Emulation differences versus bias when emulating randomized clinical trials

Highlights from the RCT DUPLICATE* initiative

Sebastian Schneeweiss, MD, ScD
Professor of Medicine and Epidemiology

Shirley Wang, PhD
Associate Professor of Medicine

Jessica Franklin, PhD
Associate Professor of Medicine

Division of Pharmacoepidemiology and Pharmacoeconomics, Department of Medicine
Brigham and Women's Hospital, Harvard Medical School, Boston

*Randomized Controlled Trials Duplicated Using Prospective Longitudinal Insurance Claims: Applying Techniques of Epidemiology

This study was funded by FDA under contracts HHSF223201710186C and HHSF223201710146C

Disclosures

Dr. Schneeweiss

- PI, Sentinel Innovation Center (FDA)
- Co-Chair, Mass General Brigham Center for Integrated Healthcare Data Research
- PI of grants and contracts from NIH, AHRQ, PCORI, FDA, IMI, Arnold Foundation
- Investigator of research grants awarded to BWH by Boehringer Ingelheim, UCB
- Consulting fees from Aetion, Inc. (incl. equity)

Dr. Wang

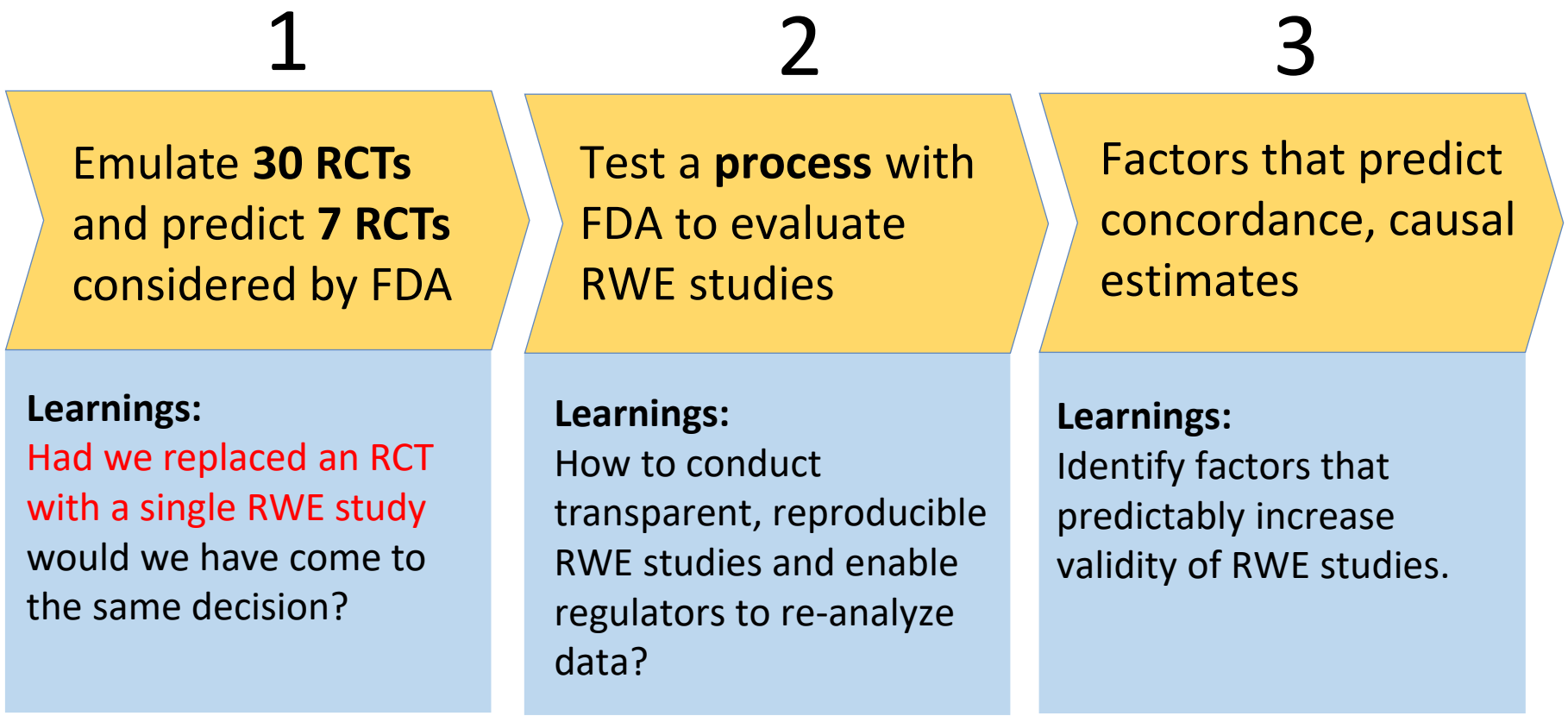
- Supported by grants from FDA Sentinel, NHLBI, NIA, NICHD

Dr. Franklin

- Current employment by Optum, Inc.

RCT-DUPLICATE: A methods demonstration project

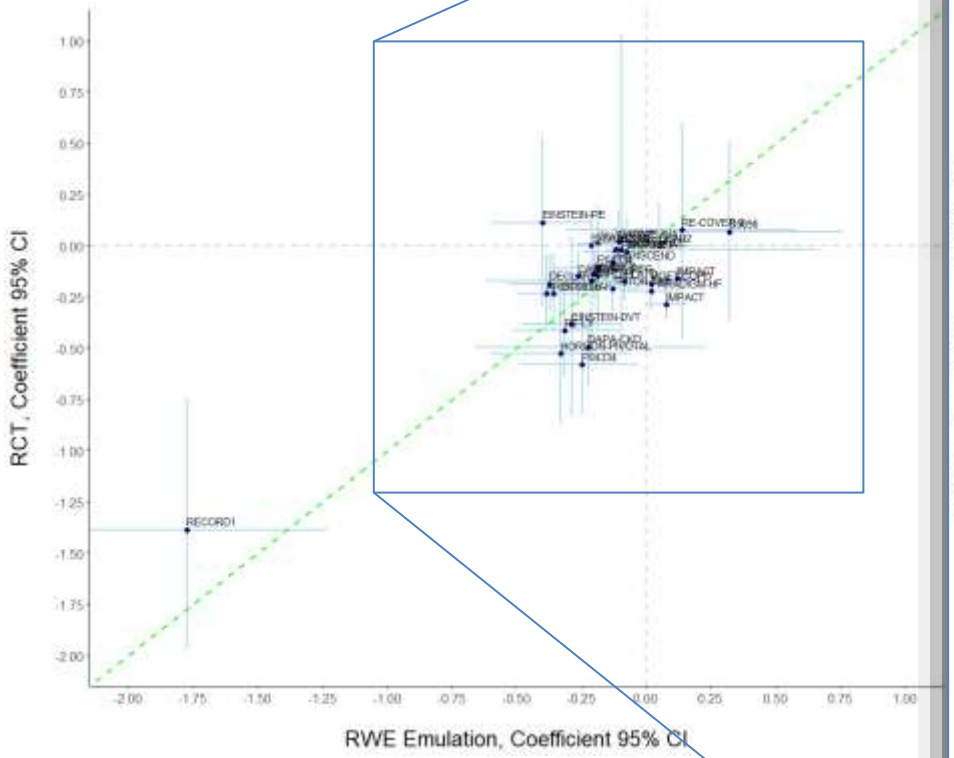
Objective: To understand and improve the validity of RWE studies to support regulatory decision making



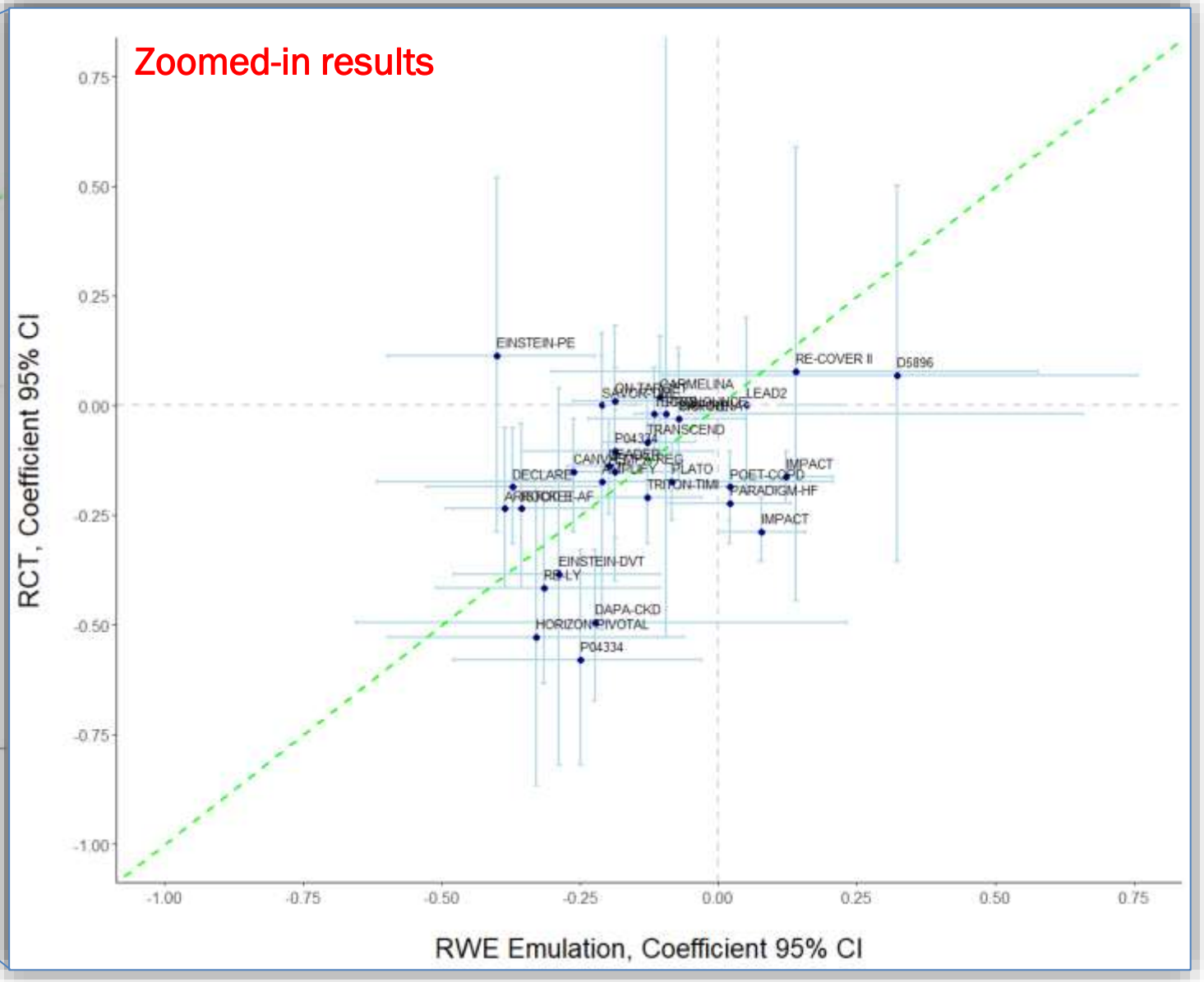
Franklin, Pawar, Martin, Glynn, Levenson, Temple, Schneeweiss. CPT 2020

32 RCT-RWE emulation results

Pearson's overall = 0.80; 0.62-0.90



Zoomed-in results



Emulation differences vs Bias

Are we asking a different question?

Emulation Differences

Differences between RCT and RWE

Population

- Inclusion-exclusion
- Run in

Treatme

- Loadir
  allowa
- Placebo

Outcome ascertainment

- Measurement definition
- Primary vs secondary data collection

Follow up

- Time-varying hazard
- Measures to maximize adherence

Case studies

1. Time varying effects
2. Chance or other factors

Bias

Differences between RWE treatment arms

Confounders

Outcome ascertainment

- Differential surveillance
- Misclassification

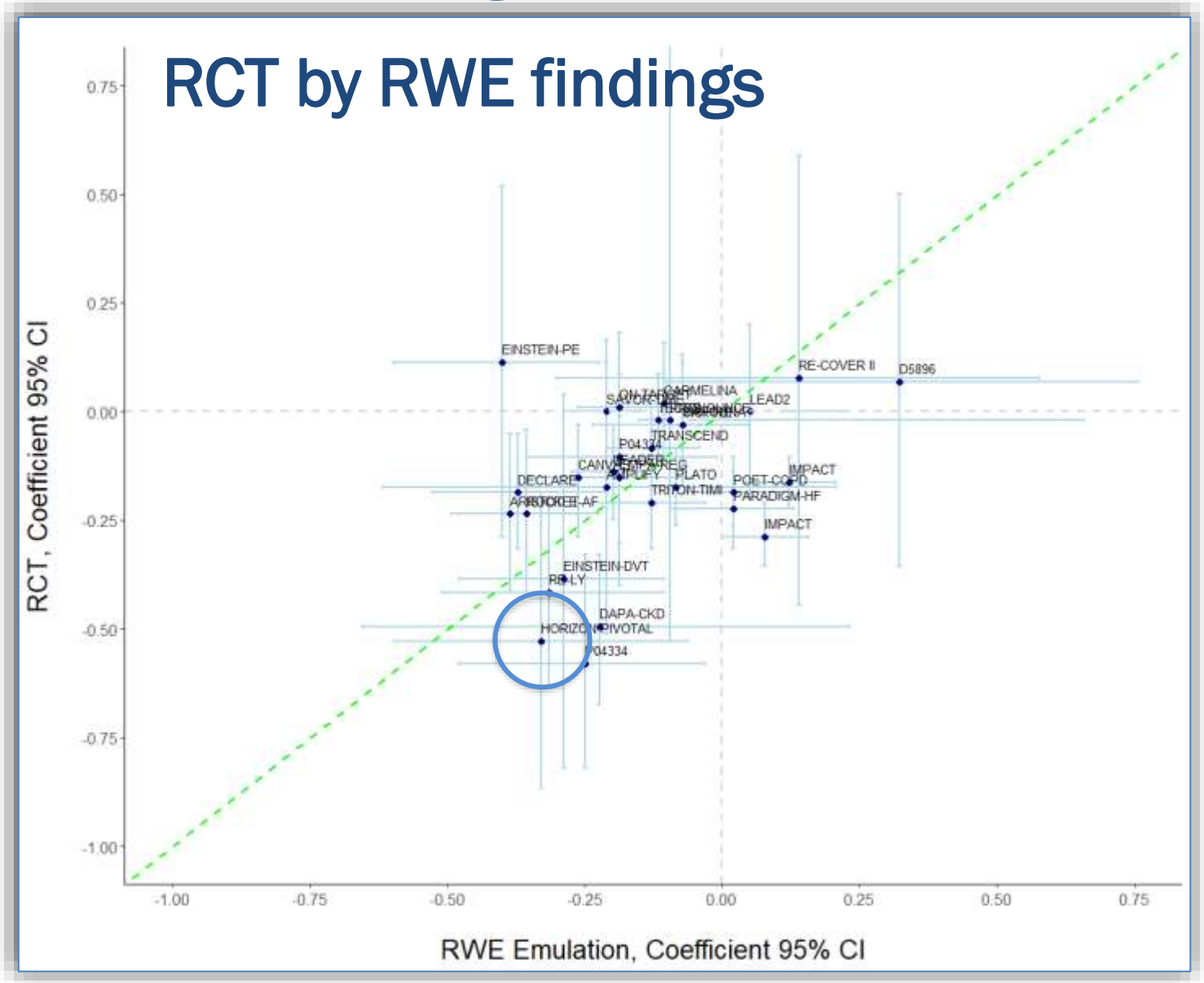
Follow up

- Differential duration
- Informative censoring

Emulation Differences vs. Biases When Calibrating Real-World Evidence Findings Against Randomized Controlled Trials

Jessica M. Franklin[1,*], Robert J. Glynn[1], Samy Suissa[2] and Sebastian Schneeweiss[1]

1. Time varying treatment effects

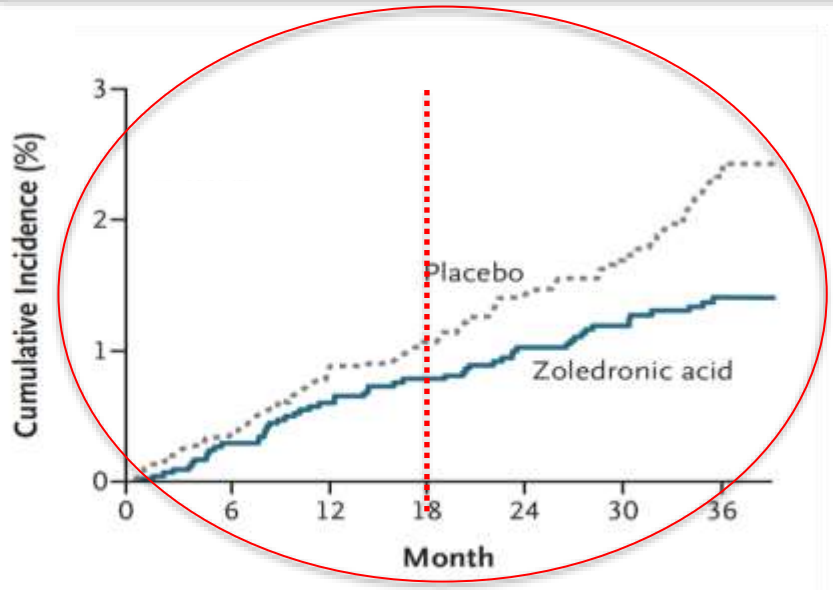


HORIZON-PIVOTAL
RCT: zoledronic acid vs placebo
RWE: zoledronic acid vs ***raloxifene***
Outcome: hip fracture

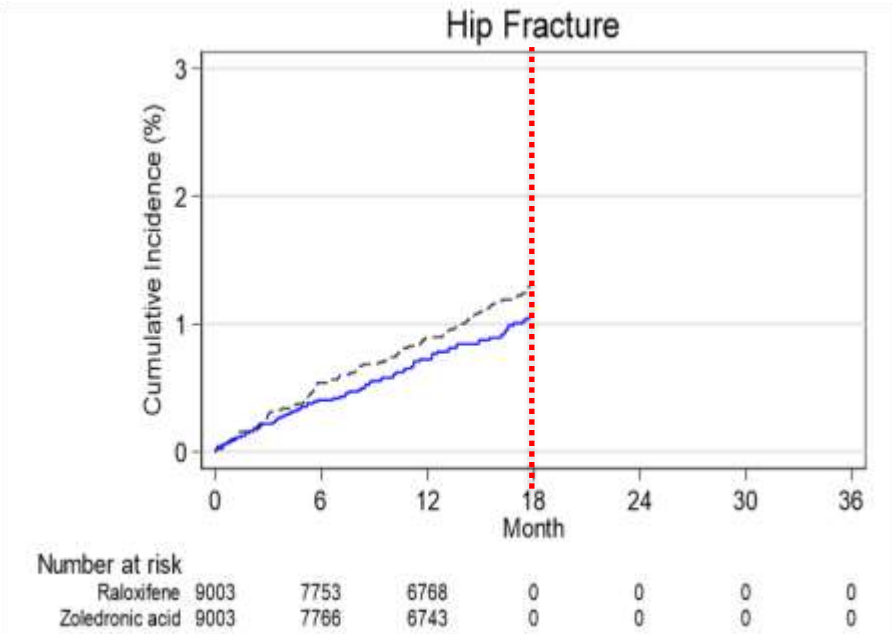
1. Time varying treatment effects

HORIZON-PIVOTAL (osteoporosis, hip fracture)

RCT	RWE
HR_{36mo} = 0.59 (0.42, 0.83)	HR_{36mo} = ??
HR_{18mo} = 0.75	HR_{18mo} = 0.75 (0.58, 0.97)

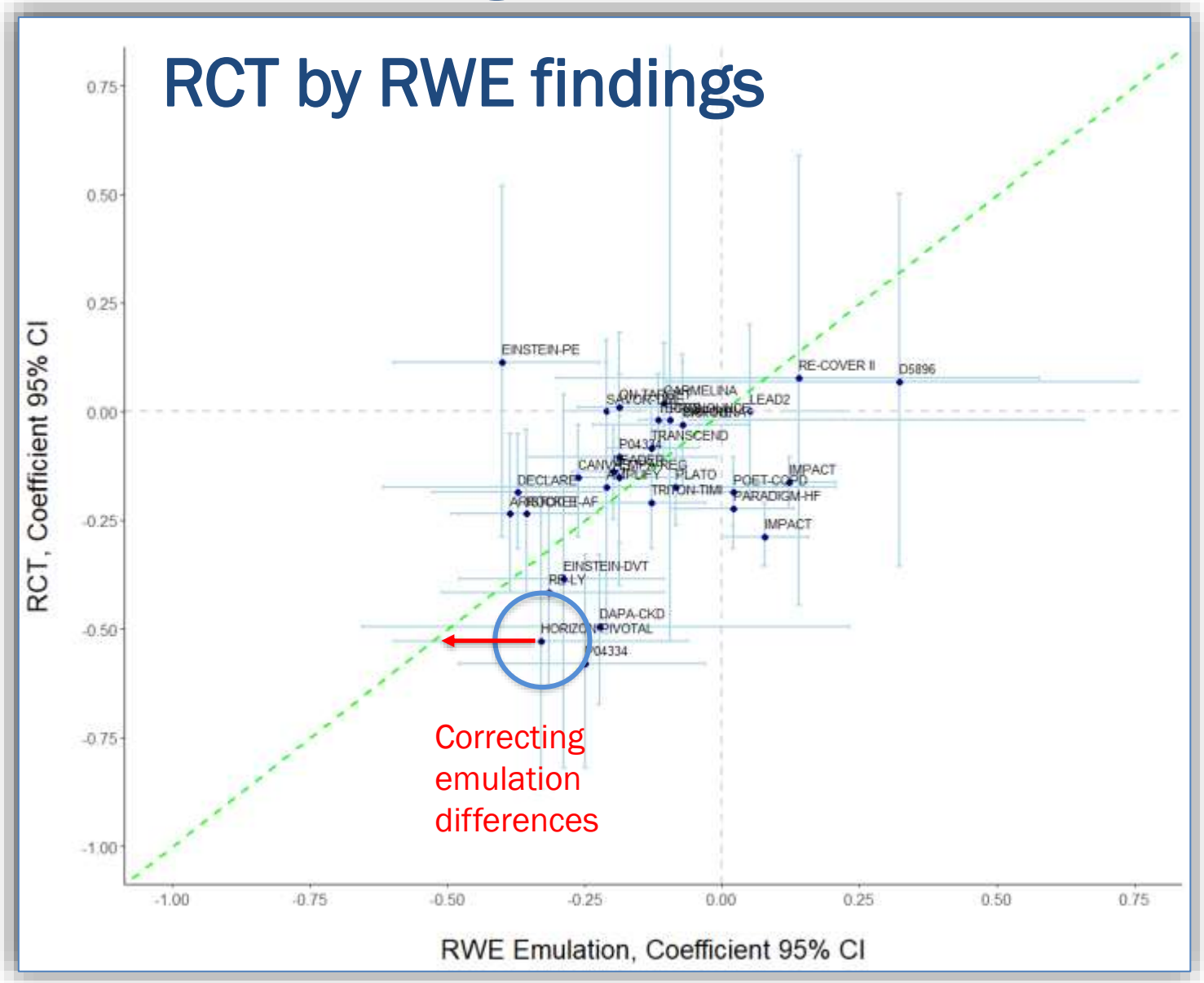


No. at Risk							
Zoledronic acid	3875	3807	3674	3553	3494	3387	3161
Placebo	3861	3806	3694	3577	3499	3397	3144



Number at risk							
Raloxifene	9003	7753	6768	0	0	0	0
Zoledronic acid	9003	7766	6743	0	0	0	0

1. Time varying treatment effects



Take home points:

- Challenging to replicate trial findings when effect is delayed
- Clinical practice patients may not experience full benefit seen in explanatory trial

Challenges with emulation of trial design expected to shift the target question for RWE study vs RCT

a) Start of follow up in hospital (hospital Rx data not available in claims, but may be available in linked data)
b) Run-in that selects responders to one treatment arm
c) Mixing effect of randomization and discontinuation of baseline maintenance therapy
d) Delayed effect over long follow up
e) Differences in population distribution coupled with effect modification
f) Inadequate emulation of the exposure or outcome

Few emulation challenges = None of { a, b, c, d } AND comparator and outcome emulation are at least moderate, with >1 classified as good

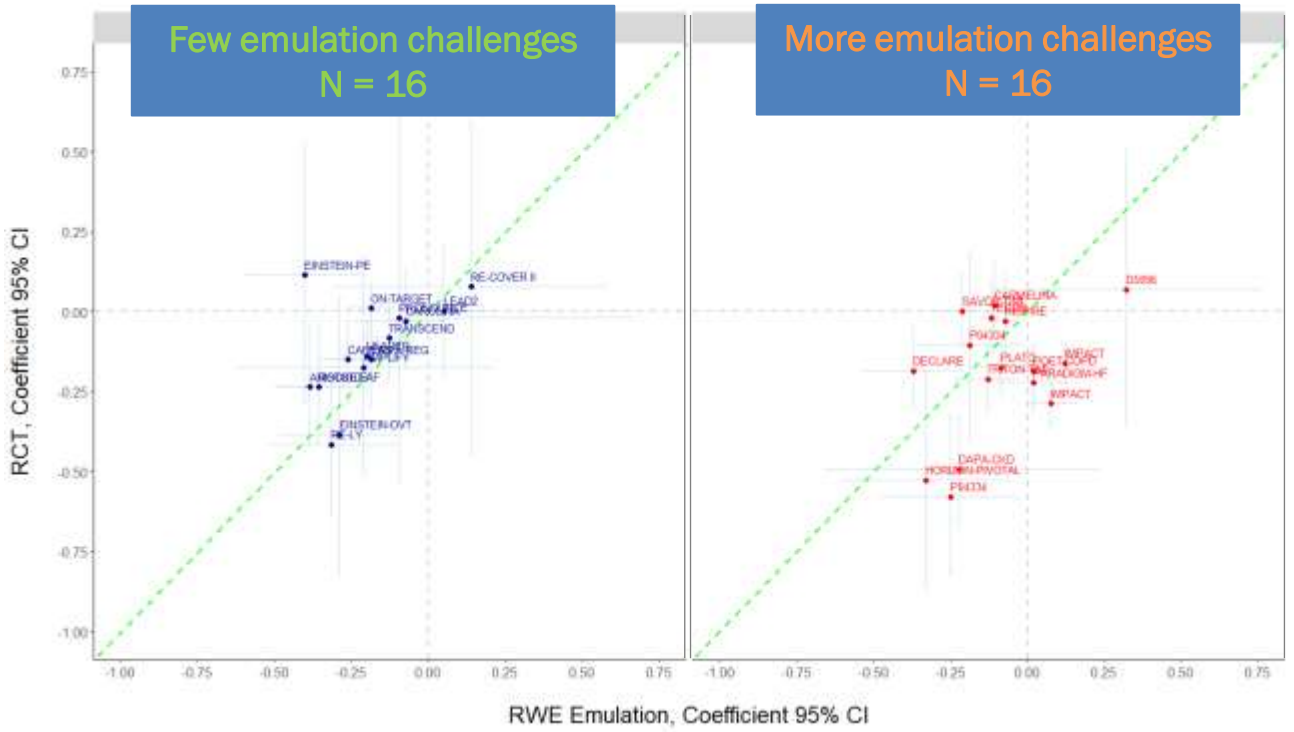
More emulation challenges = a OR b OR c OR d OR poor comparator emulation OR neither comparator and outcome emulation are classified as good

Few emulation challenges vs more emulation challenges

Pearson's overall = 0.80; 0.62-0.90

	Few emulation challenges N = 16	More emulation challenges N = 16
Pearson's	**0.93** (0.79, 0.97)	0.46 (-0.05, 0.78)
ICC, 95% CI	0.89 (0.68, 0.96)	0.41 (-0.03, 0.73)
RA*	16 (100%)	9 (56%)
EA	14 (88%)	7 (44%)
SD	14 (88%)	10 (63%)

ICC = intraclass correlation coefficient; CI = confidence interval; RA = regulatory agreement; EA = estimate agreement; SD = standardized difference agreement



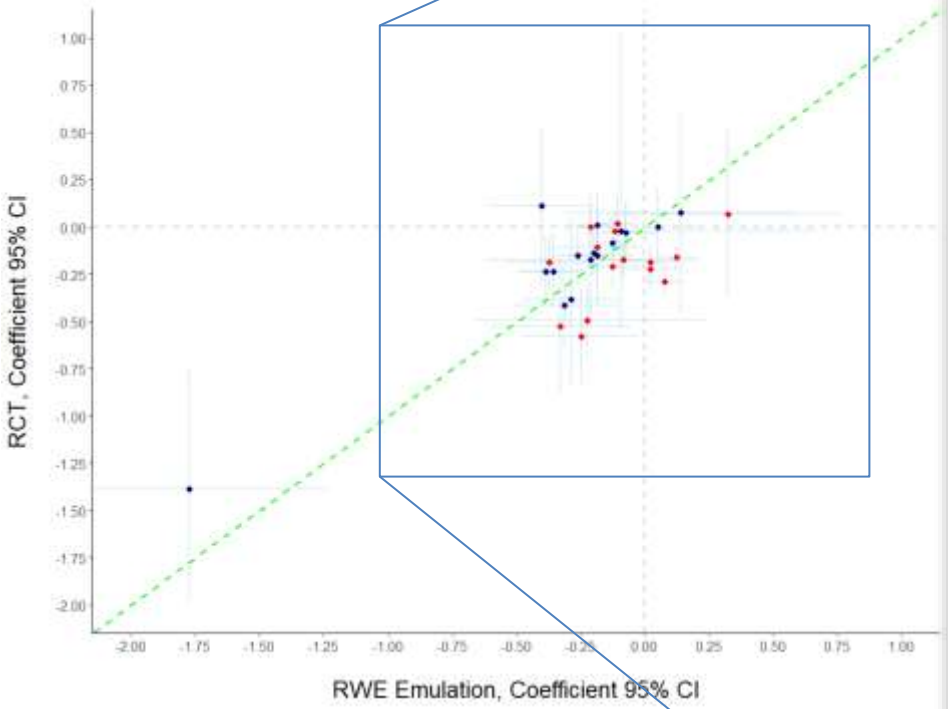
Take-home points:

Recall: For this methods project, the goal was to emulate published RCTs as closely as possible:

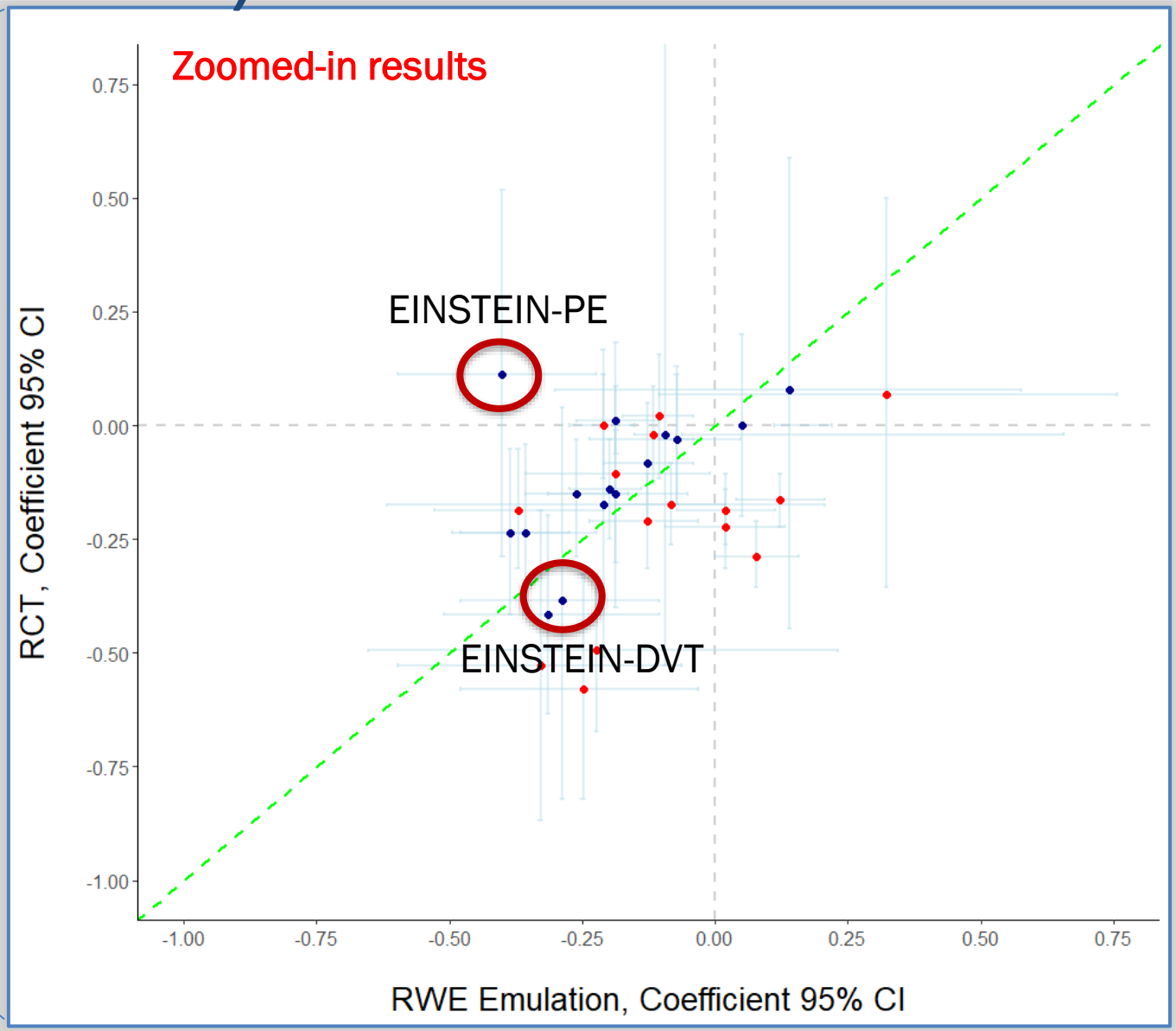
- Few emulation challenges → closer agreement in effect estimates
- More emulation challenges → less agreement in RCT/RWE effect estimates: diverge on target question/pop^{n}? Different answers may be correct.

Total # trials = 32, includes 30 emulations of completed trials and 2 predictions of ongoing trials.
* Proportions represent full or partial regulatory agreement. Full agreement only was 75% vs 38%)

2. Chance? (or other factors)



Zoomed-in results



- Few emulation challenges
- More emulation challenges

2. Chance? (or other factors)

VTE

Trial name	Comparator	Endpoint	RCT	RWE	Stand. Diff.	Test	Agreement			Indication
EINSTEIN-DVT	Rivaroxaban vs Enoxaparin/VKA	VTE	**0.68** (0.44, 1.04)	**0.75** (0.63, 0.89)	-0.42	NI	*	EA	SD	DVT
EINSTEIN-PE	Rivaroxaban vs Enoxaparin/VKA	VTE	**1.12** (0.75, 1.68)	**0.68** (0.58, 0.81)	2.21	NI	-	-	-	PE

- Both met non-inferiority criteria
- P-value for homogeneity 0.09

Meta-analysis of 6 trials* finds no heterogeneity of effects in patients presenting with DVT or PE.

**Dentali F, et al. Intern Emerg Med. 2015*

Good

Moderate

Poor

Take-home points

1. RWE studies come to the same conclusions as RCTs when we are able to emulate well, i.e. target the same question

2. There is more nuance to evaluation of replicability of trial results with RWE than can be found in binary agreement metrics.
    - Residual bias, random error
    - Efficacy vs effectiveness
    - Single trial as reference standard

3. In evaluating when and how RWE studies complement RCTs, we should think about the hypothetical target trial that would address the need/question of end users (ideal vs pragmatic)

With data that are fit-for-purpose and proper design and analysis, non-randomized real-world evidence studies can come to similar conclusions about a drug's treatment effect as randomized trials

Harvard study team:

Faculty: Drs. Schneeweiss, Franklin, Wang, Glynn, Patorno, Desai, Choudhry, Huybrechts, Fischer, Feldman, Gagne, Bykov

Research Staff: Bessette, Dr. D'Andrea, Chin, Gautham, Dr. Gopalakrishna, Jawaid, Jin, Lee, Dr. Mahesri, Dr. Pawar, Sears, Tesfaye, Umarje, York, Zabotka, Zakoul

FDA colleagues:

Drs. Martin, Quinto, Concato, Corrigan-Curay, Paraoan

Aetion team:

Drs. Garry, Rassen, and Isaman, Gibbs, Gilpin

Expert advisor panel:*

Drs. Steve Goodman, Stanford; Miguel Hernan, Harvard; Wayne Ray, Vanderbilt; Samy Suissa, McGill; Alan Brookhart, Duke

*While we are most grateful for the advice we received, the authors are solely responsible for the presented work