# New initiatives in critical care: distinguishing hype from hope

John L Moran and Patricia J Solomon


## ABSTRACT

- Recent viewpoints on critical care have expressed frustration at the slow development of new therapeutic agents and the failure of investigator-initiated trials. Several new directions have been proposed: personalised medicine and the embracing of "omic" technologies, resolving the heterogeneity of treatment effects, and adaptive trial designs. We examine these approaches in the context of analysis of randomised controlled trials (RCTs).
- The curse of treatment effect heterogeneity is found not only in critical care but also in cancer oncology. We find the uncritical appeal to personalised medicine to be misplaced because such treatments are not identified at the personal level, but at the group or stratified level. The analysis of RCTs has foundered over the problem of accounting for the centre effect and rejecting the random effects approach. Enthusiasm for adaptive trial designs has been articulated at the rhetorical, not the substantive, level.


Crit Care Resusc 2016; 18: 141-147

A recently published viewpoint of Marini and colleagues outlines potential new research strategies (the evolving toolkit) and treatments for the critically ill,[1] with several themes highlighted in the article, including personalised critical care, low rates of attempted replication, individual predisposition, and novel study designs (dynamic adaptive) and analysis (pre-specified examination for heterogeneity of treatment effects). It will not surprise many that such concerns are current in the broader medical literature,[2] but the debate about personalised medicine has pre-dated the current view from critical care by nearly 20 years.[3] Importantly, heterogeneity of treatment effects (HTE) (not to be confused with subject-level heterogeneity[4]) is not a property that is peculiar to critical care trials.

In 1980, Simon, from within the cancer paradigm, observed that "clinical trials are characterised by heterogeneity in patients treated and by variability in responses obtained",[5] a view restated in 2014 by Venook and colleagues:

> it is critical to understand why the phase III trial results are so often not as expected. The heterogeneity of cancer across and within a disease is likely one of the reasons.[6]

In the sepsis paradigm, at least 100 Phase II and Phase III trials have yet to show "a dramatic improvement in survival for patients",[7] and a lead editorial has characterised the field of sepsis research as a "quagmire".[8] This may mandate new directions but we should not forget that, despite the failures of specific sepsis therapeutics, mortality in patients with sepsis has declined, in Australia at least, over the period 2000–2012.[9]

### Abbreviations

| | |
|---|---|
| AD | adaptive design |
| ANOVA | analysis of variance |
| CONSORT | Consolidated Standards of Reporting Trials |
| FDA | Food and Drug Administration |
| FE | fixed effect |
| GEE | generalised estimating equations |
| HTE | heterogeneity of treatment effects |
| ICC | intraclass correlation coefficient |
| OR | odds ratio |
| RCT | randomised controlled trial |
| RE | random effect |
| SE | standard error |
| US | United States |

Our purpose in this review is to consider the above recurrent themes from a methodological viewpoint, in the context of a discussion of past and present approaches to the analysis of randomised controlled trials (RCTs). We also aim to suggest where we may be on firm footing, or otherwise, in embracing these "new" initiatives. We additionally mention here that it is somewhat galling to read that five anti-tumour necrosis factor agents have been approved by the United States and European drug licensing authorities for use in treating a variety of rheumatic diseases; these approvals involve 200 RCTs and global drug sales estimated at US$200 billion (estimated to 2013).[10]

What is meant by personalised medicine is difficult to define. A historical review has suggested that it is not the case that personalised medicine has always existed, but that there have been different historical forms relating to the tensions between universalism and specificity in Western medicine.[11] Smith, in a review of the implications of epigenetics ("non-Mendelian inheritance of a trait located somewhere between 'genotype' and 'phenotype'"[12]), has reminded us that with respect to "the most celebrated cases of supposedly personalised medicine … treatments are not personalised; rather they are stratified".[13] This is an

updated expression of a search begun in 1977, initially from a regression modelling perspective:

> We are interested in the question "which treatment is best for which kinds of patients?" rather than the classical question, "which treatment is best overall?" … By studying the combinations of covariates which lead to selection of the various treatments as optimal, we make recommendations of treatment for different kinds of patients.[14]

Updated methods for investigation of HTE include application of trial evidence to an independent target sample, for example, registry data;[15] and estimation of upper and lower bounds of treatment response heterogeneity.[16]

Estimation of individual responses to treatment in a conventional RCT is not easy. For example, Senn discusses a 70% patient response rate to an intervention,[17] which would be conventionally reported as 70% always-responders and 30% non-responders. However, the data would also be consistent with the radically different interpretation that the intervention worked for 100% of the patients 70% of the time. The difference between these two interpretations is stark: that success in treatment is a permanent patient feature (the conventional interpretation) versus that there is random variability. So how can individual response be identified, given that we do not normally observe the patient under conditions of both treatment and control? A parallel-arm trial will show between-treatment differences; a crossover trial will show both between-treatment and between-patient differences; but only a repeated-period crossover trial (patients randomised to sequences of treatments) will show between-treatment and between-patient differences, and patient-by-treatment interaction. That is, to show interaction at the individual level, we need replication at the level at which the interaction is claimed.[18] Patient–treatment interactions (if identified) would provide the upper bound of gene–treatment interaction and, unless the patient–treatment interaction is large, the gene–treatment interaction cannot be.[19]

Crossover and *n*-of-1 designs[20] are most suited to chronic and clinically stable conditions,[21] which may not be attainable in many clinical environments, including critical care. Not surprisingly, there have been attempts to derive patient–treatment interactions (or individual qualitative interactions) by analytical techniques. These attempts include using (i) a potential outcomes framework which delineates observable and non-observable heterogeneity, in which the upper and lower bounds of non-observable heterogeneity are estimated from the observable data,[22,23] and (ii) the overlap between the distribution of results from treatment and control groups or the proportion of similar response.[24,25]

## Personalised medicine

The potential for personalised medicine in 1997 seemed boundless[3] (the human genome project began in October 1990[26]) but, 10 years later, the biotech revolution was being described more cautiously:

> Rather than producing revolutionary changes, medicinal biotechnology is following a well-established pattern of slow and incremental technology diffusion[27]

and "there is as yet little evidence that biotechnology provides a more rapid route from bench to clinic".[28] By 2012, personalised medicine remained "a contested vision of the future",[11] and some wondered what had happened to it.[29] Pharmacogenomics was to be an area of promise; "Getting the right drug into the right patient"[30] and key "omics" technologies (genomics, proteomics and metabolomics) were to be used to analyse individual variations to therapeutic drugs.[31] An empirical study of market drugs with pharmacogenomic biomarker data in their labels ("either measurable DNA and/or RNA characteristics in inherited genotypes or proteins involved in oncogenesis that indicate a likely response to therapeutic intervention") suggested that such information had little clinical use.[32] A comprehensive review of pharmacogenetics similarly found that the "application of pharmacogenetics to clinical medicine cannot adequately predict drug response in individual patients".[33] Nebert and colleagues have made analogous cautionary comments on the variable genetic contribution of a gene to the overall variance of a quantitative trait (the level of drug or metabolite) and the impossibility of assigning a patient to an unequivocal phenotype (and especially an unequivocal genotype).[12,26,34] We repeat a caution on the variation in response to pharmaceuticals: clinical triallists have assumed that genetics have shown the inevitability of such variation, and geneticists have assumed that clinical trials have shown it.[35]

At the societal level, the potential of genomic personalised medicine entails a double jeopardy[31] — that is, the channelling of extreme levels of human and material resources (an example of the inverse care law[36]) and the potential for treatment of ever more healthy populations (the inverse benefit law[37]). It would be imprudent to retreat into nihilism; instead, an integrated systems biology approach (embracing complexity) is a plausible goal,[38] as recently applied to the prediction of sepsis outcome.[39]

## Randomised clinical trials: problems in analysis?

The 2010 Consolidated Standards of Reporting Trials (CONSORT) guidelines for reporting parallel group RCTs,[40] and an elaboration on their extension to non-pharmacological treatment,[41] provide brief comments on statistical methods (see Section 12 of both documents),

particularly the suggestion that clustering effects (broadly defined) have more relevance in non-pharmacological treatments (Table 3, page 304[41]). A commentary on the CONSORT 2010 guidelines documents the effect of clustering in parallel RCTs regardless of the treatment comparison, and suggests that clustering is relevant to sample size calculation, statistical methods and outcome interpretation.[42] From the perspective of pharmaceutical trials, a document from the International Conference on Harmonisation also recommends that "the main treatment effect may be investigated first using a model which allows for centre differences" (page 13[43]).

Individually randomised patients in multicentre trials being potentially subject to clustering may be a function of similar patients being present in a cluster, or of clusters being determinant of outcomes. The proportion of total variability explained by between-cluster or between-centre variability is reflected in the intraclass correlation coefficient (ICC). The ICC is defined (on the odds ratio [OR] scale) as $\sigma^2 \div (\sigma^2 + \pi^2/3)$, in which $\sigma^2$ is the between-centre variance. Non-ignorable clustering occurs when both the ICC and the correlation between treatment assignments within a cluster are non-zero.[44] In the absence of accounting for a centre effect, the standard error (SE) of the treatment effect is increased by $(1–ICC)^{-½}$ and, for a large ICC, for example 0.3, a loss in power may be substantial (up to 15%).[45] Cook and colleagues presented 45 ICCs from surgical trials and found that 42% were > 0.05, and 16% were > 0.20.[46] Empirical studies, including trial reanalyses, have documented the existence and effect of clustering at various levels, and surveys of trial reporting suggest that this effect has not been recognised or suitably acted upon.[45,47-51] Similarly, methods of balancing within strata and within centres, such as the use of permuted blocks and minimisation, lead to correlation between treatment groups. In the absence of accounting for prognostic balancing variables in analysis, these methods also lead to upward biasing of treatment-effect SE, with consequently low type I error rates and a decrease in power;[52,53] that is, balancing invalidates an unadjusted analysis.

## How should we adjust for clustering or centre effects?

In the biopharmaceutical literature, a seminal 1986 article by Fleiss addressed the question of information pooling with randomisation schedules which either considered centres by design or ignored them.[54] For random treatment assignment carried out separately and independently within centres, pooling or averaging within-centre differences was correct, as analysis was dictated by design. For randomisation unrelated to centre, pooling by "lumping the data together" is "theoretically possible" but "should generally be avoided" and it would be "advisable to summarise the data first within [centres] and then to average the treatment differences across the [centres]". An intensive debate ensued, with its origins in prior agricultural research (see Appendix).

Fleiss also considered analytical treatment of centre effects as either fixed effects (FEs) or random effects (REs). He favoured FEs as reflecting the then consensus, and on the basis of avoiding complicated analysis. A brisk response by Grizzle pointed to recent advances in computing, even in 1987, and further suggested that:

> Although the clinics are not randomly chosen, the assumption of random clinic effect will result in tests and confidence intervals that capture the variability inherent in the system more realistically than when clinic effects are considered fixed … the assumption of what is considered fixed and what is random is really a statement about how the covariance structure should be modeled for the experiment.[55]

A more recent expression of this approach is found in an article by Fedorov and Jones, who state that:

> the random-effects model is a parsimonious way of accounting for within-centre and between-centre variation [and] provides a useful way of describing typical data from a multicentre trial.[56]

In the medical literature, a highly cited article by Localio and colleagues was forthright in its recommendation:

> Multicenter designs also warrant testing and adjustment for the potential bias of confounding by center, and for the presence of effect modification or interaction by center.[57]

This was reiterated by Feaster and colleagues,[58] who concluded that ignoring site effects in multisite clinical trials is "not a viable option". For binary outcomes, Agresti and Hartzel found little difference in treatment effect estimates between FE and RE models with large stratum-specific sample sizes and in the absence of treatment–centre interaction. Including the latter would offer the safest approach with many strata and sparse data.[59]

In a trial reanalysis and simulation study, Kahan[44] compared various estimators to account for the centre effect (FE, RE, generalised estimating equations [GEE] and the Mantel–Haenszel method). With a small number of centres, all estimators performed adequately, but with a large number of centres, only the RE and GEE (with non-robust SE) gave nominal type I error rates and acceptable power. Note, however, that the OR effect of these two estimators (RE, a conditional model; and GEE, a marginal model) may differ in the presence of a treatment effect (marginal OR, closer to the null and an increased difference [$OR_{marginal}$ versus $OR_{conditional}$] as the ICC increases). There is also a need for clarification regarding the implications of an RE analysis. If centres are treated as REs (with no treatment–centre

interaction), there will be a gain in efficiency, especially with substantial centre patient number disparity (ie, a smaller SE as the between-centre information component is included in analysis).[58,60] The use of an RE analysis is often invoked on the basis of a presumed generalisability of results to non-participating centres. However, as pointed out by Senn,[60] this implies a random treatment–centre interaction model (with larger SE) in which there is treatment variability across centres and the treatment effect tests whether, on average, treatment is better than control.[58] As Gallo observed, this may involve a "tendency towards less significant results", but the end result is that the interaction would be handled in a systematic manner.[61] Further, the conventional RE meta-analytic approach models the treatment–trial interaction, not the main effect of the trial, as random.[57,60]

## Clinical trials: failures and responses

Although the critical care community may lament the failure of so many RCTs in sepsis, such disappointments are not unusual (eg, those in oncology).[62] It is instructive to contrast two overviews of RCTs from cancer oncology[63] and adult critical care.[64] Both reviews report unrealistic estimates of treatment effects by triallists in the context of poor overall success rates. Surprisingly, in oncology, the estimate was 37.5% ($n$ = 253) and in adult critical care, the estimate was 37% ($n$ = 146). According to a 2014 survey,[65] only 10.4% of all therapeutic agents entering Phase I development were approved by the US Food and Drug Administration (FDA). Successful development from Phase I to Phase II was 64.5%, and from Phase II to Phase III, 32.4%; from Phase III, success was 60.1%, and regulatory approval from Phase III was 83.2%. Overall, industry-wide productivity was thought to have decreased from previous estimates.

One response to the perceived lack of new therapeutics entering the market place has been a call for regulatory laws to be "updated to reflect patient heterogeneity in clinical trials, and allow for approval of drugs that show efficacy in only a subset of treated patients".[66] Another response has been a more focused search for novel and facilitatory trial designs, specifically flexible or adaptive design (AD),[67] in which adaptation refers to a change made in a trial or statistical procedure during the trial. These changes can be prospective (eg, adaptive randomisation, early stopping or sample size re-estimation), concurrent (eg, inclusion criteria modification or change in hypotheses or endpoints) or retrospective (eg, statistical analysis plan changes or unblinding treatment codes).[68] In ADs, triallists use accumulating data to guide ongoing trial modification without prejudicing the trial validity or integrity.[69] This developmental process may occur according to various imperatives and at different levels and time scales; for example, theoretical,[70] pharmaceutical,[69] regulatory[71,72] and investigator-initiated[73] adaptations. The slow uptake of AD by clinical triallists (at least in the United Kingdom) has recently been commented on.[74]

Criticisms have accompanied the ongoing development of the AD paradigm. The FDA has categorised AD variants as well understood (eg, conventional group sequential designs and adaptations to maintain study power based on blinded interim analysis of aggregate data) and less well understood (eg, adaptations for dose selection studies and adaptive randomisation). Statistical concerns identified in the less well understood ADs included preservation of type I and type II error rates; statistical bias in estimation of treatment effects; the role of clinical trial simulation in planning and evaluation ("using simulations to demonstrate control of type I error rate is, however, controversial and not fully understood"[71]); and the requirement for prospective analysis plans.[71] These generic disquiets have been robustly restated[75,76] and summarised by Chang.[77] Chow and Corey have further suggested that implementation of AD methods "should proceed with caution".[78]

In addition to the previously discussed commentary and RCT overviews,[1,63,64] AD has been advocated in two other current reviews.[79,80] However, given the diversity of ADs (Chow and Chang detail at least 10 categories[68]), it seems that ADs are being recommended in a rhetorical, rather than a substantive, manner, as a novel solution.[81] When some detail of AD has been provided, the generic definition of AD[79] appears to be that of an outcome-adaptive or response-adaptive randomisation design. This is described by the FDA as less well understood; it is not necessarily Bayesian[82] and is controversial.[83-86]

What, then, is the role of AD? The sentiments of the editorial that accompanies the article by Gan and colleagues[63] seem astute:

> we are at risk of losing our focus. Conducting larger trials … using adaptive trial designs are not the solutions … We do not need more marginal results that are then pronounced "new treatment paradigms" or a "new standard of therapy." What we need are meaningful goals and better drugs.[62]

## Competing interests

None declared.

## Author details

**John L Moran**, Associate Professor[1]

**Patricia J Solomon**, Professor of Statistical Bioinformatics[2]

1 Department of Intensive Care Medicine, The Queen Elizabeth Hospital, Adelaide, SA, Australia.

2 School of Mathematical Sciences, University of Adelaide, Adelaide, SA, Australia.

*Correspondence:* john.moran@adelaide.edu.au

## References


1 Marini JJ, Vincent JL, Annane D. Critical care evidence — new directions. *JAMA* 2015; 313: 893-4.

2 Flores L. Therapeutic inferences for individual patients. *J Eval Clin Pract* 2015; 21: 440-7.

3 Marshall A. Laying the foundations for personalized medicines. *Nat Biotechnol* 1997; 15: 954-7.

4 Longford NT. Selection bias and treatment heterogeneity in clinical trials. *Stat Med* 1999; 18: 1467-74.

5 Simon R. Patient heterogeneity in clinical trials. *Cancer Treat Rep* 1980; 64: 405-10.

6 Venook AP, Arcila ME, Benson AB III, et al. NCCN Working Group Report: designing clinical trials in the era of multiple biomarkers and targeted therapies. *J Natl Compr Canc Netw* 2014; 12: 1629-49.

7 Marshall JC. Why have clinical trials in sepsis failed? *Trends Mol Med* 2014; 20: 195-203.

8 Ward PA. What's new in the quagmire of sepsis? *Trends Mol Med* 2014; 20: 189-90.

9 Kaukonen K, Bailey M, Suzuki S, et al. Mortality related to severe sepsis and septic shock among critically ill patients in Australia and New Zealand, 2000–2012. *JAMA* 2014; 311: 1308-16.

10 Ioannidis JPA, Karassa FB, Druyts E, et al. Biologic agents in rheumatology: unmet issues after 200 trials and $200 billion sales. *Nat Rev Rheumatol* 2013; 9: 665-73.

11 Tutton R. Personalizing medicine: futures present and past. *Soc Sci Med* 2012; 75: 1721-8.

12 Nebert DW, Zhang G, Vesell ES. From human genetics and genomics to pharmacogenetics and pharmacogenomics: past lessons, future directions. *Drug Metab Rev* 2008; 40: 187-224.

13 Smith GD. Epidemiology, epigenetics and the 'gloomy prospect': embracing randomness in population health research and practice. *Int J Epidemiol* 2011; 40: 537-62.

14 Byar DP, Corle DK. Selecting optimal treatment in clinical-trials using covariate information. *J Chronic Dis* 1977; 30: 445-59.

15 Weiss CO, Segal JB, Varadhan R. Assessing the applicability of trial evidence to a target sample in the presence of heterogeneity of treatment effect. *Pharmacoepidemiol Drug Saf* 2012; 21: 121-9.

16 Kaiser KA, Gadbury GL. Estimating the range of obesity treatment response variability in humans: methods and illustrations. *Hum Hered* 2013; 75: 127-35.

17 Senn S. Individual response to treatment: is it a valid assumption? *BMJ* 2004; 329: 966-8.

18 Senn S. Being efficient about efficacy estimation. *Stat Biopharm Res* 2013; 5: 204-10.

19 Senn S. Individual therapy: new dawn or false dawn? *Drug Inf J* 2001; 35: 1479-94.

20 Schork NJ. Time for one-person trials. *Nature* 2015; 520: 609-11.

21 Collette L, Tombal B. N-of-1 trials in oncology. *Lancet Oncol* 2015; 16: 885-6.

22 Poulson RS, Gadbury GL, Allison DB. Treatment heterogeneity and individual qualitative interaction. *Am Stat* 2012; 66: 16-24.

23 Zhang Z, Wang C, Nie L, Soon G. Assessing the heterogeneity of treatment effects via potential outcomes of individual patients. *J R Stat Soc Ser C Appl Stat* 2013; 62: 687-704.

24 Acion L, Peterson JJ, Temple S, Arndt S. Probabilistic index: an intuitive non-parametric approach to measuring the size of treatment effects. *Stat Med* 2006; 25: 591-602.

25 Rom DM, Hwang E. Testing for individual and population equivalence based on the proportion of similar responses. *Stat Med* 1996; 15: 1489-505.

26 Nebert DW. Given the complexity of the human genome, can "personalised medicine' or "individualised drug therapy" ever be achieved? *Hum Genomics* 2009; 3: 299-300.

27 Nightingale P, Martin P. The myth of the biotech revolution. *Trends Biotechnol* 2004; 22: 564-9.

28 Hopkins MM, Martin PA, Nightingale P, et al. The myth of the biotech revolution: an assessment of technological, clinical and organisational change. *Res Policy* 2007; 36: 566-89.

29 What happened to personalized medicine? [editorial] *Nat Biotechnol* 2012; 30: 1.

30 Marshall A. Getting the right drug into the right patient. *Nat Biotechnol* 1997; 15: 1249-52.

31 James JE. Personalised medicine, disease prevention, and the inverse care law: more harm than benefit? *Eur J Epidemiol* 2014; 29: 383-90.

32 Tutton R. Pharmacogenomic biomarkers in drug labels: what do they tell us? *Pharmacogenomics* 2014; 15: 297-304.

33 Shah RR, Shah DR. Personalized medicine: is it a pharmacogenetic mirage? *Br J Clin Pharmacol* 2012; 74: 698-721.

34 Nebert DW, Zhang G. Personalized medicine: temper expectations. *Science* 2012; 337: 910.

35 Senn S. Mastering variation: variance components and personalised medicine. *Stat Med* 2016; 35: 966-77.

36 Hart JT. The inverse care law. *Lancet* 1971; 1: 405-12.

37 Brody H, Light DW. The inverse benefit law: how drug marketing undermines patient safety and public health. *Am J Public Health* 2011; 101: 399-404.

38 Monte AA, Brocker C, Nebert DW, et al. Improved drug therapy: triangulating phenomics with genomics and metabolomics. *Hum Genomics* 2014; 8: 16.

39 Langley RJ, Tsalik EL, van Velkinburgh JC, et al. An integrated clinico-metabolomic model improves prediction of death in sepsis. *Sci Transl Med* 2013; 5: 195ra95.

40 Moher D, Hopewell S, Schulz KF, et al. CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. *J Clin Epidemiol* 2010; 63: e1-37.

41 Boutron I, Moher D, Altman DG, et al. Extending the CONSORT statement to randomized trials of nonpharmacologic treatment: explanation and elaboration. *Ann Intern Med* 2008; 148: 295-309.

42 Cals JW, van Amelsvoort LG, Kotz D, Spigt MG. CONSORT 2010 Statement — unfinished update? *J Clin Epidemiol* 2011; 64: 579-82.

43 International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use.

International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use: ICH harmonised tripartite guideline E9, Step 4: statistical principles for clinical trials. Geneva: ICH, 1998. http://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Efficacy/E9/Step4/E9_Guideline.pdf (accessed Jun 2016).

44 Kahan BC. Accounting for centre-effects in multicentre trials with a binary outcome — when, why, and how? *BMC Med Res Methodol* 2014; 14: 20.

45 Kahan BC, Morris TP. Assessing potential sources of clustering in individually randomised trials. *BMC Med Res Methodol* 2013; 13: 58.

46 Cook JA, Bruckner T, MacLennan GS, Seiler CM. Clustering in surgical trials — database of intracluster correlations. *Trials* 2012; 13: 2.

47 Biau DJ, Halm JA, Ahmadieh H, et al. Provider and center effect in multicenter randomized controlled trials of surgical specialties: an analysis on patient-level data. *Ann Surg* 2008; 247: 892-8.

48 Lee KJ, Thompson SG. Clustering by health professional in individually randomised trials. *BMJ* 2005; 330: 142-4.

49 Biau DJ, Porcher Rl, Boutron I. The account for provider and center effects in multicenter interventional and surgical randomized controlled trials is in need of improvement: a review. *J Clin Epidemiol* 2008; 61: 435-9.

50 Tangri N, Kitsios GD, Su SH, Kent DM. Accounting for center effects in multicenter trials. *Epidemiology* 2010; 21: 912-3.

51 Oltean H, Gagnier JJ. Use of clustering analysis in randomized controlled trials in orthopaedic surgery. *BMC Med Res Methodol* 2015; 15: 17.

52 Kahan BC, Morris TP. Improper analysis of trials randomised using stratified blocks or minimisation. *Stat Med* 2012; 31: 328-40.

53 Kahan BC, Morris TP. Reporting and analysis of trials using stratified randomisation in leading medical journals: review and reanalysis. *BMJ* 2012; 345: e5840.

54 Fleiss JL. Analysis of data from multiclinic trials. *Control Clin Trials* 1986; 7: 267-75.

55 Grizzle JE. Analysis of data from multiclinic trials [letter to the editor]. *Control Clin Trials* 1987; 8: 392-3.

56 Fedorov V, Jones B. The design of multicentre trials. *Stat Methods Med Res* 2005; 14: 205-48.

57 Localio AR, Berlin JA, Ten Have TR, Kimmel SE. Adjustments for center in multicenter studies: an overview. *Ann Intern Med* 2001; 135: 112-23.

58 Feaster DJ, Mikulich-Gilbertson S, Brincks AM. Modeling site effects in the design and analysis of multi-site trials. *Am J Drug Alcohol Abuse* 2011; 37: 383-91.

59 Agresti A, Hartzel J. Strategies for comparing treatments on a binary response with multi-centre data. *Stat Med* 2000; 19: 1115-39.

60 Senn S. A note regarding "random effects". *Stat Med* 2014; 33: 2876-7.

61 Gallo PP. Center-weighting issues in multicenter clinical trials. *J Biopharm Stat* 2000; 10: 145-63.

62 Amiri-Kordestani L, Fojo T. Why do Phase III clinical trials in oncology fail so often? *J Natl Cancer Inst* 2012; 104: 568-9.

63 Gan HK, You B, Pond GR, Chen EX. Assumptions of expected benefits in randomized Phase III trials evaluating systemic treatments for cancer. *J Natl Cancer Inst* 2012; 104: 590-8.

64 Harhay MO, Wagner J, Ratcliffe SJ, et al. Outcomes and statistical power in adult critical care randomized trials. *Am J Respir Crit Care Med* 2014; 189: 1469-78.

65 Hay M, Thomas DW, Craighead JL, et al. Clinical development success rates for investigational drugs. *Nat Biotechnol* 2014; 32: 40-51.

66 Embracing patient heterogeneity [editorial]. *Nat Med* 2014; 20: 689.

67 Pong A, Chow SC. Handbook of adaptive designs in pharmaceutical and clinical development. Boca Raton, Fla: CRC Press, 2011.

68 Chow SC, Chang M. Adaptive design methods in clinical trials — a review. *Orphanet J Rare Dis* 2008; 3: 11.

69 Gallo P, Chuang-Stein C, Dragalin V, et al. Adaptive designs in clinical drug development — an executive summary of the PhRMA Working Group. *J Biopharm Stat* 2006; 16: 275-83.

70 Bauer P, Bretz F, Dragalin V, et al. Twenty-five years of confirmatory adaptive designs: opportunities and pitfalls. *Stat Med* 2015; 35: 325-47.

71 Food and Drug Administration. Guidance for industry: adaptive design clinical trials for drugs and biologics, draft guidance. Silver Spring, Md: FDA, 2010. http://www.fda.gov/downloads/Drugs/../Guidances/ucm201790.pdf (accessed Jun 2016).

72 European Medicines Agency. Reflection paper on methodological issues in confirmatory clinical trials with flexible design and analysis plan. London: EMA, 2007. http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500003617.pdf (accessed Jun 2016).

73 Dulhunty JM, Starr T, Bellomo R, Lipman J. Randomised controlled trials: the long hard climb to the summit — is there another way in the 21st century? *Crit Care Resusc* 2014; 16: 87-9.

74 Dimairo M, Boote J, Julious SA, et al. Missing steps in a staircase: a qualitative study of the perspectives of key stakeholders on the use of adaptive designs in confirmatory trials. *Trials* 2015; 16: 430.

75 Burman CF, Sonesson C. Are flexible designs sound? *Biometrics* 2006; 62: 664-9.

76 Emerson SS, Fleming TR. Adaptive methods: telling "the rest of the story". *J Biopharm Stat* 2010; 20: 1150-65.

77 Chang AM. Adaptive trial design. In: Modern issues and methods in biostatistics. New York: Springer Science+Business Media, 2011: 87-112.

78 Chow S-C, Corey R. Benefits, challenges and obstacles of adaptive clinical trial designs. *Orphanet J Rare Dis* 2011; 6: 79.

79 Opal SM, Dellinger RP, Vincent JL, et al. The next generation of sepsis clinical trial designs: what is next after the demise of recombinant human activated protein C? *Crit Care Med* 2014; 42: 1714-21.

80 Cohen J, Vincent JL, Adhikari NKJ, et al. Sepsis: a roadmap for future research. *Lancet Infect Dis* 2015; 15: 581-614.

81 Gallo P. Good practices for adaptive clinical trials. In: Pong A, Chow SC, editors. Handbook of adaptive designs in pharmaceutical and clinical development. Boca Raton, Fla: CRC Press, 2011: chapter 27: 1-12.

82 Hu F, Hu Y, Ma W, et al. Statistical inference of adaptive randomized clinical trials for personalized medicine. *Clin Invest* 2015; 5: 415-25.

83 Korn EL, Freidlin B. Outcome-adaptive randomization: is it useful? *J Clin Oncol* 2011; 29: 771-6.

84 Korn EL, Freidlin B. Are outcome-adaptive allocation trials ethical? [commentary] *Clin Trials* 2015; 12: 122-4.

85 Lee JJ, Chen N, Yin G. Worth adapting? revisiting the usefulness of outcome-adaptive randomization. *Clin Cancer Res* 2012; 18: 4498-507.

86 Thall P, Fox P, Wathen J. Statistical controversies in clinical research: scientific and ethical problems with adaptive randomization in comparative clinical trials. *Ann Oncol* 2015; 26: 1621-8.

87 Herr DG. On the history of ANOVA in unbalanced, factorial designs: the first 30 years. *Am Stat* 1986; 40: 265-70.

88 Senn S. Some controversies in planning and analysing multi-centre trials. *Stat Med* 1998; 17: 1753-65.

89 Landsheer JA, van den Wittenboer G. Unbalanced 2 x 2 factorial designs and the interaction effect: a troublesome combination. *PLoS One* 2015; 10: e0121412.

90 Hector A, von Felten S, Schmid B. Analysis of variance with unbalanced data: an update for ecology & evolution. *J Anim Ecol* 2010; 79: 308-16.

91 Langsrud Y. ANOVA for unbalanced data: use Type II instead of Type III sums of squares. *Stat Computing* 2003; 13: 163-7.

92 Jones B, Teather D, Wang J, Lewis JA. A comparison of various estimators of a treatment difference for a multi-centre clinical trial. *Stat Med* 1998; 17: 1767-77.

93 Kallen A. Treatment-by-center interaction: what is the isuue? *Drug Inf J* 1997; 31: 927-36.

94 Worthington H. Methods for pooling results from multi-center studies. *J Dent Res* 2004; 83: C119-C121.

95 Schwemer G. General linear models for multicenter clinical trials. *Control Clin Trials* 2000; 21: 21-9.

96 Lin ZN. An issue of statistical analysis in controlled multi-centre studies: how shall we weight the centres? *Stat Med* 1999; 18: 365-73.

97 Senn S. Multicenter trials. In: Statistical issues in drug development, 2nd ed. Chichester, UK: John Wiley & Sons; 2007: 213-33.

98 Peduzzi P, Henderson W, Hartigan P, Lavori P. Analysis of randomized controlled trials. *Epidemiol Rev* 2002; 24: 26-38. ❑

## Appendix

After publication of the article by Fleiss,[54] the debate was formulated initially in terms of normally distributed (continuous) data, factorial analysis of variance (ANOVA[87]) for unbalanced designs (in this case, variable site patient number, the norm for multicentre trials[88]) and the type of sums of squares.[61] For unbalanced designs, in which explanatory variables may also become correlated (non-orthogonal) because of unequal numbers of subjects in various groups,[89] the values of the (sequential) sums of squares for individual variables depend on the ordering of the variables in the fitted ANOVA model (type I sum of squares[90]), although the total sum of squares is invariant. A sensible analytical strategy to assess the effect of a variable would entail first fitting all other effects.[88] Under conditions of centre–treatment interaction, the decision may be made to fit interactive effects initially and also weight the centre treatment effects equally (type III [adjusted] sums of squares), or to exclude the interaction effect and weight centre treatment effects by precision (type II [adjusted] sums of squares). In this circumstance, adjusted sums of squares do not depend on model term ordering.[91] Problematic issues identified were:

- the power of the test for treatment-by-centre interaction was low, so operational *P* levels were suggested at 0.1[54,92,93]
- the status of small *n* centres, which may result in loss of power and precision.

Under a full model (type III) when no interaction was actually present or identified, a loss of power may result,[94] and pre-specified site-pooling algorithms were recommended, such that the sample size ratio of largest to smallest centre should be < 2:1.[95] Several commentators have favoured type II sums of squares;[61,88,91,92,96-98] others type III;[93,95] and some have been more circumspect, recommending dependence on specific analytical circumstances, mainly the possibility of treatment–centre interaction.[89,90,94] Thus, the question of using type II or type III sums of squares appears to be unresolved. We note the conclusion of a current investigation of unbalanced factorial designs and interaction effects: "The translation from the simulations in this study towards individual cases of 2 x 2 datasets is complex".[89]