DOI: 10.1111/ppe.12475

COMMENTARY

WILEY Paediatric and Perinatal Epidemiology

The statistical joys—and added complications—of twin studies

James A. Hanley

Department of Epidemiology, Biostatistics, and of Occupational Health, McGill University, Montreal, QC, Canada

Correspondence
James A. Hanley, Department of Epidemiology, Biostatistics, and Occupational Health, McGill University, Montreal, QC, Canada.
Email: james.hanley@mcgill.ca

Just like twins themselves, studies—experimental and non-experimental—involving twins are special. They allow sharper insights than we could achieve in other, noisier settings. Examples, involving just a pair of twins, are Einstein's thought experiment on the effect of space travel on ageing, and—a century later—NASA's real one on its effects on DNA, and the body more generally. Another, involving 1162 monozygous and heterozygous twins, is a trial that measured the frequency of short-term adverse events caused by childhood vaccines.[1] Yet another, with a substantial impact, is an "experiment" designed by Nature that involved 115 twin pairs born to HIV-infected mothers.[2,3]

In his critique of a "milk-feeding" experiment on 20 000 schoolchildren that went awry,[4] William Gosset ("Student") made a case for the statistical efficiency of the "split-plot" design used in agricultural experimentation. For the milk vs no milk comparison, he suggested "pairs of the same age group and sex, and as similar in height, weight and especially physical condition (ie well or ill nourished) as possible, and divided into 'controls' and 'feeders' by tossing a coin for each pair." For the "raw vs pasteurised" milk comparison, "the error of the comparison may be relied upon to be so small that 50 pairs of [identical twins] would give more reliable results than the 20 000 with which we have been dealing. [Thus,] it would be possible to obtain much greater certainty at an expenditure of perhaps 1-2 percent of the 7500 pounds [500 000 today] and less than 5 percent of the trouble."

Gains in efficiency can also be achieved when (as in the perinatal trial[5] re-analysed by David Cox in *Biometrika*) the comparison is within day/season of the year rather than family. Greater precision can also be achieved with "within-the-same-subject" or "crossover" trials. However, as Barr and colleagues[6] note, not all crossover trials are as easy, or as elegant, as we imagined.

Likewise, not all trials involving twins/related individuals (or a mix of related and unrelated individuals) are as statistically efficient (joyful) as those involving only unrelated individuals. In some settings, all the related individuals in the same unit (eg not yet born twins of the same woman) have to be allocated to the same "arm." In these situations, as with cluster randomised trials in general, the (usually) positive correlation in the responses of those in the same unit/cluster, and the fact that these related individuals are "on the same side" of the comparison, make for a *larger* standard error. Indeed, in such settings, were it not for the cost saving from recruiting "*m* for the price of fewer", it would be (statistically) more efficient to enrol one individual from each of *m* units, rather than *m* related individuals from the same unit. The planning inputs and software tools Yelland and colleagues describe in this issue of *Paediatric and Perinatal Epidemiology*[7] will make it easier for researchers in such situations to count these mixed blessings, and plan accordingly.

Before commenting on these, a word about the prevailing practice of "determining" or "estimating" *the* "required" sample size. A colleague once asked me: when you attend religious worship, are you required to put a specified amount in the collection box, or is it rather that every donation contributes-like in a meta-analysis-to the overall amount collected? Instead, how about the term sample-size "considerations"?

Yelland and colleagues[7] are to be thanked for putting (most of) the sample-size considerations for studies involving related individuals, or a mix of related and unrelated individuals, in one place. By switching between designs by toggling between "Cluster," "Individual", and "Opposite," the online tool can also be used to plan different types of observation-only (non-experimental) studies involving singletons and pairs. Their earlier *Statistics in Medicine* article addressing the general case can be used for any mix of clusters of different sizes.

Their "use all available data" approach will help investigators move away from one statistically cruel way to avoid non-independent responses. Our 2003 "GEE" article described a non-experimental study (examining the benefits of families eating meals together) where 1 in 4 subjects had a sibling in the study. But the authors split up these siblings (using randomisation) and removed the data from one of them!

Another welcome aspect is Yelland et al's use of a traditional correlation, with range −1 to +1, rather than the intraclass correlation ("fraction") with range 0 to +1. The latter is still used in most sample-size considerations for studies/trials involving clusters. In most applications, the two correlations will coincide. But in special

situations, the within-cluster *sum* may be constrained or shared, possibly because of space or other structural factors. Such limitations can lead to a genuinely negative intracluster correlation.[8] A random effects ("hierarchical") model, with its two additive variance components, cannot describe such situations, whereas the minimalist, and thus more flexible, GEE approach, with its broader range of possible intrarelationships, can. Even in the usual (positive correlation) contexts, the GEE approach is also an alternative for correlated binary responses: these are sometimes difficult to fit via hierarchical models, and variance components on a logit scale are difficult to visualise.

But how to come up with an ICC for planning purposes? All I can recommend is "carefully": if it will be working against us, then better to overestimate its magnitude; if with us, then better to underestimate the help it brings. In any case, since it is rare that a single study is definitive, we should accept that one study merely contributes to a meta-analysis, even if it may not have been as precise as we planned.

The authors are to be thanked for their work in providing ICC estimates for several relevant paediatric and perinatal outcomes, and contexts. Like all correlations, they are "range dependent." I was struck by how much smaller the ICCs were for standardised birthweight scores than birthweight itself. If we condition on a sufficient number of important determinants, such as gestational age, might the ICC even become negative,[8] as it can for birthweights of animals in the same litter?

Although their article, and the tool it offers, is limited to twins, the formulae in their earlier *Statistics in Medicine* paper accommodate a generic "cluster size" *m*, and can be applied to cluster randomised trials. When *m* is very large, even a tiny ICC leads to a large variance inflation, and a considerable loss in statistical power/precision. Having underestimated the ICCs for triceps skinfold thickness and verbal IQ in the PROBIT trial, my colleagues now warn others to ensure that assessors are spread over multiple units.[9] (The Lanarkshire trial also employed a cluster randomisation of the 67 schools, and the measuring of the initial and final height and weight required "the whole time of 5 doctors and 17 nurses" for 2 weeks at each end.)

Initially, the formula in their earlier statistical article seemed too simple to require a "black box" calculator,[10] but I came to appreciate the heuristics the online tool offers. Start, say, with a sample size of 100 unrelated subjects and then consider a cluster randomisation setting. Begin with say 1 pair of twins and an ICC of 0.01, then move to an ICC of 0.5 or 0.99 and each time see what happens. One can infer what would happen if say, the ICC is 0.4, and if instead of a pair of twins, we had one set of *triplets*. The first of the three contributes the same statistical information as a singleton, and each of the others contributes as much as 0.6 of a singleton. So, their total *contribution* is $1 + 2 \times (1-\text{ICC})$, or in general, $1 + (m-1) \times (1-\text{ICC})$. I wonder if, rather than approaching sample-size considerations using the *variance* for an estimator, we should work with its reciprocal, *information* [which is additive] instead. Thus, in the cluster randomisation context, think of (1-ICC) as the (*reduced*) contribution, after the first, from each *additional* individual in the cluster. Every 100 individuals comprising 89 singletons, 4 pairs of twins, and 1 set of triplets, are the "singleton-equivalent" of $89 + 4 + 1 + (4 + 2) \times (1-\text{ICC})$, or $94 + 6 \times (1-\text{ICC})$ *independent* units of statistical information.

As for the "opposite" randomisation design, where a pair of twins contribute $1/(1-\text{ICC})$ times *more* statistical information than two unrelated individuals, readers are encouraged to read Gosset's classic piece, still relevant today. The online tool should allow them to back-calculate how large an ICC and how much "greater certainty" he had in mind when he reduced the number of children in the "raw vs pasteurised" comparison from 20 000 to 100.

ABOUT THE AUTHOR

James Hanley is a biostatistician who began his career in clinical trials in oncology. Since joining McGill University, Canada, he has taught extensively, and collaborated widely, from pediatrics to geriatrics. He has also developed an interest in the history of epidemiology and statistics.

ORCID

James A. Hanley http://orcid.org/0000-0003-3486-5291

REFERENCES

1. Virtanen M, Peltola H, Paunio M, Heinonen OP. Day-to-day reactogenicity and the healthy vaccine effect of measles-mumps-rubella vaccination. *Pediatrics*. 2000;106:e62.
2. Duliège AM, Amos CI, Felton S, Biggar RJ, Goedert JJ. Birth order, delivery route, and concordance in the transmission of human immunodeficiency virus type 1 from mothers to twins. International Registry of HIV-Exposed Twins. *J Pediatr*. 1995;126:625-632.
3. Campbell H, Hanley JA. Twin data that made a big difference, and that deserve to be better-known and used in teaching. *J Stat Educ*. 2017;25:131-136.
4. Student. The Lanarkshire milk experiment. *Biometrika*. 1931;23:398-406.
5. Gordon T, Foss BM. The role of stimulation in the delay of onset of crying in the newborn infant. *Q J Exp Psychol*. 1996;18:79-81.
6. Barr RG, Wooldridge J, Hanley J. Effects of formula change on intestinal hydrogen production and crying and fussing behavior. *J Dev Behav Pediatr*. 1991;12:248-253.
7. Yelland L, Price D, McPhee A, Lee K. Accounting for twin births in sample size calculations for randomised trials. *Paediatr Perinat Epidemiol*. 2018; EPUB ahead of print.
8. Hanley JA, Negassa A, Edwardes MD. GEE analysis of negatively correlated binary responses: a caution. *Stat Med*. 2000;19:715-722.
9. Kramer MS, Martin RM, Sterne JAC, Shapiro S, Dahhou M, Platt RW. The double jeopardy of clustered measurement and cluster randomization. *BMJ*. 2009;339:503-505.
10. Hanley JA, Moodie EEM. Sample size, precision and power calculations: a unified approach. *J Biomet Biostat*. 2011;2:5.