ARE ASSESSMENT TOOLS GOOD PREDICTORS OF EMPIRICAL BIAS?

Hugh Sharma Waddington, LSHTM

Cochrane Bias Methods

Global Evidence Summit 2024

Prague

DOES EMPIRICAL BIAS FACTOR IN THE CREATION OF RISK ASSESSMENT TOOLS?

DECLARATION OF INTEREST

To the best of my knowledge, I declare that I, and/or any of my co-authors, and any of my close family members, have not had employment, received research support or other funding from, or had any other professional relationship with, an entity directly involved in the production, manufacture, distribution or sale of tobacco, tobacco products, weapons or arms, or have represented the interests of any such entities in any way.

My perspective is that, in non-clinical topic areas where blinding of study participants is often impossible, we need to recognize, and provide reviewers with the tools to appraise, the full range of credible quasi-experimental designs (QEDs) and non-randomized studies of interventions (NRSI).

I am not aware of any other conflict of interest in relation to this presentation.

HOW CAN WE QUANTIFY BIAS IN INDIVIDUAL STUDIES?

Two ways of quantifying bias empirically:

1. Measure distance from benchmark study in **internal replication** (within-study comparison)

2. Measure the relationship between predicted bias and distance from benchmark in **external replication**

TWO TYPES OF INTERNAL REPLICATION STUDY
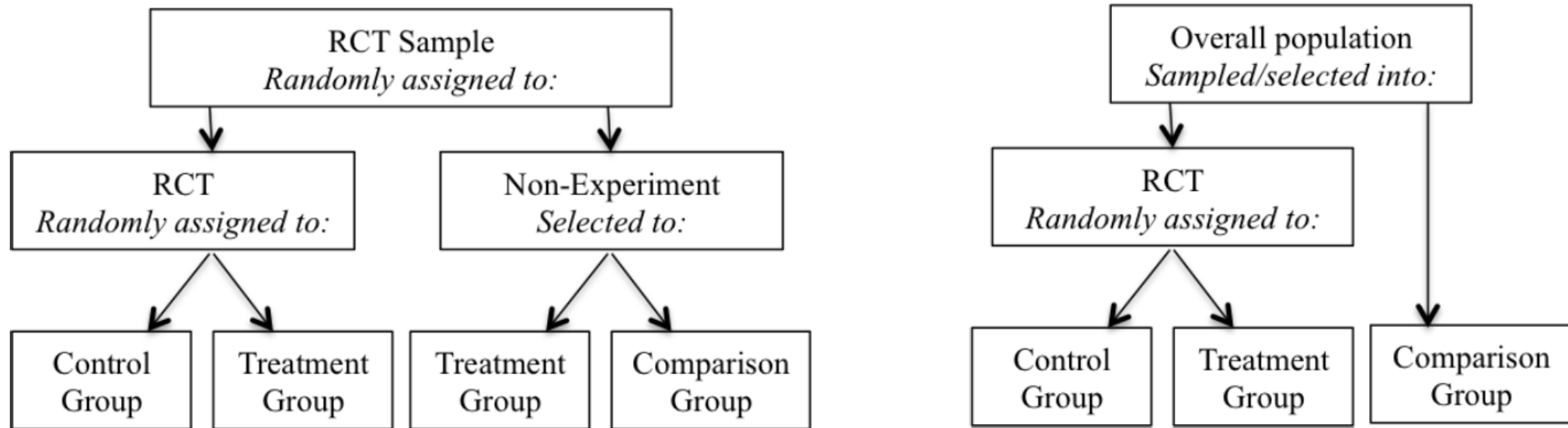
Specific adjustments for discontinuity designs to ensure comparable populations

Source: Wong and Steiner (2017).

WHY MIGHT RESULTS FROM RCTS AND NRSI DIFFER?

1. Bias in parameter estimate (internal validity)
   - Main sources of bias in RCTs: subversion of randomisation, attrition
   - Main sources of bias in NRSI: confounding, selection bias, selection of the reported result

2. Sampling bias (external validity)
   - Discrepancy in treatment effect estimand due to differences in the target population in each study (e.g., ITT versus per protocol effect; PATE versus LATE)

INTERNAL REPLICATIONS IN INTERNATIONAL DEVELOPMENT

Internal replication study	Intervention	Benchmark	NRSI
Buddelmeyer and Skoufias (2004)	Cash transfer	RCT	Regression discontinuity (RD)
Diaz and Handa (2006)	Cash transfer	RCT	Adjusted regression, matching
Handa and Maluccio (2010)	Cash transfer	RCT	Matching
McKenzie et al. (2010)	Immigration entitlement	Randomised policy experiment	Difference-in-difference, instrumental variables, matching
Galiani and McEwan (2013)	Cash transfer	RCT	RD
Barrera-Osorio et al. (2014)	Scholarship	RCT	RD
Chaplin et al. (2017)	Subsidy	RCT	Matching
Galiani et al. (2017)	Cash Transfer	RCT	Geographical discontinuity design

Source: Sharma Waddington, Fenton and Valentine 2022, *Evaluation Review*.

TWO MAIN TYPES OF NRSI

Distinguish NRSI according to the approach used to address *confounding*:

- ***selection on observables****:* mainly analysis-based (e.g., non-randomised cohort, case-control, cross-section with propensity score matching)

- ***selection on unobservables****:* mainly design-based (e.g., discontinuity design, difference-in-differences, interrupted time series) and some analysis-based (e.g., instrumental variables)

DISTANCE MEASURES OF THE CORRESPONDENCE BETWEEN NRSI AND RCT

Standardised bias $|D_i| = \dfrac{\left|\bar{Y}^C_{NRS} - \bar{Y}^C_{RCT}\right|}{s_{RCT}}$

Mean squared error $MSE_i = D_i^2 + s_i^2$

% bias removed $|D_R| = \left(1 - \dfrac{\bar{Y}^C_{NRS} - \bar{Y}^C_{RCT}}{|\bar{Y}^C_{PF} - \bar{Y}^C_{RCT}|}\right) x\, 100$

Source: Steiner and Wong (2017). Note: PF *prima facie* mean from unadjusted NRSI

FIXED EFFECT META-ANALYSIS OF BIAS ESTIMATES

Estimator	Standardised bias	Mean squared error	Percent of bias removed	Num. of estimates
Adjusted regression (cross-section data)	0.29	0.18	34%	10
Baseline adjustment (DID, PSM)	0.05	0.01	56%	17
Discontinuity design (RDD, GDD)	0.01	0.00	95%	173
Interrupted time series	-	-	-	-
Instrumental variables (strong instrument)	0.01	0.00	95%	1
Instrumental variables (weak instrument)	0.31	0.14	-92%	2
Matching (nearest neighbour)	0.13	0.07	52%	59
Matching (kernel)	0.28	0.15	34%	70

Standardised bias $|D_i| = \frac{|\bar{Y}_{NRS}^c - \bar{Y}_{RCT}^c|}{S_{RCT}}$; Mean squared error $MSE_i = D_i^2 + s_i^2$; % bias removed $|D_R| = \left(1 - \frac{\bar{Y}_{NRS}^c - \bar{Y}_{RCT}^c}{|\bar{Y}_{PF}^c - \bar{Y}_{RCT}^c|}\right) x\ 100$
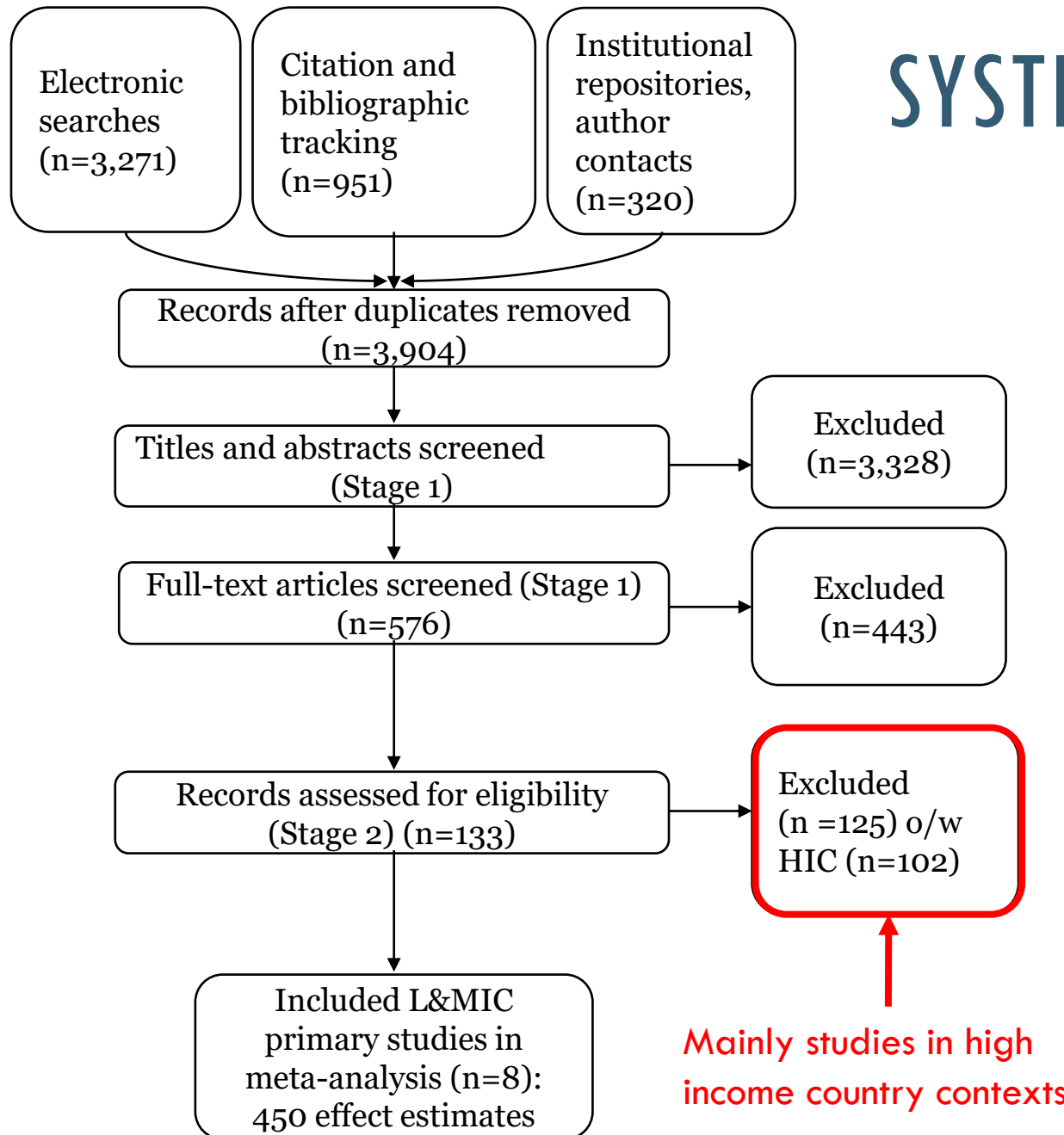
SYSTEMATIC REVIEW CONDUCT

Electronic searches (n=3,271)

Citation and bibliographic tracking (n=951)

Institutional repositories, author contacts (n=320)

Records after duplicates removed (n=3,904)

Titles and abstracts screened (Stage 1) → Excluded (n=3,328)

Full-text articles screened (Stage 1) (n=576) → Excluded (n=443)

Records assessed for eligibility (Stage 2) (n=133) → Excluded (n =125) o/w HIC (n=102)

Included L&MIC primary studies in meta-analysis (n=8): 450 effect estimates

Mainly studies in high income country contexts

First systematic review and meta-analysis of these types of studies; registered with *Campbell Collaboration*

All 450 bias estimates collected, appraised and categorise in duplicate

Meta-analysis weights adjusted for dependent effect sizes

RISK OF BIAS IN BENCHMARK AND DISTANCE METRIC

Within study comparison	Buddelmeyer and Skoufias (2004)	Diaz and Handa (2006)	Handa and Maluccio (2010)	McKenzie et al. (2010)	Barrera-Osorio et al. (2014)	Galiani and McEwan (2013); Galiani et al. (2017)	Chaplin et al. (2017)
Confounding bias due to randomisation process	Some concerns	Some concerns	Some concerns	Some concerns	Low risk	Low risk	Some concerns
Selection bias in recruitment	Some concerns	Some concerns	Some concerns	Some concerns	Low risk	Low risk	Some concerns
Attrition bias due to missing outcome data	High risk	Some concerns	Some concerns	Some concerns	Some concerns	Low risk	Low risk
Departures from intended intervention^	Some concerns	Some concerns	Low risk	Low risk	Low risk	Low risk	Some concerns
Bias in measurement of the outcome	Some concerns	Some concerns	Some concerns	Some concerns	Low risk	Low risk	Low risk
Selective analysis and reporting	Low risk	Low risk	Low risk	Low risk	Low risk	Low risk	Low risk
Bias in NRS estimate	Low risk	Some concerns	Some concerns	Low risk	Low risk	Some concerns	Low risk
Overall bias in within-study comparison	**High risk**	**Some concerns**	**Some concerns**	**Some concerns**	**Some concerns**	**Some concerns**	**Some concerns**

Note: ^ blinding of study participants not conducted in any benchmark randomized controls.

# HOW CAN WE QUANTIFY BIAS IN INDIVIDUAL STUDIES?

Two ways of quantifying bias empirically:

1. Measure distance from benchmark study in **internal replication** (within-study comparison)

2. Measure the relationship between predicted bias and distance from benchmark in **external replication** (e.g., meta-epidemiology)

"META-SOCIOECONOMICS": CAMPBELL DEVELOPMENT SYSTEMATIC REVIEWS CONTAINING RCTS AND NRSI

Review author	Sector	Outcomes	# RCTs	# NRS
Baird et al. (2013)	Education	School attendance	15	27
Brody et al. (2015)	Micro-finance	Women's empowerment	5	18
Carr-Hill et al. (2016)	Education	Drop-outs, test scores	9	17
Chinen et al. (2017)	Vocational training	Employment, earnings	26	9
Hemming et al. (2018)	Agriculture	Adoption, yield, income	2	13
Molina et al. (2016)	Governance	Health outcomes	10	5
Piza et al. (2016)	Vocational training and finance	Firm performance, employment	6	23
Stone et al. (2019)	Education	Literacy	9	7
Tripney et al. (2013)	Vocational training	Employment, income	3	23
Vaessen et al. (2014)	Micro-finance	Women's empowerment	4	21
Waddington et al. (2019)	Governance	Community engagement, service access, service use	19	16

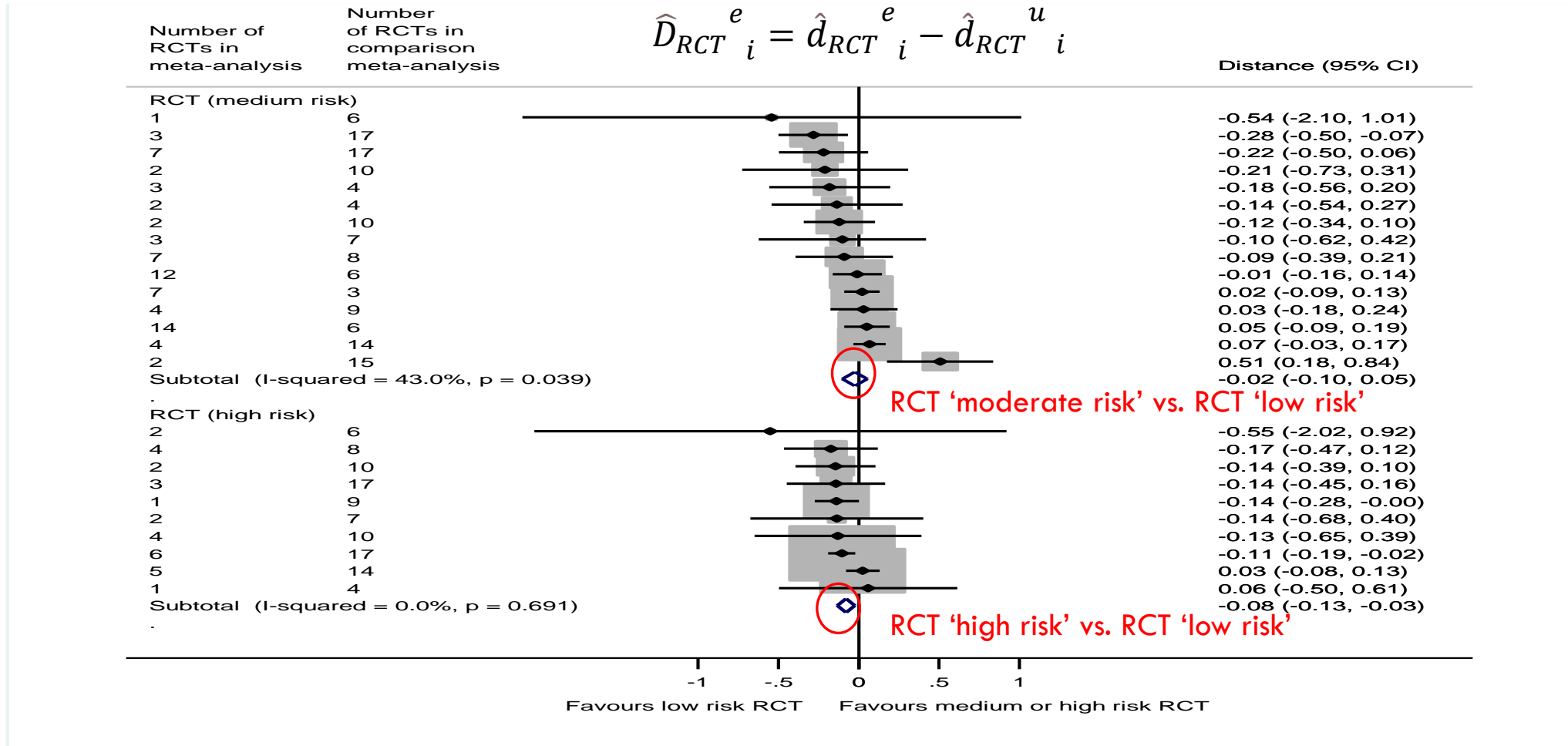# ESTIMATES BY STUDY DESIGN AND RISK OF BIAS
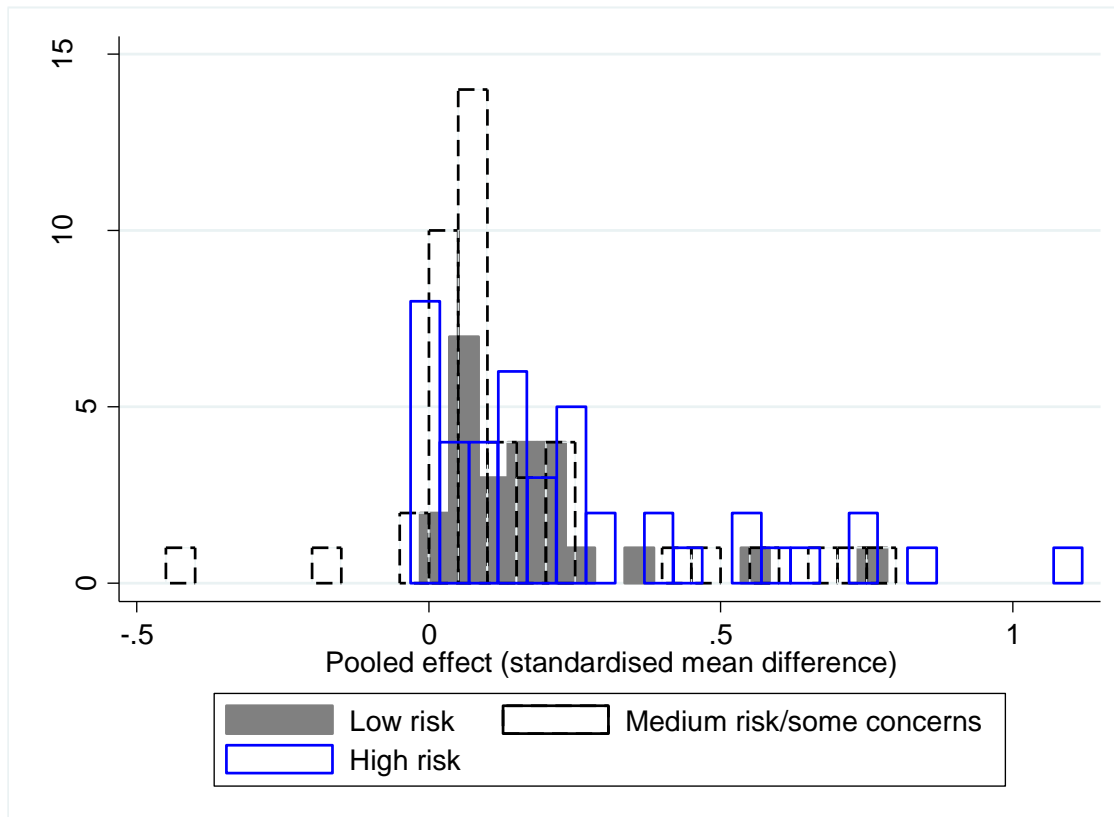
Numerical difference in standardised effect:

$$\widehat{D}_{NRS_i} = \hat{d}_{NRS\ i} - \hat{d}_{RCT\ i}$$

$$se(\widehat{D})_i = \sqrt{s^2_{NRS_i} + s^2_{RCT_i}}$$

All values calculated such that D>0 represents desirable change in outcome

'HIGH RISK' NRSI OVERSTATE EFFECTS WITH GREATER UNCERTAINTY

'HIGH RISK' RCTS UNDERESTIMATE EFFECTS?

$$\widehat{D}_{RCT}{}^{e}{}_{i} = \hat{d}_{RCT}{}^{e}{}_{i} - \hat{d}_{RCT}{}^{u}{}_{i}$$

Number of RCTs in meta-analysis	Number of RCTs in comparison meta-analysis	Distance (95% CI)
RCT (medium risk)		
1	6	-0.54 (-2.10, 1.01)
3	17	-0.28 (-0.50, -0.07)
7	17	-0.22 (-0.50, 0.06)
2	10	-0.21 (-0.73, 0.31)
3	4	-0.18 (-0.56, 0.20)
2	4	-0.14 (-0.54, 0.27)
2	10	-0.12 (-0.34, 0.10)
3	7	-0.10 (-0.62, 0.42)
7	8	-0.09 (-0.39, 0.21)
12	6	-0.01 (-0.16, 0.14)
7	3	0.02 (-0.09, 0.13)
4	9	0.03 (-0.18, 0.24)
14	6	0.05 (-0.09, 0.19)
4	14	0.07 (-0.03, 0.17)
2	15	0.51 (0.18, 0.84)
Subtotal (I-squared = 43.0%, p = 0.039)		-0.02 (-0.10, 0.05)
.		
RCT (high risk)		
2	6	-0.55 (-2.02, 0.92)
4	8	-0.17 (-0.47, 0.12)
2	10	-0.14 (-0.39, 0.10)
3	17	-0.14 (-0.45, 0.16)
1	9	-0.14 (-0.28, -0.00)
2	7	-0.14 (-0.68, 0.40)
4	10	-0.13 (-0.65, 0.39)
6	17	-0.11 (-0.19, -0.02)
5	14	0.03 (-0.08, 0.13)
1	4	0.06 (-0.50, 0.61)
Subtotal (I-squared = 0.0%, p = 0.691)		-0.08 (-0.13, -0.03)
.		

RCT 'moderate risk' vs. RCT 'low risk'

RCT 'high risk' vs. RCT 'low risk'

-1   -.5   0   .5   1

Favours low risk RCT   Favours medium or high risk RCT

Note: no randomized control trial participants were blinded due to the nature of the interventions.

SUMMARY OF RANDOM EFFECTS META-ANALYSES

| Comparison	D	95% confidence interval		I²	Tau²	N
NRSI – RCT	**0.045**	**0.010**	**0.080**	68%	0.004	28
NRSI (low) – RCT	0.002	-0.056	0.060	0%	0.000	6
NRSI (moderate) – RCT	0.010	-0.027	0.048	0%	0.000	15
NRSI (high) – RCT	**0.171**	**0.065**	**0.278**	78%	0.033	18
RCT (moderate) – RCT (low)	-0.024	-0.102	0.053	43%	0.008	15
RCT (high) – RCT (low)	**-0.080**	**-0.135**	**-0.026**	0%	0.000	10
NRSI (low) – RCT (low)	-0.001	-0.044	0.042	0%	0.000	4
NRSI (moderate) – RCT (low)	-0.013	-0.060	0.034	0%	0.000	12
NRSI (high) – RCT (low)	**0.130**	**0.008**	**0.253**	53%	0.021	13

CONCLUSIONS

Study design probably the most important factor determining bias in NRSI (see also Tom Cook, Will Shadish & Vivian Wong):

- Knowledge about the process determining selection into treatment (e.g., discontinuity designs)

- Incorporating pre-test measurement (panel data)

Study conduct also matters, especially where designs rely on strong assumptions (e.g., instrumental variables, propensity score matching)

Inverse relationship – low risk of bias leads to greater effect magnitude – when RCTs are open (unblinded), possibly due to "site selection bias"

Great scope to learn from internal replication literature using systematic methods in health, to inform risk-of-bias tools
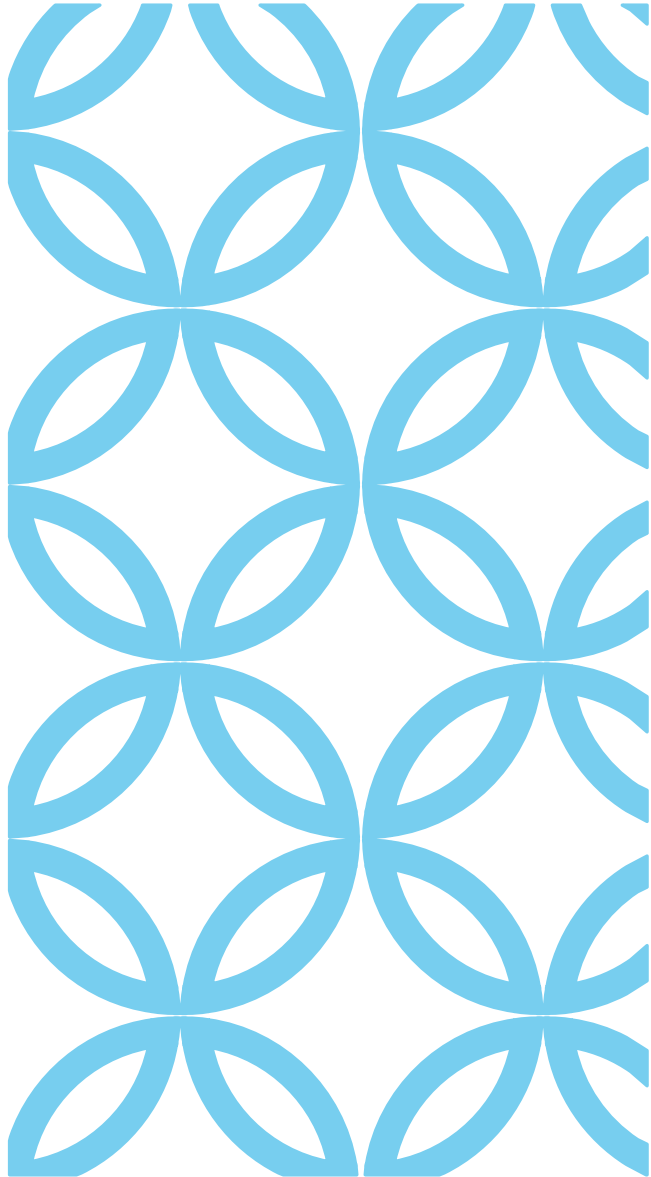
THANKS

Hugh.Waddington@lshtm.ac.uk