJECTIVE: define a research question and design that test the hypotheses using existing data.

Our focus

Types of data:

- small/simple survey, (e.g., public opinion poll, for social or political attitudes)
- extensive/complex data, such as population-based surveys/reporting (*see Secondary Datasets Handout for examples*)
- administrative and/or clinical data

CHALLENGE: ***assure data appropriately address your question to avoid altering hypothesis to fit the data.***

COPRH Con
Colorado Pragmatic Research in Health Conference
ACCORDS
CCTSI

Using Population-based, Survey Data

QUESTIONS: *(ask about these factors when considering secondary data sources and your hypothesis)*

- appropriateness of the study's unit of analysis (e.g., patient, provider, system)
- sampling methods
- variables
- values, and
- levels of measurement.

e

Finding the Data (*general characteristics*)

	Details about:
www.cdc.gov	➡ behaviors/diseases/ outcomes
www.cdc.gov/nchs/	➡ care/outcomes
www.ahrq.gov	➡ cost/quality/safety
www.census.gov	➡ base population

COPRH Con
Colorado Pragmatic Research in Health Conference
ACCORDS
CCTSI

Finding the Data

Process:

- less emphasis on computer or technical knowledge,
- still must locate data that suits research needs,
- time-consuming and meticulous process,
- use on-line tools and knowledge and experience of others to assist in this process, and
- beware, for every step of finding the right data, there are potential "*gotchas*"
- give yourself plenty of time not only to locate your data but to review and archive any associated information (e.g., technical documentation and codebooks).

Accessing the Data

CODEBOOK or manual describing a particular study or data collection typically describes:

- data collection and sampling design,
- variables contained in the data,
- surveys, instrument or questionnaire used to solicit answers from respondent and coded values of each question,
- location and format of variable within the raw data file.

Most codebooks have multiple sections:

- **how to read** the codebook
- **data dictionary** listing variables and column locations
- **data collection instrument(s)**

Some National Center for Health Statistics Datasets

Population Surveys

- National Health and Nutrition Examination Survey
- National Health Interview Survey
- National Survey of Family Growth

Vital Records

- National Vital Statistics System
- National Death Index
- Vital Statistics Rapid Release

Provider Surveys

- National Health Care Surveys
- National Ambulatory Medical Care Survey
- National Electronic Health Records Survey
- National Hospital Ambulatory Medical Care Survey
- National Hospital Care Survey
- National Post-acute and Long-term Care Study

Some Agency for Healthcare Research and Quality Datasets

Compendium of U.S. Health Systems, 2016
Consumer Assessment of Healthcare Providers and Systems (CAHPS®)
HCUPnet
Kids' Inpatient Database (KID)
Medical Expenditure Panel Survey (MEPS)
National (Nationwide) Inpatient Sample (NIS)
National Healthcare Quality and Disparities Report
Nationwide Emergency Department Sample (NEDS)
Nationwide Readmissions Database (NRD)
State Ambulatory Surgery and Services Databases (SASD)
State Emergency Department Databases (SEDD)
State Inpatient Databases (SID)
Trends in Opioid-Related Hospitalizations
United States Health Information Knowledgebase (USHIK)

Some Centers for Disease Control and Prevention Datasets

National Center for Health Statistics (see other slide)
Injury and Violence
National Notifiable Diseases Surveillance System
Vaccinations
Smoking and Tobacco Use (*Behavioral Risk Factor Surveillance System – BRFSS*)
Pregnancy and Vaccination
Disability and Health

Examples of Using an Existing Database for Retrospective and Prospective Data Collection

- Immunization Information Systems (IIS) aka “Immunization Registries”
- Confidential, population-based, computerized databases that record Iz doses administered by participating providers within a given geopolitical area (often state or region)
- IISs capable of bilateral exchange of Iz information with immunization healthcare providers to consolidate records
- 62 State or regional IIS in the U.S.!!! (all states and a few extras..)
- Each state or regional jurisdiction makes decisions about how CDC standards are interpreted and implemented

Immunization Information Systems (IIS)

- State laws and regulations differ with respect to privacy, mandated reporting of vaccinations for different age groups, inclusion rules (implicit consent, explicit consent, opt-in and opt-out) and statutes about outreach
  - Implications for completeness of data
  - Implications for potential interventions using IIS

- IISs differ with respect to HL7 connectivity with practice sites: 30% report they receive 75-100% of data via real-time HL7
  - Implications for completeness of data
  - Implications for how to access data

Immunization Information Systems (IISs): Policies around Reminder/Recall (R/R)

- Reminder/recall (R/R) is the use of different modalities (e.g. auto-dialer, mail, text, e-mail to remind patients about upcoming Izs due or recall them for Izs past due
- 27% of states have specific legal mandate allowing IIS or health department to send out R/R to all populations, 11% only among certain populations, and 62% no legal mandate exists to allow R/R
- Federal Communication Commission (FCC) implemented the Telephone Consumer Protection Act (TCPA) in 1991--several revisions have created additional restrictions on calls and texts going to cell phones
- Interpretations of TCPA vary by state and within states, therefore IIS needs to determine what if any risks incurred by R/R if phone methods used

Colorado's IIS (CIIS): Is it a Useful Secondary Dataset? (Depends on the Question….)

- How complete? For children very complete!
  - 85% of vaccinations to persons 0-18 years (2019)
  - All children <6 years have Izs in CIIS
  - Opt-out policy for children and birth VS used to populate CIIS
  - Adult data from private sector far less complete

- How UTD? All public sites and majority of private sites have HL7 connectivity with rapid or immediate upload of Izs

- Can it be used for outreach interventions? CIIS has statutory authority to conduct R/R to increase Iz rates

- Publicly available? NO, unless you are a provider, therefore partnership is key! CIIS Leadership interested in assisting with research <u>if it is consistent with their mission and doesn't increase workload!</u>

COPRH Con
Colorado Pragmatic Research in Health Conference

ACCORDS

Benefits to IIS of research collaborations

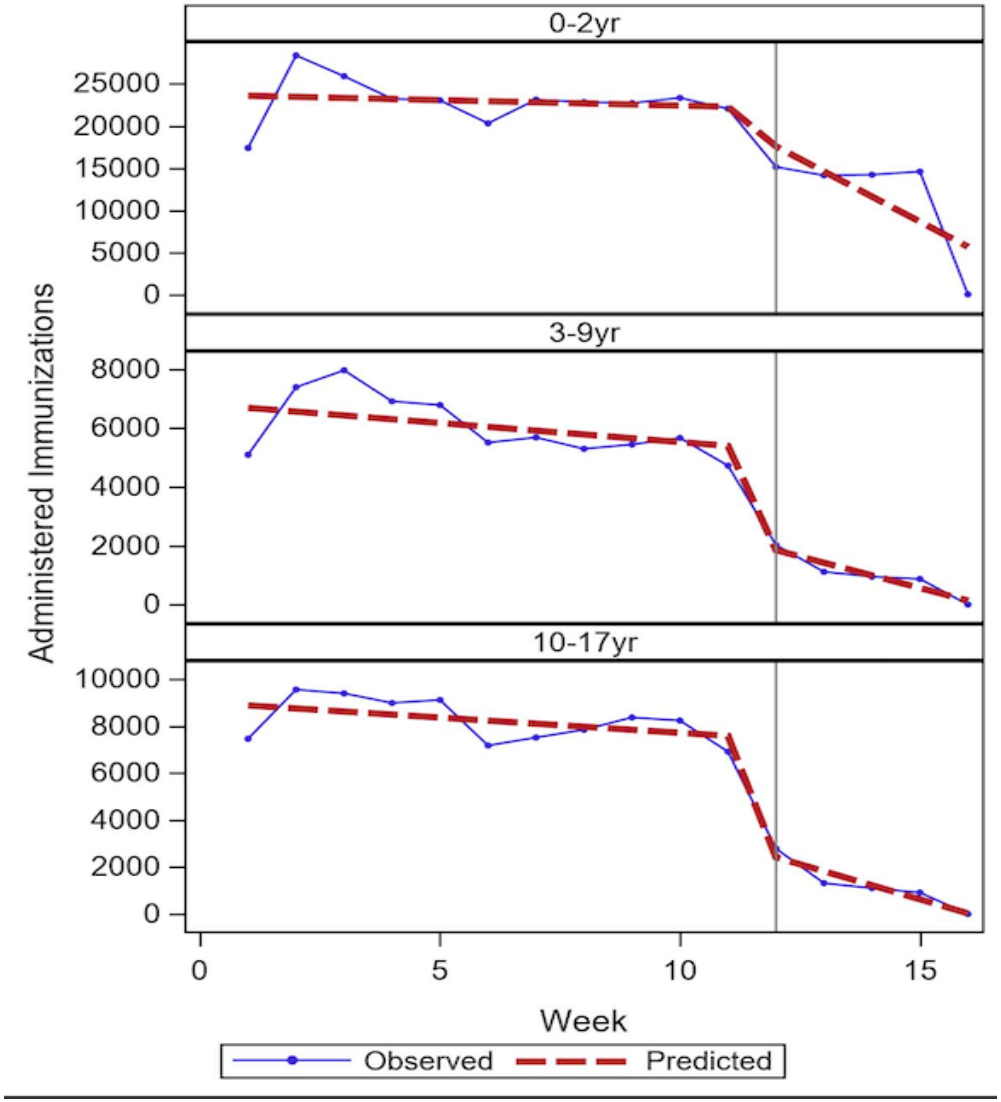
- Populating the IIS with more patient information
- Increasing the accuracy of contact and immunization information
- Increasing data on new groups (e.g., adolescents and adults)
- Increasing participation of new providers, especially adult providers
- Potentially increasing sources of funding for IIS
- Direct financial support for IIS personnel
- National reputation, recognition as leaders in IIS innovation

A Retrospective Secondary Data Analysis Using CIIS—Example 1

Number of Childhood and Adolescent Vaccinations Administered Before and After the COVID-19 Outbreak in Colorado

O'Leary ST, Trefren L, Roth H, Moss A, Severson R, Kempe A

JAMA Pediatr 2021



A Large Multi-County Pragmatic Trial—Example #2

- To increase immunization rates in children before they enter Kindergarten

- To answer a question that had both local and national significance from PUBLIC HEALTH perspective:
  - Is IIS-based recall more effective in increasing population immunization rates when conducted by individual practices or centrally by state health department?

COPRH Con
Colorado Pragmatic Research in Health Conference

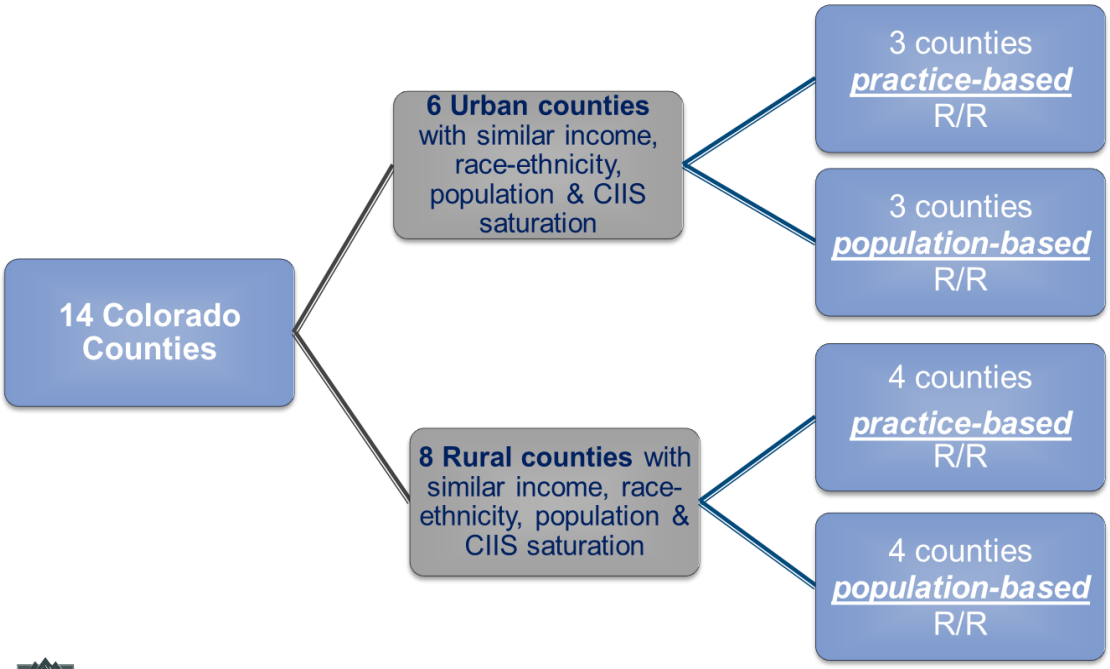
ACCORDS

CCTSI

Project Overview

- Population of interest: 19-35 month olds overdue for one or more immunizations

- Study Design: a group-randomized trial at the level of the county comparing two approaches to recalling children overdue for Izs
  - ***Practice-based recall*** at individual primary care practices
  - ***Population-based*** (geographic-based) recall at the level of the county

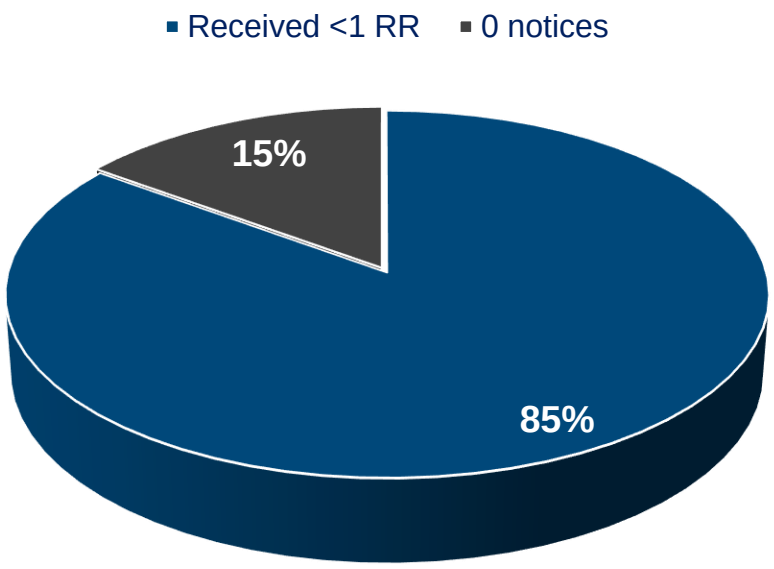


COPRH Con
Colorado Pragmatic Research in Health Conference

ACCORDS

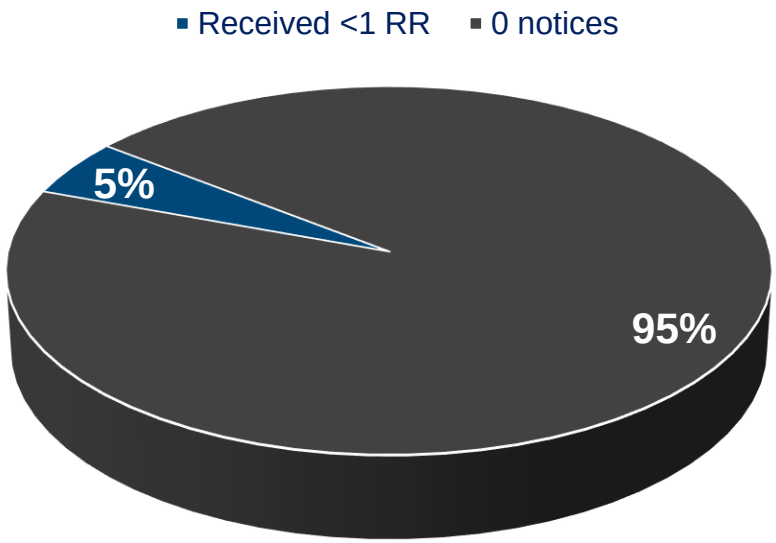
Reach of Intervention in Population-based Counties versus Practice-based Counties

Population-based Counties
188 practice sites
12,832 children eligible



*Assuming 85% receipt

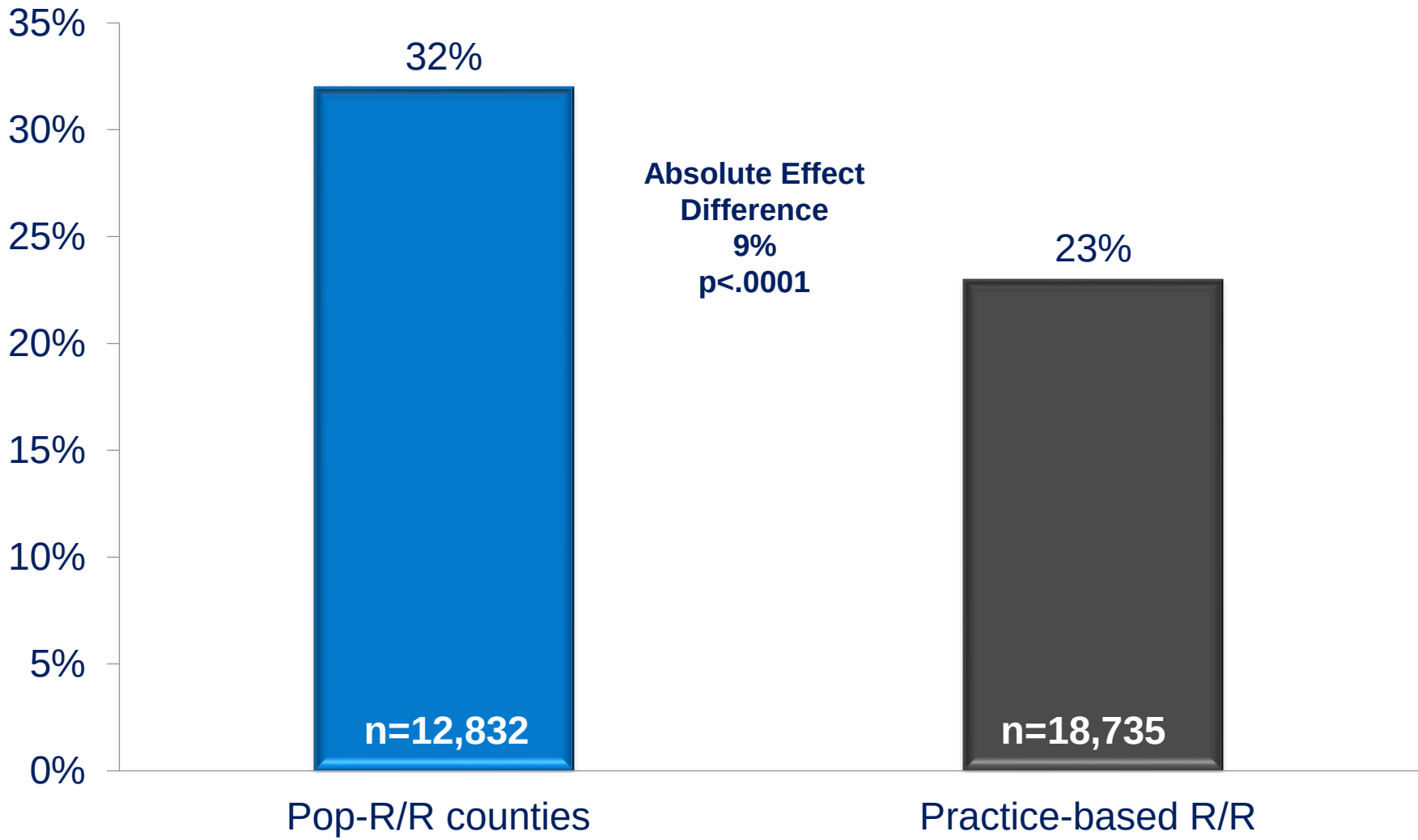
Practice-based Counties
195 practice sites—10 did recall
18,735 children eligible



**Assuming 100% receipt

COPRH Con
Colorado Pragmatic Research in Health Conference
ACCORDS
CCTSI

Percent Receiving Any Vaccine within 6 months

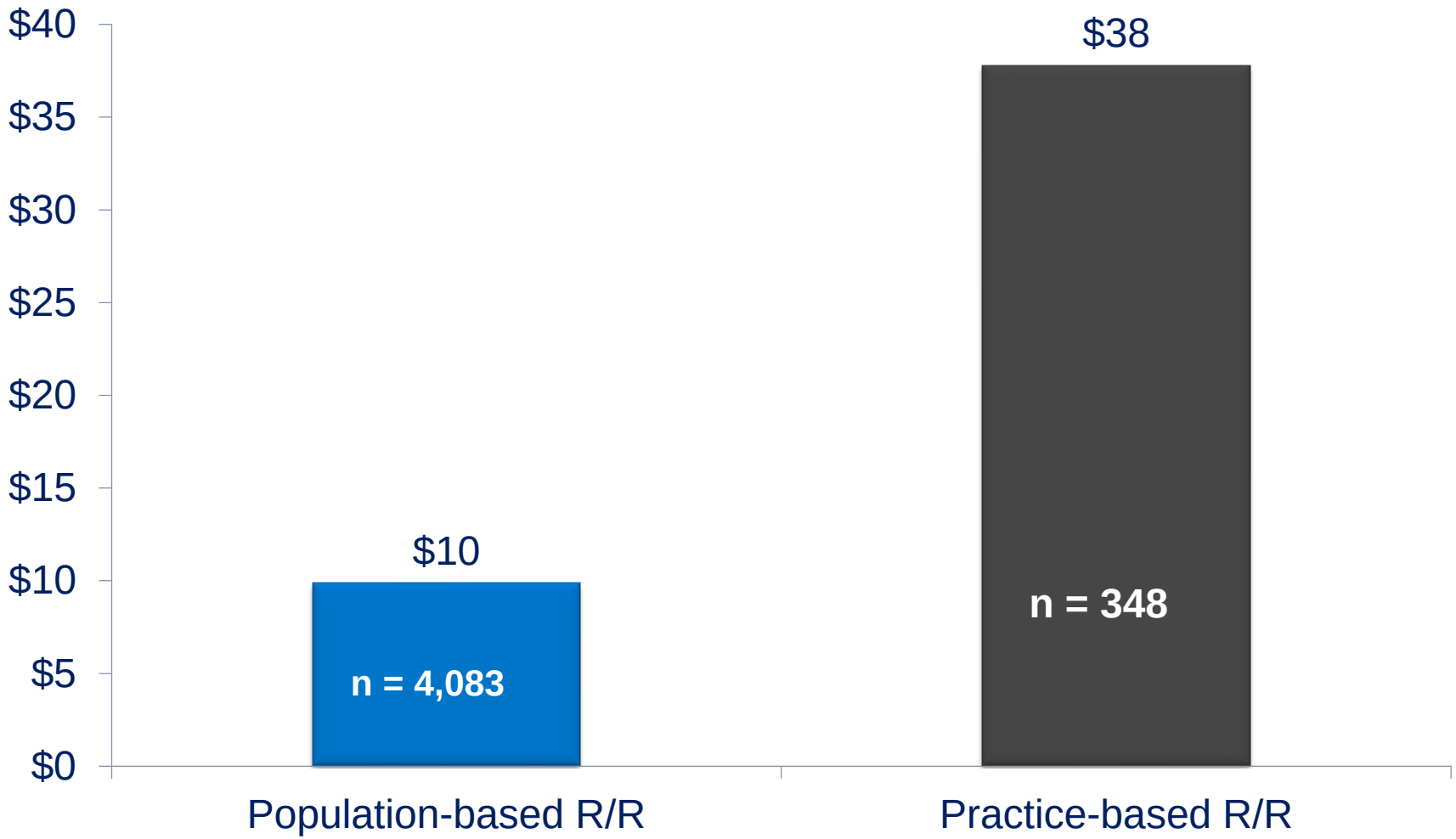


COPRH Con
Colorado Pragmatic Research in Health Conference

ACCORDS

CCTSI

Cost of R/R Per Child who Received ≥1 Vaccine



COPRH Con
Colorado Pragmatic Research in Health Conference

ACCORDS

Additional Centralized R/R Studies

- Practices and IIS collaborated by including practice names on centralized R/R messages showed better results
- Centralized R/R effective for adult vaccines IF data base can be populated with adult vaccines
- Centralized R/R for other vaccines in NY and CO—R/R for HPV and Influenza were less effective than for childhood vaccines because of higher degrees of hesitancy
- Centralized R/R within an ACO/HMO—more effective if choice-driven

Could We Have Used the NIS for These Examples?

- Example #1
  - Required longitudinal daily/weekly data pulls providing “snap shot” of Iz delivery as the pandemic progressed
  - NIS is yearly at the level of the state and substantial delay between survey and availability of data
- Example #2
  - Required Iz levels at the level of the county AND at the level of the practice
  - Data pulls needed to be timed with respect to the trial
  - NIS not flexible with respect to unit of analysis or timing of data pulls

How to work with your IIS

- Must be a collaborative relationship and benefit both partners!
- Need to assess completeness of data for population of interest
  - Mandatory reporting?
  - IIS assessment of completeness of population capture
  - How is data uploaded? What percent is HL7 automatic uploads

Need to deal with legal and data privacy issues

  - Is there statutory protection for what you want to do?
  - Does IIS have IRB? Can they do a Data Use Agreement (DUA) with your institution?

Interactive Session
Considerations for your research question:

1. What key features of a population-based database are needed to answer this question?
   - Think of the unit of analysis, frequency of data collection, granularity and/or longitudinality.
2. Are the data public availability (yes/no)?
3. What barriers might exist to access these data?
4. Who would be key partners to assist with access?
5. What regulatory or legal issues should an investigator weigh to obtain and then analyze these types of data?