Zhang *et al.*
BMC Medical Research Methodology (2024) 24:287
https://doi.org/10.1186/s12874-024-02412-1

BMC Medical Research Methodology

RESEARCH Open Access

Three new methodologies for calculating the effective sample size when performing population adjustment

Landan Zhang[1], Sylwia Bujkiewicz[2] and Dan Jackson[3*]

Abstract

Background The concept of the population is of fundamental importance in epidemiology and statistics. In some instances, it is not possible to sample directly from the population of interest. Weighting is an established statistical approach for making inferences when the sample is not representative of this population.

Methods The effective sample size (ESS) is a descriptive statistic that can be used to accompany this type of weighted statistical analysis. The ESS is an estimate of the sample size required by an unweighted sample that achieves the same level of precision as the weighted sample. The ESS therefore reflects the amount of information retained after weighting the data and is an intuitively appealing quantity to interpret, for example by those with little or no statistical training.

Results The conventional formula for calculating ESS is derived under strong assumptions, for example that outcome data are homoscedastic. This is not always true in practice, for example for survival data. We propose three new approaches to compute the ESS, that are valid for any type of data and weighted statistical analysis, and so can be applied more generally.

Conclusion We illustrate all methods using an example and conclude that our proposals should accompany, and potentially replace, the existing approach for computing the ESS.

Keywords Weighted statistical analysis, Propensity score, Inverse probability weighting, Survey weights, Indirect treatment comparisons

Background

The concept of the population is of fundamental importance in epidemiology and statistics. For example, it is the ‘P’ in PICOS statements [1, 2] and is an essential attribute of an estimand [3]. When making statistical inferences we usually assume that the sample is representative of the population of interest. However, this is not always the case, and if the sample is clearly unrepresentative and statistical analysis is to proceed, methods for population adjustment can help alleviate this issue. Statistical methods for population adjustment broadly fall under two main categories: regression-based methods [4–7] and weighting-based approaches [6–10]. We focus on the latter, where subjects in our sample are weighted to provide the population of interest. Both types of statistical methods require assumptions; for example, when using weighting, we require that the covariates’ sample space for the population of interest is contained in our sample. This is so that, despite the fact that our sample is unrepresentative, we are able to allocate subject weights

*Correspondence:
Dan Jackson
daniel.jackson1@astrazeneca.com
[1] Medical Affairs Statistics, Bayer plc, 400 S Oak Way, Reading RG2 6AD, UK
[2] Biostatistics Research Group, Department of Population Health Sciences, University of Leicester, University Rd, Leicester LE1 7RH, UK
[3] Statistical Innovation, AstraZeneca, 136 Hills Rd, Cambridge CB2 8PA, UK

to reproduce (or more pragmatically approximate) the population of interest. For an especially simple and articial example, suppose that a population of interest is comprised of equal numbers of biomarker positive and negative patients. Further, suppose that biomarker positive patients are more likely to be present in the sample, where we sample twice as many patients of this type. In order to obtain a sample that is representative of the whole population, we can weight biomarker negative patients by twice as much as biomarker positive patients.

Statistical methods for population adjustment using weighting methods raise two main concerns: 1) additional and potentially strong assumptions will in general be needed to justify them; 2) they may incur a loss of information. The first concern is usually more obvious, and often the most pressing. This is because the assumptions required to perform the adjustment may be the most difficult to defend in practice. More specifically, the statistical modelling usually required to compute the weights will often require strong assumptions, for example that appropriate covariate effects are included. We however focus on the second, more subtle, issue. Therefore, we will assume that the population adjustment method is acceptable, but there are concerns about the precision of the resulting statistical analysis. This is because it would likely have been more efficient to sample directly from the population of interest, rather than from another population and perform the necessary adjustment. If this loss of efficiency is severe then this can result in imprecise statistical inferences, for example wide confidence intervals and hypothesis tests with low power.

The weighting approach to population adjustment is established in a wide variety of contexts. For example, weighting is used when the proportions of survey respondents in some subject groups are different to the proportions in the survey design [11–13]. Here we weight the survey results in one or more subject groups so that they represent the expected numbers of respondents in each group and so the population of interest. Propensity score weighting is another commonly used weighting approach in statistical analysis, which allows for reduction of the bias in estimates due to confounding and performs estimation in particular populations [14–16]. Inverse probability weighting [17, 18] is a commonly used approach to deal with missing data, where weighting is used to reproduce the general population who may, or may not, provide outcome data. This approach weights the complete cases using the reciprocal of the estimated probability of providing data given the covariates, so that they are representative of the population. Inverse probability of censoring weighting (IPCW) is a closely related technique developed to eliminate bias arising from dependent censoring [19–21]. Here weighting is applied so that the observed survival data are representative of the population, despite the fact that some subjects are censored in this way. IPCW has also been successfully applied in treatment switching in clinical trials [22, 23], where patients in the control group can switch to the active treatment group at some point during the follow-up period (e.g. when the disease progresses). These control group patients are censored at the time of switching, and control group patient data are weighted by the inverse of their probabilities of not switching, again so that the resulting censored data are representative of the population.

The weighted data result in a weighted likelihood based analysis where correct standard errors can be obtained using robust sandwich standard errors or bootstrapping. It is much less obvious how to obtain correct posterior distributions in a Bayesian analysis when using weighted samples, so that weighting based approaches are most amenable to frequentist analyses. Uncertainty in the weights can be taken into account using bootstrapping, where subjects' weights are estimated within each bootstrap replication. Alternatively the uncertainty in the weights could be ignored, so that they are treated as fixed constants, which is a further approximation that might be made to simplify analysis.

The effective sample size (ESS) has been proposed as a descriptive statistic that gives an indication of the number of subjects contributing to the analysis after using weighting to perform population adjustment [8, 24, 25]. The ESS compares the variances of weighted (population-adjusted) and unweighted (unadjusted) estimates. More specifically, the ESS is computed as the size of a smaller, hypothetical, unweighted sample that produces the same level of precision for the sample mean as the weighted sample [6, 10]. Smaller values of ESS occur when there is more variability in the weights used for the population adjustment, reflecting a greater population adjustment [7]. The effective sample size is easily computed, and so has the merits of simplicity and transparency. However, it makes assumptions that are likely to be violated in practice, and in particular, it assumes the outcome data are homoscedastic. This assumption will be clearly violated in some contexts, such as for survival data, and will rarely, if ever, be exactly true.

In this paper, we propose three new methodologies for ESS calculation when using weighting approaches for population adjustment. The first new method compares the variances of weighted (population adjusted) and unweighted (unadjusted) estimates in the same way as in the conventional calculation, but where the variance of both the weighted and unweighted estimates are computed in a valid way irrespective of the type of data and statistical model. This method is almost as transparent as the conventional approach, resolves concerns when the conventional approach makes incorrect assumptions, and could be used as a quick check that the usual ESS calculation is reasonable. The conceptual difficulty with our first new method is that its derivation involves no direct argument relating to the size of a smaller hypothetical unweighted sample, which might reasonably be considered to be intrinsic to any definition of the ESS.

The second new method uses re-sampling to calculate the variance of estimates from an unweighted sample, where we sequentially reduce the sample size until it is small enough to provide a greater variance for the estimate than the population-adjusted analysis. We can use linear interpolation between the final two sample sizes to obtain the required ESS. This method is computationally intensive and subject to Monte Carlo error, but its advantage is that it has the sample size of a smaller unweighted sample at its conceptual basis, and so avoids the indirect nature of the first new method described above. The third new method can be used when a closed variance formula for the estimate from the unweighted analysis exists, where this formula depends directly upon easily computed counts or event rates and possibly other parameters that may be approximated with their estimates. Then by applying a scale factor to these counts or rates, we can calculate the size of the sample that would be needed to give the same level of precision (variance of estimates) as the weighted analysis, in a similar way to the second new method.

The rest of this paper is structured as follows. In Methods section, we explain the conventional ESS formula and derive our three new methods to calculate ESS. In Results section, we illustrate the use of all four methods in a numerical example and present the results. Our example involves a weighting based analysis for population adjustment, namely Matching-Adjusted Indirect Comparison (MAIC) [6–9]. We summarize with a Discussion section.

Methods

In this section, we describe all four methods to calculate ESS, including the conventional method and three new methods. We assume throughout that we have a sample of size n, where population adjustment weights, $\hat{w}_j$, $j = 1, 2, \cdots, n$, have been calculated. We also assume that the method for computing the weights $\hat{w}_j$ is satisfactory so that the weighted sample is representative of the population of interest. Our main concern is that the population-adjusted statistical analysis, where subjects' data are weighted by their $\hat{w}_j$, incurs a loss of information relative to an unadjusted analysis. We seek to quantify the amount of information that is retained after weighting using the ESS. In general, the weights will be estimated from a statistical model, and we use the notation $\hat{w}_j$ to emphasise this.

In some applications of weighted analyses we may have different subjects present, and weights $\hat{w}_j$ calculated at different time points, for example, when using IPCW. This complicates matters and the conventional ESS is, in any case, hard to apply unless each subject has a fixed population adjustment weight. We leave the extension of our methods to more complicated settings such as these as further work, and we return to this issue in the discussion.

Existing approach: The conventional ESS formula

The conventional formula to calculate ESS, which has been recommended to use after performing population adjustment [8, 24, 25], is derived by comparing the variance of the weighted sample mean and the unweighted sample mean. The resulting ESS is calculated as

$$ESS = \frac{(\sum_{j=1}^{n} \hat{w}_j)^2}{\sum_{j=1}^{n} \hat{w}_j^2}. \quad (1)$$

We now explain how Eq. (1) is derived to make its assumptions explicit. We assume that the outcome data are homoscedastic, so that $Var(Y_j) = \sigma^2$ for $j = 1, \ldots, n$, and that the estimand of interest is the population mean. For a weighted sample, the variance of the corresponding estimate, the weighted sample mean, is calculated as

$$Var(\bar{Y}_w) = Var\left(\frac{\sum_{j=1}^{n} \hat{w}_j Y_j}{\sum_{j=1}^{n} \hat{w}_j}\right) = \left(\frac{1}{\sum_{j=1}^{n} \hat{w}_j}\right)^2 \sum_{j=1}^{n} \hat{w}_j^2 \sigma^2. \quad (2)$$

where in the derivation of (2), we have treated the $\hat{w}_j$ as fixed constants, and so we have ignored any uncertainty (and association with the outcome data) in them. For an unweighted sample, the variance of the sample mean is calculated as

$$Var(\bar{Y}_u) = Var\left(\frac{\sum_{j=1}^{n} Y_j}{n}\right) = \frac{\sigma^2}{n}. \quad (3)$$

By equating Eq. (2) with Eq. (3), we solve for $n = ESS$ to calculate the required sample size of a hypothetical unweighted sample (i.e. the ESS) to achieve the same level of precision as the weighted sample (so that $Var(\bar{Y}_w) = Var(\bar{Y}_u)$. This almost immediately results in (1).

This derivation of the conventional ESS descriptive statistic (1) clarifies the main assumptions required when presenting it: it assumes independent and homoscedastic outcome data, that the estimand of interest is the (unadjusted) population mean, and it treats the $\hat{w}_j$ as fixed constants. This motivates the development of alternative methods for computing the ESS below that will be valid for other data types and statistical models. When using Eq. (1), we will always obtain $ESS < n$, unless $\hat{w}_j = c$ for all i, where c is the constant. In this case no population adjustment is required and $ESS = n$. Furthermore, the same ESS is obtained for all statistical analyses, for example, for different outcomes that use the same set of weights. These properties may be appealing due to their simplicity but may also be misleading. We return to this issue in the next section.

First new method: comparing the variances of adjusted and unadjusted estimates

From Eqs. (1), (2) and (3), we can see that an equivalent definition of the ESS in (1) is

$$ESS = \frac{n \times Var(\bar{Y}_u)}{Var(\bar{Y}_w)}, \tag{4}$$

We propose generalising Eq. (4) by replacing $\bar{Y}_u$ with $\hat{\theta}_u$, the unweighted (i.e. the unadjusted) estimate of the estimand of interest, and $\bar{Y}_w$ with $\hat{\theta}_w$, the corresponding weighted (i.e. the population adjusted) estimate. Hence, more generally, Eq. (4) becomes

$$ESS = \frac{n \times Var(\hat{\theta}_u)}{Var(\hat{\theta}_w)}, \tag{5}$$

where, unlike Eqs. 4 and 5 is suitable for any type of outcome data, estimand, and statistical model. This is because, having performed valid unadjusted and adjusted analyses and so computed variances of the two estimates in (5), this more general definition of ESS is immediately applicable. It is also very easily computed.

For example, the estimates $\hat{\theta}_u$ and $\hat{\theta}_w$, and their variances could be computed using appropriate survival models for time-to-event data, logistic regression for binary outcomes or even using novel or very sophisticated statistical methods. In general, the estimate $\hat{\theta}_w$ is obtained in the same way as $\hat{\theta}_u$, but where subjects are weighted by $\hat{w}_j$. This weighting will usually be easily implemented because typically, both estimates will be from regression models, for which standard implementations allow weights to be specified. The only potential difficulty is that $Var(\hat{\theta}_w)$ must be computed in an appropriate way that respects the fact that the $\hat{w}_j$ are not 'replication' or 'case' weights; for example, a weight of three does not mean that we have three subjects with the same data, rather one subject has received this weight in the population adjustment. Where available, sandwich/robust standard errors will be required to compute $Var(\hat{\theta}_w)$ [6, 10], or bootstrapping can be used [26].

Different models and outcomes will provide different values of $Var(\hat{\theta}_u)$ and $Var(\hat{\theta}_w)$ when using the same dataset and method for population adjustment. Hence one consequence of using Eq. (5) is that different values of ESS will then be obtained. Furthermore, Eq. (5) provides no guarantee that $ESS \leq n$. Our position is that these consequences of using (5) are entirely appropriate because the same set of weights may have dissimilar consequences for estimation precision in different statistical analyses of the same data. Furthermore, subjects with outlying or influential outcomes may receive little weight in population-adjusted analyses, where these subjects would otherwise have detrimental consequences for precision of estimates from a statistical model. In that case, this adjustment may in fact increase precision. However, we suspect that this will rarely occur in practice, and if it happens, then a special investigation would be needed to provide an explanation. Outliers can create challenges for all types of statistical analysis. In particular they can present issues for analyses that use weighting-based methods for population adjustment, because outliers may receive unusually large weights.

Despite all its advantages, there is a conceptual concern when using Eq. (5). This is because its derivation made no direct appeal to the size of a hypothetical unweighted sample. The conventional formula (1) is however directly based on the consideration of such a sample. The derivation of Eq. (5) instead took advantage of an alternative interpretation of the conventional ESS, involving the variance of estimates from weighted and unweighted analyses, and generalised this to other settings. This approach was adopted because this interpretation readily generalises. However, since Eq. (5) does not require the notion of a hypothetical smaller unweighted sample, some may prefer to interpret it as providing a type 'pseudo' ESS, that is equivalent to the conventional ESS under the strong assumptions it requires, that measures something more abstract in other settings.

Second new method: re-sampling with reduced sample size

As explained in First new method: comparing the variances of adjusted and unadjusted estimates section, our first new method is easily computed but makes no direct appeal to the size of a hypothetical unweighted sample. Interpreting the resulting quantity as a measure of ESS is therefore potentially problematic. In this section, we propose a computationally intensive alternative approach based on a smaller sample that overcomes this potential concern. In this approach, we re-sample the data multiple times to provide robust variance calculations. Specifically, we follow the procedure below:

1. We start by re-sampling multiple datasets from the entire sample size. This step simply produces bootstrap samples. Subjects can be randomly re-sampled within each treatment group to maintain the original randomisation ratio. We then perform the unweighted analysis for each re-sampled datasets, producing bootstrap replications of unweighted estimates of interest $\hat{\theta}_{1,u}$, $\hat{\theta}_{2,u}$, ..., $\hat{\theta}_{B,u}$, where B is the number of bootstrap replications. The sample variance of the boostrap replications gives an estimate of $Var(\hat{\theta}_u)$. Step 1 is simply the use of bootstrapping to estimate the variance of $\hat{\theta}_u$.
2. In most cases where $Var(\hat{\theta}_u) < Var(\hat{\theta}_w)$, we gradually reduce the sample size of the bootstrap samples created in the step 1. We implement this by removing a random selection of subjects from each treatment arm of the re-sampled datasets. Because the bootstrap samples are randomly re-sampled prior to undertaking this procedure, more simply, we may deterministically remove subjects at the start (or end) of the bootstrap replications at each iteration.

   (a) Remove k observations in each arm in each bootstrap sample. For example, if there are two treatment groups and $k = 5$, then we reduce the bootstrap sample size by 10.
   (b) Perform unweighted analysis using the bootstrap sample with reduced sample size and compute the resulting bootstrap replications $\hat{\theta}_{1,-k,u}$, $\hat{\theta}_{2,-k,u}$, ..., $\hat{\theta}_{B,-k,u}$.
   (c) The sample variance of the bootstrap replications $\hat{\theta}_{1,-k,u}, \hat{\theta}_{2,-k,u}, \ldots, \hat{\theta}_{B,-k,u}$ gives an estimate of $Var(\hat{\theta}_{-k,u})$, where $\hat{\theta}_{-k,u}$ is the unweighted estimate of interest with k observations removed from each arm in the sample.

   Together, the steps (a) to (c) are simply standard bootstrapping to compute $Var(\hat{\theta}_{-k,u})$, where before performing the unweighted analysis, we reduce the sample size of the re-sampled datasets. When implementing this approach, we must select an appropriate amount to reduce the sample size at each iteration. In general, the size of the incremental decrease in sample size should have a small but noticeable impact on precision. For example, decreasing the sample size by 1-2% of the original sample size at each iteration is likely to be appropriate. In this method, the observations are randomly removed in each treatment group and the re-sampled dataset preserves the overall distribution and characteristics of the original dataset. Therefore, the sampling approach we proposed is a form of sub-sampling, stratified by treatment group.

3. We repeat steps (a) to (c) but with additional k observations being removed from each treatment arm of each boostrap sample. We recommend using the same bootstrap samples as in steps 1 and 2, and ensuring that observations removed from the bootstrap samples at previous iterations are subsets of those removed at subsequent iterations, to help ensure the resulting variances are monotonic in the sample size. This iterative process continues until $Var(\hat{\theta}_{-l,u}) > Var(\hat{\theta}_w)$, where l observations have been removed from each arm. The value of $Var(\hat{\theta}_w)$ is then located between the last two values of variance obtained.
4. Finally, we use linear interpolation between the two corresponding sample sizes and their variances to calculate the required sample size for a hypothetical unweighted sample that gives the same level of precision as the weighted sample. This sample size is interpreted as the ESS.

In the unlikely event that this variance of unweighted entire sample $Var(\hat{\theta}_u)$ is greater than $Var(\hat{\theta}_w)$, we could conclude that the population adjustment results in no concerns about loss of precision if no further description is considered necessary or a gain in precision is considered implausible; otherwise, we could use an approach similar to the one above. In the latter approach we sequentially increase the size of the unweighted re-sampled datasets in the re-sampling procedure, and in a similar approach, find the size of sample size needed to provide an unweighted variance that is equal to $Var(\hat{\theta}_w)$, and report this sample size as the ESS.

The advantages of this method are that it directly appeals to the size of a hypothetical unweighted sample and so results in a quantity that can be unequivocally interpreted as the ESS, and it is widely applicable. Disadvantages include its computational complexity and sensitivity to both Monte Carlo error and the reduction in sample size used. In practice, it may be desirable to use different increment sizes and random seeds to assess these potential sensitivities.

There are different ways of implementing this type of approach. A simple version which avoids the computational complexity of re-sampling is to randomly remove subjects from the original sample, until the variance of unweighted sample with reduced sample size is greater than the variance of the weighted sample $Var(\hat{\theta}_w)$. Then in a similar way, use the interpolation to calculate the required sample size for a hypothetical unweighted sample that produces the value of variance as $Var(\hat{\theta}_w)$. However, we recommend repeating procedures multiple times, and in particular using the proposed re-sampling approach, to produce a more robust estimate of the variance.

Third new method: Scaling the unweighted variance formula with reduced sample size

This section describes our final approach, which has a mechanism similar to the previous one. In this section, we assume that a simple formula for $Var(\hat{\theta}_u)$ is available that depends upon sample counts and rates and possibly other parameters that may be approximated with their estimates. Such a formula will often be available, for example, when the type of outcome data and the statistical model used when computing $\hat{\theta}_u$ are of a simple, standard form.

Then, instead of using re-sampling as in the previous section, we apply a scaling factor p to the counts and rates in the variance formula. We interpret p as the percentage of the original sample size retained in a hypothetical smaller sample. We then compute the value of the scaling factor that is needed to obtain $Var(\hat{\theta}_u) = Var(\hat{\theta}_w)$, where $Var(\hat{\theta}_u)$ is obtained after applying the scaling factor and $Var(\hat{\theta}_w)$ is computed using the whole sample, and call this value $\hat{p}$. In situations where the variance formula is monotonic in the sizes of the counts and rates, as will commonly be the case, the necessary scaling factor $\hat{p}$ will be unique, and we assume this is the case. We then define the $ESS = n\hat{p}$, where $n\hat{p}$ is interpreted as the sample size of a hypothetical smaller sample that provides the same variance as the weighted sample.

For example, suppose that the outcome is binary, and the estimand is the marginal log odds ratio comparing two treatment groups A and B. Suppose that the outcome for the jth patient in the first treatment group is $Y_{A,j}$, where $Y_{A,j} = 1$ if the event occurs and $Y_{A,j} = 0$ if this does not occur. Similarly, assume that the outcome for the jth patient in the second treatment group is $Y_{B,j}$. Also, assume that the total number of patients in each treatment group is n_A and n_B, respectively. Then the standard variance formula for the estimated log odds ratio is available as the sum of the reciprocals of the entries in the resulting two-by-two table and is

$$Var(\hat{\theta}_u) = \frac{1}{\sum Y_{A,j}} + \frac{1}{n_A - \sum Y_{A,j}} + \frac{1}{\sum Y_{B,j}} + \frac{1}{n_B - \sum Y_{B,j}}. \quad (6)$$

where $\hat{\theta}_u$ is now an estimated, unweighted log odds ratio. This variance formula depends upon counts from the data, and is monotonic in these counts, as required. Upon applying the scaling factor p to (6), we obtain

$$Var(\hat{\theta}_u) = \frac{1}{p\sum Y_{A,j}} + \frac{1}{p(n_A - \sum Y_{A,j})} + \frac{1}{p\sum Y_{B,j}} + \frac{1}{p(n_B - \sum Y_{B,j})}. \quad (7)$$

We can then use root finding numerical methods to solve (7) for p so that $Var(\hat{\theta}_u) = Var(\hat{\theta}_w)$, and define $ESS = n\hat{p}$.

As another example, the variance of an estimated log hazard from a Cox model may be approximated by four divided by the total number of events [27]. This is another formula for $Var(\hat{\theta}_u)$ that is amenable to this method, where we can scale the number of events by p and find the value that equates $Var(\hat{\theta}_u)$ to $Var(\hat{\theta}_w)$. Although this variance formula is just an approximation, it will likely be acceptable in applications.

The advantages of this method, compared to the previous one that it is conceptually very similar to, are that it is not sensitive to Monte Carlo error or the increments used. Furthermore, it is not so computationally intensive. However its use relies on there being an appropriate formula for $Var(\hat{\theta}_u)$ to use. In situations where such a formula is available, we suggest that our third method is likely to be considered preferable to the second, but we retain our second method because it is widely applicable.

Results

We compute all four measures of ESS methods using the numerical example from the NICE DSU Technical Support Document 18: Methods for population-adjusted indirect comparisons in submissions to NICE [6]. This simulated example contains two data sets, each corresponding to a randomised controlled trial. The first dataset is from the 'company's trial', with treatments A and B, and the second is from the 'competitor's trial', with treatments A and C. The overall aim is to compare treatments B and C, where an indirect comparison is necessary because these two treatments are not compared in the same trial. We have individual patient data available from the company's trial, but only aggregate-level information is available from the competitor's trial, and the populations of the two trials differ in an important way. Hence a population-adjusted indirect treatment comparison is required to make this more equitable. 500 patients are enrolled in the AB trial and 300 in the AC trial, and both trials use 1:1 randomisation.

This example involves two covariates: age and sex. Patients have ages taking integer values, uniformly distributed between 45 to 75 in the company's trial and 45 to 55 in the competitor's trial. The proportion of females is 0.64 in the company's trial and 0.8 in the competitor's trial. Binary outcome data were simulated using the logistic model

$$logit(p_{it}) = 0.85 + 0.12male_{it} + 0.05(age_{it} - 40) + (\beta_t - 0.08(age_{it} - 40))I(t \neq A). \quad (8)$$

where $\beta_B = -2.1$ and $\beta_C = -2.5$ are the conditional (on age and sex) treatment effect for a 40-year-old patient. From Eq. (8), we can see that age is an effect modifier, because the term $(\beta_t - 0.08(age_{it} - 40))I(t \neq A)$ describes how the treatment effects of B and C, relative to treatment A, depend on age. Furthermore, the patients' ages substantively differ across the two trials, so a standard, unadjusted, indirect treatment comparison using Bucher's method [28] would not be equitable.

Some form of population adjustment is therefore required to fairly indirectly compare treatments B and C, where an additional difficulty is that only aggregate-level data are available from the competitor's trial. MAIC [6–9] is an established method for performing population adjustment in this situation, where weights are calculated for the company's trial to match its population to that of the competitor trial. A weighted analysis can then be performed using the MAIC weights to estimate the effect of treatment B, relative to A, in the competitor trial's population. This population-adjusted estimate can then be used in a standard indirect treatment comparison [28], with the estimate of treatment effect reported by the competitor trial, to make inferences in the competitor trial's population.

This MAIC has been described several times previously [6, 10] and so we only give brief details here. It is said to be anchored [10, 29], because both trials include a common comparator (treatment A). We follow the recommendation to match only on effect modifiers [6], and so on age, where we also match on its squared value so that its variance is also matched on [6, 10]. Briefly, this MAIC matches the means of both age and age squared, whilst requiring that the weights are calculated as the exponent of a linear predictor that includes both these covariates.

We show the histogram of the resulting 500 weights $\hat{w}_j$ which has been scaled to sum to 1 in Fig. 1, where we have one weight for each patient in the company's trial. There is considerable variation in the weights but there are no obvious outliers. The weights shown in Fig. 1 were calculated using the conventional methods described in Signorovitch et al. [8, 9] and Jackson et al. [10], to balance the mean covariate values across the two populations. The the covariates of age and sex differ notably across the two populations, but these population differences are not enormous, and the effects of both covariates in this matching can both be described as moderate. A population-adjusted estimate of the log odds ratio, $\hat{\theta}_w = -3.2151$, comparing treatments B and A, is obtained using a logistic regression of the binary outcome data on treatment group using the company's trial data, where weights are specified as the MAIC weights. A valid sandwich standard error, whose squared value is $Var(\hat{\theta}_w)$, was computed giving $Var(\hat{\theta}_w) = 0.1628$.

The unadjusted estimated log odds ratio, $\hat{\theta}_u = -3.5717$, is similarly obtained using an (unweighted) logistic regression on the treatment group, where $Var(\hat{\theta}_u) = 0.0653$ is reported by standard software. The

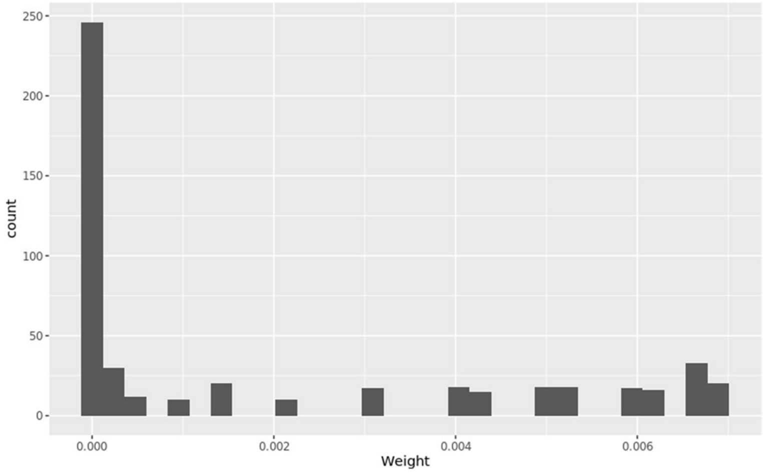

Fig. 1 Histogram of resulting 500 weights for the patients enrolled in the AB trial, using MAIC for population adjustment. The weights shown have been normalised to sum to 1

variance of the population-adjusted estimate $Var(\hat{\theta}_w)$ is notably greater than this, indicating that substantial information loss has been incurred when performing the population adjustment for the company's trial data. Our intention is now to use our ESS statistics to help us conceptualise how severe this information loss is.

Application of the existing approach

Having computed the 500 MAIC weights $\hat{w}_j$ for each patient in the company's trial (Fig. 1), it is straightforward to compute the conventional effective sample size using the formula (1). This calculation provides ESS=185.6451, indicating that after the population adjustment, the weighted sample is 'worth' around 186 patients.

Although this conventional ESS is easily computed, a number of its assumptions are clearly false. Firstly, the outcome data are not homoscedastic: from Eq. (8), we can see that different patients have different probabilities of an event, so that their binary outcome data do not have a common variance. Furthermore, the estimand is not the sample mean, rather it is a log odds ratio, estimated using a logistic regression. The conventional ESS may therefore be misleading.

Application of approach 2: Comparing the variances of estimates

Having computed $Var(\hat{\theta}_u) = 0.0653$ and $Var(\hat{\theta}_w) = 0.1628$, and noting that the company's trial includes n=500 patients, Eq. (5) is easily computed as $ESS = \frac{n \times Var(\hat{\theta}_u)}{Var(\hat{\theta}_w)} = 200.5176$. This indicates that, after the population adjustment, the weighted sample is 'worth' around 201 patients.

Application of approach 3: Re-sampling with reduced sample size

We note that the robust standard error resulted in $Var(\hat{\theta}_w) = 0.1628$, whereas the unadjusted analysis gave $Var(\hat{\theta}_u) = 0.0653$ so that the adjusted analysis incurs a loss of precision. However, this variance is calculated by the standard variance formula for binary outcomes reported in the logistic regression model in software. The first step of performing this approach is to re-estimate the $Var(\hat{\theta}_u)$ using bootstrapping as described in Step 1 of Second new method: re-sampling with reduced sample size section. To ensure numerical accuracy, we use $B = 500$ bootstrap samples. We then sequentially reduce the sample size until an unweighted analysis results in less precision (larger variance) than the weighted analysis, and we use interpolation to caclulate the required sample size as explained in Step 2 to 4.

There are 500 patients in the company's trial, so with 1:1 randomisation, there are 250 patients in each treatment group. In each step, we randomly removed multiples of five observations from both groups of the company's trial until we obtained a variance larger than the variance of estimates in the weighted sample. Hence we reduce the sample size by 2% at each iteration. We tabulate the variances for the unweighted sample with sequentially reduced sample sizes in Table 1. Furthermore, we use Fig. 2 to illustrate the observed trends between sample size and variance. As explained in Second new method: re-sampling with reduced sample size section, we estimate $Var(\hat{\theta}_u)$ using bootstrapping, giving $Var(\hat{\theta}_u) = 0.0769$. This produces a slightly different value compared to the variance $Var(\hat{\theta}_u) = 0.0653$ obtained from the initial unadjusted analysis.

From Table 1 and Fig. 2, we can conclude that the variance generally increases as the sample size decreases, as expected. The variance from the weighted analysis $Var(\hat{\theta}_w) = 0.1628$, which is between the unweighted variances of 0.1572 (with a sample size of 240) and the value of 0.1652 (with a sample size of 230). Linear interpolation is used to calculate the required sample size of the unweighted sample, which produces a variance of 0.1628, as $ESS = 233.005$. This indicates that, after the population adjustment, the weighted sample is 'worth' around 233 patients. It is worth noting that the variance is not completely monotonically decreasing as the sample size decreases. This is because removing observations may change the event rate, and the variance also depends on

Table 1 This table presents the variances of unadjusted estimates calculated using re-sampling with reduced sample size, for the numerical example

Sample size	500	490	480	470	460	450	440	430	420	410
Variance	0.0769	0.0797	0.0812	0.0828	0.0842	0.0850	0.0864	0.0897	0.0910	0.0899
Sample size	400	390	380	370	360	350	340	330	320	310
Variance	0.0916	0.0950	0.0982	0.1010	0.1021	0.1010	0.1031	0.1065	0.1078	0.1115
Sample size	300	290	280	270	260	250	240	230		
Variance	0.1188	0.1230	0.1259	0.1311	0.1410	0.1476	0.1572	0.1652		

The sample size is reduced, in steps of 10, from the initial sample size of 500. The initial value of variance with the entire sample size of 500 is 0.0769, whereas $Var(\hat{\theta}_u) = 0.0653$ is provided by the unadjusted analysis, this difference is caused by using bootstrapping to calculate the variance

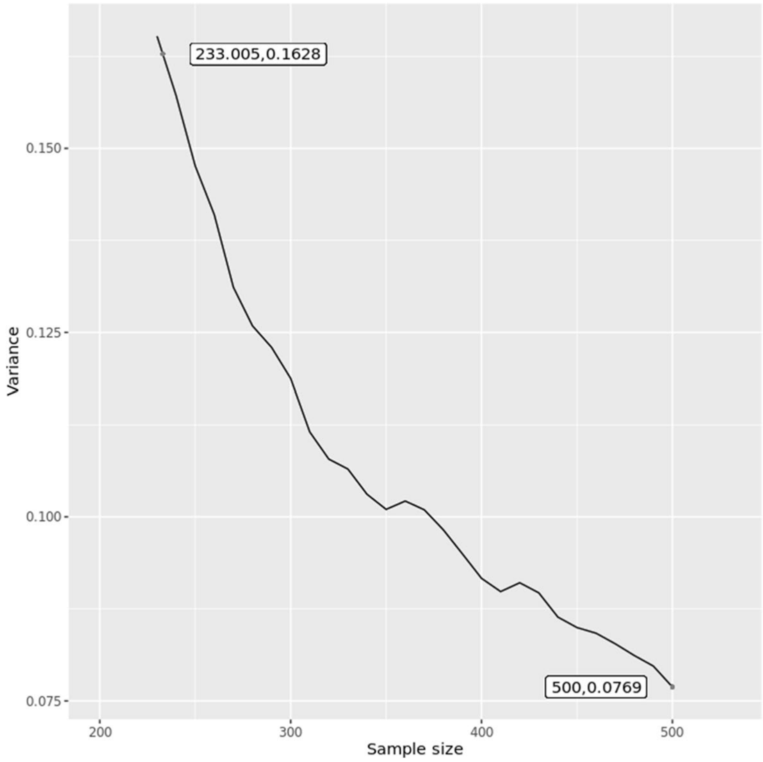

Fig. 2 This figure shows the variances of unadjusted estimates computed using resampling with progressively reduced sample sizes, it visually represents the outcomes shown in Table 1. The coordinates (500, 0.0769) show the sample size and the variance of unadjusted estimates computed with the entire sample size of 500, and the coordinates (233.005, 0.1628) show the effective sample size and variance of the weighted analysis

Table 2 Values of ESS calculated using four different approaches for the numerical example

Approach	Approach 1 Conventional ESS formula	Approach 2 Comparing the variance	Approach 3 Re-sampling	Approach 4 Scaling
ESS	185.6451	200.5176	233.005	200.5178

this rate, so there is no guarantee of monotonicity. Nevertheless, the variance of the unweighted estimator predominantly depends on the sample size, as expected.

Application of approach 4: Scaling method with reduced sample size

A closed formula for the variance of estimated log OR exists for the binary outcomes in this numerical example (Eq. 6). We use Eq. (7) and the 'uniroot' function in R, to solve for the value of p that results in a variance of $Var(\hat{\theta}_w) = 0.1628$. This process provides $\hat{p} = 0.401$, which is the proportion of observations remaining in a hypothetical unweighted sample that produces the same variance as a weighted analysis. Finally, the ESS is calculated as $ESS = n\hat{p} = 200.5178$. This indicates that, after the population adjustment, the weighted sample is 'worth' around 201 patients.

Comparing the results

We summarise values of ESS calculated using all four methods in Table 2. The ESS calculated from the conventional ESS formula produces the smallest ESS, suggesting that this conventional method might underestimate the actual ESS in this numerical example, as previously discussed in the literature [7]. The ESS values calculated using approach 2 (comparing the variance of estimates) and approach 4 (scaling method with reduced sample size) are very similar. However, all methods are in broad agreement and suggest that around 40% of the information from the company's trial is retained after performing population adjustment. This loss of information will have consequences for the precision of the indirect comparison when applying Bucher's method using the adjusted results from the company's trial.

It is hard to assess the performance of each method, for example, in terms of bias and precision, because the ESS is not a model parameter. Rather it is an intuitively appealing descriptive statistic that captures the amount of information retained after population adjustment, that is intended to be readily interpretable. However, for our example, all four methods produce similar ESS values, suggesting that all methods are broadly appropriate.

When applied to our example, the three novel methods avoid making the most seriously violated assumptions required by the conventional method. However we have not attempted to take into account the uncertainty in the weights. Since all three new methods are based upon $Var(\hat{\theta}_w)$ one way to incorporate, but not necessarily fully accommodate, this uncertainty is to estimate $Var(\hat{\theta}_w)$ using bootstrapping, where the weights are estimated within each bootstrap replication. However the uncertainty in the weights was not our primary concern in our example, rather the much more serious concerns are that the outcome data are not homoscedastic and the estimates are not sample means.

Discussion

In this paper, we have developed and applied three new methods to calculate the ESS. Our new approaches can be used in more general cases where different modelling assumptions are made. However, it is challenging to evaluate the accuracy of each method because they do not estimate a population parameter, instead they are merely intuitively appealing descriptive statistics. In the numerical example, all methods produced similar ESS values, which suggests that all methods are feasible. It is also worth noting that the 'comparing the variances method' (Application of approach 2: Comparing the variances of estimates section) and the 'scaling method' (Application of approach 4: Scaling method with reduced sample size section) produce very similar ESS values for this example. This may be because both approaches depend directly on variance calculations using standard methods, and so in general can be expected to be in good agreement.

All four measures of ESS are intended to give a guide to the amount of information available after using a weighting-based approach to performing population adjustment. In situations where the four proposed measures of ESS differ substantially, for example if one is 50% larger than another, then the reasons for this should be investigated. A likely explanation will be that the assumptions underlying the existing conventional approach are violated, rendering this method unreliable. The second approach may also result in a different ESS due to its lack of direct appeal to a smaller sample, and so its potential interpretation as a pseudo ESS. In situations where the ESS metrics differ substantially, the third and fourth methods are likely to be considered the most reliable.

Different approaches for computing the ESS have their own advantages and drawbacks. As we explained previously, the conventional ESS formula may produce misleading results when the homoscedastic assumption on the outcome is violated. The 're-sampling method' (Application of approach 3: Re-sampling with reduced sample size section) may be sensitive to the random seed used. Furthermore the number of observations being removed at each step must be determined, which introduces further sensitivity to choices made by the analyst. A more robust calculation of the ESS could be obtained by reducing the observations being removed from the sample at each step, but the method will then be more computationally expensive. The 'scaling method' (Application of approach 4: Scaling method with reduced sample size section) can only be used when a closed variance formula exists, for example, this would be challenging when using sophisticated statistical methods, where closed form variance formula may be hard to derive.

The three new methods have been applied when a MAIC was used for population adjustment, but they are applicable to the other analyses which involve weighting for this purpose. This includes, but is not limited to, survey weighting, propensity score matching, inverse probability weighting, and inverse probability of censoring weighting. The inverse probability of censoring weighting complicates matters because then each subject has a different weight at different time points. Our methods could be used to compute the ESS separately at different times, and we leave the development of ESS calculations for very complicated types of weighting schemes as further work.

Our four measures of ESS were developed to accompanying weighting-based approaches for performing population adjustment that are popular when using frequentist methods. An advantage of our second and third measures are that they could be used in Bayesian analyses that are, for example, popular in Health Technology Assessments. This is because, when applying these two methods, variances of posterior distributions could be used instead of the variances of parameter estimates from frequentist methods. Our fourth method could also be used in situations where a formula the posterior variance is available, but this is unlikely to be the case in practice.

The effective sample size is conventionally calculated when using weighting methods for population adjustment. However, the three new approaches we have proposed could also be used to calculate the effective sample size when performing this type of adjustment using a regression model. This is because our methods are based on the variances of adjusted and unadjusted estimates, which are also available when using regression-based adjustments. Hence our proposals are much more applicable than the conventional method. In general, we can expect a reduction in the effective sample size when performing population adjustment. However, there might exist unusual cases where this type of adjustment can increase the ESS. The conventional ESS formula cannot capture this, whereas the three new approaches for ESS calculation can. This may be regarded as another advantage of our proposals.

To summarise, we have developed three new approaches for calculating the ESS. They are more applicable than the conventional approach. We have illustrated all methods using an illustrative example and conclude that our proposals should accompany, and potentially replace the existing approach for computing the ESS when using statistical methods for population adjustment.

Abbreviations

ESS	Effective sample size
IPCW	Inverse probability of censoring weighting
MAIC	Matching-adjusted indirect treatment comparison

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12874-024-02412-1.

Supplementary Material 1: Example R code and datasets accompanying this article can be fond in the online supplementary materials.

Acknowledgements

LZ and DJ wrote the manuscript. LZ provided software code. SB contributed to improving the manuscript. All authors read, made contributions to and approved the final manuscript.

Code availability

The example R code is provided in the online supplementary materials.

Authors' contributions

LZ and DJ wrote the manuscript. LZ provided software code. SB contributed to improving the manuscript. All authors read, made contributions to and approved the final manuscript.

Funding

Not applicable.

Data availability

The data can be fond in the supplementary materials.

Declarations

Ethics approval and consent to participate

This is a paper about statistical methods. All data are simulated. No ethical approval for the use of these data is required.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 5 March 2024 Accepted: 11 November 2024
Published online: 20 November 2024

References

1. Richardson WS, Wilson MC, Nishikawa J, Hayward RS. The well-built clinical question: a key to evidence-based decisions. ACP J Club. 1995;123(3):A12–3.
2. Schardt C, Adams MB, Owens T, Keitz S, Fontelo P. Utilization of the PICO framework to improve searching PubMed for clinical questions. BMC Med Inform Decis Making. 2007;7:1–6.
3. International Council for Harmonisation of Technical Requirements Registration Pharmaceuticals Human Use. E9(R1) statistical principles for clinical trials: addendum: estimands and sensitivity analysis in clinical trials. 2021. https://www.fda.gov/regulatory-information/search-fda-guidance-documents. Accessed 17 Jul 2023.
4. Caro JJ, Ishak KJ. No head-to-head trial? Simulate the missing arms. Pharmacoeconomics. 2010;28(10):957–67.
5. Ishak KJ, Proskorovsky I, Benedict A. Simulation and matching-based approaches for indirect comparison of treatments. Pharmacoeconomics. 2015;33(6):537–49.
6. Phillippo DM, Ades A, Dias S, Palmer S, Abrams KR, Welton NJ. NICE DSU technical support document 18: methods for population-adjusted indirect comparisons in submissions to NICE. Report by the Decision Support Unit. 2016. https://research-information.bris.ac.uk/files/94868463/Population_adjustment_TSD_FINAL.pdf.
7. Phillippo DM, Ades AE, Dias S, Palmer S, Abrams KR, Welton NJ. Methods for population-adjusted indirect comparisons in health technology appraisal. Med Decis Mak. 2018;38(2):200–11.
8. Signorovitch JE, Wu EQ, Andrew PY, Gerrits CM, Kantor E, Bao Y, et al. Comparative effectiveness without head-to-head trials. Pharmacoeconomics. 2010;28(10):935–45.
9. Signorovitch JE, Sikirica V, Erder MH, Xie J, Lu M, Hodgkins PS, et al. Matching-adjusted indirect comparisons: a new tool for timely comparative effectiveness research. Value Health. 2012;15(6):940–7.
10. Jackson D, Rhodes K, Ouwens M. Alternative weighting schemes when performing matching-adjusted indirect comparisons. Res Synth Methods. 2021;12(3):333–46.
11. Kalton G, Flores-Cervantes I. Weighting methods. J Off Stat. 2003;19(2):81.
12. Lumley T. Analysis of complex survey samples. J Stat Softw. 2004;9:1–19.
13. Little RJ, Vartivarian S. Does weighting for nonresponse increase the variance of survey means? Surv Methodol. 2005;31(2):161.
14. D'Agostino RB Jr. Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. Stat Med. 1998;17(19):2265–81.
15. Joffe MM, Rosenbaum PR. Invited commentary: propensity scores. Am J Epidemiol. 1999;150(4):327–33.
16. Williamson E, Morley R, Lucas A, Carpenter J. Propensity scores: from naive enthusiasm to intuitive understanding. Stat Methods Med Res. 2012;21(3):273–93.
17. Seaman SR, White IR, Copas AJ, Li L. Combining multiple imputation and inverse-probability weighting. Biometrics. 2012;68(1):129–37.
18. Seaman SR, White IR. Review of inverse probability weighting for dealing with missing data. Stat Methods Med Res. 2013;22(3):278–95.
19. Robins JM, et al. Information recovery and bias adjustment in proportional hazards regression analysis of randomized trials using surrogate markers. In: Proceedings of the Biopharmaceutical Section. vol. 24. San Francisco: American Statistical Association; 1993. p. 3.
20. Robins JM, Finkelstein DM. Correcting for noncompliance and dependent censoring in an AIDS clinical trial with inverse probability of censoring weighted (IPCW) log-rank tests. Biometrics. 2000;56(3):779–88.
21. Willems S, Schat A, van Noorden M, Fiocco M. Correcting for dependent censoring in routine outcome monitoring data by applying the inverse probability censoring weighted estimator. Stat Methods Med Res. 2018;27(2):323–35.
22. Sullivan TR, Latimer NR, Gray J, Sorich MJ, Salter AB, Karnon J. Adjusting for treatment switching in oncology trials: a systematic review and recommendations for reporting. Value Health. 2020;23(3):388–96.
23. Latimer NR, Abrams KR, Lambert PC, Crowther MJ, Wailoo AJ, Morden JP, et al. Adjusting survival time estimates to account for treatment switching in randomized controlled trials-an economic evaluation context: methods, limitations, and recommendations. Med Decis Making. 2014;34(3):387–402.

24. Phillippo DM, Dias S, Elsada A, Ades A, Welton NJ. Population adjustment methods for indirect comparisons: a review of national institute for health and care excellence technology appraisals. Int J Technol Assess Health Care. 2019;35(3):221–8.
25. Remiro-Azócar A, Heath A, Baio G. Methods for population adjustment with limited access to individual patient data: a review and simulation study. Res Synth Methods. 2021;12(6):750–75.
26. Efron B. Bootstrap methods: another look at the jackknife. In: Breakthroughs in statistics: Methodology and distribution. New York: Springer; 1992. pp. 569–93.
27. Parmar MK, Torri V, Stewart L. Extracting summary statistics to perform meta-analyses of the published literature for survival endpoints. Stat Med. 1998;17(24):2815–34.
28. Bucher HC, Guyatt GH, Griffith LE, Walter SD. The results of direct and indirect treatment comparisons in meta-analysis of randomized controlled trials. J Clin Epidemiol. 1997;50(6):683–91.
29. Petto H, Kadziola Z, Brnabic A, Saure D, Belger M. Alternative weighting approaches for anchored matching-adjusted indirect comparisons via a common comparator. Value Health. 2019;22(1):85–91.

Publisher's Note