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
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”
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:
<table>
<thead>
<tr>
<th>‘Adequate and well-controlled investigation’ criteria</th>
<th>Threat to validity</th>
</tr>
</thead>
<tbody>
<tr>
<td>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.</td>
<td>Investigator bias</td>
</tr>
<tr>
<td>The study uses a design that permits a valid comparison with a control to provide a quantitative assessment of drug effect.</td>
<td>Selection bias</td>
</tr>
<tr>
<td>The method of selection of subjects provides adequate assurance that they have the disease or condition being studied.</td>
<td>Measurement error</td>
</tr>
<tr>
<td>The method of assigning patients to treatment and control groups minimizes bias and is intended to assure comparability of the groups.</td>
<td>Confounding</td>
</tr>
<tr>
<td>Adequate measures are taken to minimize bias on the part of the subjects, observers, and analysts of the data.</td>
<td>Selection bias</td>
</tr>
<tr>
<td>The methods of assessment of subjects' response are well-defined and reliable.</td>
<td>Measurement error</td>
</tr>
<tr>
<td>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.</td>
<td>Model misspecification</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>‘Adequate and well-controlled investigation’ criteria</th>
<th>Threat to validity</th>
<th>OHDSI RWE solution</th>
</tr>
</thead>
<tbody>
<tr>
<td>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.</td>
<td>Investigator bias</td>
<td>Fully and pre-specified protocol and source code made publicly available prior to study conduct. BoO: Study steps, Network research</td>
</tr>
<tr>
<td>The study uses a design that permits a valid comparison with a control to provide a quantitative assessment of drug effect.</td>
<td>Selection bias</td>
<td>Study design choice (comparative cohort vs. self-controlled). Study diagnostics: Propensity score overlap, covariate balance before adjustment. BoO: Population-level effect estimation</td>
</tr>
<tr>
<td>The method of selection of subjects provides adequate assurance that they have the disease or condition being studied.</td>
<td>Measurement error</td>
<td>Phenotype evaluation of indication. Generalizability assessment to target population. BoO: Defining population, Characterization, Clinical Validity</td>
</tr>
<tr>
<td>The method of assigning patients to treatment and control groups minimizes bias and is intended to assure comparability of the groups.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Adequate measures are taken to minimize bias on the part of the subjects, observers, and analysts of the data.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>The methods of assessment of subjects’ response are well-defined and reliable.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>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.</td>
<td>Model misspecification</td>
<td>Study diagnostics: negative control calibration Pre-specification BoO: Methods validity, Software validity</td>
</tr>
</tbody>
</table>

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

**Compare to evidence that we already trust:**
randomized clinical trials

---

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
When do two study estimates agree?
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
When do two study estimates agree?
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?
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
When do two study estimates agree?
# Comparing agreement metrics

<table>
<thead>
<tr>
<th>Agreement metric</th>
<th>Pros</th>
<th>Cons</th>
</tr>
</thead>
<tbody>
<tr>
<td>Concordance Z test</td>
<td>• It’s what you really want&lt;br&gt;• Easy to describe what it is testing</td>
<td>• You can do well with wide confidence intervals&lt;br&gt;• Can do well by guessing no effect</td>
</tr>
<tr>
<td>Study 1 agreement</td>
<td>• Easy to carry out</td>
<td>• Only a heuristic&lt;br&gt;• Can do well by guessing no effect</td>
</tr>
<tr>
<td>Study 2 agreement</td>
<td>• Mostly similar to statistical concordance&lt;br&gt;• Easy to carry out&lt;br&gt;• Tends to penalize wide confidence intervals in study 2</td>
<td>• Only a heuristic&lt;br&gt;• Can do well by guessing no effect</td>
</tr>
<tr>
<td>Significance decision agreement</td>
<td>• Easy to carry out&lt;br&gt;• Indicates what effect it could have on regulation</td>
<td>• Nearly identical studies can be discordant&lt;br&gt;• Identical studies can have mostly different results&lt;br&gt;• Can well by guessing no effect</td>
</tr>
<tr>
<td>Meta-analysis variance test</td>
<td>• Adds in value of the second study</td>
<td>• Hard to carry out&lt;br&gt;• Can do well by guessing no effect</td>
</tr>
<tr>
<td>Study Type</td>
<td>Concordance Z test</td>
<td>Study 1 agreement</td>
</tr>
<tr>
<td>------------</td>
<td>-------------------</td>
<td>-------------------</td>
</tr>
<tr>
<td>1</td>
<td>FAIL</td>
<td>FAIL</td>
</tr>
<tr>
<td>2</td>
<td>FAIL</td>
<td>FAIL</td>
</tr>
<tr>
<td>3</td>
<td>PASS</td>
<td>FAIL</td>
</tr>
<tr>
<td>4</td>
<td>PASS</td>
<td>FAIL</td>
</tr>
<tr>
<td>5</td>
<td>PASS</td>
<td>PASS</td>
</tr>
<tr>
<td>6</td>
<td>PASS</td>
<td>PASS</td>
</tr>
<tr>
<td>7</td>
<td>PASS</td>
<td>PASS</td>
</tr>
</tbody>
</table>

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


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, FA-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 Oviujele, MD, MSc, MAS, MBA, FAHA†;
Sidney C. Smith, Jr, MD, MACC, FAHA‡; Crystal C. Spencer, JD; Randall S. Stafford, MD, PhD¶;
Sandri J. Taler, MD, FAHA§§; Randall 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. Biercher, MS, PharmD, AACC; Bipkam Borkurt, MD, PhD, FACC, FAHA###;
Ralph G. Brindis, MD, MPH, MACC###; Joann 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. Goldman, 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; Dominika 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.

<table>
<thead>
<tr>
<th>COR</th>
<th>LOE</th>
<th>Recommendation</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>A&lt;sup&gt;SR&lt;/sup&gt;</td>
<td>1. For initiation of antihypertensive drug therapy, first-line agents include thiazide diuretics, CCBs, and ACE inhibitors or ARBs. S&lt;sub&gt;8.1.6-1,8.1.6-2&lt;/sub&gt;</td>
</tr>
</tbody>
</table>

SR indicates systematic review.
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).
Dissecting the comparative evidence of ACE vs THZ on AMI

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

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

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

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

**Index:**
- Time zero
Only 15 randomized trial replications of the same class-class comparison

<table>
<thead>
<tr>
<th>Outcome</th>
<th>Study 1 Target</th>
<th>Study 1 Comparator</th>
<th>Study 1 Name</th>
<th>Study 2 Target</th>
<th>Study 2 Comparator</th>
<th>Study 2 Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acute MI</td>
<td>captopril</td>
<td>atenolol</td>
<td>UKPDS 39</td>
<td>lisinopril</td>
<td>atenolol</td>
<td>Agarwal</td>
</tr>
<tr>
<td>Stroke</td>
<td>captopril</td>
<td>atenolol</td>
<td>UKPDS 39</td>
<td>lisinopril</td>
<td>atenolol</td>
<td>Agarwal</td>
</tr>
<tr>
<td>Acute MI</td>
<td>enalapril</td>
<td>nisoldipine</td>
<td>ABCD</td>
<td>fosinopril</td>
<td>amlodipine</td>
<td>FACET</td>
</tr>
<tr>
<td>Stroke</td>
<td>enalapril</td>
<td>nisoldipine</td>
<td>ABCD</td>
<td>fosinopril</td>
<td>amlodipine</td>
<td>FACET</td>
</tr>
<tr>
<td>Heart failure</td>
<td>lisinopril</td>
<td>chlorthalidone</td>
<td>ALLHAT</td>
<td>enalapril</td>
<td>hydrochlorothiazide</td>
<td>ANBP2</td>
</tr>
<tr>
<td>Stroke</td>
<td>lisinopril</td>
<td>chlorthalidone</td>
<td>ALLHAT</td>
<td>enalapril</td>
<td>hydrochlorothiazide</td>
<td>ANBP2</td>
</tr>
<tr>
<td>Acute MI</td>
<td>valsartan</td>
<td>amlodipine</td>
<td>VALUE</td>
<td>valsartan</td>
<td>amlodipine</td>
<td>VART</td>
</tr>
<tr>
<td>Heart failure</td>
<td>valsartan</td>
<td>amlodipine</td>
<td>VALUE</td>
<td>valsartan</td>
<td>amlodipine</td>
<td>VART</td>
</tr>
<tr>
<td>Stroke</td>
<td>valsartan</td>
<td>amlodipine</td>
<td>VALUE</td>
<td>valsartan</td>
<td>amlodipine</td>
<td>VART</td>
</tr>
<tr>
<td>Acute MI</td>
<td>amlodipine</td>
<td>fosinopril</td>
<td>FACET</td>
<td>nifedipine</td>
<td>lisinopril</td>
<td>JMIC-B</td>
</tr>
<tr>
<td>Acute MI</td>
<td>nisoldipine</td>
<td>enalapril</td>
<td>ABCD</td>
<td>nifedipine</td>
<td>lisinopril</td>
<td>JMIC-B</td>
</tr>
<tr>
<td>Acute MI</td>
<td>isradipine</td>
<td>hydrochlorothiazide</td>
<td>MIDAS</td>
<td>lacidipine</td>
<td>chlorthalidone</td>
<td>SHELL</td>
</tr>
<tr>
<td>Stroke</td>
<td>isradipine</td>
<td>hydrochlorothiazide</td>
<td>MIDAS</td>
<td>amlodipine</td>
<td>chlorthalidone</td>
<td>ALLHAT</td>
</tr>
<tr>
<td>Stroke</td>
<td>amlodipine</td>
<td>chlorthalidone</td>
<td>ALLHAT</td>
<td>lacidipine</td>
<td>chlorthalidone</td>
<td>SHELL</td>
</tr>
<tr>
<td>Stroke</td>
<td>isradipine</td>
<td>hydrochlorothiazide</td>
<td>MIDAS</td>
<td>lacidipine</td>
<td>chlorthalidone</td>
<td>SHELL</td>
</tr>
</tbody>
</table>
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)
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)
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?

<table>
<thead>
<tr>
<th>Study Pairs</th>
<th>concordance Z Test</th>
<th>significance Decision Agreement</th>
<th>study 1 Estimate Agreement</th>
<th>study 2 Estimate Agreement</th>
<th>meta Analysis Variance Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>RCT-RCT</td>
<td>87%</td>
<td>67%</td>
<td>67%</td>
<td>67%</td>
<td>73%</td>
</tr>
</tbody>
</table>

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

Index: Time zero
LEGEND results publicly available at data.ohdsi.org
31 randomized trial results can be directly compared with LEGEND

<table>
<thead>
<tr>
<th>RCT name</th>
<th>Target</th>
<th>Comparator</th>
<th>Outcome</th>
</tr>
</thead>
<tbody>
<tr>
<td>FACET</td>
<td>fosinopril</td>
<td>amlodipine</td>
<td>Stroke</td>
</tr>
<tr>
<td>FACET</td>
<td>fosinopril</td>
<td>amlodipine</td>
<td>Acute MI</td>
</tr>
<tr>
<td>VART</td>
<td>valsartan</td>
<td>amlodipine</td>
<td>Stroke</td>
</tr>
<tr>
<td>VART</td>
<td>valsartan</td>
<td>amlodipine</td>
<td>Acute MI</td>
</tr>
<tr>
<td>VART</td>
<td>valsartan</td>
<td>amlodipine</td>
<td>Heart failure</td>
</tr>
<tr>
<td>JMIC-B</td>
<td>nifedipine</td>
<td>lisinopril</td>
<td>Acute MI</td>
</tr>
<tr>
<td>JMIC-B</td>
<td>nifedipine</td>
<td>lisinopril</td>
<td>Hosp for heart failure</td>
</tr>
<tr>
<td>ANBP2</td>
<td>enalapril</td>
<td>hydrochlorothiazide</td>
<td>Stroke</td>
</tr>
<tr>
<td>LIFE</td>
<td>losartan</td>
<td>atenolol</td>
<td>Acute MI</td>
</tr>
<tr>
<td>ALLHAT</td>
<td>lisinopril</td>
<td>chlorthalidone</td>
<td>Stroke</td>
</tr>
<tr>
<td>ALLHAT</td>
<td>amlodipine</td>
<td>chlorthalidone</td>
<td>Stroke</td>
</tr>
<tr>
<td>ALLHAT</td>
<td>lisinopril</td>
<td>chlorthalidone</td>
<td>Heart failure</td>
</tr>
<tr>
<td>ALLHAT</td>
<td>amlodipine</td>
<td>chlorthalidone</td>
<td>Heart failure</td>
</tr>
<tr>
<td>ALLHAT</td>
<td>lisinopril</td>
<td>chlorthalidone</td>
<td>Hosp for heart failure</td>
</tr>
<tr>
<td>ALLHAT</td>
<td>amlodipine</td>
<td>chlorthalidone</td>
<td>Hosp for heart failure</td>
</tr>
<tr>
<td>Agarwal</td>
<td>lisinopril</td>
<td>atenolol</td>
<td>Acute MI</td>
</tr>
<tr>
<td>Agarwal</td>
<td>lisinopril</td>
<td>atenolol</td>
<td>Stroke</td>
</tr>
<tr>
<td>HAPPHY</td>
<td>hydrochlorothiazide</td>
<td>atenolol</td>
<td>Stroke</td>
</tr>
<tr>
<td>HAPPHY</td>
<td>hydrochlorothiazide</td>
<td>atenolol</td>
<td>Acute MI</td>
</tr>
<tr>
<td>UKPDS 39</td>
<td>captopril</td>
<td>atenolol</td>
<td>Acute MI</td>
</tr>
<tr>
<td>UKPDS 39</td>
<td>captopril</td>
<td>atenolol</td>
<td>Stroke</td>
</tr>
<tr>
<td>VALUE</td>
<td>valsartan</td>
<td>amlodipine</td>
<td>Acute MI</td>
</tr>
<tr>
<td>VALUE</td>
<td>valsartan</td>
<td>amlodipine</td>
<td>Stroke</td>
</tr>
<tr>
<td>VALUE</td>
<td>valsartan</td>
<td>amlodipine</td>
<td>Heart failure</td>
</tr>
</tbody>
</table>
RCT-LEGEND estimate comparisons
RCT-LEGEND estimate comparisons

ANBP 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)
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)
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 (0.73-1.08)
How well does LEGEND replicate RCTs?

<table>
<thead>
<tr>
<th>Study Pairs</th>
<th>concordance Z Test</th>
<th>significance Decision Agreement</th>
<th>study 1 Estimate Agreement</th>
<th>study 2 Estimate Agreement</th>
<th>meta Analysis Variance Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>RCT-RCT</td>
<td>87%</td>
<td>67%</td>
<td>67%</td>
<td>67%</td>
<td>73%</td>
</tr>
<tr>
<td>RCT-LEGEND</td>
<td>87%</td>
<td>52%</td>
<td>68%</td>
<td>74%</td>
<td>81%</td>
</tr>
</tbody>
</table>

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”?

<table>
<thead>
<tr>
<th>Study Pairs</th>
<th>concordance Z Test</th>
<th>significance Decision</th>
<th>study 1 Estimate Agreement</th>
<th>study 2 Estimate Agreement</th>
<th>meta Analysis Variance Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>RCT-RCT</td>
<td>87%</td>
<td>67%</td>
<td>67%</td>
<td>67%</td>
<td>73%</td>
</tr>
<tr>
<td>RCT-LEGEND</td>
<td>87%</td>
<td>52%</td>
<td>68%</td>
<td>74%</td>
<td>81%</td>
</tr>
<tr>
<td>“Perfect replication”</td>
<td>95%</td>
<td>55%</td>
<td>68%</td>
<td>77%</td>
<td>82%</td>
</tr>
</tbody>
</table>

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”?

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

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

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

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

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.

<table>
<thead>
<tr>
<th>COR</th>
<th>LOE</th>
<th>Recommendation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>1. For initiation of antihypertensive drug therapy, first-line agents include thiazide diuretics, CCBs, and ACE inhibitors or ARBs.</td>
</tr>
</tbody>
</table>
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

<table>
<thead>
<tr>
<th>Class</th>
<th>Drug</th>
<th>Usual Dose, Range (mg/d)</th>
<th>Daily Frequency</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Primary agents</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Thiazide or thiazide-type</td>
<td>Chlorthalidone</td>
<td>12.5–25</td>
<td>1</td>
<td>Chlorthalidone is preferred on the basis of prolonged half-life and proven trial reduction of CVD.</td>
</tr>
<tr>
<td>diuretics</td>
<td>Hydrochlorothiazide</td>
<td>25–50</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Indapamide</td>
<td>1.25–2.5</td>
<td>1</td>
<td>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.</td>
</tr>
<tr>
<td></td>
<td>Metolazone</td>
<td>2.5–5</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>
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*

<table>
<thead>
<tr>
<th>COR</th>
<th>LOE</th>
<th>Recommendations</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>C-EO</td>
<td>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.</td>
</tr>
</tbody>
</table>
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