Sample size calculations for randomised controlled trials and for prediction models

Francesca Fiorentino
Department of Surgery and Cancer, Imperial College London, London UK

Tom Treasure*
Clinical Operational Research Unit, University College London

Corresponding author
Tom Treasure: tom.treasure@gmail.com

There was no funding for this work.
The authors have no conflicts of interest to declare.

Abstract

The two study protocols are published in this issue Colorectal Disease: FALCON, a multicentre randomised controlled trial of strategies to reduce surgical site infection, and AFAR, a predictive model of atrial fibrillation after colonic resection. Both are exemplars of excellent research design that surgeon researchers should seek to emulate. Trial statisticians were involved at an early stage and the protocols have been through several rounds of peer review by trial methodologists, prior to being funded by the National Institute for Health Research (NIHR). In this article we address the important question of sample size calculations and how they should be approached for these very different forms of study.

Main Text

Most surgical procedures came into practice without randomised trials because, against a well-known experience of clinical outcomes over many years, an appropriate and well conducted operation was seen to make a dramatic and lasting difference. For example, Thomas's splint only had to be seen in use for injured farmers in north Wales, and then soldiers in the 1914-18 war, to become universally adopted. The relief of pain in the hours and days after injury was evident, followed by recovery to walk on legs of matching length, with both feet pointing forward. To generalise that process of deduction, the features that indicate that an RCT is not needed are a close temporal and mechanistic relationship between the intervention and the effect, resulting in a large and sustained benefit.(1) The Thomas's splint became the standard initial treatment, applicable to the large majority of patients with femoral fracture.

In contrast, lung metastasectomy is carried out in fewer than one in thirty of the patients who have lung metastases.(2) The outcome of importance is survival. For lung metastasectomy, results are usually given as survival rate, usually at an interval of five-years, but there are too many factors and uncertainties to conclude that metastasectomy has a survival benefit by observation alone.(3)

Calculating the sample size for a randomised controlled trial

It is wasteful of time and effort to embark on a study that is not large enough to provide a conclusive answer, or so large as to be wasteful of effort and resources.(4) To calculate a sample size, the statistician needs to know what is (a) the outcome of importance, (b) the outcome measure and (c) the clinically meaningful effect size.

For lung metastasectomy, survival beyond five years was the only outcome reported in the 51 follow-up studies found in a systematic review (5) so for our first illustration (a) survival is the outcome of importance. Survival of ~40% at five years has been consistently reported and was confirmed in a meta-analysis including 2925 patients. (6) For the illustration we will identify the survival rate at 5-years to be the outcome measure (b). The effect size depends on what would be the survival without metastasectomy. The US Society of Thoracic Surgeons based its recommendations on a consensus assumption of zero survival, but for this illustration will use the more cautious "worse than 5%" suggested by the authors of the meta-analysis. Then (c) is the absolute difference between 40% and 5%, the effect size of 35%.

The surgeons need to agree with statistician the value of *alpha*—the probability of a false positive—usually set at 5% and hence the familiar P<0.05. The value of *beta*—the probability of a false negative—is usually set at 20% or more cautiously 10%. Power is 1-*beta* so in percentage terms these are expressed as 80% or 90%, that is the power to avoid a false negative. Given these estimates a statistician can generate Table 1. This is for 1:1 randomisation and shows that 44 patients (22 in each arm) would provide 80% power for a two-sample proportion test. There are likely to be patients lost to follow up, so the target recruitment might be set at 50.

In cancer trials it is usual to use time to death (overall survival) or cancer progression (progression free survival) for (b) the outcome measure. The statistical test used for the sample size calculation is the two-sample comparison of survivor functions (log-rank test). The same assumptions can be used to do the calculation, but the statistical method takes into account the time of the event, death. It captures more information than a simple count of 5-year survivors, so it requires commensurately fewer patients. Using the log rank test, the statistician can produce Table 2. Randomisation is still 1:1 and shows that 36 patients (18 in each arm) would provide 80% power with a two-sample survivor function test. A total of42 patients would allow for loss to follow up.

In the discussion between the investigators and the statistician, all should be alert to the possibility of "back calculation". The surgeons know the number of patients available for recruitment and can tweak the effect size to give an achievable number of randomised patients. In the case of lung metastasectomy the consensus assumption of zero survival(7) had for years ruled out the possibility of randomisation at all; there was no prospect of equipoise. Also, it conveniently attributed all the credit for survival to the effect of the operation and trumps any likely effect from chemotherapy.

Tables 1 and 2 illustrate the principle, but it is not how the conversation with the statistician went in the case of the PulMiCC trial. The investigators had reason to believe that patients eligible for metastasectomy had better survival than was widely assumed. This came from a comparative study in 1980(8) and a modelling study on cancer registry data in 2006.(9) Both suggested the possibility that metastasectomy makes a much smaller difference to survival than assumed. Knowing that, the statistician asked what was the smallest clinically meaningful difference in the five-year survival that might justify lung metastasectomy. A difference from 40% survival in the treated down to 30% survival in the control was suggested (10% difference). Table 3 shows the calculation using a two-sample survivor function test.

As we said, the actual sample size calculation may be much more complicated. In fact, the PulMiCC trial was powered for *non-inferiority* of leaving the metastases unresected using time to event analysis.(10) With this smaller difference (40% and 30%) the numbers needed to power the study were commensurately higher, and in the event not achievable due to the tenacity with which cancer teams held on to the near zero assumption and its implications. (11, 12) It is also important to remember that for the sample size calculation it was important to be realistic at the planning stage, The assumptions are replaced by findings once the data are in, and the prior power calculation plays no part in the analysis or interpretation of the results.(13)

It may be important to not rely on randomisation, but to ensure that there is a balance in prognostic factors between the randomized groups, particularly if these factors might create

differences of a magnitude that compete with the treatment effect (confounding factors). For example, obesity in studies of surgical site infection(14) which might be relevant in FALCON.(15) In the case of the PulMiCC trial the unfavourable features were more than one metastasis, liver involvement, carcinoembryonic antigen elevation and shorter interval since the primary resection. In large drug trials this process is done by stratification but in trials of limited size an alternative is *minimisation* which adjusts the probability of a patient being assigned to one or other arm in order to achieve balance between the groups in the known factors, relying on randomisation to balance the unknown confounders.(16, 17) It is essential that this is done by a strict algorithm out of sight of anyone involved in the trial.

Prediction models

Prediction models are used for investigating patient outcomes in relation to patient and disease characteristics. They may be of use in surgical practice and we give three examples.

1. In the AFAR study(18) the adverse outcome to be "predicted" is the onset of new atrial fibrillation during the recovery period. Patients in the stratum more likely to have this problem can then have further planned screening or prophylactic approaches. The model is intended to target more costly and labour intensive methods to where they will achieve the greatest benefit for patients.
2. A predictive model has been developed to risk adjust postoperative mortality among patients having of colorectal cancer.(19) It allows fair comparisons to be made between hospitals, clinical teams and individual surgeons. Implementation of public reporting in 2013 was followed by a fall in the observed surgical mortality. The model allowed this to be interpreted as a real reduction in mortality without risk avoidance. (20)
3. A third example is to select patients for surgery by gaining insights into their likelihood of death or survival after surgery. We will return to an unsatisfactory example in the development of a model with this purpose as a cautionary tale.(21)

A standard approach is to use a "training" dataset and the model is then tested with a separate "validation" dataset which has been held back for the purpose. Following the same principle as the sample size calculation the statistician must be provide with the best available data, informed estimates of as yet unquantified factors and what outcome would be useful. The outcome can be a continuous scale, categorical or estimated survival (time to event).

The model developed by Walker, Finan and van der Meulen(22) used internal validation and is more sophisticated than can be described here but it illustrates the power of a collaborative effort with data available on 62,314 patients in the National Bowel Cancer Audit and collaboration with very highly skilled data analysts. Eight risk factors were included and mortality was counted up to 90 days. This captures 50% more deaths, virtually all having a relationship to treatment. The methods of "imputation" for missing data (and missingness is inevitable) and validation were at a high level of expertise.

The tables are set up to illustrate the fewest of counts that might allow for a valid model. In the case of survival (time to event outcome), the overall event rate and the mean follow-up time need to be known. In the case of binary outcome, the outcome proportion expected within the model development dataset, based on previous evidence.

Tables 4 and 5 give examples of sample size for prediction models of binary and survival outcomes prediction models.

For less common disease or particular circumstances there may still be a desire to create models to inform practice. A recent published example is of a scoring system to select patients more likely to "benefit" from lung metastasectomy for sarcoma included 135 patients.(21) The scoring system has three parameters giving scores of 0-3. What can be lauded in the report is that the authors provide the data. The figure is taken from their paper. The well-used caution "correlation does not mean causation" can be applied. The more important value of r^2 is 0.144 indicating that the scores contribute very little, <15%, to the prognosis. It is clear from the graphical display that the scores really do not discriminate usefully between lengths of survival.

Sarcoma has a predilection for metastasising to the lungs and these patients are often young so there is pressure to do something, anything, to help. The two longest survivors at over 12 years who scored 2/3 on the scale will be pointed out repeatedly on clinic visits, generating confirmation bias. What will not be recalled is the harm done by operations of unproven effectiveness on patients the larger number of patients not coming back to clinic.

Table 1

Power	Alpha	Effect size	N
80%	0.05	0.35	44
90%	0.05	0.35	56

In the case of an effect size of 35% (0.35) and using a two-sample proportions test (Pearson's chi-squared test) the variation in power and sample size can be seen, for the first scenario with assumed 35% survival gain from metastasectomy over control.

Table2

alpha	power	N	Expected events	Hazard Ratio	Survival Metastasectomy	Survival no operation
80%	0.5	36	28	3.269	40%	5%
90%	0.5	48	38	3.269	40%	5%

A sample size calculation for the two-sample comparison of survivor functions (log-rank test) using the same assumptions.

Table 3

Power	Alpha	N	Expected events	Hazard Ratio	Survival Metastasectomy	Survival no operation
80%	0.5	656	427	1.314	0.4	0.3
90%	0.5	880	571	1.314	0.4	0.3

A sample size calculation for the two-sample comparison of survivor functions (log-rank test) raising the control estimate from 5% to 30%.

Table 4

Predictor parameters	Outcome prevalence	Minimum Sample Size	Number of events
10	10%	348	35
20	10%	695	70
30	10%	1042	105
10	40%	369	148
20	40%	519	208
30	40%	778	312

Example for a binary outcome where the expected outcome proportion is 10% or 40% with model parameters 10, 20 and 30 in Table 4.

Table 5

Predictor parameters	Overall event rate	Minimum Sample Size	Number of outcome events
10	6.5%	1715	231
20	6.5%	3429	462
30	6.5%	5143	692

Example for a survival outcome where the mean follow-up is 2 years, the overall event rate is 0.065 and the time for model prediction is 2 years with model parameters 10, 20 and 30 in Table 5.

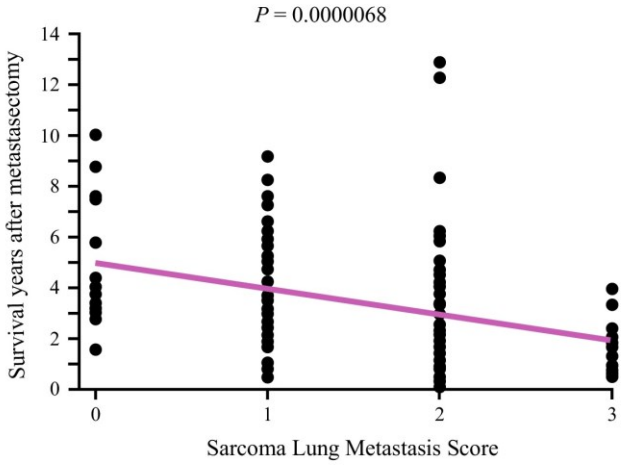

FIG. 4 Relationship between the survival period after lung metastasectomy and the Sarcoma Lung Metastasis Score. Pearson's product-moment correlation coefficient was calculated between the Sarcoma Lung Metastasis Score and survival time after lung metastasectomy, and a significant correlation was observed between these two factors ($r = -0.38$, $p = 0.0000068$)

1. Glasziou P, Chalmers I, Rawlins M, McCulloch P. When are randomised trials unnecessary? Picking signal from noise. BMJ. 2007;334(7589):349-51.
2. Fenton H, Finan PJ, Milton R, Shackcloth M, Taylor J, Treasure T, et al. National variation in pulmonary metastasectomy for colorectal cancer. Colorectal Disease. 2020;Accepted for Publication.
3. Milosevic M, Edwards J, Tsang D, Dunning J, Shackcloth M, Batchelor T, et al. Pulmonary Metastasectomy in Colorectal Cancer: updated analysis of 93 randomized patients - control survival is much better than previously assumed. Colorectal Dis. 2020;22(10):1314-24.
4. Glasziou P, Chalmers I. Research waste is still a scandal-an essay by Paul Glasziou and Iain Chalmers. BMJ. 2018;363:k4645.
5. Fiorentino F, Hunt I, Teoh K, Treasure T, Utley M. Pulmonary metastasectomy in colorectal cancer: a systematic review and quantitative synthesis. J R Soc Med. 2010;103(2):60-6.
6. Gonzalez M, Poncet A, Combescure C, Robert J, Ris HB, Gervaz P. Risk factors for survival after lung metastasectomy in colorectal cancer patients: a systematic review and meta-analysis. Ann Surg Oncol. 2013;20(2):572-9.
7. Handy JR, Bremner RM, Crocenzi TS, Detterbeck FC, Fernando HC, Fidias PM, et al. Expert Consensus Document on Pulmonary Metastasectomy. Ann Thorac Surg. 2019;107(2):631-49.
8. Aberg T, Malmberg KA, Nilsson B, Nou E. The effect of metastasectomy: fact or fiction? Ann Thorac Surg. 1980;30(4):378-84.
9. Utley M, Treasure T, Linklater K, Moller H. Better out than in? The resection of pulmonary metastases from colorectal tumours. In: Xie X, Lorca F, Marcon E, editors. Operations Research for Health Care Engineering: Proceedings of the 33rd International Conference on Operational Research Applied to Health Services. Saint-Etienne: Publications de l'Universitaire de Saint-Etienne; 2008. p. 493-500.
10. Milosevic M, Edwards J, Tsang D, Dunning J, Shackcloth M, Batchelor T, et al. Pulmonary Metastasectomy in Colorectal Cancer (PulMiCC): Updated analysis of 93 randomised patients - control survival is much better than previously assumed. Colorectal Dis. 2020;https://onlinelibrary.wiley.com/doi/full/10.1111/codi.15113.
11. Treasure T, Farewell V, Macbeth F, Monson K, Williams NR, Brew-Graves C, et al. Pulmonary Metastasectomy versus Continued Active Monitoring in Colorectal Cancer (PulMiCC): a multicentre randomised clinical trial. Trials. 2019;20(1):718.
12. Macbeth F, Fallowfield L. The myth of pulmonary metastasectomy. Br J Cancer. 2020;123(4):499-500.
13. Cox DR. Planning of Experiments: Wiley; 1958.
14. Wilson AP, Livesey SA, Treasure T, Gruneberg RN, Sturridge MF. Factors predisposing to wound infection in cardiac surgery. A prospective study of 517 patients. Eur J Cardiothorac Surg. 1987;1(3):158-64.
15. Nepogodiev D, Bahngu A. Pragmatic multicentre factorial randomised controlled trial testing measures to reduce surgical site infection in low- and middle-income countries: study protocol of the FALCON trial. Colorectal Disease. 2021;IN PRESS.
16. Treasure T, MacRae KD. Minimisation: the platinum standard for trials? BMJ. 1998;317(7155):362-3.
17. Altman DG, Bland JM. Treatment allocation by minimisation. BMJ. 2005;330(7495):843.
18. Lee MJ, Hawkins DJ, Bradburn MJ, Lee J, Brown SR, Wilson MJ. Atrial Fibrillation After Resection (AFAR) A progress III Study. Colorectal Disease. 2021;IN PRESS.
19. Walker K, Finan PJ, van der Meulen JH. Model for risk adjustment of postoperative mortality in patients with colorectal cancer. Br J Surg. 2015;102(3):269-80.

20. Vallance AE, Fearnhead NS, Kuryba A, Hill J, Maxwell-Armstrong C, Braun M, et al. Effect of public reporting of surgeons' outcomes on patient selection, "gaming," and mortality in colorectal cancer surgery in England: population based cohort study. BMJ. 2018;361:k1581.
21. Yamamoto H, Yamamoto H, Soh J. A Simple Prognostic Benefit Scoring System for Sarcoma Patients with Pulmonary Metastases: Sarcoma Lung Metastasis Score. Annals of Surgical Oncology. 2020.
22. Vallance AE, Van Der Meulen J, Kuryba A, Braun M, Jayne DG, Hill J, et al. Socioeconomic differences in selection for liver resection in metastatic colorectal cancer and the impact on survival. Eur J Surg Oncol. 2018;44(10):1588-94.