Sparse data bias: a problem hiding in plain sight

Sander Greenland,[1] Mohammad Ali Mansournia,[2] Douglas G Altman[3]

[1] Department of Epidemiology and Department of Statistics, University of California, Los Angeles, CA, USA

[2] Department of Epidemiology and Biostatistics, School of Public Health, Tehran University of Medical Sciences, PO box 14155-6446, Tehran, Iran

[3] Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, University of Oxford, Oxford, UK

Correspondence to: M A Mansournia mansournia_ma@yahoo.com

Additional material is published online only. To view please visit the journal online.

Cite this as: *BMJ* 2016;353:i1981
http://dx.doi.org/10.1136/bmj.i1981

Accepted: 07 March 2016

Effects of treatment or other exposure on outcome events are commonly measured by ratios of risks, rates, or odds. Adjusted versions of these measures are usually estimated by maximum likelihood regression (eg, logistic, Poisson, or Cox modelling). But resulting estimates of effect measures can have serious bias when the data lack adequate case numbers for some combination of exposure and outcome levels. This bias can occur even in quite large datasets and is hence often termed sparse data bias. The bias can arise or be worsened by regression adjustment for potentially confounding variables; in the extreme, the resulting estimates could be impossibly huge or even infinite values that are meaningless artefacts of data sparsity. Such estimate inflation might be obvious in light of background information, but is rarely noted let alone accounted for in research reports. We outline simple methods for detecting and dealing with the problem focusing especially on penalised estimation, which can be easily performed with common software packages.

Ratio measures such as odds ratios, rate ratios, and risk ratios are commonly used to quantify the effect of a treatment or other factor on an event outcome. Adjusted versions of these measures are usually estimated by maximum likelihood regression (eg, logistic regression for odds ratios, Poisson regression, or Cox regression for rate ratios). These methods assume that the number of events observed is sufficient at all treatment levels to result in well behaved adjusted estimates. Unfortunately, when the data lack adequate case numbers for some combination of risk factor and outcome levels, the resulting estimates of the regression coefficients can have bias away from the null (downward when the estimate is below 1, upward when it is above 1).

This bias is sometimes called a “small sample bias” but in fact can occur in quite large datasets and thus is better termed sparse data bias.[1] The problem is worsened by the fact that estimated ratios are found by taking the antilogs (exponentiation) of the coefficients, which adds a further upward bias. The consequences can be quite serious when one is trying gauge the size of the effects under study after regression adjustment for potentially confounding variables, and could be worse than the bias removed by the adjustment. In the extreme, the resulting ratio estimates could be impossibly huge or even take on infinite values that are meaningless artefacts of failure of the program to converge. Such estimate inflation may be obvious in light of background information, but is rarely noted let alone accounted for in research reports. We outline simple methods for detecting and dealing with the problem.

Preliminary checks for sparse data bias

Data sparsity is usually unrecognised when the total sample size is large, as evidenced by the fact that authors and discussants rarely comment on the plausibility of huge estimates. As an extreme example, one case-control study reported unadjusted odds ratios (95% confidence intervals) for the association of ever smoked, cemented arthroplasty, and general anaesthesia with intensive care unit admission after total joint arthroplasty of 10.63 (5.58 to 20.26), 1.02 (0.63 to 1.64), 4.49 (2.44 to 8.26), respectively.[2] After logistic regression involving 12 explanatory variables and 120 outcome events (cases), these odds ratios (95% confidence interval) were 65.13 (6.31 to 672.09), 55.75 (1.64 to 1893.70), and 45.22 (1.10 to 1851.81), respectively. The smoking relation is especially implausible, and provides a warning that the increase in estimates after adjustment is likely to reflect sparse data bias rather than removal of confounding.

SUMMARY POINTS

Maximum likelihood estimates (MLEs) of odds ratios, rate ratios, and risk ratios can have considerable upward bias when there are few or no study participants at key combinations of the outcome, exposure, and covariates, often known as sparse data bias

The hallmark sign of sparse data bias in multivariable analysis is that the model coefficients estimates get further and further from the null as more variables are used for stratification or added to the regression model

Factors contributing to sparse data bias include low event per variable (EPV), categorical covariates with very low or high prevalence, and narrowly distributed continuous predictors

Several rules based on EPV have been proposed to detect or avoid sparse data bias. The most direct approach however is to apply a method that removes or limits sparse data bias. We illustrate the use of bias adjustments and penalised estimation for that purpose. Penalisation can be easily performed with common software packages

Penalisation is a form of shrinkage estimation, in which external (or prior) information is used to improve accuracy over repeated studies. In this goal it differs from Bayesian analyses

Such examples illustrate how extreme instances of sparse data bias might be spotted in published reports: the results are far out of line with sensible expectations. More effort is required to spot less dramatic bias. For example, one well known case-control study reported an odds ratio of 15.92 (95% confidence interval 1.38 to 184.13) for the association between phenylpropanolamine use and stroke, calculated from a conditional logistic regression with four covariates but only one exposed non-case.[3] The unadjusted odds ratio was 11.85 (the odds ratio adjusted for the matching factors only was not given, but it can be derived as 12 based on the reported numbers). Given the indications for phenylpropanolamine use, however, we should expect higher background risk for those who use suppressants, in which case the estimates should have become lower on further risk adjustment. Instead an increase was seen, suggesting worsened bias due to the regression adjustment.

To detect problems before embarking on our own analyses, it is useful to examine basic data for features leading to the bias. The following features (which are themselves intertwined) contribute to sparse data bias in maximum likelihood regression analyses of disease outcomes:

- Few outcome events per variable (EPV), as measured by the number of failures per variable for Cox proportional hazards and Poisson regression, and the minimum of the numbers of cases and non-cases per variable for logistic regression (for conditional logistic regression, only the numbers within discordant matched sets should be counted)
- Variables with narrow distributions or with categories that are very uncommon
- Variables that together almost perfectly predict the outcome (eg, if a combination of discrete covariate levels is found only among the study participants with outcome)
- Variables that together almost perfectly predict the exposure (eg, if a combination of discrete covariate levels is found only among the study participants who are exposed).

Most modelling guidelines are based on EPV; for example, some authors recommend an EPV of at least 10 for developing prediction models,[4 5] while others recommend at least five EPV if the model is only being used for confounding adjustment.[6] Although such guidelines are useful, they are not infallible and thus we recommend more direct checking as well.

The simplest supplementary diagnostic method for sparse data problems is detailed tabular examination of the basic data, including unadjusted and simple stratified estimates. This basic format also allows one to quickly see the effect of small data changes and of simple bias reduction methods. Consider a case-control study of childhood leukaemia in the vicinity of nuclear reprocessing.[7] Table 1 shows the cross tabulation of the data based on the outcome (1: leukaemia, 0: local controls) and paternal exposure (1: ≥100 mSv, 0: <100 mSv). There was only one unexposed case and one exposed non-case. The unadjusted rate ratio estimate from these data is the sample odds ratio (3×19)÷(1×1)=57, which appears severely inflated based on what is reasonable to expect given what is known about leukaemia risk factors.

Suspicions that the odds ratio of 57 is an artefact of the small numbers are reinforced by noting that by reclassifying one of the three exposed cases as unexposed, the estimate drops by two thirds to (2×19)÷(2×1)=19. One method for bias reduction adds 1 to each denominator count cell;[8] in doing so, the estimate drops by three quarters to (3×19)÷(2×2)=14.25. These checks corroborate background information indicating that the original estimate of 57 must be a gross exaggeration relative to any real effect.

Beyond these basic numerical checks, we recommend direct comparison of the ordinary estimates against estimates more robust to the bias, which we describe below.

Traditional solutions

The analysis plan should set forth a priori the variables that are to be considered potential confounders and thus subject to adjustment (for example, using causal diagrams[9]), and try to adjust for these variables if they meet prespecified importance criteria. Nonetheless, on stratification to adjust for confounders, sparse data artefacts become even more severe, with an increased risk of zero cells leading to infinite or undefined estimates. As a consequence, most analyses turn to model based adjustment, in which problems of sparse data are often hidden.

Suppose, however, that an analyst has checked and determined that ordinary regression analysis cannot support inclusion of all the available variables—for example, because the data appear too sparse using the aforementioned checks. The usual response would be to remove some variables from the model based on significance testing. Unfortunately, such test based selection of variables can easily worsen bias in effect estimates because it may drop important confounders.[10 11] Also, if (as usual) no accounting is made for the variable selection, the final P values will be too small and confidence intervals will be too narrow for the coefficients of the remaining variables (table 2).[12 13] These problems worsen as the P value cutoff point for selection (the selection α level) becomes smaller, leading some authors to advise using high P value cutoff points, as high as 0.20.[14 15]

Other authors have advised using exact logistic regression to avoid sparse data bias and related problems.[16] Although such analyses can provide a useful perspective on other results, for the situations of concern here (regression with discrete outcomes and multiple confounders), exact P values tend to be too

Table 1 | Paternal radiation exposure (≥100 mSv *v* <100 mSv) and childhood leukaemia

Exposure ≥100 mSv	Childhood leukaemia		
	Yes	No	Total
Yes	3	1	4
No	1	19	20
Total	4	20	24

Table 2 | Advantages and disadvantages of approaches to sparse data bias

Approach	Advantages	Disadvantages
Stepwise variable selection procedures	Simple and available in almost all statistical software	Inflated estimates, confidence intervals too narrow, and P values too small for the selected variables; can severely bias effect estimates because it may drop important confounders
Exact statistical methods (eg, exact logistic regression)	No sample size requirement	Computationally intensive, especially with large sample sizes or many covariates; not available or feasible for all models or measures (eg, risk differences, risk ratios); P values too large and confidence intervals too wide
Exposure or treatment modelling (eg, propensity scoring, inverse-probability-of-treatment weighting)	Can be more accurate than outcome modelling in cohort studies if the exposure or treatment is common but the outcome is rare	Can be less accurate than outcome modelling, especially if the outcome is more common than the exposure or the exposure is well predicted by variables in the score; prone to statistical artefacts in case-control studies
Penalisation	Produces the most accurate estimates given the information in the penalty; data augmentation version is simple and feasible in all statistical software; can be used as a diagnostic tool for sparse data bias	The penalty factor must be determined by the research team based on background information (Firth adjustment does not require this but is not the most accurate form of penalisation)

large and exact confidence intervals tend to be unnecessarily wide (conservative),[17] thus understating the information provided by the data and potentially misleading the analyst and reader. Additionally, they are computationally intensive to the point that they cannot handle very large datasets or very large numbers of covariates, which are precisely the situations under which sparse data bias is most likely to go unnoticed.

As illustrated earlier, conventional bias reduction methods for tabular data add small numbers to cell counts, which can be thought of as pseudodata conveying information that the true association is probably smaller than what was observed. A common traditional choice (which reduces bias on the log-ratio scale) adds ½ to each cell, which in the leukaemia example yields an odds ratio estimate of (3.5×19.5)÷(1.5×1.5)=30.3. For regression analysis, one generalisation of adding ½ to each cell is the Firth bias adjustment,[18] which is a form of penalised estimation available as the option FIRTH in SAS and the command firthlogit in Stata.[19 20] Although less biased than the usual (maximum likelihood) estimate, these methods impose certain implausible background assumptions on the model coefficients and could leave unacceptably large bias in ratio estimates.[21-23] These problems can be avoided by other types of penalisation, as discussed next.

Penalisation: diagnostic tool for and solution to sparse data bias

Ordinary estimation methods can be modified to make them more resistant to sparse data bias even if all variables are left in the model. This can be done with common statistical software by adding artificial data records that penalise (shrink) coefficient estimates in proportion to their size, thus incorporating background information that the true effects are not extreme.[19 20] This penalisation arguably produces the best estimates available given that background information. A basic diagnostic method for regression results is to repeat the analysis using mild penalisation; important changes warn of serious bias in the original (unpenalised) estimates.

Penalisation is mathematically identical to what is known as Bayesian analysis. Although its goal differs insofar as it is oriented toward improving the calibration (frequency) properties of the resulting statistics, its Bayesian interpretation guides the degree of penalisation based on background information. Considering the leukaemia example, the conventional odds ratio of 57 can be viewed as a result of the following assumption: before seeing the data, we had no idea what to expect and would not have been surprised if paternal exposure was a sufficient protectant (exposed children never get leukaemia, corresponding to causal rate ratios below 10^{-6}), or a sufficient cause (exposed children always get leukaemia, corresponding to causal rate ratios over 10^{6}). This assumption is absurd, and thus huge effect estimates should be taken as signalling huge error rather than huge effects.

To reduce (shrink) huge estimates, one may begin by specifying an interval which encodes the idea that the true effect is not huge. For example, presumably everyone would be almost sure that the causal rate ratio falls between 1/40 and 40. If it was below 1/40, paternal exposure would be a near-perfect protectant; but if it was over 40, there probably would have been a massive leukaemia outbreak among the exposed people. Thus it would be uncontroversial to say we are at least 95% certain the true value is between 1/40 and 40. The interval of 1/40 to 40 would then become a conservative 95% prior interval for the effect, where "conservative" means that the 95% is a minimum certainty and "prior" means the limits were derived from background information rather than the study data.

There are many ways this prior interval can be incorporated into the analysis, some of which are quite complex, but all tend to produce qualitatively similar estimates. We thus focus on the data augmentation method, which translates prior intervals into simple prior data (pseudodata) that are added to the actual data; details and SAS and Stata code can be found elsewhere.[19 20] Briefly, instead of adding pseudorecords directly to cell counts, data augmentation treats them as distinct data records which, if analysed by proper methods, would reproduce the prior interval for each variable as a confidence interval. Appending this prior data set to the actual data thus incorporates the information contained in the prior intervals, pulling extreme estimates into a more reasonable range.

Figure 1 illustrates penalisation using data augmentation in the leukaemia example. To allow easy extension to regression analysis, instead of adding 1 to each count in the odds ratio denominator, augmentation can be done by adding a pair of artificial patient records representing one exposed case and one exposed non-case, which have all other variables set to zero (including the regression "constant" variable).[19 20] This pair of records encodes a 95% prior odds ratio interval of exactly 1/39 to 39 (derivable from an *F*(2,2) prior distribution for the odds ratio). The overwhelming majority of effects subject to epidemiological study fall well within this range, and in the present example this interval extends far beyond what background information suggests as reasonable for the true ratio effect. When the records are appended to the leukaemia data and the new augmented dataset is analysed with logistic regression, the resulting penalised estimate is 11.57—far more consistent with background information than the original estimate of 57 or the Firth estimate of 30.33.

Penalty functions and prior distributions for regression coefficients can be incorporated into the analysis using data augmentation. Data equivalent to the penalty function or prior distribution (prior data) are constructed and appended to the actual study data, after which conventional maximum likelihood regression is performed on the augmented (actual+prior) data. We illustrate this method using the data in table 1.

Genuine (G)	Exposure (X)	Disease (Y)	Frequency weight
1	1	1	3
1	1	0	1
1	0	1	1
1	0	0	19

The lead variable, G, is an indicator that is 1 for data records representing genuine (actual) observations. Consider the logistic model in which probability of disease (Y=1) depends on exposure via the equation:

$$\text{logit}(\pi) = \beta_0 + \beta_1 X$$

where logit(π) is the natural logarithm of the odds $\pi \div (1-\pi)$ and the odds ratio relating exposure to disease is $\exp(\beta_1)$. Fitting the model to these data yields the estimated odds ratio of 57 (95% Wald confidence interval: 2.76 to 1177; 95% profile-likelihood interval: 3.90 to 2392). We can, however, get identical results by fitting the model:

$$\text{logit}(\pi) = \beta_0 G + \beta_1 X$$

using the option to prevent the program from automatically adding the intercept β_0 (eg, via the NOINT option in SAS, the noconstant option in Stata, or adding "-1" in the model formula in R).

A penalised analysis using a *F*(2,2) prior distribution for the odds ratio (which implies a prior odds ratio interval of 1/39 to 39) can be done by adding a pair of pseudo-records representing one exposed case (X=1, Y=1) and one exposed non-case (X=1, Y=0), with G=0 in both records to indicate they are not genuine observations:

Genuine (G)	Exposure (X)	Disease (Y)	Frequency weight
1	1	1	3
1	1	0	1
1	0	1	1
1	0	0	19
0	1	1	1
0	1	0	1

Fitting the model $\text{logit}(\pi) = \beta_0 G + \beta_1 X$ to the augmented data (making sure the program does not add an intercept) produces an estimated odds ratio of 11.57 (95% Wald interval 1.24 to 108; 95% profile likelihood interval 1.46 to 146).

For regressions involving multiple variables, any variable can be given its own pair of penalising pseudo-records (although it is not necessary to do so; eg, one might not penalise estimates for age and sex effects). In this pair, the entries for the variable represented by the pair should be set to 1, with Y=1 in one record and Y=0 in the other. All other variables in the pseudo records should be set to zero—even variables that cannot be zero in actual data, such as blood pressure. The total of the frequency weight for the pair is the degrees of freedom M for the *F*(M,M) prior distribution for the odds ratio that the pair represents, with larger M corresponding to narrower distributions.

The above method assumes that a 1 unit change in the variable X is a sensible degree of change for the variable. For an indicator (0,1) variable this requirement is automatically satisfied, but for quantities one may need to change units. For example, blood pressure is ordinarily measured in mm, but 1 mm is too small a unit to be meaningful, so we would advise switching to cm by dividing the original variable by 10. Such rescaling will also help with coefficient interpretation.

The prior data method can be extended to impose normal, skewed or non-null centred priors,[24,25] and can be applied to other regression models including conditional logistic, Poisson, and Cox proportional hazards regression.[19]

Fig 1 | Penalisation using data augmentation

For variables whose effects can be very large (ratios below 1/40 or above 40) a weaker penalisation may be used. For example, if the program allows weighting of data records (as do SAS and Stata), then setting all the actual data records to weight 1 while down weighting the augmenting records to ½ encodes the 95% prior interval of 1/648 to 648. In the leukaemia example, this penalty produces an estimate of 22.22. Table 3 provides correspondences between simple weighting factors *w* and prior 95% intervals for the odds ratio (based on an *F*(2*w*,2*w*) prior distribution).[24 25]

However, we caution that including variables with actual effects far beyond the ratio range of 1/40 to 40 can disrupt typical fitting methods by creating very sparse data in certain categories; thus such effects are better controlled using restriction or matching rather than modelling.[26] On the other hand, for variables whose effects are known to be weak, a stronger penalisation can be used by increasing the weight of the augmenting records. For example, using a weight of 4.5 for both records corresponds to a 95% prior interval of ¼ to 4 for the ratio,[19 20] and in the example produces an estimate of 2.40.

Penalisation by data augmentation is also easily performed for multiple regression analyses. Each variable in the regression can be given its own data with its own weight determined by background (contextual) information. If desired, some variables may be left unpenalised by adding no record for those variables; this treatment may be the simplest course for variables whose effects are incontrovertibly large (typically age, sex, and (for oral and respiratory disease studies) smoking). Detailed descriptions of penalised regression are available along with SAS and Stata software,[19 20] and an R procedure is provided in the web appendix.

Interval estimates and P values

In addition to inflating estimates, data sparsity also invalidates the usual method for computing confidence

Table 3 | Weight *w* for prior record to impose indicated 95% prior limits to 3 digit accuracy*

Prior 95% limits for odds ratio	Weight *w*
1/50 to 50	0.918
1/40 to 40	0.991
1/32 to 32	1.08
1/25 to 25	1.19
1/20 to 20	1.32
1/16 to 16	1.47
1/10 to 10	1.95
⅛ to 8	2.28
1/5 to 5	3.50
¼ to 4	4.54
⅓ to 3	6.92
½ to 2	16.6
⅔ to 3/2	47.3
3/4 to 4/3	93.4
4/5 to 5/4	155
5/6 to 6/5	232

*Based on *F*(2*w*,2*w*) prior distribution for odds ratio.

intervals (called the Wald method) in which the 95% confidence interval for the ratio is found from the estimated log rate ratio or regression coefficient *b* and its standard error *s* as $e^{b\pm1.96s}$, which assumes the estimate *b* comes from a normal (Gaussian) distribution. In the leukaemia example, *b* is ln(57)=4.043 with *s*=1.545, so the 95% Wald interval is $e^{4.043\pm1.96(1.545)}$=2.76 to 1177. When instead we compute the interval using the much more accurate profile likelihood method (available in R using the confint(fit) option), we obtain an interval of 3.90 to 2392; the pllf command in Stata gives 3.90 to 2391 and the <clodds=pl> option in the model statement in SAS gives 3.90 to ">999.999." Similarly, penalised likelihood gives *b* as ln(11.57)=2.448 with *s*=1.140 so the 95% Wald interval is $e^{2.448\pm1.96(1.140)}$=1.24 to 108, whereas the profile likelihood interval is 1.46 to 146. Wald P values (computed from the *Z* score *b*/*s*) will also be distorted by data sparsity. We thus strongly advise using profile likelihood intervals and P values if there is any concern about sample size or sparsity, as in the above examples.

Degree of penalisation: relation of penalisation to Bayesian methods

We recommend setting the prior 95% interval to encompass every remotely reasonable possibility for the actual size of the ratio being estimated in light of previous studies. An acceptable prior interval would thus have the property that readers from the research community would assign at least 95% probability to the true ratio being in the interval—even though they might differ considerably about which values within the interval are probable or improbable, and would not be in conflict with any available estimate. The goals of this requirement are to avoid controversy about the penalty and to allay concerns that its use will meaningfully bias the final results.

The second goal can be recast as stating that the primary purpose of penalisation is to improve the calibration of statistical results, in the conventional frequentist sense of providing more accuracy on average across studies. We thus depart from some Bayesian teaching and practice in which the goal is to form inferences (posterior distributions) that represent a combination of data evidence with the opinions of available experts as summarised in a prior distribution. We particularly disfavour exclusive reliance on opinion based prior intervals for effects, since expert opinion often appears biased, overconfident (overly precise), or otherwise misinformed when critically evaluated against actual studies.

The prior interval we recommend represents instead the broadest possible consensus, taking into account possibly conflicting and biased literature as well as expert opinions about that literature. This means, for example, that justification for a narrow interval will require critical examination of meta-analyses, since those studies incorporate many biases of their own (including the biases in their constituent studies). Thus, although penalisation does make use of background information (as all good modelling should), one should not regard penalisation as a solely Bayesian method. The sparse data problem it addresses arises in conventional frequentist analyses, and frequentists may use their background knowledge and common sense to spot and address the problem.

More generally, penalisation is a type of shrinkage estimation, in which the goal is to use external information to produce estimators that have better calibration (better accuracy on average over repeated studies) than do traditional estimators (box 1). Other shrinkage methods similar or mathematically identical to penalisation include Stein, empirical Bayes, and partial Bayes (semi-Bayes) estimation, as well as random coefficient regression and ridge regression. These are all legitimate frequentist methods, and none requires the elaborate computing machinery (such as Markov chain Monte Carlo) demanded by many strict Bayesians.

Discussion and recommendations

Although we have focused on effect estimation, penalisation can also be used to improve the accuracy of risk prediction, and is thus worthwhile even when model coefficients are not of direct interest.[27-29] Penalised prediction is however often conducted using what are known as lasso penalties, which we do not advise when causal effects are the target because they may delete important confounders from the regression, thus adding another source of bias.

There are several strategies to avoid or minimise sparse data bias. For example, propensity scoring or other exposure or treatment modelling methods are sometimes advised, along with (or instead of) outcome modelling to

Box 1: Glossary

- Maximum likelihood regression: a regression model (such as logistic regression, Poisson regression, and Cox regression) whose coefficients are estimated by maximising the likelihood function (the probability of observed data, expressed as the function of the unknown model coefficients).
- Sparse data bias: the bias in estimates when the data lack adequate numbers of observations for some combination of risk factor and outcome levels, which may arise even if the total sample size appears large.
- Events per variable (EPV): a rough measure of the effective sample size for estimating coefficients in the regression model. The number of failures per variable for Cox proportional hazards and Poisson regression, and the minimum of the numbers of cases and non-cases per variable for logistic regression.
- Exact logistic regression: logistic regression in which the coefficients are estimated by exact sampling distributions rather than approximations as in maximum likelihood. The exact 95% confidence intervals resulting from this method are only exact in that they are derived from exact P values; they may be excessively wide and as a result cover the true coefficient much more than 95% of the time.
- Stepwise selection procedures: a variable selection algorithm that adds or deletes explanatory variables from the regression model based on statistical significance. Produces highly distorted estimates and tests.
- Penalisation: the methods which add a penalty (adjustment) factor to the original likelihood of the actual data and shrink the final estimates away from the original estimates towards the values specified in the penalty factor. They reduce the mean squared error (= bias^2 + variance) of the estimates whenever they reduce variance more than they increase bias or whenever they reduce bias. A type of shrinkage estimation method.
- Firth bias adjustment: a type of penalisation based on the data which reduces the bias of maximum likelihood coefficient estimates but does not use background information.
- Penalisation via data augmentation: a method for computing penalised estimates in which the data equivalent to the penalty is constructed and added to the actual study data, and then conventional analysis is done on the augmented data.

control confounding.[26] Unfortunately, serious complications can arise from propensity scoring in case-control studies,[30] and even cohort studies can have severe problems with propensity score matching.[31] We thus recommend matching directly on strong risk factors to reduce reliance on modelling to remove confounding by such factors. Nonetheless, after matching there might still be small numbers of cases or non-cases at different exposure or treatment levels, leading to sparse data bias. Fortunately, penalised regression can be applied to matched samples both to reduce sparse data bias while achieving finer confounding adjustments, and to study variation in effect measures (interactions) across subgroups.

It is easy to spot extreme real examples of sparse data bias, but the literature suggests that less dramatic examples are common and unnoticed. For example, some studies report subgroup effect estimates that, although not implausible, nonetheless increase in size for smaller subgroups which are then singled out as being of possible special risk, even though the increase is as easily explained by increased bias or random error due to the reduced numbers in those groups.

We thus strongly recommend that basic data numbers within treatment or exposure and outcome categories be examined and presented, and that adjustment methods such as penalisation be applied whenever the numbers of events per covariate fall below four or five.

The weighting (degree of penalisation) for each variable is best determined so that the implied prior interval encompasses the full range of reasonable possibilities for the effect of the variable. Table 3 facilitates conversion of this range to a weight, but determining the range will require contextual (subject matter) input about what is reasonable. We believe this input is a strength of the approach. If there is any ambiguity or doubt about the range to choose, we advise one to err on the side of wider prior intervals and thus weaker penalties, because such errors of caution will not reduce the coverage rates of the interval estimates.

As a final, technical point, we note that what we have termed "sparse data bias" in logistic regression is closely related to a problem sometimes called non-collapsibility of odds ratios. For reviews of this problem and its relation to confounding (with which it is often confused) and confounder adjustments, see references [32] and [33].

Contributors: All authors were involved in the concept and design of the study. SG wrote the first draft of the manuscript, and all authors contributed to subsequent revisions of the manuscript and approved the final version. SG is the guarantor.

Funding: None declared.

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

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

1 Greenland S, Schwartzbaum JA, Finkle WD. Problems due to small samples and sparse data in conditional logistic regression analysis. *Am J Epidemiol* 2000;151:531-9. doi:10.1093/oxfordjournals.aje.a010240.

2 AbdelSalam H, Restrepo C, Tarity TD, Sangster W, Parvizi J. Predictors of intensive care unit admission after total joint arthroplasty. *J Arthroplasty* 2012;27:720-5. doi:10.1016/j.arth.2011.09.027.

3 Kernan WN, Viscoli CM, Brass LM, et al. Phenylpropanolamine and the risk of hemorrhagic stroke. *N Engl J Med* 2000;343:1826-32. doi:10.1056/NEJM200012213432501.

4 Harrell FE Jr, Lee KL, Califf RM, Pryor DB, Rosati RA. Regression modelling strategies for improved prognostic prediction. *Stat Med* 1984;3:143-52. doi:10.1002/sim.4780030207.

5 Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. *J Clin Epidemiol* 1996;49:1373-9. doi:10.1016/S0895-4356(96)00236-3.

6 Vittinghoff E, McCulloch CE. Relaxing the rule of ten events per variable in logistic and Cox regression. *Am J Epidemiol* 2007;165:710-8. doi:10.1093/aje/kwk052.

7 Clayton D, Hills M. *Statistical models in epidemiology.* Oxford University Press, 1993.

8 Jewell NP. On the bias of commonly used measures of association for 2 x 2 tables. *Biometrics* 1986;42:351-8. doi:10.2307/2531055.

9 Glymour MM, Greenland S. Causal diagrams. In: Rothman KJ, Greenland S, Lash TL, eds. *Modern Epidemiology.* 3rd ed. Lippincott Williams & Wilkins, 2008: 183-209.

10 Greenland S, Neutra R. Control of confounding in the assessment of medical technology. *Int J Epidemiol* 1980;9:361-7. doi:10.1093/ije/9.4.361.

11 Greenland S. Comment: cautions in the use of preliminary test estimators. *Stat Med* 1989;8:669-73. doi:10.1002/sim.4780080606.

12 Greenland S, Pearce N. Statistical foundations for model-based adjustments. *Annu Rev Public Health* 2015;36:89-108. doi:10.1146/annurev-publhealth-031914-122559.

13 Hurvich CM, Tsai CL. The impact of model selection on inference in linear regression. *Am Stat* 1990;44:214-7.

14 Mickey RM, Greenland S. The impact of confounder selection criteria on effect estimation. *Am J Epidemiol* 1989;129:125-37.

15 Maldonado G, Greenland S. Simulation study of confounder-selection strategies. *Am J Epidemiol* 1993;138:923-36.

16 Hirji K. *Exact analysis of discrete data.* CRC Press/Chapman and Hall, 2006.

17 Agresti AA. *Categorical data analysis.* 3rd ed. Wiley, 2013.

18 Firth D. Bias reduction of maximum likelihood estimates [correction in: *Biometrika* 1995;82:667]. *Biometrika* 1993;80:27-38.

19 Sullivan SG, Greenland S. Bayesian regression in SAS software. *Int J Epidemiol* 2013;42:308-17. doi:10.1093/ije/dys213.

20 Discacciati A, Orsini N, Greenland S. Approximate Bayesian logistic regression via penalized likelihood by data augmentation. *Stata J* 2015;15:712-36.

21 Greenland S. Simpson's paradox from adding constants in contingency tables as an example of Bayesian noncollapsibility. *Am Stat* 2010;64:340-4. doi:10.1198/tast.2010.10006.

22 Lyles RH, Guo Y, Greenland S. Reducing bias and mean squared error associated with regression-based odds ratio estimators. *J Stat Plan Inference* 2012;142:3235-41. doi:10.1016/j.jspi.2012.05.005.

23 Greenland S, Mansournia MA. Penalization, bias reduction, and default priors in logistic and related categorical and survival regressions. *Stat Med* 2015;34:3133-43. doi:10.1002/sim.6537.

24 Greenland S. Prior data for non-normal priors. *Stat Med* 2007;26:3578-90. doi:10.1002/sim.2788.

25 Greenland S. Appendix to Bayesian perspectives for epidemiologic research. III. Bias analysis via missing-data methods [correction in: *Int J Epidemiol* 2010;39;1116]. *Int J Epidemiol* 2009;38:1662-73.

26 Rubin DB. *Matched sampling for causal effects.* Cambridge University Press, 2006. doi:10.1017/CBO9780511810725.

27 Harrell F. *Regression modeling strategies.* Springer, 2001. doi:10.1007/978-1-4757-3462-1.

28 Steyerberg EW. *Clinical prediction models: a practical approach to development, validation, and updating.* Springer, 2008.

29 Pavlou M, Ambler G, Seaman SR, et al. How to develop a more accurate risk prediction model when there are few events. *BMJ* 2015;351:h3868. doi:101136/bmj.h3868.

30 Månsson R, Joffe MM, Sun W, Hennessy S. On the estimation and use of propensity scores in case-control and case-cohort studies. *Am J Epidemiol* 2007;166:332-9. doi:10.1093/aje/kwm069.

31 King G, Nielsen R. Why propensity scores should not be used for matching. Vers. 2 Feb. 2016. http://gking.harvard.edu/publications/why-Propensity-Scores-Should-Not-Be-Used-Formatching

32 Greenland S, Robins JM, Pearl J. Confounding and collapsibility in causal inference. *Stat Sci* 1999;14:29-46. doi:10.1214/ss/1009211805.

33 Greenland S, Pearl J. Adjustments and their consequences - collapsibility analysis using graphical models. *Int Stat Rev* 2011;79:401-26. doi:10.1111/j.1751-5823.2011.00158.x.

Web appendix: Penalised logistic regression via data augmentation in R