OHDSI
OBSERVATIONAL HEALTH DATA SCIENCES AND INFORMATICS

Objective study validity diagnostics: a framework requiring pre-specified, empirical verification to increase trust in the reliability of real-world evidence

Mitchell M Conover, Patrick B Ryan, Yong Chen,
Marc A Suchard, George Hripcsak, Martijn J Schuemie

Journal of the American Medical Informatics Association, Volume 32, Issue 3, March 2025, Pages 518–525 https://doi.org/10.1093/jamia/ocae317

Conflicts of Interest

- Mitch Conover, Patrick Ryan, and Martijn Schuemie are employees and shareholders of Johnson & Johnson

- Marc Suchard receives grants and contracts from US Food & Drug Administration and Johnson & Johnson

Framework for objective diagnostics

How to assess the reliability of RWE studies?

- Diagnostics (e.g. covariate balance: standardized difference of means < 0.1)

Building on LEGEND framework: objective diagnostic measures should be used to evaluate/report validity of observational findings by either:

1. interpreting objective diagnostic results before unblinding study results
2. only unblinding results from analyses for which all objective diagnostics pass *pre-specified* thresholds

Diagnostic failures should be reported alongside unblinded results

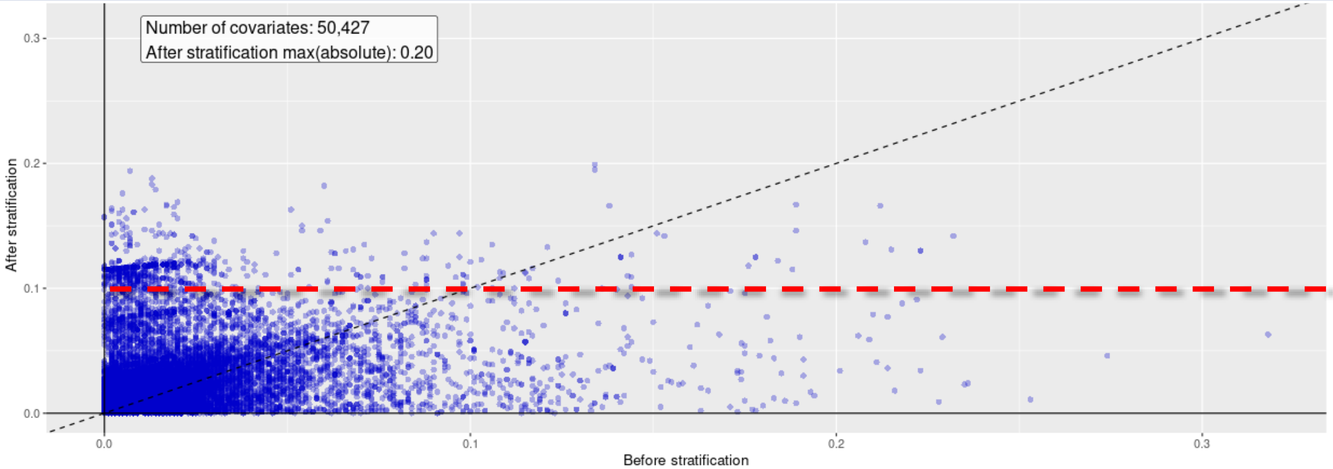
Study objective

- Six diagnostic metrics for comparative cohort studies:
  1. Covariate balance: maximum standardized difference of means (SDM)
  2. Empirical equipoise
  3. Expected absolute systematic error (EASE)
  4. Generalizability standardized difference of means
  5. Minimum detectable relative risk (MDRR)

- We provide conceptual overviews of each, the key assumption it tests, considerations or references when pre-specifying diagnostic thresholds

Covariate balance:
maximum standardized difference of means (SDM)

Threat to validity	Metric calculation	Threshold guidance
Confounding bias[26–28]	The SDM compares the proportion or mean of exposed and unexposed, scaled to the pooled standardized deviation. The maximum SDM is the largest SDM measured across all observed baseline variables. $SDM = \frac{(\bar{x}_T - \bar{x}_C)}{\sqrt{\frac{s_T^2 + s_C^2}{2}}}$ for continuous variables $SDM = \frac{(\hat{p}_T - \hat{p}_C)}{\sqrt{\frac{\hat{p}_T(1-\hat{p}_T) + \hat{p}_C(1-\hat{p}_C)}{2}}}$ for dichotomous variables T=target, C=comparator	$SDM_{max} > 0.10$ conventionally interpreted to indicate the presence of confounding bias based on Austin et al. heuristic.[26–29]



Re-using LEGEND-HTN Negative Control Experiments

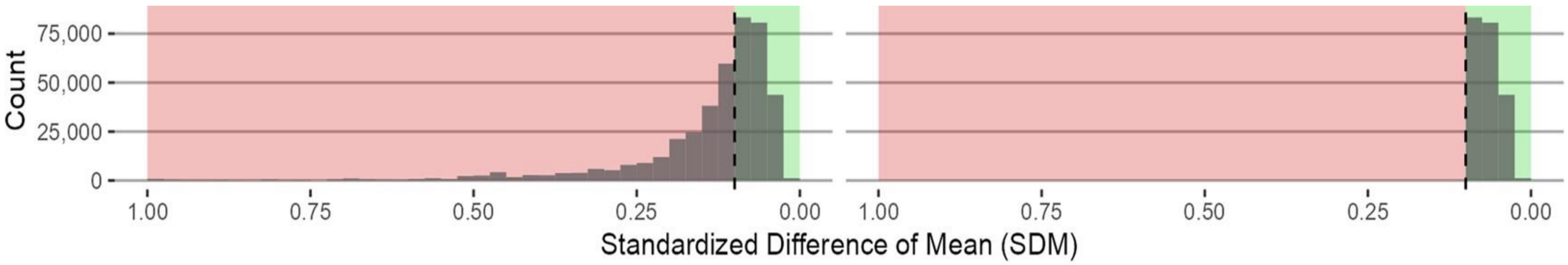
- On-treatment comparisons of the effect of various monotherapy antihypertensive treatments
- Six administrative claims databases and three electronic health record databases
- Large-scale propensity score (LSPS) adjustment (stratification and variable-ratio matching) was used to control confounding
- Empirical calibration used to account for residual systematic error
    - 11,716 negative control exposure-comparator-outcome triplets

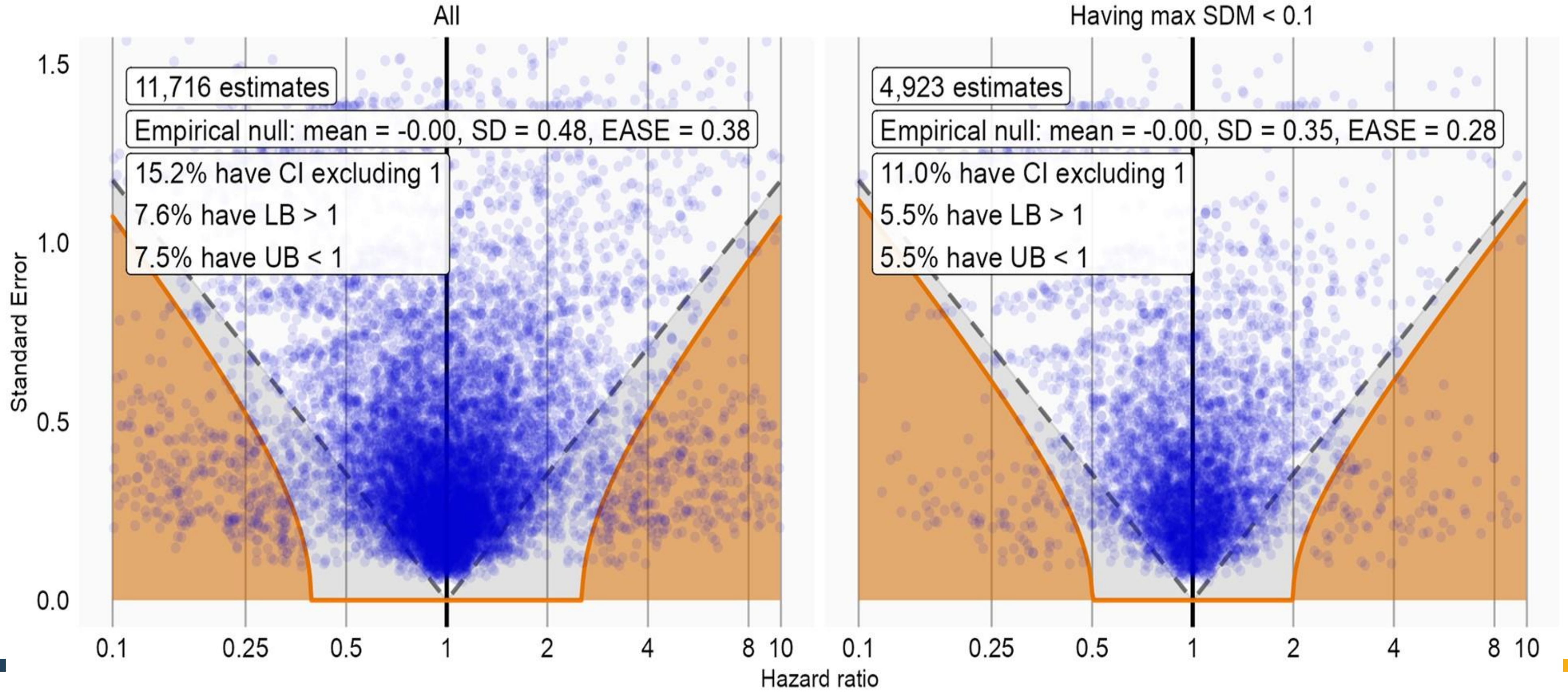
Re-using LEGEND-HTN Negative Control Experiments

- For each negative control analysis, we implemented various diagnostic thresholds:
  - Covariate balance SDM < 0.10
  - Empirical equipoise ≥ 0.50
  - Systematic error (EASE) ≤ 0.25
  - Generalizability SDM ≤ 0.25
  - MDRR≤10

- We computed the distribution of diagnostics across 11,716 LEGEND-HTN negative control studies

Covariate balance SDM < 0.1





LEGEND Negative Control Results For Selected Diagnostics

Diagnostic threshold(s)	N (% satisfied)	EASE	EASE_{Δ}
None	11,716 (100.0%)	0.38	-
Covariate balance SDM < 0.1	4,923 (42.0%)	0.28	-0.10
Equipoise > 0.5	2,792 (23.8%)	0.02	-0.36
Equipoise > 0.1	10,010 (85.4%)	0.33	-0.05
All[*]	1,633 (13.9%)	0.00	-0.38

Some diagnostics dramatically reduce systematic error but only by excluding a large share of (potentially valid) studies

[*] MDRR≤10, equipoise ≥ 0.50, covariate balance SDM < 0.10, generalizability SDM ≤ 0.25, systematic error (EASE) ≤ 0.25

Key take-aways

- Objective diagnostics are crucial for evaluating and communicating the reliability of evidence generated by observational studies

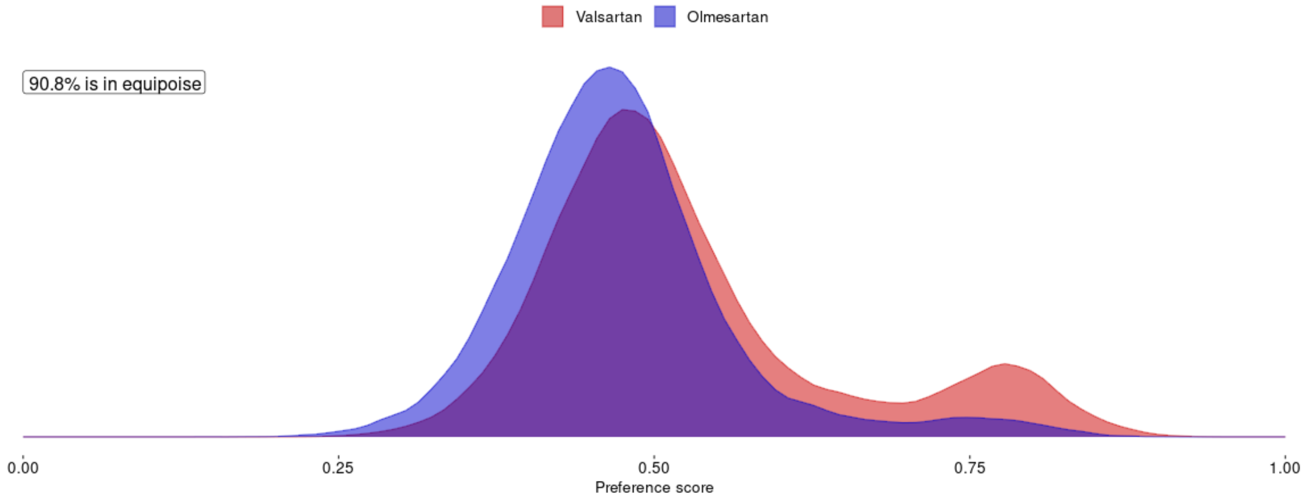
- More work is needed to identify new diagnostics, establish their use across study designs (e.g. SCCS), and provide guidance for diagnostic thresholds

BACKUP SLIDES

Empirical equipoise

Threat to validity	Metric calculation	Threshold guidance

Confounding[24]
Non-positivity[23]
	$\ln(\frac{F}{(1-F)} = \ln\left(\frac{S}{1-S}\right) - \ln(\frac{P}{1-P})$
	F=Preference score
	S=Propensity score
	P=Fraction of people receiving target
		0.3 ≤ F ≤ 0.7 in more than half of patients[24]



90.8% is in equipoise

Objective Diagnostic	Threat to validity	Metric calculation	Threshold guidance
Minimum detectable relative risk	Misinterpreting wide effect estimates from grossly underpowered studies	Compute the minimum detectable relative risk (MDRR) metric and expected standard error (SE) for a given study population, using the actual observed sample size and number of outcomes (after analytic approaches have been applied).[17] $mdrr = e^{\sqrt{\frac{\left(Z_{\beta}+Z_{1-\frac{\alpha}{2}}\right)^{2}}{totalEvents * P_A * P_B}}}$	We propose MDRR < 10, although there is debate whether power calculations have utility in studies using pre-existing observational data.[18–21]
Empirical equipoise	Confounding[24] Non-positivity[23]	$\ln(\frac{F}{(1-F)} = \ln\left(\frac{S}{1-S}\right) - \ln(\frac{P}{1-P})$ F=preference score S=Propensity score for receiving target P=Fraction of people receiving target	0.3 ≤ F ≤ 0.7 in more than half of patients[24]
Covariate balance maximum standardized difference of means (SDM)	Confounding bias[26–28]	The SDM compares the proportion or mean of exposed and unexposed, scaled to the pooled standardized deviation. The maximum SDM is the largest SDM measured across all observed baseline variables. $SDM = \frac{(\bar{x}_T - \bar{x}_C)}{\sqrt{\frac{s_T^2 + s_C^2}{2}}}$ for continuous variables $SDM = \frac{(\hat{p}_T - \hat{p}_C)}{\sqrt{\frac{\hat{p}_T(1-\hat{p}_T) + \hat{p}_C(1-\hat{p}_C)}{2}}}$ for dichotomous variables T=target, C=comparator	$SDM_{max} > 0.10$ conventionally interpreted to indicate the presence of confounding bias based on Austin et al. heuristic.[26–29]
Generalizability maximum SDM	Selection bias[31]	Same calculation as covariate balance SDM, comparing analytic vs. target population	$SDM_{max} < 0.25$ suggested as a rule of thumb to indicate that the population is “like a random sample”[31,32]
Expected Absolute Systematic Error (EASE)	Systematic error (selection, confounding, misclassification bias)[1]	$EASE = average(\lvert\ln(HR_{estimate}) - \ln(HR_{truth})\rvert)$ across negative control outcome studies	A current rule of thumb is EASE < 0.25.

Full Results Table

	LEGEND studies	LEGEND negative control studies				
Diagnostic threshold(s)	N (% satisfied)	N (% satisfied)	log-HR$_\mu$ (SD)*	EASE	EASE$_\Delta$	CIs excl. null (%)
None	471,321 (100.0%)	11,716 (100.0%)	0.00 (0.48)	0.38	-	15.2%
All†	54,358 (11.5%)	1,633 (13.9%)	0.00 (0.00)	0.00	-0.38	3.9%
MDRR < 10	447,445 (94.9%)	11,233 (95.9%)	0.00 (0.48)	0.38	0.00	15.7%
Equipoise > 0.5	136,405 (28.9%)	2,792 (23.8%)	0.00 (0.02)	0.02	-0.36	4.7%
Equipoise > 0.1	413,489 (87.7%)	10,010 (85.4%)	0.00 (0.41)	0.33	-0.05	13.5%
Covariate balance SDM < 0.1	204,758 (43.4%)	4,923 (42.0%)	0.00 (0.35)	0.28	-0.10	11.0%
Generalizability SDM < 0.25	203,986 (43.3%)	4,942 (42.2%)	0.03 (0.47)	0.37	-0.01	13.9%
EASE < 0.25	394,953 (83.8%)	9,718 (82.9%)	0.00 (0.44)	0.35	-0.03‡	14.3%