



# Linking Multiple Data Sources to Enrich RWD based Study or to Extend Clinical Trial

21st Annual ASA CT Chapter Mini Conference

Tianyu Sun, April 14, 2023

# I Acknowledgment

This presentation is the author's opinion based on published literatures and personal experiences. It does **NOT** represent the official view or standing point of Moderna.

This presentation will be focusing on conducting studies for research and publication purpose, it does **NOT** apply to scenarios engaging regulatory for NDA or post-marketing commitment.

The case studies included in this presentation were led by the Smith Center for Outcomes Research at Beth Israel Deaconess Medical Center, department of cardiology. They were supported by NIH grants: **R01HL136708** (EXTEND Study, Yeh); **R01HL157530** (AHA COVID registry, Yeh, Gerszten, Kazi).

# I Agenda

- **Introduction: why linkage and how?**
- **Case 1: conduct linkage to enrich RWD sources**
- **Case 2: conduct linkage for clinical trials**
- **Summary**

# Introduction: why linkage?

- Different type of studies and data sources have their pros and cons.
- The evidence-based medicine ranks the randomized controlled studies higher than observational studies (such as cohort or case-control).
- Observational studies using large databases have broad sample and comprehensive longitudinal data.

	Strength	Limitations
Clinical trial	Randomization theoretically eliminates measured/unmeasured confounding; Specific measurements with high validity and reliability	Burden on patients, healthcare professionals, and sponsors Small sample size; Shorter period of follow-up; Restricted population;
Administrative claims	Captures insurance reimbursed healthcare utilizations; A broad population; Comprehensive history	Confounding; Measurement errors; Lack of granularity of data
Registries	Specific measurements with better validity and reliability	Confounding; May have limited population; May lack long period longitudinal data
Electronic Health Records (EHRs)	Chance to retrieve more details of medical/clinical history than claims	Data quality and missingness; Challenges of free-text and varied data structure Resources needed for review and extraction of information

Gliklich RE, Dreyer NA, Leavy MB, editors. Registries for Evaluating Patient Outcomes: A User's Guide [Internet]. 3rd edition. Rockville (MD): Agency for Healthcare Research and Quality (US); 2014 Apr. Table 6–1, Key data sources—strengths and limitations. Available from: <https://www.ncbi.nlm.nih.gov/books/NBK208611/table/ch6.t1/>  
Rosner, Anthony L. "Evidence-based medicine: revisiting the pyramid of priorities." *Journal of Bodywork and Movement Therapies* 16.1 (2012): 42-49.

# I Administrative claims and registry

Administrative claims:

Good generalizability;  
Comprehensive longitudinal data

Measurement error and lack of  
granularity

Confounding, lost-to-follow-up

Link with registry to gain  
more variables and  
specific measurements;

Causal inference  
methods;

Proper study design

Robust estimation of  
effectiveness minimizing  
confounding, selection  
bias, measurement error.

The generalizability could  
be assessed comparing  
the linked and unlinked.

Better understanding of  
specific risk factors/  
biomarkers.

# Link randomized clinical trials with RWDs

Randomized trials:

Randomization;  
Valid endpoint;  
Specific measurements

Relatively short study period

Restricted patient population

Link with RWD: to obtain  
comprehensive and  
longitudinal data;

Causal inference  
methods;

Proper study design

Explore additional  
endpoints

Longer period of  
follow-up with  
minimum burden for  
patients/clinicians

# How to conduct the linkage

- **Deterministic linkage:**

***“whether record pairs agree or disagree on a given set of identifiers, where agreement on a given identifier is assessed as a discrete—“all-or-nothing”—outcome.”***

Social security number, beneficiary ID, etc. (NCHS with Medicare FFS)

In situation when lacking direct identifiers, some highly reliable index events could be used as the key/anchor: an inpatient event at a given hospital for a certain disease, or a surgical procedure at a given hospital/physician office. With other demographic identifiers: DOB, sex, etc, a deterministic linkage could be conducted.

- **Probabilistic link:**

***“The likelihood that two records are a true match based on whether they agree or disagree on the various identifiers.”***

In situation when lacking direct identifiers, limited information about index events, results in a one:multiple, multiple:one linkage.

- **Tokenization**

Mask the private information via encryption non-reversible tokenization (preserve privacy and could be applied to de-identified data satisfying HIPAA).

Reliable linkage: high precision (TP + TN out of all), accuracy (TP out of linked), and specificity.

With higher flexibility: Handle typos, missingness of data elements

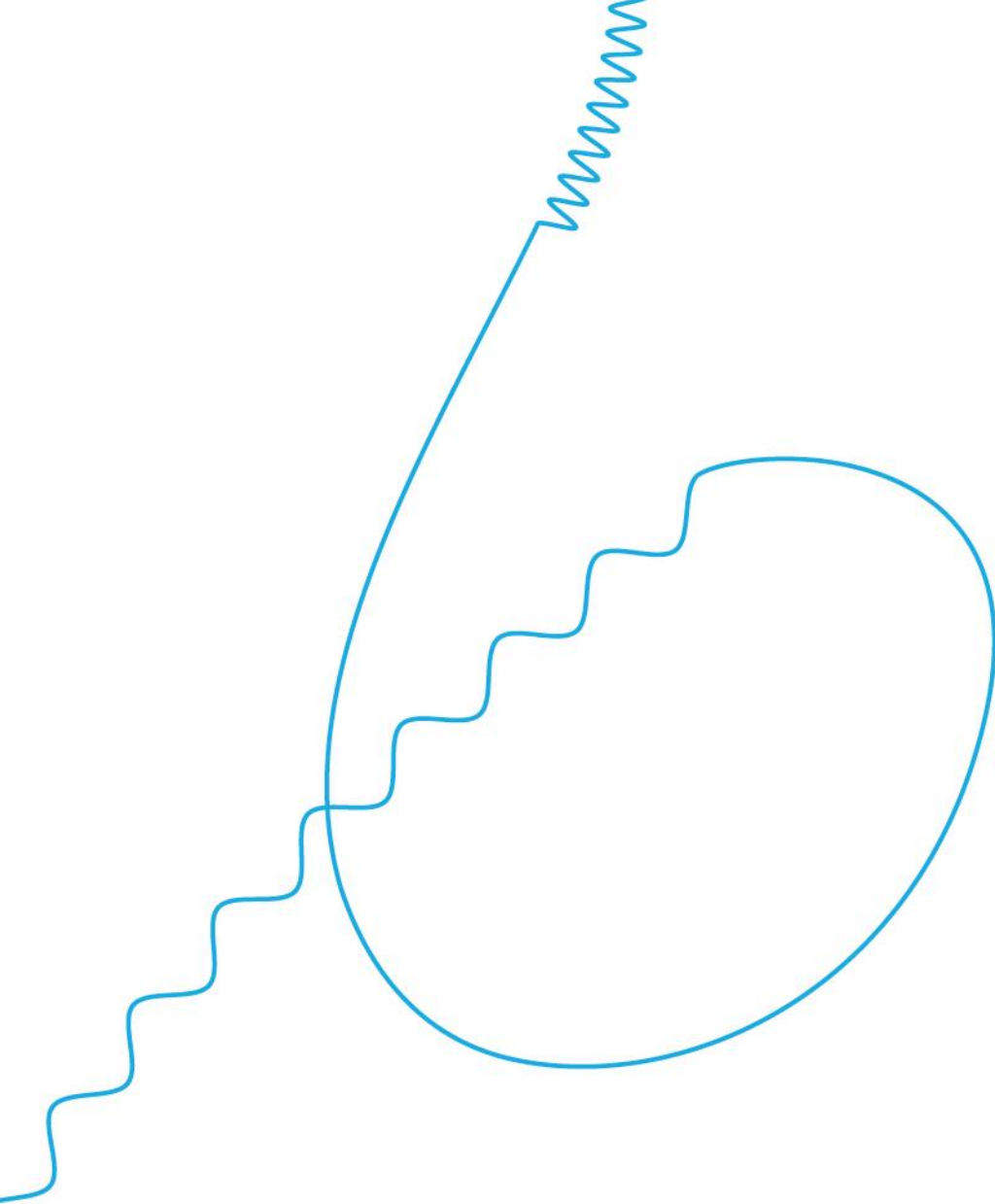
Dusetzina SB, Tyree S, Meyer AM, et al. Linking Data for Health Services Research: A Framework and Instructional Guide [Internet]. Rockville (MD): Agency for Healthcare Research and Quality (US); 2014 Sep. 4, An Overview of Record Linkage Methods. Available from: <https://www.ncbi.nlm.nih.gov/books/NBK253312/>

The Linkage of National Center for Health Statistics Survey Data to Medicare Enrollment, Claims/Encounters and Assessment Data (2014-2018): [https://www.cdc.gov/nchs/data-linkage/cms/nchs\\_medicare14\\_18\\_linkage\\_methodology\\_and\\_analytic\\_considerations.pdf](https://www.cdc.gov/nchs/data-linkage/cms/nchs_medicare14_18_linkage_methodology_and_analytic_considerations.pdf)  
<https://www.cdc.gov/nchs/tutorials/nhanes-cms/orientation/data-linkage.htm>

Hammill BG, Hernandez AF, Peterson ED, Fonarow GC, Schulman KA, Curtis LH. Linking inpatient clinical registry data to Medicare claims data using indirect identifiers. *Am Heart J.* 2009;157:995–1000. doi:10.1016/j.ahj.2009.04.002

Newgard, Craig D. "Validation of probabilistic linkage to match de-identified ambulance records to a state trauma registry." *Academic Emergency Medicine* 13.1 (2006): 69-75.

EVALUATING THE PERFORMANCE OF PRIVACY PRESERVING RECORD LINKAGE SYSTEMS (PPRLS)–PART ONE : <https://surveillance.cancer.gov/reports/TO-P2-PPRLS-Evaluation-Report-Part1.pdf>



---

# Case Study 1: Enrich the data elements of RWD

***Link the AHA COVID Registry to  
Medicare FFS and other data sources***

A brief introduction of the deterministic  
linkage



# Enrich the data elements of RWD











From national COVID registry to administrative claims data

- **The American Heart Association (AHA) COVID-19 registry was established to collect a wide range of information of COVID-19 hospitalizations.**
  - Lab results for a collection of biomarkers (creatinine, troponin, etc.)
  - Vital sign
  - ICU/ECMO utilization details
- **Limited information:**
  - Comprehensive medical history
  - Health outcomes post-discharge
  - Generalizability of the participants

Journal of the American Heart Association

## ORIGINAL RESEARCH

### Enriching the American Heart Association COVID-19 Cardiovascular Disease Registry Through Linkage With External Data Sources: Rationale and Design

Andrew S. Oseran , MD, MBA\*; Tianyu Sun , PhD\*; Rishi K. Wadhera , MD, MPP, MPhil; Issa J. Dahabreh , MD, ScD; James A. de Lemos , MD; Sandeep R. Das , MD, MPH, MBA; Christine Rutan , BS; Aarti H. Asnani , MD; Robert W. Yeh , MD, MSc†; Dhruv S. Kazi , MD, MSc, MS†

<https://www.heart.org/en/professional/quality-improvement/covid-19-cvd-registry>

Oseran, A. S., Sun, T., Wadhera, R. K., Dahabreh, I. J., de Lemos, J. A., Das, S. R., ... & Kazi, D. S. (2022). Enriching the American Heart Association COVID-19 Cardiovascular Disease Registry Through Linkage With External Data Sources: Rationale and Design. *Journal of the American Heart Association*, 11(18), e7743.

# Deterministic linkage across data sources (individual and aggregate level)

Social vulnerability Index	Rural-Urban Commuting Area Codes	Medicare FFS	American Hospital Association Survey data	American Heart Association COVID registry
		Demographic info	Demographic info	Demographic info
Geographic info	Geographic info	Geographic info	Geographic info	Geographic info
		NPI	NPI	Hospital ID
			Hospital ID	Hospital ID
			Hospital level information	Lab data
		Longitudinal healthcare utilization		COVID-19 hospitalization
		Death info		
	Population density, commuting, etc.			
Social vulnerability index: potential negative effects on communities caused by external stresses on human health				

Price of linkage is a potential compromised generalizability when selecting linked only. [moderna](https://www.moderna.com)

# The enriched dataset and assess representativeness of AHA elderly patients

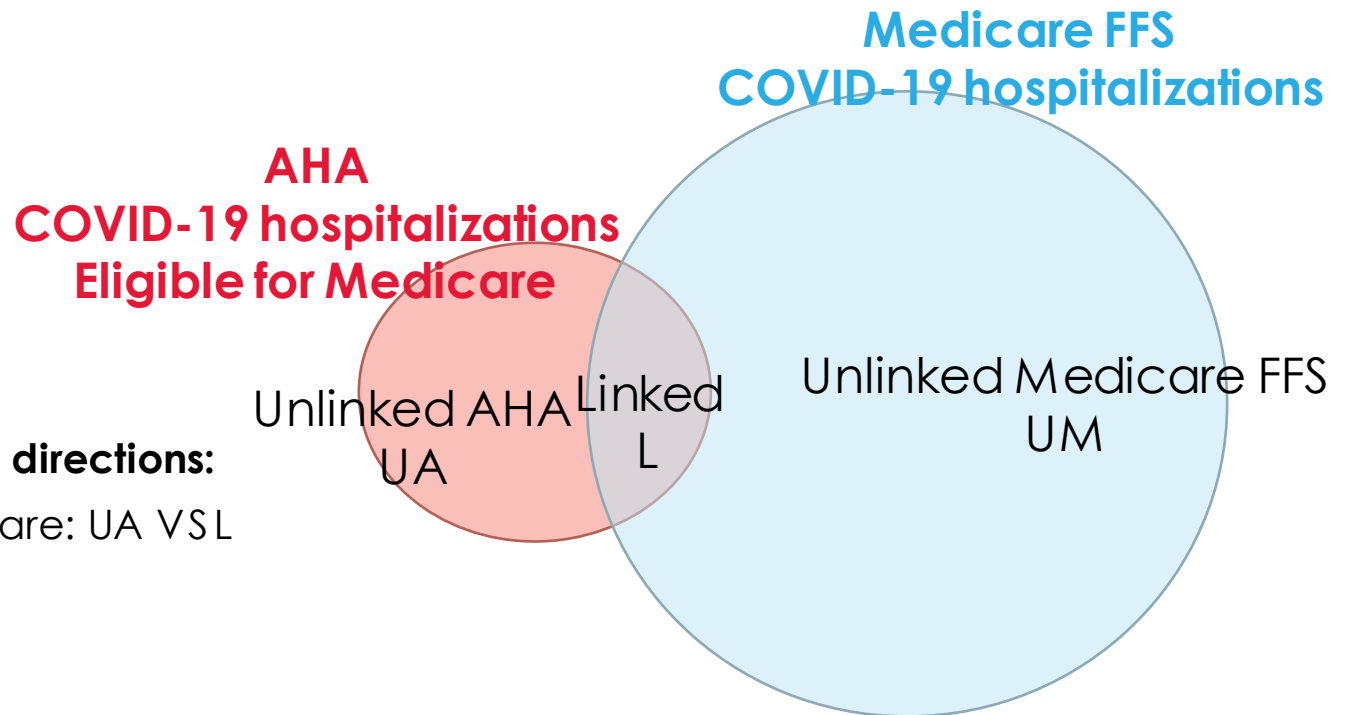
- **The enriched data elements:**

Social demographic, index event information, hospital level information, long-term (pre- and post- index event) medical information.

- **The representativeness is assessed in 2 directions:**

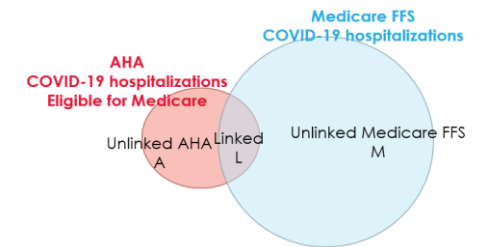
AHA COVID patients eligible for Medicare: UA VSL

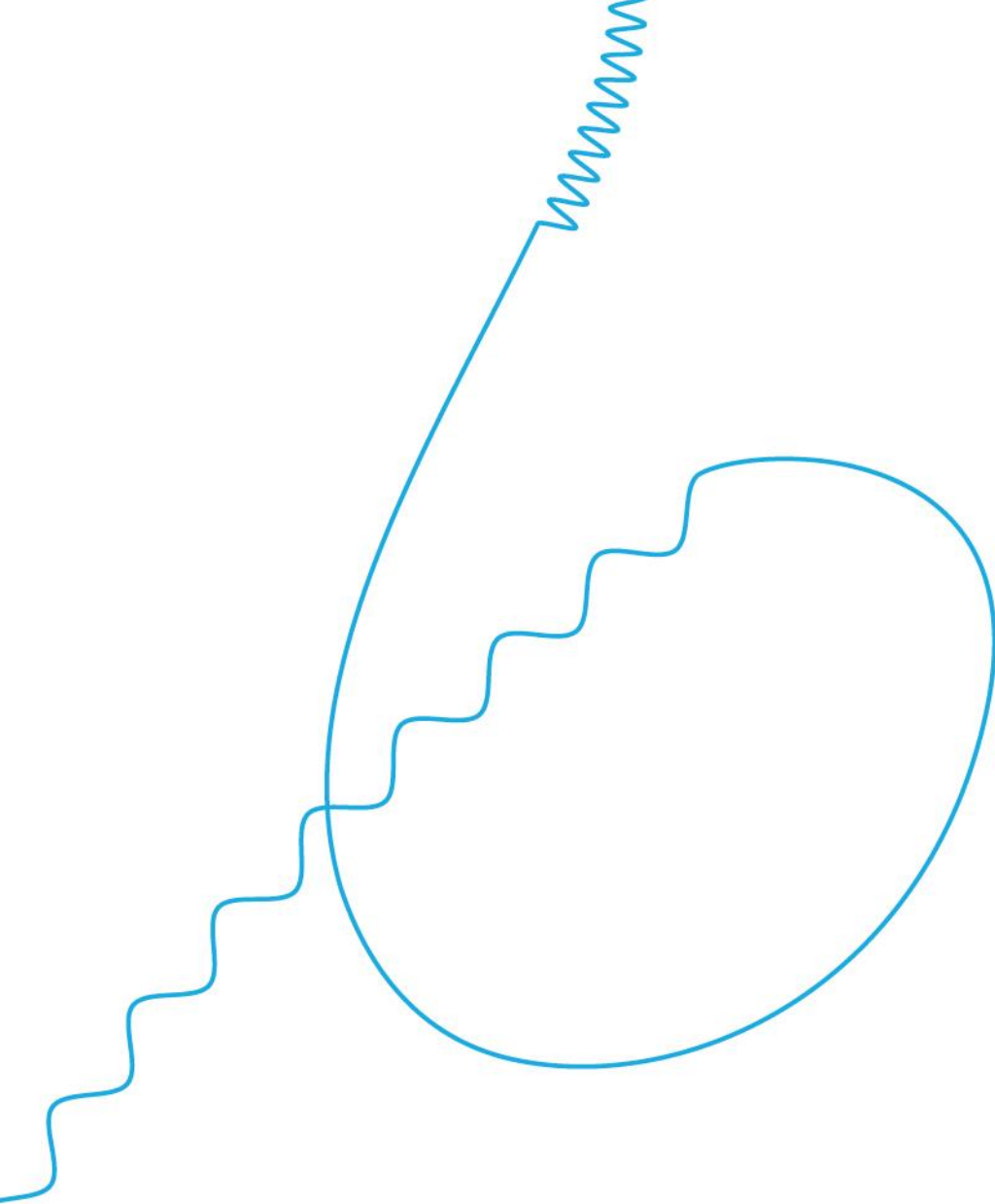
Medicare FFS COVID patients: UM VSL



# Assess the representativeness of the AHA COVID registry elderly patients to a broader Medicare FFS COVID population

- **The patient characteristics around the time of index hospitalization was compared between these 2 groups to assess the representativeness and generalizability of AHA elderly patients:**
  - Demographic information: age, race/ethnicity, sex;
  - Geographic information: 4 US regions
  - Chronic comorbidities (more than 25): Diabetes, Hypertension, dementia, cancer, etc.
  - Hospital level information: hospital size and resources
  - Community level information: SVI and URCA
  - Length of the index hospitalization stay.
  - Discharge status
- **Medicare FFS patients linked with AHA Registry were broadly representative of the general population of FFS patients with a COVID-19 hospitalization**
  - Measured sociodemographic characteristics and comorbidity burden.
  - Similar in-hospital outcomes
- **Deviation among the community level information and hospital-level characteristics was observed**
  - Linkable registry patients tended to be from larger, major teaching hospitals in metropolitan area.





---

## Case Study 2: Link existing clinical trial to Medicare FFS

Judicate endpoints collected by  
clinical trial and validate  
outcomes from RWD

## Introduction: 2 RCTs linked with Medicare in the EXTEND study

RCT	HiR	SURTAVI
Arms	TAVR vs SAVR	
Target patient population	Severe aortic stenosis and heart-failure with high risk of surgery.	Severe aortic stenosis with intermediate risk of surgery.
Primary endpoint	All-cause mortality	Composite of death and stroke
Follow-up time	1 year	2 year
Shared secondary endpoints:	All-cause mortality, stroke+TIA	
Other secondary endpoints	Bleeding, acute kidney injury Aortic valve reintervention, etc.	MACE (death, MI, stroke, aortic valve reintervention)

HiR= US CoreValve Pivotal High Risk

SURTAVI= The Surgical or Transcatheter Aortic Valve Replacement in Intermediated-Risk Patients

SAVR = surgical aortic valve replacement; TAVR = transcatheter aortic valve replacement

MACE = Major adverse cardiovascular and cerebrovascular events

Strom, J. B., Faridi, K. F., Butala, N. M., Zhao, Y., Tamez, H., Valsdottir, L. R., ... & Yeh, R. W. (2020). Use of administrative claims to assess outcomes and treatment effect in randomized clinical trials for transcatheter aortic valve replacement: findings from the EXTEND study. *Circulation*, 142(3), 203-213.

Adams DH, Popma JJ, Reardon MJ, Yakubov SJ, Coselli JS, Deeb GM, Gleason TG, Buchbinder M, Hemiller J Jr, Kleiman NS, et al; U.S. CoreValve Clinical Investigators. Transcatheter aortic-valve replacement with a self-expanding prosthesis. *N Engl J Med*. 2014;370:1790–1798. doi: 10.1056/NEJMoa1400590

Reardon MJ, Van Mieghem NM, Popma JJ, Kleiman NS, Søndergaard L, Mumtaz M, Adams DH, Deeb GM, Maini B, Gada H, et al; SURTAVI Investigators. Surgical or transcatheter aortic-valve replacement in intermediate-risk patients. *N Engl J Med*. 2017;376:1321–1331. doi: 10.1056/NEJMoa1700456

# Linkage: from trials to administrative claims

- Deterministic linkage algorithms were used for linking both trials with Medicare claims data
- The two medical device RCTs linked to Medicare FFS directly

**Exclude** patients who were not eligible for Medicare: non-US participants, <65 years old, and VA-hospitals:

- HiR lost 15/750 (2%);
- SURTAVI lost 355/1660 (21%)

**Successfully linked** with Medicare FFS

- HiR with 82% linkage rate (600/735);
- SURTAVI with 77% linkage rate (1004/1305)

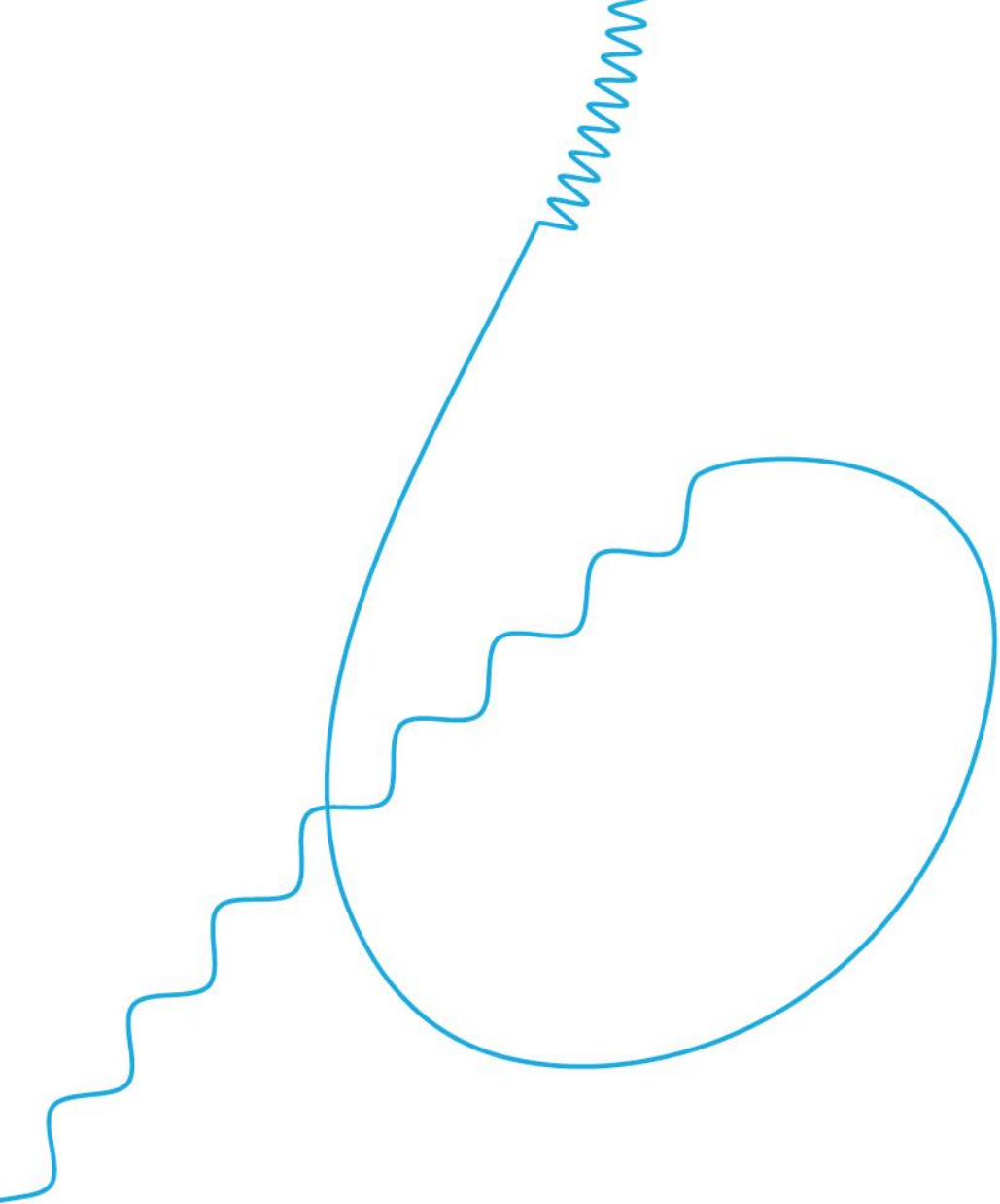
# Results

Study	HiR N = 600		SURTAVI N = 1004	
Linked Patient characteristics between arms (SAVR vs TAVR)	Demographic, cardiac risk factor, measured medical conditions are similar			
Patient characteristics (linked vs unlinked)	Linked patients were older on average than unlinked Other characteristics were similar.			
All-cause mortality	Almost Identical result: exactly same # of death, following results are HR with 95%CI with SAVR as the reference group			
	RCT 0.84 (.65, 1.09)	RWD 0.86 (.66, 1.11)	RCT 1.28 (0.86, 1.91)	RWD 1.28 (0.86, 1.91);
Stroke + TIA	0.78 (0.52, 1.16)	0.59(0.37, 0.95);	0.91 (0.61–1.34)	0.80 (0.50–1.28)
MACE	N/A		1.22 (0.90–1.64);	1.17 (0.87–1.57);
Bleeding	0.93 (0.75–1.14)	0.65 (0.55–0.78)	N/A	
Acute kidney injury	0.35 (0.21–0.59)	0.46 (0.31–0.69)	N/A	



# I Summary of the 2 linked RCTs

- **Linkage**
  - The linkage of these 2 RCTs with Medicare FFS among the eligible patients were high.
- **Descriptive results after linkage showed no evidence of breaking balanced baseline characteristic by conducting linkage**
  - Among the linked patients, the baseline demographic variables were similar across SVAR and TVAR arms.
  - The baseline variables between the linked and unlinked were comparable.
- **RWD ascertained endpoints**
  - Clinically severe endpoints were very similar (stroke and MACE): all-cause mortality was almost identical.
  - Other endpoints showed some discrepancy between RCT and RWD ascertained.



---

# Summary

## Summary: enrichment of elements and validity of outcome ascertain

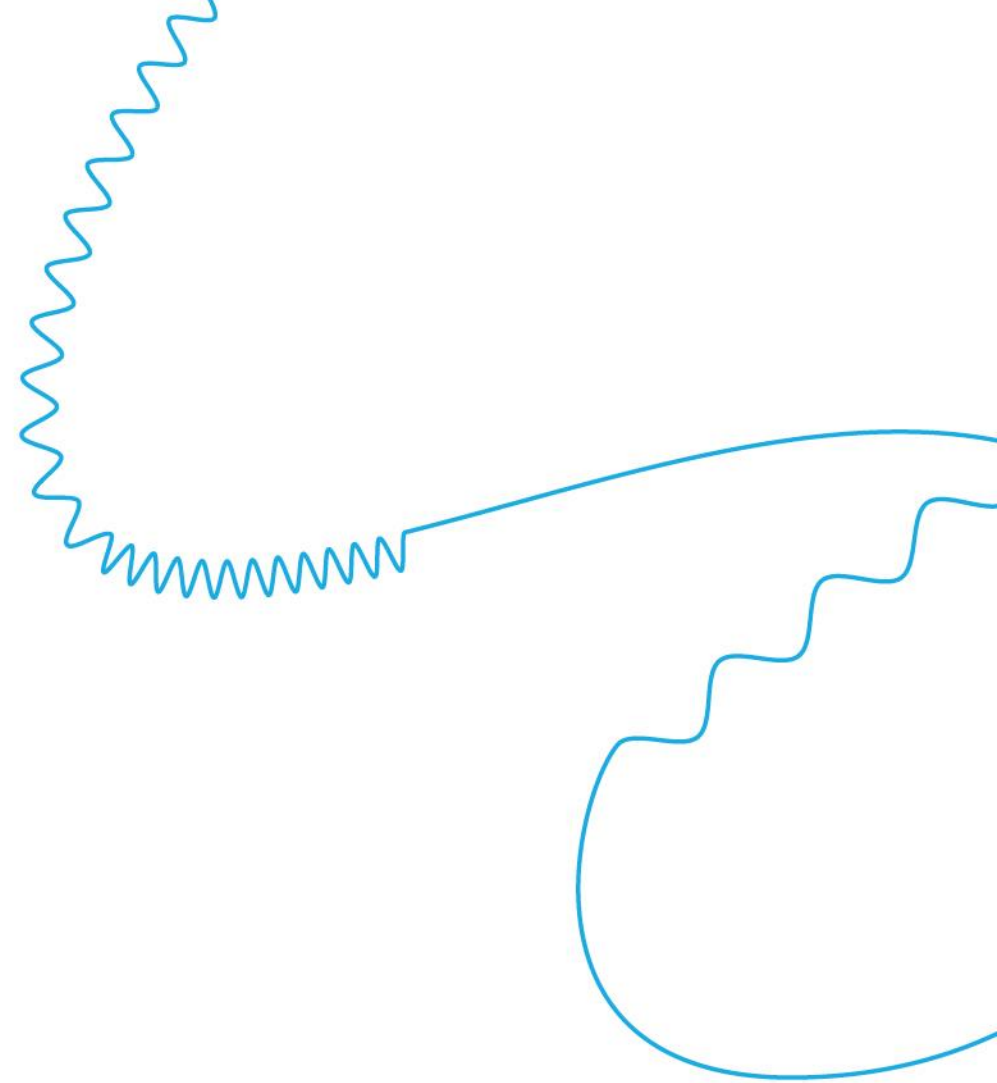
- **Linking registries, administrative claims, and other types of RWD**
  - Harvest a wide range of data variables to help better control confounding and answer questions engaging specific measurements.
  - Linking relatively limited sample in registries to a more generalized population in administrative claims could help assess the representativeness using empirical methods.
- **Linking RCTs to RWD:**
  - Explore additional endpoints: health outcomes, healthcare utilizations, and cost
  - Explore endpoints during a longer follow-up period than the trial.

# I Limitations

- **Indirect linkage and tokenization:**
  - Not guarantee precision and accuracy and in many scenarios without gold standard to cross validate.
- **Linking different RWDs: administrative claims, EHRs, registries may lead to a smaller sample**
  - A trade of data elements, completeness, and granularity versus generalizability.
- **Linking RCTs with RWD:**
  - Preserve the magic of randomization?? (Free of measured/unmeasured confounding).
  - Many endpoints identified by diagnosis code algorithm might not be as reliable as others.
- **Linking with Medicare (fee-for-service vs Medicare Advantage)**

---

**Thank you**



moderna®