REPEAT
Reproducible Evidence: Practices to Enhance and Achieve Transparency

www.repeatinitiative.org

Disclosures

- This work was funded by the Laura and John Arnold Foundation

- At the time that this work was conducted, Dr. Wang was principal investigator on other grants from:
  - National Institute of Aging
  - Laura and John Arnold Foundation
  - FDA Sentinel Initiative
  - Investigator initiated grants to Brigham and Women's Hospital from Novartis, J & J, Boehringer Ingelheim for unrelated work
- She is a consultant to Aetion Inc, for unrelated work

2

Reproducibility

What is reproducibility in database studies?

	Data Source	Methods
Analytic reproduction *Re-running the same code on same data*	Same	Same
Direct replication *Independent implementation of a specific study*	Same	Same
Conceptual replication (robustness) *Implementing a study of the same exposure (and comparator), outcome and estimand of interest*	**Different**	Same
	Same	**Different**
	Different	**Different**

(row label at left, rotated: Reproducibility)

3

Reproducibility

What is reproducibility in database studies?

	Data Source	Methods
Analytic reproduction *Re-running the same code on same data*	Same	Same

Hazard ratio = 2.0
Hazard ratio = 2.0
Hazard ratio = 2.0

Important but not transparent by itself

Thousands of lines of code to create a temporally anchored analytic cohort from raw longitudinal data streams
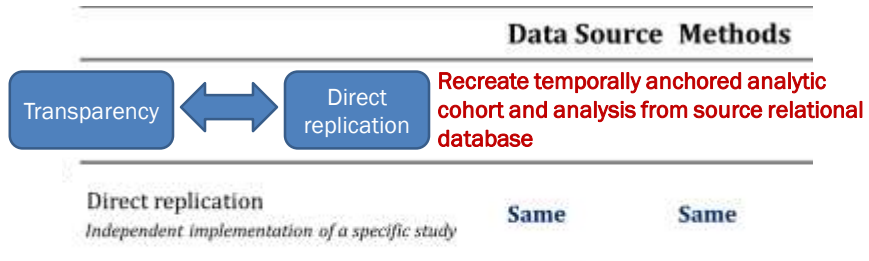
What scientific decisions is the code implementing? Agree with the validity and/or relevance for the question of interest?

4

2

Reproducibility

What is reproducibility in database studies?

	Data Source	Methods
Transparency ↔ **Direct replication** — Recreate temporally anchored analytic cohort and analysis from source relational database		
Direct replication *Independent implementation of a specific study*	Same	Same

Transparency in design and implementation decisions are a necessary pre-condition for direct replication

∴ Ability to directly replicate a study is a proxy for transparency of study methodology

Need transparency to assess validity and relevance of evidence

Reproducibility

What is reproducibility in database studies?

	Data Source	Methods
Most common, most interesting?	Need transparency to understand	
	• Subtle design/implementation differences	
Why do results differ or converge?	• Differences in data	
	• Differences in population	
Conceptual replication (robustness) *Implementing a study of the same exposure (and comparator), outcome and estimand of interest*	Different	Same
	Same	Different
	Different	Different

Important point to keep in mind

Transparency facilitates **assessment** of validity, relevance, replicability

Transparent and replicable	≠	Scientifically valid and robust

Study quality

		High	Low
Study Transparency	**High**	✔	🚫
	Low	?	?

Aim 1. To quantify the current state of healthcare database study reproducibility via direct replication
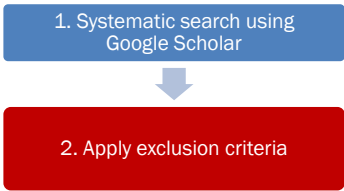
1. Systematic search using Google Scholar

➡ Top h-5 clinical, epidemiology journals
- Published after Jan 1, 2011
- "cohort" + "claims" + database name

Aim 1. To quantify the current state of healthcare database study reproducibility via direct replication

1. Systematic search using Google Scholar

2. Apply exclusion criteria

Top h-5 clinical, epidemiology journals
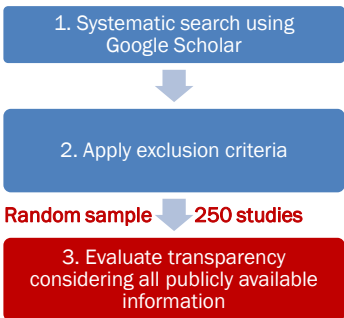- Published after Jan 1, 2011
- "cohort" + "claims" + database name

CONSORT style diagram
- Include descriptive, comparative safety/effectiveness cohort studies
- Exclude if data source mismatch, PDF unavailable, methods study, etc.

Aim 1. To quantify the current state of healthcare database study reproducibility via direct replication

1. Systematic search using Google Scholar

2. Apply exclusion criteria

Random sample → 250 studies

3. Evaluate transparency considering all publicly available information

Top h-5 clinical, epidemiology journals
- Published after Jan 1, 2011
- "cohort" + "claims" + database name

CONSORT style diagram
- Include descriptive, comparative safety/effectiveness cohort studies
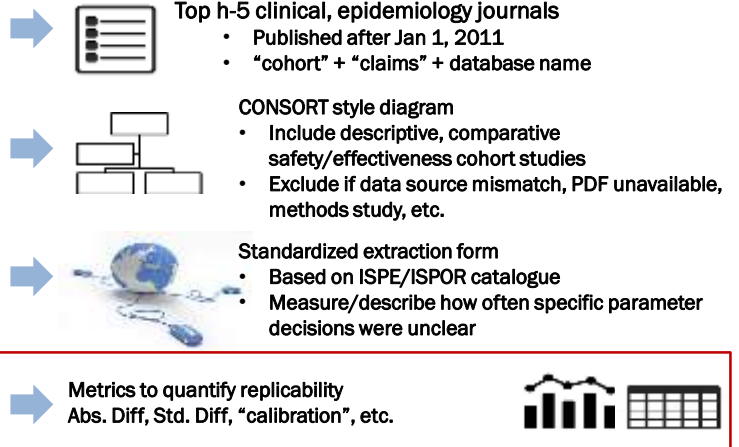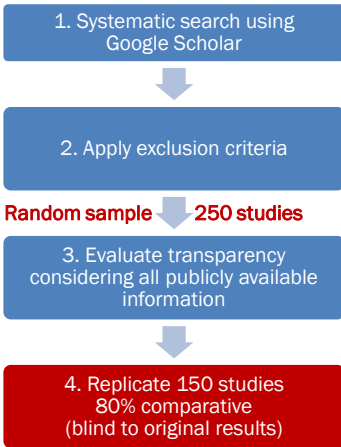- Exclude if data source mismatch, PDF unavailable, methods study, etc.

Standardized extraction form
- Based on ISPE/ISPOR catalogue
- Measure/describe how often specific parameter decisions were unclear

Aim 1. To quantify the current state of healthcare database study reproducibility via direct replication

1. Systematic search using Google Scholar

2. Apply exclusion criteria

Random sample → 250 studies

3. Evaluate transparency considering all publicly available information

4. Replicate 150 studies 80% comparative (blind to original results)

Top h-5 clinical, epidemiology journals
- Published after Jan 1, 2011
- "cohort" + "claims" + database name

CONSORT style diagram
- Include descriptive, comparative safety/effectiveness cohort studies
- Exclude if data source mismatch, PDF unavailable, methods study, etc.
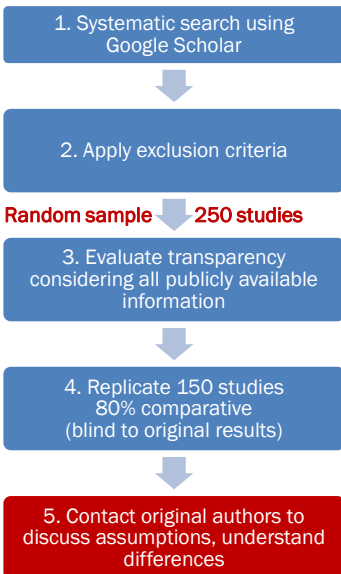
Standardized extraction form
- Based on ISPE/ISPOR catalogue
- Measure/describe how often specific parameter decisions were unclear

Metrics to quantify replicability
Abs. Diff, Std. Diff, "calibration", etc.
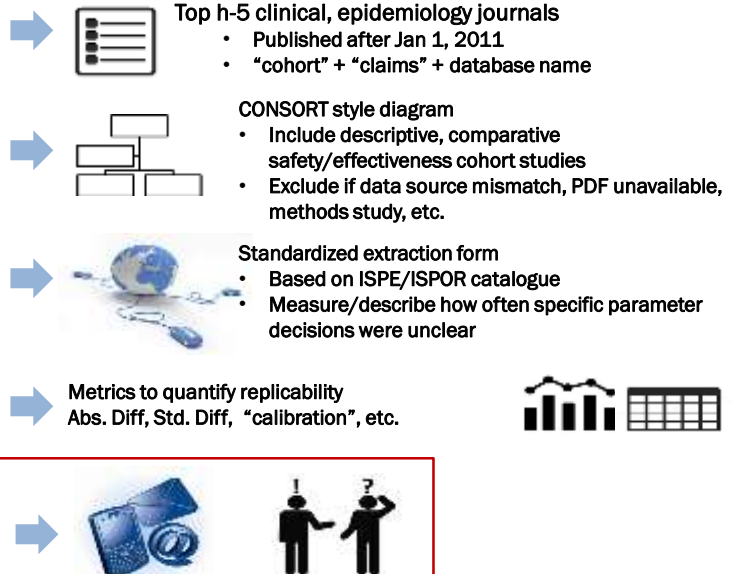
Aim 2. To evaluate the robustness of evidence currently found in healthcare database studies

Involve original investigators

1. Identify sample of 50 comparative studies

- Closely replicated
- Noted design/analysis issue
- Implementation parameters ≠ intended question?

13

Aim 2. To evaluate the robustness of evidence currently found in healthcare database studies

Involve original investigators

1. Identify sample of 50 comparative studies

- Closely replicated
- Noted design/analysis issue
- Implementation parameters ≠ intended question?
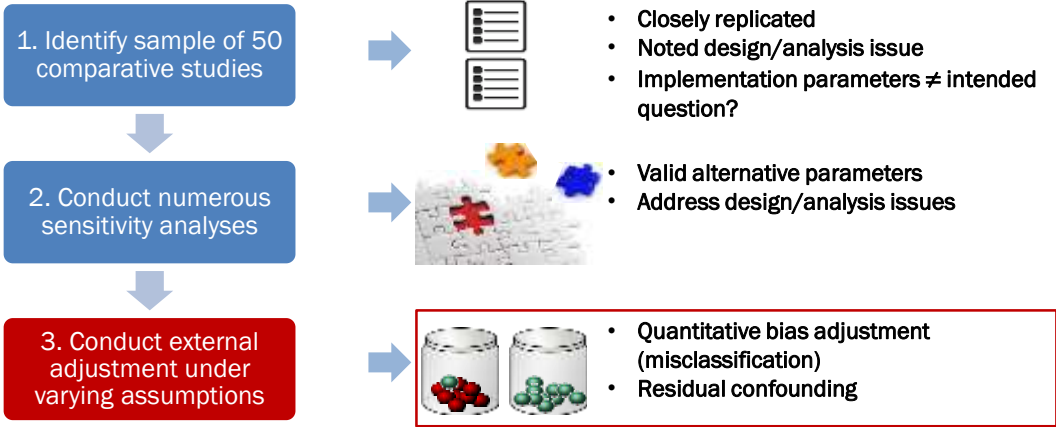
2. Conduct numerous sensitivity analyses

- Valid alternative parameters
- Address design/analysis issues

14

Aim 2. To evaluate the robustness of evidence currently found in healthcare database studies
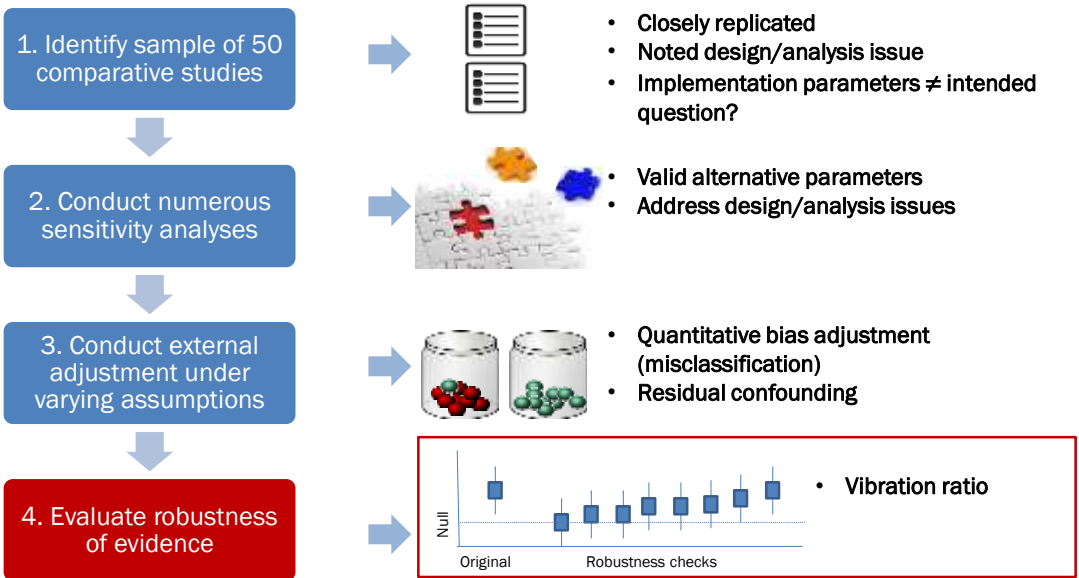
Involve original investigators

1. Identify sample of 50 comparative studies

- Closely replicated
- Noted design/analysis issue
- Implementation parameters ≠ intended question?

2. Conduct numerous sensitivity analyses

- Valid alternative parameters
- Address design/analysis issues

3. Conduct external adjustment under varying assumptions

- Quantitative bias adjustment (misclassification)
- Residual confounding

Aim 2. To evaluate the robustness of evidence currently found in healthcare database studies

Involve original investigators

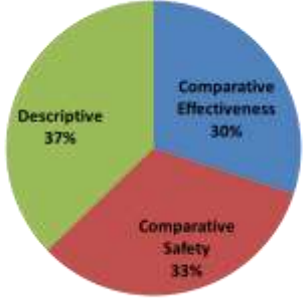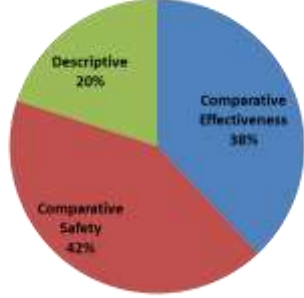**1. Identify sample of 50 comparative studies**

- Closely replicated
- Noted design/analysis issue
- Implementation parameters ≠ intended question?

2. Conduct numerous sensitivity analyses

- Valid alternative parameters
- Address design/analysis issues

3. Conduct external adjustment under varying assumptions

- Quantitative bias adjustment (misclassification)
- Residual confounding

4. Evaluate robustness of evidence

- Vibration ratio

Null

Original   Robustness checks

Random Sample of Peer-Reviewed, Published Database Studies

Study types for transparency (N = 250)

- Descriptive 37%
- Comparative Effectiveness 30%
- Comparative Safety 33%

Study types for replication (N = 150)

- Descriptive 20%
- Comparative Effectiveness 38%
- Comparative Safety 42%

Study types for robustness (N = 50)

- Comparative Safety 48%
- Comparative Effectiveness 52%
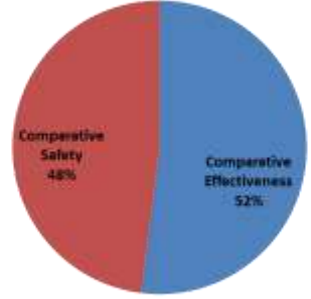
Current progress

Transparency Evaluation	Replication	Author Contacts	Robustness
of 250	of 150	of 150	of 50

ISPOR/ISPE Joint Task Force catalogue of specific parameters

Same data source, Same methods
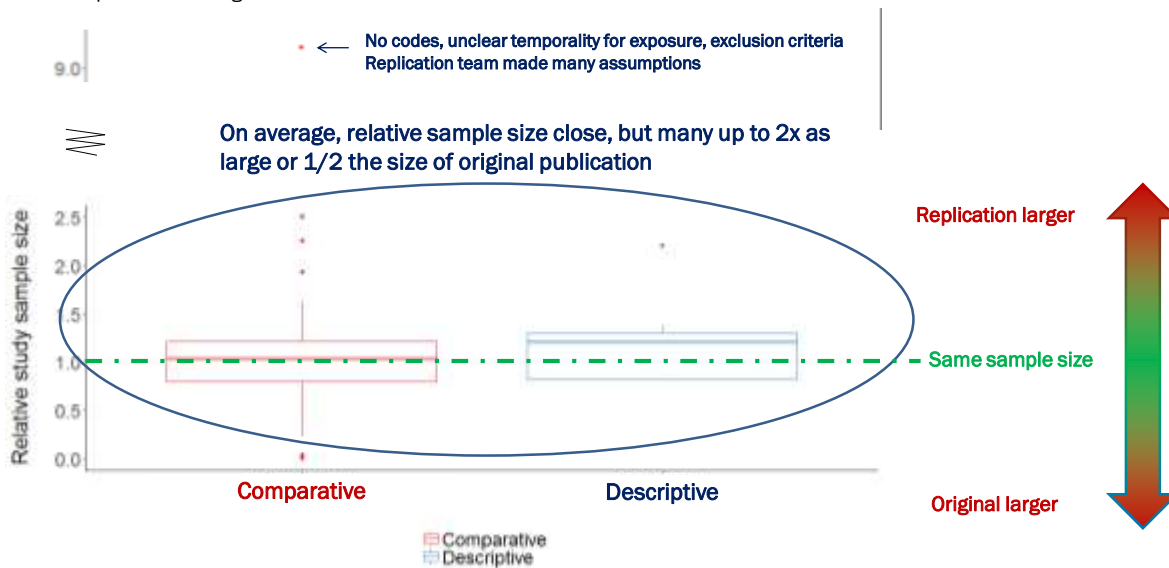
(15 attempted contacts)

Valid alternative specifications

INTERIM RESULTS
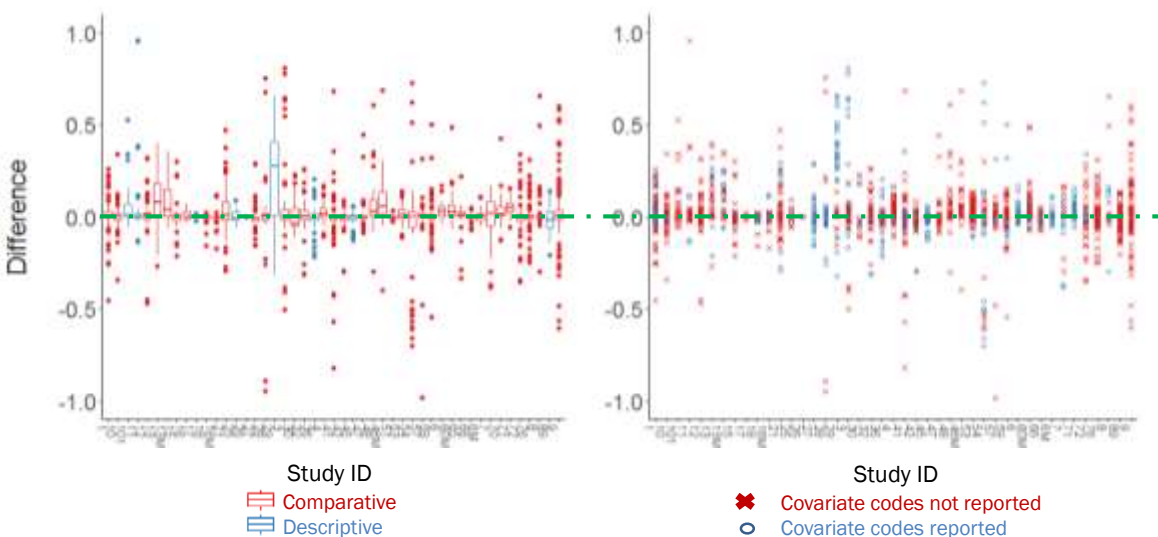## Relative sample size of replication versus original
$(N_{replication}/N_{original})$

No codes, unclear temporality for exposure, exclusion criteria
Replication team made many assumptions

On average, relative sample size close, but many up to 2x as large or 1/2 the size of original publication

Replication larger

Same sample size

Original larger

Comparative   Descriptive

Comparative
Descriptive

Relative study sample size

INTERIM RESULTS
Difference in baseline characteristics* of cohort
(% original – % replication)

Difference

Study ID

Comparative
Descriptive

Study ID

Covariate codes not reported
Covariate codes reported

* binary/categorical

INTERIM RESULTS
Difference in baseline characteristics* of cohort
(% original – % replication)

Study ID — Comparative / Descriptive

Study ID — Covariate codes not reported / Covariate codes reported

85% of baseline characteristics were within ± 10% points

* binary/categorical

INTERIM RESULTS
Difference in baseline characteristics* of cohort
(% original – % replication)

Study ID — Comparative / Descriptive

Study ID — Covariate codes not reported / Covariate codes reported
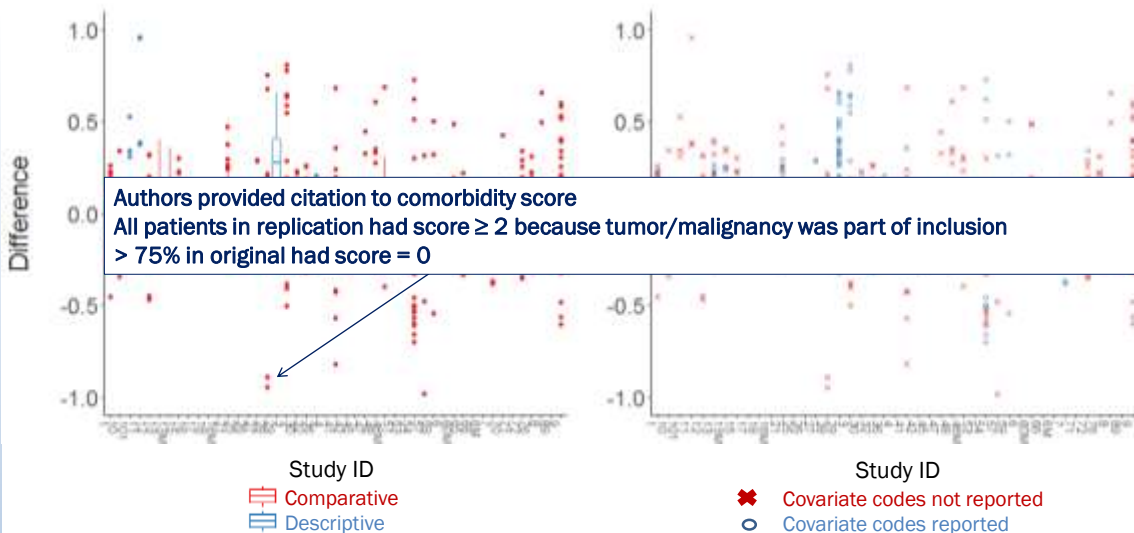
95% of baseline characteristics were within ± 25% points

* binary/categorical

INTERIM RESULTS
### *Why did the replication differ so much from the original for some baseline characteristics?*

Authors provided citation to comorbidity score
All patients in replication had score ≥ 2 because tumor/malignancy was part of inclusion
> 75% in original had score = 0

Study ID
- Comparative
- Descriptive

Study ID
- ✖ Covariate codes not reported
- ⊙ Covariate codes reported

* binary/categorical

INTERIM RESULTS
Calibration of effect estimates* for original versus replication
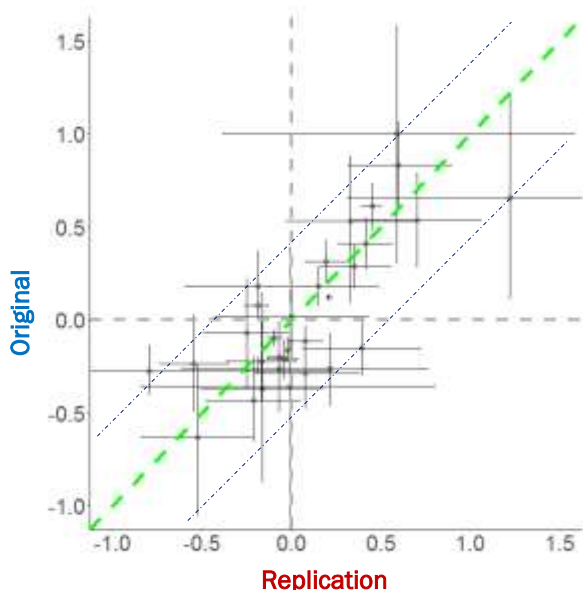

Replication = Original
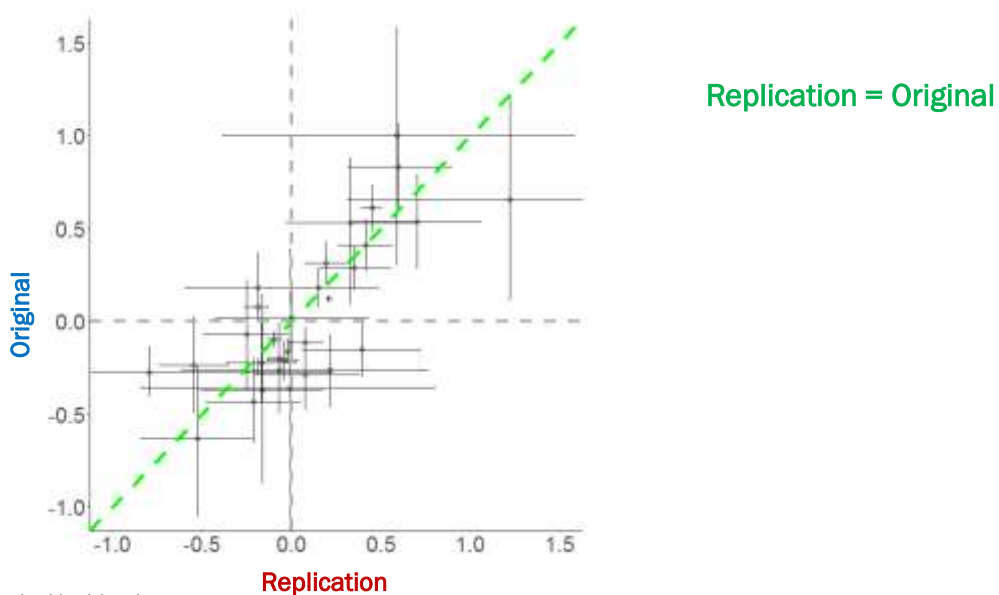
* Log hazard, odds, risk ratio

INTERIM RESULTS
Calibration of effect estimates* for original versus replication

* Log hazard, odds, risk ratio

Estimates follow diagonal

Same side of null?
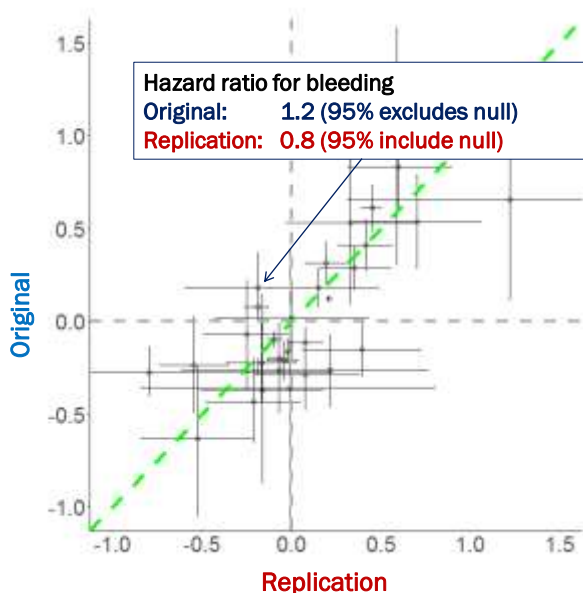
84% of effect estimates were on the same side of null
    16% were not

52% of effect estimates *and* confidence intervals were on same side of null

---

INTERIM RESULTS
Why are the effect estimates on opposite sides of null?

Hazard ratio for bleeding
Original: 1.2 (95% excludes null)
Replication: 0.8 (95% include null)

* Log hazard, odds, risk ratio

Notes from replication team:

- Assumptions regarding algorithms for exclusion, covariates
  - Codes? Care setting? Dx position?
  - Day 0 in assessment window?

- Outcome algorithm provided

- Assumptions about follow up
  - Censoring criteria, exposure stockpiling, bridging, extension

Sample size and characteristics:

- Replication cohort was 30% larger

- Over half of baseline characteristics differed by more than 10% points

Work in progress...

Transparency
Reproducibility
Assessment of validity

Investigator burden
Reviewer burden
Information overload

- **Empirical evaluation**
  - Describe frequency of reporting specific parameters
  - Model impact of transparency of specific parameters on replicability (std. diff effect estimate)
    - → Help focus reporting guidance on underreported parameters with larger influence
- **General comment**
  - Hard to replicate results if unable to replicate analytic cohort
  - Exclusion criteria often mentioned in passing without detail
  - Majority of internal debate over vague prose on temporality (slower timeline for replication)
- **How much do alternative decisions/assumptions for specific parameters matter?**
  - Context dependent, robustness next...

27

REPEAT Core Team (alphabetical)

6 groups working in parallel on different studies (1+ faculty, 2+ research staff)

- Adrian Ortiz Santiago BS
- Ajinkya Pawar PhD MS
- Elisabetta Patorno MD DrPH
- Elizabeth M. Garry PhD MPH
- Emma Payne BS
- Jessica Franklin PhD
- Joshua Gagne PharmD ScD
- Krista Huybrechts PhD MS
- Kristina Stefanini BA
- Lily Bessette BS
- Mimi Zakarian BS
- Monica L. Gierrada MPH
- Mufaddal Mahresi MD MPH
- Nileesa Gautam BS
- Sebastian Schneeweiss MD ScD
- Shirley V Wang PhD ScM
- Sushama Kattinakere MBBS MSPH
- Yinzhu Jin MS MPH

REPEAT
Reproducible Evidence: Practices to Enhance and Achieve Transparency
www.repeatinitiative.org

Scientific Advisory Board (alphabetical)

Regulators, HTA, delivery systems, patients, payers, industry, journals, research societies...

- Jeffrey Brown PhD
- Alison Bourke MSc FRPharm.S
- Amr Makady PharmD PhD
- Andrew Bate PhD
- Brian Bradbury DSc
- Brian Nosek PhD
- Christine Laine MD MPH FACP
- David Martin MD MPH
- Deborah Zarin MD
- Dick Willke PhD
- Dorothee Bartels MSc PhD
- Elizabeth Loder MD MPH
- Frank de Vries PharmD PhD
- Hans-Georg Eichler MD, MSc
- Henrik Toft Sørensen MD PhD
- Javier Jimenez MD MPH
- Jesper Hallas MD PhD
- Joanne Waldstreicher MD
- John Ioannidis MD DSc

- John Seeger PharmD DrPh
- K. Arnold Chan MD ScD
- Karen Burnett MBA MS
- Kris Kahler PhD
- Laura Happe PharmD MPH
- Liam Smeeth PhD
- Lisa Freeman
- Michael Nguyen MD
- Nam-Kyong Choi B. Pharm PhD
- Páll Jónsson PhD Mres
- Peter Arlett BSc MBBS MRCP FFPM
- Peter Tugwell MSc MD FRCPC
- Richard Platt MD MSc
- Sarah Priddy Alwardt PhD
- Sean Hennessy PharmD, PhD
- Troyen Brennan MD
- Will Shrank MD
- Wolfgang Winkelmayer MD MPH ScD FASN
- Yoshiaki Uyama PhD