Sample size calculations in randomized trials: common pitfalls

Penelope M.A. Brasher PhD,* Rollin F. Brant PhD†

MANY articles have appeared in medical journals over the years providing guidance on the calculation of sample sizes for clinical trials. For the most part, these articles deal only with the mechanics of sample size calculation and do not provide guidance on a number of practical issues. Ignorance of these issues can lead to flawed calculations and underpowered trials. In this editorial, we discuss a number of issues that have recurred in our statistical consultations with investigators and authors and in journal manuscript reviews.

Panel 1: Sample size formula for a continuous outcome, assuming $\alpha = 0.05$, power = 0.90, and equal sample sizes in the two groups.

n = the sample size in each of the groups
 μ_1 = population mean in treatment group 1
 μ_2 = population mean in treatment group 2
 σ^2 = population variance
 1.96 = two-sided normal cut-off for $\alpha = 0.05$
 1.28 = one-sided normal cut-off for power = 0.90

$$n = \frac{(1.96 + 1.28)^2 \sigma^2}{(\mu_1 - \mu_2)^2}$$

The sample size calculation

Panel 1 illustrates the sample size calculation for a hypothesis test to compare the averages of two populations where the outcome is continuous. Four components are required: the type I error (α), power, the variance of the outcome measure (σ^2), and the difference the investigator wishes to detect ($\mu_1 - \mu_2$). The interested reader is referred to the 2005 article by Schulz and Grimes¹ for an excellent introduction to sample size calculations.

Choosing parameter values for the calculations

We can see from the equation in Panel 1 that the sample size calculation is based on using the *population* variance, σ^2 . However, this is an unknown quantity and so instead an estimate of the variance must be employed. Frequently, investigators will use an estimate obtained from a pilot study in the patient population of interest. However, unless the pilot study was large, this sample variance is an unreliable estimator of σ^2 and its use may result in the actual power of the planned study being less than the planned power.² To mitigate this problem Browne suggests calculating an upper one-sided confidence limit for the population variance from the pilot data in accordance with the planned power (e.g., an 80% upper limit for power of 80%) and then using this upper limit in the sample size calculation for the planned trial.

For the same reason, it is unwise to base a sample size calculation on the *observed* difference in a pilot study. As shown in Panel 1, the sample size calculation is based on the difference in the *population* means. The observed difference from the pilot study is an *estimate* of this difference and is subject to *sampling error*; not accounting for the imprecision of this estimate in the sample size calculation will lead to underpowered studies.³ While one could take account of the imprecision in the sample size calculation, replication of a difference observed in a small pilot study is not the best approach.

Instead, the trial should be powered to detect a minimal clinically important difference (MCID). Jaeschke, Singer and Guyatt⁴ defined the minimal important difference as “the smallest difference in score in the domain of interest which patients perceive as beneficial and which would mandate, in the absence of troublesome side-effects and excessive cost, a change in the patient’s management”. Determining the MCID is often not straightforward. For clinical

CAN J ANESTH 2007 / 54: 2 / pp 103–106

From the Centre for Clinical Epidemiology and Evaluation, Vancouver Coastal Health Research Institute,* and the Department of Statistics,† University of British Columbia, and the Centre for Community Child Health Research, Child and Family Research Institute, Vancouver, British Columbia, Canada.

Address correspondence to: Dr. Penelope M.A. Brasher, Centre for Clinical Epidemiology and Evaluation, Vancouver Coastal Health Research Institute, 828 West 10th Avenue, Vancouver, British Columbia V5Z 1L8, Canada. Phone: 604-875-4111, ext 68167; Fax: 604-875-5179; E-mail: brasher@interchange.ubc.ca

outcomes one could consider differences that have led to adoption of new therapies in the past. Another approach is to survey experts in the field to determine what difference would need to be demonstrated for them to adopt the new therapy in view of its costs and risks. Spiegelhalter and Freedman describe such an approach in determining the MCID for the CHART trials.⁵ If the outcome of interest is health-related quality of life the MCID may have been determined during the development of the instrument. For example, Juniper and colleagues determined the MCID for the Asthma Quality of Life Questionnaire during its initial development.⁶

Some authors have argued that specification of variability and treatment differences can be sidestepped by noting that the calculation depends on the *standardized treatment difference* $(\mu_1 - \mu_2)/\sigma$. This is merely mathematical sleight of hand, since a meaningful choice of the standardized difference depends crucially on specifying $(\mu_1 - \mu_2)$.⁷

Panel 2: Sample size formula based on the expected width of a 95% confidence interval (CI).

$$n = 2 \frac{(1.96)^2 4\sigma^2}{(\text{width})^2}$$

derived from 95% CI for $(\mu_1 - \mu_2)$:

$$\Delta \pm 1.96\sqrt{2\sigma^2/n}$$

Sample sizes based on the expected width of the confidence interval (CI)

A less common approach to sample size calculations for randomized controlled trials is based on a CI approach (Panel 2). That is, the investigator does not wish to test a hypothesis that one treatment is better than another but rather wishes to estimate the difference between treatments. This approach calculates the sample size required so that the *expected width* of the CI is no more than some pre-specified value. The *actual* width of the CI from the trial, however, may be either larger or smaller than the *expected* width since it will incorporate the *sample variance*, i.e., $2 \times 1.96\sqrt{2s^2/n}$. We can improve the sample size calculation by accounting for the stochastic nature of s^2 . That is, we wish to ensure that with a certain probability (power) the width of the CI is no more than a pre-specified value. This can be achieved by including a *tolerance probability* in the calculation.⁸

It is important to note that this approach is not appropriate when the aim is to exclude *pre-specified differences* in treatment efficacy as in non-inferiority or equivalence trials. In these cases the difference in the population means must also be incorporated in the calculation⁹ yielding calculations identical to those based on power.

Power after the fact

An issue not directly related to *a priori* sample size calculations but which also warrants comment is *post hoc power*. Such calculations are misleading, since calculating power is essentially a pre-trial planning step.^{7,10} Once a trial is completed, CIs convey the range of possible effects consistent with the data.

In many trials the outcomes may not be continuous, as above, but instead may be dichotomous (e.g., success of therapy) or survival (e.g., time to failure). While the details of calculations differ, the substantial issues raised here are still relevant. Whatever the case, while the wide availability of relevant software has greatly eased the technical burden of sample size calculations, such issues should be addressed at the earliest possible phase of trial design. Leaving them as an afterthought to be addressed in the last minute scramble of submitting the proposal may lead to a rejected proposal, or worse, could result in expenditure of time, effort and funding on a scientifically flawed study.

Calculs de la taille des échantillons dans les études randomisées : pièges fréquents

Au fil des années, de nombreux articles ont été publiés dans les revues médicales afin de fournir des pistes quant au calcul de la taille des échantillons pour les études cliniques. La plupart de ces articles ne traitent que du processus de calcul de la taille des échantillons ; ils ne fournissent pas de conseils sur nombre de questions pratiques. La méconnaissance de ces questions peut entraîner des erreurs de calcul et des études manquant de puissance. Dans cet éditorial, nous abordons un certain nombre de questions qui reviennent dans nos consultations statistiques auprès des chercheurs et

des auteurs, ainsi que dans l'évaluation des manuscrits des revues médicales.

Le calcul de la taille de l'échantillon

Le tableau 1 illustre le calcul de la taille de l'échantillon pour un test d'hypothèse comparant les moyennes de deux populations, où la variable est continue. Quatre éléments sont nécessaires : l'erreur de type I (α), la puissance, la variance de la variable mesurée (σ^2), et la différence que l'on souhaite détecter ($\mu_1 - \mu_2$). Le lecteur intéressé peut consulter l'article de Schulz et Grimes¹ de 2005, une excellente introduction au calcul des tailles d'échantillon.

Tableau 1: Formule du calcul de la taille d'échantillon pour une variable continue, avec $\alpha = 0,05$, puissance = 0,90, et tailles d'échantillon égales dans les deux groupes.

n = taille de l'échantillon dans chaque groupe
 μ_1 = moyenne de population dans le groupe de traitement 1
 μ_2 = moyenne de population dans le groupe de traitement 2
 σ^2 = variance de la population
 1,96 = seuil pour test normal bilatéral avec $\alpha = 0,05$
 1,28 = seuil pour test normal unilatéral pour une puissance = 0,90

$$n = \frac{(1.96 + 1.28)^2 \sigma^2}{(\mu_1 - \mu_2)^2}$$

Le choix des valeurs des paramètres pour les calculs

L'équation du tableau 1 nous montre que le calcul de la taille de l'échantillon est basé sur l'utilisation de la variance de la *population*, σ^2 . Cette quantité est néanmoins inconnue ; c'est pourquoi une estimation de la variance est employée. Les chercheurs ont fréquemment recours à une estimation obtenue dans une étude pilote de la population de patients en question. Toutefois, sauf s'il s'agissait d'étude pilote d'envergure, cette variance de l'échantillon est un estimateur imprécis de σ^2 , et son utilisation peut résulter en une perte de puissance de l'étude.² Dans le but de réduire ce problème, Browne suggère de calculer une limite de confiance supérieure unilatérale pour la variance de population issue des données pilote selon la puissance voulue (par ex., une limite supérieure de 80% pour une puissance de 80%), et d'ensuite se servir de cette limite supérieure dans le calcul de la taille de l'échantillon pour l'étude planifiée.

C'est également la raison pour laquelle il est déconseillé de baser un calcul de la taille de l'échantillon sur la différence *observée* dans une étude pilote. Comme nous le montrons dans le tableau 1, le calcul de la taille de l'échantillon est basé sur la différence des moyennes de *population*. La différence observée de l'étude pilote est une *estimation* de cette différence et est dès lors sujette à une *erreur d'échantillonnage* ; le fait de ne pas prendre en compte le côté imprécis de cette estimation dans le calcul de la taille de l'échantillon résultera en des études de faible puissance.³ Bien qu'il soit possible de tenir compte de l'imprécision du calcul de la taille de l'échantillon, il est préférable de ne pas chercher à reproduire une différence observée dans une petite étude pilote.

L'étude devrait plutôt être dirigée vers la recherche de la différence minimale cliniquement importante (DMCI). Jaeschke, Singer et Guyatt⁴ définissent la différence minimale importante comme « la plus petite différence dans le domaine d'intérêt qui est considérée comme avantageuse par les patients et qui requerrait, en l'absence d'effets secondaires gênants et de coûts excessifs, un changement dans la prise en charge du patient ». La détermination de la DMCI est souvent complexe. Dans le cas de résultats cliniques, il est possible de prendre en compte les différences ayant par le passé mené à l'adoption de nouvelles thérapies. Une autre approche consiste à interroger les experts du domaine afin de déterminer quelle différence leur paraît importante pour l'adoption de la nouvelle thérapie au vu de ses coûts et risques. Cette approche est décrite par Spiegelhalter et Freedman pour la détermination de la DCMI dans les études CHART.⁵ Si le résultat recherché est une qualité de vie liée à la santé, la DCMI a peut-être été définie pendant le processus de développement de l'instrument. Par exemple, dans une étude sur l'asthme Juniper *et coll.* ont déterminé la DCMI pour le questionnaire sur la qualité de vie pendant son développement initial.⁶

Certains auteurs ont défendu qu'on n'avait pas besoin de spécifier la variabilité et les différences entre les traitements en remarquant que le calcul dépend de la différence standardisée des traitements $(\mu_1 - \mu_2)/\sigma$. Ceci n'est qu'une astuce mathématique puisqu'un choix significatif de la différence standardisée dépend de manière cruciale de la spécification de $(\mu_1 - \mu_2)$.⁷

Tailles d'échantillon basées sur l'envergure escomptée de l'intervalle de confiance (IC)

Une approche moins courante dans le calcul de la taille de l'échantillon pour des études randomisées et contrôlées se fonde sur une approche basée sur

l'IC (tableau 2). Ici, le chercheur ne souhaite pas tester l'hypothèse qu'un traitement est meilleur qu'un autre, mais plutôt estimer la différence entre les traitements. Cette approche calcule la taille de l'échantillon nécessaire afin que *l'envergure escomptée* de l'IC ne soit pas supérieure à une valeur spécifiée au préalable. L'envergure *véritable* de l'IC de l'étude peut toutefois être supérieure ou inférieure à l'envergure *escomptée*, étant donné qu'elle englobe la variance de l'échantillon, c'est à dire $2 \times 1.96 \sqrt{2s^2/n}$. Le calcul de la taille de l'échantillon peut être amélioré en tenant compte de la nature stochastique de s^2 . Ainsi, nous souhaitons nous assurer qu'avec une certaine probabilité (puissance) l'envergure de l'IC ne sera pas plus qu'une valeur spécifiée au préalable. Ce résultat peut être atteint en incluant une *probabilité de tolérance* dans le calcul.⁸

Tableau 2: Formule du calcul de la taille de l'échantillon basée sur l'envergure escomptée d'un intervalle de confiance (IC) de 95%.

$$n = 2 \frac{(1.96)^2 4\sigma^2}{(\text{envergure})^2}$$

dérivée de IC 95% pour $(\mu_1 - \mu_2)$:

$$\Delta \pm 1.96 \sqrt{2\sigma^2/n}$$

Il faut souligner que cette approche ne convient pas lorsque l'objectif est d'exclure des *différences pré-spécifiées* dans l'efficacité d'un traitement, comme lors d'études de non infériorité ou d'équivalence. Dans ces cas-là, la différence dans les moyennes de population doit également être incluse dans le calcul⁹, résultant dans des calculs identiques à ceux basés sur la puissance.

Puissance après coup

Une question qui n'est pas directement liée au calcul de la taille de l'échantillon *a priori*, mais qui légitime un approfondissement, est la *puissance post hoc*. De tels calculs sont trompeurs, puisque le calcul de la puissance est essentiellement une étape de planification ayant lieu avant l'étude.^{7,10} Une fois l'étude terminée, la fourchette d'effets possibles et cohérents avec les données est mise en évidence par les IC.

Dans de nombreuses études, les résultats peuvent ne pas être continus, comme c'est le cas ci-dessus, mais plutôt dichotomiques (par ex. le succès de la thérapie)

ou de survie (par ex. le temps jusqu'à l'échec). Bien que les détails des calculs diffèrent, les questions essentielles posées ici sont tout aussi pertinentes. Quel que soit le cas de figure, et bien que l'accès à des logiciels pertinents ait grandement facilité le fardeau technique des calculs de taille d'échantillon, de telles questions doivent être posées le plus tôt possible lors de la conception d'une étude. Si ces calculs sont laissés de côté jusqu'à la dernière minute avant la soumission du projet, le projet pourrait être rejeté ou pire, avoir pour conséquence des dépenses de temps, d'énergie et de fonds pour une étude scientifiquement erronée.

References

- 1 Schulz KF, Grimes DA. Sample size calculations in randomised trials: mandatory and mystical. *Lancet* 2005; 365: 1348-53.
- 2 Browne RH. On the use of a pilot sample for sample size determination. *Stat Med* 1995; 14: 1933-40.
- 3 Goodman SN. A comment on replication, p-values and evidence. *Stat Med* 1992; 11: 875-9.
- 4 Jaeschke R, Singer J, Guyatt GH. Measurement of health status. Ascertaining the minimal clinically important difference. *Control Clin Trials* 1989; 10: 407-15.
- 5 Spiegelhalter DJ, Freedman LS. A predictive approach to selecting the size of a clinical trial, based on subjective clinical opinion. *Stat Med* 1986; 5: 1-13.
- 6 Juniper EF, Guyatt GH, Willan A, Griffith LE. Determining a minimal important change in a disease-specific quality of life questionnaire. *J Clin Epidemiol* 1994; 47: 81-7.
- 7 Lenth RV. Some practical guidelines for effective sample size determination. *Am Stat* 2001; 55: 187-93.
- 8 Kupper LL, Hafner KB. How appropriate are popular sample size formulas? *Am Stat* 1989; 43: 101-5.
- 9 Makuch R, Simon R. Sample size requirements for evaluating a conservative therapy. *Cancer Treat Rep* 1978; 62: 1037-40.
- 10 Goodman SN, Berlin JA. The use of predicted confidence intervals when planning experiments and the misuse of power when interpreting the results. *Ann Intern Med* 1994; 121: 200-6.