A journey toward real-world evidence for regulatory decision-making:

Proving reliable *real-world evidence*: Replicating RCTs using LEGEND

Patrick Ryan, PhD, Vice President of Observational Health Data Analytics at Janssen Research & Development; Adjunct Assistant Professor of Biomedical Informatics at Columbia University

George Hripcsak, MD, MS, Vivian Beaumont Allen Professor and Chair of Biomedical Informatics at Columbia University Irving Medical Center; Director of Medical Informatics Services at NewYork-Presbyterian Hospital/Columbia Campus

OHDSI

Observational Health Data Sciences and Informatics

Guidance for Industry

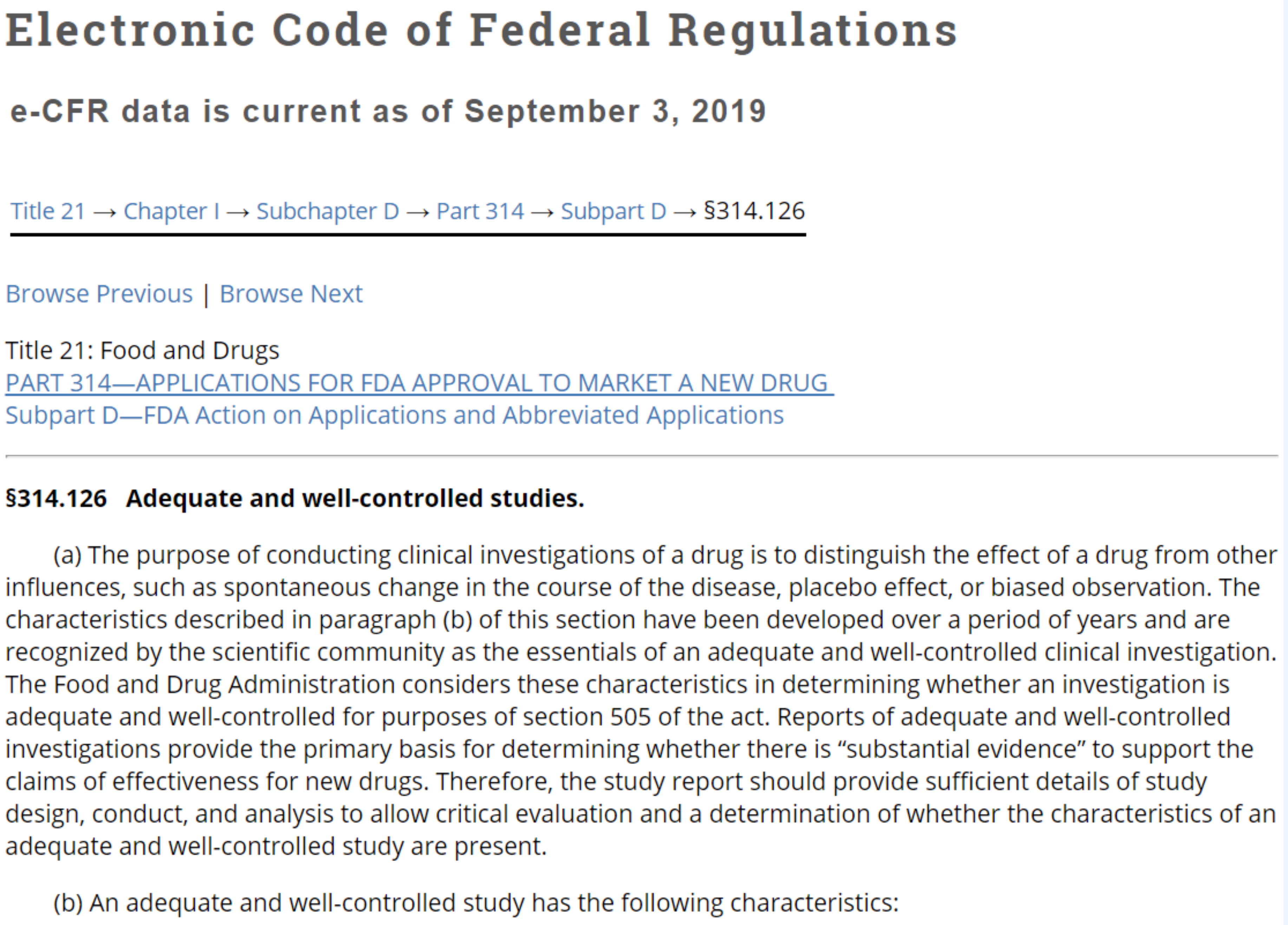
Providing Clinical Evidence of Effectiveness for Human Drug and Biological Products

U.S. Department of Health and Human Services
Food and Drug Administration
Center for Drug Evaluation and Research (CDER)
Center for Biologics Evaluation and Research (CBER)
May 1998
Clinical 6

1962 FDC Act section 505(d)
Substantial evidence:
"evidence consisting of **adequate and well-controlled investigations**, including clinical investigations, by experts qualified by scientific training and experience to evaluate the effectiveness of the drug involved, on the basis of which it could fairly and responsibly be concluded by such experts that the drug will have the effect it purports or is represented to have under the conditions of use prescribed, recommended, or suggested in the labeling or proposed labeling thereof"

1997 FDAMA section 115(a)
"If the Secretary determines, based on relevant science, that data from **one adequate and well-controlled clinical investigation and confirmatory evidence** (obtained prior to or after such investigation) are sufficient to establish effectiveness, the Secretary may consider such data and evidence to constitute substantial evidence"

https://www.fda.gov/media/71655/download



Electronic Code of Federal Regulations

e-CFR data is current as of September 3, 2019

Title 21 → Chapter I → Subchapter D → Part 314 → Subpart D → §314.126

Browse Previous | Browse Next

Title 21: Food and Drugs
PART 314—APPLICATIONS FOR FDA APPROVAL TO MARKET A NEW DRUG
Subpart D—FDA Action on Applications and Abbreviated Applications

§314.126 Adequate and well-controlled studies.

(a) The purpose of conducting clinical investigations of a drug is to distinguish the effect of a drug from other influences, such as spontaneous change in the course of the disease, placebo effect, or biased observation. The characteristics described in paragraph (b) of this section have been developed over a period of years and are recognized by the scientific community as the essentials of an adequate and well-controlled clinical investigation. The Food and Drug Administration considers these characteristics in determining whether an investigation is adequate and well-controlled for purposes of section 505 of the act. Reports of adequate and well-controlled investigations provide the primary basis for determining whether there is "substantial evidence" to support the claims of effectiveness for new drugs. Therefore, the study report should provide sufficient details of study design, conduct, and analysis to allow critical evaluation and a determination of whether the characteristics of an adequate and well-controlled study are present.

(b) An adequate and well-controlled study has the following characteristics:

https://www.ecfr.gov/cgi-bin/text-idx?SID=d029468a9e69025b8d25795d2585ece9&mc=true&node=se21.5.314_1126&rgn=div8

'Adequate and well-controlled investigation' criteria	Threat to validity
There is a clear statement of the objectives of the investigation and a summary of the methods of analysis in the protocol for the study.	Investigator bias
The study uses a design that permits a valid comparison with a control to provide a quantitative assessment of drug effect.	Selection bias
The method of selection of subjects provides adequate assurance that they have the disease or condition being studied.	Measurement error
The method of assigning patients to treatment and control groups minimizes bias and is intended to assure comparability of the groups.	Confounding
Adequate measures are taken to minimize bias on the part of the subjects, observers, and analysts of the data.	Selection bias
The methods of assessment of subjects' response are well-defined and reliable.	Measurement error
There is an analysis of the results of the study adequate to assess the effects of the drug. The report of the study should describe the results and the analytic methods used to evaluate them, including any appropriate statistical methods.	Model misspecification

https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfcfr/cfrsearch.cfm?fr=314.126

‘Adequate and well-controlled investigation’ criteria	Threat to validity	OHDSI RWE solution
There is a clear statement of the objectives of the investigation and a summary of the methods of analysis in the protocol for the study.	Investigator bias	Fully and pre-specified protocol and source code made publicly available prior to study conduct. BoO: Study steps, Network research
The study uses a design that permits a valid comparison with a control to provide a quantitative assessment of drug effect.	Selection bias	Study design choice (comparative cohort vs. self-controlled). Study diagnostics: Propensity score overlap, covariate balance before adjustment. BoO: Population-level effect estimation
The method of selection of subjects provides adequate assurance that they have the disease or	Measurement error	Phenotype evaluation of indication. Generalizability t to target population. BoO: Defining aracterization, Clinical Validity
The method of assigning patients to trea groups minimizes bias and is intended t the groups.		nostics: propensity score overlap, covariate gative control calibration lation-level effect estimation
Adequate measures are taken to minimi the subjects, observers, and analysts of t		nostics: covariate balance after adjustment, ntrol calibration. lation-level effect estimation
The methods of assessment of subjects' defined and reliable.		evaluation of outcome. quality, Clinical validity
There is an analysis of the results of the study adequate to assess the effects of the drug. The report of the study should describe the results and the analytic methods used to evaluate them, including any appropriate statistical methods.	Model misspecification	Study diagnostics: negative control calibration Pre-specification BoO: Methods validity, Software validity

How do we know if we can trust real-world evidence?

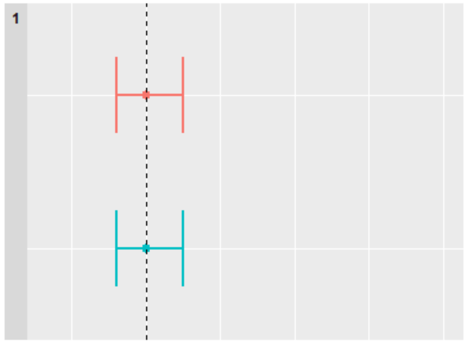
Compare to evidence that we already trust:
randomized clinical trials

https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfcfr/cfrsearch.cfm?fr=314.126

Even if RWE is reliable, we wouldn't necessarily expect it to match RCTs

- Greater sample size
- Real world practice: effectiveness vs. efficacy
- Generalizable populations

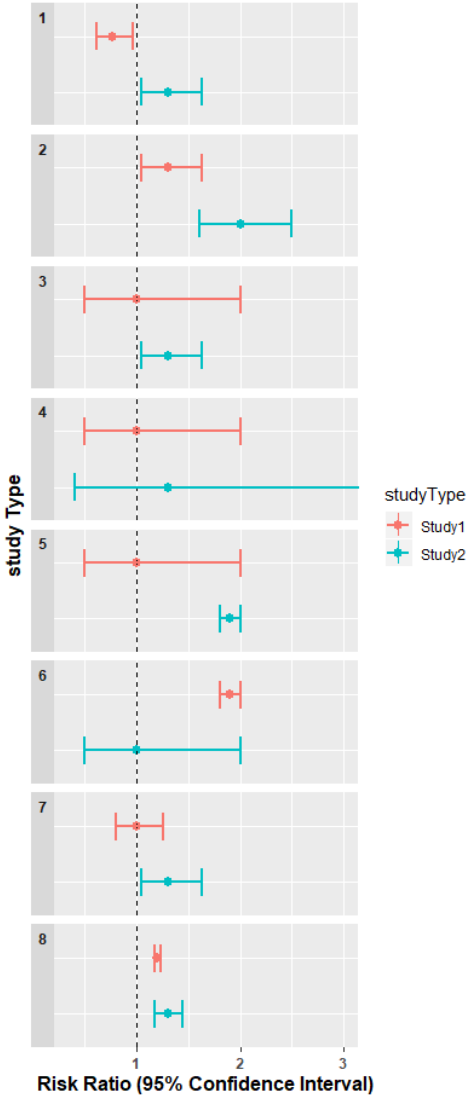
When do two study estimates agree?



If Study 1 produces an effect estimate of RR=1.00 (0.80-1.25), and a second study replicating the first produces the effect estimate of RR=1.00 (0.80-1.25), would you conclude the two studies are in agreement?

What if Study 1 produces a statistically significant ***decreased*** risk and the Study 2 yields a statistically significant ***increased*** risk would you conclude the two studies are in agreement?

When do two study estimates agree?



What if both estimates are statistically significant?

What if only one study is significant, but one study confidence interval is subsumed by the other?

What if study 2 has more uncertainty?

What if study 2 has tight confidence interval which far from study 1 point estimate?

What if study 2 occurred first?

What if confidence intervals partially overlap?

What if both studies are low variance?

How to measure concordance

- What has been suggested
  - Statistical concordance z test
  - Study 2 estimate agreement
  - Study 1 estimate agreement
  - Statistical decision agreement
  - Meta-analysis variance test
- Others we looked at
  - Cross entropy, KL-divergence, max log likelihood, …

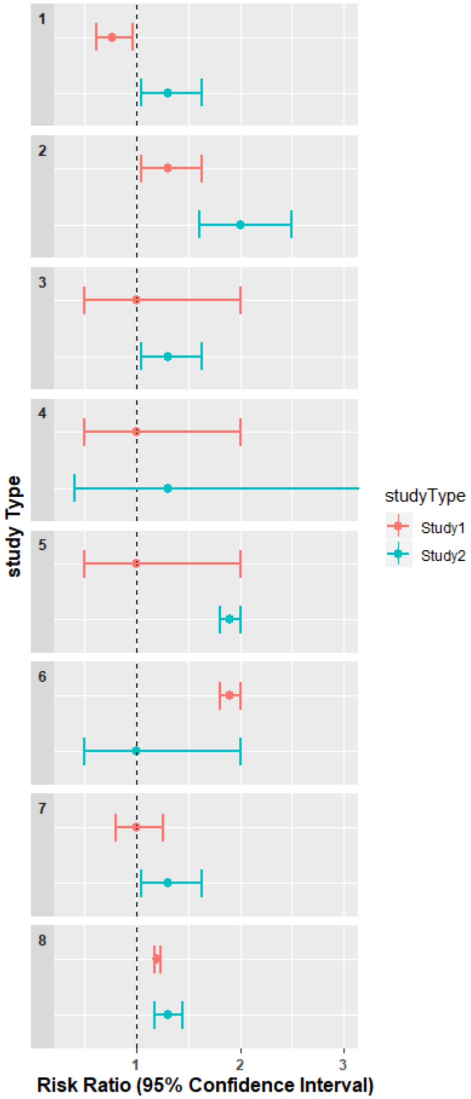
Statistical concordance

Statistical concordance using a z-test measures what we want to know, do two study results differ

- We are comparing one statistical analysis to another (RCT to observational)
  - We seem to trust statistics, so use it!
- Null hypothesis
  - Two studies are identical, differ only in #subjects
    - E.g., split a real study into two (unequal) parts
- z-test whether the two results differ
  - Points estimates and variance
  - Under the null, find a difference 5% of the time
    - Don't get perfect concordance even when identical

Altman BMJ 2003

When do two study estimates agree?



Concordance
Z test

FAIL

FAIL

PASS

PASS

PASS

PASS

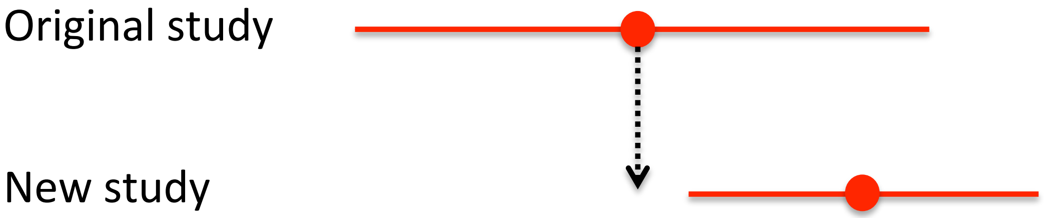
PASS

PASS

Study 1 estimate agreement

A heuristic using to test whether the original effect size is within the 95% confidence interval of the effect size estimate from the replication.

- Bad if second study has more subjects
- Don’t get perfect concordance even when identical

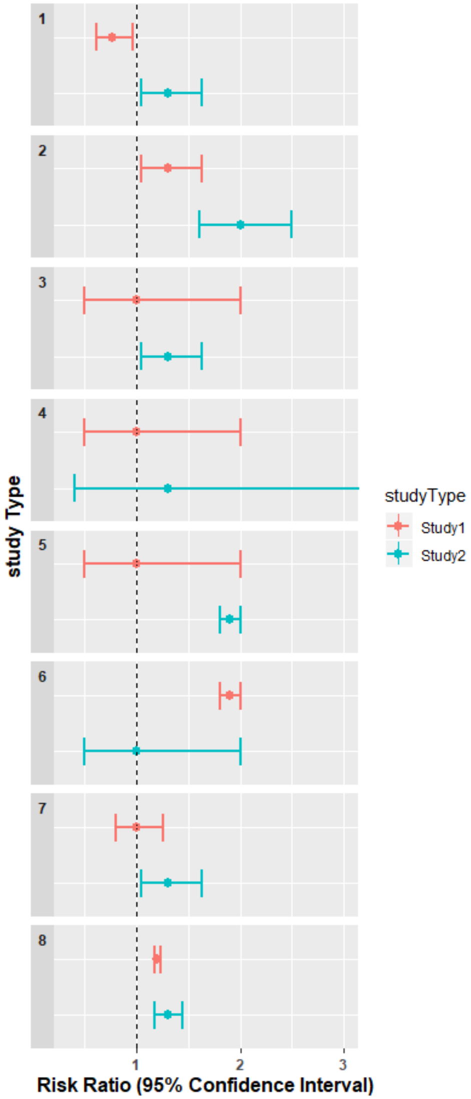


Original study

New study

Nosek Science 2015

When do two study estimates agree?



Study 1 agreement

FAIL

FAIL

FAIL

PASS

FAIL

PASS

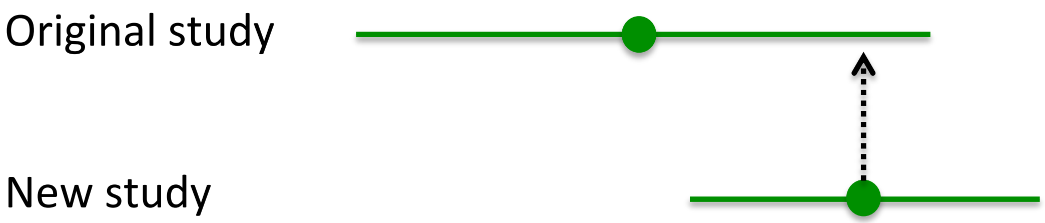
FAIL

PASS

Study 2 estimate agreement

A heuristic using the first study as a gold standard

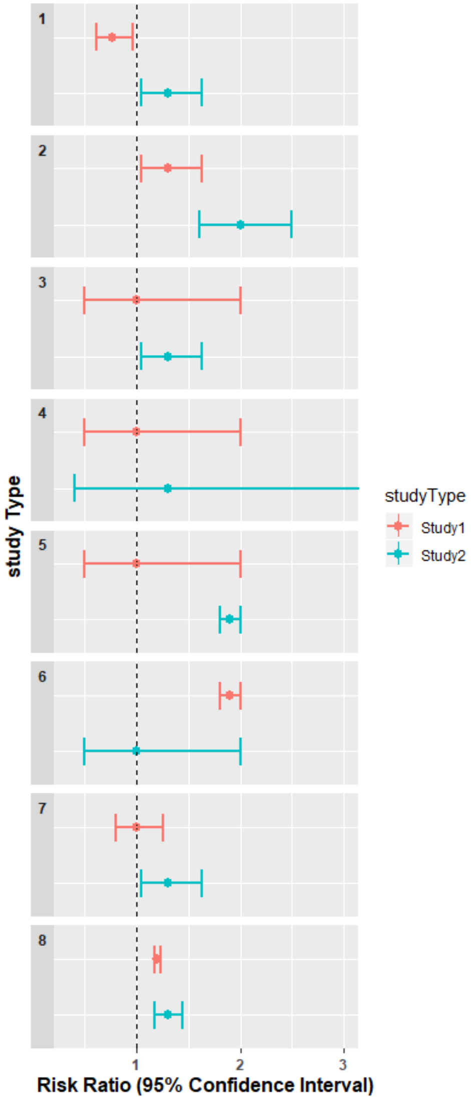
- Is the second study’s estimate in the first study’s confidence interval?
- Not account for variance in the study 2 estimate
- Don’t get perfect concordance even when identical



Original study

New study

When do two study estimates agree?



Study 2 agreement

FAIL

FAIL

PASS

PASS

PASS

FAIL

FAIL

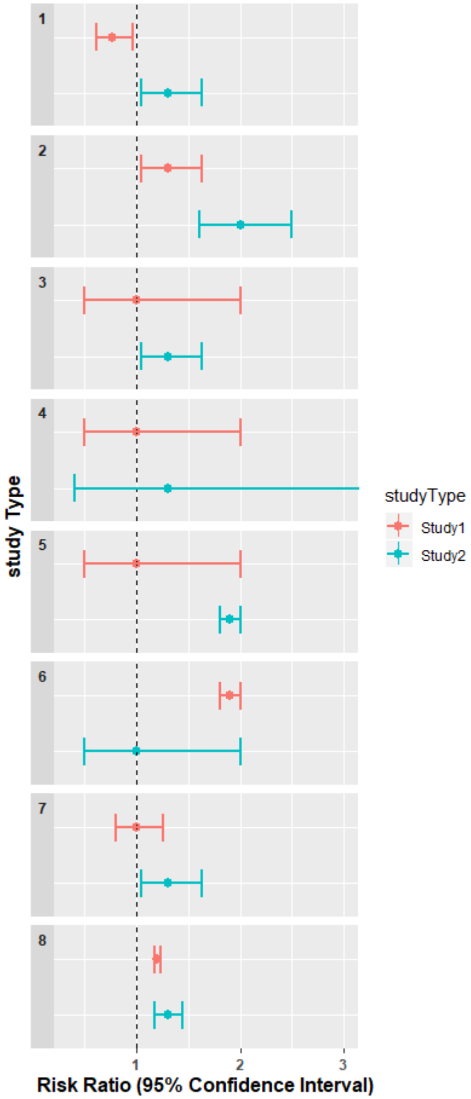
FAIL

Statistical decision agreement

Do you get the same headline:
A causes B, A prevents B, or no effect

- Do the studies agree in significance
  - Both statistically significantly greater than 1
  - Both statistically significantly less than 1
  - Both non-significant
- What effect on regulatory decision?
- Uses the *false* interpretation that non-significant means no effect
  - Difference in sample size for otherwise identical studies can lead to discordance most of the time

When do two study estimates agree?



Significance agreement

FAIL

PASS

FAIL

PASS

FAIL

FAIL

FAIL

PASS

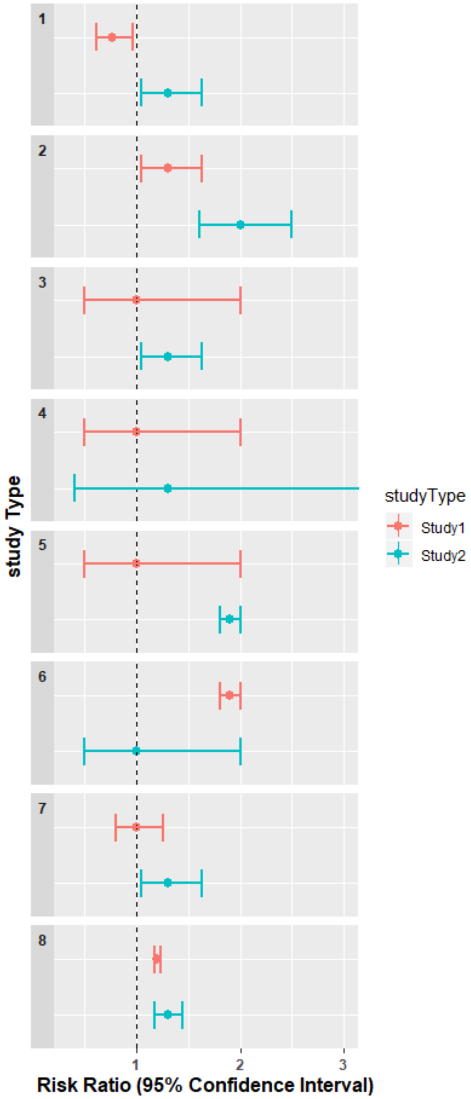
Meta-analysis variance test

Does the second study add useful information

- Carry out meta-analysis of the two studies
  - If the confidence interval of the combination is smaller than that of the original study, then information that is concordant has been added

Nosek Science 2015

When do two study estimates agree?



Meta-analysis variance test

FAIL

FAIL

PASS

PASS

PASS

FAIL

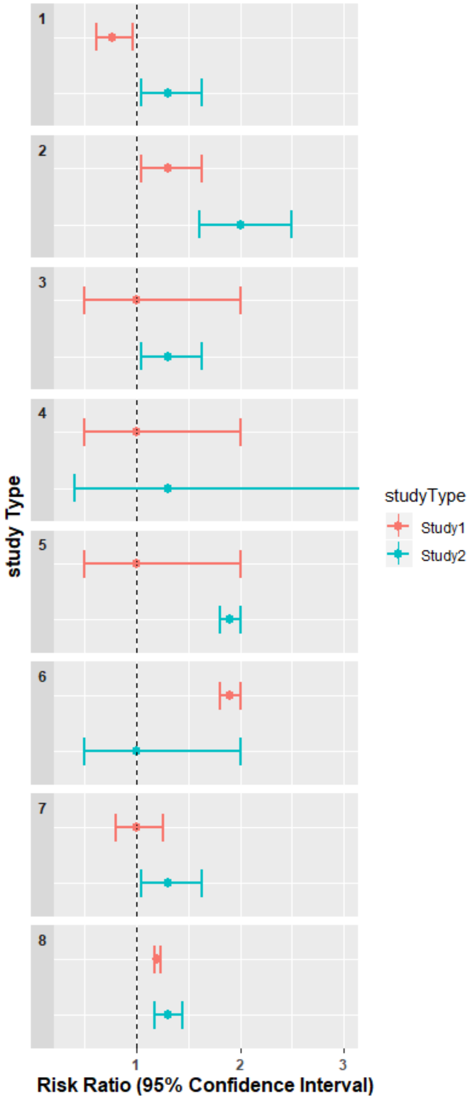
FAIL

FAIL

Comparing agreement metrics

Agreement metric	Pros	Cons
Concordance Z test	• It’s what you really want • Easy to describe what it is testing	• You can do well with wide confidence intervals • Can do well by guessing no effect
Study 1 agreement	• Easy to carry out	• Only a heuristic • Can do well by guessing no effect
Study 2 agreement	• Mostly similar to statistical concordance • Easy to carry out • Tends to penalize wide confidence intervals in study 2	• Only a heuristic • Can do well by guessing no effect
Significance decision agreement	• Easy to carry out • Indicates what effect it could have on regulation	• Nearly identical studies can be discordant • Identical studies can have mostly different results • Can well by guessing no effect
Meta-analysis variance test	• Adds in value of the second study	• Hard to carry out • Can do well by guessing no effect

When do two study estimates agree?

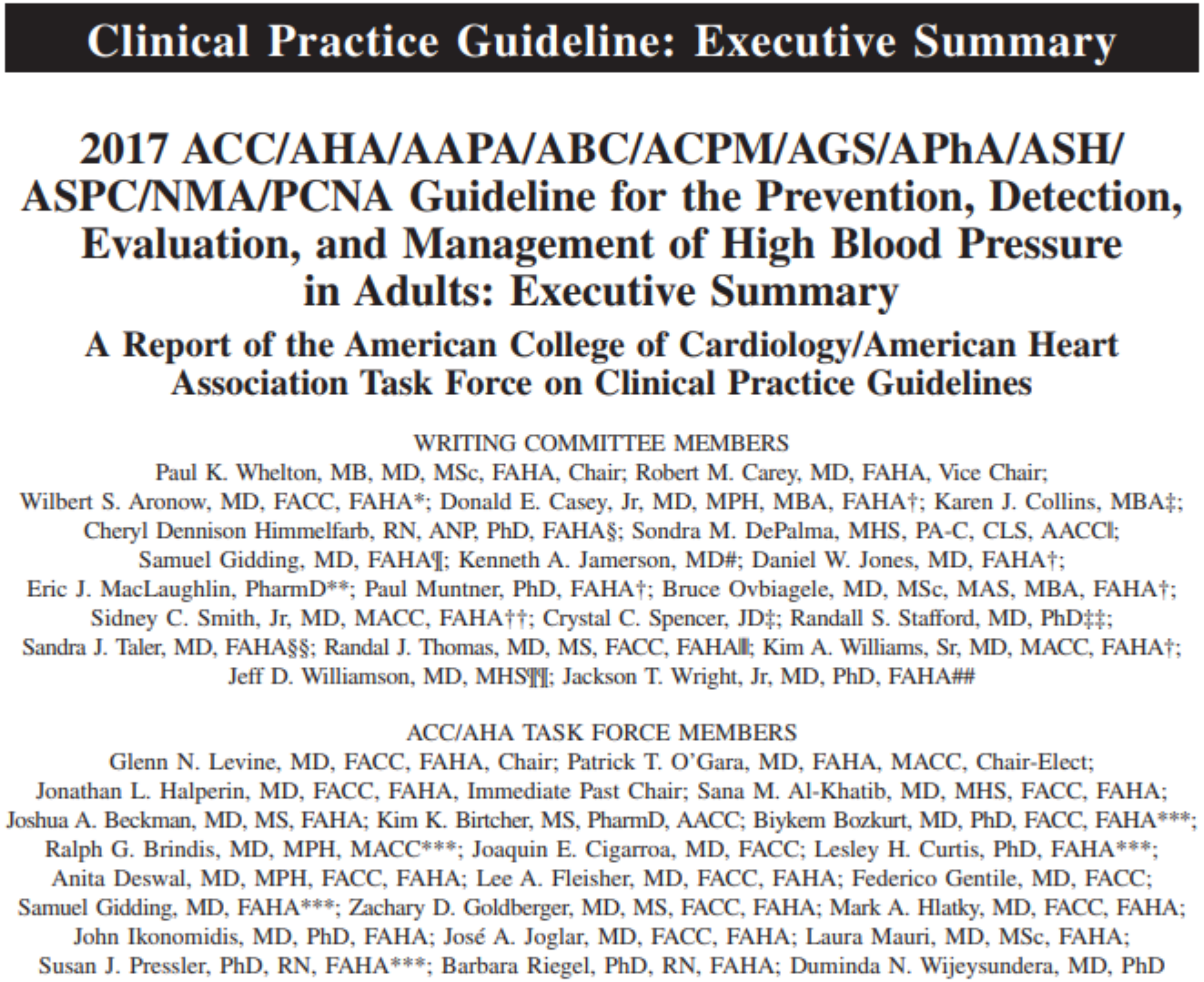


Concordance Z test	Study 1 agreement	Study 2 agreement	Significance agreement	Meta-analysis variance test
FAIL	FAIL	FAIL	FAIL	FAIL
FAIL	FAIL	FAIL	PASS	FAIL
PASS	FAIL	PASS	FAIL	PASS
PASS				PASS
PASS				PASS
PASS				FAIL
PASS				FAIL
PASS	PASS	FAIL	PASS	FAIL

Different ‘agreement metrics’ will produce different results using the same set of studies.

So if you want to compare studies, you need to benchmark how well your ‘agreement metrics’ perform

Revisiting the Hypertension guideline



Clinical Practice Guideline: Executive Summary

2017 ACC/AHA/AAPA/ABC/ACPM/AGS/APhA/ASH/ASPC/NMA/PCNA Guideline for the Prevention, Detection, Evaluation, and Management of High Blood Pressure in Adults: Executive Summary

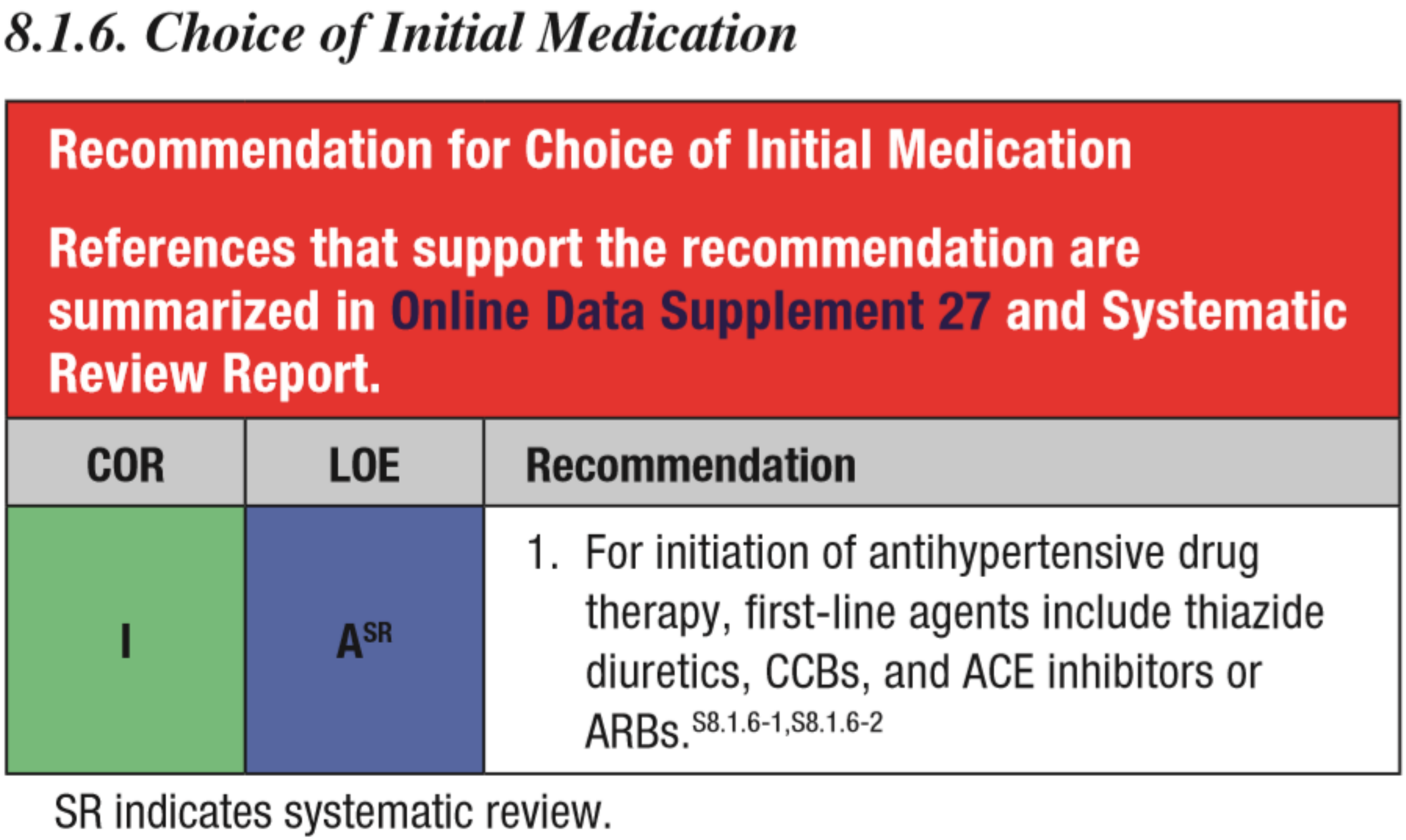
A Report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines

WRITING COMMITTEE MEMBERS

Paul K. Whelton, MB, MD, MSc, FAHA, Chair; Robert M. Carey, MD, FAHA, Vice Chair;
Wilbert S. Aronow, MD, FACC, FAHA[*]; Donald E. Casey, Jr, MD, MPH, MBA, FAHA[†]; Karen J. Collins, MBA[‡];
Cheryl Dennison Himmelfarb, RN, ANP, PhD, FAHA[§]; Sondra M. DePalma, MHS, PA-C, CLS, AACC[‖];
Samuel Gidding, MD, FAHA[¶]; Kenneth A. Jamerson, MD[#]; Daniel W. Jones, MD, FAHA[†];
Eric J. MacLaughlin, PharmD[**]; Paul Muntner, PhD, FAHA[†]; Bruce Ovbiagele, MD, MSc, MAS, MBA, FAHA[†];
Sidney C. Smith, Jr, MD, MACC, FAHA[††]; Crystal C. Spencer, JD[‡]; Randall S. Stafford, MD, PhD[‡‡];
Sandra J. Taler, MD, FAHA[§§]; Randal J. Thomas, MD, MS, FACC, FAHA[‖]; Kim A. Williams, Sr, MD, MACC, FAHA[†];
Jeff D. Williamson, MD, MHS[¶¶]; Jackson T. Wright, Jr, MD, PhD, FAHA[##]

ACC/AHA TASK FORCE MEMBERS

Glenn N. Levine, MD, FACC, FAHA, Chair; Patrick T. O'Gara, MD, FAHA, MACC, Chair-Elect;
Jonathan L. Halperin, MD, FACC, FAHA, Immediate Past Chair; Sana M. Al-Khatib, MD, MHS, FACC, FAHA;
Joshua A. Beckman, MD, MS, FAHA; Kim K. Birtcher, MS, PharmD, AACC; Biykem Bozkurt, MD, PhD, FACC, FAHA[***];
Ralph G. Brindis, MD, MPH, MACC[***]; Joaquin E. Cigarroa, MD, FACC; Lesley H. Curtis, PhD, FAHA[***];
Anita Deswal, MD, MPH, FACC, FAHA; Lee A. Fleisher, MD, FACC, FAHA; Federico Gentile, MD, FACC;
Samuel Gidding, MD, FAHA[***]; Zachary D. Goldberger, MD, MS, FACC, FAHA; Mark A. Hlatky, MD, FACC, FAHA;
John Ikonomidis, MD, PhD, FAHA; José A. Joglar, MD, FACC, FAHA; Laura Mauri, MD, MSc, FAHA;
Susan J. Pressler, PhD, RN, FAHA[***]; Barbara Riegel, PhD, RN, FAHA; Duminda N. Wijeysundera, MD, PhD



8.1.6. Choice of Initial Medication

Recommendation for Choice of Initial Medication

References that support the recommendation are summarized in Online Data Supplement 27 and Systematic Review Report.

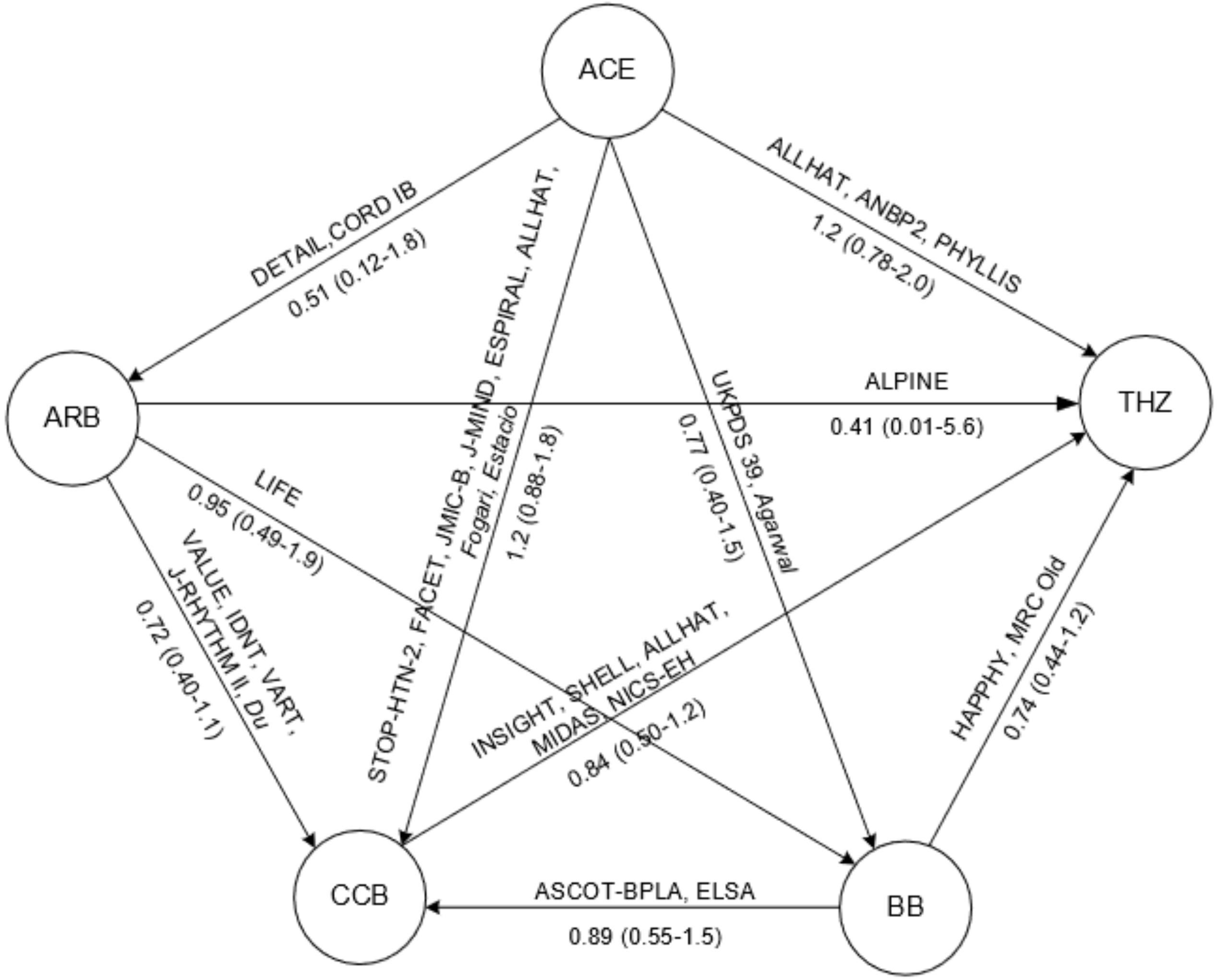
COR	LOE	Recommendation
I	A^{SR}	1. For initiation of antihypertensive drug therapy, first-line agents include thiazide diuretics, CCBs, and ACE inhibitors or ARBs.[S8.1.6-1,S8.1.6-2]

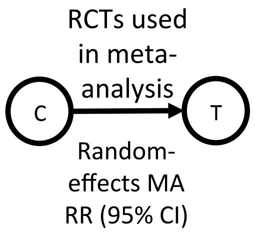
SR indicates systematic review.

Whelton et al., Hypertension 2018

RCT evidence about comparative effectiveness for myocardial infarction

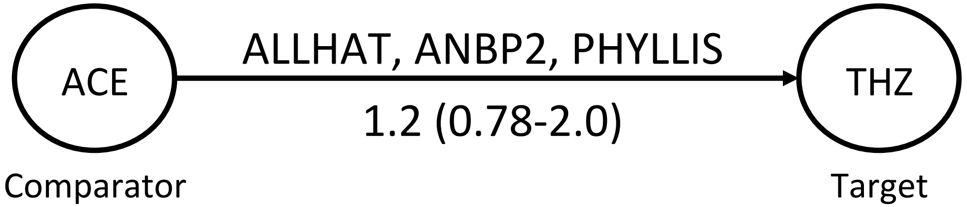
Figure 3.3 Network of clinical trials of antihypertensive drug classes in which myocardial infarction was reported (N=29). *





Reboussin et al., Hypertension 2018

Dissecting the comparative evidence of ACE vs THZ on AMI



ACE

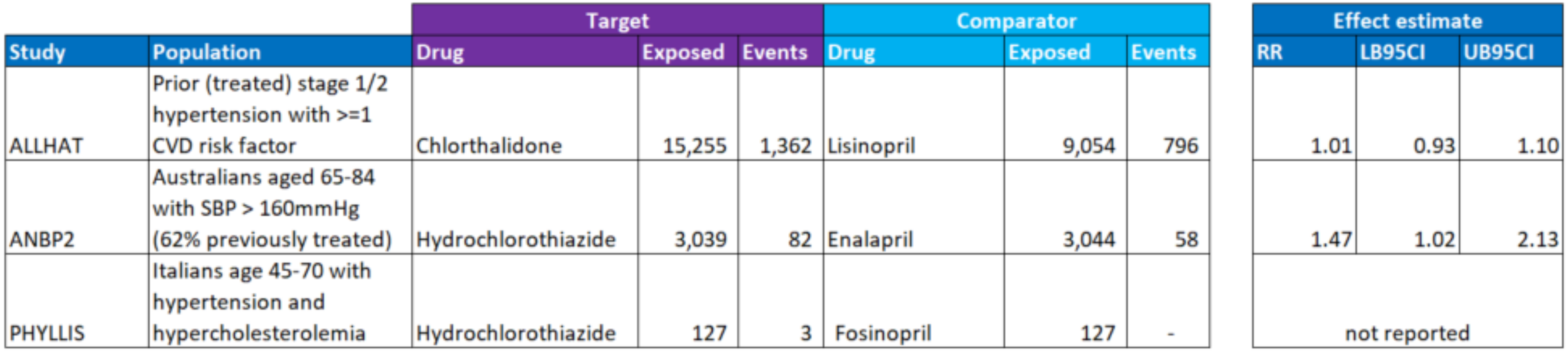
Comparator

ALLHAT, ANBP2, PHYLLIS

1.2 (0.78-2.0)

THZ

Target



		Target			Comparator			Effect estimate		
Study	Population	Drug	Exposed	Events	Drug	Exposed	Events	RR	LB95CI	UB95CI
ALLHAT	Prior (treated) stage 1/2 hypertension with >=1 CVD risk factor	Chlorthalidone	15,255	1,362	Lisinopril	9,054	796	1.01	0.93	1.10
ANBP2	Australians aged 65-84 with SBP > 160mmHg (62% previously treated)	Hydrochlorothiazide	3,039	82	Enalapril	3,044	58	1.47	1.02	2.13
PHYLLIS	Italians age 45-70 with hypertension and hypercholesterolemia	Hydrochlorothiazide	127	3	Fosinopril	127	-	not reported		

New question: How 'consistent' is the current RCT evidence about the effects of antihypertensive medications?

What would the 'target trial' look like to compare efficacy of two initial therapies?

Treatment strategies:

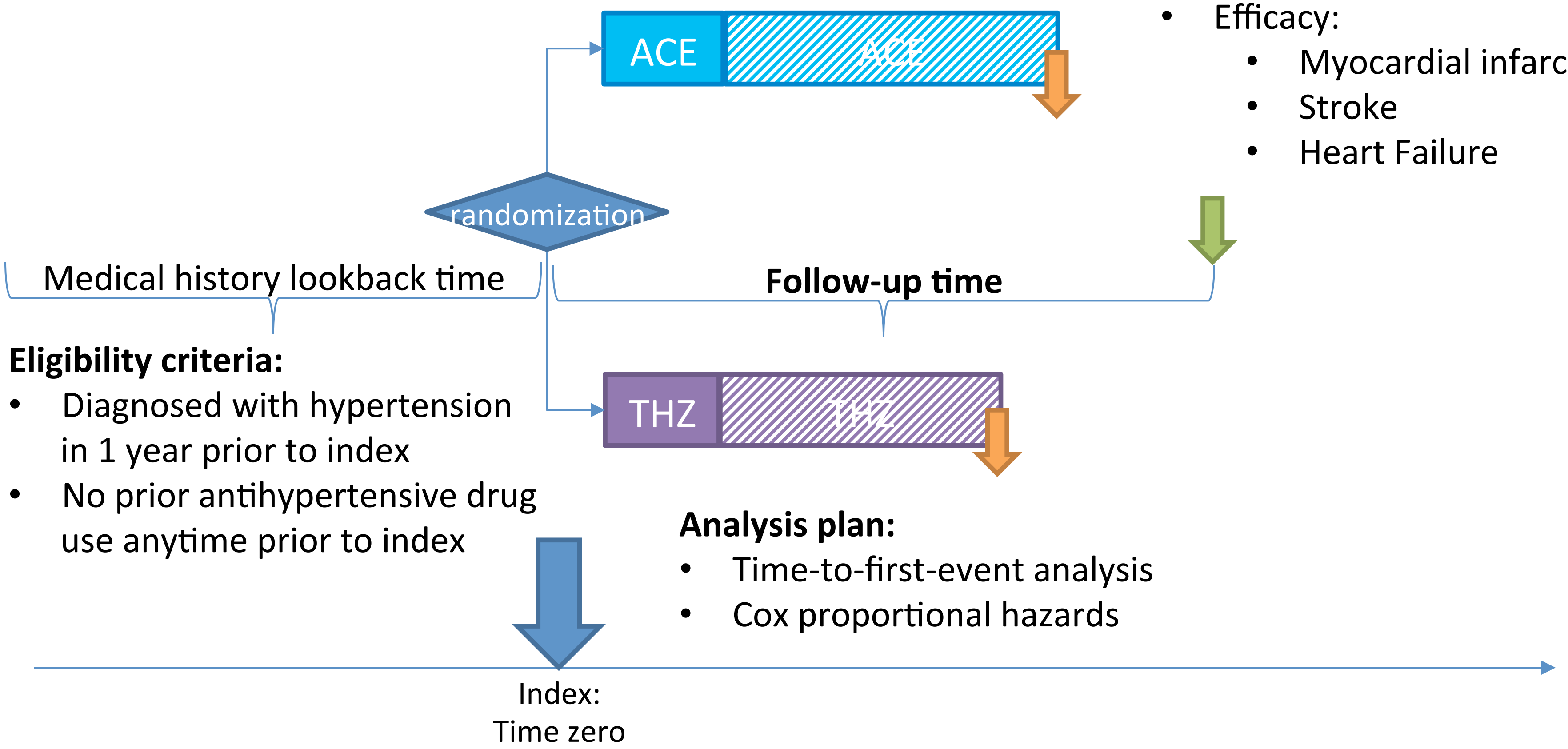
- Monotherapy with ACE
- Monotherapy with THZ

Causal contrasts of interest:

- Intent-to-treat effect
- On-treatment effect

Outcomes:

- Efficacy:
  - Myocardial infarction
  - Stroke
  - Heart Failure



Eligibility criteria:

- Diagnosed with hypertension in 1 year prior to index
- No prior antihypertensive drug use anytime prior to index

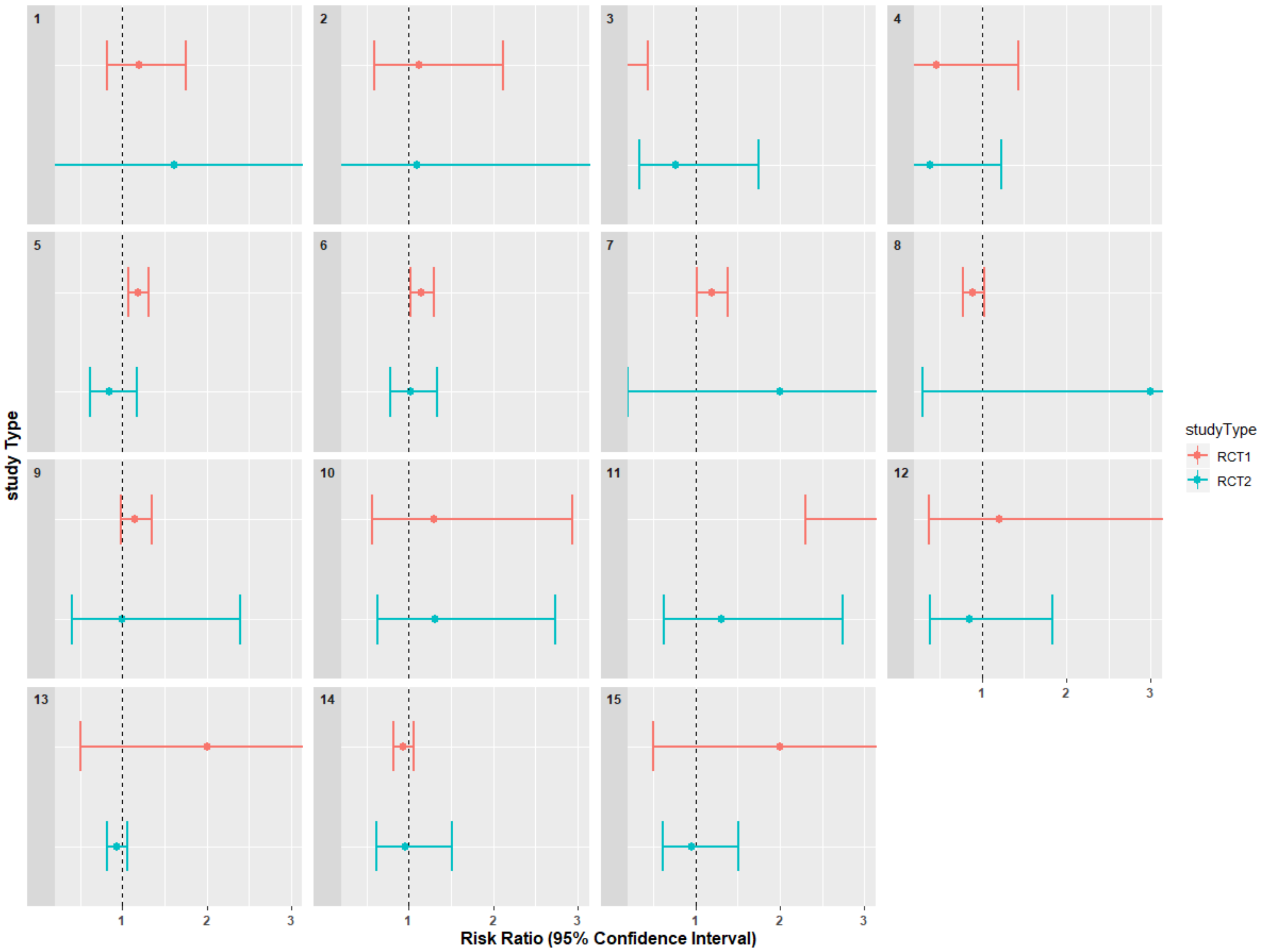
Analysis plan:

- Time-to-first-event analysis
- Cox proportional hazards

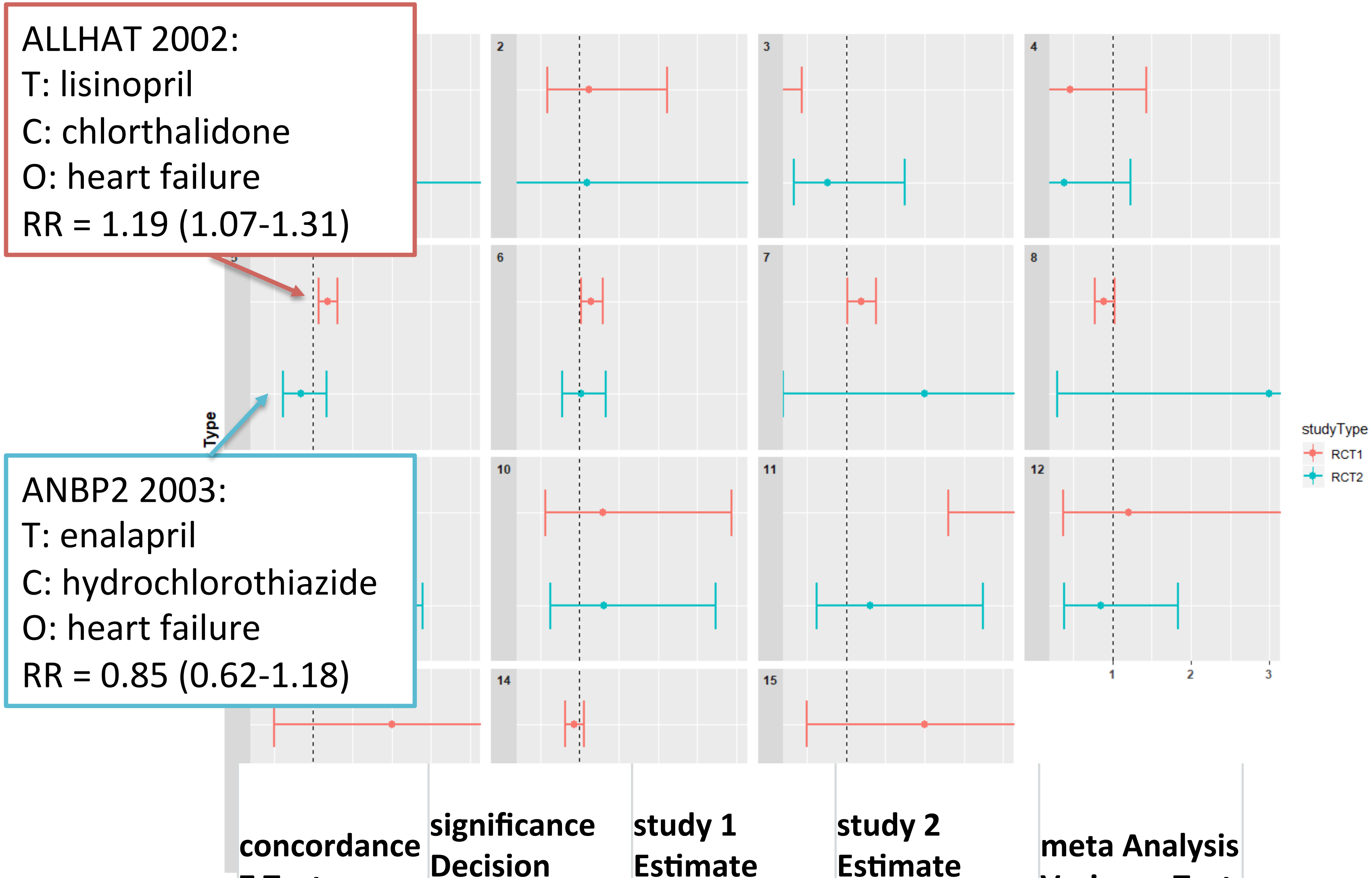
Only 15 randomized trial replications of the same class-class comparison

Outcome	Study 1 Target	Study 1 Comparator	Study 1 Name	Study 2 Target	Study 2 Comparator	Study 2 Name
Acute MI	captopril	atenolol	UKPDS 39	lisinopril	atenolol	Agarwal
Stroke	captopril	atenolol	UKPDS 39	lisinopril	atenolol	Agarwal
Acute MI	enalapril	nisoldipine	ABCD	fosinopril	amlodipine	FACET
Stroke	enalapril	nisoldipine	ABCD	fosinopril	amlodipine	FACET
Heart failure	lisinopril	chlorthalidone	ALLHAT	enalapril	hydrochlorothiazide	ANBP2
Stroke	lisinopril	chlorthalidone	ALLHAT	enalapril	hydrochlorothiazide	ANBP2
Acute MI	valsartan	amlodipine	VALUE	valsartan	amlodipine	VART
Heart failure	valsartan	amlodipine	VALUE	valsartan	amlodipine	VART
Stroke	valsartan	amlodipine	VALUE	valsartan	amlodipine	VART
Acute MI	amlodipine	fosinopril	FACET	nifedipine	lisinopril	JMIC-B
Acute MI	nisoldipine	enalapril	ABCD	nifedipine	lisinopril	JMIC-B
Acute MI	isradipine	hydrochlorothiazide	MIDAS	lacidipine	chlorthalidone	SHELL
Stroke	isradipine	hydrochlorothiazide	MIDAS	amlodipine	chlorthalidone	ALLHAT
Stroke	amlodipine	chlorthalidone	ALLHAT	lacidipine	chlorthalidone	SHELL
Stroke	isradipine	hydrochlorothiazide	MIDAS	lacidipine	chlorthalidone	SHELL

RCT-RCT estimate comparisons



RCT-RCT estimate comparisons

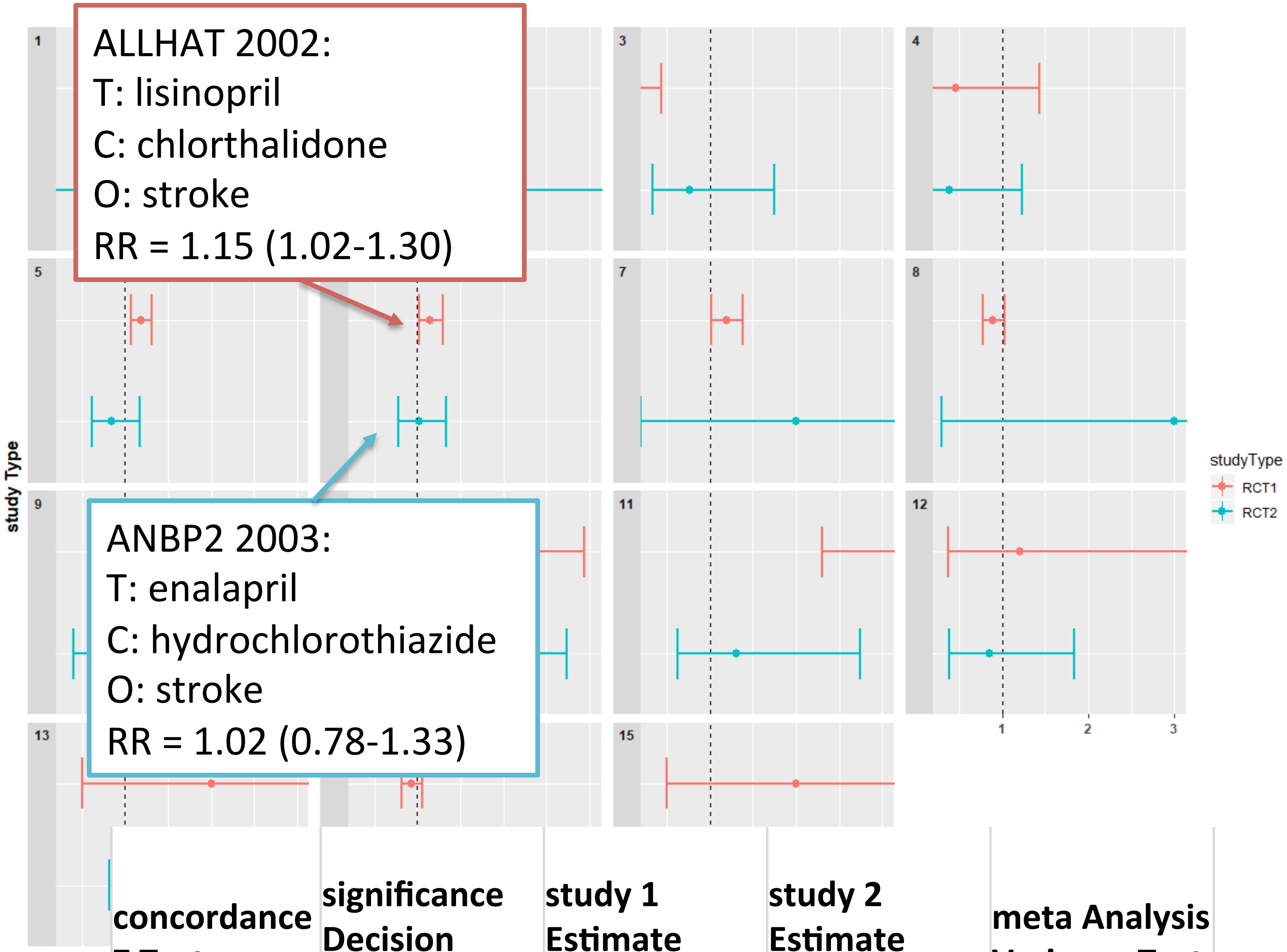


ALLHAT 2002:
T: lisinopril
C: chlorthalidone
O: heart failure
RR = 1.19 (1.07-1.31)

ANBP2 2003:
T: enalapril
C: hydrochlorothiazide
O: heart failure
RR = 0.85 (0.62-1.18)

concordance Z Test	significance Decision Agreement	study 1 Estimate Agreement	study 2 Estimate Agreement	meta Analysis Variance Test
PASS	FAIL	FAIL	FAIL	FAIL

RCT-RCT estimate comparisons

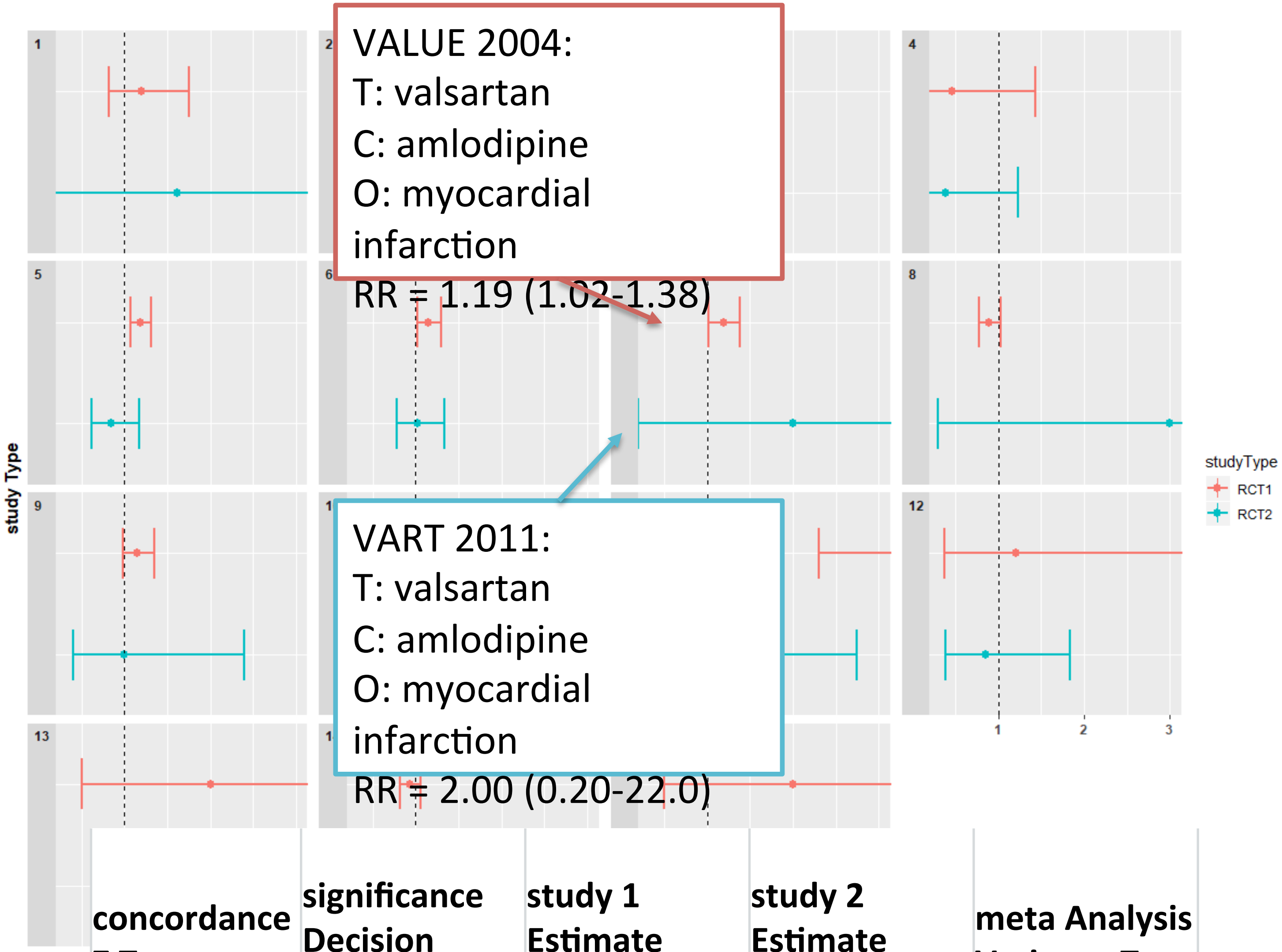


ALLHAT 2002:
T: lisinopril
C: chlorthalidone
O: stroke
RR = 1.15 (1.02-1.30)

ANBP2 2003:
T: enalapril
C: hydrochlorothiazide
O: stroke
RR = 1.02 (0.78-1.33)

concordance Z Test	significance Decision Agreement	study 1 Estimate Agreement	study 2 Estimate Agreement	meta Analysis Variance Test
PASS	FAIL	PASS	PASS	PASS

RCT-RCT estimate comparisons



VALUE 2004:
T: valsartan
C: amlodipine
O: myocardial infarction
RR = 1.19 (1.02-1.38)

VART 2011:
T: valsartan
C: amlodipine
O: myocardial infarction
RR = 2.00 (0.20-22.0)

concordance Z Test	significance Decision Agreement	study 1 Estimate Agreement	study 2 Estimate Agreement	meta Analysis Variance Test
PASS	FAIL	PASS	FAIL	PASS

How well do RCTs ‘replicate’ each other, according to our metrics?

Study Pairs	concordance Z Test	significance Decision Agreement	study 1 Estimate Agreement	study 2 Estimate Agreement	meta Analysis Variance Test
RCT-RCT	87%	67%	67%	67%	73%

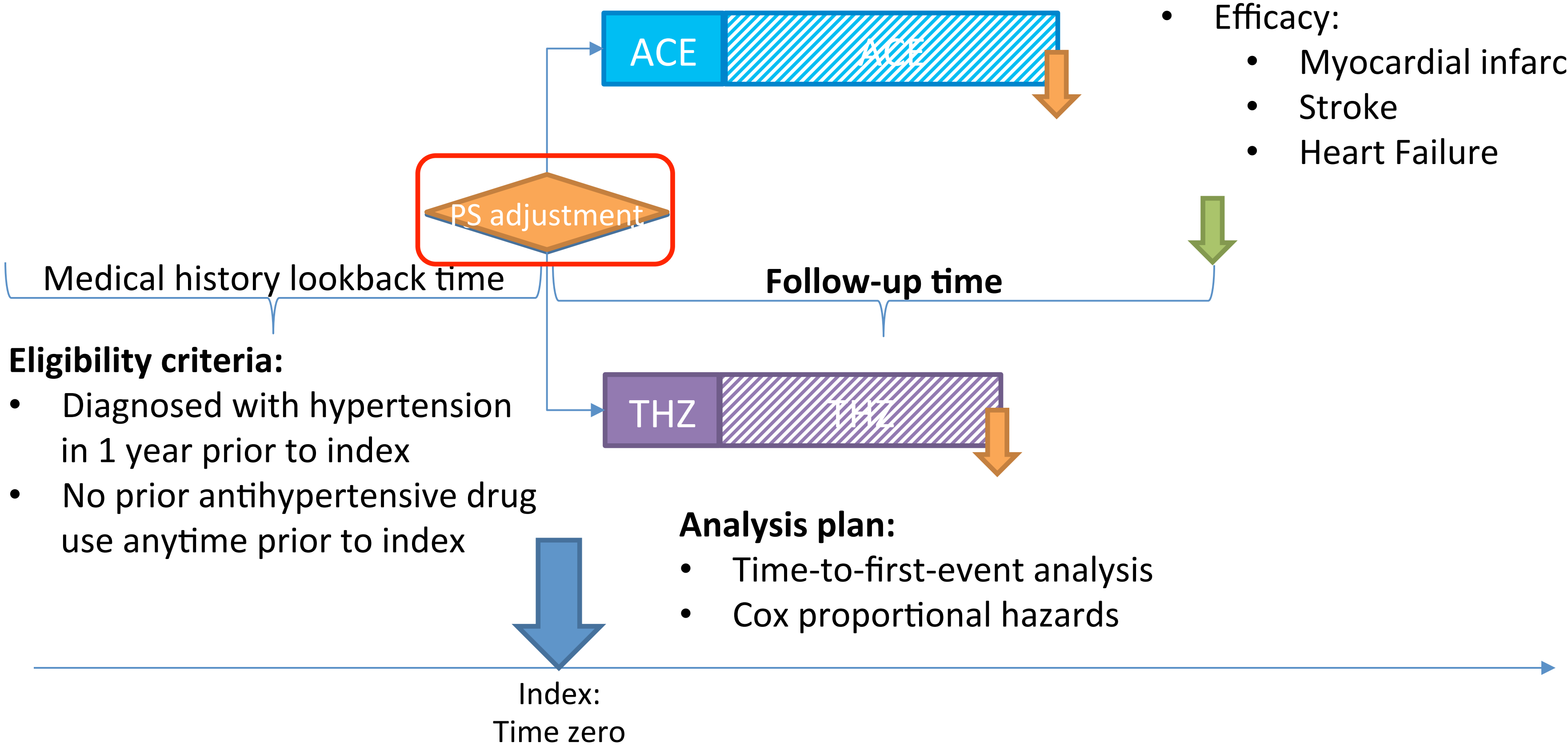
Insight: if we consider RCTs our gold standard, then we shouldn’t expect ‘perfect’ agreement under any of our evaluation metrics when considering the consistency of RWE to RCTs

What does LEGEND look like to compare effectiveness of two initial therapies?

Treatment strategies:
- Monotherapy with ACE
- Monotherapy with THZ

Causal contrasts of interest:
- Intent-to-treat effect
- On-treatment effect

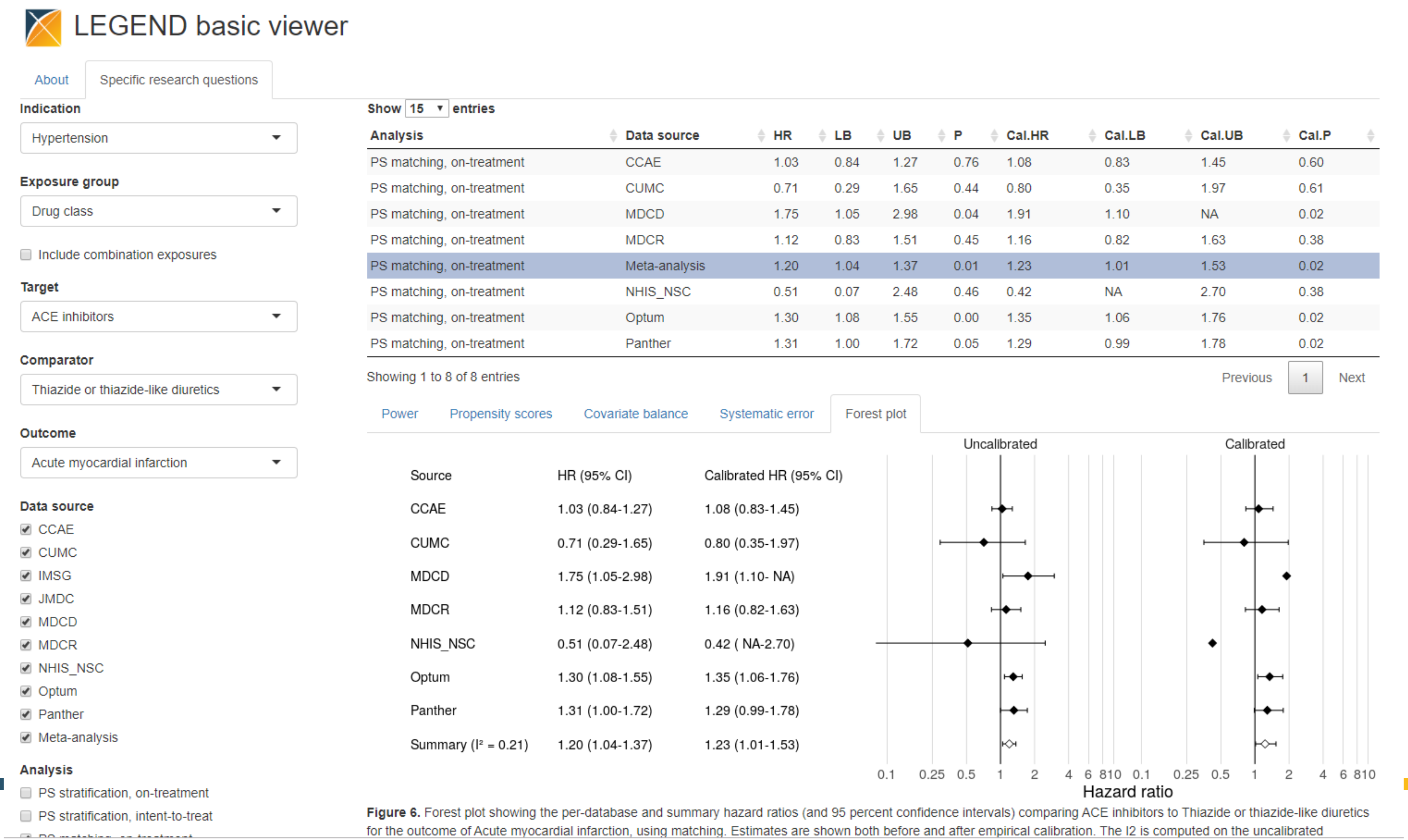
Outcomes:
- Efficacy:
  - Myocardial infarction
  - Stroke
  - Heart Failure



Eligibility criteria:
- Diagnosed with hypertension in 1 year prior to index
- No prior antihypertensive drug use anytime prior to index

Analysis plan:
- Time-to-first-event analysis
- Cox proportional hazards

LEGEND results publicly available at data.ohdsi.org



LEGEND basic viewer

About | Specific research questions

Indication
Hypertension

Exposure group
Drug class

□ Include combination exposures

Target
ACE inhibitors

Comparator
Thiazide or thiazide-like diuretics

Outcome
Acute myocardial infarction

Data source
☑ CCAE
☑ CUMC
☑ IMSG
☑ JMDC
☑ MDCD
☑ MDCR
☑ NHIS_NSC
☑ Optum
☑ Panther
☑ Meta-analysis

Analysis
□ PS stratification, on-treatment
□ PS stratification, intent-to-treat
□ PS matching, on-treatment

Show 15 entries

Analysis	Data source	HR	LB	UB	P	Cal.HR	Cal.LB	Cal.UB	Cal.P
PS matching, on-treatment	CCAE	1.03	0.84	1.27	0.76	1.08	0.83	1.45	0.60
PS matching, on-treatment	CUMC	0.71	0.29	1.65	0.44	0.80	0.35	1.97	0.61
PS matching, on-treatment	MDCD	1.75	1.05	2.98	0.04	1.91	1.10	NA	0.02
PS matching, on-treatment	MDCR	1.12	0.83	1.51	0.45	1.16	0.82	1.63	0.38
PS matching, on-treatment	Meta-analysis	1.20	1.04	1.37	0.01	1.23	1.01	1.53	0.02
PS matching, on-treatment	NHIS_NSC	0.51	0.07	2.48	0.46	0.42	NA	2.70	0.38
PS matching, on-treatment	Optum	1.30	1.08	1.55	0.00	1.35	1.06	1.76	0.02
PS matching, on-treatment	Panther	1.31	1.00	1.72	0.05	1.29	0.99	1.78	0.02

Showing 1 to 8 of 8 entries

Previous 1 Next

Power | Propensity scores | Covariate balance | Systematic error | Forest plot

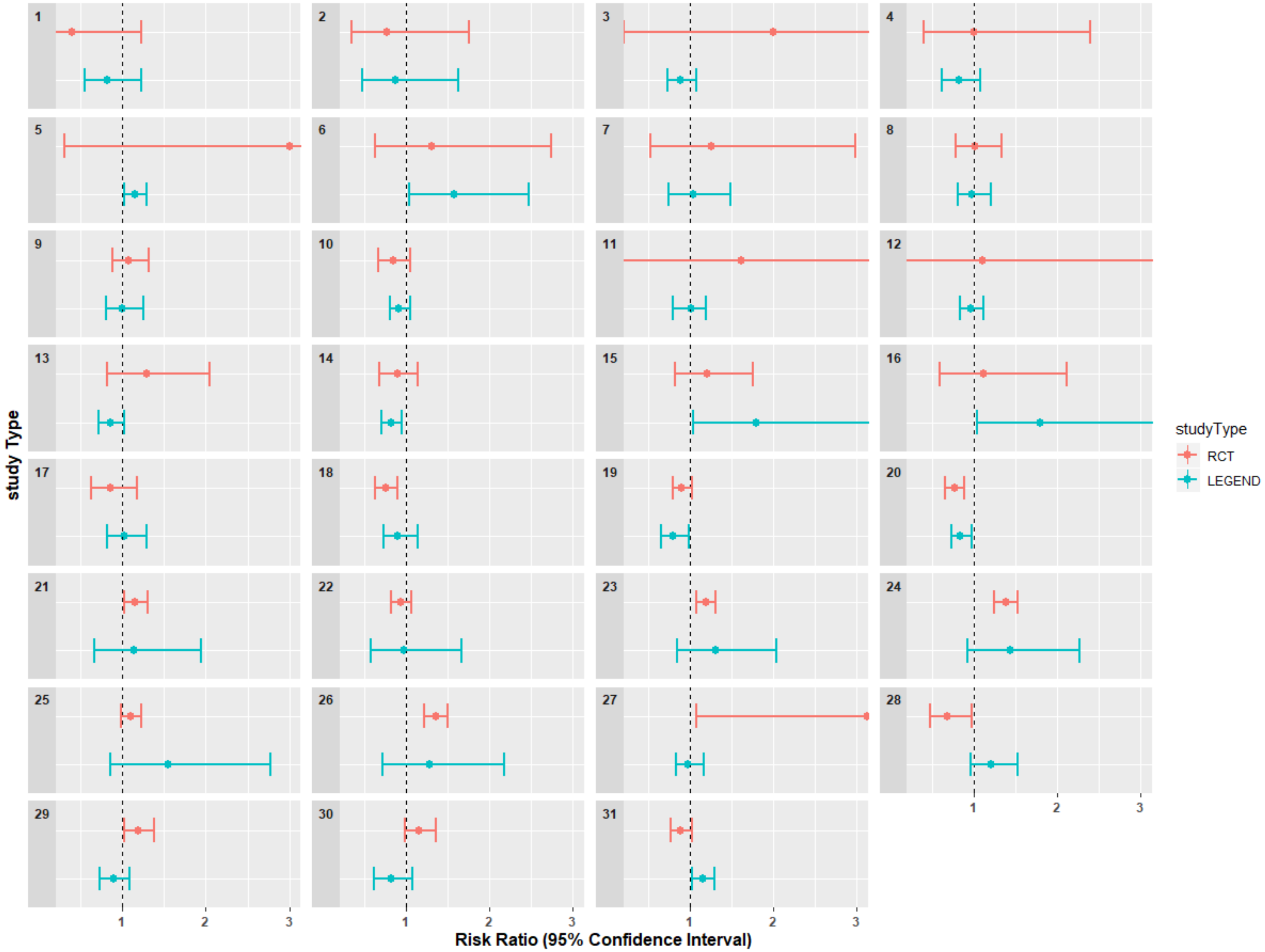
Source	HR (95% CI)	Calibrated HR (95% CI)
CCAE	1.03 (0.84-1.27)	1.08 (0.83-1.45)
CUMC	0.71 (0.29-1.65)	0.80 (0.35-1.97)
MDCD	1.75 (1.05-2.98)	1.91 (1.10- NA)
MDCR	1.12 (0.83-1.51)	1.16 (0.82-1.63)
NHIS_NSC	0.51 (0.07-2.48)	0.42 (NA-2.70)
Optum	1.30 (1.08-1.55)	1.35 (1.06-1.76)
Panther	1.31 (1.00-1.72)	1.29 (0.99-1.78)
Summary (I² = 0.21)	1.20 (1.04-1.37)	1.23 (1.01-1.53)

Figure 6. Forest plot showing the per-database and summary hazard ratios (and 95 percent confidence intervals) comparing ACE inhibitors to Thiazide or thiazide-like diuretics for the outcome of Acute myocardial infarction, using matching. Estimates are shown both before and after empirical calibration. The I2 is computed on the uncalibrated

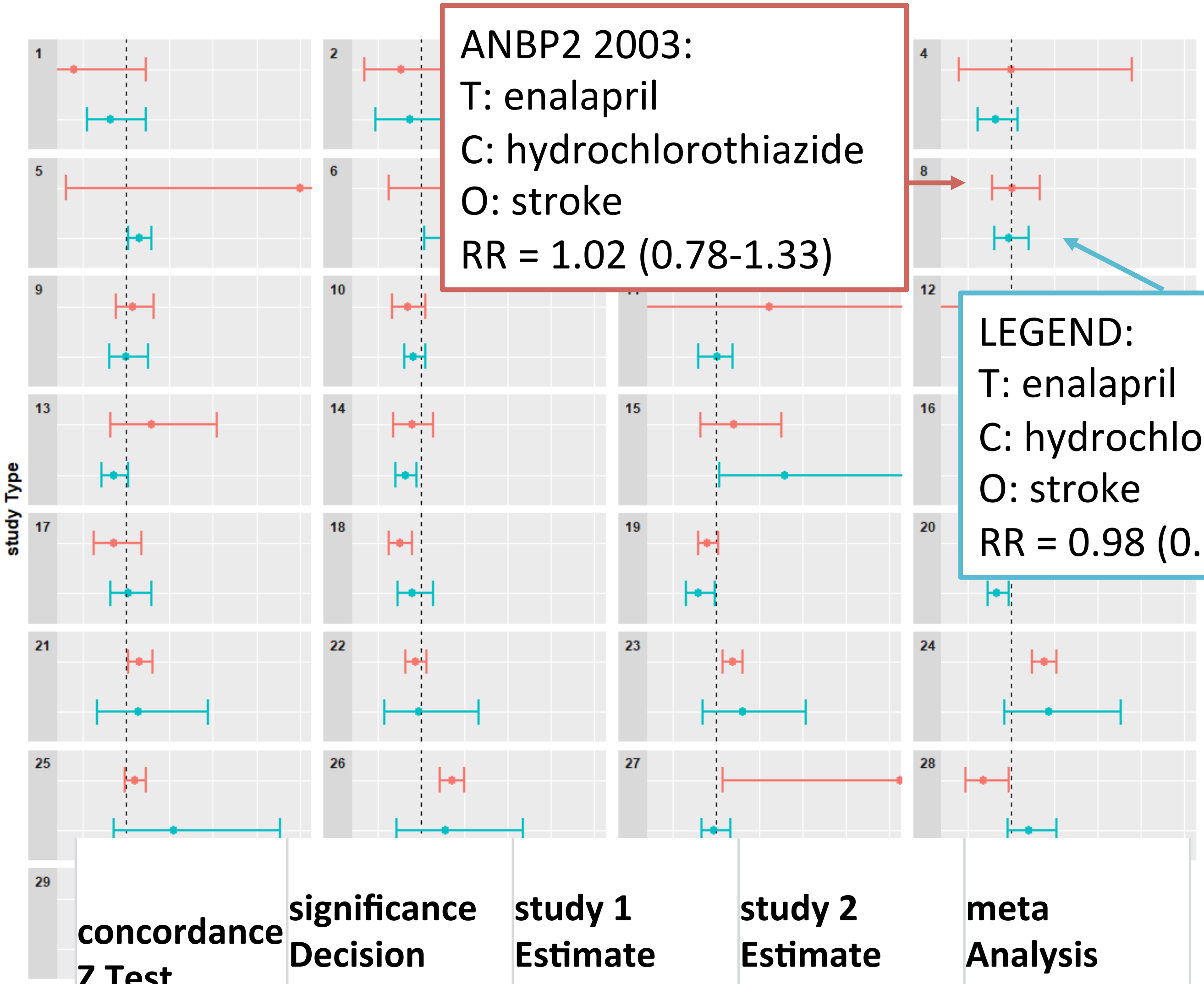
31 randomized trial results can be directly compared with LEGEND

RCT name	Target	Comparator	Outcome
FACET	fosinopril	amlodipine	Stroke
FACET	fosinopril	amlodipine	Acute MI
VART	valsartan	amlodipine	Acute MI
VART	valsartan	amlodipine	Stroke
VART	valsartan	amlodipine	Heart failure
JMIC-B	nifedipine	lisinopril	Acute MI
JMIC-B	nifedipine	lisinopril	Hosp for heart failure
ANBP2	enalapril	hydrochlorothiazide	Stroke
LIFE	losartan	atenolol	Acute MI
ASCOT-BPLA	amlodipine	atenolol	Heart failure
Agarwal	lisinopril	atenolol	Acute MI
Agarwal	lisinopril	atenolol	Stroke
HAPPHY	hydrochlorothiazide	atenolol	Stroke
HAPPHY	hydrochlorothiazide	atenolol	Acute MI
UKPDS 39	captopril	atenolol	Acute MI
UKPDS 39	captopril	atenolol	Stroke
ANBP2	enalapril	hydrochlorothiazide	Heart failure
LIFE	losartan	atenolol	Stroke
ASCOT-BPLA	amlodipine	atenolol	Acute MI
ASCOT-BPLA	amlodipine	atenolol	Stroke
ALLHAT	lisinopril	chlorthalidone	Stroke
ALLHAT	amlodipine	chlorthalidone	Stroke
ALLHAT	lisinopril	chlorthalidone	Heart failure
ALLHAT	amlodipine	chlorthalidone	Heart failure
ALLHAT	lisinopril	chlorthalidone	Hosp for heart failure
ALLHAT	amlodipine	chlorthalidone	Hosp for heart failure
Agarwal	lisinopril	atenolol	Heart failure
ANBP2	enalapril	hydrochlorothiazide	Acute MI
VALUE	valsartan	amlodipine	Acute MI
VALUE	valsartan	amlodipine	Stroke
VALUE	valsartan	amlodipine	Heart failure

RCT-LEGEND estimate comparisons



RCT-LEGEND estimate comparisons

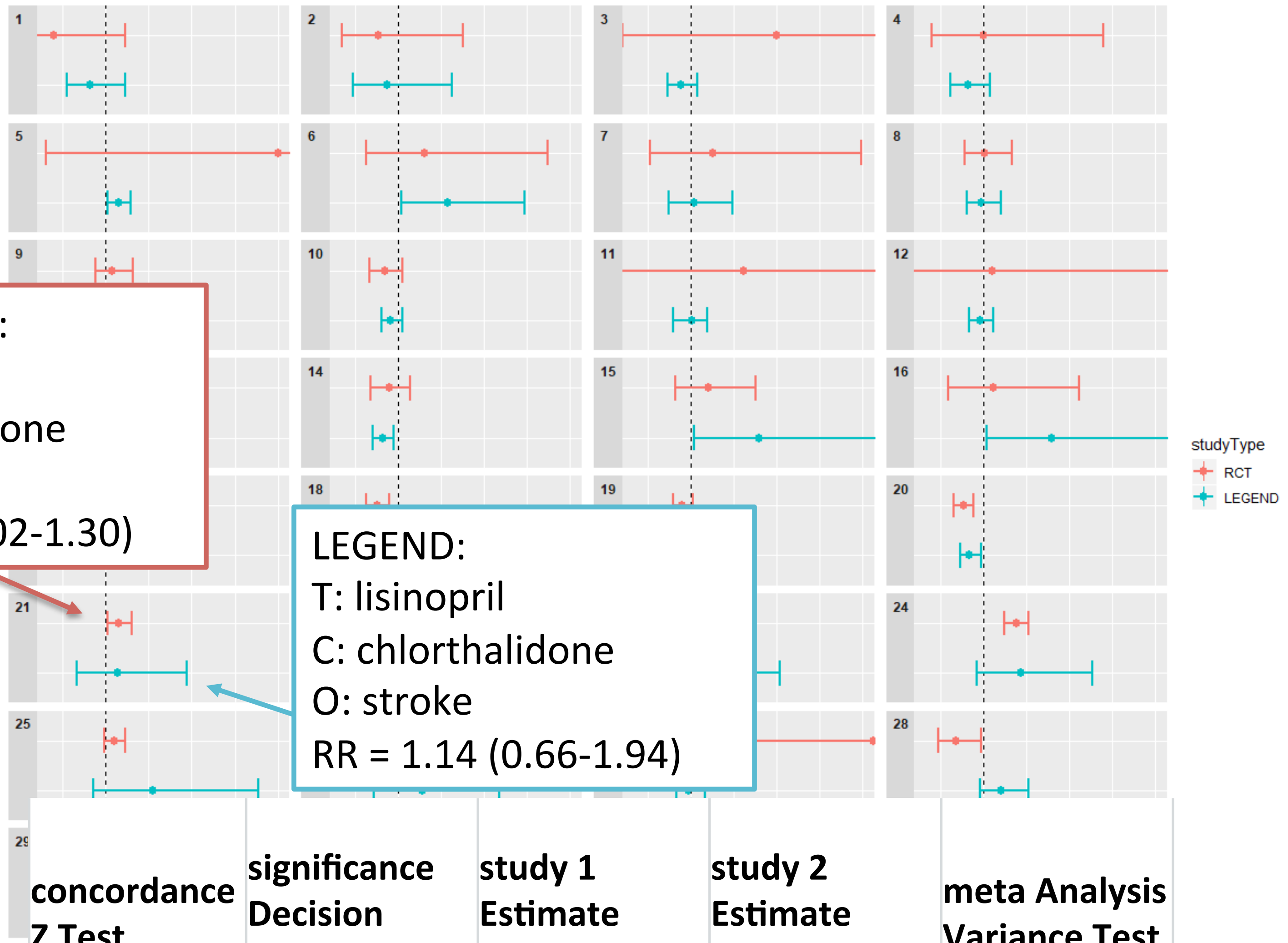


ANBP2 2003:
T: enalapril
C: hydrochlorothiazide
O: stroke
RR = 1.02 (0.78-1.33)

LEGEND:
T: enalapril
C: hydrochlorothiazide
O: stroke
RR = 0.98 (0.81-1.21)

concordance Z Test	significance Decision Agreement	study 1 Estimate Agreement	study 2 Estimate Agreement	meta Analysis Variance Test
PASS	PASS	PASS	PASS	PASS

RCT-LEGEND estimate comparisons

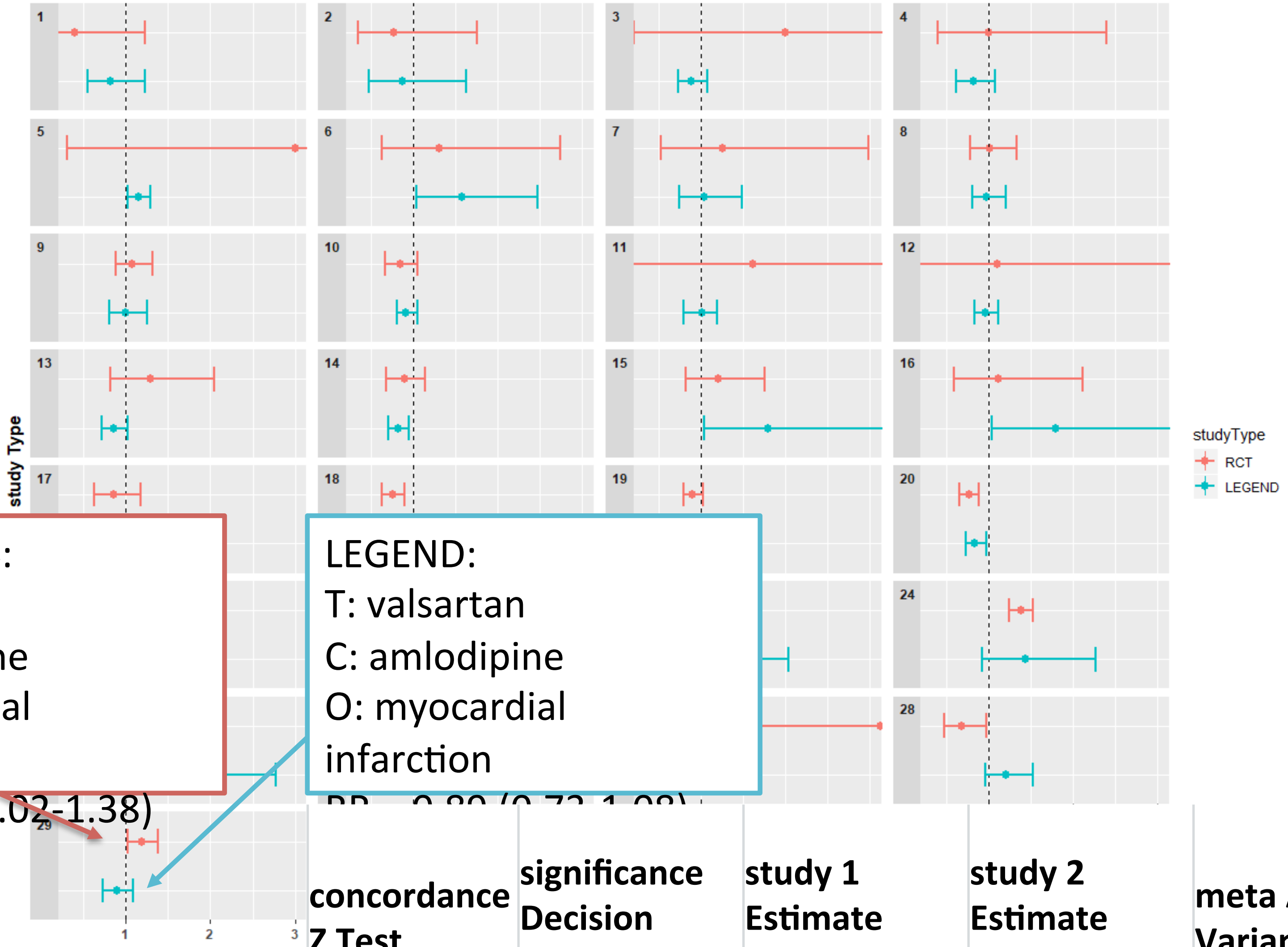


ALLHAT 2002:
T: lisinopril
C: chlorthalidone
O: stroke
RR = 1.15 (1.02-1.30)

LEGEND:
T: lisinopril
C: chlorthalidone
O: stroke
RR = 1.14 (0.66-1.94)

concordance Z Test	significance Decision Agreement	study 1 Estimate Agreement	study 2 Estimate Agreement	meta Analysis Variance Test
PASS	FAIL	PASS	PASS	PASS

RCT-LEGEND estimate comparisons



VALUE 2004:
T: valsartan
C: amlodipine
O: myocardial infarction
RR = 1.19 (1.02-1.38)

LEGEND:
T: valsartan
C: amlodipine
O: myocardial infarction
RR = 0.89 (0.72-1.08)

concordance Z Test	significance Decision Agreement	study 1 Estimate Agreement	study 2 Estimate Agreement	meta Analysis Variance Test
FAIL	FAIL	FAIL	FAIL	FAIL

How well does LEGEND replicate RCTs?

Study Pairs	concordance Z Test	significance Decision Agreement	study 1 Estimate Agreement	study 2 Estimate Agreement	meta Analysis Variance Test
RCT-RCT	87%	67%	67%	67%	73%
RCT-LEGEND	**87%**	**52%**	**68%**	**74%**	**81%**

Across all metrics, evidence from LEGEND is concordant with RCTs to the same extent that RCTs are concordant with each other

What if the first study was exactly repeated (ex: same protocol, same sites, same time) with other subjects drawn from the same original population…

what do our metrics show in this "perfect replication"?

- We can create that result statistically through simulation using the RCT and LEGEND results

What was the expected performance for a “perfect replication”?

Study Pairs	concordance Z Test	significance Decision Agreement	study 1 Estimate Agreement	study 2 Estimate Agreement	meta Analysis Variance Test
RCT-RCT	87%	67%	67%	67%	73%
RCT-LEGEND	**87%**	**52%**	**68%**	**74%**	**81%**
“Perfect replication”	95%	55%	68%	77%	82%

Across all metrics, evidence from LEGEND is concordant with RCTs to the same extent as would be expected in a “perfect replication”

What if the first study was exactly repeated (ex: same protocol, same sites, same time) but the second study had some defined bias …
what do our metrics show in this "biased replication"?

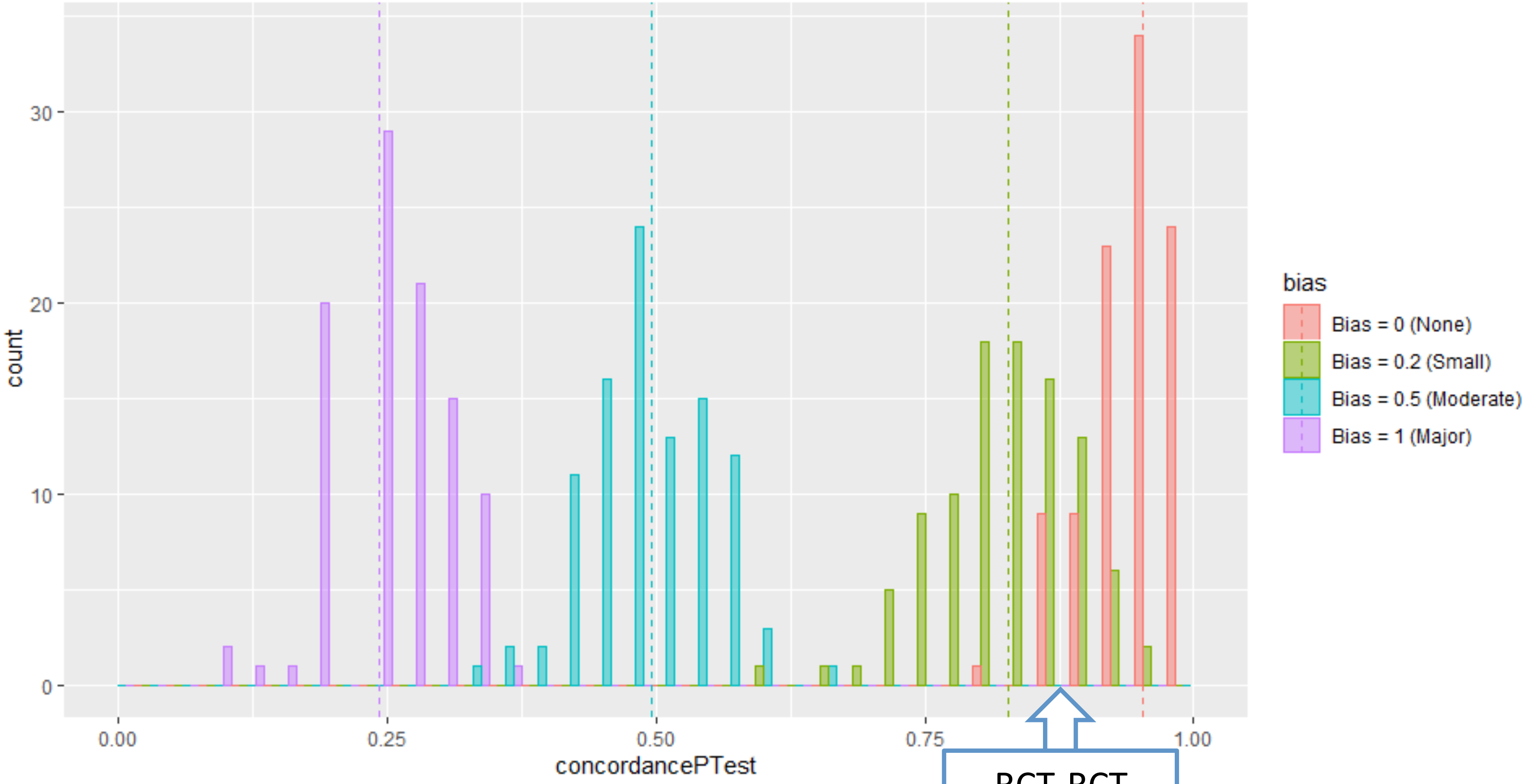
- We can create that result statistically through simulation using the RCT and LEGEND results

What was the expected performance for a “biased replication”?

Study Pairs	concordance Z Test	significance Decision Agreement	study 1 Estimate Agreement	study 2 Estimate Agreement	meta Analysis Variance Test
RCT-RCT	87%	67%	67%	67%	73%
RCT-LEGEND	**87%**	**52%**	**68%**	**74%**	**81%**
“Perfect replication”	95%	55%	68%	77%	82%
“Biased replication” = 0.2	83%	45%	51%	61%	66%
“Biased replication” = 0.5	50%	30%	24%	35%	37%
“Biased replication” = 1.0	24%	23%	5%	21%	22%

- Bias makes all agreement metrics worse
- RCTs agree with each other with performance between “perfect” and “minimal bias”
- LEGEND performs similarly

Expected performance is a distribution based on the number of study comparisons



RCT-RCT
RCT-LEGEND

What if one just guessed RR=1 for all studies?

Study Pairs	concordance Z Test	significance Decision Agreement	study 1 Estimate Agreement	study 2 Estimate Agreement	meta Analysis Variance Test
RCT-RCT	87%	67%	67%	67%	73%
RCT-LEGEND	**87%**	**52%**	**68%**	**74%**	**81%**
“Perfect replication”	95%	55%	68%	77%	82%
“Biased replication” = 0.2	83%	45%	51%	61%	66%
“Biased replication” = 0.5	50%	30%	24%	35%	37%
“Biased replication” = 1.0	24%	23%	5%	21%	22%
RCT- RF	[illegible]	[illegible]	[illegible]	[illegible]	90%

- All metrics are sensitive to the distribution of the initial RCT estimates and to the variance of the replication study.
- We need to be careful that we don’t incentivize unpowered studies or conflate non-significant effects with evidence of no effect.

Lesson to inform real-world evidence for regulatory decision-making

Real-world evidence from LEGEND is as consistent with RCTs as RCTs are with each other, according to any agreement metrics

Methodological lessons about evaluating consistency between studies

- It is unreasonable to expect any set of studies to achieve ‘perfect’ replication using any of the published metrics
- 31 RCT-RWE replications is not enough to make a definitive conclusion, but what precision is needed for regulatory acceptance?
- Sample size of each study matters when establishing expected performance
- Prior knowledge of the studies to be replicated can be used to game the evaluation

Replicating past studies is a means to and end, not an end in itself: our goal is to determine how we can make real-world evidence reliable enough to be used as “adequate and well-controlled investigations” and “confirmatory evidence”

‘Adequate and well-controlled investigation’ criteria	Threat to validity	OHDSI RWE solution
There is a clear statement of the objectives of the investigation and a summary of the methods of analysis in the protocol for the study.	Investigator bias	Fully and pre-specified protocol and source code made publicly available prior to study conduct. BoO: Study steps, Network research
The study uses a design that permits a valid comparison with a control to provide a quantitative assessment of drug effect.	Selection bias	Study design choice (comparative cohort vs. self-controlled). Study diagnostics: Propensity score overlap, covariate balance before adjustment. BoO: Population-level effect estimation
The method of selection of subjects provides adequate assurance that they have the disease or condition being studied.	Measurement error	Phenotype evaluation of indication. Generalizability assessment to target population. BoO: Defining cohorts, Characterization, Clinical Validity
The method of assigning patients to treatment and control groups minimizes bias and is intended to assure comparability of the groups.	Confounding	Study diagnostics: propensity score overlap, covariate balance, negative control calibration BoO: Population-level effect estimation
Adequate measures are taken to minimize bias on the part of the subjects, observers, and analysts of the data.	Selection bias	Study diagnostics: covariate balance after adjustment, negative control calibration. BoO: Population-level effect estimation
The methods of assessment of subjects' response are well-defined and reliable.	Measurement error	Phenotype evaluation of outcome. BoO: Data quality, Clinical validity
There is an analysis of the results of the study adequate to assess the effects of the drug. The report of the study should describe the results and the analytic methods used to evaluate them, including any appropriate statistical methods.	Model misspecification	Study diagnostics: negative control calibration Pre-specification BoO: Methods validity, Software validity

https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfcfr/cfrsearch.cfm?fr=314.126

Recall: RWE is different from RCTs…

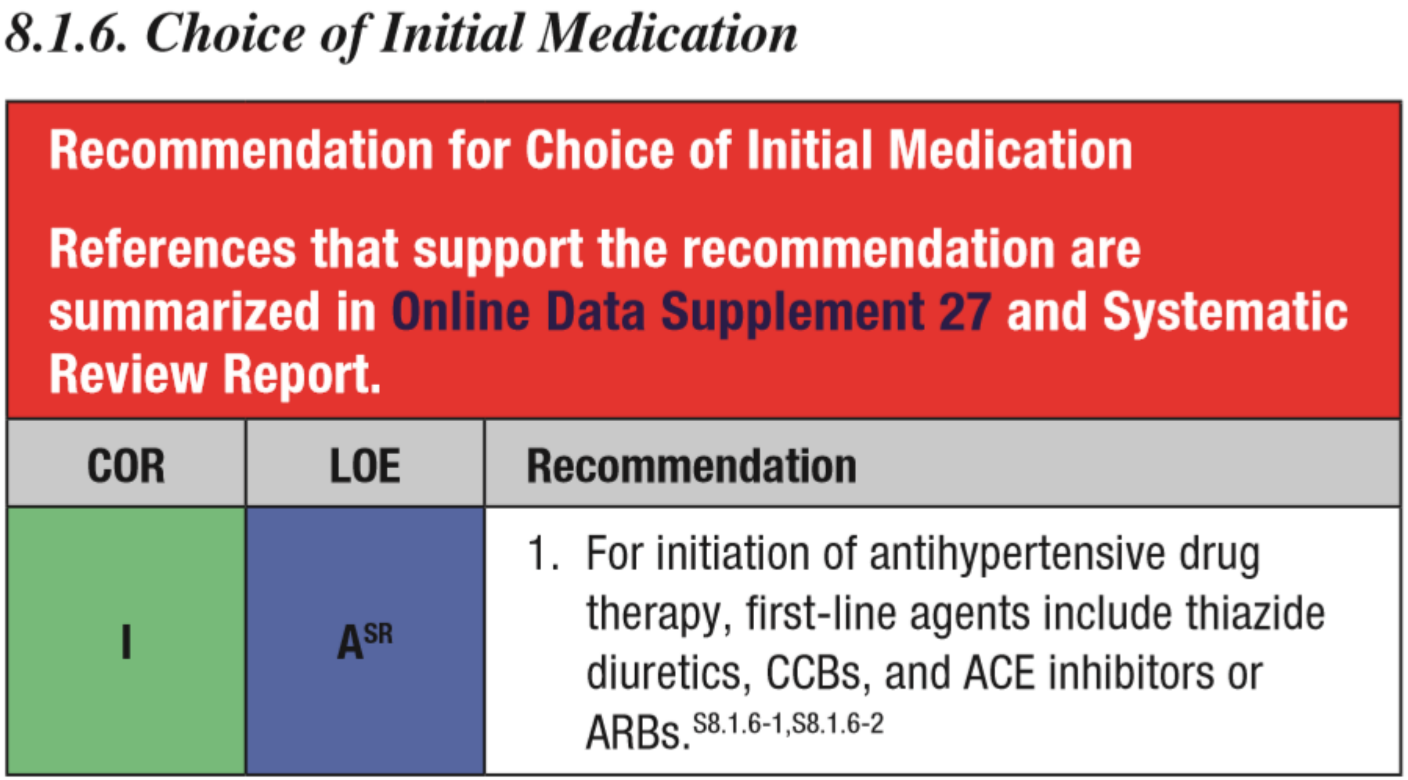
- Greater sample size
- Real world practice: effectiveness vs. efficacy
- Generalizable populations

Clinical scenarios where RWE can contribute to current evidence for medical decision-making from LEGEND

- Resolving uncertainty from RCTs can uncover significant differences between treatments
  - ACE vs. THZ
- Resolving uncertainty from RCTs can increase comfort by bounding the potential effect size
  - ACE vs. ARB
- Real-world evidence fills gaps where no RCTs exist
  - Chlorthalidone vs. hydrochlorothiazide
  - Mono vs. combination therapy

Clinical scenarios where RWE can contribute to current evidence for medical decision-making from LEGEND

- Resolving uncertainty from RCTs can uncover significant differences between treatments
  - **ACE vs. THZ**



8.1.6. Choice of Initial Medication

Recommendation for Choice of Initial Medication

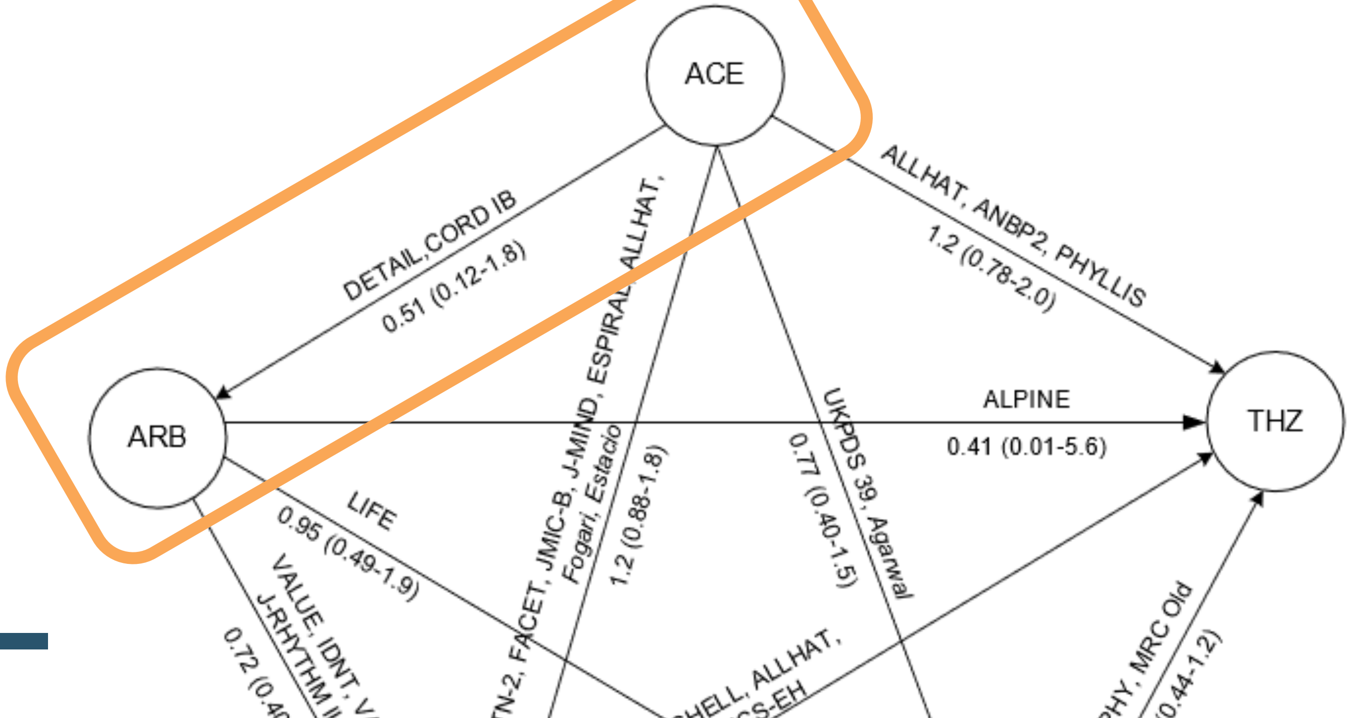
References that support the recommendation are summarized in Online Data Supplement 27 and Systematic Review Report.

COR	LOE	Recommendation
I	A^{SR}	1. For initiation of antihypertensive drug therapy, first-line agents include thiazide diuretics, CCBs, and ACE inhibitors or ARBs.[S8.1.6-1,S8.1.6-2]

Clinical scenarios where RWE can contribute to current evidence for medical decision-making from LEGEND

- Resolving uncertainty from RCTs can increase comfort by bounding the potential effect size
  - **ACE vs. ARB**

Figure 3.3 Network of clinical trials of antihypertensive drug classes in which myocardial infarction was reported (N=29). *



Clinical scenarios where RWE can contribute to current evidence for medical decision-making from LEGEND

- Real-world evidence fills gaps where no RCTs exist
  - **Chlorthalidone vs. hydrochlorothiazide**

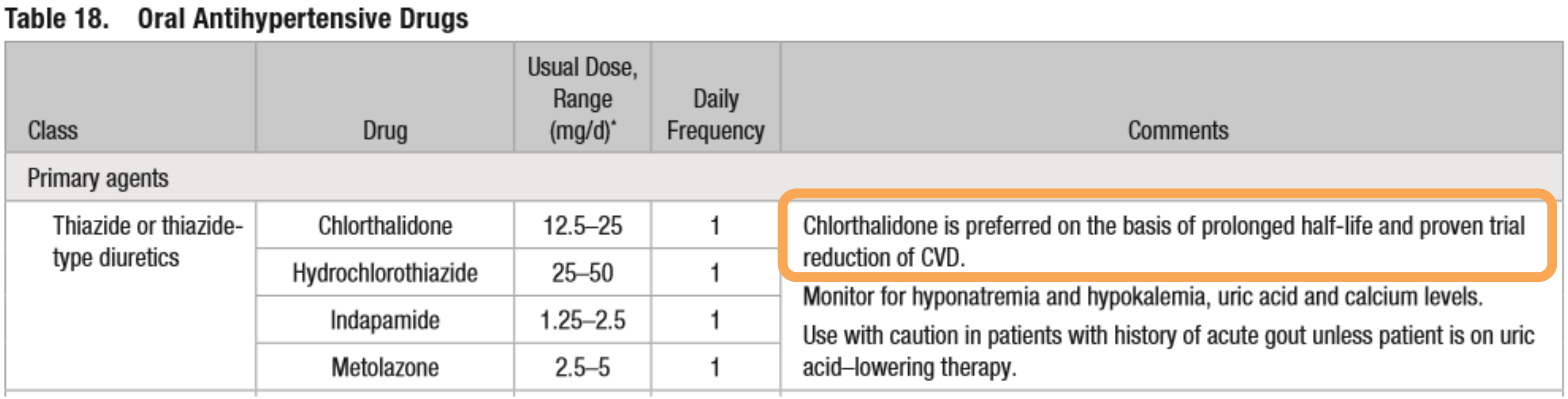
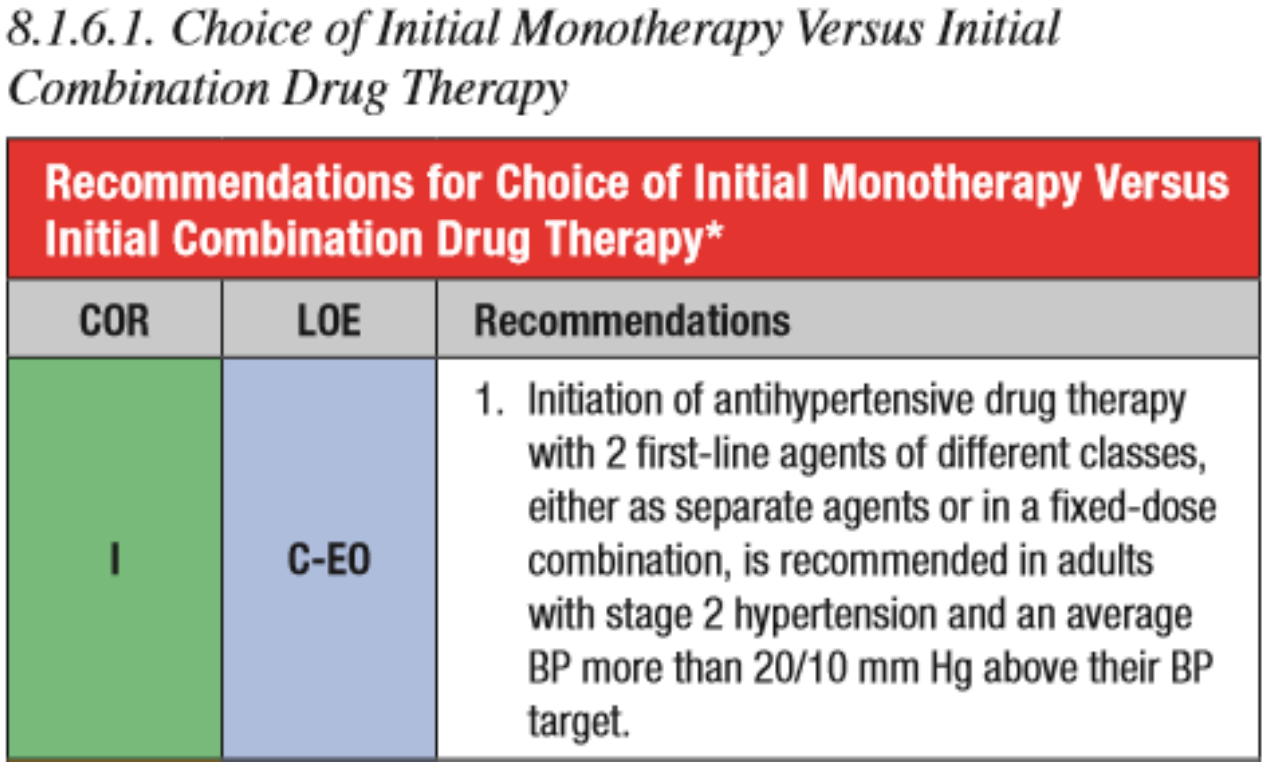


Table 18. Oral Antihypertensive Drugs

Class	Drug	Usual Dose, Range (mg/d)*	Daily Frequency	Comments
Primary agents				
Thiazide or thiazide-type diuretics	Chlorthalidone	12.5–25	1	Chlorthalidone is preferred on the basis of prolonged half-life and proven trial reduction of CVD. Monitor for hyponatremia and hypokalemia, uric acid and calcium levels. Use with caution in patients with history of acute gout unless patient is on uric acid–lowering therapy.
	Hydrochlorothiazide	25–50	1	
	Indapamide	1.25–2.5	1	
	Metolazone	2.5–5	1	

Clinical scenarios where RWE can contribute to current evidence for medical decision-making from LEGEND

- Real-world evidence fills gaps where no RCTs exist
  - **Mono vs. combination therapy**



8.1.6.1. Choice of Initial Monotherapy Versus Initial Combination Drug Therapy

Recommendations for Choice of Initial Monotherapy Versus Initial Combination Drug Therapy*

COR	LOE	Recommendations
I	C-EO	1. Initiation of antihypertensive drug therapy with 2 first-line agents of different classes, either as separate agents or in a fixed-dose combination, is recommended in adults with stage 2 hypertension and an average BP more than 20/10 mm Hg above their BP target.

A journey toward real-world evidence for regulatory decision-making

- Building confidence in ***real-world data***:

  Data quality reporting

- Establishing scientific best practices for ***real-world analysis***:

  Book Of OHDSI

- Proving reliable ***real-world evidence***:

  Replicating RCTs using LEGEND