How to Become a Data Synthesis Expert in IBD: A Clinician's Guide to Cutting Through the Noise

Author:

Morris Gordon, School of Medicine, University of Central Lancashire, UK, mgordon@uclan.ac.uk

Corresponding Author:

Professor Morris Gordon
Harrington Building HA340,
University of central Lancashire
Fylde Road,
Preston,
PR1 2HE
Telephone 01772 201201
Email: mgordon@uclan.ac.uk

Competing interests

None declared / No, there are no competing interests

Study materials

There are none

Writing assistance

The author used OpenAI's ChatGPT (GPT-4) to assist with drafting and refining this invited editorial, based solely on original content and dictation provided by the author. The AI tool was employed for language editing, structuring, and clarity enhancement only. No content was generated without direct author input, and all final decisions and revisions were made by the author. The AI tool was not used to generate citations, evidence summaries, or new scientific content.

Author Contributions

M. Gordon was the sole author and produced this work with no external input

Abstract

Research and evidence are evolving daily in the field of Inflammatory bowel disease (IBD). This represents a significant problem for the IBD clinical team or researchers in the field. Data synthesis is a useful tool to identify, select, appraise, and analyse research data to help solve this issue. However, it can seem to be a highly complex and therefore challenging form of research to read and apply. In this article, I have identified four key areas of awareness or learning to support IBD clinical and research teams in interpreting and using data synthesis. These include appraisal of clinical trials using structured bias domains, reading between the lines of forest plots, recognizing and understanding how certainty of findings is assessed and finally highlighting a key area of misunderstanding in network meta-analysis. Whilst by no means an exhaustive digest, this article should support in key areas and enhance uptake of evidence based IBD care.

Key points

- Appraisal of trial reports though the use of a risk of bias tool has high utility not just in systematic reviews but for the IBD team using such studies
- Forest plots give useful data on heterogeneity that is often overlooked
- When considering meta-analysis and forest plots, imprecision is also often overlooked and I propose clinical framing of the results to consider precision in clinical, not statistical, terms.
- GRADE analysis of certainty of outcomes is the pre-eminent technique and for synthesis of clinical trials in IBD is vital to allow findings to be used.
- Network meta-analyses commonly focus on ranking. This is a major flaw, not in the technique, but in its execution
- Instead, the certainty of findings is vital in network meta-analysis and this should be the priority over ranking alone.

Introduction: When More Is Less Clear

The volume of published research in medicine is expanding at an exponential rate—and inflammatory bowel disease (IBD) is no exception. Across a widening spectrum of focus, from advanced therapeutics and diagnostics to qualitative and health economic research, the sheer scale of information being generated is unprecedented. Added to this is the proliferation of journals and conferences, each with differing editorial standards and peer review quality.

In earlier eras, a more tightly focused community, coupled with fewer publication outlets, led to implicit, if sometimes unjustified, trust in what was published. That trust was grounded in community-based scrutiny—where most experts were actively involved in reviewing nearly every major output. Inadequate studies would be swiftly challenged. That system, however, has largely dissolved. Papers now appear in outlets with highly variable editorial oversight. The result is an ecosystem where poor reporting often goes unchallenged and misleading conclusions go unchecked.

In IBD, these concerns are acute. A recent study [1] found that only 1 in 10 IBD randomised controlled trials (RCTs) published in high-impact journals over the past three years were free of significant issues in reporting quality.

Compounding the problem, systematic reviews are themselves growing rapidly in number—and ignoring the wastefulness of this ,[2] they often contradict one another due to differences in study selection, analysis methods, and synthesis techniques.[3]

This article outlines key practical tools and critical concepts that any IBD clinician or clinical researcher can use to better appraise primary trials, systematic reviews, forest plots, and network meta-analyses. The goal is simple: not to turn every reader into a statistician, but to empower better clinical decisions with clearer evidence interpretation.

1. Reading Randomised Trials: A Crash Course in Risk of Bias

The Cochrane Risk of Bias (RoB) tool [4] —though now updated in newer versions which are as yet not widely adopted—remains an accessible and powerful starting point for evaluating trial quality. For the everyday IBD practitioner, mastering this tool can be transformative and allows for a rapid but meaningful critique of trials—especially those underpinning treatment decisions.

The RoB tool evaluates several domains of bias, which are critical to the internal validity of trial results. These are summarised and shown in Figure 1 and detailed below:

- Random Sequence Generation (Selection Bias):

The first key consideration is whether the randomisation method used in the trial was truly random. Acceptable methods include computer-generated random lists, random number tables, or low-tech solutions like dice, cards, or drawing lots. These methods ensure that each participant has an equal chance of being allocated to any group, maintaining the integrity of the comparison.

Conversely, quasi-random methods such as allocation based on hospital numbers, date of birth, clinic attendance date, or alternate allocation are not truly random. These introduce a risk of selection bias, as they may be predictable or manipulable.

If the trial describes a valid randomisation method, it can be considered at low risk for this element . However, if the method is not described at all or ambiguously noted as “random” without explanation, it should raise immediate concern.

- Allocation Concealment (Selection Bias):
Often confused with blinding, allocation concealment refers to the process by which the generated random sequence is protected during patient enrolment. This prevents clinicians or investigators from knowing or influencing the next assignment before a patient is entered into the trial. Historical and seminal work [5] identify that failure to describe this in a clear way was associated with A 17% inflation in the estimate of effects for the intervention. More recently, a larger and methodologically gold standard study [6] found compelling evidence to demonstrate that overestimation is probably still occurring in this situation suggesting that this is key in all published IBD trials. It is therefore disappointing that in IBD trials in the 5 years from 2016-20, 60% of studies were at unclear risk of bias with respect to allocation concealment.[1]

Effective methods include central randomisation (e.g. via telephone, web portal, or pharmacy-controlled allocation), or the use of sealed opaque envelopes or identical packaging. Without proper concealment, even in an open-label study, allocation can be manipulated or subconsciously influenced—compromising the balance intended by randomisation. This is key for all methods but is especially obvious with low tech methods. The presence or absence of this detail is a vital clue in determining the validity of the trial’s structure.

- Blinding (Performance and Detection Bias):
This domain is split into two: (1) Blinding of participants and personnel (performance bias), and (2) Blinding of outcome assessors (detection bias). Blinding has many effects but its impact on influencing perception of benefit and reporting of adverse events is key.

A trial should clearly describe how blinding was implemented—not just whether it was claimed. Was sham therapy used? Were identical containers, packaging, or infusion methods employed? If the intervention group receives intravenous therapy and the comparator receives oral tablets without a sham infusion, blinding even if proposed would

not appear feasible and should be considered with concern.

Similarly, outcome assessors must be blinded—especially when the outcome is subjective. If a pathologist or endoscopist is unaware of the treatment allocation, risk of detection bias is reduced. Patient-reported outcomes are inherently identical for both blinding types as the participant is the outcome assessor.

Importantly, the lack of blinding does not automatically imply poor quality if the outcome is objective (e.g. death, lab values), but unexplained or implausible claims of blinding should be viewed critically.

- Incomplete Outcome Data (Attrition Bias):
A high-quality trial will account for all patients, reporting flow through the study (e.g. CONSORT diagram), and providing reasons for drop-outs in each group. Balanced attrition between groups with non-differential reasons is generally acceptable. Of interest from the perspective of a data synthesist is that we often don't get concerned with the issues of Intention to treat or per protocol analysis as reported by a study author team. Instead, we will always consider intention to treat where those that do not have outcome data are assumed to be non-responders or sufferers of adverse events. This consideration therefore requires clear reporting of such data to allow the reader to make such interpretations.

However, significant or unexplained loss to follow-up—especially if unequal between groups—can bias results. Were patients lost because of poor treatment response or adverse effects? Without clarity, this domain should be treated with caution.
- Selective Reporting (Reporting Bias):
This domain addresses whether all prespecified outcomes (both efficacy and safety) were reported in the final manuscript. A trial that deviates from its protocol or omits key endpoints—especially negative or safety-related ones—may be cherry-picking results to favour a conclusion. Of more concern is the addition of post-hoc analyses that were not specified, often done when pre-specified findings are negative. Such analyses may not have appropriate statistical power, but more than that are a significant source of bias.

The presence of a pre-registered trial protocol or registry entry (e.g. ClinicalTrials.gov) helps determine whether the reported outcomes match the original plan.

- Other Bias:
This category captures imbalances or anomalies not covered by the prior domains. A common example is significantly different baseline characteristics between groups. If these differences are unexplained, they may reflect failed randomisation or introduce confounding or even be a reflection of inappropriate sample size,[7,8] which is not assessed independently.

Overall, the RoB tool is not just for systematic reviewers. When understood and applied, it

allows the everyday IBD clinician to quickly identify red flags, improving the ability to interpret trial findings meaningfully. Figure 1 is a quick tool to support IBD clinicians in considering these items when reading a trial.

2. Forest Plots: Seeing What Others Miss

Forest plots are often treated as the central visual output of a meta-analysis. They offer a striking, accessible representation of pooled results, showing the direction and size of effects across studies. But to interpret them meaningfully, we must look beyond the summary diamond and understand the broader context — particularly **heterogeneity**.

Heterogeneity describes the variability between studies. It can be caused by and conceptualised in a number of ways:

- **Clinical heterogeneity** refers to differences in populations, interventions, comparators, or settings.
- **Methodological heterogeneity** covers variations in study design, risk of bias, or outcome measurement.
- **Statistical heterogeneity**, often quantified using I^2 or Chi-squared test within a forest plot and meta-analysis and indicates whether variation in results is greater than would be expected by chance.

A forest plot does not explain these forms of heterogeneity on its own. It shows them — often as visually divergent confidence intervals or high I^2 values. —. Figure 2 exemplifies this showing both a high I^2 which is visually clear when considering the individual study estimates. but it doesn't help the reader understand why they exist or what to do about them. That's why interpreting a forest plot without considering the underlying reasons for heterogeneity can lead to misleading conclusions.

Moreover, visual symmetry or overlap in the forest plot may give a false sense of certainty. It's important to remember that consistency in visual alignment doesn't guarantee consistency in study quality or relevance. For example, combining high-risk-of-bias trials with low-risk ones might yield an overall impactful effect estimate, , but the true implication for practice may be unclear or even unhelpful because of heterogeneity of clinical or methodological origins. High quality systematic reviews account for this in transparent reproducible methods to extract data in a way to only perform analyses when this has been considered. They will also pre-plan sensitivity analyses for plausible and possibly expected sources of heterogeneity, such as concomitant therapies or study publication. When looking at a forest plot, always be reminded to think of clinical and methodological sources stated above.

Another key element seen in forest plots is the concept of precision. This is best seen within the overall estimate diamond, with the left and right of the diamond representing the

confidence intervals and thereby precision of the results. Figure 2 also shows this with a green arrow representing the width of the diamond and also indicating the underlying numbers that it refers to. Often the focus is simply on whether it crosses the line of no effect, but confidence intervals should be read as ranges of plausible effect sizes — and when they are wide, uncertainty is high, regardless of statistical significance. We would ask you to consider this in clinical terms within Figure 2 as an example. , The mean difference is 0.61 pain episodes per week. The reader could consider the clinical significance of this alone. But if the lower confidence interval of 0.24 pain episodes is considered, this is a much smaller difference. Would that be clinically significant to your patients? If not, this range of plausible effects within the confidence intervals would indicate imprecision. This approach of considering not just whether the result is 'significant' but the clinical interpretation of the confidence intervals is a highly useful approach for the clinician reader.

In summary, a forest plot is not the answer — it's part of the question. It's a starting point for deeper scrutiny: what's causing variation? Are studies truly combinable? What are the implications of observed inconsistency? Without considering these questions, the simplicity of a forest plot becomes a false comfort rather than a tool for insight.

3. The non-negotiables in systematic reviews: the right tools, properly used

There are some elements of systematic reviews that are simply not optional. Regardless of topic or scope, these are the minimum methodological standards that must be in place — otherwise, a review cannot be considered truly systematic. They are also vital to the reader to interpret the findings.

The first is the use of validated tools to appraise individual studies, one of which we have already discussed, the Risk of bias tool for RCTs.[4]. Other tools appropriate to appraise study design include:

- For non-randomised intervention studies: ROBINS-I.[9]
- For observational studies [10] assessing exposure or prognosis: ROBINS-E.[11]

These tools provide a structured framework to assess the strengths and limitations of each included study. Without this step, any synthesis of findings risks putting poor-quality or biased evidence on equal footing with high-quality research — distorting conclusions and misleading readers. Of particular note is often reviews will justify the lack of such tools as the 'evidence was of low quality, but this is akin to using a 'fractured lens to look at broken glass' and is completely invalid.[12]

Second, and equally essential, is the use of GRADE (Grading of Recommendations, Assessment, Development and Evaluation) to assess the certainty of evidence across studies for each outcome. GRADE has recently been updated and essentially relaunched as CORE-GRADE.[13] This is particularly useful development as the seven article series is very clear

and useful. It is out of the scope of this article to go into the specific details of GRADE and the use of GRADE without those with appropriate understanding should be discouraged as much as ignoring GRADE. However, I would particularly highlight on its clarification that GRADE includes two symbiotically linked but separate elements – GRADE assessment of certainty of outcomes assessed within systematic review and then the GRADE process to make recommendations in guidelines. For the purposes of this article, we are discussing the former element only.

GRADE forces reviewers to look beyond p-values and confidence intervals, asking whether the body of evidence is actually robust, consistent, direct, and precise enough to support real-world decisions. It introduces transparency into the interpretation process and should not be skipped — especially in reviews intending to influence policy or practice.

Unfortunately, many published reviews skip these steps or include only superficial appraisals. Merely stating that a tool was used, without explaining how the results were applied to interpretation, is not sufficient. These tools are not tick-boxes and certainly it is not a likert scale to be selected based on overall feeling or consensus decision making. It is a scientific and objective part of the synthesis, no different to the meta-analysis in that sense. GRADE is fundamental to making sense of what the evidence actually tells us. An example can be seen in a recent cochrane review on iron deficiency anaemia treatments in IBD.[14 An analysis of Oral Ferric Maltol to placebo found a relative risk of 73 (95% CI 4.6 – 1164.4) in favour of active therapy. This may lead you to conclude this is a highly clinically effective option. However, it is very low certainty on GRADE assessment due to serious risk of bias concerns and very serious imprecision. Without GRADE an inappropriate conclusion would be drawn.

Without these components — proper risk of bias assessment and a GRADE certainty judgement — a review is little more than a literature summary. These are the non-negotiables that uphold the rigour, credibility, and decision-making value of a systematic review.

4. Network meta-analyses: ranking isn't the goal

Network meta-analyses (NMAs) are powerful tools that allow comparisons across multiple treatments, even when direct head-to-head trials are lacking. They are particularly useful in areas where many interventions exist, but not every combination has been tested which explains their exponential growth in IBD. However, with this complexity comes a high risk of misinterpretation and misuse.

Too often, NMAs are reduced to league tables — simplistic rankings of treatments from "best" to "worst" — without critical interpretation. This overlooks the core purpose of an NMA, which is to compare interventions in a statistically coherent and clinically meaningful way, while fully acknowledging uncertainty.

It is not enough to rank interventions. A robust NMA also requires attention to transitivity (i.e. whether studies are similar enough to justify indirect comparisons).

It must also include a judgement of certainty for each comparison. That means using a tool like GRADE for network meta-analysis,[15] which assesses the credibility of the relative effects across the network. Without this, apparent differences may be statistical aberrations, driven by indirectness, imprecision, or inconsistency. Essentially, being ranked high in an NMA is irrelevant if the finding is so uncertain that it cannot be interpreted. A recent example is an NMA used by the British society of gastroenterology colorectal cancer screening guideline committee on the modalities of imaging.[16] This suggests that virtual chromoendoscopy is ranked bottom. However, this finding is very uncertain and this means that ranking cannot be used to draw any conclusions on chromoendoscopy. This therefore cannot be interpreted as the worst performing modality.

In short, if a network meta-analysis does not include a certainty judgement and focuses only on ranking, it is not fit for purpose. Clarity, transparency, and cautious interpretation are not optional in this complex form of evidence synthesis. Without them, NMAs risk becoming impressive-looking outputs that lack meaningful insight

Conclusion: Asking the Right Questions

For the busy IBD clinician—or the clinical researcher under pressure—data synthesis doesn't need to be a black box. By learning to:

- Appraise trials using structured bias domains,
- Read between the lines of forest plots,
- Evaluate systematic reviews for their appraisal and certainty claims,
- And critically reframe rankings in network meta-analysis…

you can protect yourself and your patients from misleading conclusions dressed in statistical authority (these are summarized in Figure 3).

Evidence synthesis is not about knowing everything. It's about asking the right questions—even when the answers look polished. This guide was designed to help you do just that.

References

1. Gordon M, Khudr J, Sinopoulou V, et al. Quality of reporting inflammatory bowel disease randomised controlled trials: a systematic review. BMJ Open Gastroenterol. 2024;11:e001337. doi: 10.1136/bmjgast-2023-001337.
2. Siontis K C, Hernandez-Boussard T, et al. Overlapping meta-analyses on the same topic: survey of published studies BMJ 2013;347:f4501. doi:10.1136/bmj.f4501
3. Doundoulakis I, Antza C, Apostolidou-Kiouti F, et al. Overview of systematic reviews of non–vitamin K Oral anticoagulants in atrial fibrillation: evidence of publication overlap. Circulation. 2018;11(12):e004769. https://doi.org/10.1161/CIRCOUTCOMES.118.004769.
4. Higgins JP, Altman DG, Gøtzsche PC, et al; Cochrane Bias Methods Group; Cochrane Statistical Methods Group. The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. BMJ. 2011 Oct 18;343:d5928. doi: 10.1136/bmj.d5928.
5. Wood L, Egger M, Gluud LL, et al. Empirical evidence of bias in treatment effect estimates in controlled trials with different interventions and outcomes: meta-epidemiological study. BMJ. 2008 Mar 15;336(7644):601-5. doi: 10.1136/bmj.39465.451748.AD.
6. Wang Y, Parpia S, Couban R, et al. Compelling evidence from meta-epidemiological studies demonstrates overestimation of effects in randomized trials that fail to optimize randomization and blind patients and outcome assessors. J Clin Epidemiol. 2024 Jan;165:111211. doi: 10.1016/j.jclinepi.2023.11.001.
7. Gordon M, Lakunina S, Sinopoulou V, Akobeng A. Minimum sample size estimates for trials in inflammatory bowel disease: A systematic review of a support resource. World J Gastroenterol. 2021 Nov 21;27(43):7572-7581. doi: 10.3748/wjg.v27.i43.7572. PMID: 34887650; PMCID: PMC8613748.
8. Iheozor‐Ejiofor Z, Lakunina S, Gordon M, et al. Sample‐size estimation is not reported in 24% of randomised controlled trials of inflammatory bowel disease: A systematic review UEG Journal, 2021;9:47-53
9. Sterne J A, HernÃ¡n M A, Reeves B C, et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions BMJ 2016;355 :i4919. doi:10.1136/bmj.i4919
10. Akobeng AK, Grafton-Clarke C, Abdelgadir I, et al. Gastrointestinal manifestations of COVID-19 in children: a systematic review and meta-analysis. Frontline Gastroenterology 2021;12:332-337.
11. Higgins JPT, Morgan RL, Rooney AA et al. A tool to assess risk of bias in non-randomized follow-up studies of exposure effects (ROBINS-E). Environment International 2024. doi: 10.1016/j.envint.2024.108602.
12. Gordon M, Guyatt G. Assessment of Evidence Quality in Inflammatory Bowel Disease Guidance: The Use and Misuse of GRADE. Gastroenterology. 159;4:1209-15
13. Guyatt G, Agoritsas T, Brignardello-Petersen R, et al. Core GRADE 1: overview of the Core GRADE approach BMJ 2025; 389 :e081903 doi:10.1136/bmj-2024-081903
14. Gordon M, Sinopoulou V, Iheozor-Ejiofor Z et al. Interventions for treating iron deficiency anaemia in inflammatory bowel disease. Cochrane Database of Systematic Reviews 2021;1: CD013529. DOI: 10.1002/14651858.CD013529.pub2.

15. Iheozor‐Ejiofor Z, Gordon M, et al. Interventions for maintenance of surgically induced remission in Crohn's disease: a network meta‐analysis. Cochrane Database of Systematic Reviews 2019;9:CD013210. DOI: 10.1002/14651858.CD013210.pub2. Accessed 16 July 2025.
16. Sinopoulou V, Nigam GB, Gordon M, et al. Comparative Efficacy and Safety of Endoscopic Modalities for Colorectal Cancer Screening in Inflammatory Bowel Disease: A Systematic Review and Network Meta-analysis. Clin Gastroenterol Hepatol. 2024:S1542-3565(24)01073-5. doi: 10.1016/j.cgh.2024.11.008.

Figure Legends

Figure 1. Summary tool for considering risk of bias in randomised controlled trials

Figure 2. Example forest plot with highlighted heterogeneity and imprecision elements discussed in text.

Figure 3. Infographic describing the 4 data synthesis key points to consider by IBD clinicians