Statistically Speaking: Demystifying Methods

Before and after studies and historical controls. Is the proof in the pudding?

Delivering definitive airway management using rapid sequence induction in a pre-hospital setting can indeed be a daunting task with little margin for error. Minor infractions at any step in the sequence can rapidly culminate to create an airway catastrophe. Nowhere else would the aggregation of marginal gains [1] reap more benefit than in such adverse conditions. In this issue of *Anaesthesia*, Angerman and colleagues [2] evaluated the effect of introducing a videolaryngoscope and a bougie into a standardised protocol for airway management. They demonstrated > 12% increase in first-attempt tracheal intubation success in a 'before and after' comparison. As it is widely accepted that airway complications increase with the increased number of attempts, this would represent a potentially significant improvement in patient care [3].

As tempting as it may be to attribute this success to the introduction of novel equipment, upon closer scrutiny of the intervention, it is clear that the investigators are really comparing a new airway management protocol that included several "soft elements" in teamwork communication, with an older, less standardised protocol (though this is somewhat understated in the title). It is difficult to tease out the marginal contribution that the pieces of equipment made without access to more information such as the specific reasons for the failed first tracheal intubation attempts in the cohort before the new protocol was introduced, such as an unfavourable Cormack-Lehane grade or difficulty encountered when manoeuvring the tracheal tube into the correct position. The before and after study design does not lend itself to such comparisons, and perhaps a prospective randomised trial is the only way to resolve this issue. The more important question is whether a prospective randomised controlled trial is worthwhile? If the pudding tastes great, does it matter which ingredient contributed to its success?

There are situations when conducting randomised trials is neither feasible nor ethical, and comparing airway management in the pre-hospital setting might fall into this category. Using data from epidemiological studies, administrative databases and previously conducted clinical trials to estimate the effect of a placebo or standard of care, is one approach to assessing the benefits of a new therapy or protocol. However, the use of such 'historical controls' is not without its criticisms. Studies comparing a new treatment or method to historical controls may be unreliable as a mixture of retrospective and prospectively collected data will be used, and also because of unknown sources of bias. These sources of bias include:

a. changes in patient characteristics over time, improvements in diagnostic tests, or changes in clinical rating assessments may change the staging of patients receiving treatment;
b. changes in the use of concomitant medications;
c. lack of blinding during data collection; and
d. operator-dependent elements (for e.g. do the operators' skill improve over time?)

For example, in a study looking at the trials of six drugs for which data from both randomised trials and historical controlled trials are available, Sacks et al [4] found that 44 of the 56 trials which used historical controls showed that the newer therapy was better than the control regimen, whilst only 10 of the 50 randomised controlled trials agreed with this outcome. In situations where historical controls must be used, covariate adjustment, or matching baseline characteristics between the control and study group should be considered when estimating treatment effects [5].

Patient outcomes are likely to have some correlation with variables that can be measured before random assignment; in the case of Angerman et al, the variable which may have affected first-pass success rate is ‘threatening airway obstruction’, which was present in 3.8% of the historical control group, but only 1.3% of the current study group (difference between two groups p = 0.0005 using Fisher’s Exact test). Angerman et al could consider matching the two groups for this variable (rather than using all available data), allowing the isolation of the effect of the new regime on first-pass success rate with greater precision and power. The investigators decided not to use propensity matching in the current study due to the fact that some patient characteristics (specifically the anatomical features) were not included in the data set.

Although retrospective studies and the use of historical control data is unavoidable in some situations, we should do well to remember that frequentist statistical analytical methods are used (most of the time) to analyse data resulting from these studies. Frequentist statistical theory is based on the idea of random sampling. If the patients in the study were randomly allocated to either treatment, the differences between the treatment groups would behave like the differences between random samples taken from a single population. Because we know how random samples are supposed to behave, we can then compare the observations between treatment groups with what we would expect to see if there was no difference between the two treatments. The difference in the primary outcome of first-pass success in the current study was analysed using Fisher’s Exact Test, based on frequentist theory and used to determine if there are non-random associations between two categorical variables. Continuous variables were analysed using the Mann-Whitney test, which is another example of a statistical test based on frequentist theory. Unfortunately, there is no consensus on which statistical tests should be used in situations where data were not obtained using random methods. Though there are, however, some precautions which investigators considering before and after studies or the use of historical controls can take into account. These precautions include:

1. *Data collection timeframe and seasonal effects:*

The timeframe for data collection should be the same. Take for example, Miles et al's study investigating general anaesthesia versus conscious sedation for transfemoral aortic valve implantation [6]. Data were collected over a 12-month period on those patients who underwent surgery under general anaesthesia, data collected during a six-month transition period when the newer technique using conscious sedation was introduced was not used in the final analysis to avoid cross-over effects. Finally, data was collected and used for analysis over another, consecutive 12-month period when conscious sedation became the standard technique. Although Angerman et al [2] did factor in a three-month cross-over period when data collected was not analysed, the data collection periods before and after the new protocol was established differed by ten months (data from 12 months was used in the 'before' cohort, whereas data from 22 months was used in the 'after' cohort). Matching times of data collection is important so that seasonal variations in event rates can be accounted for. For example, major trauma requiring pre-hospital setting intubation and airway management may be more prevalent during certain seasons amenable to extreme outdoor sports.

2. *Participant matching (propensity score or one to one matching)*

Patient characteristics should be matched as far as possible. For example, Miles [6] used a propensity matching method, matching on body mass index and EuroSCORE-2. The EuroSCORE-2 was used as a surrogate measure of perioperative risk and is an aggregate risk score from a wide range of variables. In the case of Angerman, propensity (or one to one) matching could have been performed on Cormack-Lehane grade or Mallampati class [3], if data on anatomical features were available.

3. *Time bias:*

   A pertinent issue regarding time bias in Angerman's study is the experience level of the personnel involved. For example, first-pass success might have improved in the later cohort not because of a protocol change or additional use of a bougie, but simply because the intubation team (which was the same in both time periods) have attained a further two years' experience in the field and are simply better at their job. Controlling for experience in the data analysis is therefore essential [7].

4. *Sample size calculation:*

   Unlike sample size requirements for randomised controlled studies, there is considerably less research conducted on calculation of sample size for propensity score matching. This is because the number of participants required for matching largely depends on the magnitude of differences between the two cohorts, as well as differences expected in the primary outcome. There are no easy to apply formulas for calculating sample size and investigators may have to use rules of thumb such as those described for regression analysis [8].

The take-home message from Angerman's before and after study is that combining videolaryngoscopy and bougie with a standardised rapid sequence induction protocol leads to a higher first attempt tracheal intubation success rate, but as with all studies investigating a 'bundle of processes' [9] it is not possible to quantify the effect of each variable on the final outcome [5], regardless of whether the study was a prospective, randomised controlled study or not. This study has shown that before and after studies can be informative in difficult pre-hospital setting, and it is perhaps useful to take a more prospective viewpoint on the use of historical controls, by always viewing data collected today as potentially being controls for trials conducted in the future.

References

1. Lumb AB, McLure HA. AAGBI recommendations for standards of monitoring during anaesthesia and recovery 2015 – a further example of 'aggregation of marginal gains'. *Anaesthesia* 2016; **71:** 3-6.

2. Ångerman S, Kirves H, Nurmi J. A before-and-after observational study of a protocol for use of the C-MAC videolaryngoscope with a Frova introducer in pre-hospital rapid sequence intubation. *Anaesthesia***:** n/a-n/a.

3. Cook TM. Strategies for the prevention of airway complications – a narrative review. *Anaesthesia* 2018; **73:** 93-111.

4. Sacks H, Chalmers TC, Smith H, Jr. Randomized versus historical controls for clinical trials. *American Journal of Medicine* 1982; **72:** 233-40.

5. Smith AF, Choi SW. Major trauma and the need for massive transfusion. *Anaesthesia* 2017; **72:** 1299-301.

6. Miles LF, Joshi KR, Ogilvie EH, et al. General anaesthesia vs. conscious sedation for transfemoral aortic valve implantation: a single UK centre before-and-after study. *Anaesthesia* 2016; **71:** 892-900.

7. Ho AMH, Dion PW, Ng CSH, Karmakar MK .Understanding immortal time bias in observational cohort studies. *Anaesthesia* 2013; **68:** 126-30.

8. Choi SW, Lam DMH. Regression: How much data do I really need? *Anaesthesia* 2017; **72:** 1029-30.

9. Moore JA, Conway DH, Thomas N, Cummings D, Atkinson D. Impact of a peri-operative quality improvement programme on postoperative pulmonary complications. *Anaesthesia* 2017; **72:** 317-27.

Acknowledgements

SC is statistical advisor to Anaesthesia. No other competing interests declared.

S. W. Choi

Assistant Research Officer

Department of Anaesthesiology,

The University of Hong Kong,

Hong Kong, HKSAR

Email: htswchoi@hku.hk

G. T. C. Wong

Clinical Associate Professor,

Department of Anaesthesiology,

Queen Mary Hospital,

Hong Kong, HKSAR