CHRISTMAS 2022: THE SCIENTIST

On the 12th Day of Christmas, a Statistician Sent to Me . . .

The BMJ's statistical editors relish a quiet Christmas, so make their wish come true and pay attention to the list of common statistical faux pas presented here by Riley and colleagues

Richard D Riley, [1] Tim J Cole, [2] Jon Deeks, [1] Jamie J Kirkham, [3] Julie Morris, [4] Rafael Perera, [5] Angie Wade, [6] Gary S Collins[7]

[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, University of Manchester, Manchester Academic Health Science Centre, Manchester, UK

[4] University of Manchester, Manchester, UK

[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, UK

Correspondence to: R D Riley
r.d.riley@bham.ac.uk

Cite this as: *BMJ* 2022;379:e072883
http://dx.doi.org/10.1136/bmj-2022-072883
Published: 20 December 2022

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 finding time to finish off statistical analyses, draft manuscripts, and respond to reviewers' comments. This activity leads to a plethora of submissions to journals such as *The BMJ* in December, so that researchers can finish the year with a sense of academic achievement and enjoy the festivities with their loved ones. Indeed, with optimism fuelled by mulled wine and mince pies, researchers may even anticipate their article's acceptance by early January, at the end of the 12 days of Christmas.

A collective, however, works against this season of publication goodwill and cheer—a small but influential group of statisticians with very shiny noses for detail, seeking "all is right" rather than "all is bright" and emphasising no, no, no rather than ho, ho, ho. The statisticians' core belief is that a research article is for life, not just for Christmas, and they deliver statistical reviews that promote high standards of methodological rigour and transparency. So you can imagine how busy they are during the Christmas period with its influx of submissions—even before they can eat, drink, and be merry, these individuals are working tirelessly 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!

Each year *The BMJ*'s statistical editors review more than 500 articles. For about 30 years, the statistical team was led by Martin Gardner and Doug Altman,[1 2] both of whom saw similarities between statisticians and the Christmas star, with the statisticians lighting a path of research integrity, promoting methodology over metrics,[3 4] and encouraging statistical principles to "save science and the world."[5]

To elicit the most common issues encountered during statistical peer review, an internal survey was administered to *The BMJ*'s statistical editors. Twelve items were identified, and each are described here. There is one item for each of the 12 days of Christmas, the period between 25 December and 5 January when the statisticians conduct their reviews in the mindset of the Grinch,[6] but with the kind heart of *Miracle On 34th Street*.

Advent

Every December *The BMJ*'s statistical editors meet for a day, when they discuss common statistical concerns, problematic submissions (including those that slipped through the net, the so-called sin bin articles), and how to improve the review process, before unwinding at *The BMJ's* Christmas party. At the meeting on 18 December 2019, the statisticians agreed that an article showcasing common statistical issues would be helpful for authors of future article submissions, and an initial set of items was discussed. When reminded about this article at subsequent Christmas meetings on 17 December 2020 and 16 December 2021, the statisticians explained that progress was being delayed, ironically because of the number of statistical reviews that needed to be prioritised in *The BMJ*'s system.

After further procrastination, on 28 June 2022 a potential list of items was shared among the statistical editors by email, and everyone was asked to include any further issues they regularly encountered during statistical review. The findings were collated and discussed (by email) and a final list of the most important items agreed for wider dissemination. Twelve items were selected, to match the number of days of Christmas in the well known song (and thereby increase the chance of publication in *The BMJ*'s Christmas issue). Sensitivity analyses, including shallow and deep learning approaches, led to the same 12 items being selected. An automated artificial intelligence algorithm quickly identified that all the statistical editors were guilty of similar statistical faux pas in some of their own research articles, and so are not whiter than snow.

The 12 days of statistical review

To help drive them home for Christmas, the 12 identified items are briefly explained. Consider them as stocking fillers for you, *The BMJ* reader and potential future author. Allowing for sizeable Christmas meals, digest one item each day between 25 December and 5 January and make a New Year's resolution to follow the guidance.

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

Clarify the research question

Christmas is a time for reflection on the meaning of life and future expectations. Similarly, in their reviews, statisticians will often encourage authors to reflect on their research question and clarify their

objectives. As an example, in an observational study, the authors may need to clarify the extent to which their research is descriptive or causal, prognostic factor identification or prediction model development, or exploratory or confirmatory. For causal research, authors may be asked to express the underlying premise (causal pathway or model), for example, in terms of a directed acyclic graph. In systematic reviews of intervention studies, authors might need to state their research question using the Population, Intervention, Comparison, and Outcome system—the PICO structure.

A related request would be to clarify the estimand—the study's target measure for estimation.[7] In a randomised trial, for example, the estimand is a treatment effect, but a statistician might request better definitions for the population, treatments being compared, outcomes, summary measure (eg, risk ratio or risk difference, conditional or marginal effect), and other features.[7 8] Similarly, in a meta-analysis of randomised trials the estimand must be defined in the context of potential heterogeneity of study characteristics. In a meta-analysis of hypertension trials with different lengths of follow-up, for example, if the estimand is a treatment effect on blood pressure, clarity is needed about whether this relates to one time point (eg, one year), each of multiple time points (eg, one year and five years), or some average across a range of time points (eg, six months to two years).

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

Focus on estimates, confidence intervals, and clinical relevance

Just as with under-cooked turkeys being sent back so will articles that focus solely on P values and "statistical significance" to determine whether a finding is crucial. It is important to consider the estimates (eg, mean differences, risk ratios, or hazard ratios corresponding to the specified estimands from the first day of Christmas), corresponding 95% confidence intervals, and potential clinical relevance of findings. Statistical significance often does not equate to clinical significance—if, as an example, a large trial estimates a risk ratio of 0.97 and a 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]—here's an example; if a small trial estimates a risk ratio of 0.70 and a 95% confidence interval of 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, the statistical editors will ask authors 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] to express probabilistic statements (eg, there is a probability of 0.85 that the risk ratio is <0.9).

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

Carefully account for missing data

Missing values occur in all types of medical research,[11] both for covariates and for outcomes. Authors need to not only acknowledge the completeness of their data but also to quantify and report the amount of missing data and explain how such data were handled in analyses. It is spooky how many submissions fail to do this—the ghost of Christmas articles past, present, and future.

If it transpires participants with missing data were simply excluded (ie, a complete case analysis was carried out), then authors may be asked to revise their analyses by including those participants, 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 detailed interrogation here. For example, strategies for handling missing baseline values in randomised trials might include replacing with the mean value (for continuous variables), creating a separate category of a categorical predictor to indicate the presence of a missing value (ie, 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. Under a missing at random assumption, this involves missing values being 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, the methods used to do this need to be described, including the set of variables used in the imputation process. An introduction to multiple imputation is provided elsewhere,[12] and there are textbooks dedicated to missing data.[18]

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

Do not dichotomise continuous variables

Santa likes dichotomisation (you are either naughty or nice), but statisticians would be appalled if authors chose to dichotomise continuous variables, such as age and blood pressure, by splitting them into two groups defined by being above and below some arbitrary cut point, such as a systolic blood pressure of 130 mm Hg. Dichotomisation should be avoided,[19 20] as it wastes information and is rarely justifiable compared with analysing continuous variables on their continuous scale (see the stocking filler for the fifth day of Christmas). Why should an individual with a value just below the cut point (in this instance 129 mm Hg) be considered completely different from an individual with a value just above it (131 mm Hg)? Conversely, the values for two individuals within the same group may differ greatly (let us say 131 mm Hg and 220 mm Hg) 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 covariate values violates this agreement.

Dichotomisation also reduces statistical power to detect associations between a continuous covariate and the outcome,[19-21] and it 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] whereas in another example, 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 and 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 result in misleading conclusions.[24 25] A good example is a randomised trial in which the required sample size was reduced from 800 to 88 after the outcome (Beck score) changed from being analysed as dichotomised to being analysed on its continuous scale.[26]

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

Consider non-linear relationships

At Christmas dinner, some family relationships are simple to handle, whereas 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), whereas others have a more complex non-linear relationship. A linear relationship (association) assumes that a 1 unit increase in the covariate has the same effect on the outcome across the entire range of the covariate's values. The assumption being, for example, that the impact of a change in age from 30 to 31 years is the same as a change in age from 90 to 91 years. 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 years may be important. The two most common approaches to non-linear modelling are cubic splines and fractional polynomials.[27-32]

Aside from categorisation, most submissions to *The BMJ* only consider linear relationships. The statistical reviewers therefore may ask the researchers to consider non-linear relationships, to avoid important associations not being fully captured or even missed.[33] The study by Johannesen and colleagues is an example of non-linear relationships being examined.[34] The authors used 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 of Denmark. Figure 1 illustrates the findings for the overall population, and for subgroups defined by use of lipid lowering treatment, with the relationship strongest in those not receiving treatment.

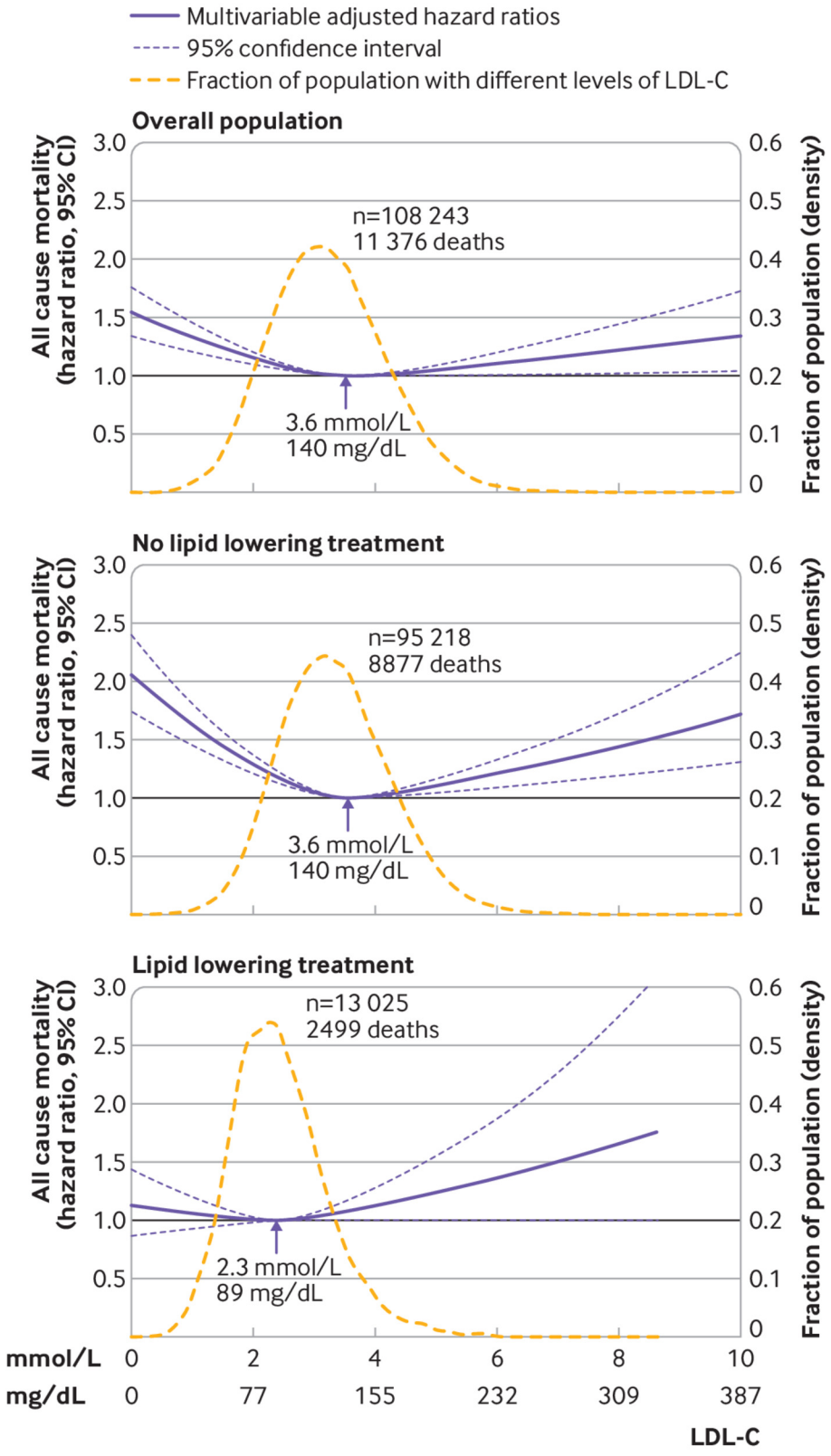

Fig 1 | Non-linear association derived using restricted cubic splines of individuals from the Copenhagen General Population Study followed for a mean 9.4 years, from Johannesen et al.[34] Multivariable adjusted hazard ratios for all cause mortality are shown according to levels of low density lipoprotein cholesterol (LDL-C) on a continuous scale. 95% confidence intervals are derived from restricted cubic spline regressions with three knots. Reference lines for no association are shown at a hazard ratio of 1.0. Arrows indicate concentration of LDL-C associated with the lowest risk of all cause mortality. Analyses were adjusted for baseline age, sex, current smoking, cumulative number of cigarette pack years, systolic blood pressure, lipid lowering treatment, diabetes, cardiovascular disease, cancer, and chronic obstructive pulmonary disease

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

Quantify differences in subgroup results

Many submitted articles include results for subgroups, such as defined by sex or gender, or those who do and do not eat Brussels sprouts. A common mistake is to conclude that the results for one subgroup are different from the results of another subgroup, without actually quantifying the difference. Altman and Bland considered this eloquently,[35] showing treatment effect results for two subgroups, the first of which was statistically significant (risk ratio 0.67, 95% confidence interval 0.46 to 0.98; P=0.03), whereas the second was not (0.88, 0.71 to 1.08; P=0.2). A naïve interpretation is to conclude that the treatment is beneficial for the first subgroup but not for the second subgroup. However, actually comparing the results between the two subgroups reveals a wide confidence interval (ratio of risk ratios 0.76, 95% confidence interval 0.49 to 1.17; P=0.2), which suggests further research is needed before concluding a subgroup effect. A related mistake is to make conclusions about whether subgroups differ based solely on if their separate 95% confidence intervals overlap or not.[36] Hence, if researchers examine subgroups in their study, the statistical editors will check for quantification of differences in subgroup results, and, if not done, ask for this to be addressed. Even when genuine differences exist between subgroups, the (treatment) effect may still be important for each subgroup, and therefore this should be recognised in study conclusions.

Examining differences between subgroups is complex, and a broader topic is the modelling of interactions between (treatment) effects and covariates.[37] Problems include the scale used to measure the effect (eg, 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 for the fourth day and fifth day of Christmas).[40]

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

Consider accounting for clustering

At *The BMJ*'s Christmas party, the statistical editors tend to cluster in a corner, avoiding interaction and eye contact with non-statisticians whenever possible for fear of being asked to conduct a postmortem examination of rejected work. Similarly, a 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 from multiple studies.[47] Sometimes the analysis does not account for this clustering, which can 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 (eg, 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, in the data analysis, a submitted article ignores obvious clustering that needs to be captured or considered, the statistical editors will ask for justification of this or for a reanalysis accounting for clustering using an approach suitable for the estimand of interest (see our stocking filler for the first day of Christmas).[52-54] A multilevel or mixed effects model might be recommended, for example, 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, a statistician sent to me:

Interpret I^2 and meta-regression appropriately

Systematic reviews and meta-analyses are popular submissions to *The BMJ*. Most of them include the I^2 statistic[55] but interpret it incorrectly, which gives the statisticians 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 it is 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 (ie, 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 τ^2). For example, if all the included 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. Statistical reviews will ask authors 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 the extent to which study level covariates (eg, mean age, dose of treatment, risk of bias rating) explain between study heterogeneity, but generally the statistical editors will ask authors to interpret meta-regression results cautiously.[57] Firstly, the number of trials are often small, and then meta-regression is affected by 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 might 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 (eg, mean age, proportion men) with the overall treatment effect should not be used to make inferences about how values of participant level covariates (eg, 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] as shown in figure 2.

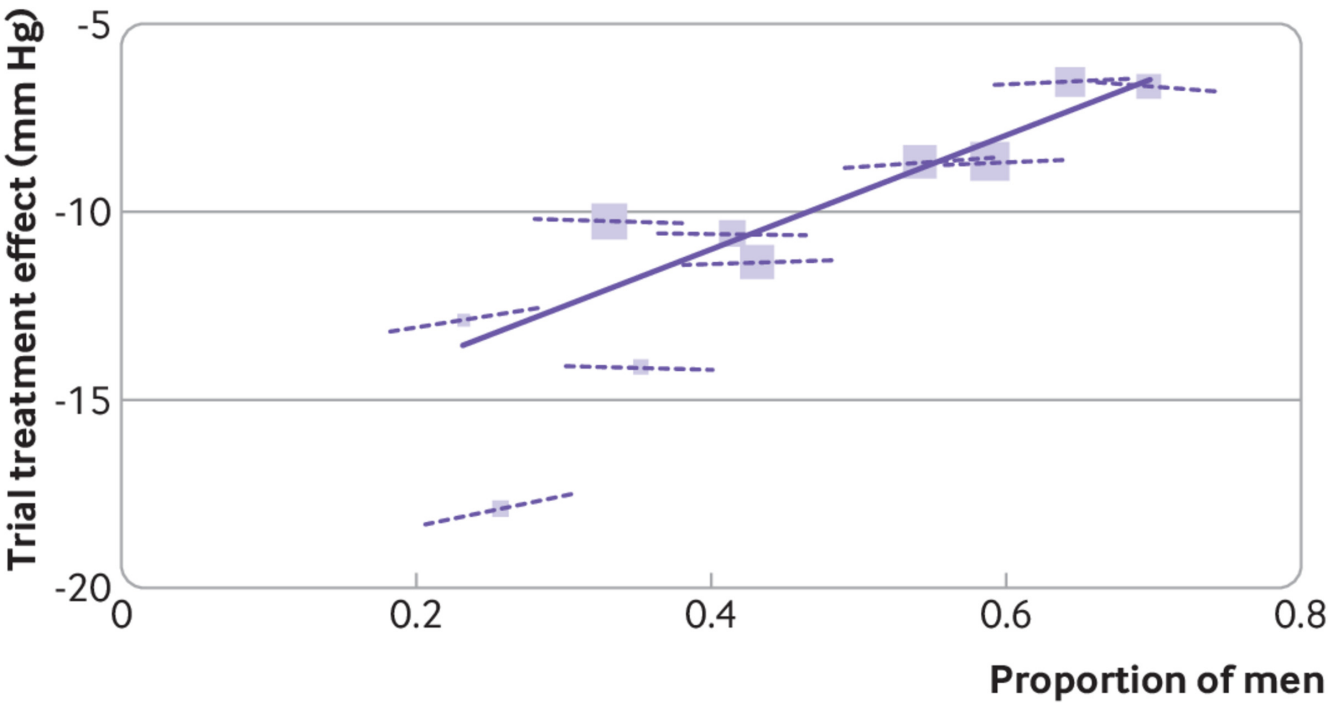

Fig 2 | Aggregation bias when using meta-regression of study level results rather than individual participant data meta-analysis of treatment-covariate interactions. The research question was whether blood pressure lowering treatment is more effective among women than men. Evidence is shown from a meta-analysis of 10 trials of antihypertensive treatment, comparing the across trial association of treatment effect and proportion men (solid line)—which is steep and statistically significant—with participant level interactions of sex and treatment effect in each trial (dashed lines) —which are flat and neither clinically nor statistically important. This case study is based on previous work.[47 58 60] Each block represents one trial, with block size proportional to trial size. Across trial association is denoted by gradient of solid line, derived from a meta-regression of the trial treatment effects against proportion of men, which suggests a large effect of a 15 mm Hg (95% confidence interval 8.8 to 21 mm Hg) greater reduction in systolic blood pressure in trials with only women compared with only men. 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 mm Hg (−0.5 to 2.1 mm Hg) greater treatment effect for women than for men, which is neither clinically nor statistically significant

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

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 can have important consequences because inaccurate predictions can lead to incorrect decisions and harmful communication to patients, such as 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 elsewhere.[61-63]

However, the majority of submissions focus only on model discrimination (as quantified by, for example, the C statistic or area under the curve[28])—when this is done, an incomplete impression is created, just as with that unfinished 1000 piece jigsaw from last Christmas. Figure 3 shows a published calibration plot 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 and 0.2.[64] This miscalibration may impact the clinical utility of the model, especially if decisions, such as about treatment or monitoring strategies, are dictated by risk thresholds in that range of predicted risks, 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 when it occurs in relation to decision thresholds.

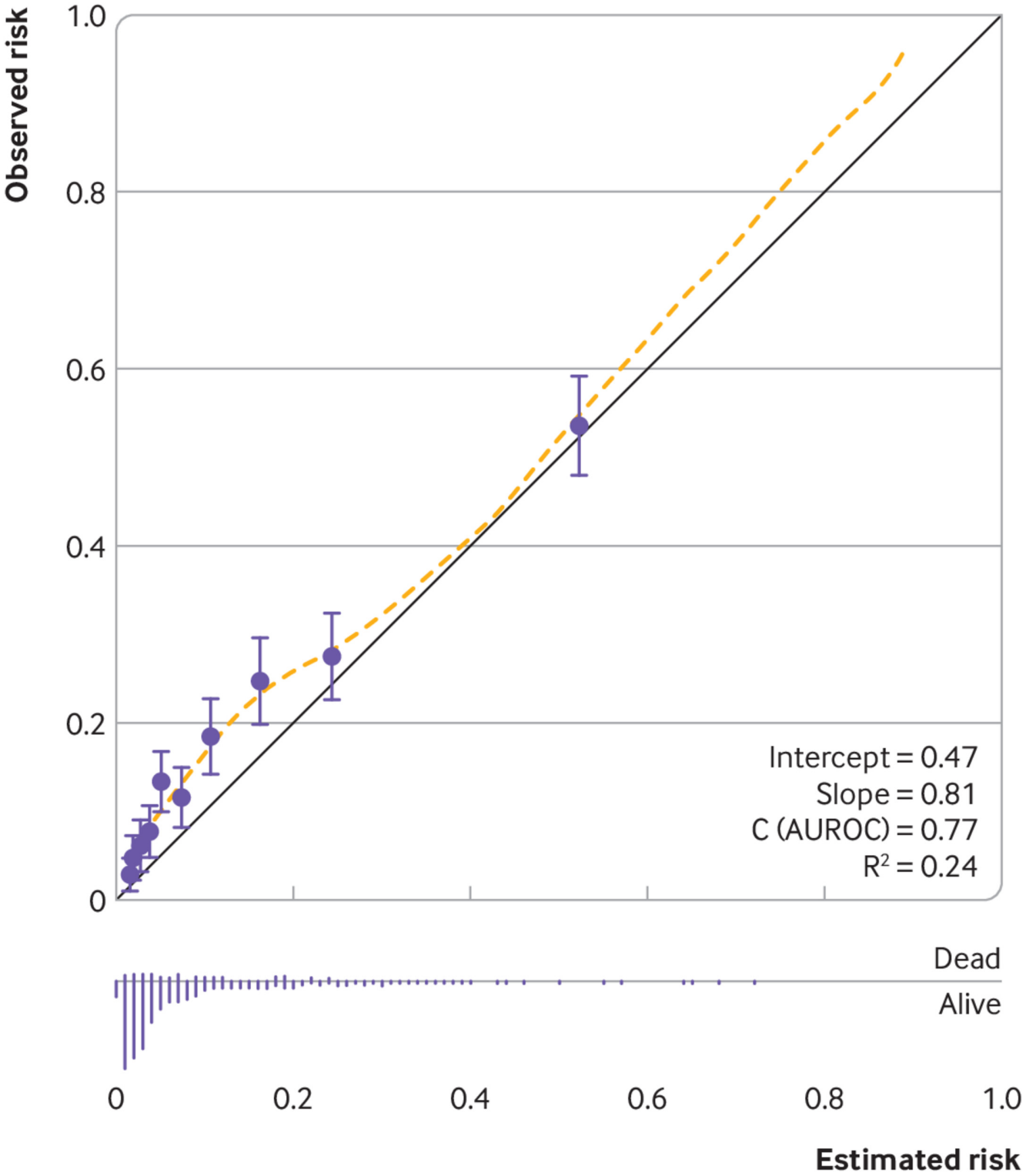

Fig 3 | Example of a calibration plot to examine agreement between observed risks and estimated (predicted) risks from a prediction model.[64] The study developed prediction models to estimate the risk of mortality in individuals who experienced subarachnoid haemorrhage from ruptured intracranial aneurysm. Circles are estimated and observed risks grouped by 10ths of estimated risks, and the yellow dashed line is a loess smoother to capture agreement across the range of estimated risks. AUROC=area under the receiving operator characteristic

Statistical editors may also suggest that researchers of model development studies undertake a reanalysis using penalisation or shrinkage methods (eg, 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 non-prediction situations

(eg, randomised trials estimating treatment effects) with sparse data, as standard methods (such as logistic regression) may give biased effect estimates in this situation.[69]

On the 10th day of Christmas, a statistician sent to me:

Carefully consider the variable selection approach

A common area of criticism in statistical reviews is the use of variable selection methods (eg, selection of covariates based on the statistical significance of their effects).[70] If these methods are used, statistical editors will ask authors for justification. Depending on the study, statistical editors might even suggest authors avoid these approaches entirely, just as you would 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 directed acyclic graphs (with consideration of potential mediators between covariates and outcome[72]), not statistical significance based on automated selection methods.

In the development of clinical prediction models, variable selection (through 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, when 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, because in practice the relevant predictors are used (by healthcare professionals and patients) in combination. When, for example, a prognostic model was being developed for risk of recurrent venous thromboembolism, the researchers 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 11th day of Christmas, a statistician sent to me:

Assess the impact of any assumptions

Everyone agrees that *It's A Wonderful Life* is a Christmas movie, but whether this applies to *Die Hard* is debatable. Similarly, statistical editors might debate authors' die-hard analysis assumptions, and even ask them to examine whether results change if the assumptions change (a sensitivity analysis). For example, in submitted trials with time-to-event data, such as 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 this assumption is not justified in an article, authors may be asked 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 may still be influential. In this situation, authors may be asked to demonstrate how results change when other plausible prior distributions are chosen.

On the 12th day of Christmas, a statistician sent to me:

Use reporting guidelines and avoid overinterpretation

Altman once 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 reported research is indefensible and creates confusion, just as with those unlabelled presents under the Christmas tree. Readers need to know the rationale and objectives of a reported study, the study design, methods used, participant characteristics, results, certainty of evidence, research implications, and so forth. If any of these elements are missing, authors will be asked to clarify them.

Make use of reporting guidelines. They provide a checklist of items to be reported (Santa suggests checking this 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] Table 1 shows examples, including the CONSORT statement for randomised trials[79] and the TRIPOD guideline for prediction model studies.[80 81] *The BMJ* requires authors to complete the checklist within the relevant guideline (and include it with a submission), indicating on which page of the submitted manuscript each item has been reported.

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 (consolidated standards of reporting trials)	Cluster trials (CONSORT-Cluster), multi-arm trials, non-inferiority or equivalence trials (CONSORT non-inferiority), harms (CONSORT-HARMS), pilot and feasibility trials, adaptative designs (ACE statement), artificial intelligence (CONSORT-AI), interventions (TIDieR, template for intervention description and replication)
Observational studies	STROBE (strengthening the reporting of observational studies in epidemiology)	Genetic associations (STREGA, strengthening the reporting of genetic association studies), molecular epidemiology (STROBE-ME), infectious diseases (STROBE-ID), nutritional epidemiology (STROBE-Nut), mendelian randomisation (STROBE-MR)
Systematic reviews	PRISMA (preferred reporting items for systematic reviews and meta-analyses)	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 (standards for reporting diagnostic accuracy studies)	Abstracts (STARD-Abstracts), artificial intelligence (STARD-AI)
Prediction model studies	TRIPOD (transparent reporting of a multivariable prediction model for individual prognosis or diagnosis)	Abstracts (TRIPOD-Abstracts), individual participant data meta-analysis or clustered data (TRIPOD-Cluster),[78] systematic reviews (TRIPOD-SRMA), machine learning (TRIPOD-AI)

Another common part of the statistical editors review process, related to reporting, is to query overinterpretation of findings—and even spin,[82] such as 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 (eg, deciles are the nine cut points used to create 10 equal sized groups called 10ths).[83]

Epiphany

This list of 12 statistical issues routinely encountered during peer review of articles submitted to *The BMJ* will hopefully help authors of future submissions. Last Christmas statistical editors tweeted this list, but the very next day they got poor submissions anyway. This year, to save them from tears, they've tailored it for someone special—you, *The BMJ* reader.

Authors should address this list before rushing to submit papers to *The BMJ* next Christmas, in order to bring joy to the world by reducing the length of statistical reviews and allowing the statistical editors to spend more time with their significant (yes, pun intended) others over the festive period. If authors did adhere to this guidance, the "On the 12th Day of Christmas" song would change to the very positive "On the 12th Day of Christmas Review" with lyrics reflecting feedback from a happy statistician (perhaps join in using the song sheet in figure 4).

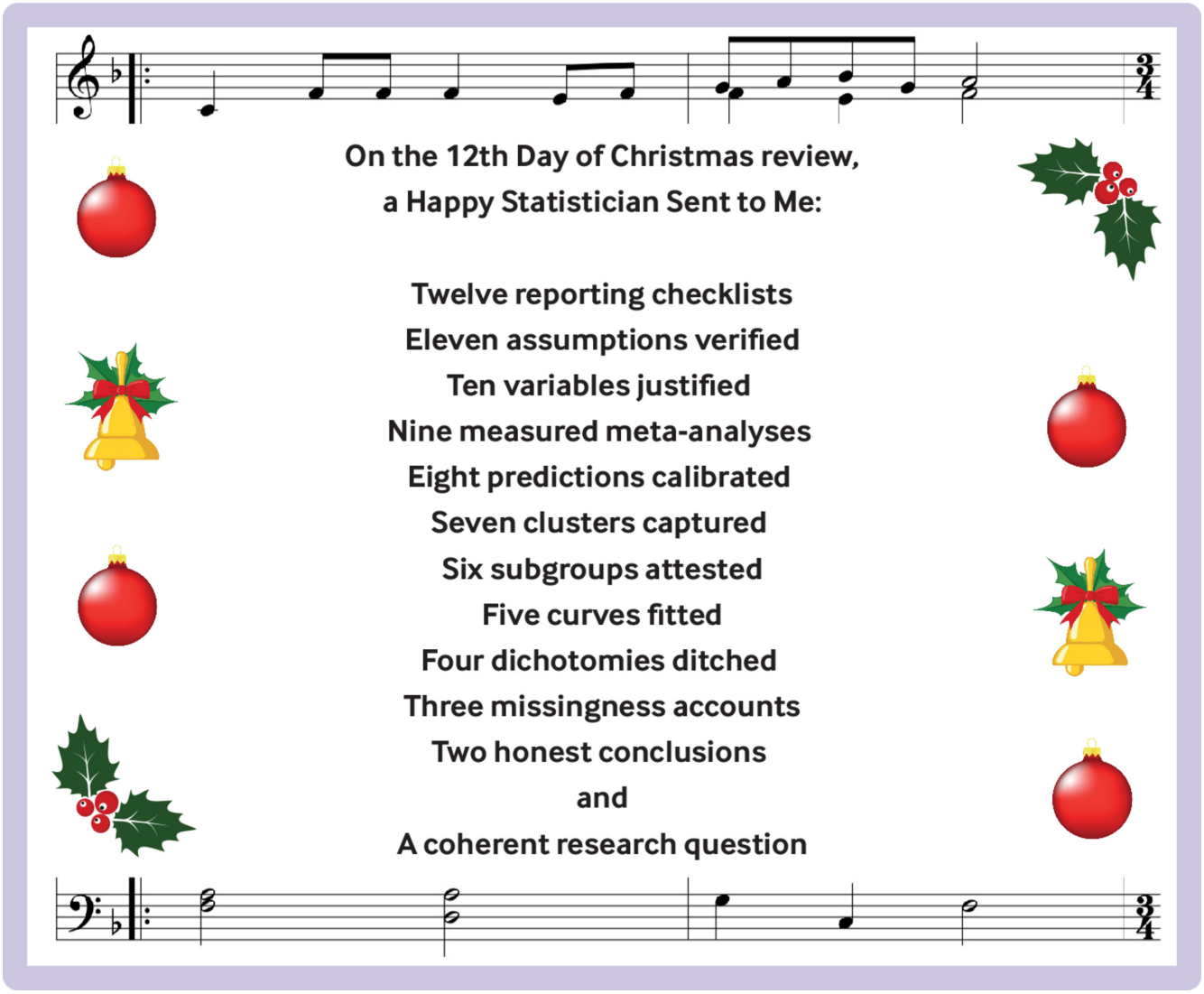

Fig 4 | Song sheet for "On the 12th Day of Christmas Review, a Happy Statistician sent to me . . ."

Ultimately, *The BMJ* wants 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 readers are directed to the BMJ Statistics Notes series (written mainly by Doug Altman and Martin Bland), the Research Methods and Reporting section of *The BMJ*,[84] and other overviews of common statistical mistakes.[85 86]

Contributors: RR conceived the paper and generated the Christmas theme and initial set of 12 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 on the basis of reviewer comments, followed by suggestions and final approval by all authors.

Competing interests: We have read and understood the BMJ Group policy on declaration of interests and declare: none.

Provenance and peer review: Not commissioned; externally peer reviewed.

This article is dedicated to Doug Altman and Martin Gardner, who led by example as chief statistical editors at *The BMJ* for over 30 years. We also thank the researchers who have responded politely to our statistical reviews over many years.

1 Sauerbrei W, Bland M, Evans SJW, etal. Doug Altman: Driving critical appraisal and improvements in the quality of methodological and medical research. *Biom J* 2021;63:-46. doi: 10.1002/bimj.202000053 pmid: 32639065

2 Osmond C. Professor Martin Gardner (1940-93). *J R Stat Soc Ser A Stat Soc* 1993;156:-9doi: 10.1111/j.1467-985X.1993.tb00518.x.

3 Van Calster B, Wynants L, Riley RD, van Smeden M, Collins GS. Methodology over metrics: current scientific standards are a disservice to patients and society. *J Clin Epidemiol* 2021;138:-26. doi: 10.1016/j.jclinepi.2021.05.018 pmid: 34077797

4 Altman DG. The scandal of poor medical research. *BMJ* 1994;308:-4. doi: 10.1136/bmj.308.6924.283 pmid: 8124111

5 Ioannidis JP. Errors (my very own) and the fearful uncertainty of numbers. *Eur J Clin Invest* 2014;44:-8. doi: 10.1111/eci.12277 pmid: 24785138

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://database.ich.org/sites/default/files/E9-R1_Step4_Guideline_2019_1203

8 Kahan BC, Morris TP, White IR, Carpenter J, Cro S. Estimands in published protocols of randomised trials: urgent improvement needed. *Trials* 2021;22:. doi: 10.1186/s13063-021-05644-4 pmid: 34627347

9 Altman DG, Bland JM. Absence of evidence is not evidence of absence. *BMJ* 1995;311:. doi: 10.1136/bmj.311.7003.485 pmid: 7647644

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

11 Little JA, Rubin DB. *Statistical Analysis with Missing Data*. John Wiley & Sons, 2002doi: 10.1002/9781119013563.

12 Sterne JA, White IR, Carlin JB, etal. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ* 2009;338:. doi: 10.1136/bmj.b2393 pmid: 19564179

13 White IR, Thompson SG. Adjusting for partially missing baseline measurements in randomized trials. *Stat Med* 2005;24:-1007. doi: 10.1002/sim.1981 pmid: 15570623

14 Sullivan TR, White IR, Salter AB, Ryan P, Lee KJ. Should multiple imputation be the method of choice for handling missing data in randomized trials?*Stat Methods Med Res* 2018;27:-26. doi: 10.1177/0962280216683570 pmid: 28034175

15 Groenwold RH, White IR, Donders AR, Carpenter JR, Altman DG, Moons KG. Missing covariate data in clinical research: when and when not to use the missing-indicator method for analysis. *CMAJ* 2012;184:-9. doi: 10.1503/cmaj.110977 pmid: 22371511

16 Hughes RA, Heron J, Sterne JAC, Tilling K. Accounting for missing data in statistical analyses: multiple imputation is not always the answer. *Int J Epidemiol* 2019;48:-304. doi: 10.1093/ije/dyz032 pmid: 30879056

17 Lee KJ, Tilling KM, Cornish RP, etalSTRATOS initiative. Framework for the treatment and reporting of missing data in observational studies: The Treatment And Reporting of Missing data in Observational Studies framework. *J Clin Epidemiol* 2021;134:-88. doi: 10.1016/j.jclinepi.2021.01.008 pmid: 33539930

18 van Buuren S. Flexible Imputation of Missing Data (2nd ed). Boca Raton, FL, 2018.

19 Altman DG, Royston P. The cost of dichotomising continuous variables. *BMJ* 2006;332:. doi: 10.1136/bmj.332.7549.1080 pmid: 16675816

20 Royston P, Altman DG, Sauerbrei W. Dichotomizing continuous predictors in multiple regression: a bad idea. *Stat Med* 2006;25:-41. doi: 10.1002/sim.2331 pmid: 16217841

21 Altman DG, Lausen B, Sauerbrei W, Schumacher M. Dangers of using "optimal" cutpoints in the evaluation of prognostic factors. *J Natl Cancer Inst* 1994;86:-35. doi: 10.1093/jnci/86.11.829 pmid: 8182763

22 Collins GS, Ogundimu EO, Cook JA, Manach YL, Altman DG. Quantifying the impact of different approaches for handling continuous predictors on the performance of a prognostic model. *Stat Med* 2016;35:-35. doi: 10.1002/sim.6986 pmid: 27193918

23 MacCallum RC, Zhang S, Preacher KJ, Rucker DD. On the practice of dichotomization of quantitative variables. *Psychol Methods* 2002;7:-40. doi: 10.1037/1082-989X.7.1.19 pmid: 11928888

24 Senn S, Julious S. Measurement in clinical trials: a neglected issue for statisticians?*Stat Med* 2009;28:-209. doi: 10.1002/sim.3603 pmid: 19455540

25 Senn S. Individual response to treatment: is it a valid assumption?*BMJ* 2004;329:-8. doi: 10.1136/bmj.329.7472.966 pmid: 15499115

26 Chilvers C, Dewey M, Fielding K, etalCounselling versus Antidepressants in Primary Care Study Group. Antidepressant drugs and generic counselling for treatment of major depression in primary care: randomised trial with patient preference arms. *BMJ* 2001;322:-5. doi: 10.1136/bmj.322.7289.772 pmid: 11282864

27 Durrleman S, Simon R. Flexible regression models with cubic splines. *Stat Med* 1989;8:-61. doi: 10.1002/sim.4780080504 pmid: 2657958

28 Harrell FEJr. *Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis*. 2nd ed. Springer, 2015doi: 10.1007/978-3-319-19425-7.

29 Nieboer D, Vergouwe Y, Roobol MJ, etalProstate Biopsy Collaborative Group. Nonlinear modeling was applied thoughtfully for risk prediction: the Prostate Biopsy Collaborative Group. *J Clin Epidemiol* 2015;68:-34. doi: 10.1016/j.jclinepi.2014.11.022 pmid: 25777297

30 Sauerbrei W, Royston P. Building multivariable prognostic and diagnostic models: transformation of the predictors by using fractional polynomials. *J R Stat Soc [Ser A]* 1999;162:-94doi: 10.1111/1467-985X.00122.

31 Royston P, Sauerbrei W. *Multivariable Model-Building - A pragmatic approach to regression analysis based on fractional polynomials for modelling continuous variables*. Wiley, 2008doi: 10.1002/9780470770771.

32 Royston P, Altman DG. Regression Using Fractional Polynomials of Continuous Covariates: Parsimonious Parametric Modelling. *J R Stat Soc Ser C Appl Stat* 1994;43:-67.

33 Sauerbrei W, Royston P, Bojar H, Schmoor C, Schumacher MGerman Breast Cancer Study Group (GBSG). Modelling the effects of standard prognostic factors in node-positive breast cancer. *Br J Cancer* 1999;79:-60. doi: 10.1038/sj.bjc.6690279 pmid: 10206288

34 Johannesen CDL, Langsted A, Mortensen MB, Nordestgaard BG. Association between low density lipoprotein and all cause and cause specific mortality in Denmark: prospective cohort study. *BMJ* 2020;371:. doi: 10.1136/bmj.m4266 pmid: 33293274

35 Altman DG, Bland JM. Interaction revisited: the difference between two estimates. *BMJ* 2003;326:. doi: 10.1136/bmj.326.7382.219 pmid: 12543843

36 Austin PC, Hux JE. A brief note on overlapping confidence intervals. *J Vasc Surg* 2002;36:-5. doi: 10.1067/mva.2002.125015 pmid: 12096281

37 VanderWeele Tyler J, Knol Mirjam J. A Tutorial on Interaction. *Epidemiol Methods* 2014;3:.

38 Shrier I, Pang M. Confounding, effect modification, and the odds ratio: common misinterpretations. *J Clin Epidemiol* 2015;68:-4. doi: 10.1016/j.jclinepi.2014.12.012 pmid: 25662008

39 Williamson SF, Grayling MJ, Mander AP, etal. Subgroup analyses in randomised controlled trials frequently categorised continuous subgroup information. *J Clin Epidemiol* 2022 1 Jul. doi: 10.1016/j.jclinepi.2022.06.017.

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:-25. doi: 10.1002/sim.1815 pmid: 15287081

41 Peters TJ, Richards SH, Bankhead CR, Ades AE, Sterne JA. Comparison of methods for analysing cluster randomized trials: an example involving a factorial design. *Int J Epidemiol* 2003;32:-6. doi: 10.1093/ije/dyg228 pmid: 14559762

42 Bland JM. Cluster randomised trials in the medical literature: two bibliometric surveys. *BMC Med Res Methodol* 2004;4:. doi: 10.1186/1471-2288-4-21 pmid: 15310402

43 Lee KJ, Thompson SG. The use of random effects models to allow for clustering in individually randomized trials. *Clin Trials* 2005;2:-73. doi: 10.1191/1740774505cn082oa pmid: 16279138

44 Steyerberg EW, Bossuyt PM, Lee KL. Clinical trials in acute myocardial infarction: should we adjust for baseline characteristics?*Am Heart J* 2000;139:-51. doi: 10.1016/S0002-8703(00)90001-2 pmid: 10783203

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:-60. doi: 10.1016/j.jclinepi.2003.09.014 pmid: 15196615

46 Turner EL, Perel P, Clayton T, etalCRASH trial collaborators. Covariate adjustment increased power in randomized controlled trials: an example in traumatic brain injury. *J Clin Epidemiol* 2012;65:-81. doi: 10.1016/j.jclinepi.2011.08.012 pmid: 22169080

47 Riley RD, Tierney JF, Stewart LA, eds. *Individual Participant Data Meta-Analysis: A Handbook for Healthcare Research*. Wiley, 2021doi: 10.1002/9781119333784.

48 Abo-Zaid G, Guo B, Deeks JJ, etal. Individual participant data meta-analyses should not ignore clustering. *J Clin Epidemiol* 2013;66:-873.e4. doi: 10.1016/j.jclinepi.2012.12.017 pmid: 23651765

49 Greenland S, Robins MR, Pearl J. Confounding and collapsibility in causal inference. *Stat Sci* 1999;14:-46doi: 10.1214/ss/1009211805.

50 Robinson LD, Jewell NP. Some surprising results about covariate adjustment in logistic regression models. *Int Stat Rev* 1991;59:-40doi: 10.2307/1403444.

51 Gail MH, Wieand S, Piantadosi S. Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates. *Biometrika* 1984;71:-44doi: 10.1093/biomet/71.3.431.

52 Gardiner JC, Luo Z, Roman LA. Fixed effects, random effects and GEE: what are the differences?*Stat Med* 2009;28:-39. doi: 10.1002/sim.3478 pmid: 19012297

53 Kahan BC, Morris TP. Assessing potential sources of clustering in individually randomised trials. *BMC Med Res Methodol* 2013;13:. doi: 10.1186/1471-2288-13-58 pmid: 23590245

54 Kahan BC, Li F, Copas AJ, Harhay MO. Estimands in cluster-randomized trials: choosing analyses that answer the right question. *Int J Epidemiol* 2022;dyac131. doi: 10.1093/ije/dyac131 pmid: 35834775

55 Higgins JP, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. *BMJ* 2003;327:-60. doi: 10.1136/bmj.327.7414.557 pmid: 12958120

56 Rücker G, Schwarzer G, Carpenter JR, Schumacher M. Undue reliance on I(2) in assessing heterogeneity may mislead. *BMC Med Res Methodol* 2008;8:. doi: 10.1186/1471-2288-8-79 pmid: 19036172

57 Thompson SG, Higgins JPT. How should meta-regression analyses be undertaken and interpreted?*Stat Med* 2002;21:-73. doi: 10.1002/sim.1187 pmid: 12111920

58 Riley RD, Debray TPA, Fisher D, etal. 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:-37. doi: 10.1002/sim.8516 pmid: 32350891

59 Fisher DJ, Carpenter JR, Morris TP, Freeman SC, Tierney JF. Meta-analytical methods to identify who benefits most from treatments: daft, deluded, or deft approach?*BMJ* 2017;356:. doi: 10.1136/bmj.j573 pmid: 28258124

60 Riley RD, Lambert PC, Abo-Zaid G. Meta-analysis of individual participant data: rationale, conduct, and reporting. *BMJ* 2010;340:. doi: 10.1136/bmj.c221 pmid: 20139215

61 Steyerberg EW, Moons KG, van der Windt DA, etalPROGRESS Group. Prognosis Research Strategy (PROGRESS) 3: prognostic model research. *PLoS Med* 2013;10:e1001381. doi: 10.1371/journal.pmed.1001381 pmid: 23393430

62 Altman DG, Vergouwe Y, Royston P, Moons KG. Prognosis and prognostic research: validating a prognostic model. *BMJ* 2009;338:. doi: 10.1136/bmj.b605 pmid: 19477892

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:. doi: 10.1136/bmj.i6 pmid: 26810254

64 Jaja BNR, Saposnik G, Lingsma HF, etalSAHIT collaboration. Development and validation of outcome prediction models for aneurysmal subarachnoid haemorrhage: the SAHIT multinational cohort study. *BMJ* 2018;360:. doi: 10.1136/bmj.j5745 pmid: 29348138

65 Vickers AJ, van Calster B, Steyerberg EW. A simple, step-by-step guide to interpreting decision curve analysis. *Diagn Progn Res* 2019;3:. doi: 10.1186/s41512-019-0064-7 pmid: 31592444

66 Copas JB. Using regression models for prediction: shrinkage and regression to the mean. *Stat Methods Med Res* 1997;6:-83. doi: 10.1177/096228029700600206 pmid: 9261914

67 Van Houwelingen JC. Shrinkage and penalized likelihood as methods to improve predictive accuracy. *Stat Neerl* 2001;55:-34doi: 10.1111/1467-9574.00154.

68 Firth D. Bias reduction of maximum likelihood estimates. *Biometrika* 1993;80:-38doi: 10.1093/biomet/80.1.27.

69 Greenland S, Mansournia MA, Altman DG. Sparse data bias: a problem hiding in plain sight. *BMJ* 2016;352:. doi: 10.1136/bmj.i1981 pmid: 27121591

70 Heinze G, Wallisch C, Dunkler D. Variable selection - A review and recommendations for the practicing statistician. *Biom J* 2018;60:-49. doi: 10.1002/bimj.201700067 pmid: 29292533

71 Riley RD, van der Windt D, Croft P, etal, eds. *Prognosis Research in Healthcare: Concepts, Methods and Impact*. Oxford University Press, 2019doi: 10.1093/med/9780198796619.001.0001.

72 Richiardi L, Bellocco R, Zugna D. Mediation analysis in epidemiology: methods, interpretation and bias. *Int J Epidemiol* 2013;42:-9. doi: 10.1093/ije/dyt127 pmid: 24019424

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:-16. doi: 10.1016/0895-4356(96)00025-X pmid: 8699212

74 Ensor J, Riley RD, Jowett S, etalPIT-STOP collaborative group. 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:-xxxiii, 1-190. doi: 10.3310/hta20120 pmid: 26879848

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:. doi: 10.1186/1745-6215-15-314 pmid: 25098243

76 Altman DG. Better reporting of randomised controlled trials: the CONSORT statement. *BMJ* 1996;313:-1. doi: 10.1136/bmj.313.7057.570 pmid: 8806240

77 Simera I, Moher D, Hirst A, Hoey J, Schulz KF, Altman DG. Transparent and accurate reporting increases reliability, utility, and impact of your research: reporting guidelines and the EQUATOR Network. *BMC Med* 2010;8:. doi: 10.1186/1741-7015-8-24 pmid: 20420659

78 Debray TPA, Collins GS, Riley RD, etal. Transparent reporting of multivariable prediction models developed or validating using clustered data: TRIPOD-Cluster checklist. *BMJ* 2023;380:in press.

79 Schulz KF, Altman DG, Moher DCONSORT Group. CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. *BMJ* 2010;340:. doi: 10.1136/bmj.c332 pmid: 20332509

80 Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): the TRIPOD statement. *Ann Intern Med* 2015;162:-63. doi: 10.7326/M14-0697 pmid: 25560714

81 Moons KG, Altman DG, Reitsma JB, etal. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. *Ann Intern Med* 2015;162:. doi: 10.7326/M14-0698 pmid: 25560730

82 Boutron I, Altman DG, Hopewell S, Vera-Badillo F, Tannock I, Ravaud P. Impact of spin in the abstracts of articles reporting results of randomized controlled trials in the field of cancer: the SPIIN randomized controlled trial. *J Clin Oncol* 2014;32:-6. doi: 10.1200/JCO.2014.56.7503 pmid: 25403215

83 Altman DG, Bland JM. Quartiles, quintiles, centiles, and other quantiles. *BMJ* 1994;309:. doi: 10.1136/bmj.309.6960.996 pmid: 7950724

84 Groves T. Research methods and reporting. *BMJ* 2008;337:. doi: 10.1136/bmj.a2201.

85 Assel M, Sjoberg D, Elders A, etal. Guidelines for Reporting of Statistics for Clinical Research in Urology. *J Urol* 2019;201:-604. doi: 10.1097/JU.0000000000000001 pmid: 30633111

86 Stratton IM, Neil A. How to ensure your paper is rejected by the statistical reviewer. *Diabet Med* 2005;22:-3. doi: 10.1111/j.1464-5491.2004.01443.x pmid: 15787658