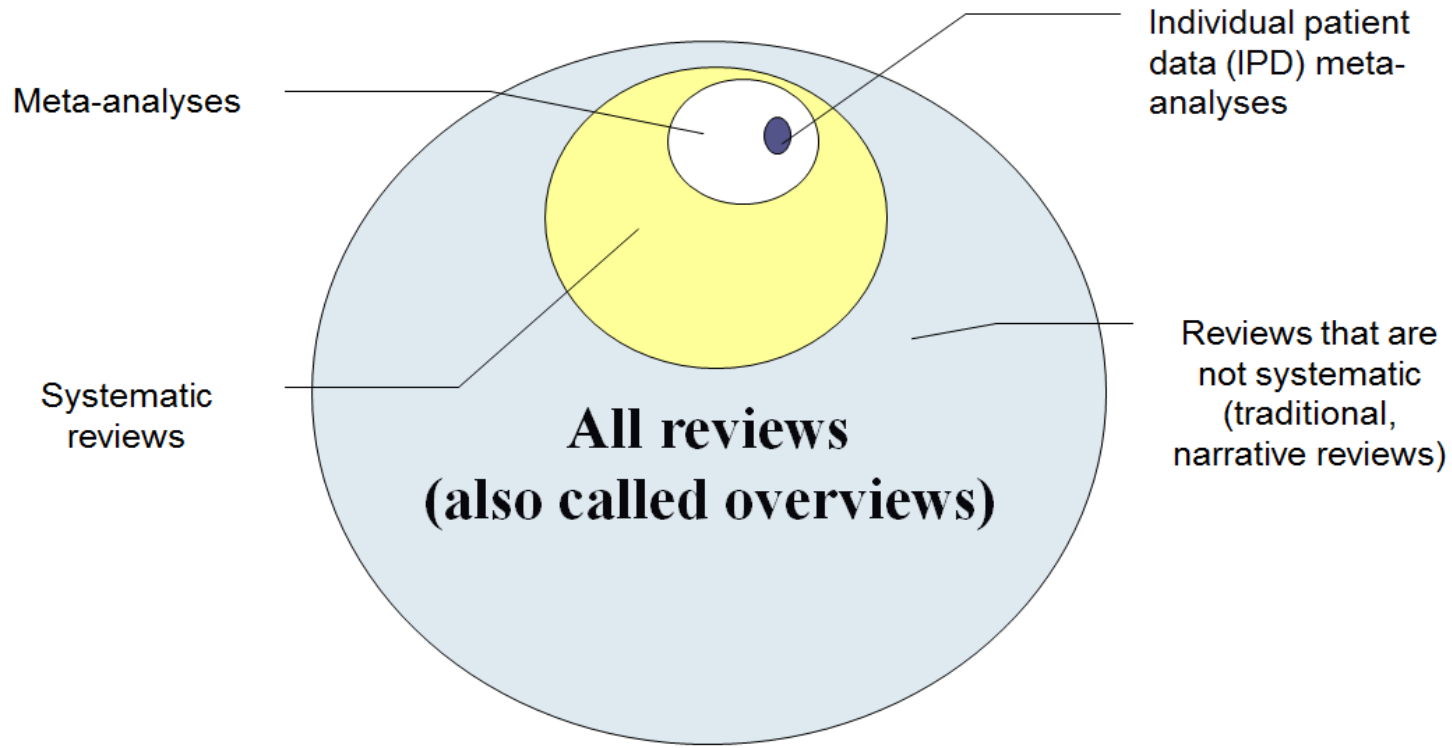
Individual participant data meta analyses

NOVEMBER 2025

ANDREA BENEDETTI, PHD
MCGILL UNIVERSITY

What are IPD meta-analyses

Reviews, MA, IPD-MA



Pai M, et al. Systematic reviews and meta-analyses: An illustrated, step-by-step guide. *Natl Med J India* 2004;17(2):86-95.

Figure : Relationship between Reviews, Systematic Reviews, Meta-analyses and Individual Patient Data Meta-analyses

Conventional Meta Analysis

Collect summary measures from investigators/papers

- Odds ratios, risk ratios, risks, etc.
- From the existing report!

"Aggregate data"

Individual Participant Data Meta Analysis (IPD-MA)

Collect line by line data for each subject from investigators

Individual participant data: "IPD"

"Gold standard"

Aggregate Data:

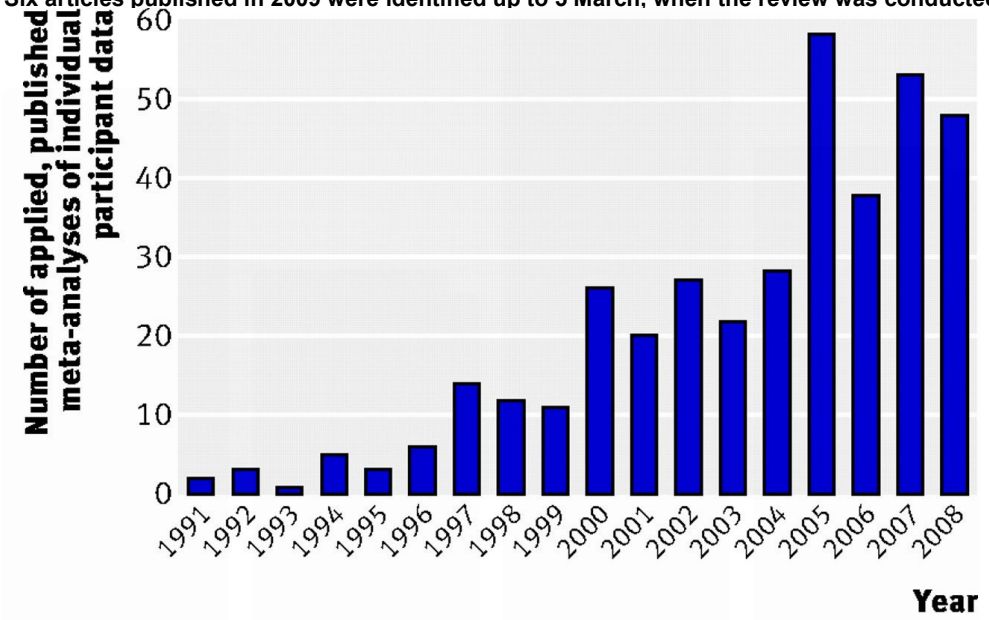
Study	Cutoff	Sens	Spec	N	N Depress	Mean Age	%Male	Setting
1	10	85	86	266	57	67.1	48	Inpatient
2	8	88	88	1053	204	22.6	52	Outpatient
3	9	78	89	…	…	…	…	Outpatient
4	11	74	85	…	…	…	…	…
5	10	82	87	…	…	..	…	…
…..								

Individual Participant Data:

Study	PatientID	PHQ Score	Depressed?	Age	Sex	…
1	1	10	Y	19	M	
1	2	6	N	37	F	
1	3	4	N	26	F	
1	4	15	N	42	F	
1	5	8	Y	58	M	
…						
2						
2						
2						

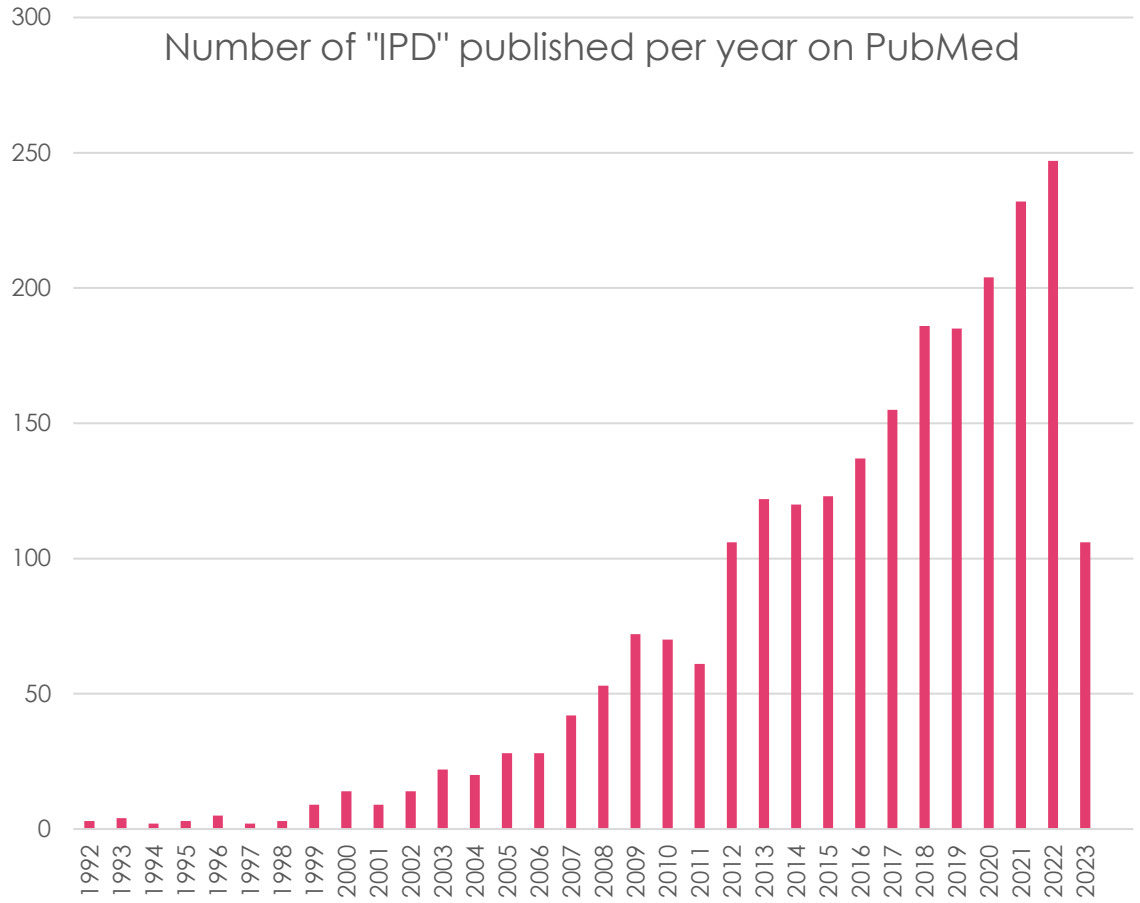
Increasing role of IPD meta analyses

Fig 1 Number of distinct, applied meta-analyses of individual participant data published up to March 2009,* as identified by a systematic review of Medline, Embase, and the Cochrane Library. *Six articles published in 2009 were identified up to 5 March, when the review was conducted.

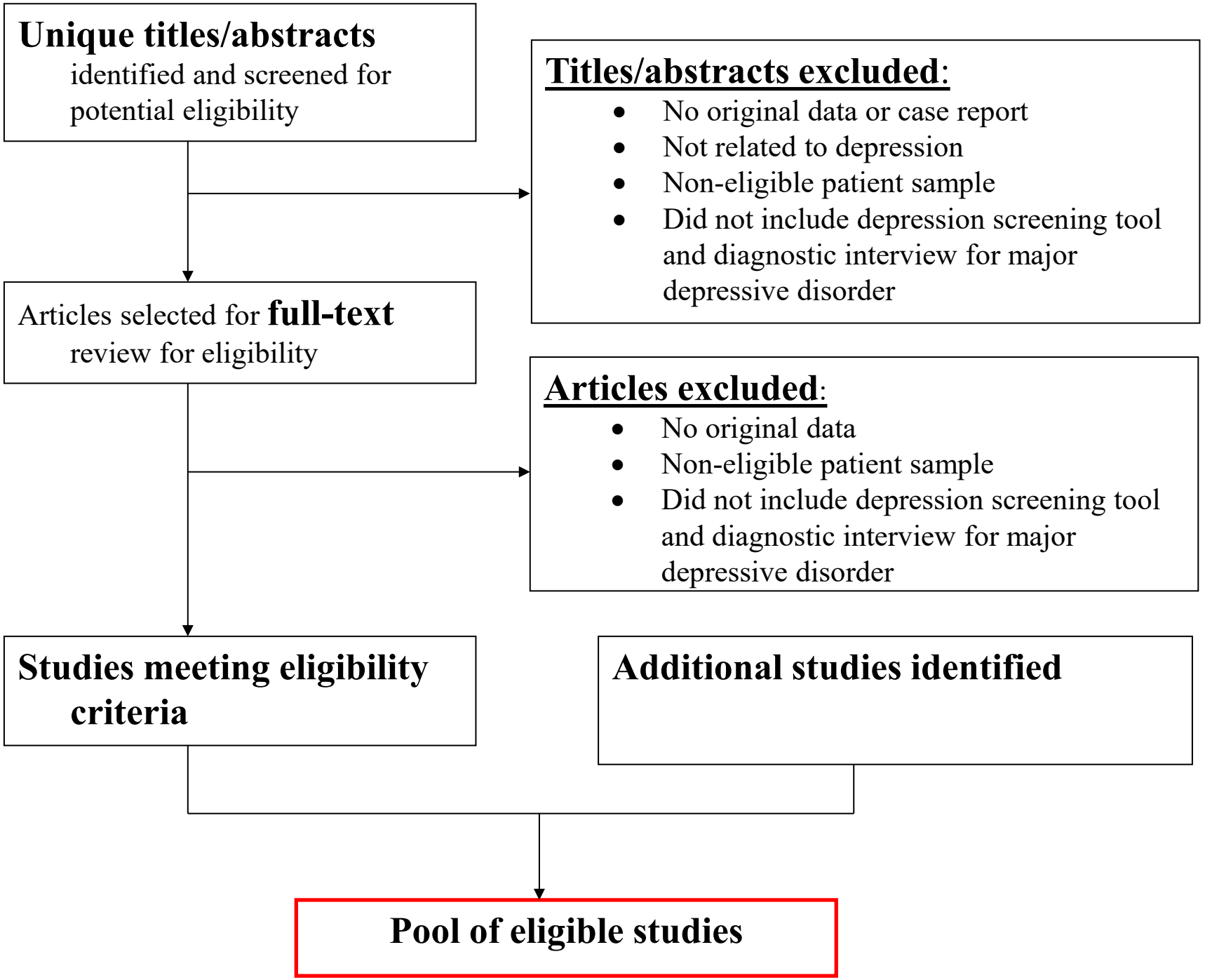


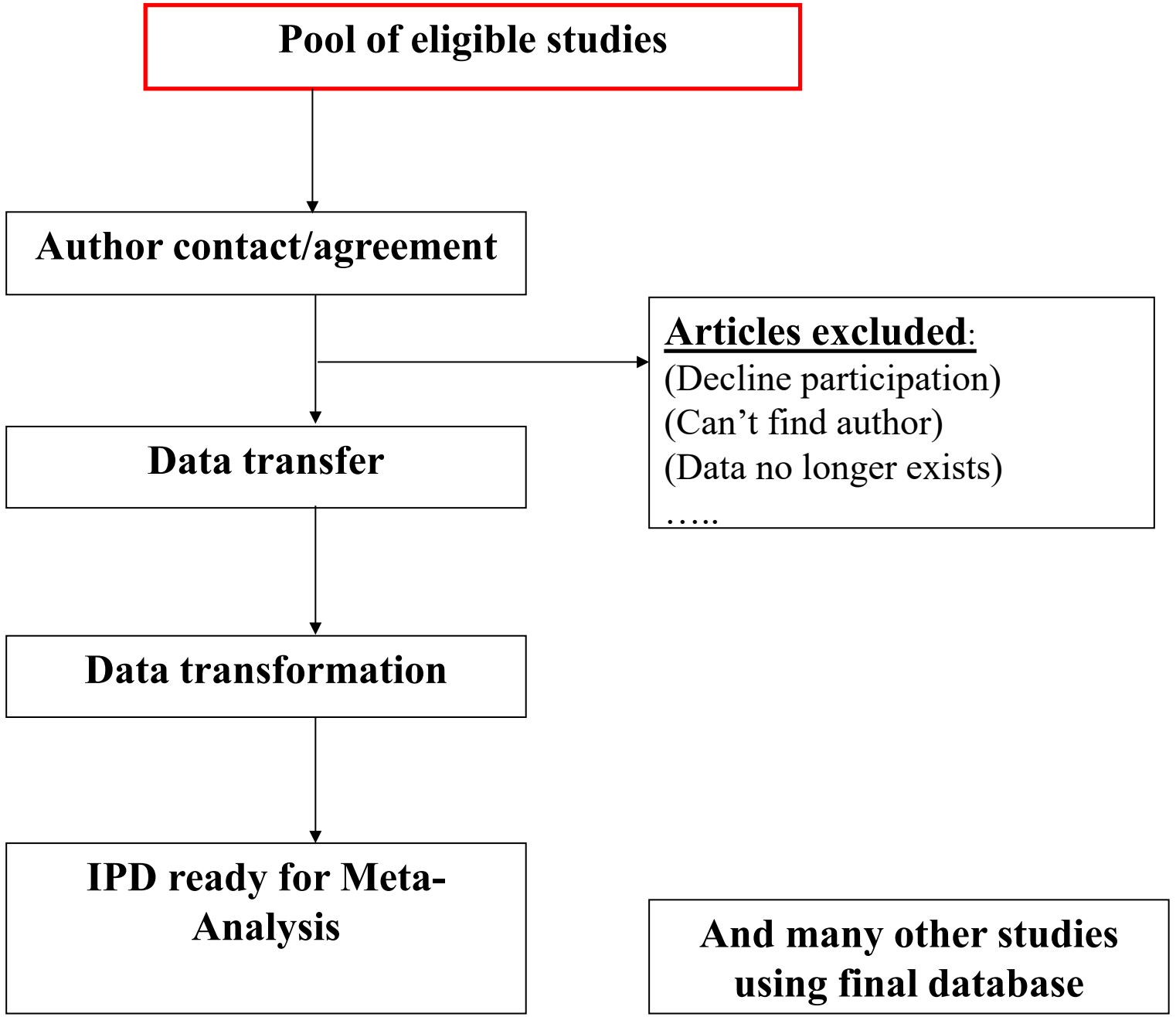
Riley R D et al. BMJ 2010;340:bmj.c221

BMJ



Flow of an IPD Meta-Analysis


Unique titles/abstracts
identified and screened for potential eligibility
Titles/abstracts excluded:
No original data or case report
Not related to depression
Non-eligible patient sample
Did not include depression screening tool and diagnostic interview for major depressive disorder
Articles selected for full-text review for eligibility
Articles excluded:
No original data
Non-eligible patient sample
Did not include depression screening tool and diagnostic interview for major depressive disorder
Studies meeting eligibility criteria
Additional studies identified
Pool of eligible studies


Pool of eligible studies
Author contact/agreement
Articles excluded:
(Decline participation)
(Can't find author)
(Data no longer exists)
…..
Data transfer
Data transformation
IPD ready for Meta-Analysis
And many other studies using final database

IPDMA - Disadvantages

- Time consuming!
- Expensive – very labour intensive!

- More prone than aggregate data MA to excluding studies?
  - You need to get people to agree to participate…
  - But methods exist to combine IPD with Aggregate Data
- Diplomacy needed!
  - Now you are managing a big group of people who may participate in article writing, etc.

IPDMA Advantages - overview

- May reduce heterogeneity
- Better for investigating heterogeneity including subgroup differences
- Very flexible with respect to the type of analyses that can be run
- Builds large collaborative groups

IPDMA can decrease heterogeneity

- **Heterogeneity: treatment effects that vary across studies**
  - Different study populations
  - Different study procedures
- **Can standardize:**
  - inclusion/exclusion criteria
  - exposure & outcome definition
  - investigate the same subgroups
  - the analytic method
    - the model, confounders, missing data, etc.

IPDMA advantages: Subgroup analyses

- Meta-analyses often attempt to explain how patient-level covariates (e.g. age, sex, drug dosage) might modify the treatment effect
  - (i.e. to estimate treatment-covariate interactions).
  - Does the effect of interest differ across subgroups (e.g. older subjects, men, etc.)?

Subgroup analyses in Traditional MA

- Meta-regression
  - Study specific estimates are regressed on the covariate(s) of interest
    - Covariates of interest: study level covariates or aggregated participant level information
    - But… prone to ecological bias, and to confounding from other variables (that are not included in the model)
  - Low power! Since most MA include a small number of studies and …
- Estimating subgroup specific effects
  - e.g. the effect in HIV+ subjects
  - Relies on those effects having been reported by the original studies
  - Often available in few studies

Subgroup effects in IPD-MA

- Not prone to ecological bias
  - because one does not make inferences about individuals based on aggregated data
- Higher power!
  - than meta-regression to detect the effect of an interaction between covariates and treatment
  - power to detect interactions depends on:
    - variation of the covariate within each study in an IPD-MA
    - variation in mean covariate values across studies in AD-MA

Example

- MDR-TB results
- 21 studies included
- Compared surgical + drug vs. drug only

INT J TUBERC LUNG DIS 17(1):6–16

http://dx.doi.org/10.5588/ijtld.12.0198

REVIEW ARTICLE

Surgical interventions for drug-resistant tuberculosis: a systematic review and meta-analysis

M. T. Marrone,*† V. Venkataramanan,* M. Goodman,† A. C. Hill,‡ J. A. Jereb,* S. R. Mase*

*Centers for Disease Control and Prevention, Division of Tuberculosis Elimination, Atlanta, Georgia, †Rollins School of Public Health, Emory University, Atlanta, Georgia, ‡University of California, San Francisco, California, USA

Adjusting for confounders?

- How often was an unadjusted OR used?
- Note: confounding by indication is a major concern here!
- IPD-MA would allow better control of confounding

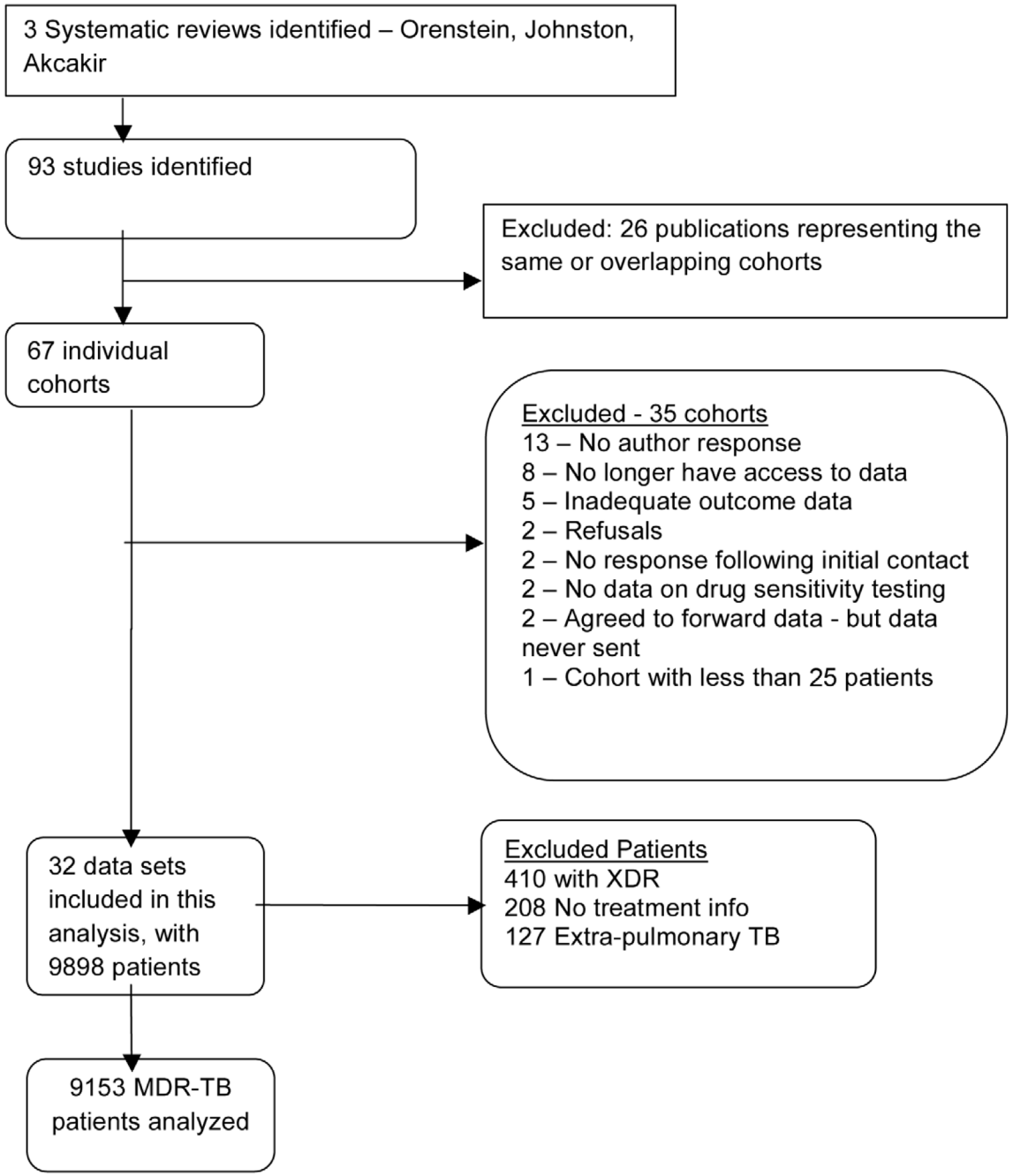
thors of two studies replied when contacted for additional data for the meta-analysis.[12,16] When reported, adjusted ORs comparing successful treatment outcomes between surgical and non-surgical patients were used.

Subgroup analysis for studies reporting treatment outcomes

Study characteristic	MDR-TB treatment outcomes			
	Studies *n*	OR (95%CI)*	$I^{2\dagger}$	*P* value‡
Length of treatment, months				
⩾18	7	2.27 (1.49–3.45)	9	0.36
<18	5	2.09 (1.03–4.24)	55	0.07
Not reported	9	2.41 (1.58–3.66)	17	0.29
Drugs in regimen				
⩾5	10	2.42 (1.80–3.26)	2	0.42
<5	5	2.24 (1.36–3.69)	0	0.67
Not reported	6	2.35 (0.99–5.56)	61	0.03
Average resistance§				
⩾4.7	8	2.50 (1.67–3.74)	0	0.56
<4.7	4	1.90 (1.01–3.55)	26	0.26
Not reported	9	2.33 (1.41–3.84)	46	0.06

Collaborative Group for Meta-Analysis of Individual Patient Data in Multi-Drug Resistant Tuberculosis (IPD-MDRTB) (V1)

- individual patient data on treatment outcomes in MDR-TB patients
  - founded by Dick Menzies
  - 32 observational studies
    - (about half those approached)
  - over 9000 individual patients
  - demographic information (age, sex, country, etc.), medical information (HIV status, previous history of TB, etc.), treatment and treatment outcomes


3 Systematic reviews identified – Orenstein, Johnston, Akcakir
93 studies identified
Excluded: 26 publications representing the same or overlapping cohorts
67 individual cohorts
Excluded - 35 cohorts
13 – No author response
8 – No longer have access to data
5 – Inadequate outcome data
2 – Refusals
2 – No response following initial contact
2 – No data on drug sensitivity testing
2 – Agreed to forward data - but data never sent
1 – Cohort with less than 25 patients
32 data sets included in this analysis, with 9898 patients
Excluded Patients
410 with XDR
208 No treatment info
127 Extra-pulmonary TB
9153 MDR-TB patients analyzed

Primary study results related to treatment effects of various drugs for treating MDR-TB

OPEN ACCESS Freely available online

PLOS | MEDICINE

Multidrug Resistant Pulmonary Tuberculosis Treatment Regimens and Patient Outcomes: An Individual Patient Data Meta-analysis of 9,153 Patients

Shama D. Ahuja[1], David Ashkin[2], Monika Avendano[3], Rita Banerjee[4], Melissa Bauer[5], Jamie N. Bayona[6], Mercedes C. Becerra[7,8], Andrea Benedetti[5], Marcos Burgos[9], Rosella Centis[10], Eward D. Chan[11], Chen-Yuan Chiang[12], Helen Cox[13], Lia D'Ambrosio[10], Kathy DeRiemer[14], Nguyen Huy Dung[15], Donald Enarson[16], Dennis Falzon[17], Katherine Flanagan[18], Jennifer Flood[19], Maria L. Garcia-Garcia[20], Neel Gandhi[21], Reuben M. Granich[17], Maria G. Hollm-Delgado[5], Timothy H. Holtz[22], Michael D. Iseman[23], Leah G. Jarlsberg[24], Salmaan Keshavjee[7], Hye-Ryoun Kim[25], Won-Jung Koh[26], Joey Lancaster[27], Christophe Lange[28], Wiel C. M. de Lange[29], Vaira Leimane[30], Chi Chiu Leung[31], Jiehui Li[32], Dick Menzies[5]*, Giovanni B. Migliori[10], Sergey P. Mishustin[33], Carole D. Mitnick[7], Masa Narita[34], Philly O'Riordan[35], Madhukar Pai[5], Domingo Palmero[36], Seung-kyu Park[37], Geoffrey Pasvol[38], Jose Peña[39], Carlos Pérez-Guzmán[40], Maria I. D. Quelapio[41], Alfredo Ponce-de-Leon[42], Vija Riekstina[30], Jerome Robert[43], Sarah Royce[24], H. Simon Schaaf[44], Kwonjune J. Seung[45], Lena Shah[5], Tae Sun Shim[46], Sonya S. Shin[45], Yuji Shiraishi[47], José Sifuentes-Osornio[48], Giovanni Sotgiu[49], Matthew J. Strand[23], Payam Tabarsi[50], Thelma E. Tupasi[41], Robert van Altena[29], Martie Van der Walt[27], Tjip S. Van der Werf[29], Mario H. Vargas[51], Pirett Viiklepp[52], Janice Westenhouse[53], Wing Wai Yew[54], Jae-Joon Yim[55], on behalf of the Collaborative Group for Meta-Analysis of Individual Patient Data in MDR-TB[¶]

Diplomacy!

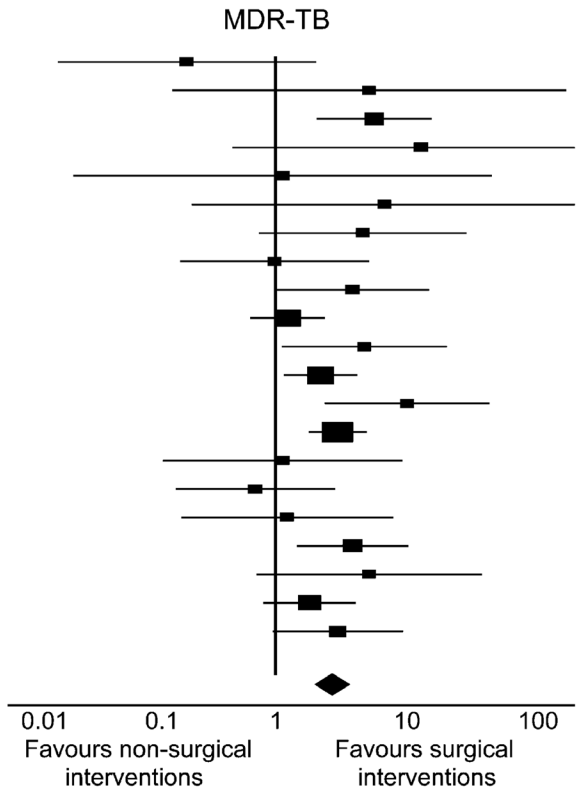
Results

- Treatment success, compared to failure/relapse, was associated with use of:
  - later generation quinolones, (adjusted odds ratio [aOR]: 2.5 [95% CI 1.1–6.0])
  - ofloxacin (aOR: 2.5 [1.6–3.9])
  - ethionamide or prothionamide (aOR: 1.7 [1.3–2.3])
  - use ≥4 likely effective drugs in the initial intensive phase (aOR: 2.3 [1.3–3.9])
  - use ≥3 likely effective drugs in the continuation phase (aOR: 2.7 [1.7–4.1])

Study: Su

- Drug treatment
  poor outcomes
- Surgical lung rese
  therapy
- Results (adjusted
  - Partial lung resection was associated with improved treatment success (OR=3.0, 95%CI: 1.5-5.9)
  - but pneumonectomy was not (OR=1.1, 95%CI: 0.3-2.3)

Avendano et al. 2000[19]	4	6	30	34	0.27	(0.04–1.96)
Burgos et al. 2005[20]	3	3	28	45	4.30	(0.21–88.3)
Chan et al. 2004[21]	99	108	38	54	4.63	(1.89–11.4)
Eker et al. 2008[23]	6	6	99	171	9.47	(0.53–171)
Escudero et al. 2006[24]	2	2	19	23	1.15	(0.05–28.4)
Geerligs et al. 2000[25]	6	6	27	38	5.44	(0.28–105)
Goble et al. 1993[26]	7	9	71	150	3.89	(0.78–19.4)
Jeon et al. 2011[27]	3	8	72	194	1.02	(0.24–4.38)
Keragoz et al. 2009[29]	—	35	—	107	3.33	(1.02–10.9)†
Keshavjee et al. 2008[30]	—	56	—	552	1.24	(0.69–2.24)†
Kwon et al. 2008[12]	19	21	65	106	4.00	(1.11–14.3)
Leimane et al. 2010[32]	62	77	635	950	2.05	(1.15–3.66)
Mishin et al. 2000[33]	14	21	6	29	7.67	(2.14–27.5)
Mitnick et al. 2008[16]	47	87	158	516	2.66	(1.68–4.22)
Narita et al. 2001[34]	3	5	43	76	1.15	(0.18–7.29)
Palmero et al. 2004[35]	5	11	68	130	0.76	(0.22–2.61)
Park et al. 2004[36]	3	6	61	136	1.23	(0.24–6.31)
Shean et al. 2008[37]	21	28	218	460	3.33	(1.39–7.99)
Tao et al. 2006[38]	5	7	18	49	4.31	(0.76–24.5)
Törün et al. 2007[39]	55	66	138	186	1.74	(0.84–3.59)
Yoshiyama et al. 2005[40]	28	34	32	42	2.67	(0.97–7.35)
Summary totals		602		4408	2.27	(1.73–2.97)
Heterogeneity	I^2 = 21%					
FSN	208					



Fox et al. Clin Infect Dis. (2016)
doi: 10.1093/cid/ciw002

Statistical Analysis of IPDMA

- Any approach needs to take into account the clustering of participants by study
  - Participants in the same study are more alike than participants in different studies
    - Similar study procedures
    - Similar patient population
- Similarly to AD-MA: common effects or random
- Two-step or one-step

Notation

- I studies, i=1, … I
- n_i=number of subjects in study i
- θ_i= the TRUE effect measure from study i
- $\widehat{\theta_i}$ = the estimated effect from study i
- $\widehat{S_i}$ = the estimated standard error of $\widehat{\theta_i}$

Common vs. Random

- Common (fixed) effects model:
  - assumes that the estimated effect is the same across all studies ($\theta_i = \theta$)
  - there is little or no heterogeneity
  - Differences across studies due to chance
- Random effects model:
  - assume that θ_i may not all be equal – HETEROGENEITY
  - assumes that the estimated effect varies across the studies
  - due to differences in patient population, study procedures, study quality, etc.

Random effects

- How do we operationalize this?
- Why can't we just estimate the θ_i?
  - Would we have a summary estimate?
  - Problems with the number of parameters?
- Proposal: $\theta_i = \theta + u_i$ where u_i has some distribution
  - EG: $u_i \sim N(0, \tau^2)$

u_i belongs to the study

u_i describes how the study specific effect is above or below the pooled effect

u_i is the random effect

Interpretation

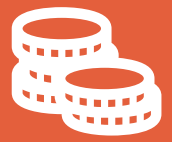
- What does the point estimate mean?
  - Fixed effect: it means the common effect for all studies
  - Random effect: it means the AVERAGE effect – not necessarily the effect in any one study
- Random effect: the effect in the next study is a sample from the normal distribution

Estimating the pooled effect from a IPD MA

- Two step analysis
  - Estimate the effect of interest in each study separately
  - Then, use standard meta-analytic techniques to pool the study specific measures
    - Still benefits from all the advantages related to consistency
- One step analysis
  - use one statistical model, while accounting for the clustering among participants in the same study, to estimate an overall effect
  - fixed and random effects analyses are possible
  - Usually: a mixed model

What else? Heterogeneity

An important output of any MA or IPD-MA



How much does the effect vary across studies?

Why does the effect vary across studies?

Heterogeneity metrics

- τ^2:
  - variance of the random effect
  - Varies from 0 to infinity
  - Interpretability?
- I^2:
  - % heterogeneity
  - Varies from 0-1
  - Interpretability?
- Prediction interval:
  - 95% range of true effects to be expected in similar studies

Why do an IPD meta-analysis?

When a traditional MA is not sufficient to answer your research question because…

Want to do a more fine-tuned analysis

- Additional exclusion criteria
- Subgroups of interest

Suspect bias

- E.g., published reports only showing one part of the story

Want to go beyond the original research question/analysis

- Have alternative/advanced analyses that you would like to apply to answer the same research question
- Have a different research question you could answer using same data

To explore/reduce heterogeneity

IPD-MA Examples

- Collaborative Group for the Meta-Analysis of Individual Patient Data in MDR-TB
  - Observational cohort studies
- Zika Virus Individual Participant Data Consortium
  - Various study designs
- DEPRESSD: an international collaboration to collect depression screening data
  - Diagnostic accuracy studies
  - Each study provides gold standard diagnosis of depression, a screening score, and other participant level information

The DEPRESSD Project

*Evaluate the diagnostic accuracy of the most commonly used depression screening tools

*Diagnostic accuracy studies:
Compare screening questionnaires to a reference standard

*Sensitivity (P[T+ | D+])
*Specificity (P[T- | D-])

Patient Health Questionnaire (PHQ)
105 studies, 46277 participants

Hospital Anxiety and Depression Scale (HADS)
102 studies, 31679 participants

Edinburgh Postnatal Depression Scale (EPDS)
59 studies, 15593 participants

Geriatric Depression Scale (GDS): 21 studies, 5876 participants

Why DEPRESSD?

- Many diagnostic accuracy studies include subjects already diagnosed or being treated for depression
  - Standardize inclusion criteria
- Heterogenous reference standards combined in AD-MA
- Lack of subgroup analyses
- Selective cutoff reporting
  - Only well-performing cutoffs!

Ordinal screening tests & selective cutoff reporting

Negative screen (Likely not depressed)

Cutoff threshold

Positive screen (Likely depressed)

0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27

Screening score

Many cutoff thresholds can be assessed in terms of diagnostic accuracy

Some authors only report cutoffs with high accuracy estimates

Example: conventional meta analysis

CMAJ RESEARCH

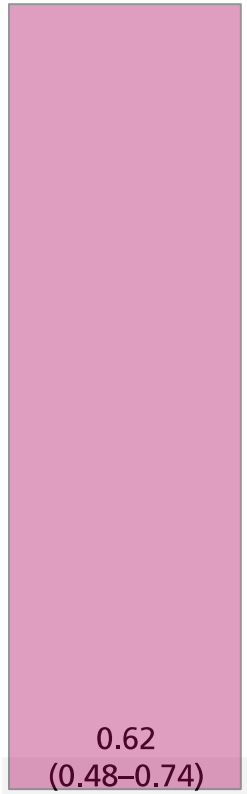
Optimal cut-off score for diagnosing depression with the Patient Health Questionnaire (PHQ-9): a meta-analysis

Laura Manea MSc, Simon Gilbody PhD, Dean McMillan PhD

18 studies

Each study reported sensitivity and specificity for one or more cutoffs

Results from traditional MA



- Sensitivity increases between cutoff 8 to cutoff 11
- For standard cutoff of 10, missing 897 cases (13%)
- For cutoffs of 7-9 and 11, missing 52-58% of data

Manea et al., CMAJ, 2012

What is published?

Study (Author, year)	PHQ-9 cutoff												
	4	5	6	7	8	9	10	11	12	13	14	15	16
Stafford, 2007		O											
Thombs, 2008		O											
Azah, 2005		O											
Lamers, 2008				O									
Lotrakul, 2008						O							
Gjerdingen, 2009							O						
Williams, 2005							O						
Wittkampf, 2009							O						
Osoria, 2009							O	O					
Grafe, 2004								O					
Fann, 2005									O				
Gilbody, 2007									O				
Yeung, 2008												O	

published cutoff O optimal cutoff

Levis et al, AJE 2017

Selective Reporting!

We compared…

Usual aggregate data MA:

- takes available data from every study and pools them at each cutoff
- the number of studies and which studies vary by cutoff

VS

IPD-MA:

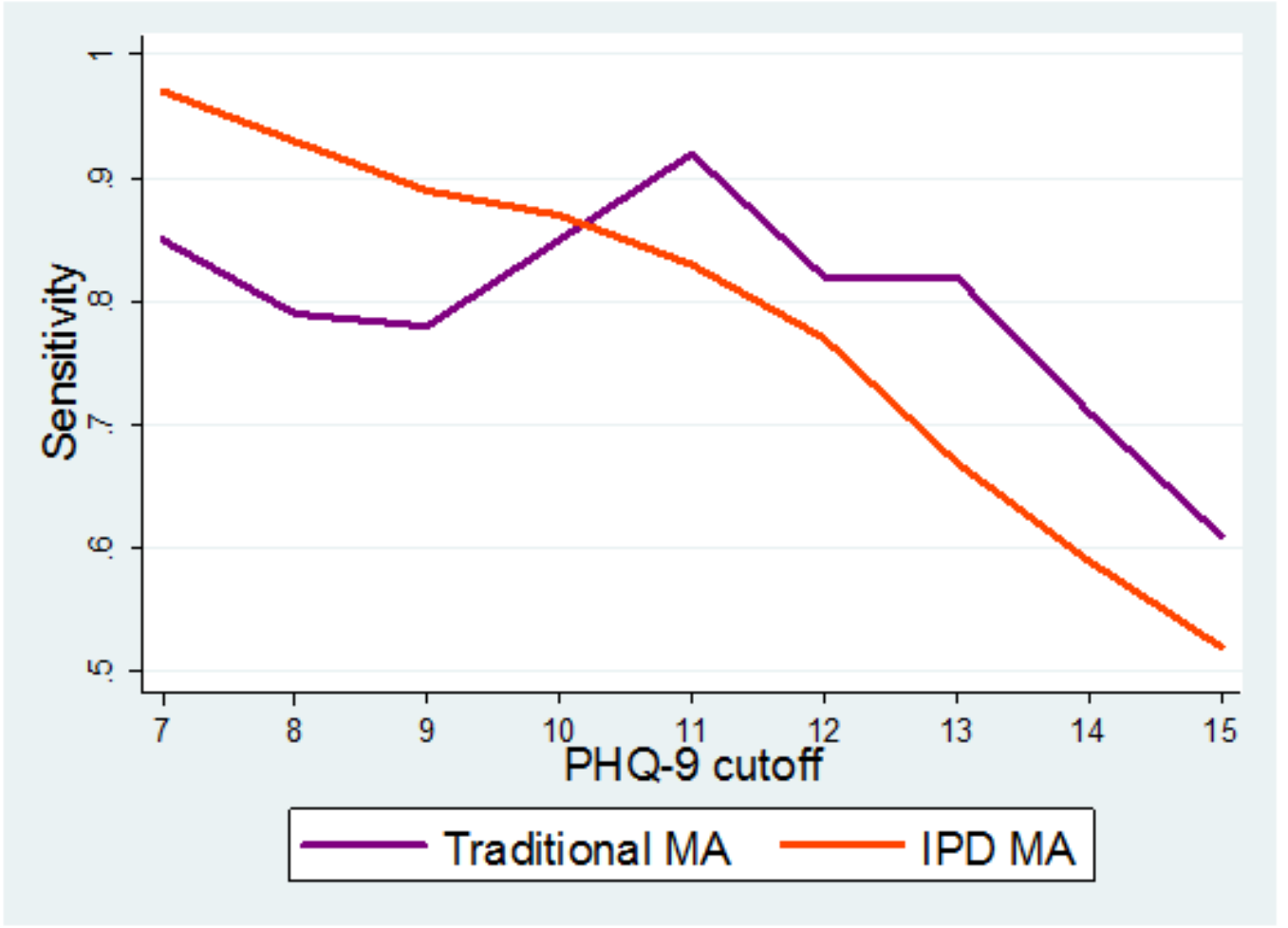
- takes data for all cutoffs
  - *Every* study gives information for *every* cutoff
  - Solves the problem of selective cutoff reporting

What is published?

Table 3. Discrepancies in Sensitivity and Specificity Across Cutoff Points Fro
Screening Studies Published in 2004–2009

Cutoff	% Included in Published Results		Differences Between Analy Analyses Using All D	
			Sensitivity	
	Patients	Cases	Estimate	95% CI
7	46	53	−0.12	−0.30, −0.01
8	46	53	−0.14	−0.33, −0.01
9	34	30	−0.11	−0.37, 0.03
10	83	70	−0.02	−0.09, 0.03
11	27	21	0.09	−0.15, 0.18
12	30	25	0.05	−0.09, 0.16
13	23	18	0.15	0.01, 0.25
14	21	14	0.12	−0.10, 0.25
15	23	19	0.09	−0.09, 0.25

Abbreviations: CI, confidence interval; IPD, individual patient data; PHQ-9, P



Levis et al, AJE 2017

Other results & current preoccupations….

OPEN ACCESS

Check for updates

Accuracy of the Patient Health Questionnaire-9 for screening to detect major depression: updated systematic review and individual participant data meta-analysis

Zelalem F Negeri,[1,2] Brooke Levis,[3] Ying Sun,[1] Chen He,[1] Ankur Krishnan,[1] Yin Wu,[1,4] Parash Mani Bhandari,[1,2] Dipika Neupane,[1,2] Eliana Brehaut,[1] Andrea Benedetti,[2,5,6] Brett D Thombs,[1,2,4,5,7,8,9] on behalf of the Depression Screening Data (DEPRESSD) PHQ Group

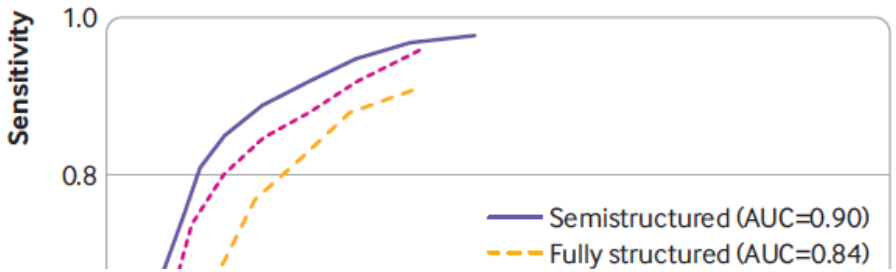


Table 3 | Comparison of sensitivity (95% confidence interval) and specificity (95% confidence interval) estimates among semistructured, full structured, and MINI reference standards

	Semi structured reference standard*		Fully structured reference standard†		MINI reference standard‡	
Cut-off score	Sensitivity (95% CI)	Specificity (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)
5	0.98 (0.95 to 0.99)	0.53 (0.49 to 0.58)	0.91 (0.85 to 0.95)	0.61 (0.51 to 0.69)	0.96 (0.93 to 0.97)	0.60 (0.55 to 0.64)
6	0.97 (0.94 to 0.98)	0.61 (0.57 to 0.65)	0.88 (0.80 to 0.93)	0.69 (0.60 to 0.76)	0.92 (0.89 to 0.95)	0.68 (0.63 to 0.72)
7	0.95 (0.92 to 0.98)	0.68 (0.64 to 0.72)	0.82 (0.73 to 0.89)	0.75 (0.67 to 0.82)	0.88 (0.83 to 0.92)	0.74 (0.70 to 0.78)
8	0.92 (0.88 to 0.95)	0.74 (0.70 to 0.77)	0.77 (0.66 to 0.86)	0.81 (0.74 to 0.86)	0.85 (0.79 to 0.89)	0.80 (0.76 to 0.83)
9	0.89 (0.84 to 0.92)	0.80 (0.76 to 0.82)	0.69 (0.59 to 0.78)	0.85 (0.79 to 0.90)	0.80 (0.73 to 0.85)	0.85 (0.82 to 0.88)
10	0.85 (0.79 to 0.89)	0.85 (0.82 to 0.87)	0.64 (0.53 to 0.74)	0.88 (0.83 to 0.92)	0.74 (0.67 to 0.79)	0.89 (0.86 to 0.91)
11	0.81 (0.75 to 0.86)	0.88 (0.85 to 0.90)	0.57 (0.46 to 0.67)	0.91 (0.87 to 0.94)	0.67 (0.60 to 0.73)	0.91 (0.89 to 0.93)
12	0.75 (0.69 to 0.80)	0.90 (0.88 to 0.92)	0.52 (0.41 to 0.63)	0.93 (0.89 to 0.95)	0.61 (0.54 to 0.68)	0.93 (0.91 to 0.95)
13	0.67 (0.61 to 0.72)	0.93 (0.91 to 0.94)	0.45 (0.35 to 0.56)	0.95 (0.92 to 0.97)	0.55 (0.47 to 0.62)	0.95 (0.93 to 0.96)
14	0.61 (0.55 to 0.67)	0.94 (0.93 to 0.96)	0.39 (0.30 to 0.50)	0.96 (0.94 to 0.97)	0.47 (0.41 to 0.54)	0.96 (0.95 to 0.97)
15	0.52 (0.46 to 0.58)	0.96 (0.94 to 0.97)	0.32 (0.24 to 0.41)	0.97 (0.95 to 0.98)	0.40 (0.35 to 0.46)	0.97 (0.96 to 0.98)

MINI=Mini International Neuropsychiatric Interview.
*Number of studies=47; number of participants=11 234; number of participants with major depression=1528.
†Number of studies=20; number of participants=17 167; number of participants with major depression=1352.
‡Number of studies=33; number of participants=16 102; number of participants with major depression=1661.

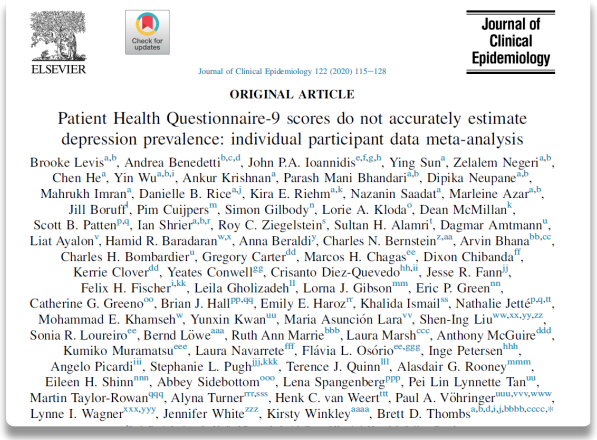
Fig 1 | Flow diagram of study selection process. MINI=Mini International Neuropsychiatric Interview; PHQ=Patient Health Questionnaire

Di
of PHQ

Making these results useable…

- http://www.depressionscreening100.com/phq/

Can we use screening tools to estimate prevalence of depression?



ELSEVIER

Journal of Clinical Epidemiology 122 (2020) 115–128

Journal of Clinical Epidemiology

ORIGINAL ARTICLE

Patient Health Questionnaire-9 scores do not accurately estimate depression prevalence: individual participant data meta-analysis

Brooke Levis[a,b], Andrea Benedetti[b,c,d], John P.A. Ioannidis[e,f,g,h], Ying Sun[a], Zelalem Negeri[a,b], Chen He[a], Yin Wu[a,b,i], Ankur Krishnan[a], Parash Mani Bhandari[a,b], Dipika Neupane[a,b], Mahrukh Imran[a], Danielle B. Rice[a,j], Kira E. Riehm[a,k], Nazanin Saadat[a], Marleine Azar[a,b], Jill Boruff[l], Pim Cuijpers[m], Simon Gilbody[n], Lorie A. Kloda[o], Dean McMillan[k], Scott B. Patten[p,q], Ian Shrier[a,b,r], Roy C. Ziegelstein[s], Sultan H. Alamri[t], Dagmar Amtmann[u], Liat Ayalon[v], Hamid R. Baradaran[w,x], Anna Beraldi[y], Charles N. Bernstein[z,aa], Arvin Bhana[bb,cc], Charles H. Bombardier[a], Gregory Carter[dd], Marcos H. Chagas[ee], Dixon Chibanda[ff], Kerrie Clover[dd], Yeates Conwell[gg], Crisanto Diez-Quevedo[hh,ii], Jesse R. Fann[jj], Felix H. Fischer[j,kk], Leila Gholizadeh[ll], Lorna J. Gibson[mm], Eric P. Green[nn], Catherine G. Greeno[oo], Brian J. Hall[pp,qq], Emily E. Haroz[rr], Khalida Ismail[ss], Nathalie Jetté[p,q,tt], Mohammad E. Khamseh[w], Yunxin Kwan[uu], Maria Asunción Lara[vv], Shen-Ing Liu[ww,xx,yy,zz], Sonia R. Loureiro[ee], Bernd Löwe[aaa], Ruth Ann Marrie[bbb], Laura Marsh[ccc], Anthony McGuire[ddd], Kumiko Muramatsu[eee], Laura Navarrete[fff], Flávia L. Osório[ee,ggg], Inge Petersen[hhh], Angelo Picardi[iii], Stephanie L. Pugh[jjj,kkk], Terence J. Quinn[lll], Alasdair G. Rooney[mmm], Eileen H. Shinn[nnn], Abbey Sidebottom[ooo], Lena Spangenberg[ppp], Pei Lin Lynnette Tan[uu], Martin Taylor-Rowan[qqq], Alyna Turner[rrr,sss], Henk C. van Weert[ttt], Paul A. Vöhringer[uuu,vvv,www], Lynne I. Wagner[xxx,yyy], Jennifer White[zzz], Kirsty Winkley[aaaa], Brett D. Thombs[a,b,d,j,bbbb,cccc,*]

- Comparison of PHQ9$_{\geq 10}$ prevalence and SCID major depression prevalence
- % participants with PHQ9≥10:
  - ranged from 5.3% to 64.8% (in 44 studies)
  - pooled prevalence: 24.6% (95% CI: 20.8%, 28.9%; τ^2: 0.505)
- % participants with SCID major depression:
  - ranged from 0.6% to 56.4%
  - pooled prevalence: 12.1% (95% CI: 9.6%, 15.2%; τ^2 : 0.703)
- Δ prevalence (PHQ9≥10-SCID):
  - ranged from 6.0% to 46.9%
  - pooled difference: 11.9% (95% CI: 9.3%, 14.6%; τ^2 : 0.007)

How else could we have done this?

- Prevalence matching
  - Pooled Δ prevalence (PHQ9≥14-SCID): 0.5% (95% CI: -1.7, 2.6, τ^2 :0.005)
  - But… high heterogeneity!

  - 95% prediction interval: -13.6%, 14.5%
  - ranged from -18.7% to 29.7% (in 44 studies)

Estimating depression prevalence from screening questionnaires

- PHQ-9 ≥ 10 is often used to estimate depression prevalence
- **BUT** it overestimates major depression prevalence substantially!
- too much heterogeneity to correct statistically in individual studies

- Estimates of depression prevalence should be based on validated diagnostic interviews

Small studies, data driven cutoff selection…

- Primary studies on depression screening tool accuracy are small
  - Recent review of 172 studies found median sample size=194 with median number of depression cases=20
  - Most use a data driven approach to identify the optimal cutoff
  - **76% identified an optimal score that differed from the standard score**

- In our IPDMA database on the PHQ:
  - Median sample size (IQR): 194 (134-386)
  - Median number of depression cases (IQR): 28 (14-60)
  - Median optimal cutoff: 10 (range: 3-18)

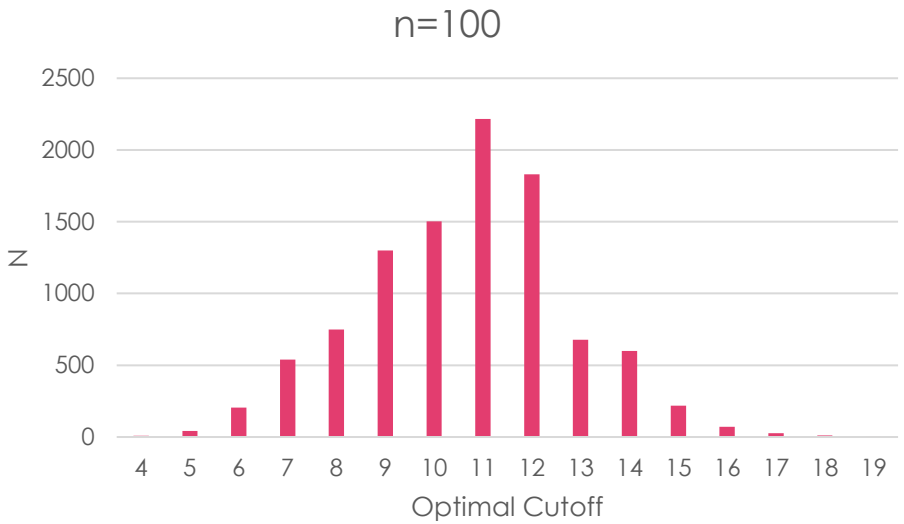
Resampling study

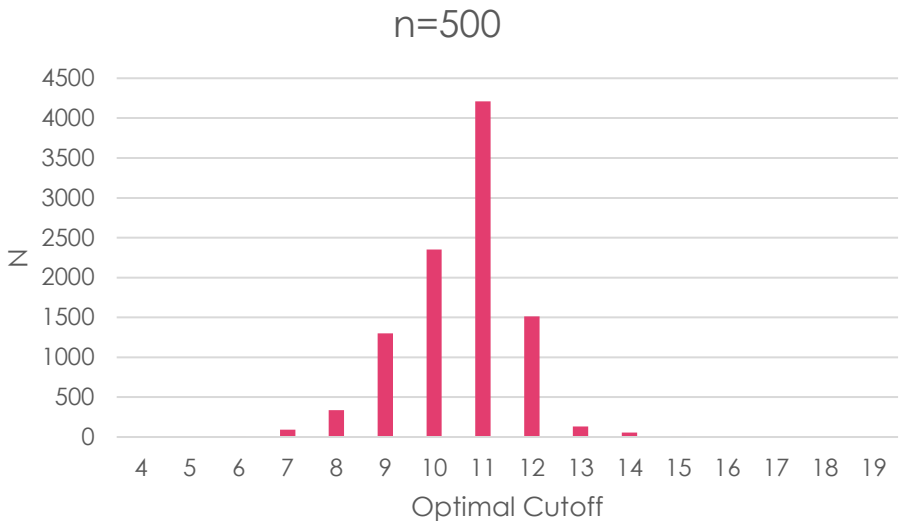
Objectives:

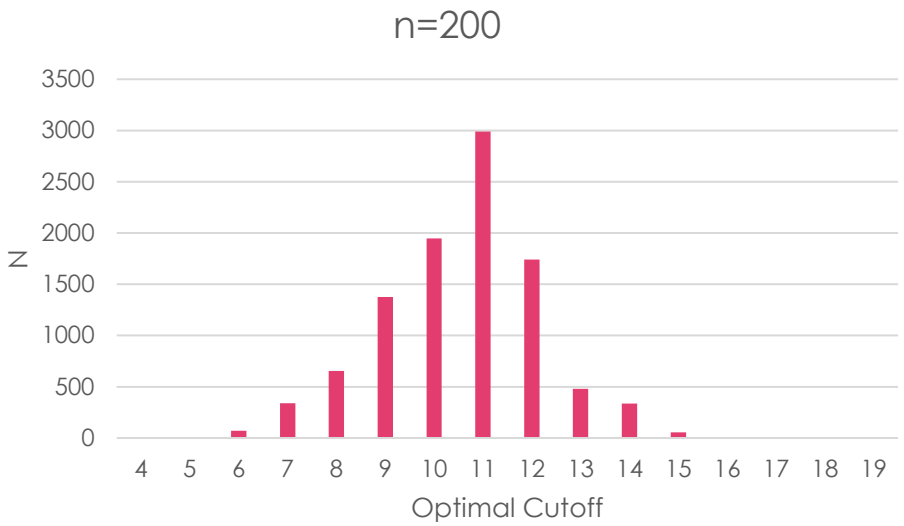
- To demonstrate variability of data-driven optimal cutoffs.
- To estimate bias of accuracy parameters from data-driven optimal cutoffs.
- To evaluate how bootstrap validation could reduce bias of data-driven optimal cutoffs.

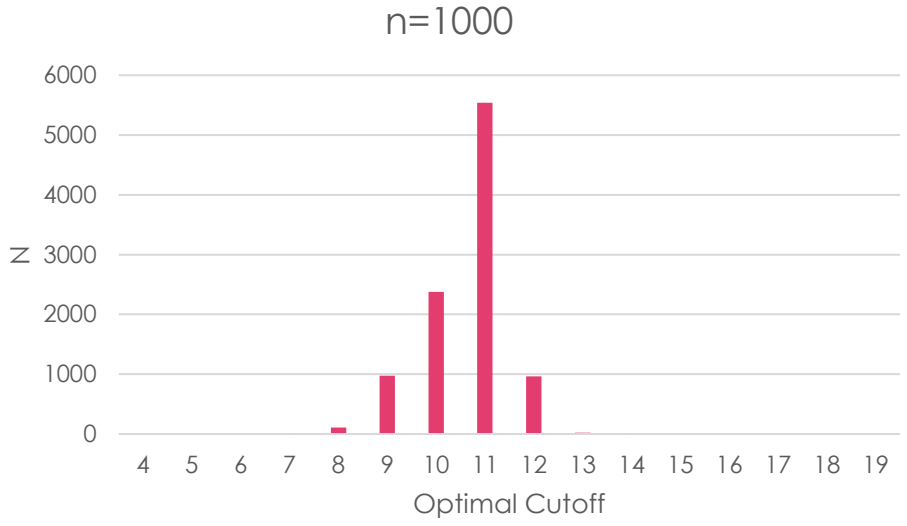
- We considered the IPDMA data a hypothetical population (100 studies, 44503 participants, 4541 cases of depression)
- We drew samples with replacement of size 100, 200, 500 and 1000
- In each sample, we found the cutoff that maximized the sum of sensitivity & specificity (Youden's J)
- We ignored clustering, sampling weights and different reference standards
- Repeated 10000 times

Frequency of Optimal Cutoff by Sample Size









Bias

- We compared sensitivity and specificity at the sample-specific optimal cutoff to sensitivity and specificity in the entire IPDMA at that cutoff

- **Overall bias in accuracy estimates from data-driven optimal cutoffs in 10,000 resampled datasets of size 100, 200, 500 and 1,000**

	Sensitivity				Specificity			
	N = 100	N = 200	N = 500	N = 1000	N = 100	N = 200	N = 500	N = 1000
Bias	7.3%	4.1%	1.8%	1.0%	0.9%	0.6%	0.3%	0.2%

Can we use bootstrap validation to correct for this?

- When building a prediction model:
  - the best way to evaluate it is to use the prediction model on a new data set – to validate it.
- Bootstrap validation:
  - offers an approach to estimate the overfit engendered by estimating and evaluating the prediction model on the same data.
  - And to correct estimates.

- Can we use this approach to get less biased accuracy estimates from data driven optimal cutoffs?

How does the bootstrap validation work?

- Start with a dataset – find the optimal cutoff ci and $sens_{D,ci}$, $spec_{D,ci}$
- Take a bootstrap sample
  - Find the optimal cutoff (**bi**)
  - Estimate sensitivity and specificity in the bootstrap sample at **bi**: $sens_{B,bi}$, $spec_{B,bi}$
  - Estimate sensitivity and specificity in the dataset at bi: $sens_{D,bi}$ $spec_{D,bi}$
- Estimate optimism as:
  - $optimism_{sens} = sens_{B,bi} - sens_{D,bi}$
  - $optimism_{spec} = spec_{B,bi} - spec_{D,bi}$

  …Repeat…
- $sens_{corrected} = sens_{D,ci} - mean\ (optimism_{sens})$
- $spec_{corrected} = spec_{D,ci} - mean\ (optimism_{spec})$

	Sensitivity				Specificity			
	N = 100	N = 200	N = 500	N = 1000	N = 100	N = 200	N = 500	N = 1000
Bias	7.3%	4.1%	1.8%	1.0%	0.9%	0.6%	0.3%	0.2%
Residual bias after correction for optimism	1.9%	0.9%	0.1%	0.1%	0.2%	0.2%	0.1%	0.0%
Correction for optimism	5.5%	3.3%	0.9%	0.9%	0.6%	0.5%	0.3%	0.2%

Conclusions

- Beware small samples!
- Often, primary studies conclude that alternate cutoffs are necessary in their "special" population
- Probably, more likely due to small samples.
- Boostrap validation can help reduce bias in diagnostic accuracy estimates in small samples.

Other cool things we have done:

- Diagnostic accuracy for other versions of the PHQ (PHQ2, algorithm-based, etc.)
- Equivalence of PHQ8 and PHQ9
- Minimal clinically important differences
- Differences in reference standards
- Modeling approaches vs IPDMA to avoid bias from selective cutoff reporting

What next?

- Head-to-head comparisons of some screening questionnaires
- Anxiety!

Acknowledgements

Dr. Brett Thombs

&

Many students

Fonds de recherche Québec

CIHR IRSC

Canadian Institutes of Health Research | Instituts de recherche en santé du Canada