Data-sharing and trustworthiness of trials evaluating cervical ripening in induction of labour: a meta-epidemiological study of randomised controlled trials

Malitha Patabendige,[a,b,*] Daniel Lorber Rolnik,[a,b] Wanlin Li,[a] Ben Willem Mol,[a,b,d,e,**] and Wentao Li[c,a,d]

[a]Department of Obstetrics and Gynaecology, Monash University, Monash Medical Centre, Clayton, Victoria, 3168, Australia
[b]Women's and Newborn, Monash Health, Victoria, Australia
[c]National Perinatal Epidemiology and Statistics Unit (NPESU), Centre for Big Data Research in Health, and School of Clinical Medicine, Faculty of Medicine, University of New South Wales, Sydney, Australia

eClinicalMedicine
2025;85: 103346
Published Online 8 July 2025
https://doi.org/10.1016/j.eclinm.2025.103346

Summary

Background Induction of labour is a common medical intervention in obstetrics, but its uptake is highly variable in well-resourced and low-resourced settings. Cervical ripening is the initial phase of prepping the cervix when it is not 'ready'. Approximately 30% of women are induced in high-income countries, while this proportion is generally lower in low-resource settings. One way to assess the trustworthiness of randomised controlled trials (RCTs) is by examining individual participant data (IPD), but data sharing remains a significant challenge. It is still unclear whether there are substantial differences in trial characteristics, trustworthiness, effect sizes, and the certainty of evidence between RCTs that share data and those that do not.

Methods In this meta-epidemiological study of RCTs comparing different methods of induction of labour, we included nine IPD meta-analyses. RCTs in the shared and non-shared groups were assessed on the following criteria: trial characteristics, trial registration, trustworthiness (independently evaluated via the Trustworthiness in RAndomised Controlled Trials [TRACT] checklist by two investigators), excessive similarity or difference in baseline characteristics beyond what would be expected by chance, and statistical analysis results that cannot be reproduced with summary data. The TRACT checklist aims to identify and triage RCTs at risk of trustworthiness issues, and includes seven domains: governance, author group, plausibility of intervention usage, timeframe, drop-out rates, baseline characteristics, and outcomes. We performed random-effects meta-analyses separately for the two groups, followed by the GRADE approach, and compared their effect estimates using ratio of odds ratios (ROR). This study was prospectively registered at the Center for Open Science OSF registries network (WU93A) before data extraction.

Findings Of 265 eligible RCTs (65,115 women), 64 (24.2%) trials shared data, while 201 (75.8%) did not. Adequate trial registration (after 2010) was found in 44.0% (22/50) of shared RCTs vs 14.1% of (14/99) non-shared RCTs ($p < 0.001$). In the shared group, 84.4% (54/64) were considered to have no trustworthiness concerns, compared to 49.1% (87/177) in the non-shared RCTs ($p < 0.001$). Risk of bias assessment revealed that 20.3% (13/64) of shared RCTs had an overall high risk, compared to 28.4% (57/201) of non-shared RCTs ($p = 0.20$). In the shared RCTs, simulation-generated p-value distributions of baseline characteristics were likely consistent with the expected uniform distribution ($p = 0.50$), but not in the non-shared RCTs ($p = 0.006$). Proportions of RCTs with at least one inconsistently reported p-value that cannot be reproduced with summary data were not significantly different between shared and non-shared groups (for baseline characteristics: 17.2% vs 25.4%; for caesarean delivery: 42.2% vs 30.8%; for uterine hyperstimulation: 17.2% vs 12.9%, respectively). Non-shared RCTs showed exaggerated effect estimates compared to the shared group, with statistically significant RORs ranging from 2.36 to 1.29 in three out of nine induction of labour comparisons. The GRADE assessment showed higher certainty of evidence for nine effect estimates and equal certainty for seven when comparing the shared groups to non-shared groups across all IPD meta-analysis projects on caesarean section and hyperstimulation. This was more pronounced among RCTs without trustworthiness concerns (than in RCTs with trustworthiness concerns), demonstrating greater certainty of evidence of RCTs without such trustworthiness issues.

*Corresponding author. Department of Obstetrics and Gynaecology, Monash University, Monash Medical Centre, Clayton, Victoria, 3168, Australia.
**Corresponding author. Department of Obstetrics and Gynaecology, Monash University, Monash Medical Centre, Clayton, Victoria, 3168, Australia.
E-mail addresses: malitha.madavitapatabendige@monash.edu (M. Patabendige), ben.mol@monash.edu (B.W. Mol), daniel.rolnik@monash.edu (D.L. Rolnik), wanlin.Li@monash.edu (W. Li), wt.li@unsw.edu.au (W. Li).
[d]Contributed equally.
[e]Full professor.

Interpretation RCTs on labour induction without IPD-sharing are more likely to have lower quality, more trustworthiness concerns, and exaggerated effect estimates than those with shared IPD. IPD-sharing and IPD meta-analyses should be encouraged. Trustworthiness and quality assessment should be prioritised whenever using RCTs for evidence synthesis and clinical guidelines development to ensure better evidence informs clinical practice.

Funding This study is supported by three NHMRC Investigator Grants for BWM, WenL, and DLR.

Keywords: Induction of labour; Cervical ripening; Trustworthiness; Quality of research; Data sharing

Research in context

Evidence before this study

Methods and strategies used during the conduct of randomised controlled trials (RCT) minimise the risk of bias, thus providing evidence that is believed to be close to the true effect. RCTs, either directly or through systematic reviews and meta-analyses, inform clinical guidelines and clinical practice. However, there is increasing concern about the trustworthiness of some RCTs. We searched the Ovid Medline database for the last 10 years, from January 1, 2014 to April 1, 2025, for articles published in English using the search terms "quality of research" OR "trustworthiness" AND "individual participant data" OR "data sharing" to understand the use of individual participant data, data-sharing and its impact on the trustworthiness of evidence. Studies or reviews that focused explicitly on data sharing and trustworthiness were identified. We identified 19 eligible studies out of 94 search results, spanning various medical disciplines, but no studies evaluating labour induction. Overall, the literature suggests that de-identified shared data are valuable for generating new research, validating original results, and conducting evidence synthesis through individual participant data meta-analyses. It also helps identify problematic studies, conduct subgroup analyses, and offers greater statistical flexibility. Sharing de-identified data is crucial for adding value to original research and minimising study repetition. There is a dearth of studies examining the actual impact of data sharing on the results of evidence synthesis.

Added value of this study

We performed a head-head comparison of a large number of RCTs evaluating cervical ripening before induction of labour, stratified by data-sharing practices, evaluating their methodological characteristics, risk of bias, GRADE approach, trial registration, trustworthiness (assessed via the Trustworthiness in RAndomised Controlled Trials [TRACT] checklist), outcome effect estimates, randomisation, and statistical analysis results that cannot be reproduced with summary data. Our findings indicate that RCTs that do not share individual participant data have lower quality, more trustworthiness concerns, are less likely to be properly randomised, and have poorer reproducibility than those with data-sharing. Non-shared RCTs exhibited larger treatment effects compared to the shared group, potentially leading to exaggerated effect estimates in meta-analyses. The GRADE assessment revealed more or equal certainty of evidence in the shared groups compared to non-shared groups, which was more evident in RCTs without trustworthiness concerns. This study addresses the gap in understanding how data-sharing is associated with the quality and trustworthiness of RCTs and how it affects evidence synthesis. The findings provide further rationale for promoting data-sharing practices to enhance the quality and reliability of medical research.

Implications of all the available evidence

Among RCTs evaluating methods of cervical ripening before induction of labour, the availability of individual participant data and willingness to share are strong indicators of reproducibility, quality, and trustworthiness. Encouraging data-sharing and individual participant data meta-analyses can lead to higher-quality and more trustworthy data, provide opportunities to identify problematic RCTs, and transform the evidence synthesis ecosystem in the induction of labour. Assessment of trustworthiness and quality should be a priority when using RCTs for systematic reviews, meta-analyses, and clinical guidelines, to ensure that better evidence informs clinical practice. As research informs clinical care, future research into the detection and management of trustworthiness issues in RCTs is urgently needed to restore confidence in trial findings and the evidence synthesis they underpin.

Introduction

Well-conducted randomised controlled trials (RCTs) are important because they are considered the most rigorous research design for evaluating the effectiveness of healthcare interventions.[1] Compared to other study designs assessing interventions, findings yielded by RCTs are considered closer to the true effects of treatments. These RCTs, either directly or through systematic reviews and meta-analyses, inform guidelines and clinical practice.

However, there is growing evidence raising concerns about data integrity in RCTs, with current findings potentially representing just the '*tip of the iceberg*'.[2,3] More importantly, these compromised RCTs have affected evidence synthesis. For example, meta-analyses have suggested benefits of sildenafil for various pregnancy complications, but these conclusions are based on studies that have been retracted or identified as problematic.[4–6] Similarly, a meta-analysis on the use of foetal pillow during caesarean delivery was retracted following the earlier retraction of an underlying RCT.[7,8] This issue extends beyond women's health. In the context of COVID-19, several systematic reviews suggested that ivermectin reduced mortality, only for it to be revealed that these systematic reviews included fabricated RCTs and recently retracted RCTs on hydroxychloroquine and azithromycin.[9,10] Other examples are vitamin K for fracture prevention and tranexamic acid for postpartum haemorrhage prevention.[11,12]

The Cochrane defines a problematic, or untrustworthy, study as 'any published or unpublished study where there are serious questions about the trustworthiness of the data or findings, regardless of whether the study has been formally retracted'.[13] In practice, we consider research trustworthy when it is true, based on real events, conducted honestly and ethically, with transparency at all stages of the research life cycle: proposal, conduct, analysis, reporting, and post-publication. Criteria for trustworthiness include statistical checks of data and results, aspects of research governance such as ethical approval, the presence of plagiarised content, the plausibility of the study conduct, and the track record of the research team.[14]

The current recommended approach of systematic reviews is all-inclusive, without a routine to scrutinise the trustworthiness of underlying RCTs, potentially leading to compromised systematic reviews. Importantly, the risk of bias tool, which assesses the quality of studies, cannot effectively assess trustworthiness issues. This all-inclusive approach perpetuates the survival of problematic RCTs in evidence-based medicine. Unlike traditional systematic reviews that rely on data extracted from published studies and are subject to the methodological limitations of aggregate data, individual participant data (IPD) meta-analysis assesses and uses raw data. IPD meta-analysis can more easily identify problematic studies, allow for subgroup analyses, offer wider statistical flexibility, and harmonise exposures and outcomes.[15] Therefore, IPD sharing is increasingly recognised as a major requirement of clinical research culture.[16] However, despite clinical trial authors claiming a high willingness to share IPD, actual IPD-sharing remains suboptimal.[17] A knowledge gap exists on whether IPD-sharing represents better quality and trustworthiness of RCTs. Also, it is important to understand how the inclusion of RCTs with trustworthiness concerns affects the results of evidence synthesis. This understanding is crucial to determine whether future evidence synthesis should move towards the IPD approach.

We obtained IPD from RCTs included in IPD meta-analyses evaluating the effectiveness and safety of various methods of induction of labour, compiling a large, homogeneous dataset.[18–24] As a widely performed obstetric procedure, induction of labour provides a relevant case study to explore these knowledge gaps. Notably, a significant number of trialists did not share IPD for various reasons. Therefore, we aimed to assess whether RCTs that provided IPD ('IPD provided') differed from those that did not ('IPD not provided') in terms of methodological rigour, trustworthiness concerns, effect estimates, and certainty of evidence.

Methods

Study design

This meta-epidemiological study was prospectively registered at the Center for Open Science OSF registries network (available at https://doi.org/10.17605/OSF.IO/WU93A) before data extraction.[25] The study was conducted according to a pre-defined protocol (Appendix 1). Ethical approval was granted by the Monash Health Human Research Ethics Committee (RES-23-0000-017Q), Victoria, Australia, on 13/01/2023, with no requirement for written informed consent from study participants.

Data sources

We built on a series of IPD meta-analyses comparing the effectiveness and safety of various induction of labour methods (Supplementary Table S1). Six of these IPD meta-analyses have been published.[18–24] The others are ongoing and have completed author enquiry for IPD requests.[26–28] We conducted an exhaustive inquiry to gain access to IPD. Our approach included sequential outreach to corresponding authors, trial co-authors, co-authors of recent publications, institutions, academic contacts in various countries, and journal editors. Initially, invitations were sent to corresponding authors every fortnight for the first few months. Thereafter, all other potential sources were regularly contacted for at least one year, ensuring a thorough process until a consensus-based conclusion was reached for cases of non-response. Apart from emails, we have utilised all available means to contact trial authors, including phone, messaging apps, and other digital channels, such as social media platforms.

Data collection

We extracted information from the publications on various aspects, including authors, year of publication, country, randomisation process, summary statistics of baseline characteristics and univariable analyses, outcome effect estimates and other important characteristics of RCT design. These characteristics include the number of participants and study centres, study arms, conclusions, statement of ethics approval with details about consent, description of participant recruitment, reporting of dropouts, timeline, use of a flow chart and data sharing statement. The following assessments and analyses were carried out separately in the shared and non-shared groups of RCTs. More details of these assessments can be found in previous publications by our group and others.[29–32] The first author, MP, was directly involved in seeking IPD for 6 of 9 IPD meta-analysis projects.[19,21–24,26,33] For the other 3 IPD meta-analyses where the author MP was not directly involved,[18,20,21] new requests were not sent as the consensus was reached after a minimum of one year of a well-structured (for more than two years for most RCTs) IPD-seeking exercise under the supervision of the same senior researchers for all these projects (BWM and WenL).

Statistical analysis

Characteristics of RCTs

We compared geographical regions, data-sharing statements, and additional concerns, such as the absence of a statement confirming ethics approval, use of the CONSORT (Consolidated Standards of Reporting Trials) flow diagram, description of informed written consent, and recruitment or study timelines that are not feasible between the start and end of recruitment. We also looked for any other internal inconsistencies in the publications.

Trial registration

The International Committee of Medical Journal Editors (ICMJE) introduced in 2004 a requirement for the prospective registration of clinical trials for consideration for publication in its member journals, applying to any clinical trial starting enrolment after July 1, 2005.[34] This policy was required by most of the journals after 2010. Therefore, RCTs published after January 1, 2010, were included in the assessment of the trial registration criterion. Trial registration status was categorised into one of four groups, with registrations occurring: 1. prior to commencing recruitment or within two weeks of initiating recruitment (adequate), 2. after two weeks from starting recruitment but before the completion of recruitment (late), 3. after completing recruitment (retrospective), 4. not registered at all (absent).

Trustworthiness assessment

We used the Trustworthiness in RAndomised Controlled Trials (TRACT) checklist (Appendix 2) to evaluate RCTs' trustworthiness.[35] This screening tool aims to identify and triage RCTs at risk of trustworthiness issues. The checklist includes seven domains: governance, author group, plausibility of intervention usage, timeframe, drop-out rates, baseline characteristics, and outcomes. Two investigators (MP, WanL) independently assessed RCTs published in full-text form and disagreements were resolved by consensus. The first author, MP, primarily conducted the trustworthiness assessment for all the RCTs independently. Only the second reviewer (WanL) was blinded to the data sharing status in the trustworthiness assessment. This was followed by a consensus in cases of disagreement. Item-specific discrepancy rate was around 30% (3 in every 10 RCTs), and there were no major disagreements on overall judgement. We adopted a scoring system for the TRACT checklist items, assigning scores as follows: No Concerns: 0, Some Concerns/No Information: 1, and Major Concerns: 2. These scores were summed up to produce a total score for the RCT. An RCT was categorised as 'not-meeting trustworthiness criteria'; if: 1) Its total TRACT score was 4 or more and above the median of the total score of RCTs for the induction of labour comparison of interest, or 2) There were one or more major concerns for specific TRACT checklist 'veto power' items including 3.1, 4.2, 6.2, 6.3, and 7.2 items (Appendix 2). Otherwise, it was categorised as 'meeting trustworthiness criteria'. If an RCT with shared IPD was initially categorised as 'not-meeting trustworthiness criteria' according to TRACT but a check on the IPD was satisfactory, it was recategorised as 'meeting trustworthiness criteria' or vice versa. This process has also been summarised as an algorithm in the Supplementary Figure S1.

As shown in the algorithm (Supplementary Figure S1), each RCT was initially assessed independently of its comparison group. However, since TRACT scores varied across comparisons, we accounted for this by requiring a score above the comparison-specific median and a minimum total score of 4 to classify a trial as 'not meeting trustworthiness criteria'. To reduce false negatives, we applied 'veto power' items - high-weight TRACT criteria indicating serious concerns. This combined approach balanced the risks of false positives and negatives, similar to diagnostic decision-making. For this IPD check, the received data were examined for missing data, errors, internal consistency, consistency with the publication, reproducing the baseline characteristics table, pattern of treatment allocation and data presentation. The identified issues were communicated to the RCT investigators for a solution. If the response was not resolving the issues we found, RCT remained categorised as 'not-meeting trustworthiness criteria' or vice versa.

Risk of bias assessment

The Cochrane risk of bias tool version 2.0 (RoB 2.0) was applied for all the RCTs.[36] The risk of bias was independently assessed by two or more researchers unblinded to data-sharing status (MP for all RCTs and separately by the researchers managing each IPD meta-analysis project). However, these reviewers were blinded to the outcomes of the trustworthiness assessment. Research group consensus was reached in case of a discrepancy to minimise observer variations. The discrepancy rate was around 10% (1 in every 10 RCTs) for the risk of bias assessment. Apart from comparing shared and non-shared groups, we also compared RCTs according to trustworthiness categorisation.

Probability of random sampling according to baseline characteristics

Monte Carlo simulations (computational algorithms that use random sampling to generate numerical results) were used to generate p-values for excessive similarity or differences in baseline continuous variables between randomised groups. This method has been used previously by Carlisle et al., 2015 and similar works by our team.[29–32] According to Carlisle's works, the resulting distribution of p-values for individual baseline variables closely resembles the distribution expected from randomly sampled, normally distributed, continuous variables, resulting in a uniform [0, 1] distribution.[31,32] A dominance of simulation-generated p-values close to either 0 or 1 would cast doubt on the assumption that the randomisation was appropriately performed. Using the Kolmogorov–Smirnov test, we compared the distribution of simulation-generated p-values of continuous baseline variables against the expected uniform distribution in the shared and non-shared groups separately. We also plotted the cumulative distribution of simulation-generated p-values for baseline variables. Analysis was further stratified by trustworthiness categorisation. Whenever possible, subgroup analyses were performed in the non-shared group according to the reasons for not sharing IPD (data loss, not willing, and other reasons).

Statistical analysis results that cannot be reproduced with summary data

We attempted to reproduce univariable statistical analyses using the summary statistics and methods provided in publications. For comparing the means of continuous baseline variables, we used independent-sample t-tests and one-way analysis of variance (ANOVA), reproducing the p values with the corresponding means, standard deviations (with or without equal variances), and the number of participants reported in the tables and texts. We also reproduced analyses using the absolute numbers in tables presenting the cross-tabulation of categorical data, using Pearson's Chi-squared or Fisher's Exact tests, and calculating unadjusted odds ratios and risk ratios. We evaluated the consistency between the computed results and those published, reporting inconsistencies only when calculated p-values deviated from the reported values by more than 0.03 to consider the possible rounding of the reported summary data in the publications. Additionally, we plotted the cumulative distribution of a ratio (calculated p-value/reported p-value) for these inconsistencies (p-values that cannot be reproduced with summary data), separating baseline characteristics and outcomes in the shared and non-shared groups, as well as trustworthiness assessment groups.

Data synthesis for outcome effect estimates

Random effects models using aggregate data were applied for two pre-defined outcomes (caesarean delivery and uterine hyperstimulation) in the shared and non-shared RCT groups across nine comparisons of different methods of induction of labour. Analysis was further stratified by trustworthiness categorisation. These two outcomes are commonly used in all nine IPD meta-analyses and are clinically important in assessing the effectiveness and safety of induction of labour. Both are undesirable, and a lower incidence indicates improved effectiveness or safety of induction of labour methods. Odds ratios with 95% confidence intervals (CI) were computed along with statistical heterogeneity using the I^2 statistic. All analyses were performed following the intention-to-treat principle. Odds ratios between shared and non-shared groups were compared using ratios of odds ratios (ROR), as shown by Tsujimoto et al., 2019. R coding for ROR and 95% CIs have been added as Appendix 3.[37] We applied the GRADE (Grading of Recommendations, Assessment, Development and Evaluations) approach for the pre-defined outcomes in the aggregate data meta-analyses.[38,39] A post-hoc sensitivity analysis was conducted by categorising RCTs based on their publication date: before 2010 versus on or after 2010 and comparing shared and non-shared groups.

Statistical analyses were performed using the R statistical environment (R Foundation for Statistical Computing, Vienna, Austria).[40] The "meta" package was used for aggregate data meta-analysis, and the 'sim_distr' package was used for Monte Carlo simulations.[31,41,42] Pearson's Chi-squared or Fisher's Exact tests were used to compare categorical data and Mann–Whitney U test was used to compare continuous variables between the shared and non-shared groups.

Patient and public involvement

This study used deidentified and aggregated data provided in the published randomised controlled trials or trial registries studied. No patients or the public were involved in the development of the research question, outcome measures, study design, or implementation of the study, as it is meta-epidemiological research. Although no patients or the public were directly involved in this paper, the research idea was generated directly from the impact on the relevant clinical context. Results will be disseminated to the public through media and a press release written using layman's terms.

Role of the funding source

The funder of the study had no role in the study design, data collection, data analysis, interpretation of results, or writing of the report.

Results

We identified 265 RCTs, including 65,115 women, in the nine IPD meta-analysis projects for induction of labour methods (range 7–60 RCTs per project). Characteristics of RCTs evaluated in this study are shown in Table 1. IPD were provided from 64 RCTs (shared group, 24.2%) and not provided from 201 RCTs (non-shared group, 75.8%). The reasons for not sharing IPD are given in Supplementary Table S1. In the non-shared group, only 83 of 201 (41.3%) responded, of whom 16 (8.0%) authors were not willing to share, 46 (22.9%) claimed that the data have been destroyed, and 21 (10.4%) did not respond after an initial positive response. The median (range) number of participants per RCT in the shared and non-shared groups were 204.0 (126.7–363.0) and 132.5 (100.0–210.0), respectively ($p < 0.001$). Recent RCTs are more likely to share IPD than older ones (Supplementary Figure S1). As shown in Table 1, trial characteristics, including ethical approval statement, consent, description of participant recruitment, reporting of dropouts, use of a study flow chart and data-sharing statement, were more likely to be unsatisfactory in the non-shared group. Among shared RCTs, Europe contributed the most, accounting for 28.1% (18/64). In contrast, South-East Asia had the highest number of non-shared RCTs, making up 31.3% (63/201). Characteristics of non-shared RCTs with regard to the publication format (full-text and abstracts/unpublished) and trustworthiness category are shown in Table 2. Non-shared RCTs with trustworthiness concerns were more likely to have unsatisfactory trial characteristics than were RCTs without trustworthiness concerns.

Adequate trial registration (assessed only for studies published on or after 2010) was found in 22/50 (44.0%) of RCTs in the shared group versus 14/99 (14.1%) in the non-shared group ($p < 0.001$).

Twenty-nine abstracts (7 shared and 22 non-shared) and four unpublished (two each for shared and non-shared) RCTs were not included in the trustworthiness assessment, leaving 232 RCTs published in full-text form (55 shared and 177 non-shared RCTs) for trustworthiness assessment using TRACT checklist (Supplementary Table S2). In the shared group, 4/55 (7.3%) RCTs did not meet pre-defined trustworthiness criteria based on the TRACT checklist.[43–46] In the non-shared group, 87/177 (49.1%) RCTs met pre-defined trustworthiness criteria, while 90/177 (50.9%) did not (Table 1). The median (range) TRACT score was 4 (0–8) for the shared group and 8[5–13] for the non-shared group ($p < 0.001$), with a higher score expressing more trustworthiness concerns.

Following the TRACT assessment, an IPD check was performed in the shared group, a further six RCTs which had acceptable TRACT assessment, were identified as having trustworthiness concerns based on IPD checking.[47–52] Nine shared RCTs (seven published as abstracts and two unpublished reports) were considered to be without trustworthiness concerns based on IPD checking. Therefore, out of the total 64 shared RCTs, 54 (84.4%) met the pre-defined trustworthiness criteria, while 10 were eventually deemed to have trustworthiness concerns.[43–52] The authors of 10 RCTs with trustworthiness concerns were informed regarding the issues. After informing the authors, only one out of these 10 RCTs has been retracted,[51] another one was corrected by omitting 14 duplicate observations in a sample size of 236.[49]

Risk of bias assessment (Fig. 1) showed that 13/64 (20.3%) RCTs in the shared group vs 57/201 (28.4%) RCTs in the non-shared group were classified as overall high-risk of bias (Odds ratio 0.65, 95% CI 0.32–1.26; $p = 0.20$). In addition, 22/141 (15.6%) RCTs meeting trustworthiness criteria vs 47/100 (47.0%) RCTs not meeting trustworthiness criteria were deemed overall high risk of bias (Odds ratio 0.21, 95% CI 0.11–0.38; $p < 0.001$).

In the shared group, continuous baseline characteristics were not reported in 14 RCTs. In the 50 RCTs that did report them, the distribution of simulation-generated p-values was consistent with the expected uniform distribution (166 variables, $p = 0.50$) (Fig. 2a). In the non-shared group, 138 out of 201 RCTs reported continuous baseline characteristics and their distribution significantly deviated from the expected uniform distribution (466 variables, $p = 0.006$) (Fig. 2b). When further stratified by trustworthiness, the distributions in RCTs meeting trustworthiness criteria were consistent with the expected uniform distribution in both the shared (137 variables from 41 RCTs, $p = 0.34$) and non-shared groups (288 variables from 79 RCTs, $p = 0.37$) (Fig. 2c). However, among RCTs not meeting trustworthiness criteria, the distribution significantly deviated from the expected uniform distribution in the non-shared group (175 variables from 58 RCTs, $p = 0.003$) but remained consistent in the shared group of RCTs (29 variables from nine RCTs, $p = 0.92$) (Fig. 2d). Subgroup analysis of RCTs in the non-shared group according to reasons for not sharing did not show significant inconsistency (128 variables from 35 RCTs, $p = 0.05$ for RCTs with data destroyed, and 40 variables from 12 RCTs, $p = 0.08$ for RCTs with authors not willing to share), but these subgroup results are underpowered (Supplementary Figure S2).

Characteristics	N = 265 RCTs		p-value[b]
	Shared group, N = 64/265 (24.2%)	Non-shared group, N = 201/265 (75.8%)	
Median year of publication (IQR)	2016 (2010–2019)	2008 (2002–2015)	<0.001
Median author number (IQR)	5.0 (4.0–8.0)	4.0 (3.0–6.0)	0.003
Median randomised participants in each RCT (IQR)	204.0 (126.7–363.0)	132.5 (100.0–210.0)	<0.001
Median number of recruitment centres (IQR)	1.0 (1.0–3.0)	1.0 (1.0–1.0)	<0.001
Median number of study arms (IQR)	2.0 (2.0–2.0)	2.0 (2.0–2.0)	0.046
Region of publication[c]			
Africa	2 (3.1)	16 (8.0)	0.10
North- and South America	17 (26.6)	46 (22.9)	
Eastern Mediterranean	3 (4.7)	27 (13.4)	
Europe	18 (28.1)	33 (16.4)	
South-East Asia	17 (26.6)	63 (31.3)	
Western Pacific	6 (9.4)	15 (7.5)	
Taiwan	1 (1.6)	1 (0.5)	
Year of publication			
Published before 2000	3 (4.7)	28 (13.9)	<0.001
2000–2010	12 (18.7)	78 (38.8)	
After 2010	49 (76.6)	95 (47.3)	
Trial registration (on or after 2010 CTs only)[d]			
Adequate	22/50 (44.0)	14/99 (14.1)	<0.001
Retrospective	10/50 (20.0)	7/99 (7.1)	
Late	0/50	1/99 (1.0)	
Not registered/absent	18/50 (36.0)	77/99 (77.8)	
General conclusion of the trial			
Positive	29 (45.3)	119 (59.2)	0.21
Negative	5 (7.8)	10 (5.0)	
Inconclusive	28 (43.8)	67 (33.3)	
No details found/unpublished	2 (3.1)	5 (2.5)	
Ethics approval statement[e] (%)	54 (84.4)	136 (67.7)	0.01
Informed written consent (%)	39 (60.9)	88 (43.8)	0.02
Clear description of participant recruitment (%)	46 (71.9)	99 (49.2)	0.001
Poor reporting of dropouts (%)	10 (15.6)	91 (45.3)	<0.001
Feasible timeline[f] (%)			
Yes	50 (78.1)	129 (64.2)	0.08
No	1 (1.6)	14 (7.0)	
Not stated	13 (20.3)	58 (28.8)	
Use of CONSORT flow chart (%)	16 (25.0)	12 (6.0)	<0.001
Use of any study flow chart (%)	39 (60.9)	40 (19.9)	<0.001
Data sharing statement (%)	8 (12.5)	6 (3.0)	0.006
Median TRACT Score (IQR)	4 (0–8)	8 (5–13)	<0.001
RCTs meeting pre-defined trustworthiness criteria (%)	54/64 (84.4)	87/177 (49.1)[a]	<0.001

TRACT: Trustworthiness in RAndomised Controlled Trials (TRACT checklist); RCT: Randomised controlled trial; NR: Not reported; CONSORT: Consolidated Standards of Reporting Trials. [a]Non-shared group assessments were made only for the RCTs published as full-text. [b]For comparison between groups (Pearson's Chi-squared test/Fisher's Exact test for categorical data and Mann–Whitney U test for continuous data). [c]According to WHO regions of member countries (https://www.who.int/countries/). [d]Trial registration was assessed for trials published in or after 2010. [e]We considered reporting the name of the ethics committee, approval number/date and informed written consent to meet the full ethics requirement. [f]The study timeline is another important aspect to consider when appraising RCTs. The recruitment rate and the time between the end of recruitment to the submission of the manuscript have been identified by our study group as critical timeframes. Feasible timeline was assessed whether there is unrealistic fast recruitment of participants within the study time (especially single-centre studies) or an impossible time frame between the end of recruitment and submission of the paper; in either case, a feasible timeline was not observed.

Table 1: **Characteristics of randomised controlled trials with regard to the individual participant data-sharing status.**

The proportions of RCTs with at least one inconsistently reported p-values that cannot be reproduced with summary data p-value were not significantly different between the two groups (for baseline characteristics: 11/64 (17.2%) in shared vs 51/201 (25.4%) in the non-shared group, p = 0.18; for caesarean delivery: 27/64 (42.2%) in shared vs 62/201 (30.8%) non-shared group, p = 0.09; for uterine hyperstimulation: 11/64 (17.2%) in shared vs 26/201 (12.9%) in non-shared group, p = 0.39). Among RCTs with inconsistently

Characteristics	Non-shared group (201 RCTs)		
	TRACT-based trustworthiness assessment in non-shared group[a], N = 177/265 (66.8%)		
	'Meeting trustworthiness criteria', n = 87/177 (49.1%)	'Not-meeting trustworthiness criteria', n = 90/177 (50.9%)	p-value[b]
Responded to author inquiry, n (%)	51/87 (58.6)	26/90 (28.9)	0.01
Median year of publication (IQR)	2004 (2001–2012)	2013 (2005–2018)	<0.001
Median author number (IQR)	5.0 (4.0–6.0)	3.0 (3.0–5.0)	<0.001
Median randomised participants in each RCT (IQR)	155.5 (106.5–214.5)	112.0 (92.5–200.0)	<0.001
Median number of recruitment centres (IQR)	1.0 (1.0–1.0)	1.0 (1.0–1.0)	0.002
Median number of study arms (IQR)	2.0 (2.0–2.0)	2.0 (2.0–2.0)	0.62
Region of publication[c]			
Africa	10 (11.5)	6 (6.7)	<0.001
North- and South America	28 (32.2)	5 (5.7)	
Eastern Mediterranean	7 (8.1)	19 (21.8)	
Europe	23 (26.4)	7 (8.1)	
South-East Asia	11 (12.6)	47 (54.0)	
Western Pacific	7 (8.1)	6 (6.7)	
Taiwan	1 (1.1)	0	
Year of publication			
Published before 2000	17 (19.5)	6 (6.7)	<0.001
2000–2010	45 (51.7)	28 (31.1)	
After 2010	25 (28.7)	56 (62.2)	
Trial registration (on or after 2010 RCTs only)[d]			
Adequate	11/26 (42.3)	1/56 (1.8)	<0.001
Retrospective	2/26 (7.7)	5/56 (8.9)	
Late	1/26 (3.8)	0/56	
Not registered/absent	12/26 (46.1)	50/56 (89.3)	
General conclusion of the trial			
Positive	48 (55.2)	63 (70.0)	0.14
Negative	5 (5.7)	3 (3.3)	
Inconclusive	34 (39.1)	24 (26.7)	
No details found/unpublished	0	0	
Ethics approval statement[e] (%)	75 (86.2)	61 (67.8)	0.004
Informed written consent (%)	49 (56.3)	39 (43.3)	0.08
Clear description of participant recruitment (%)	70 (80.5)	29 (32.2)	<0.001
Poor reporting of dropouts (%)	17 (19.5)	74 (82.2)	<0.001
Feasible timeline[f] (%)			
Yes	74 (85.1)	55 (61.1)	0.004
No	5 (5.7)	9 (10.0)	
Not stated	8 (9.2)	23 (25.6)	
Use of CONSORT flow chart (%)	9 (10.3)	3 (3.3)	0.08
Use of any study flow chart (%)	34 (39.1)	6 (6.7)	<0.001
Data sharing statement (%)	2 (2.3)	2 (2.2)	1.0
Median TRACT Score (IQR)	6 (5–8)	14 (11–17)	<0.001
RCTs meeting pre-defined trustworthiness criteria (%)	–	–	–

TRACT: Trustworthiness in RAndomised Controlled Trials (TRACT checklist); RCT: Randomised controlled trial; NR: Not reported; CONSORT: Consolidated Standards of Reporting Trials. [a]Non-shared group assessments were made only for the RCTs published as full-text. [b]For comparison between groups (Pearson's Chi-squared test/Fisher's Exact test for categorical data and Mann–Whitney U test for continuous data). [c]According to WHO regions of member countries (https://www.who.int/countries/). [d]Trial registration was assessed for trials published in or after 2010. [e]We considered reporting the name of the ethics committee, approval number/date and informed written consent to meet the full ethics requirement. [f]The study timeline is another important aspect to consider when appraising RCTs. The recruitment rate and the time between the end of recruitment to the submission of the manuscript have been identified by our study group as critical timeframes. Feasible timeline was assessed whether there is unrealistic fast recruitment of participants within the study time (especially single-centre studies) or an impossible time frame between the end of recruitment and submission of the paper; in either case, a feasible timeline was not observed.

***Table 2:* Characteristics of individual participant data non-shared randomised controlled trials with regard to the publication format and trustworthiness category.**

reported p-values that cannot be reproduced with summary data, median (IQR) rate of inconsistently reported p-values per trial for continuous baseline characteristics (number of inconsistently reported p-values/total reported p-values) were not significantly different between shared and non-shared groups [75% (50–100) vs 66.7% (50–100), p = 0.75]. The proportions of RCTs with at least one inconsistently reported effect estimate were not significantly different between the shared and non-shared groups either (for caesarean delivery: 9/21 (42.9%) in shared vs 7/17 (41.2%) in non-shared group, p = 0.92; for uterine hyperstimulation: 7/15 (46.7%) in shared vs 6/10 (60.0%) non-shared group, p = 0.69) (Supplementary Table S3).

The shared group had more calculated p-value/reported p-value ratios of <1 than those of >1 for inconsistently reported continuous baseline characteristics (13/22, 59.1% vs 9/22, 40.9%). This indicates a tendency to report p-values that are larger than they should be. However, the opposite was found for outcomes (6/43, 14.0% vs 37/43, 86.0%), suggesting that most inconsistently reported p-values were due to reported p-values being smaller than they should be (Fig. 3a). For the non-shared group, we found similar patterns, where more ratios were <1 than >1 for continuous baseline characteristics (57/94, 60.6% vs 37/94, 39.4%) and the opposite for outcomes (17/56, 30.4% vs 39/56, 69.6%) (Fig. 3b). In the group meeting trustworthiness criteria, the proportions of ratios that were <1 and >1 for inconsistently reported continuous baseline characteristics were nearly equal (40/78, 51.3% vs 38/78, 48.7%, respectively), suggesting no direction of error (Fig. 3c). However, for the outcomes in this group, more ratios were >1 (<1, 10/80, 12.5% vs >1, 70/80, 87.5%) (Fig. 3c). In the group not meeting trustworthiness criteria, more ratios for continuous baseline characteristics were <1 (<1, 36/57, 63.2% vs >1, 21/57, 36.8%), while more ratios for outcomes were >1 (<1, 6/38, 15.8% vs >1, 32/38, 84.2%).

The results of the aggregate data meta-analysis for the shared and non-shared groups in each of the nine comparisons of the induction of labour methods and accompanying GRADE assessments are given in Table 3. For caesarean delivery, when comparing the shared group to the non-shared group, three out of nine comparisons had RORs that were significantly >1.0, showing a more extreme effect estimate in the non-shared group than the shared group. This was not observed for hyperstimulation between shared and non-shared groups. The GRADE assessment showed higher certainty of evidence for nine and equal certainty of evidence for seven effect estimates in the shared groups compared to non-shared groups. This was more pronounced in the group of RCTs without trustworthiness concerns than in group of RCTs with trustworthiness concerns for all available 13 effect estimates, showing higher certainty in effect estimates from those RCTs

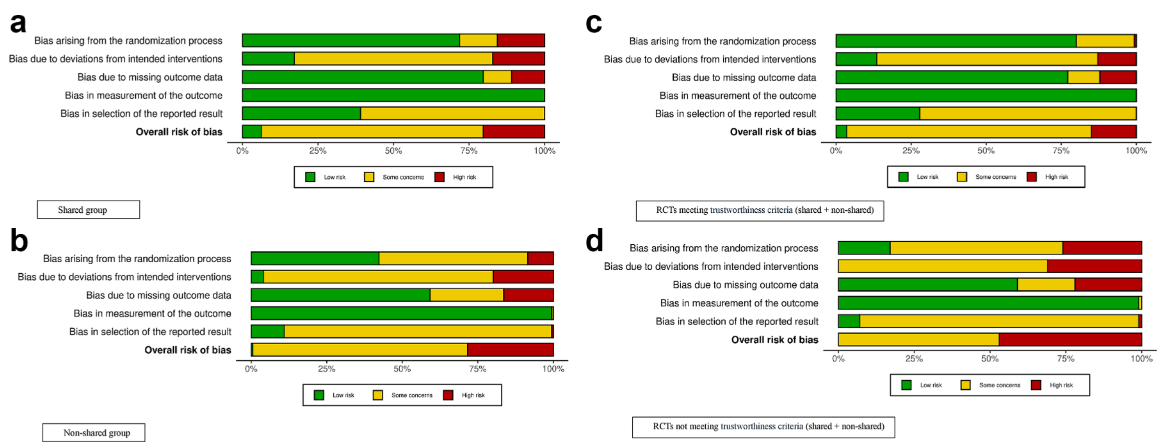

Fig. 1: Risk of bias among randomised controlled trials evaluated in this study. a. Shared group; b. Non-shared group; c. RCTs meeting trustworthiness criteria (shared + non-shared); d. RCTs not meeting trustworthiness criteria (shared + non-shared). RCT: Randomised controlled trials. The Cochrane risk of bias-2 tool was used.

without such trustworthiness issues (Supplementary Table S4).

A post-hoc sensitivity analysis was conducted by categorising RCTs based on their publication date: before 2010 versus on or after 2010 and comparing shared and non-shared groups (Supplementary Table S6 and Figure S4). Regardless of whether the RCTs were published before or after 2010, issues related to trustworthiness, risk of bias assessments, and the probability of random sampling based on continuous baseline characteristics were more prevalent in the non-shared group.

Discussion

In this meta-epidemiological study, we assessed the characteristics, certainty of evidence, randomisation, trustworthiness, and effect estimates of RCTs evaluating cervical ripening in induction of labour, comparing those that shared IPD versus those that did not. RCTs that did not share IPD have lower quality, more trustworthiness concerns, and less reproducibility than those that shared IPD. While most shared RCTs met trustworthiness criteria, only half of non-shared RCTs were considered to be without trustworthiness concerns. Non-shared RCTs were generally less likely to have been properly randomised, whereas shared RCTs were more likely to have proper randomisation, reinforcing the reliability of IPD for evidence synthesis. Concerningly, these differences in quality and trustworthiness impact the effect estimates. As shown by the RORs of data synthesis, in most comparisons, non-shared RCTs showed exaggerated effect estimates compared to the shared group, which may have distorted effect estimates in previous meta-analyses, leading to misleading conclusions and influencing medical practice around induction of labour.

The prevalence of inconsistently reported p-values and effect estimates in a large proportion of RCTs is another significant concern, and the shared group is not exempt from this issue. Multiple factors contribute to inconsistently reported results in research, including the failure to adequately consider the null hypothesis, selective reporting of data, and biases in experimental design and data handling. Statistical flaws are particularly notable as a major source of irreproducibility across biomedical research, highlighting the importance of robust statistical expertise in the conduct of RCTs.[53] We used the calculated p-value/reported p-value ratio to infer possible reasons for inconsistently reported p-values that cannot be reproduced with summary data. If the distribution of these ratios (<1 and >1) were balanced, it would suggest that inconsistently reported results are likely due to unintentional errors. However, a predominance of ratios <1 for baseline characteristics suggests a tendency to report p-values larger than they should be, potentially to obscure small p-values that "indicate" inadequate randomisation. This reflects a persistent misuse of statistical tests to assess randomisation adequacy, alongside the flawed assumption that a p-value <0.05 for baseline characteristics indicates a failure of randomisation. For outcome measures, we hypothesised that some authors may manipulate p-values to make them smaller, seeking to generate positive findings. This would lead to more ratios >1, and unfortunately, this pattern was evident in our analysis, including within the shared group. These findings highlight the urgent need for improved statistical rigour, transparency, and oversight in clinical research. More importantly, they emphasise

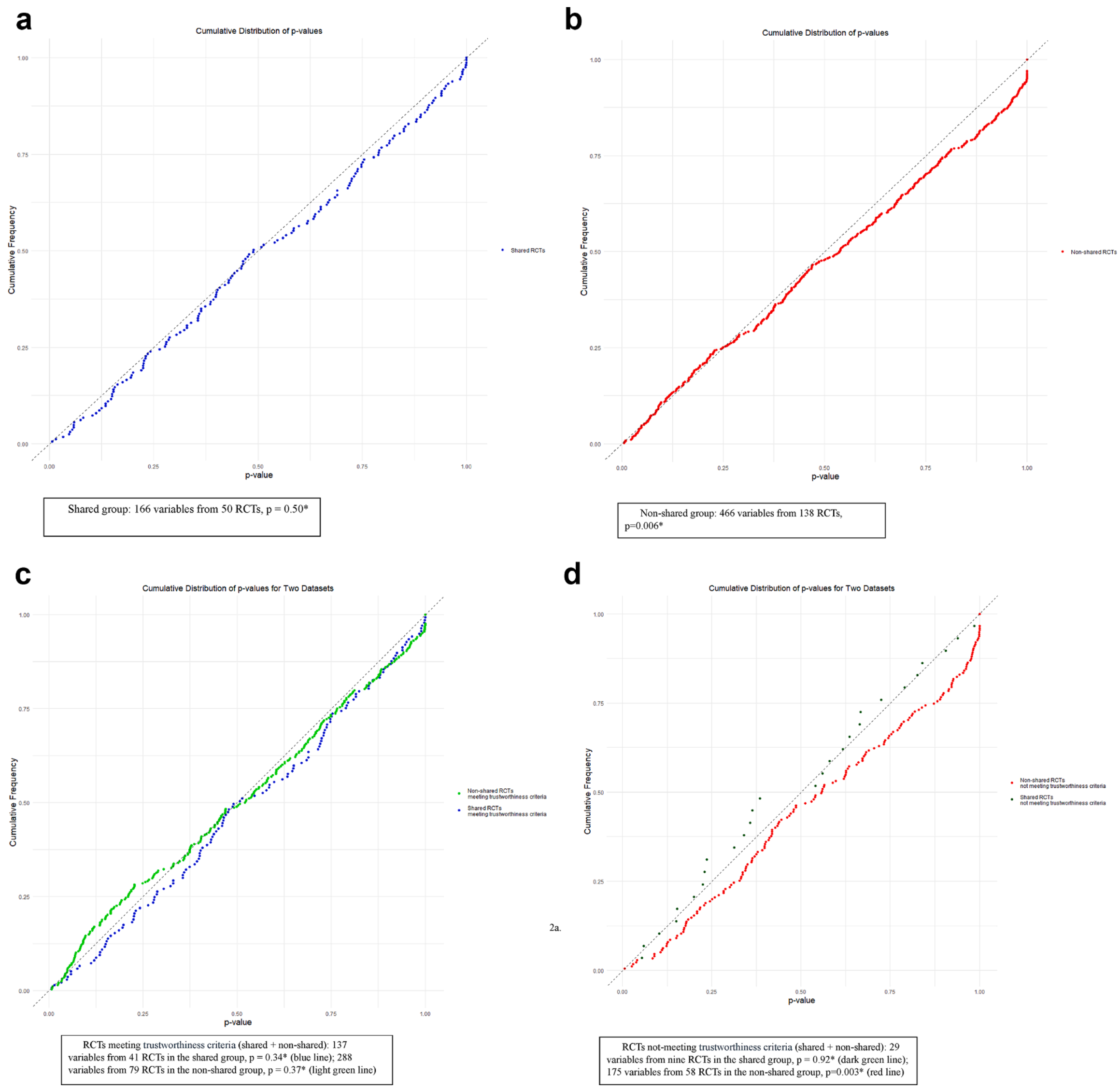

Fig. 2: The cumulative distribution of Monte Carlo simulations generated p-values for continuous baseline characteristics. a. Shared group: 166 variables from 50 RCTs, p = 0.50*; b. Non-shared group: 466 variables from 138 RCTs, p = 0.006*; c. RCTs meeting trustworthiness criteria (shared + non-shared): 137 variables from 41 RCTs in the shared group, p = 0.34* (blue line) and 288 variables from 79 RCTs in the non-shared group, p = 0.37* (light green line); d. RCTs not-meeting trustworthiness criteria (shared + non-shared): 29 variables from nine RCTs in the shared group, p = 0.92* (dark green line) and 175 variables from 58 RCTs in the non-shared group, p = 0.003* (red line). RCT: randomised controlled trials. *According to the Kolmogorov–Smirnov test.

the necessity of re-evaluating RCT data in evidence synthesis, with IPD offering the most robust approach to address these issues and ensure reproducibility in research.

Data-sharing for evidence synthesis is increasingly recognised as a major requirement of clinical research culture.[16,54] The ICMJE requires a data-sharing statement from manuscript submissions reporting clinical trials effective from July 1, 2018.[54] Beyond enhancing data validity, it maximises the value of research investments by enabling comprehensive utilisation of existing data, helping prevent unnecessary duplication of studies, and fostering more efficient and ethical research practices. Despite the efforts, in reality, IPD retrieval for evidence synthesis is challenging and data are not always available. Evidence suggests that data-sharing status is unsatisfactory despite journal policies for data-sharing in two leading biomedical journals (*The BMJ and PLOS Medicine*).[55] IPD are unavailable due to multiple and often complex reasons, including a lack

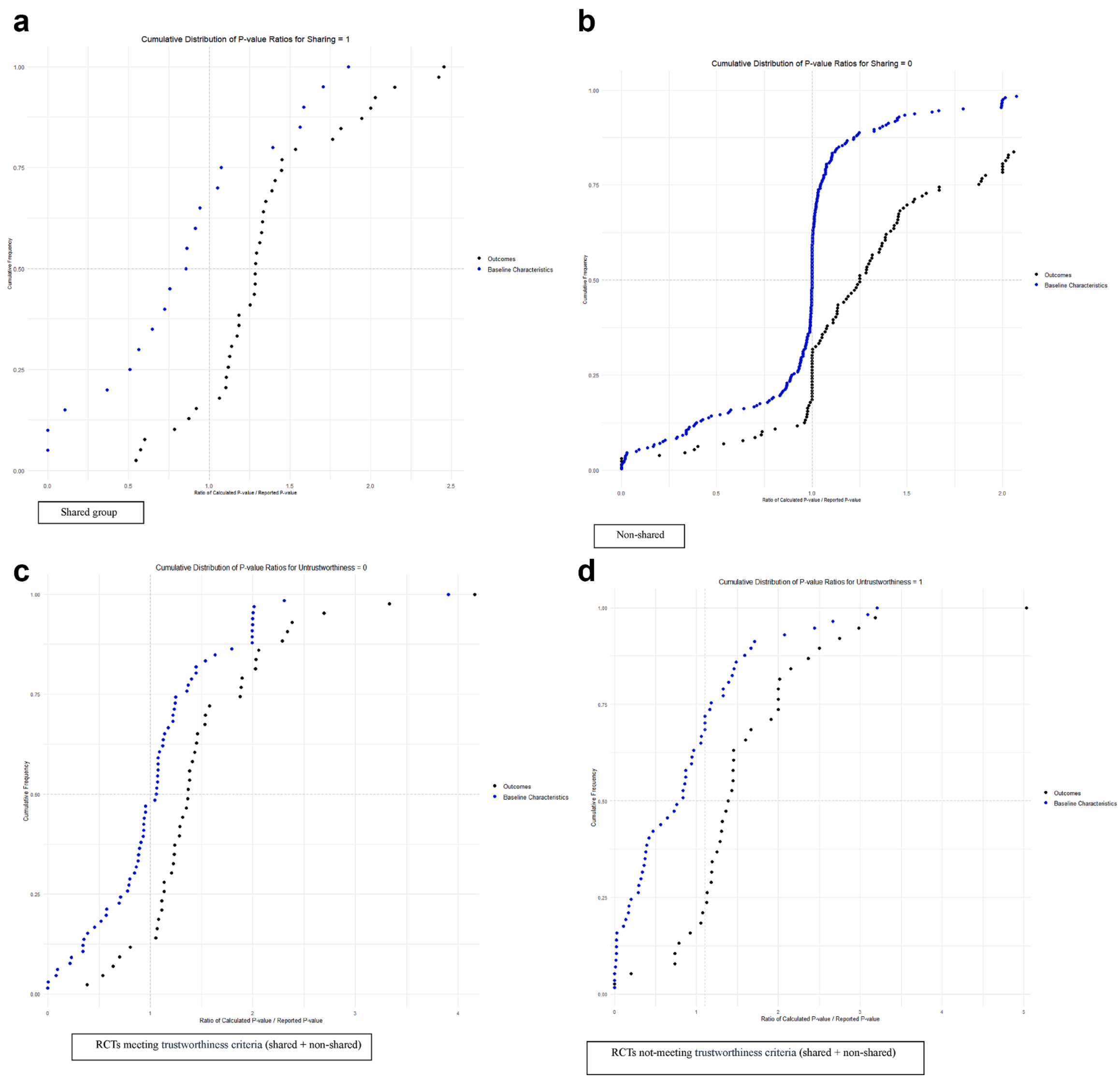

Fig. 3: The cumulative distribution of the calculated p-value/reported p-value ratio for inconsistently reported baseline characteristics and outcomes. a. Shared group; b. Non-shared group; c. RCTs meeting trustworthiness criteria (shared + non-shared); d. RCTs not-meeting trustworthiness criteria (shared + non-shared). RCT: Randomised controlled trials.

of informed consent for sharing, participant confidentiality, poor governance for IPD-sharing requests, navigating administrative barriers, and misunderstanding the concept of data-sharing.[56–58] It remains unknown whether the value of medical research increases with data-sharing practices.[17]

There is now an emphasis on data-sharing to expedite the translation of research results into knowledge and practice for the collective good, rather than individual or corporate benefit.[59] We acknowledge that there are cases where participant confidentiality or the failure to obtain explicit consent for sharing hampered data-sharing. In our exercises of data seeking, where we reached some of the ethical committees in these instances, the ethical committees were mostly supportive in arranging a feasible pathway for sharing; however, the trial authors were not willing to proceed. Data custodians of publicly funded health research and health-related datasets have the responsibility to support and facilitate access to data for research purposes when such requests are aligned with relevant legislation, policies, and guidelines for ethical practice in human research.[60] Trialists, journals, and research organisations should explore the potential archiving of raw data in an accessible data repository soon after trial completion. To our knowledge, only a few journals (e.

Comparison (*Project status*)	No of RCTs[a]		Outcome	Group	Number of trials (women)	OR (95% CI)	p-value	ROR (95% CI)	I^2	GRADE assessment
	Shared as IPD	Non-shared								
Balloon catheter versus low-dose oral misoprostol (*Published*)[15]	4	3	Caesarean delivery rate	Shared	4 (2815)	1.30 (1.08; 1.56)	0.02	1.91 (0.08; 43.6)	0	Moderate-certainty evidence[b]
				Non-shared	2 (108)	0.68 (0.03; 15.42)	0.36		0	Very low-certainty evidence[f]
			Uterine hyperstimulation	Shared	4 (2815)	0.75 (0.40; 1.40)	0.24	–	0	Moderate-certainty evidence[b]
				Non-shared	No data	–	–		–	–
Balloon catheter versus vaginal prostaglandins (*Published*)[16]	13	47	Caesarean delivery rate	Shared	13 (5549)	1.09 (0.93; 1.28)	0.25	1.04 (0.79; 1.37)	20	Moderate-certainty evidence[d]
				Non-shared	28 (3728)	1.05 (0.84; 1.30)	0.67		29	Very low-certainty evidence[d,f]
			Uterine hyperstimulation	Shared	10 (4188)	0.34 (0.15; 0.79)	0.02	1.00 (0.24; 4.09)	42	Moderate-certainty evidence[d]
				Non-shared	14 (2023)	0.34 (0.11; 1.10)	0.07		37	Very low-certainty evidence[d,f]
Double balloon versus single balloon catheter (*Published*)[17]	3	5	Caesarean delivery rate	Shared	3 (688)	1.39 (0.76; 2.52)	0.14	2.36 (1.05; 5.30)	0	Low-certainty evidence[b,e]
				Non-shared	5 (606)	0.59 (0.34; 1.01)	0.05		7	Very low-certainty evidence[e,f]
			Uterine hyperstimulation	Shared	No data	–	–	–	–	–
				Non-shared	No data	–	–		–	–
Vaginal misoprostol versus vaginal dinoprostone (*Published*)[18]	9	42	Caesarean delivery rate	Shared	9 (4341)	1.09 (0.92; 1.30)	0.27	1.43 (1.05; 1.95)	2	High-certainty evidence
				Non-shared	24 (5360)	0.76 (0.59; 0.97)	0.03		65	Low-certainty evidence[b,d]
			Uterine hyperstimulation	Shared	6 (3726)	1.07 (0.74; 1.54)	0.65	0.86 (0.49; 1.51)	0	High-certainty evidence
				Non-shared	20 (4559)	1.24 (0.81; 1.92)	0.30		20	Moderate-certainty evidence[d]
Outpatient versus inpatient methods (*Awaiting publication*)	11	8	Caesarean delivery rate	Shared	11 (2577)	1.16 (0.95; 1.41)	0.13	1.29 (0.95; 1.75)	0	Moderate-certainty evidence[b]
				Non-shared	7 (1703)	0.90 (0.71; 1.15)	0.34		0	Moderate-certainty evidence[b]
			Uterine hyperstimulation	Shared	3 (1636)	0.39 (0.01; 12.05)	0.36	0.29 (0.01; 14.42)	53	Very low-certainty evidence[b,c,d]
				Non-shared	4 (1228)	1.33 (0.21; 8.43)	0.57		0	Very low-certainty evidence[b,c]
Oral misoprostol versus vaginal dinoprostone (*Author inquiry completed*)	8	10	Caesarean delivery rate	Shared	8 (2563)	0.98 (0.82; 1.16)	0.79	1.29 (1.06; 1.57)	0	Moderate-certainty evidence[b]
				Non-shared	10 (8162)	0.76 (0.69; 0.84)	<0.01		0	Very low-certainty evidence[f]
			Uterine hyperstimulation	Shared	5 (1759)	0.59 (0.48; 0.73)	<0.01	0.82 (0.35; 1.89)	0	Low-certainty evidence[b,e]
				Non-shared	9 (7866)	0.72 (0.32; 1.63)	0.36		62	Very low-certainty evidence[f]
Oral misoprostol versus vaginal misoprostol (*Author inquiry completed*)	12	48	Caesarean delivery rate	Shared	10 (2510)	1.02 (0.67; 1.55)	0.91	1.03 (0.65; 1.62)	55	Low-certainty evidence[b,d]
				Non-shared	41 (7334)	0.99 (0.83; 1.18)	0.93		44	Low-certainty evidence[b,d]
			Uterine hyperstimulation	Shared	8 (2140)	0.47 (0.20; 1.11)	0.08	0.70 (0.26; 1.86)	51	Low-certainty evidence[b,d]
				Non-shared	30 (5788)	0.67 (0.42; 1.05)	0.08		51	Low-certainty evidence[b,d]
Vaginal misoprostol versus intracervical dinoprostone (*Author inquiry completed*)	5	28	Caesarean delivery rate	Shared	4 (770)	0.77 (0.51; 1.15)	0.13	0.84 (0.54; 1.29)	0	Low-certainty evidence[b,e]
				Non-shared	26 (3633)	0.92 (0.78; 1.10)	0.35		0	Low-certainty evidence[b,e]
			Uterine hyperstimulation	Shared	4 (770)	1.14 (0.09; 14.78)	0.85	0.64 (0.05; 8.61)	0	Very low-certainty evidence[b,c,e]
				Non-shared	18 (1945)	1.79 (1.12; 2.84)	0.02		0	Very low-certainty evidence[c,e,f]
Balloon catheter plus concurrent low-dose vaginal misoprostol versus low-dose vaginal misoprostol alone (*Author inquiry completed*)	8	15	Caesarean delivery rate	Shared	8 (1294)	0.96 (0.68; 1.34)	0.77	1.25 (0.74; 2.09)	24	Low-certainty evidence[b,c,e]
				Non-shared	13 (1863)	0.77 (0.52; 1.13)	0.16		44	Very low-certainty evidence[b,c,e,d,f]
			Uterine hyperstimulation	Shared	5 (726)	0.28 (0.10; 0.81)	0.03	0.35 (0.11; 1.14)	24	Very low-certainty evidence[c,e,f]
				Non-shared	8 (1039)	0.79 (0.49; 1.27)	0.21		0	Very low-certainty evidence[c,e,f]

IPD: Individual participant data; RCT: Randomised controlled trial; 95% CI: 95% confidence interval; ROR: Ratio of odds ratios; GRADE, Grading of Recommendations, Assessment, Development and Evaluations; I^2: statistics of heterogeneity; NA, not applicable. [a]These nine comparisons included 265 trials. Some trials contributed to multiple comparisons. [b]Downgraded one level due to some concerns on risk of bias. [c]Downgraded one level for imprecision. [d]Downgraded one level for inconsistency. [e]Downgraded one level for concerns about data completeness. [f]Downgraded by three levels due to major concerns on risk of bias and serious imprecision.

***Table 3:* The results of aggregate data meta-analyses in the shared and non-shared groups.**

g., BMJ, PLOS Medicine) require the immediate release of raw data accompanying publication. Promoting such practice could foster research integrity, minimise research waste, and facilitate evidence synthesis. The results of this study provide empirical evidence to support initiatives that promote actual research data-sharing and proper conduct of RCTs. Moving forward, the use of IPD for evidence synthesis should be recommended to enhance the reliability and validity of research findings in evidence synthesis.

The concept of trustworthiness in RCTs is a relatively new and evolving area. To strengthen the integrity of the evidence base, we should focus not only on eliminating studies with trustworthiness concerns from the system but also on preventing their entry in the first place. This approach could pave the way for an 'evidence ecosystem' that acts as a filter, removing 'pollutants' beyond the limitations of the traditional evidence pyramid.[61] Apart from the TRACT checklist used in this study, there are several other similar activities in this regard[35,62–66] Alfirevic et al., in their recent review, have proposed a menu of domains and questions to be considered in a more formal trustworthiness assessment.[67] They summarised 77 questions under five domains (governance, plausibility, plagiarism, reporting, and statistics). The INSPECT-SR project is an ongoing tool development project to assess the trustworthiness of RCTs in systematic reviews of healthcare-related interventions.[14] None of these tools has been formally validated. Assessment of trustworthiness of RCTs and labelling a published work as having trustworthiness concerns could be challenging. Still, some outstanding issues could potentially malfunction this assessment process. These include variations in the tools used and their level of validity, missing theory behind some tools, the level of experience in assessing trustworthiness, the additional burden on the assessment, feasibility, and the limited capacity to address the relatively new problem of AI-generated fraudulent research.[67,68] Our tool, the TRACT checklist, is also a living tool and has its own limitations.[68]

Assuming that every study is valuable and all studies are trustworthy, the so-called data unavailability bias stemming from the inability to source IPD from all identified RCTs in IPD meta-analysis is believed to impede research transparency, reduce statistical power, increase the risk of bias, and even jeopardise study completion.[69,70] However, our findings suggest this traditional belief may be incorrect in many instances. We propose that the right approach is to assess the trustworthiness and quality of all eligible RCTs to identify problematic ones, regardless of whether they provided IPD. The traditional 'all-inclusive' approach, which pools all RCTs into a single estimate without evaluating their trustworthiness or quality, risks contaminating the evidence base. To address this concern in meta-analysis, we recommend that all eligible data still be reported but stratified into four subgroups based on trustworthiness and data availability.[19,33,71,72] This enables a clearer understanding of how these factors may impact the findings. The four subgroups are: (1) IPD meeting trustworthiness criteria; (2) non-IPD meeting trustworthiness criteria; (3) IPD not meeting trustworthiness criteria; and (4) non-IPD not meeting trustworthiness criteria. Interpretation for clinical implications should be based on a holistic overview of all subgroup results. The assumption that research is trustworthy by default is increasingly being challenged, and the concept of data unavailability bias could be misleading if the underlying reasons for not sharing data are not analysed.

The study's strengths include the use of a predefined protocol and analysis plan, a systematic search for IPD meta-analysis projects, and rigorous efforts to obtain IPD through author enquiries. Additionally, the study used established methodologies such as the RoB-2 for risk of bias, the GRADE tool for certainty in evidence, the TRACT checklist for trustworthiness evaluation, and Monte Carlo simulations for randomisation assessment. This study also has limitations. Not all induction of labour methods were included; however, we believe the comparisons included represent the most clinically relevant and significant ones. We could not assess the trustworthiness of abstracts if IPD were not available. Also, the TRACT checklist has not formally been validated. We acknowledge the possibility that if non-IPD RCTs had shared their data, some may have been reclassified into different trustworthiness categories. However, this assumption could not be verified. Our primary aim is not to ensure a perfectly balanced comparison between shared and non-shared RCTs, but rather to enhance the accuracy of evidence synthesis. This involves maximising the use of available data, particularly valuable IPD, while minimising the risk of contamination from studies with trustworthiness concerns, which, as demonstrated in our analysis, are more commonly associated with non-shared RCTs. In our data enquiry, we have made extensive efforts to contact the authors of papers that did not share IPD. The fact that we could not get a response is another indication of challenges in the trustworthiness of the non-shared data. We welcome feedback from authors of non-shared RCTs to enhance our assessment process and refine our algorithm. A larger number of non-shared RCTs compared to shared RCTs could have affected statistical power in the relevant calculation, and caution should be taken when interpreting the results. As data sharing policies have been different for earlier RCTs, the year of publication could also have a role in the findings. However, we did not find trustworthiness issues vary between those published before or after 2010 in the sensitivity analysis. The use of multiple univariate statistical tests could potentially cause multiplicity issues. Borderline significant p-values should be interpreted with caution.

In conclusion, in studies evaluating labour induction methods, RCTs without IPD-sharing have lower quality and more trustworthiness concerns than those with shared IPD. The availability of IPD and willingness to share data are good indicators of the quality and trustworthiness of RCTs. Trustworthiness and quality should be considered whenever using RCTs for evidence synthesis. This would help foster the credibility of the evidence for clinical guidelines and clinical practice.

Contributors

MP principally managed the whole project and carried out the design, statistical analysis plan, data collection, data synthesis, writing, editing, finalising, and submitting the manuscript. BWM, WenL, and DLR

conceived the research idea, interpreted data, revised the manuscript and were responsible for overseeing all aspects of conduct. WenL contributed to the statistical strategy. WanL contributed to the trustworthiness assessment as the second reviewer. BWM and DLR provided clinical oversight. MP designed and conducted all the literature searches. MP, BMW, DLR, and WenL have accessed and verified the underlying data. All authors were involved in the decision to submit the manuscript.

Data sharing statement
Datasets used in this study, data dictionary, and trustworthiness assessment data will be made available upon a reasonable request from the corresponding author. However, the trial investigators who shared individual participant data for the purposes of the meta-analysis retain ownership of their trial data and any requests for access to individual participant data should be made directly to them.

Declaration of interests
MP is supported by a Research Training Stipend, provided by the Australian Government. BWM declared grants from NHMRC, personal fees from Merck, Organon and Norgine and grants from Merck, outside the submitted work. DLR declared grants from NHMRC, fees to participate in Advisory boards from Alexion Pharmaceuticals, and travel support and lecture fees from the International Society of Ultrasound in Obstetrics and Gynecology (ISUOG), unrelated to this work. WenL declared a grant from NHMRC that supports this work, a Ramaciotti Health Investment Grant unrelated to this work, and travel support from Merck. The remaining authors report no conflict of interest.

Acknowledgements
This study is supported by three NHMRC Investigator Grants (GNT1176437 for BWM, GNT2016729 for WenL, and GNT2025670 for DLR). These funding sources had no role in the design, execution, analyses, or data interpretation of this research. MP is supported by a Research Training Stipend provided by the Australian Government. We also thank all the individual trial participants and investigators for their time and effort in conducting the 265 randomised controlled trials evaluated in this study. Our sincere gratitude to Professor Toshi Furukawa, Professor Emerit, Kyoto University Graduate School of Medicine/School of Public Health, Japan, for his valuable input for the R code to calculate 95% confidence interval for ratio of odds ratios (ROR).

Appendix A. Supplementary data
Supplementary data related to this article can be found at https://doi.org/10.1016/j.eclinm.2025.103346.

References
1 Akobeng AK. Understanding randomised controlled trials. *Arch Dis Child*. 2005;90:837–840.
2 Li W, van Wely M, Gurrin L, Mol BW. Integrity of randomized controlled trials: challenges and solutions. *Fertil Steril*. 2020;113(6):1113–1119.
3 Liu Y, Thornton JG, Li W, van Wely M, Mol BW. Concerns about data integrity of 22 randomized controlled trials in women's health. *Am J Perinatol*. 2023;40(3):279–289.
4 Ferreira RDDS, Negrini R, Bernardo WM, Simões R, Piato S. The effects of sildenafil in maternal and fetal outcomes in pregnancy: a systematic review and meta-analysis. *PLoS One*. 2019;14(7):e0219732.
5 Rakhanova Y, Almawi WY, Aimagambetova G, Riethmacher D. The effects of sildenafil citrate on intrauterine growth restriction: a systematic review and meta-analysis. *BMC Pregnancy Childbirth*. 2023;23(1):409.
6 Manouchehri E, Makvandi S, Razi M, Sahebari M, Larki M. Efficient administration of a combination of nifedipine and sildenafil citrate versus only nifedipine on clinical outcomes in women with threatened preterm labor: a systematic review and meta-analysis. *BMC Pediatr*. 2024;24(1):106.
7 Cornthwaite K, Bahl R, Winter C, et al. Management of impacted fetal head at caesarean birth: scientific impact paper no. 73. *BJOG*. 2023;130(12):e40–e64.
8 Grey A, Avenell A, Bolland MJ, Thornton JG. The fetal pillow deflates-lessons for all. *BJOG*. 2024;132(4):429–432.
9 Lawrence JM, Meyerowitz-Katz G, Heathers JAJ, Brown NJL, Sheldrick KA. The lesson of ivermectin: meta-analyses based on summary data alone are inherently unreliable. *Nat Med*. 2021;27(11):1853–1854.
10 Gautret P, Lagier JC, Parola P, et al. RETRACTED: hydroxychloroquine and azithromycin as a treatment of COVID-19: results of an open-label non-randomized clinical trial. *Int J Antimicrob Agents*. 2020;56(1):105949.
11 Ker K, Shakur H, Roberts I. Does tranexamic acid prevent postpartum haemorrhage? A systematic review of randomised controlled trials. *BJOG*. 2016;123(11):1745–1752.
12 Grey A, Avenell A, Bolland M. Revised meta-analysis of vitamin K and fractures. *JAMA Intern Med*. 2018;178(8):1135.
13 Cochrane database of systematic reviews: editorial policies. Available from: https://www.cochranelibrary.com/cdsr/editorial-policies. Accessed May 20, 2025.
14 Wilkinson J, Heal C, Antoniou GA, et al. Protocol for the development of a tool (INSPECT-SR) to identify problematic randomised controlled trials in systematic reviews of health interventions. *BMJ Open*. 2024;14(3):e084164.
15 Higgins JP, Thomas J, Chandler J, et al. *Cochrane Handbook for Systematic Reviews of Interventions*. Hoboken, NJ, USA: John Wiley & Sons; 2019.
16 Nosek BA, Alter G, Banks GC, et al. SCIENTIFIC STANDARDS. Promoting an open research culture. *Science*. 2015;348(6242):1422–1425.
17 Ohmann C, Moher D, Siebert M, Motschall E, Naudet F. Status, use and impact of sharing individual participant data from clinical trials: a scoping review. *BMJ Open*. 2021;11(8):e049228.
18 Jones MN, Palmer KR, Pathirana MM, et al. Balloon catheters versus vaginal prostaglandins for labour induction (CPI Collaborative): an individual participant data meta-analysis of randomised controlled trials. *Lancet*. 2022;400(10364):1681–1692.
19 Tan SY, Au N, Peel MD, et al. Oral misoprostol versus vaginal dinoprostone (PGE2) for labour induction: an individual participant data meta-analysis of randomised controlled trials. *Ultrasound Obstet Gynecol*. 2025 [In press].
20 Peel MD, Croll DMR, Kessler J, et al. Double-vs single-balloon catheter for induction of labor: systematic review and individual participant data meta-analysis. *Acta Obstet Gynecol Scand*. 2023;102(11):1440–1449.
21 Kemper JI, Li W, Goni S, et al. Foley catheter vs oral misoprostol for induction of labor: individual participant data meta-analysis. *Ultrasound Obstet Gynecol*. 2021;57(2):215–223.
22 Chan F, Patabendige M, Wise MR, et al. Inpatient vaginal dinoprostone vs outpatient balloon catheters for cervical ripening in induction of labor: an individual participant data meta-analysis of randomized controlled trials. *Acta Obstet Gynecol Scand*. 2025;104(6):1041–1055.
23 Patabendige M, Chan F, Wise MR, et al. Outpatient or inpatient setting for cervical ripening before induction of labour: an individual participant data meta-analysis. *BJOG*. 2025:1–17. https://doi.org/10.1111/1471-0528.18253.
24 Patabendige M, Chan F, Vayssiere C, et al. Vaginal misoprostol versus vaginal dinoprostone for cervical ripening and induction of labour: an individual participant data meta-analysis of randomised controlled trials. *BJOG*. 2024;131(9):1167–1180.
25 Center for Open Science 'OSF REGISTRIES' open registries network [Internet]. Available from: https://accounts.osf.io/login?campaign=institution&institutionId=&service=https://osf.io/login/?next%3D%252Fregistries%252Fosf%252Fnew.
26 Kemper JI, Palmer K, Mol BW, Li W. Individual participant data meta-analysis (IPD-MA) comparing oral misoprostol versus vaginal misoprostol for the induction of labour (CRD42020141111). In: *PROSPERO International Prospective Register of Systematic Reviews*. UK: Centre for Reviews and Dissemination, University of York; 2020.
27 Chan F, Li W, Palmer K, Mol B, Jones M. Individual Participant data meta-analysis (IPD-MA) comparing intracervical PGE-2 to balloon catheter and vaginal misoprostol. CRD42022303249. In: *PROSPERO International Prospective Register of Systematic Reviews*. UK: Centre for Reviews and Dissemination, University of York; 2022.
28 Patabendige M, Li W, Mol BW, Rolnik DL. Balloon catheter plus concurrent low-dose vaginal misoprostol versus low-dose vaginal misoprostol alone for induction of labour: individual participant

data meta-analysis of randomised controlled trials (CRD42023429368). In: *PROSPERO International Prospective Register of Systematic Reviews*. UK: Centre for Reviews and Dissemination, University of York; 2023.
29 Li W, Suke S, Wertaschnigg D, et al. Randomised controlled trials evaluating endometrial scratching: assessment of methodological issues. *Hum Reprod*. 2019;34(12):2372–2380.
30 Bordewijk EM, Wang R, van Wely M, et al. To share or not to share data: how valid are trials evaluating first-line ovulation induction for polycystic ovary syndrome? *Hum Reprod Update*. 2020;26 (6):929–941.
31 Carlisle JB, Loadsman JA. Evidence for non-random sampling in randomised, controlled trials by Yuhji Saitoh. *Anaesthesia*. 2017;72 (1):17–27.
32 Carlisle JB, Dexter F, Pandit JJ, Shafer SL, Yentis SM. Calculating the probability of random sampling for continuous variables in submitted or published randomised controlled trials. *Anaesthesia*. 2015;70(7):848–858.
33 Patabendige M, Chan F, Al-Ibraheemi Z, et al. Balloon catheter plus concurrent vaginal misoprostol versus vaginal misoprostol alone for cervical ripening in induction of labour: an individual participant data meta-analysis of randomised controlled trials. *Authorea*. 2024. https://doi.org/10.22541/au.173287343.33301642/v1. Pre-print. Accessed May 10, 2025.
34 De Angelis C, Drazen JM, Frizelle FA, et al. Clinical trial registration: a statement from the International committee of medical journal editors. *N Engl J Med*. 2004;351:1250–1251.
35 Mol BW, Lai S, Rahim A, et al. Checklist to assess Trustworthiness in RAndomised Controlled Trials (TRACT checklist): concept proposal and pilot. *Res Integr Peer Rev*. 2023;8(1):6.
36 Sterne JAC, Savović J, Page MJ, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. *BMJ*. 2019;366:l4898.
37 Tsujimoto Y, Fujii T, Onishi A, et al. No consistent evidence of data availability bias existed in recent individual participant data meta-analyses: a meta-epidemiological study. *J Clin Epidemiol*. 2020;118:107–114.e5.
38 Schünemann H, Brożek J, Guyatt G, Oxman A, eds. *GRADE Handbook for Grading Quality of Evidence and Strength of Recommendations. The GRADE Working Group*. Canada: McMaster University; 2013.
39 Schünemann HJ, Neumann I, Hultcrantz M, et al. GRADE guidance 35: update on rating imprecision for assessing contextualized certainty of evidence and making decisions. *J Clin Epidemiol*. 2022;150:225–242.
40 R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, AustriaR Foundation for Statistical Computing; 2022.
41 Sousa-Pinto B, Cerqueira JJ, Costa-Santos C, Carlisle JB, Loadsman JA, Goncalves H. *Package 'simdistr'. Version 1.0*. London: R Foundation for Statistical Computing; 2023. https://doi.org/10.32614/CRAN.package.simdistr.2023.
42 Balduzzi S, Rücker G, Schwarzer G. How to perform a meta-analysis with R: a practical tutorial. *Evid Based Ment Health*. 2019;22:153–160.
43 Ilyas A, Mir MK, Hanif S. Comparison of oral misoprostol with Pge2 gel for induction of labour in PROM at term with unfavourable Bishop score. *Pakistan J Med Health Sci*. 2016;10(2):409–412.
44 Kadu NA, Shiragur S. Comparison of intracervical Foley's catheter with vaginal misoprostol versus intravaginal misoprostol alone for cervical ripening and induction of labor. *Cureus*. 2023;15(9):e44772.
45 Kalambe M, Jungari ML, Nair PP, Shrivastava D. Combination of misoprostol with transcervical Foley's catheter compared to misoprostol alone for cervical ripening at term and labour induction in tertiary care hospital: a randomized trial. *Indian J Forensic Med Toxicol*. 2020;14(4):6836–6840.
46 Qazi Q, Fatima SS, Wahab S, Syed W. Efficacy of misoprostol and prostaglandin E2 gel for induction of labor in term pregnancy. *J Postgrad Med Inst*. 2019;33(4):331–336.
47 Yehia AH, Abd ElFattah IH, Bayoumy KM, et al. Titrated misoprostol versus dinoprostone for labor induction. *J Basic Clin Reprod Sci*. 2016;5(2):75–81.
48 Elkady MO. *Oral Misoprostol versus Vaginal Dinoprostone for Induction of Labour: a Randomized Controlled Clinical Trial*. Egypt: Ain Shams University; 2016.
49 Viteri OA, Tabsh KK, Alrais MA, et al. Transcervical Foley balloon plus vaginal misoprostol versus vaginal misoprostol alone for cervical ripening in nulliparous obese women: a multicenter, randomized, comparative-effectiveness trial. *Am J Perinatol*. 2021;38(S 01):e123–e128.
50 Deshmukh VL, Yelikar KA, Waso V. Comparative study of efficacy and safety of oral versus vaginal misoprostol for induction or labour. *J Obstet Gynaecol India*. 2013;63(5):321–324.
51 Souza AS, Feitosa FE, Costa AA, et al. Titrated oral misoprostol solution versus vaginal misoprostol for labor induction. *Int J Gynaecol Obstet*. 2013;123(3):207–212.
52 Chavakula PR, Benjamin SJ, Abraham A, Londhe V, Jeyaseelan V, Mathews JE. Misoprostol versus Foley catheter insertion for induction of labor in pregnancies affected by fetal growth restriction. *Int J Gynaecol Obstet*. 2015;129(2):152–155.
53 Loscalzo J. Irreproducible experimental results: causes, (mis)interpretations, and consequences. *Circulation*. 2012;125(10):1211–1214.
54 Taichman DB, Sahni P, Pinborg A, et al. Data sharing statements for clinical trials: a requirement of the International committee of medical journal editors. *Ann Intern Med*. 2017;167(1):63–65.
55 Naudet F, Sakarovitch C, Janiaud P, et al. Data sharing and reanalysis of randomized controlled trials in leading biomedical journals with a full data sharing policy: survey of studies published in the BMJ and PLOS Medicine. *BMJ*. 2018;360:k400–k.
56 Tan AC, Askie LM, Hunter KE, Barba A, Simes RJ, Seidler AL. Data sharing-trialists' plans at registration, attitudes, barriers and facilitators: a cohort study and cross-sectional survey. *Res Synth Methods*. 2021;12(5):641–657.
57 Ventresca M, Schünemann HJ, Macbeth F, et al. Obtaining and managing data sets for individual participant data meta-analysis: scoping review and practical guide. *BMC Med Res Methodol*. 2020;20(1):113.
58 Cascini F, Pantovic A, Al-Ajlouni YA, Puleo V, De Maio L, Ricciardi W. Health data sharing attitudes towards primary and secondary use of data: a systematic review. *EClinicalMedicine*. 2024;71:102551.
59 National Institutes of Health (NIH) data sharing policies. Available from: https://www.nlm.nih.gov/NIHbmic/nih_data_sharing_policies.html.
60 National Health and Medical Research Council (NHMRC). Principles for accessing and using publicly funded data for health research. Available from https://www.nhmrc.gov.au/sites/default/files/documents/reports/principles-publically-funded-data.pdf.
61 Wang R. 'And Then There Were None'—the shrinkage of trials in the evidence ecosystem. *Hum Reprod*. 2025;40(5):785–787.
62 OBGYN Editors' Integrity Group (OGEIG). Trustworthiness criteria for meta-analyses of randomized controlled studies: OBGYN journal guidelines. *Int J Gynaecol Obstet*. 2024;167(2):475–478.
63 Weibel S, Popp M, Reis S, Skoetz N, Garner P, Sydenham E. Identifying and managing problematic trials: a research integrity assessment tool for randomized controlled trials in evidence synthesis. *Res Synth Methods*. 2023;14(3):357–369.
64 Hunter KE, Aberoumand M, Libesman S, et al. The individual participant data integrity tool for assessing the integrity of randomised trials. *Res Synth Methods*. 2024;15(6):917–939.
65 Grey A, Bolland MJ, Avenell A, Klein AA, Gunsalus CK. Check for publication integrity before misconduct. *Nature*. 2020;577 (7789):167–169.
66 Weeks J, Cuthbert A, Alfirevic Z. Trustworthiness assessment as an inclusion criterion for systematic reviews—what is the impact on results? *Cochrane Evid Synth Methods*. 2023;10(1):e12037.
67 Alfirevic Z, Weeks J. Trustworthiness assessment of published clinical trials: literature review of domains and questions. *Cochrane Evid Synth Methods*. 2024;2(8):e12099.
68 Wilkinson J, Tovey D. How should we assess trustworthiness of randomized controlled trials? *J Clin Epidemiol*. 2025;180:111670.
69 Nevitt SJ, Marson AG, Davie B, Reynolds S, Williams L, Smith CT. Exploring changes over time and characteristics associated with data retrieval across individual participant data meta-analyses: systematic review. *BMJ*. 2017;357:j1390.
70 Jaspers GJ, Degraeuwe PL. A failed attempt to conduct an individual patient data meta-analysis. *Syst Rev*. 2014;3:97.
71 Flanagan M, Au LS, Patabendige M, et al. Misoprostol versus oxytocin for the prevention of postpartum haemorrhage: a systematic review and meta-analysis including individual participant data. *BJOG*. 2025:1–14. https://doi.org/10.1111/1471-0528.18197.
72 Rattan A, Flanagan M, Au LS, et al. The effectiveness of oxytocin for the prevention of postpartum haemorrhage: an individual participant data meta-analysis. *BJOG*. 2025 [In press].