On the twelfth day of Christmas, the BMJ statistician sent to me …

Richard D Riley[1 #], ORCID: 0000-0001-8699-0735

Tim J Cole[2], ORCID: 0000-0001-5711-8200

Jon Deeks[1], ORCID: 0000-0002-8850-1971

Jamie J Kirkham[3], ORCID: 0000-0003-2579-9325

Julie Morris[4], ORCID: 0000-0003-4941-4645

Rafael Perera[5], ORCID: 0000-0003-2418-2091

Angie Wade[6], ORCID: 0000-0002-4823-9219

Gary S Collins[7,8], ORCID: 0000-0002-2772-2316

Author details:

[#] corresponding author – email: r.riley@keele.ac.uk twitter: @Richard_D_Riley

[1] Institute of Applied Health Research, College of Medical and Dental Sciences, University of Birmingham, Birmingham, UK.

[2] UCL Great Ormond Street Institute of Child Health, London, UK.

[3] Centre for Biostatistics, The University of Manchester, Manchester Academic Health Science Centre, Manchester, United Kingdom.

[4] Honorary Reader, University of Manchester, Oxford Road, Manchester, UK. M13 9PL.

[5] Nuffield Department of Primary Care Health Sciences, University of Oxford, Oxford, UK.

[6] UCL Great Ormond Street Institute of Child Health, London UK.

[7] Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, University of Oxford, Oxford, OX3 7LD, UK.

[8] NIHR Oxford Biomedical Research Centre, Oxford University Hospitals NHS Foundation Trust, Oxford, United Kingdom.

Word count: 4500

Acknowledgements: This article is dedicated to Doug Altman and Martin Gardner, who led by example as Chief Statistical Editors at the BMJ for over thirty years. We would also like to thank all the researchers that have responded politely to our statistical reviews over many years, and acknowledge the reviewers of this article who provided helpful comments for improvement.

Funding: This article did not receive any specific funding.

Ethics: Ethics approval was not required.

Competing interests: We have read and understood the BMJ Group policy on declaration of interests and declare we have no competing interests.

Dissemination: We plan to disseminate the published article via social media and refer to it in our statistical reviews.

Patient and Public Involvement: Patients or the public were not involved in the design, or conduct, or reporting, or dissemination of our research.

Data sharing: Data sharing not applicable as no datasets generated and/or analysed for this study

Contribution statement: RR conceived the paper, and generated the Christmas theme and initial set of twelve stocking fillers. RR and GC produced the first draft of the article, including the exploratory text for each item and examples. All authors provided comments and suggested changes, which were then addressed by RR and GC. RR revised the article following reviewer comments, followed by suggestions and final approval by all authors.

Standfirst

The *BMJ Statistical Editors* need a quiet Christmas. They review hundreds of articles each year, working tirelessly to improve standards of statistical analysis and reporting. Many articles exhibit the same problems, and therefore the Statistical Editors have come together to tell you about them. Make their wish come true. All they want for Christmas is due.

Introduction

The weeks leading up to Christmas are a magical time for medical research. The impending holiday season creates a dramatic upsurge in productivity, with researchers suddenly finding time to finish off statistical analyses, draft manuscripts, and respond to reviewers. This leads to a plethora of submissions to journals such as the *BMJ* throughout December, so that researchers can finish the year with a sense of academic accomplishment and enjoy the festivities with their loved ones. Indeed, with optimism fuelled by mulled wine and mince pies, they may even anticipate their article's acceptance by early January, at the end of the twelve days of Christmas.

However, there is a collective who work against this season of publication goodwill and cheer. A small but influential group of individuals with a very shiny nose for detail; seeking 'all is right' rather than 'all is bright'; and emphasising No, No, No rather than Ho, Ho, Ho. We call ourselves *statisticians*, and our core belief is that a research article is for life, not just for Christmas. Our key role is to deliver statistical reviews that promote high standards of methodological rigour and transparency, and we are especially busy over the Christmas period with the influx of new submissions. Indeed, before we can eat, drink and be merry, we are working flat out to detect submissions with erroneous analysis methods that should be roasting on an open fire; dubious statistical interpretations as pure as yellow snow; and half-baked reporting of study details that bring zero comfort and joy. Bah humbug!

At the *BMJ*, the team of statisticians are called the *Statistical Editors*, and each year we review over 500 articles. For about thirty years, the *BMJ* Statistical Editors were led by pioneers Professor Martin Gardner and Professor Doug Altman,[1 2] who saw statisticians to be like the Christmas star – lighting a path of research integrity, promoting methodology over metrics,[3 4] and encouraging statistical principles to "save science and the world".[5] With this vision in mind, here we present the results of an internal survey to identify some common issues encountered during statistical peer review at the *BMJ*. Twelve items are identified, one for each of the Twelve Days of Christmas, the period between 25th December to 5th January when we conduct our reviews in the critical mindset of the Grinch,[6] but with the kind heart of Miracle On 34th Street.

Methods

The *BMJ* Statistical Editors meet for the day every December, where we discuss common statistical issues, problematic submissions (including ones we refer to as 'sin-bin articles' that slipped through our net) and how to improve the review process, before unwinding at the *BMJ* Christmas party. At the meeting on 18th December 2019, the *BMJ* Statisticians agreed that an article showcasing common statistical issues would be helpful for authors of future *BMJ* submissions, and an initial set of items was discussed. The first author reminded the others about this article at subsequent Christmas meetings on 17th December 2020 and 16th December 2021, and explained that progress was being delayed, ironically due to the number of statistical reviews in the *BMJ* system that needed prioritizing.

After further procrastination, on 28th June 2022, a potential list of items was shared amongst the Statistical Editors via email, and everyone was asked to suggest additional irksome issues they encounter during statistical review. These were collated by the first author, and email discussions used to agree a final list of the most important items for wider dissemination. Twelve items were selected to match the number of days of Christmas in the well-known song, to increase the chances of publication in the *BMJ* Christmas issue. Sensitivity analyses, including shallow and deep learning approaches, led to the same twelve items being selected. An automated artificial intelligence algorithm quickly identified that all Statistical Editors were guilty of having similar statistical issues in some of their own research articles.

Results

The twelve items identified by us, the *BMJ* Statistical Editors, are listed below including a brief explanation to help drive them home for Christmas. Consider them as twelve stocking fillers. For maximum impact, we write them directed at you (the *BMJ* reader and potential co-author of future submissions). We suggest you digest one item per day, between 25th December and 5th January, and make a New Year's resolution to adhere to the guidance provided.

On the first day of Christmas, the *BMJ* statistician sent to me:

"1. Clarify the research question"

Christmas is a time to reflect on the meaning of life and to clarify your goals. Similarly, our reviews will encourage you to reflect on your research question and to clarify your objectives. For instance, your observational study may be unclear about the extent to which the focus is descriptive or causal

research; prognostic factor identification or prediction model development; exploratory or confirmatory research. For causal research, your underlying premise (causal pathway or model) may not be formally expressed (e.g., in terms of a directed acyclic graph, commonly referred to as a DAG). In systematic reviews of intervention studies, your research question might not be stated in terms of the Population, Intervention, Comparison and Outcome (PICO) system.

A related request is to clarify your estimand, which refers to your study's target measure for estimation.[7] For example, in a randomised trial the estimand is a treatment effect, but we may ask you to better define this in terms of the population, treatments being compared, outcomes, summary measure (e.g. risk ratio or risk difference; conditional or marginal effect), and other aspects.[7 8] Similarly, in a meta-analysis of randomised trials, your estimand must be defined in the context of potential heterogeneity of study characteristics. For example, in a meta-analysis of hypertension trials with different lengths of follow-up, if the estimand is a treatment effect on blood pressure, we need clarity about whether this relates to one particular time-point (e.g. 1 year), each of multiple time-points (e.g. 1 year and 5 years), or some average across a range of time-points (e.g. 6 months to 2 years).

On the second day of Christmas, the *BMJ* statistician sent to me:

"2. Focus on estimates, confidence intervals and clinical relevance"

Just like an under-cooked turkey, your article will be sent back if it focuses solely on *p*-values and 'statistical significance' to determine whether a finding is important. It is more important for you to consider estimates (e.g., of mean differences, risk ratios, or hazard ratios corresponding to the specified estimands from the first day of Christmas), their corresponding 95% confidence intervals, and the potential clinical relevance of your findings. Statistical significance often does not equate to clinical significance; for example, if a large trial gives a risk ratio of 0.97 with 95% confidence interval of 0.95 to 0.99, then the treatment effect is potentially small, even though the p-value is much less than 0.05. Conversely, absence of evidence does not mean evidence of absence;[9] for example, if a small trial estimates a risk ratio of 0.70, with 95% confidence interval from 0.40 to 1.10, then the magnitude of effect is still potentially large, even though the *p*-value is greater than 0.05. Hence, we will ask you to clarify phrases such as 'significant finding', be less definitive when confidence intervals are wide, and consider results in the context of clinical relevance or impact. A Bayesian approach may be helpful,[10] in order to express probabilistic statements (e.g. there is a probability of 0.85 that the risk ratio is less than 0.9).

On the third day of Christmas, the *BMJ* statistician sent to me:

"3. Carefully account for missing data"

Missing values occur in all types of medical research,[11] for covariates and outcomes, and are likely to be present in your research study. Therefore, we need you to not only acknowledge the completeness of your data but also quantify the amount of missing data and explain how missing data are handled in your analyses. It is incredible how many submissions fail to do this! It is the ghost of Christmas articles past, present and future.

If it transpires you simply excluded participants with missing data (i.e. you carried out a complete-case analysis), we may ask you to revise your analyses by including participants with missing values, using an appropriate approach for imputing the missing values. A complete-case analysis is rarely recommended, especially in observational research, as discarding patients usually reduces statistical power and precision to estimate relationships, and may also lead to biased estimates.[12] The best approach for imputation is context specific, and too nuanced for detailing here. For example, for randomised trials missing baseline values may be dealt with by replacing with the mean value (for continuous variables), creating a separate category of a categorical predictor to indicate the presence or a missing value (i.e. the missing indicator method) or multiple imputation performed separately by randomised group.[13 14] For observational studies examining associations, mean imputation and missing indicator approaches can lead to biased results,[15] and so a multiple imputation approach is often (though not always[16]) preferred, where missing values are imputed (on multiple occasions to reflect the uncertainty in the imputation) conditional on the observed values of other study variables.[17] When using multiple imputation, tell us the methods you used to do this, including the set of variables used in the imputation process. An introduction to multiple imputation is given in the *BMJ* by Sterne et al.,[12] and dedicated textbooks on missing data should be consulted.[18]

On the fourth day of Christmas, the *BMJ* statistician sent to me:

"4. Do not dichotomise continuous variables"

Santa likes dichotomisation (you are either naughty or nice), but we will be appalled if you choose to dichotomise continuous variables, such as age or blood pressure, by splitting them into two groups defined by being above and below some arbitrary cut-point (e.g. age 60 years, systolic blood pressure of 130 mmHg). Dichotomisation should be avoided,[19 20] as it wastes information and is rarely justifiable compared to analysing continuous variables on their continuous scale (see our stocking filler on the fifth day of Christmas). Why should an individual whose value is just below the cut-point (e.g. 129 mmHg) be considered completely different from an individual whose value is just

above it (e.g. 131 mmHg)? Conversely, two individuals within the same group may have very different values (e.g. 131 and 220 mmHg) and so why should they be considered the same? In this context, dichotomisation might be considered unethical. Study participants agree to contribute their data for research on the proviso it is used appropriately; discarding information by dichotomising their covariate values violates this agreement.

Dichotomisation also reduces statistical power to detect associations between the continuous covariate and the outcome,[19-21] and attenuates the predictive performance of prognostic models.[22] In one example, dichotomising at the median value led to a reduction in power akin to discarding a third of the data,[23] whilst in another retaining the continuous scale explained 31% more outcome variability than dichotomising at the median.[20] Cut-points also lead to data-dredging and the selection of 'optimal' cut-points to maximise statistical significance.[21] This leads to bias, lack of replication in new data, and hinders meta-analysis because different studies adopt different cut-points. Dichotomisation of continuous outcomes also reduces power and may lead to misleading conclusions.[24 25] For example, in a randomised trial published in the *BMJ*, the required sample size was reduced from 800 to 88 patients after changing the outcome (Beck score) from being analysed as dichotomised to being analysed on its continuous scale.[26]

On the fifth day of Christmas, the *BMJ* statistician sent to me:

"5. Consider non-linear relationships"

At the Christmas meal table, some family relationships are simple to handle, but others are more complex and require greater care. Similarly, some continuous covariates have a simple linear relationship with an outcome (perhaps after some transformation of the data, such as a natural log transformation) but others have a more complex non-linear relationship. A linear relationship (association) assumes that a one-unit increase in the covariate has the same effect on the outcome across the entire range of the covariate's values. For example, it assumes the impact of a change in age from 30 to 31 years is the same as a change from 90 to 91. In contrast, a non-linear association allows the impact of a 1-unit increase in the continuous covariate to vary across the spectrum of predictor values. For example, a change in age from 30 to 31 years may have little impact on risk, whereas a change in age from 90 to 91 may be very important. The two most common approaches to non-linear modelling are cubic splines and fractional polynomials.[27-32]

Aside from categorisation, most *BMJ* submissions only consider linear relationships. Our statistical reviews, therefore, may ask you to consider non-linear relationships, as otherwise important associations may not be fully captured or even missed.[33] An example of examining non-linear

relationships is by Johannesen et al.,[34] who use restricted cubic splines to show that the association between low density lipoprotein cholesterol levels and the risk of all-cause mortality is U-shaped, with low and high levels associated with an increased risk of all-cause mortality in the general population in Denmark. This is illustrated in **Figure 1** for the overall population, and in subgroups defined by use of lipid lowering treatment, with the relationship strongest in those not receiving treatment.

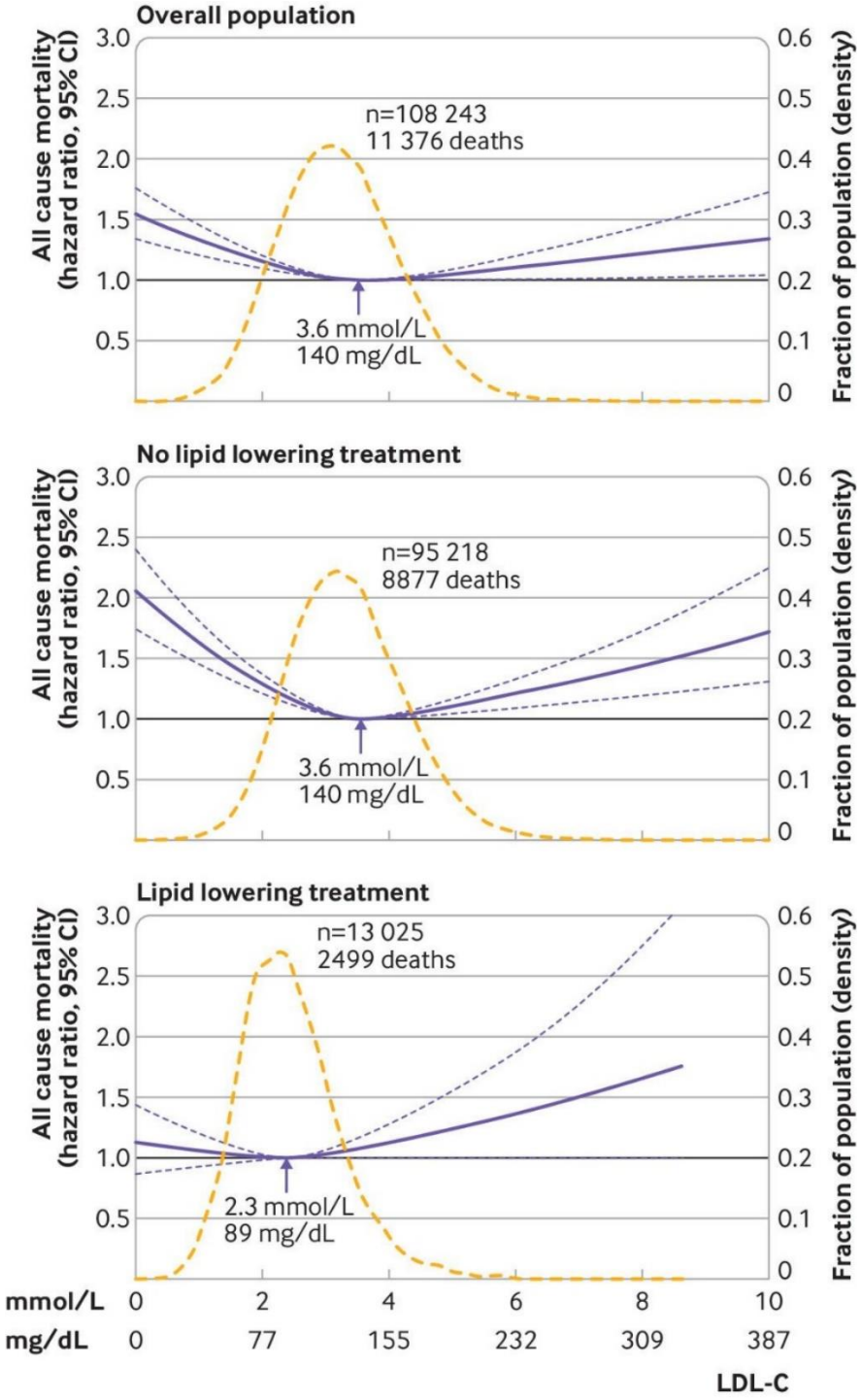
On the sixth day of Christmas, the *BMJ* statistician sent to me:

"6. Quantify differences in subgroup results"

Many submitted articles include results for subgroups, such as male and females, or those who do and do not eat Brussels sprouts. A common mistake is to conclude results for one subgroup are different from another subgroup, without actually quantifying the *difference* in their results. Altman and Bland consider this eloquently,[35] showing treatment effect results for two subgroups, the first of which is statistically significant (risk ratio = 0.67; 95% CI: 0.46 to 0.98; p-value = 0.03) while the second is not (risk ratio = 0.88, 95% CI: 0.71 to 1.08; p-value = 0.2). A naïve interpretation is to conclude the treatment is beneficial for the first subgroup but not the second. However, actually comparing the results between males and females reveals a wide confidence interval (ratio of risk ratios = 0.76; 95% CI: 0.49 to 1.17, p-value = 0.2), which suggests further research is needed before concluding a subgroup effect. A related mistake is to conclude that there is no difference in subgroups based solely on whether their 95% confidence intervals overlap.[36] Hence, if your study examines subgroups, we will check that you quantify differences in subgroup results and, if not, ask for this to be done. Even when there are genuine subgroup differences, the (treatment) effect may still be important for each subgroup and your conclusions should recognise this.

Examining subgroup differences is a complex issue, and a broader topic is the modelling of interactions between (treatment) effects and covariates.[37] Issues include the scale used to measure the effect (e.g. risk ratio or odds ratio);[38] ensuring subgroups are not arbitrarily defined by dichotomising a continuous covariate,[39] and allowing for potentially non-linear relationships (see our stocking fillers on the fourth and fifth days of Christmas).[40]

Figure 1 Example of a non-linear association derived by restricted cubic splines, as originally presented by Johannesen et al. in the *BMJ*[34] The figure shows multivariable adjusted hazard ratios for all-cause mortality according to levels of low density lipoprotein cholesterol (LDL-C) on a continuous scale. Solid blue lines are multivariable adjusted hazard ratios, with dashed blue lines showing 95% confidence intervals derived from restricted cubic spline regressions with three knots. Reference lines for no association are indicated by the solid bold lines at a hazard ratio of 1.0. Dashed yellow curves show the fraction of the population with different levels of LDL-C. Arrows indicate the concentration of LDL-C with the lowest risk of all-cause mortality. Analyses were adjusted for age, sex, current smoking, cumulative number of pack years, systolic blood pressure, lipid lowering treatment, diabetes, cardiovascular disease, cancer, and chronic obstructive pulmonary disease at baseline. Based on individuals from the Copenhagen General Population Study followed for a mean 9.4 years



On the seventh day of Christmas, the *BMJ* statistician sent to me:

"7. Consider accounting for clustering"

At the *BMJ* Christmas party, the *BMJ* Statistical Editors tend to cluster together in a corner, avoiding interaction and eye-contact with non-statisticians where possible in fear of being asked to conduct a post-mortem examination of their failed study. Similarly, your research study may contain data from multiple clusters, including observational studies that use e-health records from multiple hospitals or practices; cluster or multicentre randomised trials;[41-46] and meta-analyses of individual participant data (IPD) from multiple studies.[47] Sometimes the analysis does not account for this clustering, which may lead to biased results or misleading confidence intervals.[48-51] Ignoring clustering makes a strong assumption that outcomes for individuals within different clusters are similar to each other (e.g. in terms of the outcome risk), which may be difficult to justify when clusters such as hospitals or studies have different clinicians, procedures, and patient case-mix.

Thus, if your submitted article ignores obvious clustering in your data which needs to be captured or considered, we will ask you to either justify this, or else to re-analyse accounting for clustering using an approach suitable for your study's estimand of interest (see our stocking filler on the first day of Christmas).[52-54] For example, a multi-level or mixed-effects model may be recommended, as this allows cluster-specific baseline risks to be accounted for, and enables between-cluster heterogeneity in the effect of interest to be examined.

On the eighth day of Christmas, the *BMJ* statistician sent to me:

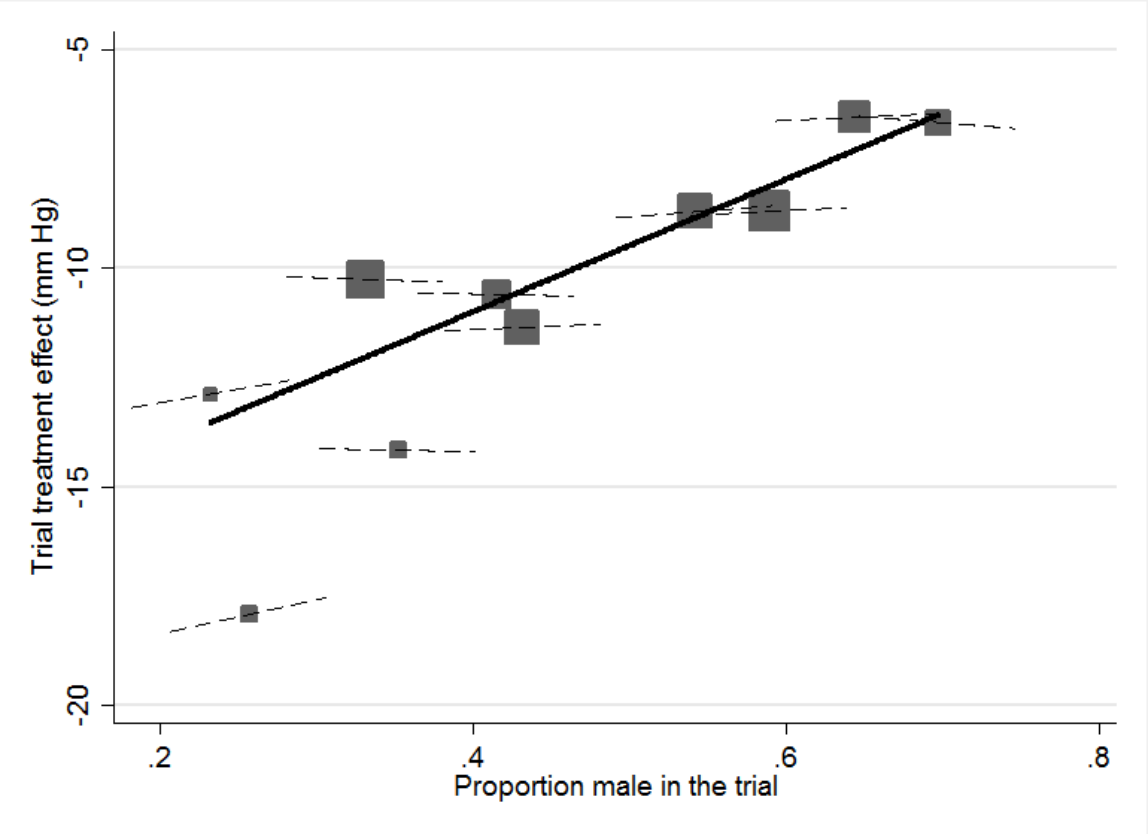
"8. Interpret I-squared and meta-regression appropriately"

Systematic reviews and meta-analyses are popular submissions to the *BMJ*. Most of them include I^2,[55] but interpret it incorrectly, which gives us a recurring Nightmare Before (and after) Christmas. I^2 describes the percentage of variability in (treatment) effect estimates that is due to between-study heterogeneity rather than chance. The impact of between-study heterogeneity on the summary treatment effect estimate is small if I^2 is close to 0%, and large if I^2 is close to 100%. A common mistake is for authors to interpret I^2 as a measure of the (absolute) amount of heterogeneity (i.e. to consider I^2 as an estimate of the between-study variance in true effects), and to erroneously use it to decide whether to use a random-effects meta-analysis model. This is unwise, as I^2 is a relative measure and depends on the size of the within-study variances of effect estimates, not just the size of the between-study variance of true effects (also known as tau-squared). For example, if all studies are small and thus within-study variances of effect estimates are large, I^2 can

be close to 0% even when the between-study variance is large and important.[56] Conversely, I^2 may be large even when the between-study variance is small and unimportant. Our reviews will ask you to correct any misuse of I^2, and to also present the estimate of between-study variance directly.

Meta-regression is often used to examine to what extent study-level covariates (e.g. mean age, dose of treatment, risk of bias rating) explain between-study heterogeneity, but generally we will ask you to interpret meta-regression results cautiously.[57] Firstly, there are often a small number of trials, and then meta-regression suffers from low power to detect study-level characteristics that are genuinely associated with changes in the overall treatment effect in a trial. Secondly, confounding across trials is likely, and so making causal statements about the impact of trial-level covariates is best avoided. For example, those trials with a higher risk of bias may also have the highest dose or be conducted in particular countries, thus making it hard to disentangle the effect of risk of bias from the effect of dose and country. Thirdly, the trial-level association of aggregated participant-level covariates (e.g. mean age, proportion male) with the overall treatment effect should not be used to make inferences about how values of participant-level covariates (e.g. age, sex, biomarker values) interact with treatment effect. Aggregation bias may lead to dramatic differences in observed relationships at the trial level from those at the participant level.[58 59] This is demonstrated in **Figure 2.**

Figure 2 Example of aggregation bias when using a meta-regression of study-level results rather than an individual participant data (IPD) meta-analysis of treatment-covariate interactions. The research question is whether blood pressure lowering treatment is more effective amongst women than men. Evidence is shown from a meta-analysis of 10 trials of anti-hypertensive treatment, comparing the across-trial association of treatment effect and proportion male (solid line) – which is steep and statistically significant – with the participant-level interactions of sex and treatment effect in each trial (dashed lines) - which are flat and neither clinically nor statistically important. Case study based on that previously reported by Riley et al.[47 58 60]



Each block represents one trial, and the block size is proportional to the size of the trial. Across-trial association is denoted by gradient of solid line (—), derived from a meta-regression of the trial treatment effects against proportion male, which suggests a large effect of a 15 mmHg (95% CI: 8.8 to 21) greater reduction in SBP in trials with only females compared to only males. However, the treatment-sex interaction based on participant-level data is denoted by gradient of dashed lines (- - -) within each trial, and on average these suggest only a 0.8 mmHg (95% CI: -0.5 to 2.1) greater treatment effect for females than males, which is neither clinically nor statistically significant.

On the ninth day of Christmas, the *BMJ* statistician sent to me:

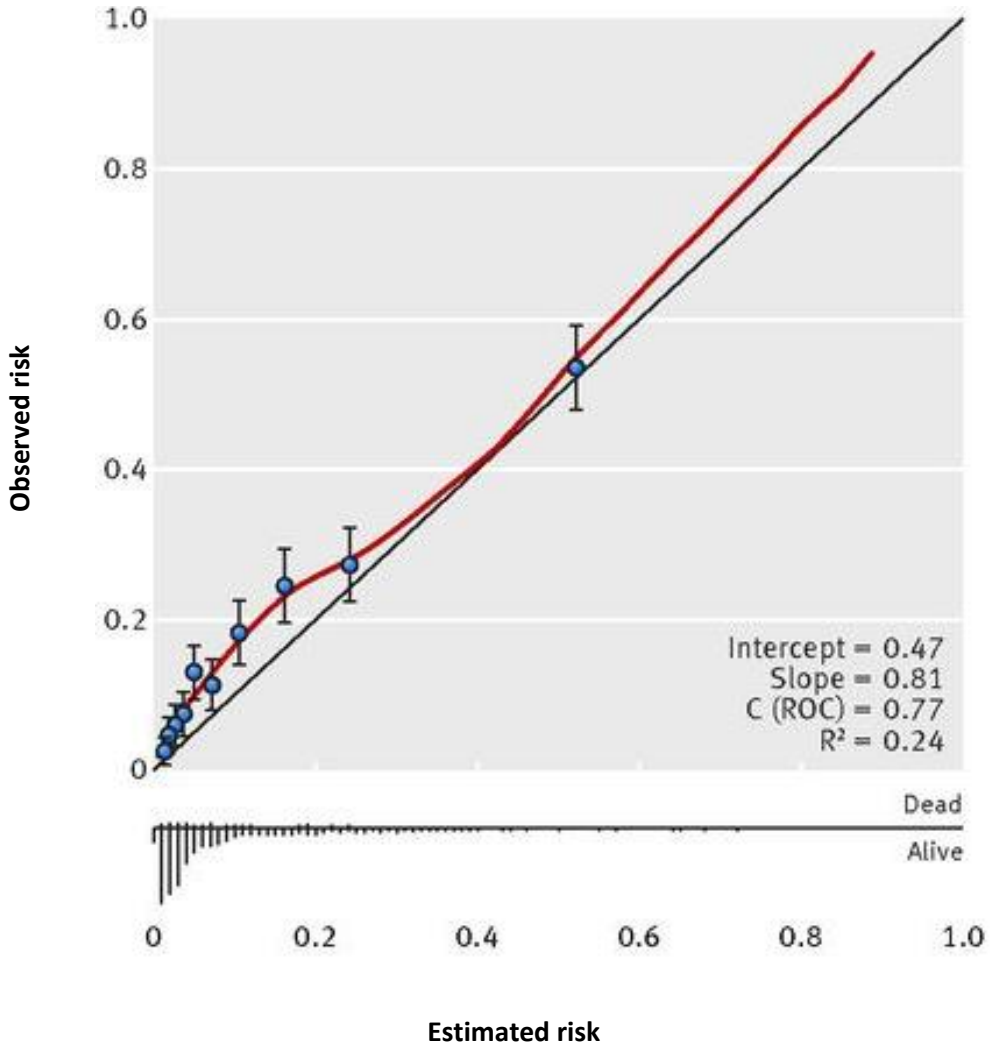
"9. Assess calibration of model predictions"

Clinical prediction models estimate outcome values (for continuous outcomes) or outcome risks (for binary or time-to-event outcomes) to inform diagnosis and prognosis in individuals. Articles developing or validating prediction models often fail to fully evaluate model performance, which is dangerous, as inaccurate predictions can lead to incorrect decisions and wrong communication to patients (e.g., giving false reassurance or hope). For models that estimate outcome risk, predictive performance should be evaluated in terms of discrimination, calibration and clinical utility, as described in previous *BMJ* articles.[61-63]

However, the overwhelming majority of submissions focus only on model discrimination (e.g. as quantified by the c-statistic or area under the curve[28]) – if you do this, it provides an incomplete picture, just like that unfinished 10000-piece jigsaw from last Christmas. For example, **Figure *3*** shows a calibration plot published in the *BMJ* for a prediction model with a promising c-statistic of 0.81, but there is clear (albeit perhaps small) miscalibration of predicted risks in the range of predicted risks between 0.05 to 0.2.[64] This miscalibration may impact upon the clinical utility of the model, especially if decisions (e.g. about treatment or monitoring strategies) are dictated by risk thresholds in the range of 0.05 to 0.2, which can be investigated in a decision curve analysis.[65] Conversely, miscalibration does not necessarily indicate the model has no clinical utility, as it depends on the magnitude of miscalibration and where it occurs in relation to decision thresholds.

We may also suggest model development studies undertake a re-analysis using penalization or shrinkage methods (e.g. ridge regression, lasso, elastic net), which reduce the potential for overfitting and help improve calibration of predictions in new data.[66 67] Penalisation methods, such as Firth's correction,[68] can also be important in trials or observational studies with sparse data, as standard methods (such as logistic regression) may give biased effect estimates in this situation.[69]

Figure 3 Example of a calibration plot to examine the agreement between the observed risks and estimated (predicted) risks from a prediction model. This figure is taken from Jaja et al.[64] who developed prediction models to estimate the risk of mortality in individuals who have suffered subarachnoid haemorrhage from ruptured intracranial aneurysm. The blue circles are the estimated and observed risks grouped by tenths of estimated risks, and the red line is a loess smoother to capture agreement across the range of estimated risks.



On the tenth day of Christmas, the *BMJ* statistician sent to me:

"10. Carefully consider the variable selection approach"

The use of variable selection methods (e.g. particularly forward selection of covariates based on the statistical significance of their effects) is a common area of criticism in our reviews.[70] If you use them, we will ask you to justify your approach. Depending on the study, we might even suggest you avoid them entirely, just like that last remaining turkey sandwich on New Year's Day.

For example, variable selection methods are best avoided in prognostic factor studies, as the typical aim is to provide an unbiased estimate of how a particular factor adds prognostic value over and above other (established) prognostic factors.[71] Therefore, a regression model forcing in all the existing factors is needed to examine the prognostic effect of the new factor after accounting for the effect of existing prognostic factors. Similarly, in causal research based on observational data, the choice of confounding factors to include as adjustment factors should be selected based on the causal pathway, for example as expressed using DAGs (with consideration of potential mediators between covariates and outcome [72]), not their statistical significance based on automated selection methods.

In the development of clinical prediction models, variable selection (via shrinkage) may be incorporated using methods such as lasso or elastic net, which start with a full model including all candidate predictors for potential inclusion. A common, but inappropriate approach is to use univariable screening, where decisions for predictor inclusion are based on *p*-values for observed unadjusted effect estimates. This is not a sensible strategy,[73] as what matters is the effect of a predictor after adjustment for other predictors, since in practice the relevant predictors are used (by healthcare professionals and patients) in combination. For example, when developing a prognostic model for risk of recurrent venous thromboembolism, Ensor et al. found that the unadjusted prognostic effect of age was not statistically significant from univariable analysis, but that the adjusted effect was significant and in the opposite direction from multivariable analysis.[74]

On the eleventh day of Christmas, the *BMJ* statistician sent to me:

"11. Assess the impact of any assumptions"

Everyone agrees that *It's A Wonderful Life* is a Christmas movie, but there is much debate about whether *Die Hard* is. Similarly, we may debate your die-hard analysis assumptions, and even ask you to examine whether results change if the assumptions change (a sensitivity analysis). For example, in submitted trials with time-to-event data (e.g. time to recurrence or death), it is common to report the hazard ratio, assuming it is a constant over the whole follow-up period. If your submission does not justify this assumption, we may ask you to address this, for example by graphically presenting how the hazard ratio changes over time (perhaps based on a survival model that includes an interaction between the covariate of interest and (log) time).[75] Another example is in submissions with Bayesian analyses, where prior distributions are labelled as 'vague' or 'non-informative', but we think they may still be influential. In this situation, we may ask you to demonstrate how results change when other plausible prior distributions are chosen.

On the twelfth day of Christmas, the *BMJ* statistician sent to me:

"12. Use reporting guidelines and avoid overinterpretation"

Doug Altman said, "*Readers should not have to infer what was probably done, they should be told explicitly. Proper methodology should be used and be seen to have been used*".[76] Incompletely reporting your research is indefensible and creates confusion, just like those unlabelled presents under the Christmas tree. We need to know your rationale and objectives, the study design, the methods used, the participant characteristics, the results, the certainty of evidence, the research implications, and so forth. If any of these aspects are missing, we will ask you to clarify them.

Make use of reporting guidelines. They provide a checklist of items to be reported (Santa suggests checking this list twice), which represent the minimum detail required to enable readers (including Statistical Editors) to understand the research and critically appraise its findings. Reporting guidelines are listed on The EQUATOR Network website, which maintains a comprehensive collection of guidelines and other materials related to health research reporting.[77] Examples are given in **Table *1***, including the CONSORT statement for randomised trials,[78] and the TRIPOD guideline for prediction model studies.[79 80] The *BMJ* requires you to complete the checklist found within the relevant guideline (and include it with your submission), indicating on which page of your submitted manuscript you have reported each item.

Another common aspect of our reviews, related to reporting, is to query overinterpretation of findings, and even spin,[81] for example, about unjustified claims of causality, generalisability of results, or immediate implications for clinical practice. Incorrect terminology is another bugbear, in particular the misuse of multivariate (rather than multivariable) to refer to a regression model with multiple covariates (variables), and the misuse of quantiles to refer to groups rather than the cut-points used to create the groups (e.g. deciles are the nine cut-points used to create 10 equal size groups called tenths).[82]

Table 1 Examples of reporting guidelines and their extensions for different study designs

Study design	Reporting guideline	Extensions available for some other common designs
Randomised trials	CONSORT	Cluster trials (CONSORT-Cluster); multi-arm trials; non-inferiority/equivalence trials (CONSORT non-inferiority); harms (CONSORT-HARMS); pilot and feasibility trials; adaptative designs (ACE Statement); artificial intelligence (CONSORT-AI); interventions (TIDieR)
Observational studies	STROBE	Genetic associations (STREGA); molecular epidemiology (STROBE-ME); infectious diseases (STROBE-ID); nutritional epidemiology (STROBE-Nut); mendelian randomization (STROBE-MR)
Systematic reviews	PRISMA	Abstracts (PRISMA-Abstracts); individual participant data (PRISMA-IPD; diagnostic test accuracy (PRISMA-DTA); harms (PRISMA-harms); network meta-analysis (PRISMA-NMA); literature searches (PRISMA-S)
Diagnostic test accuracy	STARD	Abstracts (STARD-Abstracts); artificial intelligence (STARD-AI) *
Prediction model studies	TRIPOD	Abstracts (TRIPOD-Abstracts); Individual participant data meta-analysis/clustered data (TRIPOD-Cluster)*; systematic reviews (TRIPOD-SRMA)*; machine learning (TRIPOD-AI) *

* forthcoming

Epiphany

We have provided a list of twelve issues routinely encountered during statistical peer review of articles submitted to the *BMJ*. Last Christmas we tweeted this list, but the very next day we got poor submissions anyway. This year, to save us from tears, we've tailored it for someone special – you, the *BMJ* reader.

Our hope is that you will check these twelve issues *before* rushing to submit articles to the *BMJ* next Christmas; this would bring joy to the world by reducing the length of our reviews and allowing us to spend more time with our *significant* (yes, pun-intended) others over the festive period. Indeed, if you adhere to our guidance, our song will change to the very positive "Twelve Days of Christmas Review" shown in Box 1.

Ultimately, we want the *BMJ* to publish the gold not the mould; the frankincense not the makes-no-sense; and the myrrh not the urrgghh. Many other topics could have been mentioned, and for further guidance we point readers to the *BMJ Statistics Notes* series (written mainly by Doug Altman and Martin Bland), the *Research Methods and Reporting* section of the *BMJ*,[83] and other overviews of common statistical mistakes.[84 85]

Box 1 The Twelve Days of Christmas Review, to be sung by the BMJ Statisticians when researchers improve their article submissions

On the twelfth day of Christmas, the BMJ statistician sent to me:

12. Complete reporting
11. Assumptions assessed
10. Variables verified
9. Measured meta-analysis
8. Predictions calibrated
7. Clusters captured
6. Subgroups attested
5. Curves fitted
4. Wise not dichotomised
3. Missing mattered
2. Honest interpreting

and

A coherent research question

References

1. Sauerbrei W, Bland M, Evans SJW, et al. Doug Altman: Driving critical appraisal and improvements in the quality of methodological and medical research. *Biom J* 2021;63(2):226-46.
2. Osmond C. Professor Martin Gardner (1940-93). *Journal of the Royal Statistical Society Series A (Statistics in Society)* 1993;156(3):498-99.
3. Van Calster B, Wynants L, Riley RD, et al. Methodology over metrics: current scientific standards are a disservice to patients and society. *J Clin Epidemiol* 2021;138:219-26.
4. Altman DG. The scandal of poor medical research. *BMJ* 1994;308(6924):283-4.
5. Ioannidis JP. Errors (my very own) and the fearful uncertainty of numbers. *Eur J Clin Invest* 2014;44(7):617-8.
6. Dr. Seuss. How the Grinch Stole Christmas!: Random House 1957.
7. Addendum on Estimands and Sensitivity Analysis in Clinical Trials to the Guideline on Statistical Principles for Clinical Trials [E9(R1) Final version (Step 4), Adopted on 20 November 2019]. *https://databaseichorg/sites/default/files/E9-R1_Step4_Guideline_2019_1203pdf*
8. Kahan BC, Morris TP, White IR, et al. Estimands in published protocols of randomised trials: urgent improvement needed. *Trials* 2021;22(1):686.
9. Altman DG, Bland JM. Statistics notes: Absence of evidence is not evidence of absence. *BMJ* 1995;311(7003):485.

10. Bayes T. An essay toward solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society* 1764;53:418.

11. Little JA, Rubin DB. Statistical Analysis with Missing Data. New York: John Wiley and Sons 2002.

12. Sterne JA, White IR, Carlin JB, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ* 2009;338:b2393.

13. White IR, Thompson SG. Adjusting for partially missing baseline measurements in randomized trials. *Stat Med* 2005;24(7):993-1007.

14. Sullivan TR, White IR, Salter AB, et al. Should multiple imputation be the method of choice for handling missing data in randomized trials? *Stat Methods Med Res* 2018;27(9):2610-26.

15. Groenwold RH, White IR, Donders AR, et al. Missing covariate data in clinical research: when and when not to use the missing-indicator method for analysis. *CMAJ : Canadian Medical Association journal = journal de l'Association medicale canadienne* 2012;184(11):1265-9.

16. Hughes RA, Heron J, Sterne JAC, et al. Accounting for missing data in statistical analyses: multiple imputation is not always the answer. *International Journal of Epidemiology* 2019;48(4):1294-304.

17. Lee KJ, Tilling KM, Cornish RP, et al. Framework for the treatment and reporting of missing data in observational studies: The Treatment And Reporting of Missing data in Observational Studies framework. *Journal of Clinical Epidemiology* 2021;134:79-88.

18. van Buuren S. Flexible Imputation of Missing Data (Second Edition). Boca Raton, FL2018.

19. Altman DG, Royston P. Statistics notes: The cost of dichotomising continuous variables. *BMJ* 2006;332::1080.

20. Royston P, Altman DG, Sauerbrei W. Dichotomizing continuous predictors in multiple regression: a bad idea. *Stat Med* 2006;25(1):127-41.

21. Altman DG, Lausen B, Sauerbrei W, et al. Dangers of using "optimal" cutpoints in the evaluation of prognostic factors. *J Natl Cancer Inst* 1994;86(11):829-35.

22. Collins GS, Ogundimu EO, Cook JA, et al. Quantifying the impact of different approaches for handling continuous predictors on the performance of a prognostic model. *Stat Med* 2016;35(23):4124-35.

23. MacCallum, R.C., Zhang S, Preacher KJ, et al. On the practice of dichotomization of quantitative variables. *Psychol Meth* 2002;7:19-40.

24. Senn S, Julious S. Measurement in clinical trials: a neglected issue for statisticians? *Stat Med* 2009;28(26):3189-209.

25. Senn S. Individual response to treatment: is it a valid assumption? *BMJ* 2004;329(7472):966-8.

26. Chilvers C, Dewey M, Fielding K, et al. Antidepressant drugs and generic counselling for treatment of major depression in primary care: randomised trial with patient preference arms. *Bmj* 2001;322(7289):772-5.

27. Durrleman S, Simon R. Flexible regression models with cubic splines. *Stat Med* 1989;8:551-61.

28. Harrell FE, Jr. Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis (Second Edition). New York: Springer 2015.

29. Nieboer D, Vergouwe Y, Roobol MJ, et al. Nonlinear modeling was applied thoughtfully for risk prediction: the Prostate Biopsy Collaborative Group. *J Clin Epidemiol* 2015;68(4):426-34.

30. Sauerbrei W, Royston P. Building multivariable prognostic and diagnostic models: transformation of the predictors by using fractional polynomials. *Journal of the Royal Statistical Society Series A* 1999;162:71-94.

31. Royston P, Sauerbrei W. Multivariable Model-Building - A pragmatic approach to regression analysis based on fractional polynomials for modelling continuous variables. Chichester: Wiley 2008.

32. Royston P, Altman DG. Regression Using Fractional Polynomials of Continuous Covariates: Parsimonious Parametric Modelling. *Journal of the Royal Statistical Society Series C (Applied Statistics)* 1994;43(3):429-67.

33. Sauerbrei W, Royston P, Bojar H, et al. Modelling the effects of standard prognostic factors in node-positive breast cancer. German Breast Cancer Study Group (GBSG). *Br J Cancer* 1999;79(11-12):1752-60.

34. Johannesen CDL, Langsted A, Mortensen MB, et al. Association between low density lipoprotein and all cause and cause specific mortality in Denmark: prospective cohort study. *BMJ* 2020;371:m4266.

35. Altman DG, Bland JM. Interaction revisited: the difference between two estimates. *BMJ* 2003;326(7382):219.

36. Austin PC, Hux JE. A brief note on overlapping confidence intervals. *J Vasc Surg* 2002;36(1):194-5.

37. VanderWeele Tyler J, Knol Mirjam J. A Tutorial on Interaction. *Epidemiologic methods* 2014;3(1):33.

38. Shrier I, Pang M. Confounding, effect modification, and the odds ratio: common misinterpretations. *Journal of Clinical Epidemiology* 2015;68(4):470-74.

39. Williamson SF, Grayling MJ, Mander AP, et al. Subgroup analyses in randomised controlled trials frequently categorised continuous subgroup information *Journal of Clinical Epidemiology* 2022

40. Royston P, Sauerbrei W. A new approach to modelling interactions between treatment and continuous covariates in clinical trials by using fractional polynomials. *Stat Med* 2004;23(16):2509-25.

41. Peters TJ, Richards SH, Bankhead CR, et al. Comparison of methods for analysing cluster randomized trials: an example involving a factorial design. *Int J Epidemiol* 2003;32(5):840-6.

42. Bland JM. Cluster randomised trials in the medical literature: two bibliometric surveys. *BMC Med Res Methodol* 2004;4:21.

43. Lee KJ, Thompson SG. The use of random effects models to allow for clustering in individually randomized trials. *Clin Trials* 2005;2(2):163-73.

44. Steyerberg EW, Bossuyt PM, Lee KL. Clinical trials in acute myocardial infarction: should we adjust for baseline characteristics? *Am Heart J* 2000;139(5) 745-51.

45. Hernández AV, Steyerberg EW, Habbema JD. Covariate adjustment in randomized controlled trials with dichotomous outcomes increases statistical power and reduces sample size requirements. *J Clin Epidemiol* 2004 57(5):454-60.

46. Turner EL, Perel P, Clayton T, et al. Covariate adjustment increased power in randomized controlled trials: an example in traumatic brain injury. *J Clin Epidemiol* 2012 65(5)::474-81. .

47. Riley RD, Tierney JF, Stewart LA, editors. *Individual Participant Data Meta-Analysis: A Handbook for Healthcare Research*. Chichester, West Sussex: Wiley, 2021.

48. Abo-Zaid G, Guo B, Deeks JJ, et al. Individual participant data meta-analyses should not ignore clustering. *J Clin Epidemiol* 2013;66(8):865-73 e4.

49. Greenland S, Robins MR, Pearl J. Confounding and collapsibility in causal inference. *Statistical Science* 1999;14:29-46.

50. Robinson LD, Jewell NP. Some surprising results about covariate adjustment in logistic regression models. *Int Stat Rev* 1991;58:227-40.

51. Gail MH, Wieand S, Piantadosi S. Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates. *Biometrika* 1984;71:431-44.

52. Gardiner JC, Luo Z, Roman LA. Fixed effects, random effects and GEE: what are the differences? *Stat Med* 2009;28(2):221-39.

53. Kahan BC, Morris TP. Assessing potential sources of clustering in individually randomised trials. *BMC Med Res Methodol* 2013;13:58.

54. Kahan BC, Li F, Copas AJ, et al. Estimands in cluster-randomized trials: choosing analyses that answer the right question. *Int J Epidemiol* 2022

55. Higgins JP, Thompson SG, Deeks JJ, et al. Measuring inconsistency in meta-analyses. *BMJ* 2003;327(7414):557-60.

56. Rucker G, Schwarzer G, Carpenter JR, et al. Undue reliance on I(2) in assessing heterogeneity may mislead. *BMC Med Res Methodol* 2008;8:79.

57. Thompson SG, Higgins JPT. How should meta-regression analyses be undertaken and interpreted? *Stat Med* 2002;21:1559-74.

58. Riley RD, Debray TPA, Fisher D, et al. Individual participant data meta-analysis to examine interactions between treatment effect and participant-level covariates: Statistical recommendations for conduct and planning. *Stat Med* 2020;39(15):2115-37.

59. Fisher DJ, Carpenter JR, Morris TP, et al. Meta-analytical methods to identify who benefits most from treatments: daft, deluded, or deft approach? *BMJ* 2017;356:j573.

60. Riley RD, Lambert PC, Abo-Zaid G. Meta-analysis of individual participant data: rationale, conduct, and reporting. *BMJ* 2010;340:c221.

61. Steyerberg EW, Moons KG, van der Windt DA, et al. Prognosis Research Strategy (PROGRESS) 3: prognostic model research. *PLoS Med* 2013;10(2):e1001381.

62. Altman DG, Vergouwe Y, Royston P, et al. Prognosis and prognostic research: validating a prognostic model. *BMJ* 2009;338:b605.

63. Vickers AJ, Van Calster B, Steyerberg EW. Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. *BMJ* 2016;352:i6.

64. Jaja BNR, Saposnik G, Lingsma HF, et al. Development and validation of outcome prediction models for aneurysmal subarachnoid haemorrhage: the SAHIT multinational cohort study. *BMJ* 2018;360:j5745.

65. Vickers AJ, van Calster B, Steyerberg EW. A simple, step-by-step guide to interpreting decision curve analysis. *Diagn Progn Res* 2019;3:18.

66. Copas JB. Using regression models for prediction: shrinkage and regression to the mean. *Stat Methods Med Res* 1997;6(2):167-83.

67. Van Houwelingen JC. Shrinkage and penalized likelihood as methods to improve predictive accuracy. *Statistica Neerlandica* 2001;55:17-34.

68. Firth D. Bias reduction of maximum likelihood estimates. *Biometrika* 1993;80(1):27-38.

69. Greenland S, Mansournia MA, Altman DG. Sparse data bias: a problem hiding in plain sight. *BMJ* 2016;352:i1981.

70. Heinze G, Wallisch C, Dunkler D. Variable selection - A review and recommendations for the practicing statistician. *Biom J* 2018;60(3):431-49.

71. Riley RD, van der Windt D, Croft P, et al., editors. *Prognosis Research in Healthcare: Concepts, Methods and Impact*. Oxford, UK: Oxford University Press, 2019.

72. Richiardi L, Bellocco R, Zugna D. Mediation analysis in epidemiology: methods, interpretation and bias. *Int J Epidemiol* 2013;42(5):1511-9.

73. Sun GW, Shook TL, Kay GL. Inappropriate use of bivariable analysis to screen risk factors for use in multivariable analysis. *J Clin Epidemiol* 1996;49(8):907-16.

74. Ensor J, Riley RD, Jowett S, et al. Prediction of risk of recurrence of venous thromboembolism following treatment for a first unprovoked venous thromboembolism: systematic review, prognostic model and clinical decision rule, and economic evaluation. *Health Technol Assess* 2016;20(12):i-xxxiii, 1-190.

75. Royston P, Parmar MK. An approach to trial design and analysis in the era of non-proportional hazards of the treatment effect. *Trials* 2014;15:314.

76. Altman DG. Better reporting of randomised controlled trials: the CONSORT statement. *BMJ* 1996;313(7057):570-71.

77. Simera I, Moher D, Hirst A, et al. Transparent and accurate reporting increases reliability, utility, and impact of your research: reporting guidelines and the EQUATOR Network. *BMC medicine* 2010;8:24.

78. Schulz KF, Altman DG, Moher D. CONSORT 2010 Statement: updated guidelines for reporting parallel group randomised trials. *BMJ* 2010;340:c332

79. Collins GS, Reitsma JB, Altman DG, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): The TRIPOD Statement. *Ann Intern Med* 2015;162:55-63.

80. Moons KG, Altman DG, Reitsma JB, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): explanation and elaboration. *Ann Intern Med* 2015;162(1):W1-73.

81. Boutron I, Altman DG, Hopewell S, et al. Impact of spin in the abstracts of articles reporting results of randomized controlled trials in the field of cancer: the SPIIN randomized controlled trial. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* 2014;32(36):4120-6.

82. Altman DG, Bland JM. Quartiles, quintiles, centiles, and other quantiles. *BMJ* 1994;309(6960):996.

83. Groves T. Research methods and reporting. *BMJ* 2008;337:a2201.

84. Assel M, Sjoberg D, Elders A, et al. Guidelines for Reporting of Statistics for Clinical Research in Urology. *J Urol* 2019;201(3):595-604.

85. Stratton IM, Neil A. How to ensure your paper is rejected by the statistical reviewer. *Diabet Med* 2005;22(4):371-3.