Improving the reporting and interpretation of clinical trial outcomes

Increasing demands on clinicians means that 'time-poverty' has become a very real issue.[1] Despite this, the majority of practitioners still spend some time keeping abreast of the latest guidelines, reviews, and clinical trial reports, ensuring their patients get the most effective treatments. Unfortunately this is not being made easy. Not only is the volume of medical literature increasing, but the way in which reviews and clinical trials are being reported is becoming more complicated. Largely, this is a side effect of increasing complexities in trial design, and use of increasingly sophisticated methods of analysis. This may be set to become even more challenging with the advent of value-based pricing to inform NHS pharmaceutical purchasing decisions. Nevertheless, should advancements in the underlying scientific process mean that answers to important clinical questions become less accessible to those who need them the most? Not necessarily.

Earlier this year we reported that a group of clinicians who see patients with back pain felt that clinical trials were difficult to interpret and not written with them in mind.[2] Clinicians expressed dissatisfaction and an unfamiliarity with current reporting methods and suggested that a standardised set of reporting methods, including description of individual improvements would facilitate consumption, and aid the transition of the research into practice.

THE PROBLEM

One pitfall associated with the way outcomes are currently reported may be a lack of standardisation. It can be tempting for authors to report their results in ways that make an intervention appear as attractive as possible or lends as much weight as possible to their views; it is well-documented that the use of relative terms leads to increased perceptions of treatment effectiveness.[3] For example, if the annual risk of venous thromboembolism in a non-pregnant woman, not using combined oral contraception, is 0.00005, but increases to 0.00015 after commencing combined oral contraception;[4] reporting this as tripling the risk (a risk ratio of 3.0) is likely to have more impact than reporting the absolute risk increases by one episode every 10 000 woman-years. Three times something very small, is still something small. Compounded by time restrictions one can see how, in a brief side-by-side comparison of trial reports, reporting methods may influence clinicians' judgments more than underlying differences in effectiveness. A further pitfall may be the sheer variety of reporting methods used: some clinicians may not feel confident interpreting risk ratios and confidence intervals; so more exotic reporting methods may confound and confuse even the most statistically aware practitioner.

So can we not standardise the reporting of outcomes? Is there any reason why outcomes, even if derived via relatively complicated means, cannot be reported in simpler, more accessible terms? The short answers to these questions may be that there is little to stop us, and that outcomes can be reported in more straightforward terms, so long as there is motivation do to so. It may be that authors have been forced to focus on keeping up with methodological advancements, and successful competition in getting their work into the best journals, in order to satisfy the needs of maximising their employers' research profiles. It would be wise to set a short-term aim in applied health research to improve patient-reported outcomes and how these are communicated to end-users. Losing sight of this could risk authors failing to satisfy the needs of the consumers of their research: healthcare professionals and planners.

BARRIERS TO IMPROVING INTERPRETATION

Some trial outcomes are easy to interpret. For example, large trials of cardiovascular interventions may need to do little more than count the dead bodies to understand if an intervention is worthwhile. Some explanatory trials with biological interpretations are fairly immune from misinterpretation too; if the outcome is easily understood by specialising end-users and reports will not be used to directly inform practice. However in pragmatic trials with patient-reported outcome measures, such as a visual analogue score, the interpretation of clinical importance can be much more difficult. More objective measures can be equally challenging: just how many mmHg is an important benefit from a new antihypertensive agent, or how many litres per minute quantifies a worthwhile improvement in peak flow in an asthma trial, is seldom entirely clear. Furthermore, what is important to an individual patient, may not equate to what is important at a population level. A 3–5 mmHg reduction in blood pressure may be relatively trivial at an individual level, but if one were to lower blood pressure in a population by this magnitude it would likely lead to an important reduction in the number of strokes.[5] One can see the danger in judging outcomes by what is important to the individual, rather than to a population: that the latter may be smaller than the former. At the very least, greater care needs to be taken when designing trials and when interpreting outcomes that use judgment thresholds, as inadequate distinctions between individuals and populations may be being made.[6] But we suggest that further work is needed to critically review the methods suggested for defining thresholds of population level importance.

Notwithstanding the issues surrounding the definition of importance, a nagging problem remains. The measurement error (in terms of reproducibility) of patient-reported outcome measures can exceed the minimally important change for an individual.[7,8] Using the largest of these thresholds to make clinical decisions may be impractical, since few patients are likely to achieve such magnitudes of change. The measurement error can, in some circumstances, be over two-thirds of the whole outcome measure.[7] This leaves an uncomfortable dilemma: is it appropriate to use patient-reported outcome measures to make clinical decisions about individual patients? This may be of more concern in the consulting room than in trials. For depression at least, some limitations of measuring individuals with patient-reported outcome measures have been identified; this comes as GPs, rewarded for carrying out standardised assessments, have voiced concerns surrounding other possible effects on patient care, resulting from using such outcomes in clinical practice.[9,10] However in trials, the combination of large numbers of participants and randomisation, goes a long way to help mitigate the effects of false-positive and false-negative diagnoses of individual improvements on decision-making when comparing trial groups (if we assume misclassification is non-differential). One pragmatic solution to

using patient-reported outcome measures to make decisions about individual patients in practice, and improving the accuracy of these measures in trials, may be to agree a threshold for marking improvement that is a compromise between measurement error and importance (that is, a threshold set somewhere between the two).[11]

In addition to issues of how we measure and report in trials, there have been calls to re-examine what we measure and report. For example, for some chronic pain conditions it may be that current outcomes do not capture what is important to patients and patients need to be involved much more in the development process.[12,13] Identifying more patient-centred and relevant information will enable much sharper measurement instruments to be developed. In clinical trials, these improved 'next generation' outcome measures could help us better differentiate which treatments work for which patients. Moreover, improving the reporting of outcomes could become rather immaterial if the underlying outcomes are sub-optimal.

TOWARDS CLEARER, ACCESSIBLE, AND CLINICALLY-USEFUL REPORTING

Reporting continuous patient-reported outcome measures using only mean change, or mean difference in change between groups, is a practice we suggest needs to be consigned to the past. In the case that there is a variable response to treatment, it is possible for a small, and perhaps provisionally unattractive mean difference, to mask a group of patients for whom treatment was highly successful.[14] Furthermore, mean differences reported alone can be difficult to interpret; not only because of the confusion surrounding individual and population importance, but also because for some condition-specific domains, so many patient-reported outcome measures are in use, that some generalists may be unfamiliar with all of the scales.

Appropriate thresholds for judgement of individual improvements and group benefits, permits the use of reporting methods that facilitate consumer interpretation. For example, a group of academics working in low back pain recommended a suite of reporting methods for improving the interpretation of back pain trials.[2,15] They recommend reporting the difference in the proportion of individuals improving in each group, and the number needed to treat (NNT: the number of participants that must be randomised to receive the intervention to gain, on average, one extra improvement over the control; a favourite of GPs and researchers alike). Additional reporting methods such as these help to contextualise mean differences, especially when differences are small, or typical in magnitude. Box 1 provides a summary of their full recommendations. These reporting methods have the advantage of being readily understandable, easily communicated to patients, and they do not require any specialist training, nor recall of epidemiology/statistics modules that were taken in the dim and distant past.

Consider instances where these additional reporting methods have been used. If the mean difference between groups is small and the difference in the number of individuals improving between groups is also small (or conversely NNT is large); one may conclude that the treatment is unlikely to offer much in the way of clinical advantage, assuming good internal and external validity and the intervention is not particularly inexpensive. Alternatively, if the mean difference between groups is small, or typical, but the difference in the proportion of individuals improving between groups is large (or the NNT small), then the treatment may well be attractive to clinicians and purchasers.

Given the commonalities that exist across trials of interventions of chronic conditions in terms of the shared challenges surrounding the interpretation of outcomes, authors of trial reports could consider whether adopting additional modes of reporting could aid clarity and interpretation for end-users. In particular, authors could consider whether using reporting methods based on individual improvements could facilitate interpretation. While for some conditions, this may necessitate some preliminary work, or at least some head-scratching in order to define or understand individual and population importance thresholds, the potential benefits to healthcare professionals and planners, in our view, justifies any work needed and a continued motivation to change reporting behaviour.

Box 1. Consensus statement of recommendations for future reporting of low back pain trial outcomes

Where possible, authors of future back pain trials should endeavour, when reporting a continuous primary outcome measure, to include the following reporting methods:

1. Between-group difference, with its 95% confidence interval (CI). Where possible the between-group minimally important difference (that is, what is important at a population-level) should also be specified.
2. The between-group difference in proportion, with its 95% CI, of participants improving by a score equal or larger than an established and relevant minimally important change threshold (that is, what is important at an individual-level, adjusted for measurement error).
3. The number needed to treat (NNT), with its 95% CI, for one participant to improve, on average, by a score equal or larger than an established and relevant minimally important change threshold.

Where possible the between-group difference in proportion, with its 95% CI, of participants deteriorating by a score equal or larger than an established minimally important change threshold for deterioration should be reported.

The inclusion of a contingency table should be considered. Results may additionally be reported using alternative approaches (for example, relative risk, odds ratio, standardised mean difference) according to the needs of a particular trial.

Adapted from Froud R, Eldridge S, Kovacs F, et al. *Eur J Pain 2011;* ***15(10):*** *1068–1074 (Table 2).*[15]

ADDRESS FOR CORRESPONDENCE

Robert Froud
Queen Mary University of London, Centre for Health Sciences, 2 Newark Street, Whitechapel, London, E1 2AT, UK.

E-mail: r.j.froud@qmul.ac.uk

Robert Froud,
Senior Research Fellow, Centre for Primary Care and Public Health, Queen Mary University of London, London; and Professor of Health Sciences, Norwegian University College of Health Sciences, Campus Kristiania, Oslo.

Martin Underwood,
Professor of Primary Care Research, Warwick Clinical Trials Unit, Warwick Medical School, University of Warwick, Coventry.

Sandra Eldridge,
Professor of Biostatistics, Centre for Primary Care and Public Health, Queen Mary University of London, London.

Provenance
Commissioned; not externally peer reviewed.

REFERENCES

1. Silverman J, Kinnersley P. Calling time on the 10-minute consultation. *Br J Gen Pract* 2012; **62(596):** 118–119.
2. Froud R, Underwood M, Carnes D, Eldridge S. Clinicians' perceptions of reporting methods for back pain trials: a qualitative study. *Br J Gen Pract* 2012; DOI: 10.3399/bjgp12X630034.
3. Covey J. A meta-analysis of the effects of presenting treatment benefits in different formats. *Med Decis Making* 2007; **27(5):** 638–654.
4. Faculty of Family Planning and Reproductive Health Care. *Faculty of Family Planning and Reproductive Health Care clinical guidance. First prescription of combined oral contraception.* London: Faculty of Family Planning and Reproductive Health Care, 2007.
5. Rose G. *Individuals and populations. The strategy of preventive medicine.* Oxford, UK: Oxford University Press, 1992.
6. de Vet H, Terluin B, Knol D, *et al.* Three ways to quantify uncertainty in individually applied 'minimally important change' values. *J Clin Epidemiol* 2010; **63(1):** 37–45.
7. Terwee CB, Roorda LD, Knol DL, *et al.* Linking measurement error to minimal important change of patient-reported outcomes. *J Clin Epidemiol* 2009; **62(10):** 1062–1067.
8. de Vet HC, Terwee CB, Ostelo RW, *et al.* Minimal changes in health status questionnaires: distinction between minimally detectable change and minimally important change. *Health Qual Life Outcomes* 2006; **4:** 54.
9. Leydon GM, Dowrick CF, McBride AS, *et al.* Questionnaire severity measures for depression: a threat to the doctor–patient relationship? *Br J Gen Pract* 2011; **61(583):** 117–123.
10. Moore M, Ali S, Stuart B, *et al.* Depression management in primary care: an observational study of management changes related to PHQ-9 score for depression monitoring. *Br J Gen Pract* 2012; DOI: 10.3399/bjgp12X649151.
11. Ostelo RWJG, Deyo RA, Stratford P, *et al.* Interpreting change scores for pain and functional status in low back pain: towards international consensus regarding minimal important change. *Spine (Phila Pa 1976)* 2008; **33(1):** 90–94.
12. Mullis R, Barber J, Lewis M, Hay E. ICF core sets for low back pain: do they include what matters to patients? *J Rehabil Med* 2007; **39(5):** 353–357.
13. Foster NE, Dziedzic KS, van der Windt DA, *et al.* Research priorities for non-pharmacological therapies for common musculoskeletal problems: nationally and internationally agreed recommendations. *BMC Musculoskelet Disord* 2009; **10:** 3.
14. Froud R, Eldridge S, Lall R, Underwood M. Estimating the number needed to treat from continuous outcomes in randomised controlled trials: methodological challenges and worked example using data from the UK Back Pain Exercise and Manipulation (BEAM) trial. *BMC Med Res Methodol* 2009; **9:** 35.
15. Froud R, Eldridge S, Kovacs F, *et al.* Reporting outcomes of back pain trials: a modified Delphi study. *Eur J Pain* 2011; **15(10):** 1068–1074.