MULTI-SITE DIABETES REGISTRY USING ELECTRONIC HEALTH RECORDS: IDENTIFICATION, VALIDATION, AND REPRESENTATIVENESS

Jay Desai
HealthPartners Research Foundation

EDM Forum Symposium
Orlando, FL
June 23, 2012

TEAM

- Jay Desai (HealthPartners Research Foundation)
- Pingsheng Wu (Vanderbilt University)
- Greg Nichols (Kaiser Permanente Northwest)
- Tracy Lieu (Kaiser Permanente Northern California)
- Patrick O'Connor (HealthPartners Research Foundation)
- SUPREME-DM DataLink

- Support from Academy Health, AHRQ 1RO1HS19859, and AHRQ 1RO1HS19669

CONSIDERATIONS IN DEVELOPING A REGISTRY

- The Gold Standard
- Sensitivity, Specificity, Predictive Positive Value
- Case Definition
- Confidence and Understanding
  - Understand contribution of EHD sources
  - Variation in EHD data sources
- Population Representativeness

VALIDITY CHARACTERISTICS

- Ideally want maximize
  - Sensitivity: Likelihood that everyone with condition of interest will be identified
  - Specificity: Likelihood that all those without condition will be identified
  - Positive predictive value: Of those identified, percent who are actual cases (minimize false positives)

THE GOLD STANDARD PROBLEM

Comparative validity

- Medical record documentation
- Self-report
- Claims-based diagnosis codes

THE GOLD STANDARD PROBLEM

- Biological gold standard for diabetes identification:
  - Elevated blood glucose levels
- With good care management glucose levels may be below the threshold for diabetes diagnosis
- Remission due to substantial weight loss, bariatric surgery
- Dependent upon routinely performed lab tests, diagnosis standards, diagnostic coding practices, errors of comission or omission, reimbursement practices, economy, etc.

TAILORING DIABETES CASE DEFINITION TO SPECIFIC RESEARCH QUESTIONS

High Sensitivity

- *Maximize inclusion of all potential cases*
- Population-level monitoring
- Observational studies
- Public Health Surveillance
- Population-based quality metrics
- CER
  - Attenuate results but may have broader generalizability

High Predictive Positive Value

- *Maximize certainty of true cases*
- Practical randomized interventions
- Support of clinical management
- Patient outreach
- Accountability tied to providers or systems
- CER
  - Stringent case identification
  - Potential selection bias

BUILDING CONFIDENCE IN CASE IDENTIFICATION

- Vary time frames using same case identification criteria
  - Shorter time frames: more confident but capture fewer cases
  - Longer time frames: less confident but may capture more cases
  - What is ideal, especially with no gold standard?
- Periodic recapture or rolling case identification

BUILDING CONFIDENCE IN CASE IDENTIFICATION

- Prioritizing case identification criteria
  - Assign probabilities of ‘true case’
- Prioritize data sources: Lab most important
- More independent data sources identification… more confidence.

DIABETES DATALINK

- HMO Research Network
  - Approximately 15.8 million unique member population
- SUrveillance, PREvention, and ManagEment of Diabetes Mellitus (SUPREME-DM)
  - Eleven participating health systems
- Funded by AHRQ

BUILDING THE DATALINK REGISTRY

Initial registry construction

Broad sweep:

Any indication of diabetes from any electronic health data source

Enrollees 2005-2009 with look back to 2000

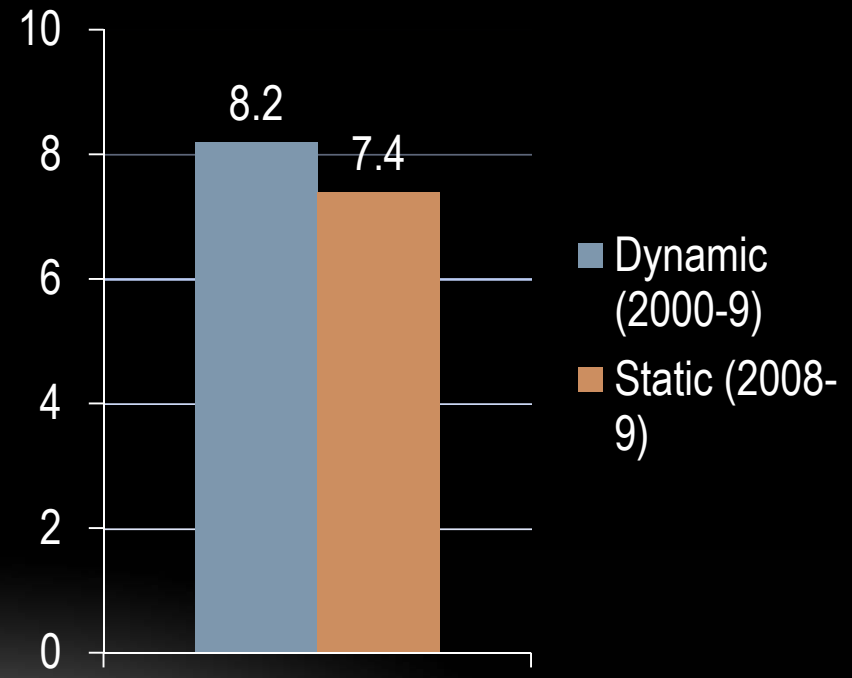
BASE DIABETES CASE DEFINITION
2-YEAR TIME FRAME

- Two elevated lab criteria on separate days, any combination
  - Fasting plasma glucose ≥ 126 mg/dL
  - Random plasma glucose ≥ 200 mg/dL
  - HbA1c ≥ 6.5%
  - OGTT ≥ 200 mg/dL (only 1 needed)
- ICD-9 250.xx, 357.2, 366.41, 362.01-07 (EMR or claim)
  - Two outpatient visit diagnosis on separate days
  - One inpatient visit diagnosis
- At least 1 diabetes drug pharmacy dispense/claim
  - Insulin
  - Oral agents (metformin & TZD must have other criteria met)
- Exclude criteria occurring within 270 days of a birth

DYNAMIC AND STATIC COHORTS

- Dynamic Cohort
  - Cumulative case identification over multiple years
  - Enter as new case or care system member
  - Leave due to death or disenrollment
- Static Cohort / Cross-sectional
  - Identification over defined time period and followed
  - Enter…none.

Figure 1. Dynamic & Static Diabetes Prevalence at one DataLink Health System (2008-9)



DIFFERENTIAL USE & CHARACTERISTICS OF ELECTRONIC HEALTH DATA SOURCES

- There is wide variation across health systems regarding the primary source for case identification.
- There may be selection bias associated with specific data sources.
- This could affect case-mix and therefore results.

DIABETES DATALINK: VARIATION IN DATA SOURCE IDENTIFICATION ACROSS SITES

Table 3. Number of Incident Diabetes Cases, Years of Membership Following Diagnosis Date, and Source of First Indication of Diabetes, by Health System, 2002–2009

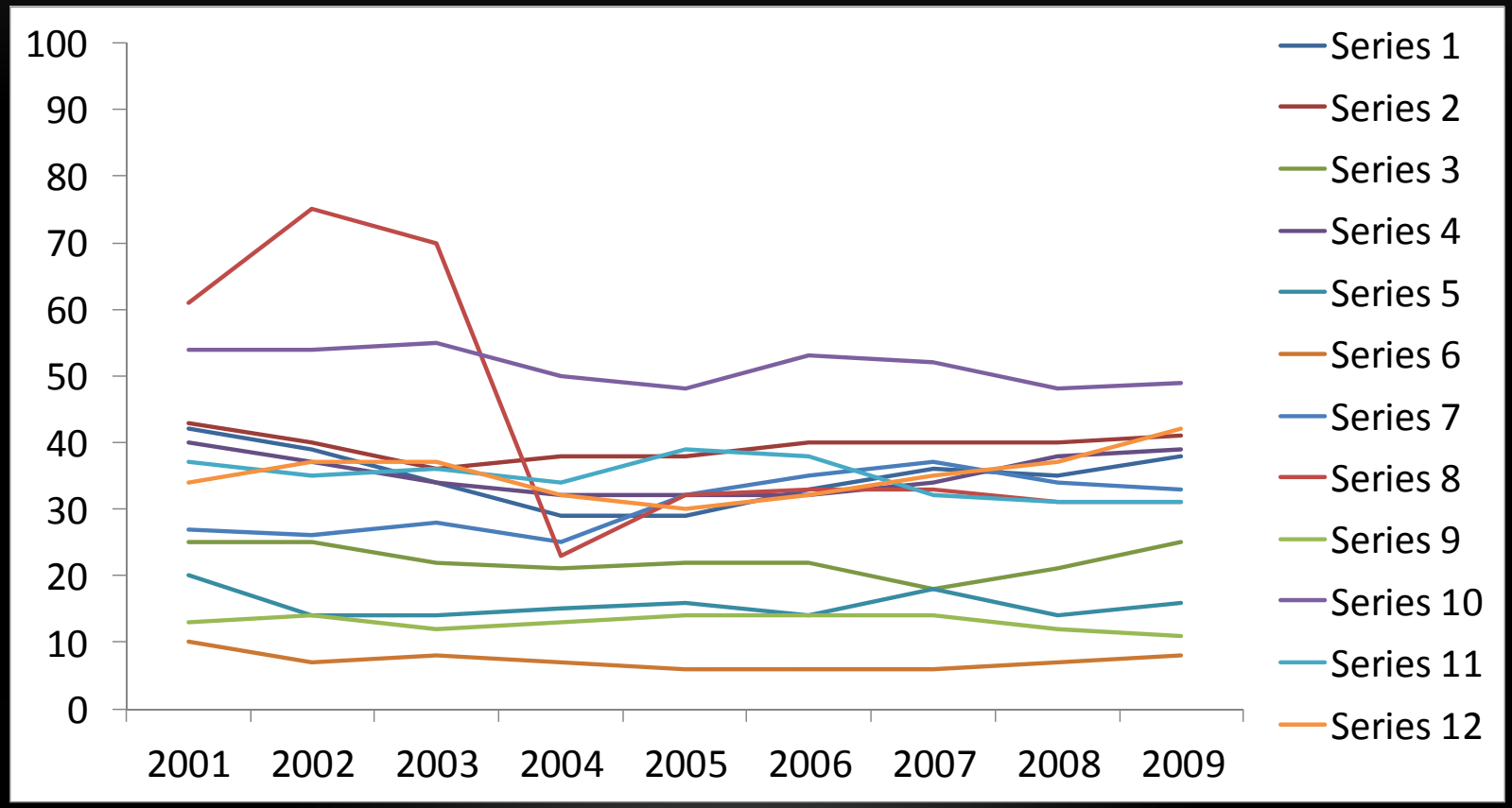
Return

Site	Incident Diabetes Cases[a]	Mean Years of Enrollment With Diabetes[b]	Source of First Indication of Diabetes[c]			
			Inpatient Diagnosis, %	Outpatient Diagnoses, %	Pharmacy Dispense, %	Outpatient Laboratory Results, %
1	18,287	3.3	11.0	17.5	16.6	54.9
2	151,177	3.6	9.6	20.0	20.1	50.2
3	7,248	3.3	14.4	36.2	12.4	37.0
4	153,473	3.4	9.8	18.9	23.2	48.1
5	10,149	3.2	15.0	26.7	26.8	31.5
6	12,062	3.5	6.1	41.4	17.0	35.5
7	8,253	2.7	16.6	40.5	11.9	31.0
8	27,096	3.1	4.3	51.2	26.7	17.8
9	11,832	3.5	8.5	13.7	10.5	67.3
10	15,059	3.3	14.2	19.0	16.3	50.5
11	13,713	3.4	12.7	16.9	16.0	54.5
Total	428,349	3.3	9.9	22.6	20.7	46.8

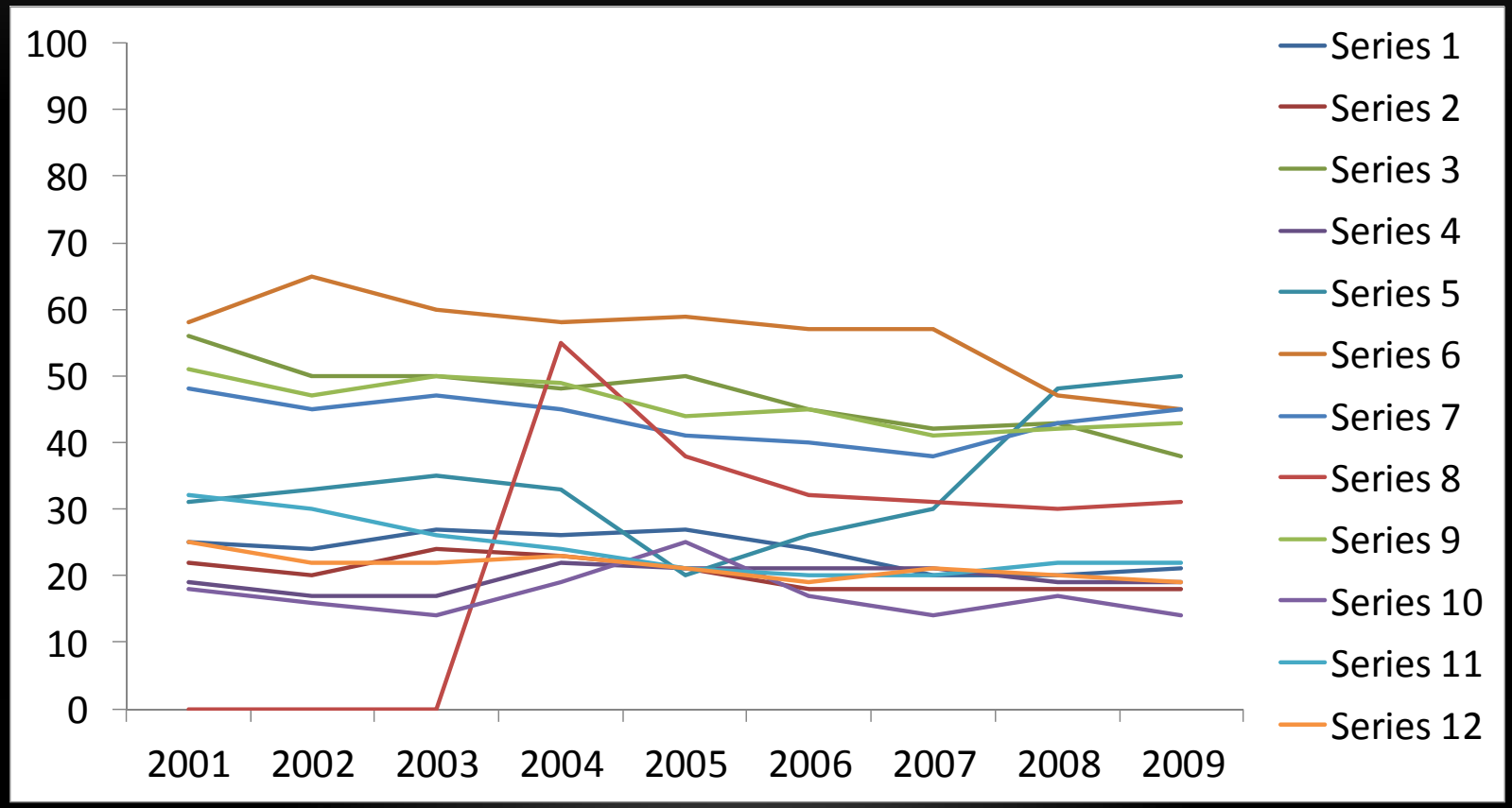
[a] Number of diabetes cases with at least 2 years of health system eligibility before first indication of diabetes.
[b] Calculated as number of years following first indication of diabetes until end of health system enrollment or December 31, 2009, whichever came first.
[c] Based on study criteria in Table 1, the source of the earliest indication of diabetes.

Nichols GA, Desai J, Elston Lafata J, Lawrence JM, O'Connor PJ, Pathak RD, et al. Construction of a Multisite DataLink Using Electronic Health Records for the Identification, Surveillance, Prevention, and Management of Diabetes Mellitus: The SUPREME-DM Project. Prev Chronic Dis 2012;9:110311. DOI: http://dx.doi.org/10.5888/pcd9.110311

INITIAL CASE IDENTIFICATION: AT LEAST 2 ELEVATED BLOOD GLUCOSE LEVELS (LAB)



INITIAL CASE IDENTIFICATION: AT LEAST 1 DIABETES DRUG PHARMACY CLAIMS



2008-9 DIABETES CASE IDENTIFICATION AT A DATALINK HEALTH SYSTEM

- Prioritize Data Sources
  - 65.1% of cases met elevated glucose criteria
- Meeting Criteria from Different of Data Sources
  - 18.1% from only 1 data source
  - 36.2% from 2 data sources
  - 45.6% from 3 data sources

DATA SOURCES: ADDED VALUE

- Insurance and pharmacy claims are routinely used for diabetes case identification.
  - Numerous validation studies against medical record or self-report

- What is the added case identification value of clinical data found in EMR's?

STEP-WISE CONTRIBUTIONS OF ELECTRONIC HEALTH DATA SOURCES TO DIABETES CASE IDENTIFICATION (2008-9) AT ONE DATALINK HEALTH SYSTEM

	A	B	C	D	E	F	G
	2 Outpatient Claims Dx	1 Inpatient Claims Dx Only	A + B	1 Diabetes Drug Claim Only	C + D	2 Elevated Blood Glucose Tests Only	E + F
Prevalence (N)	7.0% (12,916)	0.1% (111)	7.1% (13,027)	0.1% (140)	7.1% (13,167)	0.2% (422)	7.4% (13,589)
% Cases	95%	+1%		+1%		+3%	

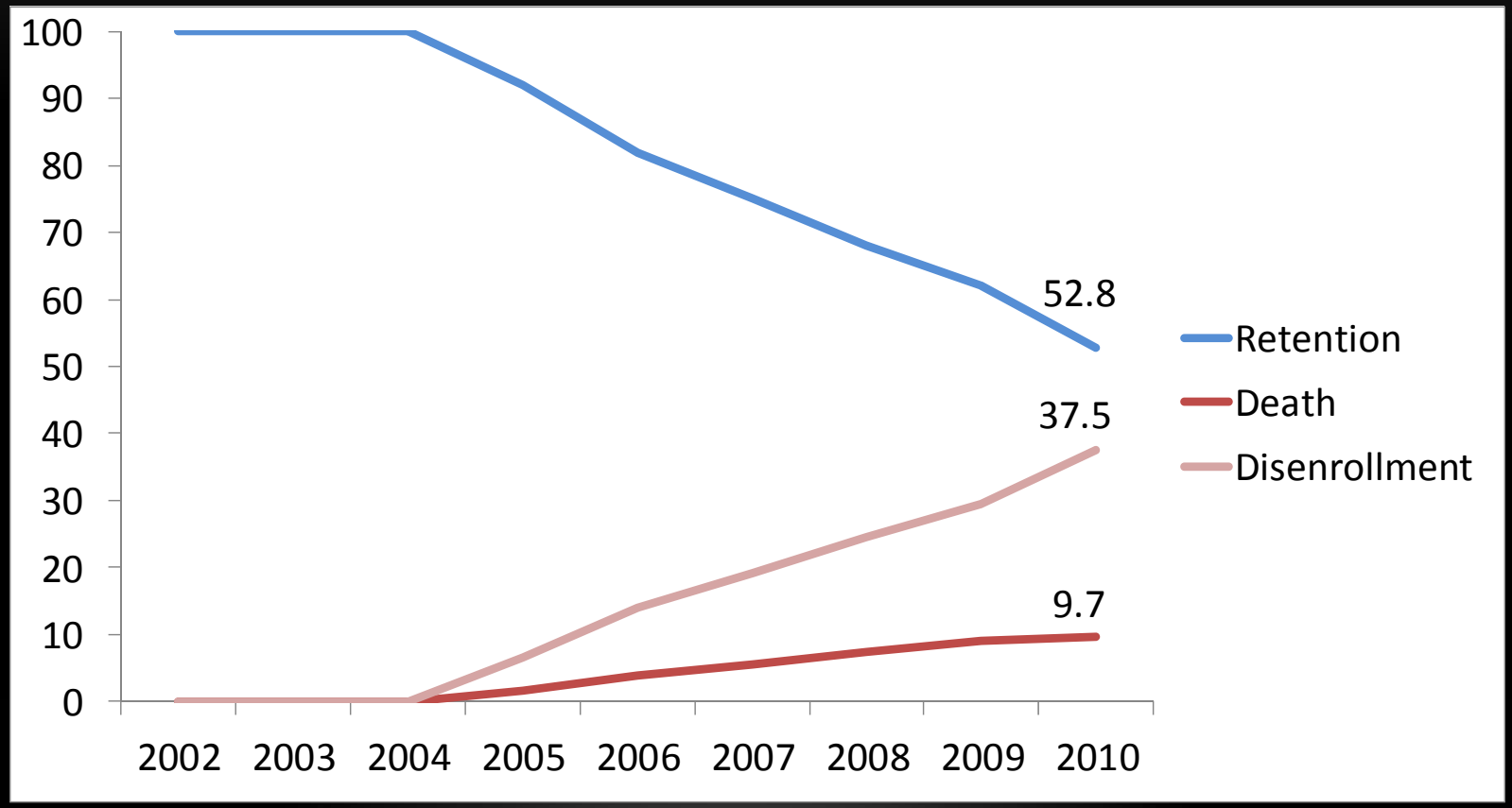
PATIENT CHARACTERISTICS BASED ON QUALIFYING CASE IDENTIFICATION EHD (2008-9) AT ONE DATALINK HEALTH SYSTEM

	A	B	C	D	E	F	G
	Outpatient Claims Dx	Inpatient Claims Dx Only	A + B	Diabetes Drug Claim Only	C + D	Blood Glucose Lab Test Only	E + F
Prevalence (N)	7.0% (12,916)	0.1% (111)	7.1% (13,027)	0.1% (140)	7.1% (13,167)	0.3% (422)	7.4% (13,589)
% Cases	91%	1%		1%		3%	
Select Patient Characteristics							
Female (%)	49	60	49	54	49	50	49
18-44 years (%)	12	26	12	31	12	8	12
HbA1c < 8 (%)	81	91	81	76	81	98	81
LDL-c < 100 (%)	74	71	74	51	73	56	73
Current smoker (%)	11	11	11	28	11	15	12

POPULATION REPRESENTATIVENESS

- CER studies?
  - Assess relative effectiveness of various treatments and systems of care in defined patient populations.
- Uninsured: highly variable across states
  - No: if population defined based on insurance claims
  - Probably: if population defined based on EMR
- Units of analysis
  - Patient, Provider, Clinic, Health System
- Large multi-site registries more likely to provide representative ‘units of analysis’
  - HIE potential to include smaller, less integrated systems

PERCENT RETENTION OF 2002 INCIDENT DIABETES COHORT AT ONE DATALINK HEALTH SYSTEM



COMPARING SELECTED CHARACTERISTICS OF 2006 INCIDENT COHORT BY RETENTION STATUS THROUGH 2010:

Baseline characteristics	Retained Cohort	Lost to disenrollment
Female	51%	50%
18-44 years	15%	27%
45-64 years	52%	55%
65+ years	32%	16%
Current smoker	15%	19%
BMI ≥ 30 kg/m^2	60%	62%
HbA1c < 8%	86%	78%
LDL-c < 100 mg/dl	51%	43%
SBP < 140 mmHg	80%	79%
DBP < 90 mmHg	94%	90%

SUMMARY

- No realistic EHD gold standard for many conditions.
- When designing a registry,
  - Think multi-purpose
  - Maximize case capture so that a variety of case definitions can be derived depending on specific study needs.
- Consider developing several case definitions with different levels of confidence [sensitivity & PPV].

SUMMARY

- For CER studies we are interested in defined patient populations, providers, clinics, health systems…
- The greater the diversity of health systems participating in a disease registry the better…the more representative.
- EMR-derived registries may include uninsured and be most representative.

SUMMARY

- Cohorts developed using insurance claims have substantial attrition due to disenrollment over time.
- Important to include demographic and clinical characteristics of retained population compared to those loss-to-follow-up
- CER studies requiring long follow-up to outcomes may be challenging if based on secondary use of EHD data
  - Improve as health systems get regionally connected so patients can be tracked across systems (HIE's, ACO's)?