



RESEARCH ARTICLE

Does evidence support the high expectations placed in precision medicine? A bibliographic review [version 1; peer review: 2 approved with reservations, 1 not approved]

Jordi Cortés ¹, José Antonio González¹, María Nuncia Medina², Markus Vogler³, Marta Vilaró⁴, Matt Elmore¹, Stephen John Senn⁵, Michael Campbell⁶, Erik Cobo¹

¹Department of Statistics and Operations Research, Universitat Politècnica de Catalunya, Barcelona, 08034, Spain

²Escuela Colombiana de Ingeniería Julio Garavito, Bogotá, 111211, Colombia

³Department of Statistics, Ludwig-Maximilians-Universität München, München, 80539, Germany

⁴Fundació lliga per a la investigació i prevenció del càncer, Reus, 43201, Spain

⁵Competence Center for Methodology and Statistics, Luxembourg Institute of Health, Strassen, 1445, Luxembourg

⁶School of Health and Related Research, University of Sheffield, Sheffield, S1 4DA, UK

v1 **First published:** 09 Jan 2018, 7:30 (<https://doi.org/10.12688/f1000research.13490.1>)
Second version: 13 Jun 2018, 7:30 (<https://doi.org/10.12688/f1000research.13490.2>)
Third version: 15 Nov 2018, 7:30 (<https://doi.org/10.12688/f1000research.13490.3>)
Latest published: 07 Mar 2019, 7:30 (<https://doi.org/10.12688/f1000research.13490.4>)

Abstract

Background: Precision medicine is the Holy Grail of interventions that are tailored to a patient’s individual characteristics. However, the conventional design of randomized trials assumes that each individual benefits by the same amount.

Methods: We reviewed parallel trials with quantitative outcomes published in 2004, 2007, 2010 and 2013. We collected baseline and final standard deviations of the main outcome. We assessed homoscedasticity by comparing the outcome variability between treated and control arms.

Results: The review provided 208 articles with enough information to conduct the analysis. At the end of the study, 113 (54%, 95% CI 47 to 61%) papers find less variability in the treated arm. The adjusted point estimate of the mean ratio (treated to control group) of the outcome variances is 0.89 (95% CI 0.81 to 0.97).

Conclusions: Some variance inflation was observed in just 1 out of 6 interventions, suggesting the need for further eligibility criteria to tailor precision medicine. Surprisingly, the variance was more often smaller in the intervention group, suggesting, if anything, a reduced role for precision medicine.

Homoscedasticity is a useful tool for assessing whether or not the premise of constant effect is reasonable.

Keywords

Constant Effect, Precision medicine, Homoscedasticity, Clinical Trial, Variability, Standard deviation, Review



Open Peer Review

Referee Status: XXX ??

	Invited Referees				
	1	2	3	4	5
version 4 published 07 Mar 2019				?	?
				report	report
version 3 published 15 Nov 2018		X			
		report			
version 2 published 13 Jun 2018		X		X	
		report		report	
version 1 published 09 Jan 2018		?	X	?	
		report	report	report	

1 **Ian R. White** , University College London, UK

2 **Erica E.M. Moodie** , McGill University, Canada

- 3 **Saskia le Cessie** , Leiden University Medical Center, The Netherlands
Leiden University Medical Center, The Netherlands
- 4 **Richard Stevens**, University of Oxford, UK
David Nunan , University of Oxford, UK
- 5 **Vance W. Berger**, National Cancer Institute, USA

Any reports and responses or comments on the article can be found at the end of the article.

Corresponding author: Jordi Cortés (jordi.cortes-martinez@upc.edu)

Author roles: **Cortés J:** Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **González JA:** Conceptualization, Formal Analysis, Methodology, Supervision, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Medina MN:** Conceptualization, Data Curation, Investigation, Writing – Original Draft Preparation, Writing – Review & Editing; **Vogler M:** Data Curation, Investigation, Validation, Writing – Review & Editing; **Vilaró M:** Data Curation, Investigation, Validation, Writing – Review & Editing; **Elmore M:** Writing – Original Draft Preparation, Writing – Review & Editing; **Senn SJ:** Conceptualization, Methodology, Validation, Writing – Original Draft Preparation, Writing – Review & Editing; **Campbell M:** Conceptualization, Methodology, Validation, Writing – Original Draft Preparation, Writing – Review & Editing; **Cobo E:** Conceptualization, Funding Acquisition, Methodology, Project Administration, Supervision, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: Partially supported by Methods in Research on Research (MiRoR, Marie Skłodowska-Curie No. 676207); MTM2015-64465-C2-1-R (MINECO/FEDER); and 2014 SGR 464.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2018 Cortés J *et al.* This is an open access article distributed under the terms of the [Creative Commons Attribution Licence](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Cortés J, González JA, Medina MN *et al.* **Does evidence support the high expectations placed in precision medicine? A bibliographic review [version 1; peer review: 2 approved with reservations, 1 not approved]** F1000Research 2018, 7:30 (<https://doi.org/10.12688/f1000research.13490.1>)

First published: 09 Jan 2018, 7:30 (<https://doi.org/10.12688/f1000research.13490.1>)

Introduction

The idea behind precision medicine is to develop prevention and treatment strategies that take into account individual characteristics. With this strong endorsement “The prospect of applying this concept broadly has been dramatically improved by recent developments in large-scale biologic databases (such as the human genome sequence), powerful methods for characterizing patients (such as proteomics, metabolomics, genomics, diverse cellular assays, and mobile health technology), and computational tools for analyzing large sets of data.”, US President Obama launched the Precision Medicine initiative in 2015 to capitalize on these developments^{1,2}. However, we are not convinced that this is a sensible idea.

Variability of a clinical trial outcome measure should interest researchers because it conveys important information about whether there is a need for precision medicine. Does variance come only from unpredictable and ineluctable sources of patient variability? Or should it also be attributed to a different treatment effect that requires more precise prescription rules³⁻⁵? Researchers assess treatment effect modifications (“interactions”) among subgroups based on relevant variables. The main problem with that methodology is that, by the usual standards of classical phase III trial, the stratification factors must be known in advance and be measurable. This in turn implies that when new variables are discovered and introduced into the causal path, new clinical trials are needed. Fortunately, one observable consequence of a constant effect is that the treatment will not affect variability, and therefore the outcome variances in both arms should be equal (“homoscedasticity”). If this homoscedasticity holds, there is no need to repeat the clinical trial once a new possible effect modifier becomes measurable.

Nevertheless, the fundamental problem of causal inference is that for each patient in a parallel group trial, we can know the response for only one of the interventions. That is, we observe their response to either the new Treatment or to the Control, but not both. By experimentally controlling unknown confounders through randomization, a clinical trial may estimate the averaged causal effect. In order to translate this population estimate into effects for individual patients, additional assumptions are needed. We try to elucidate whether the comparison of observed variances may shed some light on the non-observable individual treatment effect. See examples and references that illustrate in their interpretation in [Figure 1⁶⁻¹⁵](#).

The assumption that the average effect equals the single unit effect underlies the rationale behind the usual sample size calculation, where only a single effect is specified. As an example, the 10 clinical trials published in the *Trials Journal* in October 2017 (see [Supplementary File 1 : Table S1](#)) were designed under this scenario of a fixed, constant or unique effect in the sample size calculation.

Our objectives were, first, to compare the variability of the main outcome between different arms in clinical trials published in medical journals and, second, to provide a first, rough estimate of the proportion of studies that could potentially benefit from precision medicine. As sensitivity analysis, we explore the changes in

the experimental arm’s variability over time (from baseline to the end of the study). We also fit a random effect model to the outcome variance ratio in order to isolate studies with a variance ratio outside their expected random variability values (heterogeneity).

Methods

Population

Our target population was parallel randomized clinical trials with quantitative outcomes. Trials needed to provide enough information to assess two homoscedasticity assumptions in the primary endpoint: between arms at trial end; and baseline to outcome over time in the treated arm. Therefore, baseline and final SDs for the main outcome were necessary or, failing that, at least one measure that would allow us to calculate them (variances, standard errors or mean confidence intervals).

Data collection

Articles on parallel clinical trials from the years 2004, 2007, 2010 and 2013 were selected from the Medline database with the following criteria: “*AB (clinical trial* AND random*) AND AB (change OR evolution OR (difference AND baseline)*” [The word “difference” was paired with “baseline” because the initial purpose of the data collection, subsequently modified, was to estimate the correlation between baseline and final measurements]. The rationale behind the election of these years was to have a global view of the behavior of the studies over a whole decade. For the years 2004 and 2007, we selected all papers that met the inclusion criteria; while for the years 2010 and 2013, as we obtained a greater number of articles retrieved from the search (478 and 653, respectively), we chose a random sample of 300 papers (Section II in [Supplementary File 1](#)).

Data were collected by two different researchers (NM, MkV) in two phases: 2004/2007 and 2010/2013. Later, two statisticians (JC, MtV) verified the correctness of the data and to make them accessible to readers through a [shiny application](#) and through the [Figshare repository](#)¹⁶.

Variables

Collected variables were: baseline and outcome SDs; experimental and reference interventions; sample size in each group; medical field according to *Web of Science* (WOS) classification; main outcome; patient’s disease; kind of disease (chronic or acute); outcome type (measured or scored); intervention type (pharmacological or not); and whether or not the main effect was significant.

For studies with more than one quantitative outcome, primary endpoint was determined according to the following hierarchical criteria: (1) objective or hypothesis; (2) sample size determination; (3) main statistical method; or (4) first quantitative variable reported in results.

In the same way, the choice of the “experimental” arm was determined depending on the role in the following sections of the article: (1) objective or hypothesis; (2) sample size determination; (3) rationale in the introduction or (4) first comparison reported in results (in the case of more than two arms)

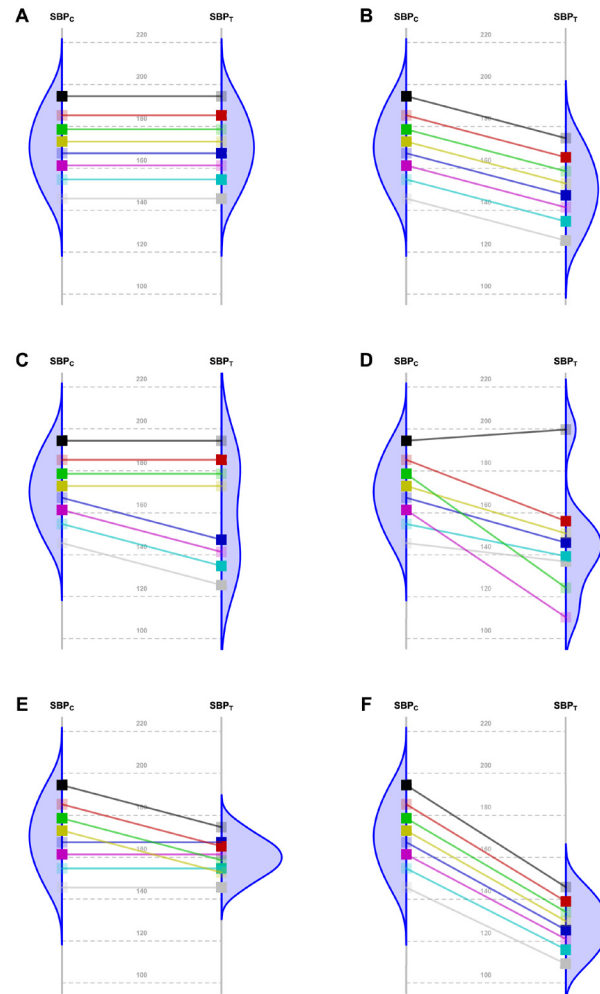


Figure 1. Scenarios representing fictional trials with 8 participants with Systolic Blood Pressure as the primary endpoint. Because of the random allocation to one of two treatment arms, we will only observe one of the two potential outcomes for each patient: either under T or under C. Fully saturated colors represent observed Systolic Blood Pressure (SBP) values, and transparent squares represent missing potential values. The line slope indicates the individual non-observable effect for each patient. Densities are the potential distributions of the outcome in each group: As both random samples come from the same target population, the average causal effect is estimable without bias.

Panel A shows the potential outcome values that we could obtain if there were not any treatment effect; as the intervention has no effect at all, both groups have the same distribution (i.e., mean and variance). **Panel B** shows the scenario of a constant effect, meaning that the intervention lowers the SBP by a single value in every patient, implying the same variability in both arms. For instance, the study from Duran-Cantolla *et al.*⁶ compared the 24-hour systolic blood pressure among 340 patients randomized to either Continuous Positive Airway Pressure (CPAP) or sham-CPAP, and it showed a greater decrease of 2.1 mmHg (95% CI from 0.4 to 3.7) in the intervention group compared to the control group. Furthermore, baseline standard deviations (SDs) were 12 and 11; and final SDs were 13 for both groups. Therefore, their results fully agree with the trial design's assumption of a constant effect (scenario B) and nothing contradicts the inference that each patient exhibits a constant reduction of 2.1mmHg, although the uncertainty of random allocation makes the results compatible with a constant effect that lies anywhere between 0.4 and 3.7. **Panel C** represents a situation with 2 different effects in 2 subpopulations ("treatment by subgroup interaction"). Although the effects are identical within them, the observable distribution in the treated arm would have higher variability. Here, we need to find finer eligibility criteria for classifying patients in those subpopulations so that a constant effect could be assumed again. In **Panel D**, the treatment has a variable effect in each patient, resulting also in greater variability within the treated arm but without any subgroup sharing a common effect, and results are poorly predictive about the effects on future patients. In the study by Kojima *et al.*⁷, the primary outcome measure was the 3-hour postprandial area under the curve of apolipoprotein B48, with outcome SDs being 0.78 and 0.16 in the treated and reference arms, respectively, thus showing a variance ratio of 23.77. This is compatible with different treatment effects that could need further refinements through precision medicine, since a greater variance in the treated arm indicates that "the interpretation of the main treatment effect is controversial"⁸. In that case, guidelines for treating new patients should be based either on additional eligibility criteria ("precision medicine", panel C) or on n-of-1 trials ("individualized medicine", panel D)⁹⁻¹³. This "treatment by patient interaction" was already highlighted by W. S. Gosset in the data of his 1908 paper proposing the Student t-distribution¹⁴. Alternatively, interactions can result in smaller variances in the treated arm. **Panel E** shows a different effect in 2 subgroups, but the variability is now reduced indicating that the best solution would be to identify the subpopulations in order to refine the selection criteria. In **Panel F**, the treatment has a stabilizing effect, with higher blood pressure falling more in severe patients, thus resulting in lower variability in the treatment arm. In the study from Kim *et al.*¹⁵, the primary endpoint was the PTSD Checklist-Civilian version (PCL-C). This scale is based on the sum of 17 Likert symptoms, ranging from 17 (perfect health) to 85 (worst clinical situation). At the end of the trial, the respective outcome SDs were 16 and 3 for the control and treated arms, meaning that variance was reduced around 28 times. This situation can correspond to scenarios E or F and merit much more statistical considerations, which is beyond the scope of this paper.

Statistical analysis

We assessed homoscedasticity between treatments and over time. Our main analysis compared, for the former, the outcome variability between Treated (T) and Control (C) arms at the trial end. For the latter, we compared the variability between Outcome (O) and its Baseline (B) value for the treated arm.

To distinguish between random variability and heterogeneity, we fitted a random mixed effects model using the logarithm of the variance ratio at the end of the trial as response with the study as random effect and the logarithm of the variance ratio at baseline as fixed effect¹⁷. An analogous model was employed to assess the homoscedasticity over time, as such a model allows the separation of random allocation variability from additional heterogeneity. To obtain a reference in the absence of treatment effect, we first modeled the baseline variance ratio as a response that is expected to have heterogeneity equal to 0 due to randomization – so long as no methodological impurities are present (e.g., consider the outcomes obtained 1 month after the start of treatment as the baseline values). This reference model allows us to know the proportion of studies in the previous models that could have additional heterogeneity which cannot be explained by the variability among studies (sections III, V and VI in [Supplementary File 1](#)).

Funnel plots, centered at zero, of the measurement of interest as a function of its standard errors are reported in order to investigate asymmetries.

As sensitivity analyses, we assessed homoscedasticity in each single study: (a) between outcomes on both arms with F-test for independent samples; and (b) between baseline and outcome in the treated arm with a specific test for paired samples¹⁸ when the variance of the paired difference was available. All tests were two-sided ($\alpha=5\%$).

Several subgroup analyses were carried out according to the statistical significance of the main effect of the study and to the different types of outcomes and interventions.

All analyses were performed with the [R statistical package](#) version 3.2.5. (The R code for the main analysis is available from <https://doi.org/10.5281/zenodo.1133609>¹⁹)

Results

Population

A total of 1214 articles were retrieved from the search. Of those papers, 542 belong to the target population and 208 (38.4%) contained enough information to conduct the analysis ([Figure 2](#)).

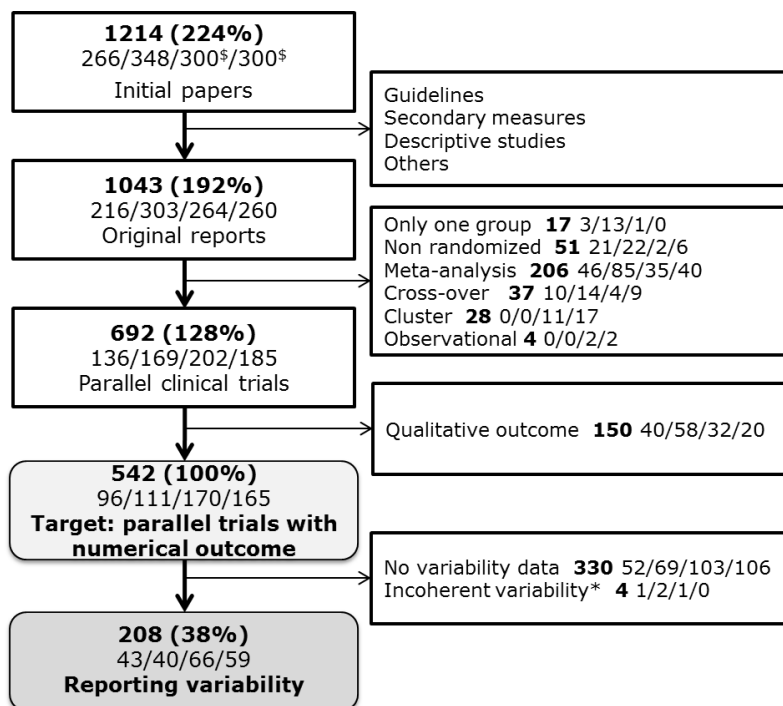


Figure 2. Flow-chart of the articles in the study. Percentages represent the quantity of papers in the target population. The number of articles for each year (2004/2007/2010/2013) is specified in the second line of each box (separated by slashes). *300 papers were randomly selected for years 2010 and 2013. *Four papers were excluded because the variance of the change over time was inconsistent with both the baseline and final variances, which would lead to impossible absolute correlation estimates greater than 1.

Mainly, the selected studies were non-pharmacological (122, 58.6%), referred to chronic conditions (101, 57.4%), had an outcome measure with units (132, 63.8%) instead of a constructed scale, and this outcome were measured rather than assessed (125, 60.1%). Regarding the primary objective of each trial, the authors found statistically significant differences between the arms in 83 (39.9%) studies. Following the WOS criteria, 203 articles (97.6%) belonged to at least one medical field. The main areas of study were: General & Internal Medicine (n=31, 14.9%), Nutrition & Dietetics (21, 10.1%), Endocrinology & Metabolism (19, 9.1%), and Cardiovascular System & Cardiology (16, 7.7%).

Homoscedasticity

There is a high average concordance between variances in the treatment and control arm, but with evidence of a smaller variability in the treated arm. At the end of the study, 113/208 (54%, 95% CI, 47 to 61%) papers showed less variability in the treated arm (Supplementary File 1 : Figure S1). Among the treated arms, 111/208 (53%, 95% CI, 46 to 60%) had less or equal variability at the end of follow-up than at the beginning (Supplementary File 1 : Figure S2).

We found statistically significant differences (at 5%) between outcome variances in 41 out of 208 (19.7%) studies: 7.2% were in favor of greater variance in the treated arm, and 12.5% in the opposite direction. A greater proportion was obtained from the comparisons over time of 95 treated arms: 16.8% had significantly greater variability at the end of the study and 23.2% at the beginning (Table 1).

Regarding the comparison between arms, the adjusted point estimate of the ratio (Treated to Control group) of the outcome variances is 0.89 (95% CI 0.81 to 0.97), indicating that treatments tend to reduce the variability of the patient’s response by about 11% on average. The comparison over time provides a similar result: the average variability at the end of the study is 14% lower than that at the beginning (Supplementary File 1 : Table S4).

According to the random model, the baseline heterogeneity was 0.31; this is a very high value which can only be explained by methodological flaws similar to those presented by Carlisle²⁰. Fortunately, the exclusion of the four most extreme papers reduced it to 0.07; one of them was the study by Hsieh *et al.*²¹ whose “baseline” values were obtained 1 month after the treatment started. When we modeled the outcome instead of the baseline variances as the response, heterogeneity was approximately doubled. We found 30 studies that compromised homoscedasticity (11 with higher variance in the treated arm and 19 with lower, Table 1). Figure 3 shows the funnel plots for both between-arm and over-time comparisons.

Subgroup analyses suggest that only significant interventions had an effect on reducing variability (Supplementary File 1 : Figures S3–S5), which has already been observed in other studies^{22,23} and in the line of other works that had found a positive correlation between the effect size and its heterogeneity^{24,25}: in fact, it is difficult to find heterogeneity when there is no overall treatment effect. The remaining subgroups analyses did not raise concerns (section V in Supplementary File 1).

Table 1. Variance comparison. Alternative possible methods to estimate the number and percentage of studies with different variances on comparisons between arms and over-time. Limits for declaring different variances come from different statistical methods; either masked specified statistical tests (F for independent outcomes or Sachs’ test¹⁸ for related samples); or sensitivity analysis about the number of studies that have to be deleted from the random mixed model in order to achieve a negligible heterogeneity (see Methods for details). [†] Only performed in studies reporting enough information to obtain the variability of the change from baseline to outcome, for example because they provide the correlation.

Comparing variances	N	Method	After treatment, variability is...		
			Increased n (%)	Decreased n (%)	Not changed n (%)
Outcome between treatment arms	208	F test	15 (7.2%)	26 (12.5%)	167 (80.3%)
		Random model	11 (5.3%)	19 (9.1%)	178 (85.6%)
Outcome versus baseline in treated arm	95 [†]	Paired test	16 (16.8%)	22 (23.2%)	57 (60.0%)
		Random model	13 (13.7%)	19 (20.0%)	63 (66.3%)

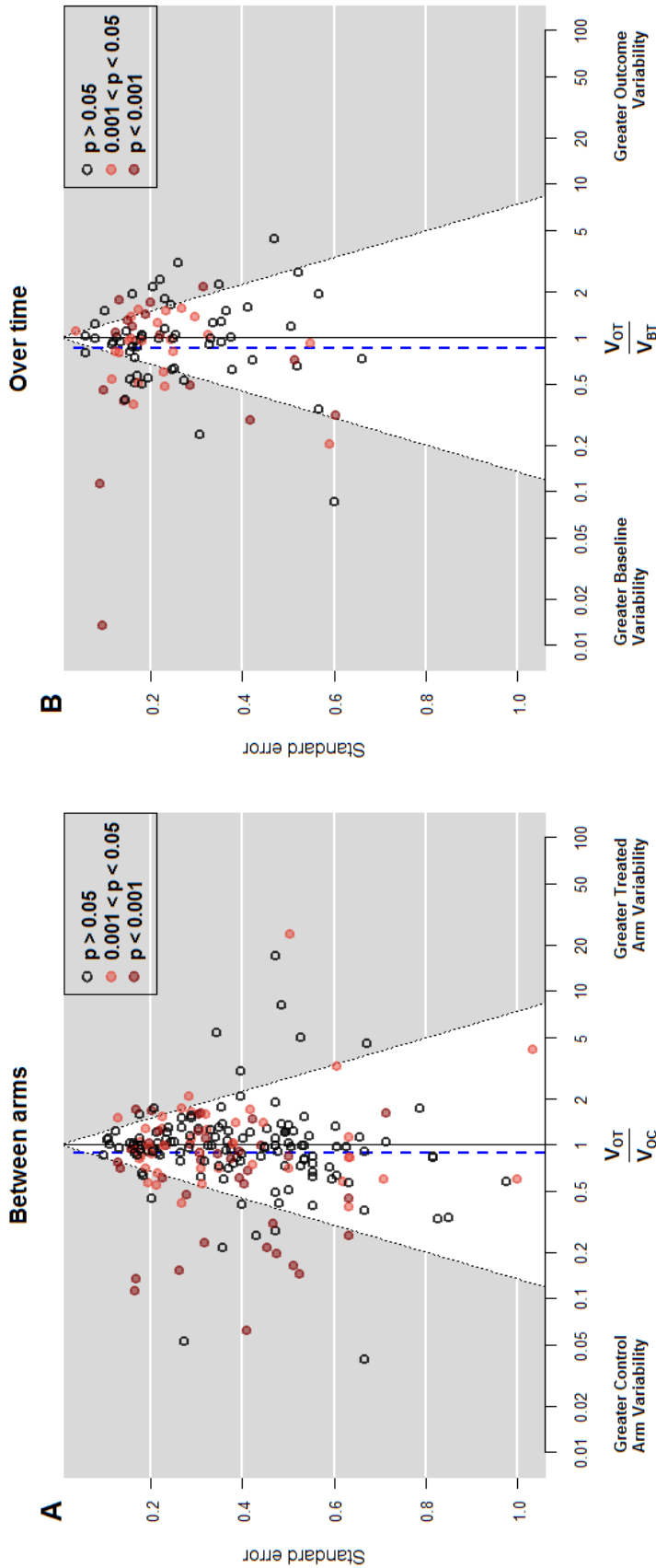


Figure 3. Funnel plots of variance ratio. Funnel plots of variance ratio between arms with the 208 studies (Panel A) and for over-time comparison with the 95 studies for which the variance of the difference between the basal and final response was available (Panel B). Vertical axis indicates precision for the comparison of variances; with points outside the triangle being statistically significant. Additionally, red points mark significant differences between the means, which correspond to the objective of each study to assess main treatment effects. In Panel A, points on the right indicate higher outcome variability for the treated individuals, as expected if there is patient-by-treatment interaction; similarly, points on the left correspond to lower variability, although this is compatible with traditional Evidence-Based Medicine. Eleven (5.2%) out of 208 studies reported exactly the same outcome variability in both arms. We observe more red points on the left, indicating that changes in the average, come with reductions in the variance. In Panel B, points on the right indicate higher variability in the experimental arm at the end of the study, as expected in a scenario of heterogeneous treatment effect; points on the left correspond to lower variability at the end, which implies a more homogenous response after treatment. The largest number of points on the left side indicates a majority of experimental interventions that reduce variability. In addition, several of these interventions yielded significant results in the main endpoint.

Discussion

Our main objective was to show that comparing variances can provide some evidence about how much precision medicine is needed. The variability seems to decrease for treatments that perform significantly better than the reference; otherwise, it remains similar. Therefore, contrary to popular belief, variability tends to be reduced on average after treatment, thus making precision medicine dispensable in most cases. This could be due to several reasons: some measurements may have “ceiling” or “floor” effects (e.g., in the extreme case, if a treatment makes a person well, no further improvement is possible); or the treatment may act proportionally rather than linearly, in which case the logarithm of the outcome would serve as a better scale.

When both arms have equal variances, then an obvious default explanation is that the treatment is equally effective for all, thus rendering the search for predictors of differential response futile. This means that treatment effects obtained by comparing the means between groups can be used to estimate both the averaged treatment effect and the non-observable patient treatment effect.

Furthermore, our second objective was to provide a rough estimate of the proportion of interventions that require a greater degree of precision medicine, and our answer is “not many”: considering the most extreme result from Table 1, we found that 1/6 interventions (16.8%) showed some variance inflation.

There are three reasons why these findings do not invalidate precision medicine in all settings. First, some additional heterogeneity is present in the outcome variances ratio, which indicates that the variability had increased between arms as well as over time. Second, the outcomes of some type of interventions such as surgeries, for example, are greatly influenced by the skills and training of those administering the intervention, and these situations could have some effect on increasing variability. And third, we focus on quantitative outcomes, which are neither time-to-event nor binary, meaning that the effect could take a different form, such as all-or-nothing.

The results rely on published articles, which raises some relevant issues. First, some of our analyses are based on Normality assumptions for the outcomes that are unverifiable without access to raw data. Second, a high number of manuscripts (61.6%, Figure 2) act contrary to CONSORT²⁶ advice in that they do not report variability. Thus, the results may be biased in either direction. Third, trials are usually powered to test constant effects and thus the presence of greater variability would lead to underpowered trials, non-significant results and unpublished papers. Fourth, the random effect model reveals additional heterogeneity in the outcome variances ratio, which may be the result of methodological inaccuracies²⁰ arising from typographical errors in data translation, inadequate follow-up, insufficient reporting, or even data fabrication. On the other hand, this heterogeneity could also be the result of relevant undetected factors interacting with the treatment, which would indeed justify the need for precision medicine. A fifth limitation is that many clinical trials are not completely randomized. For example, multicenter trials are often

blocked by center through the permuted blocks method. This means that if variances are calculated as if the trial were completely randomized (which is standard practice), the standard simple theory covering the random variation of variances from arm to arm is at best approximately true²².

The main limitation of our study arises from the fact that, although constant effect always implies homoscedasticity on the chosen scale, the reverse is not true; i.e., homoscedasticity does not necessarily imply a constant effect. Nevertheless, a constant effect is the simplest explanation for homoscedasticity. For example, the non-parsimonious situation reflected in Figure 4 indicates homoscedasticity but without a constant effect.

Heteroscedasticity may suggest the need for further refinements of the eligibility criteria or for finding an additive scale^{22,27}. Because interaction analyses cannot include unknown variables, all trials would potentially need to be repeated once any new potential interaction variable emerges (e.g., a new biomarker) as a candidate for a new subgroup analysis. Nevertheless, we have shown how homoscedasticity can be assessed when reporting trials with numerical outcomes, regardless of whether every potential effect modifier is known.

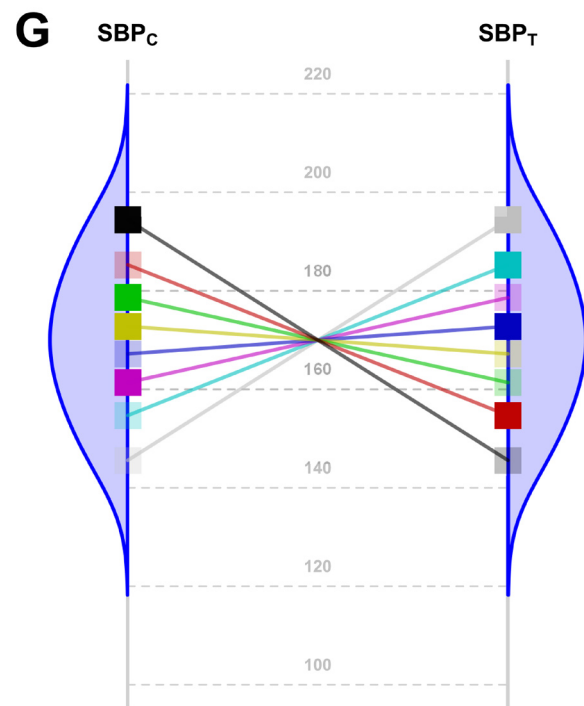


Figure 4. Scenario representing a fictional trial with 8 participants with homoscedasticity but non constant effect. SBP potential values of each patient in both groups (C: Control; T: Treated) under a highly hypothetical scenario: the treatment effect has no value if systematically applied to the whole population; but if n-of-1 trials could be performed in this situation, the best treatment strategy would be chosen for each patient and the overall health of the population would be improved.

For most trials, subjects vary little in their response to treatment, which suggests that precision medicine's scope may be less than what is commonly assumed. In the past century, Evidence-Based Medicine operated under the paradigm of a constant effect assumption, by which we learned from previous patients in order to develop practical clinical guides for treating future ones. Here, we have provided empirical insights for the rationale behind Evidence-Based Medicine. However, even where one common effect applies to all patients fulfilling the eligibility criteria, this does not imply the same decision is optimal for all patients, specifically because different patients and stakeholders may vary in their weighting not only of efficacy outcomes, but also of the harm and cost of the interventions – thus bridging the gap between common evidence and personalized decisions.

Nevertheless, in 16 trials of our sample, there was some evidence of variation arising from the treatment effect, suggesting a possible role for more tailored treatments: either with finer selection criteria (common effect within specific subgroups), or with n-of-1 trials (no subgroups of patients with a common effect). By identifying indications where the scope for precision medicine is limited, studies such as ours may free up resources for cases with a greater scope.

Our results uphold the assertion by Horwitz *et al.* that there is a “need to measure a greater range of features to determine [...] the response to treatment”²⁸. One of these features is an old friend of

statisticians: the variance. Looking only at averages can cause us to miss out on important information.

Data availability

Data is available through two sources:

- A shiny app that allows the user to interact with the data without the need to download it: http://shiny-eio.upc.edu/pubs/F1000_precision_medicine/

The Figshare repository: <https://doi.org/10.6084/m9.figshare.5552656>¹⁶

In both sources, the data can be downloaded under a Creative Commons License v. 4.0.

The code for the main analysis is available in the following link: <https://doi.org/10.5281/zenodo.1133609>¹⁹

Competing interests

No competing interests were disclosed.

Grant information

Partially supported by Methods in Research on Research (MiRoR, Marie Skłodowska-Curie No. 676207); MTM2015-64465-C2-1-R (MINECO/FEDER); and 2014 SGR 464.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Supplementary material

Supplementary File 1: The supplementary material contains the following sections:

[Click here to access the data.](#)

- Section I: Constant effect assumption in sample size rationale
- Section II: Bibliographic review
- Section III: Descriptive measures
- Section IV: Random effects models
- Section V: Subgroup analyses
- Section VI: Standard error of $\log(V_{OT}/V_{OC})$ in independent samples
- Section VII: Standard error of $\log(V_{OT}/V_{BT})$ in paired samples

References

1. Collins FS, Varmus H: **A new initiative on precision medicine.** *N Engl J Med.* 2015; **372**: 793–5.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
2. Kohane IS: **HEALTH CARE POLICY. Ten things we have to do to achieve precision medicine.** *Science.* 2015; **349**(6243): 37–8.
[PubMed Abstract](#) | [Publisher Full Text](#)
3. Schork NJ: **Personalized medicine: Time for one-person trials.** *Nature.* 2015; **520**(7549): 609–11.
[PubMed Abstract](#) | [Publisher Full Text](#)
4. Willis JC, Lord GM: **Immune biomarkers: the promises and pitfalls of personalized medicine.** *Nat Rev Immunol.* 2015; **15**(5): 323–29.
[PubMed Abstract](#) | [Publisher Full Text](#)

5. Wallach JD, Sullivan PG, Trepanowski JF, *et al.*: **Evaluation of Evidence of Statistical Support and Corroboration of Subgroup Claims in Randomized Clinical Trials.** *JAMA Intern Med.* 2017; **177**(4): 554–60.
[PubMed Abstract](#)
6. Durán-Cantolla J, Aizpuru F, Montserrat JM, *et al.*: **Continuous positive airway pressure as treatment for systemic hypertension in people with obstructive sleep apnoea: randomised controlled trial.** *BMJ.* 2010; **341**: c5991.
[PubMed Abstract](#) | [Publisher Full Text](#)
7. Kojima Y, Kaga H, Hayashi S, *et al.*: **Comparison between sitagliptin and nateglinide on postprandial lipid levels: the STANDARD study.** *World J Diabetes.* 2013; **4**(1): 8–13.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
8. International conference on harmonisation: **statistical principles for clinical trials ICH-E9.** 1998. Accessed September 14 2017.
[Reference Source](#)
9. Shamseer L, Sampson M, Bukutu C, *et al.*: **CONSORT extension for reporting N-of-1 trials (CENT) 2015: Explanation and elaboration.** *BMJ.* 2015; **350**: h1793.
[PubMed Abstract](#) | [Publisher Full Text](#)
10. Araujo A, Julious S, Senn S: **Understanding Variation in Sets of N-of-1 Trials.** *PLoS One.* 2016; **11**(12): e0167167.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
11. Senn S: **Individual response to treatment: is it a valid assumption?** *BMJ.* 2004; **329**(7472): 966–68.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
12. Senn S: **Mastering variation: variance components and personalised medicine.** *Stat Med.* 2016; **35**(7): 966–77.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
13. Wang R, Lagakos SW, Ware JH, *et al.*: **Statistics in medicine--reporting of subgroup analyses in clinical trials.** *N Engl J Med.* 2007; **357**(21): 2189–94.
[PubMed Abstract](#) | [Publisher Full Text](#)
14. Senn S, Richardson W: **The first t-test.** *Stat Med.* 1994; **13**(8): 785–803.
[PubMed Abstract](#) | [Publisher Full Text](#)
15. Kim SH, Schneider SM, Bevens M, *et al.*: **PTSD symptom reduction with mindfulness - based stretching and deep breathing exercise: randomized controlled clinical trial of efficacy.** *J Clin Endocr Metab.* 2013; **98**(7): 2984–92.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
16. Cortés J: **'review_homoscedasticity_clinical_trials'**. [Data set]. 2017.
[Publisher Full Text](#)
17. Bartlett MS, Kendall DG: **The statistical analysis of variance-heterogeneity and the logarithmic transformation.** *J R Stat Soc.* 1946; **8**(1): 128–38.
[Publisher Full Text](#)
18. Sachs L: **Applied Statistics: A Handbook of Techniques.** 2nd ed. New York: Springer-Verlag, 1984.
[Publisher Full Text](#)
19. Cortés J: **R code for analysis of homoscedasticity in clinical trials.** *Zenodo.* 2017.
[Publisher Full Text](#)
20. Carlisle JB: **Data fabrication and other reasons for non-random sampling in 5087 randomised, controlled trials in anaesthetic and general medical journals.** *Anaesthesia.* 2017; **72**(8): 944–952.
[PubMed Abstract](#) | [Publisher Full Text](#)
21. Hsieh LL, Kuo CH, Yen MF, *et al.*: **A randomized controlled clinical trial for low back pain treated by acupuncture and physical therapy.** *Prev Med.* 2004; **39**(1): 168–76.
[PubMed Abstract](#) | [Publisher Full Text](#)
22. Senn S: **controversies concerning randomization and additivity in clinical trials.** *Stat Med.* 2004; **23**(24): 3729–53.
[PubMed Abstract](#) | [Publisher Full Text](#)
23. Jamieson J: **Measurement of change and the law of initial values: A computer simulation study.** *Educ Psychol Meas.* 1995; **55**(1): 38–46.
[Publisher Full Text](#)
24. Senn S: **Trying to be precise about vagueness.** *Stat Med.* 2007; **26**(7): 1417–30.
[PubMed Abstract](#) | [Publisher Full Text](#)
25. Greenlaw N: **Constructing appropriate models for meta-analyses.** University of Glasgow, 2010. Accessed September 14, 2017.
[Reference Source](#)
26. Schulz KF, Altman DG, Moher D, *et al.*: **CONSORT 2010 Statement: updated guidelines for reporting parallel group randomised trials.** *BMJ.* 2010; **340**: c332.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
27. Rothman KJ, Greenland S, Walker AM: **Concepts of interaction.** *Am J Epidemiol.* 1980; **112**(4): 467–70.
[PubMed Abstract](#) | [Publisher Full Text](#)
28. Horwitz RI, Cullen MR, Abell J, *et al.*: **Medicine. (De)personalized medicine.** *Science.* 2013; **339**(6124): 1155–6.
[PubMed Abstract](#) | [Publisher Full Text](#)

Open Peer Review

Current Referee Status: ? X ?

Version 1

Referee Report 03 April 2018

<https://doi.org/10.5256/f1000research.14648.r31692>



Saskia le Cessie  1,2

¹ Department of Clinical Epidemiology, Leiden University Medical Center, Leiden, The Netherlands

² Section Medical Statistics, Department of Biomedical Data Science, Leiden University Medical Center, Leiden, The Netherlands

Review report. Does evidence support the high expectations placed in precision medicine? A bibliographic review. By J Cortés et al,

Summary: a review of randomised trials with continuous outcomes, measured at baseline and at follow up has been conducted. The aim was to compare the variance of the outcome measure at baseline and follow-up and to compare the variances at follow-up between the treated and the control group. The authors argue that a difference in variances may indicate a heterogeneous treatment effect.

General impression. The paper is well written and the results are of interest. The interpretation of the results is somewhat speculative but the authors discuss adequately the limitations.

Remarks

1. Abstract, Background : “However, the conventional design of randomized trials assumes that each individual benefits by the same amount.” This is not a correct statement. In a randomised trial, the average treatment effect in the population is estimated, and no assumptions about homogeneity of treatment effects are made here. The authors probably mean that many researchers implicitly assume a homogeneous treatment effect when conducting a randomised trial, interpreting the average treatment effect in the population as treatment effect at an individual level.
2. Introduction. I liked Figure 1 with the different explanations.
3. Methods and flow chart. The target population was parallel randomized clinical trials with quantitative/numerical outcomes. This is not true: trials with a survival time as outcome are also trials with a numerical (sometimes censored) outcome, but are not into the scope of your paper. So please mention that you are interested in trials with a numerical response variable which are measured both at baseline and at followup. In the Flow-chart, please check whether there were indeed 150 trials with a qualitative outcome, or whether there were 150 trials which did not satisfy the requirement of both having a baseline and a followup numerical measurement.
4. Statistical analysis. Here I got lost, the random mixed effects models should distinguish between random variability and heterogeneity, but how was unclear to me. Is adding the variance ratio at baseline needed to correct for the random variability? More details of the models and explanation of the different estimates of the model is needed, and should not only be given in the

supplementary material.

5. Did you compare the Var(change) between the treated and control group? Power to detect differences here would be larger.
6. It may be of interest to perform a subgroup analysis in the studies where control is placebo
7. Supplementary material, section 4. The model has two random effects: s_i , the heterogeneity between-study effect and e_i the within sample error with variance ν^2 . I guess this should be ν_i , as each study has its own within sample error variance, estimated from the sample sizes in the two groups (as described in the material)?
8. The supplementary material did not clearly describe which parameter(s) from the models reflected the heterogeneity. From the main text I derived that you used the mean effect μ to indicate the amount of heterogeneity. But then how to interpret the parameter τ ?
9. Supplementary Table S4. Why not put this Table in Section 4, and make one overview of all the models fitted? And I guess that e_{ij} should be e_i here.
10. Results: I did not find Figure S1 and Figure S2 very informative. Why not just give a histogram of $\log(\text{var}_{OT}/\text{var}_{CT})$ etc.
11. Table 1: How were the results from the random model obtained (the 11 increased, 19 decreased etc)?
12. Figure 3. Please explain what V_{OT} , V_{OC} etc is, as Figures should be self-explained.
13. I did not understand the second paragraph of the discussion. I guess that you want to say that the average treatment effect can be interpreted as an individual treatment effect, but I was confused at first by the words "non-observable patient treatment effect".
14. Shocking to see that so many studies do not report measures of variability.
15. The fourth limitation: "the random effect model reveals additional heterogeneity". To which result are you referring here, comparisons at baseline, followup or over time? The estimate of τ ? Why should this be the result of methodological inaccuracies?
16. Figure G is of interest because this is a situation where precision medicine is of interest: for some patients treatment T would be a better choice, for others treatment C and by performing precision medicine the subgroups with different responses could be detected and tailored prescriptions could be given. This indicates that observed homoscedasticity in a study should be interpreted with care and background knowledge of a study is needed to assess whether a situation as in Figure 4 is plausible.

Is the work clearly and accurately presented and does it cite the current literature?

Yes

Is the work clearly and accurately presented and does it cite the current literature?

Yes

Is the work clearly and accurately presented and does it cite the current literature?

Yes

Is the work clearly and accurately presented and does it cite the current literature?

Yes

Is the study design appropriate and is the work technically sound?

Yes

Is the study design appropriate and is the work technically sound?

Yes

Is the study design appropriate and is the work technically sound?

Yes

Is the study design appropriate and is the work technically sound?

Yes

Are sufficient details of methods and analysis provided to allow replication by others?

Yes

Are sufficient details of methods and analysis provided to allow replication by others?

Yes

Are sufficient details of methods and analysis provided to allow replication by others?

Yes

Are sufficient details of methods and analysis provided to allow replication by others?

Yes

If applicable, is the statistical analysis and its interpretation appropriate?

No

If applicable, is the statistical analysis and its interpretation appropriate?

No

If applicable, is the statistical analysis and its interpretation appropriate?

No

If applicable, is the statistical analysis and its interpretation appropriate?

No

Are all the source data underlying the results available to ensure full reproducibility?

Yes

Are all the source data underlying the results available to ensure full reproducibility?

Yes

Are all the source data underlying the results available to ensure full reproducibility?

Yes

Are all the source data underlying the results available to ensure full reproducibility?

Yes

Are the conclusions drawn adequately supported by the results?

Partly

Are the conclusions drawn adequately supported by the results?

Partly

Are the conclusions drawn adequately supported by the results?

Partly

Are the conclusions drawn adequately supported by the results?

Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Medical Statistics; Epidemiology; Methods for Observational studies

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 03 Jun 2018

Jordi Cortés, Universitat Politècnica de Catalunya, Spain

JOINT ANSWER to Ian White and Saskia le Cessie

This is a general response to Ian White and Saskia Le Cessie on why we stated that the standard clinical trial design and analysis assume a constant effect.

In the following, (1) we update the standard sample size rationale; and (2) we explain why inflated variances may require precision medicine in just two general cases: (a) interaction, as represented in Fig. 1, panel C; and (b) random treatment effect, Fig. 1, panel D.

1. Under the Neyman-Pearson framework to determine sample size, a single effect size value Δ is specified under the alternative hypothesis H1, assuming in that way a constant effect, as in Fig. 1, panels A (H0) and B (H1).
2. We devise two situations that, because they result in higher variance, they would need personalized medicine:
 - Interaction between treatment and a baseline variable such as, for example, gender (Fig. 1, panel C). In this scenario there are two subpopulations (e.g., men and women) with different treatment effects that require the effect to be made further “precise”.
 - Random treatment effect on each patient (Fig. 1, panel D). In this scenario, the effect size does not depend on a known patient baseline characteristic

and the only way to estimate the individual patient effect is by means of individualized trials (“n of 1” trials).

Those 2 hypothetical scenarios, lead to an increased variance. Conversely, scenarios E and F represent two similar situations (interaction and random effect) but result in reduced variance –without relevant changes on the average. Although we agree that in those two last scenarios leading to reduced variability the specific patient treatment effect may still be unknown because the outcome has reduced variability with a similar central overall position, we argue that patients in those situations were subject to “further control” (having more stable values within the boundaries of “normality”). So, the usual sample size rationale specified by statisticians in trials assumes a constant, unique effect that agrees with the clinical and legal interpretation that the effect is the same – or at least similar enough to be considered homogeneous – for all the patients fulfilling the eligibility criteria.

To illustrate this secondary “argument”, we reviewed the sample size rationale for the last (at that time) 10 protocols published in Trials, and we found that all of them defined a single effect size (100%, two-sided 95% confidence interval from 69% to 100%). In addition, we have included a new column in Table S1 with the main analysis showing that the SAP in all those cases (10 out of 10, 95%CI from 69 to 100%) was also designed to estimate a single, constant effect.

We have modified Fig. 1 (panels E and F) to show decreasing variance treatment effects, but now without affecting the average. We have also improved the 2 following sentences:

Before [Abstract]: However, the conventional design of randomized trials assumes that each individual benefits by the same amount.

After [Abstract]: However, conventional clinical trials are designed to find differences with the implicit assumption that the effect is the same in all patients within the eligibility criteria.

Before [Introduction]: The assumption that the average effect equals the single unit effect underlies the rationale behind the usual sample size calculation, where only a single effect is specified. As an example, the 10 clinical trials published in the Trials Journal in October 2017 (see Supplementary material: Table S1) were designed under this scenario of a fixed, constant or unique effect in the sample size calculation.

After [Introduction]: The assumption of homoscedasticity in the usual calculations of sample size is better interpreted under the constant effect model (Figure 1, panels A, H_0 ; and B, H_1). As an example, the 10 clinical trials published in the Trials Journal in October 2017 (Table S1 of Supplementary material) were designed with only a constant for the effect size. Furthermore, all their analyses were designed to test (and estimate) a single constant for the effect size. In other words, there was mention of neither any possible interaction with baseline variables (Figure 1, scenarios C and E), nor of any random variability for the treatment effect (Figure 1, scenarios D and F); and thus, all those trials were designed to test a constant effect.

We have also updated the legend of Figure 1 to highlight that now panels C to F show only possible individual treatment effects on variances but not on means.

We are deeply grateful to Ian White and Saskia le Cessie for highlighting the need to clarify this crucial issue.

Saskia le Cessie

From here, we'll answer specific issues

Summary: a review of randomised trials with continuous outcomes, measured at baseline and at follow up has been conducted. The aim was to compare the variance of the outcome measure at baseline and follow-up and to compare the variances at follow-up between the treated and the control group. The authors argue that a difference in variances may indicate a heterogeneous treatment effect.

General impression. The paper is well written and the results are of interest. The interpretation of the results is somewhat speculative but the authors discuss adequately the limitations.

Remarks

We are grateful to Prof. Saskia le Cessie for her suggestions, which definitively help us to improve our manuscript.

1. Abstract, Background : "However, the conventional design of randomized trials assumes that each individual benefits by the same amount." This is not a correct statement. In a randomised trial, the average treatment effect in the population is estimated, and no assumptions about homogeneity of treatment effects are made here. The authors probably mean that many researchers implicitly assume a homogeneous treatment effect when conducting a randomised trial, interpreting the average treatment effect in the population as treatment effect at an individual level.

Yes, our impression is that at least some trialists are not aware of these assumptions. But the fact that we wanted to highlight is that trials are usually designed to provide evidence for just one parameter (in our context the "effect size" collected by the difference of means) without further specification, neither in the sample size rationale nor in the analysis of the further parameters required by precision medicine. We have addressed this point in the joint answer above.

2. Introduction. I liked Figure 1 with the different explanations.

Thank you. Please note that we have now updated panels C to F to isolate changes just in variance.

3. Methods and flow chart. The target population was parallel randomized clinical trials with quantitative/numerical outcomes. This is not true: trials with a survival time as outcome are also trials with a numerical (sometimes censored) outcome, but are not into the scope of your paper. So please mention that you are interested in trials with a numerical response variable which are measured both at baseline and at follow-up. In the Flow-chart, please check whether there were

indeed 150 trials with a qualitative outcome, or whether there were 150 trials which did not satisfy the requirement of both having a baseline and a follow-up numerical measurement.

Thanks. We fully agree that discussion was introduced too late, and we have further clarified it in the Methods section and in the flow chart. The modifications are described below.

Before [Methods]: Our target population was parallel randomized clinical trials with quantitative outcomes

After [Methods]: Our target population was parallel randomized clinical trials with quantitative outcomes (not including time-to-event studies)

Before [Flow chart]: Qualitative outcome

After [Flow chart]: Categorical or time-to-event outcome

4. Statistical analysis. Here I got lost, the random mixed effects models should distinguish between random variability and heterogeneity, but how was unclear to me. Is adding the variance ratio at baseline needed to correct for the random variability? More details of the models and explanation of the different estimates of the model is needed, and should not only be given in the supplementary material.

Thanks. The model includes the (logarithm of the) baseline variances ratio because some imbalances in the initial variability between groups (after randomization) can occur simply by chance. It is foreseeable that these baseline differences in variability may influence the final differences in variability. This baseline log-ratio was highly significant ($p < 0.0001$) in the model.

All your suggestions related to the statistical analysis (4, 7, 8, 9 and 11) and the random effects model have been addressed through a clearer and longer explanation of the model in the statistical analysis section (detailed [here](#) and in the manuscript)

Nevertheless, we provide the following rule of thumb for interpreting the parameters μ (heteroscedasticity) and I^2 (heterogeneity) of the random-effects model.

**$\mu < 0$ --> On average, studies have lower variability in the experimental arm.
 $\mu > 0$ --> On average, studies have greater variability in the experimental arm.**

**$I^2 < 25\%$ --> As the point estimate of heterogeneity is not high enough, μ is constant throughout all the studies.
 $I^2 > 25\%$ --> As the point estimate of heterogeneity is high, μ does not apply to every single study.**

**$\mu < 0$ & $I^2 < 25\%$ --> Not one study requires precision medicine.
 $\mu < 0$ & $I^2 > 25\%$ --> Some studies require precision medicine.
 $\mu > 0$ & $I^2 < 25\%$ --> All studies require precision medicine.**

$\mu > 0$ & $I^2 > 25\%$ --> Most studies require precision medicine.

[The threshold of 25% for I^2 is based on PRISMA Statement [1] that considers values under this cutpoint as low.]

The estimates of these parameters in our data were $\mu = -0.12$ and $I^2 = 80.8\%$, which implies that some studies require precision medicine.

1. **Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gøtzsche P, et al. (2009) The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: Explanation and elaboration. PLoS Med 6: e1000100.**

5. Did you compare the Var(change) between the treated and control group? Power to detect differences here would be larger.

Strongly agree. For high correlations between baseline and outcome, it follows that $V(\log(V_{Ox}/V_{Bx})) < V(\log(V_{Ox}))$, as can be seen in Appendix VII of the supplementary material. However, just 95 out of 208 studies provide the Var(change) or the baseline-final correlation that would allow this analysis.

6. It may be of interest to perform a subgroup analysis in the studies where control is placebo

In fact, we performed this subgroup analysis beforehand without obtaining relevant results. We decided not to include or mention it because the distinction between a treatment called "placebo" and an "active" treatment is not clear. "Placebo" is defined as a simulator of the experimental treatment that tries to emulate its characteristics; but in some studies "control" may equal "best medical treatment", which is also provided to "treated" patients, such that "Placebo" is complemented by the standard intervention. Because of this ambiguity in the classification, we decided to omit this information. As illustrative examples, we mention the following included studies:

- Ghaleiha A, Mohammadi E, Mohammadi M, et al. Riluzole as an adjunctive therapy to risperidone for the treatment of irritability in children with autistic disorder: a double-blind, placebo-controlled, randomized trial. Paediatr Drugs 2013 15:505–514. The patients in the reference group took placebo in addition to risperidone (titrated up to 2 or 3 mg/day based on bodyweight) for 10 weeks.

- Carroll MW, Jeon D, Mountz JM, et al. Efficacy and safety of metronidazole for pulmonary multidrug-resistant tuberculosis. Antimicrob Agents Chemother 2013; 57:3903-9. The patients in the reference group took placebo for 8 weeks in addition to an individualized background regimen.

7. Supplementary material, section 4. The model has two random effects: s_i , the heterogeneity between-study effect and e_i the within sample error with variance ν^2 . I guess this should be ν_i , as each study has its own within sample error variance, estimated from the sample sizes in the two groups (as described in the material)?

Thank you. We have corrected the typo including the subscript both in the Methods section and in the supplementary material: “nu_i”

8. The supplementary material did not clearly describe which parameter(s) from the models reflected the heterogeneity. From the main text I derived that you used the mean effect μ to indicate the amount of heterogeneity. But then how to interpret the parameter τ ?

Thanks. Tau reflects the heterogeneity in the assessment of the heteroscedasticity throughout the studies. Following this suggestion and similar comment of Professor Ian White, we have tried to clarify that μ is a measure of heteroscedasticity and τ is a measure of the heterogeneity of the former throughout all the studies. See also the answer to question 4 for more clarification.

9. Supplementary Table S4. Why not put this Table in Section 4, and make one overview of all the models fitted? And I guess that e_{ij} should be e_i here.

Thank you, we have corrected the subscript typo: “e_i”

And yes, your suggestion facilitates readability. We have interspersed all the tables and figures of the supplementary material in their respective sections.

10. Results: I did not find Figure S1 and Figure S2 very informative. Why not just give a histogram of $\log(\text{var}_{OT}/\text{var}_{CT})$ etc.

We have kept Figures S1 and S2 because we believe that they provide additional information about whether or not the increase (or decrease) in the variability in the outcome of the experimental arm depends on the outcome variability of the control arm (or on the baseline variability of the experimental group). However, we have also added the histograms you mention in order to summarize the essential information. The histograms can be seen [here](#) or in the Supplementary material.

11. Table 1: How were the results from the random model obtained (the 11 increased, 19 decreased etc)?

We have obtained them as the studies that had to be removed in order to obtain heterogeneity (i.e., τ) similar to the baseline (which we expect to be null by randomization). We have tried to clarify this point in the legend of the table:

“...or (2) number of studies that have to be deleted from the random-effects model in order to achieve a negligible heterogeneity (studies with more extreme outcome were removed one by one until achieving an estimated value of τ similar to the one obtained from the reference model. See Methods section for more details...)”

12. Figure 3. Please explain what V_{OT} , V_{OC} etc is, as Figures should be self-explained.

Thanks. We have included a legend in this figure explaining these abbreviations:

V_OT: Variance of the Outcome in the Treated arm

V_OC: Variance of the Outcome in the Control arm

V_BT: Variance of the Outcome at baseline in the Treated arm

13. I did not understand the second paragraph of the discussion. I guess that you want to say that the average treatment effect can be interpreted as an individual treatment effect, but I was confused at first by the words “non-observable patient treatment effect”.

You are right. We say “non-observable” for the fundamental problem of causal inference (both potential responses are not observable in the same patient), which avoids seeing the treatment effect at the individual level. We have clarified this point:

Before: This means that treatment effects obtained by comparing the means between groups can be used to estimate both the averaged treatment effect and the non-observable patient treatment effect.

After: This means that the average treatment effect can be interpreted as an individual treatment effect (not directly observable).

14. Shocking to see that so many studies do not report measures of variability.

Yes. It is really surprising that 61.6% of studies do not report the variability either at baseline or at the end of the study. Although CONSORT advises it, this guideline does not provide the historical data on this practice with which it can be compared.

15. The fourth limitation: “the random effect model reveals additional heterogeneity”. To which result are you referring here, comparisons at baseline, followup or over time? The estimate of tau? Why should this be the result of methodological inaccuracies?

We are referring to the main analysis: comparison between arms. Nevertheless, this sentence could be applied to all analyses. Heterogeneity among studies is measured by tau (see response to question 4).

We stated that methodological inaccuracies can be derived in the presence of heterogeneity. In an ideal scenario of constant treatment effect in all the studies, the only thing that could lead to heterogeneity in the model would be methodological inaccuracies such as those mentioned in the manuscript or in the referenced paper of Carlisle: transcription errors, insufficient follow-up time for being able to observe this constant effect, or the manipulation of the results in order to achieve greater impact.

16. Figure G is of interest because this is a situation where precision medicine is of interest: for some patients treatment T would be a better choice, for others treatment C and by performing precision medicine the subgroups with different responses could be detected and tailored

prescriptions could be given. This indicates that observed homoscedasticity in a study should be interpreted with care and background knowledge of a study is needed to assess whether a situation as in Figure 4 is plausible.

Fully agree, although this is a highly sophisticated scenario that we hope will not be viewed as a frequent scenario.

Of course, we think that personalized medicine has already been demonstrated to be effective in some areas. Our point is that unless those demonstrations exist, most interventions should be routinely administered to all patients fulfilling eligibility criteria.

Competing Interests: No competing interests were disclosed.

Referee Report 23 March 2018

<https://doi.org/10.5256/f1000research.14648.r31694>



Erica E.M. Moodie 

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

The authors have performed a review of a sample of clinical trials conducted every three years from 2004-2013 to examine whether there exists post-treatment heterogeneity in participants responses with premise that lack of heterogeneity suggests that precision medicine is not warranted.

While the question is one that should be asked. However, the study carried out is not suited to answering the question as it has been conducted in randomized trials where there is typically little heterogeneity. That is, the authors have performed a perfectly reasonable analysis that cannot answer the pertinent question. It is well known that randomized trials tend to be populated by homogenous population (more white, more male, etc.) – see, for example Oh et al. (2015) Diversity in Clinical and Biomedical Research: A Promise Yet to Be Fulfilled. PLoS Med 12(12): e1001918, Caplan & Friesen P (2017) Health disparities and clinical trial recruitment: Is there a duty to tweet? PLoS Biol 15(3): e2002040 and the references therein – or indeed many other papers on this topic. This may be in part a function of recruitment strategies and also by design, as trialists (particularly those testing new therapies) often determine inclusion criteria to target the (potentially homogeneous) segment of the population who might show the greatest response to the treatment. Thus, the authors have chosen to study a population that is likely to be homogeneous and not reflective of real-world clinical care. There are numerous examples covariate-tailored treatment algorithms, from the choice of hormonal therapies for women diagnosed with estrogen-receptor-positive, HER2-negative breast cancer to the choice of ACE inhibitors vs. calcium channel blockers for hypertension, that the authors choose to overlook as cases where we have learned about previous patients *with particular characteristics* to learn about future similar patients.

Is the work clearly and accurately presented and does it cite the current literature?

Partly

Is the work clearly and accurately presented and does it cite the current literature?

Partly

Is the work clearly and accurately presented and does it cite the current literature?

Partly

Is the work clearly and accurately presented and does it cite the current literature?

Partly

Is the study design appropriate and is the work technically sound?

Partly

Is the study design appropriate and is the work technically sound?

Partly

Is the study design appropriate and is the work technically sound?

Partly

Is the study design appropriate and is the work technically sound?

Partly

Are sufficient details of methods and analysis provided to allow replication by others?

Yes

Are sufficient details of methods and analysis provided to allow replication by others?

Yes

Are sufficient details of methods and analysis provided to allow replication by others?

Yes

Are sufficient details of methods and analysis provided to allow replication by others?

Yes

If applicable, is the statistical analysis and its interpretation appropriate?

Partly

If applicable, is the statistical analysis and its interpretation appropriate?

Partly

If applicable, is the statistical analysis and its interpretation appropriate?

Partly

If applicable, is the statistical analysis and its interpretation appropriate?

Partly

Are all the source data underlying the results available to ensure full reproducibility?

Yes

Are all the source data underlying the results available to ensure full reproducibility?

Yes

Are all the source data underlying the results available to ensure full reproducibility?

Yes

Are all the source data underlying the results available to ensure full reproducibility?

Yes

Are the conclusions drawn adequately supported by the results?

No

Are the conclusions drawn adequately supported by the results?

No

Are the conclusions drawn adequately supported by the results?

No

Are the conclusions drawn adequately supported by the results?

No

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Longitudinal data analysis, adaptive treatment strategies

I have read this submission. I believe that I have an appropriate level of expertise to state that I do not consider it to be of an acceptable scientific standard, for reasons outlined above.

Author Response 03 Jun 2018

Jordi Cortés, Universitat Politècnica de Catalunya, Spain

The referee's objections can be summarized in two points:

- (1) "There are numerous examples of covariate-tailored treatment algorithms."
- (2) "Randomized trials tend to be populated by homogenous populations", which in turn does not reflect a real population's existing variability.

We understand the reviewer's comments, but we disagree with the reviewer's conclusions:

(1) Our work does not intend to completely invalidate precision medicine. We are not stating that there are not "examples of covariate-tailored treatments"; rather that (Abstract): "We found that the outcome variance was more often smaller in the intervention group, suggesting that treated patients may end up pertaining more often to reference or normality values and thus would not require further precision medicine". This was already stated in the discussion: "these findings do not invalidate precision medicine in all settings." Thus in the quite wide settings of our trials, we found little evidence that precision medicine would be of any use.

(2) The referee argues that trials have "too many" selection criteria to reflect "a real population". This is a standard criticism of explanatory clinical trials, suggesting that the selection criteria are usually "too many". And we agree, because our point is that most trials have "enough" selection criteria to provide a homogeneous effect. Furthermore, we also agree that we can only talk about "published trials with eligibility criteria". As for whether those selection criteria should be used to

define the target population in clinical guidelines, there is no further need to tailor precision medicine.

Dr. Moodie argues that our results do not answer the question that is posed. We also disagree. The issue of heterogeneity is obviously one that bedevils the generalizability of clinical trials. However, these are randomized comparisons; so, in the absence of a treatment effect we would expect the two arms to be comparable, no matter how heterogeneous the underlying population. The fact that even in the presence of a treatment effect there was little evidence of heterogeneity suggests there will be little scope for precision medicine in these populations. One might argue that with a more heterogeneous population, there is more scope to detect the few non-responders who would not form part of a general trial population. This does not invalidate our results; rather it argues for much larger trials with a more heterogeneous population. The point is that in the absence of these the evidence base for precision medicine is weak.

Competing Interests: No competing interests were disclosed.

Referee Report 02 March 2018

<https://doi.org/10.5256/f1000research.14648.r30604>



Ian R. White 

Medical Research Council Clinical Trials Unit, Institute of Clinical Trials and Methodology, University College London, London, UK

This paper considers randomised trials (RCTs) of treatment versus control with a quantitative outcome. It observes that *if* treatment effects are homogeneous (the same for all trial participants) *then* the outcome variance will be the same in both trial arms. It therefore reviews the extent to which the outcome variance is the same across trial arms in 208 published RCTs. It finds 41 RCTs with significant differences in outcome variance, and that it is more common for the outcome variance to be smaller in the treatment arm rather than larger.

My overall comment is that the analysis results are useful, but they need to be made clearer, and the interpretation should be much more cautious. Points marked * must be addressed to make the article scientifically sound (with ** the top priority).

Background

1. *Abstract, background: "The conventional design of randomized trials assumes that each individual benefits by the same amount." This is also asserted elsewhere in the paper, but it is not true. From a causal inference perspective, a RCT estimates the average causal effect, which is well defined in the presence of treatment effect heterogeneity. This is why the trials community worries so much about external generalisability: for example, if a trial treated 60% women and 40% men and showed a benefit of treatment, then a clinician treating women and men in the same ratio can be confident of giving a benefit overall, but a clinician treating women and men in a different ratio cannot be so confident. This point (repeated elsewhere) is not essential to the paper's argument, so should be removed.
2. *Similarly, the argument "The assumption that the average effect equals the single unit effect underlies the rationale behind the usual sample size calculation, where only a single effect is specified" (Introduction) is false. Sample size calculations relate only to comparisons of group averages.

Methods

The methods used appear entirely appropriate. However they are not well described.

1. *Terminology must be improved. For example, the key outcome in this study is the ratio of variances between treatment and control arms, and this (or its opposite) is variously called “homoscedasticity”, “heterogeneity”, even “concordance”. The authors should choose a term and stick with it. Similarly for the “random mixed effects model” which later becomes the “random model”. (I’m going to use “homoscedasticity” and “random-effects model”.)
2. The authors are doing a meta-analysis, even though they don’t call it that, so the term “heterogeneity” should be reserved for “variation between studies”, i.e. τ^2 in the random effects models.
3. *It’s not clear to me what the “random model” results in Table 1 are. Since this is a model across studies, how can it count individual studies? If empirical Bayes estimates of study-specific effects are being tested, this must be explained.
4. Trials that are “significant” are combined - “Subgroup analyses suggest that only significant interventions had an effect on reducing variability” - but interventions that increase the mean should be separated from those that decrease the mean. The later conclusion that “The variability seems to decrease for treatments that perform significantly better than the reference” suggests a different distinction (better/worse is not the same as larger/smaller because outcomes may be positive or negative) and is not supported by the results presented.
5. Abstract, Results: “The adjusted point estimate of the mean ratio (treated to control group) of the outcome variances” is not clear without reading the whole text. Again, defining a term (“outcome variance ratio”?) will help.
6. *Table 1, “variability is... increased”: from the text, this means “significantly increased”, which should be clarified.

Interpretation

The results may be interpreted in many ways, which are sensibly discussed by the authors. Most importantly, treatment effect homogeneity implies homoscedasticity, but the converse (“homoscedasticity implies treatment effect homogeneity”) is not true: this is demonstrated very nicely in Figure 4.

Homoscedasticity is scale-dependent: for example, it may be removed (or created) by a log transformation (mentioned in the Discussion).

1. *The authors omit one alternative explanation of homoscedasticity over time: clinical trial populations have eligibility criteria at baseline which may limit baseline variance. For example, a hypertension trial might recruit patients with baseline SBP between 140 and 159 mm Hg. In this case, variance is very likely to naturally increase over time.
2. **The authors’ conclusions ignore the alternative interpretations noted above. Here are some examples which are illogical:
 - Abstract, Conclusions: “the variance was more often smaller in the intervention group, suggesting, if anything, a reduced role for precision medicine”, and Discussion: “variability tends to be reduced on average after treatment, thus making precision medicine dispensable in most cases”. This is actually false. If a study finds smaller variance in the treated group then we DO have evidence of treatment effect heterogeneity, and indeed the treatment may be doing exactly what medicine should do - making the sickest better while not harming the less sick.
 - Introduction: “If this homoscedasticity holds, there is no need to repeat the clinical trial once a new possible effect modifier becomes measurable” - again, this wrongly assumes the converse stated above.

- Discussion: “When both arms have equal variances, then an obvious default explanation is that the treatment is equally effective for all, thus rendering the search for predictors of differential response futile”: this is illogical.
 - Discussion: “For most trials, subjects vary little in their response to treatment, which suggests that precision medicine’s scope may be less than what is commonly assumed.” : this is also illogical.
3. *In the light of the above arguments, I find the statement (Abstract, Conclusions) that “Homoscedasticity is a useful tool for assessing whether or not the premise of constant effect is reasonable” to be highly debatable. Logic suggests it gives a lower bound on the extent of usefulness of precision medicine, and the results of this study do not add any more to this.
 4. *The objectives in the Discussion should be the same as those stated in the Introduction.

Source data

1. I had trouble opening the source data both in Excel (since the csv file is in fact semi-colon-delimited) and in Stata (which was thrown by line 80). Could it be provided in a more convenient format or with some notes?

Is the work clearly and accurately presented and does it cite the current literature?

Yes

Is the work clearly and accurately presented and does it cite the current literature?

Yes

Is the work clearly and accurately presented and does it cite the current literature?

Yes

Is the work clearly and accurately presented and does it cite the current literature?

Yes

Is the study design appropriate and is the work technically sound?

Yes

Is the study design appropriate and is the work technically sound?

Yes

Is the study design appropriate and is the work technically sound?

Yes

Is the study design appropriate and is the work technically sound?

Yes

Are sufficient details of methods and analysis provided to allow replication by others?

Yes

Are sufficient details of methods and analysis provided to allow replication by others?

Yes

Are sufficient details of methods and analysis provided to allow replication by others?

Yes

Are sufficient details of methods and analysis provided to allow replication by others?

Yes

If applicable, is the statistical analysis and its interpretation appropriate?

No

If applicable, is the statistical analysis and its interpretation appropriate?

No

If applicable, is the statistical analysis and its interpretation appropriate?

No

If applicable, is the statistical analysis and its interpretation appropriate?

No

Are all the source data underlying the results available to ensure full reproducibility?

Yes

Are all the source data underlying the results available to ensure full reproducibility?

Yes

Are all the source data underlying the results available to ensure full reproducibility?

Yes

Are all the source data underlying the results available to ensure full reproducibility?

Yes

Are the conclusions drawn adequately supported by the results?

No

Are the conclusions drawn adequately supported by the results?

No

Are the conclusions drawn adequately supported by the results?

No

Are the conclusions drawn adequately supported by the results?

No

Competing Interests: No competing interests were disclosed.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 03 Jun 2018

Jordi Cortés, Universitat Politècnica de Catalunya, Spain

JOINT ANSWER to Ian White and Saskia le Cessie

This is a general response to Ian White and Saskia Le Cessie on why we stated that the standard clinical trial design and analysis assume a constant effect.

In the following, (1) we update the standard sample size rationale; and (2) we explain why inflated variances may require precision medicine in just two general cases: (a) interaction, as represented in Fig. 1, panel C; and (b) random treatment effect, Fig. 1, panel D.

- 1. Under the Neyman-Pearson framework to determine sample size, a single effect size value Δ is specified under the alternative hypothesis H1, assuming in that way a constant effect, as in Fig. 1, panels A (H0) and B (H1).**
- 2. We devise two situations that, because they result in higher variance, they would need personalized medicine:**
 - Interaction between treatment and a baseline variable such as, for example, gender (Fig. 1, panel C). In this scenario there are two subpopulations (e.g., men and women) with different treatment effects that require the effect to be made further “precise”.**
 - Random treatment effect on each patient (Fig. 1, panel D). In this scenario, the effect size does not depend on a known patient baseline characteristic and the only way to estimate the individual patient effect is by means of individualized trials (“n of 1” trials).**

Those 2 hypothetical scenarios, lead to an increased variance. Conversely, scenarios E and F represent two similar situations (interaction and random effect) but result in reduced variance –without relevant changes on the average. Although we agree that in those two last scenarios leading to reduced variability the specific patient treatment effect may still be unknown because the outcome has reduced variability with a similar central overall position, we argue that patients in those situations were subject to “further control” (having more stable values within the boundaries of “normality”).

So, the usual sample size rationale specified by statisticians in trials assumes a constant, unique effect that agrees with the clinical and legal interpretation that the effect is the same – or at least similar enough to be considered homogeneous – for all the patients fulfilling the eligibility criteria.

To illustrate this secondary “argument”, we reviewed the sample size rationale for the last (at that time) 10 protocols published in Trials, and we found that all of them defined a single effect size (100%, two-sided 95% confidence interval from 69% to 100%). In addition, we have included a new column in Table S1 with the main analysis showing that the SAP in all those cases (10 out of 10, 95%CI from 69 to 100%) was also designed to estimate a single, constant effect.

We have modified Fig. 1 (panels E and F) to show decreasing variance treatment effects, but now without affecting the average. We have also improved the 2 following sentences:

Before [Abstract]: However, the conventional design of randomized trials assumes that each individual benefits by the same amount.

After [Abstract]: However, conventional clinical trials are designed to find differences with the implicit assumption that the effect is the same in all patients within the eligibility criteria.

Before [Introduction]: The assumption that the average effect equals the single unit effect underlies the rationale behind the usual sample size calculation, where only a single effect is specified. As an example, the 10 clinical trials published in the *Trials Journal* in October 2017 (see Supplementary material: Table S1) were designed under this scenario of a fixed, constant or unique effect in the sample size calculation.

After [Introduction]: The assumption of homoscedasticity in the usual calculations of sample size is better interpreted under the constant effect model (Figure 1, panels A, H_0 ; and B, H_1). As an example, the 10 clinical trials published in the *Trials Journal* in October 2017 (Table S1 of Supplementary material) were designed with only a constant for the effect size. Furthermore, all their analyses were designed to test (and estimate) a single constant for the effect size. In other words, there was mention of neither any possible interaction with baseline variables (Figure 1, scenarios C and E), nor of any random variability for the treatment effect (Figure 1, scenarios D and F); and thus, all those trials were designed to test a constant effect.

We have also updated the legend of Figure 1 to highlight that now panels C to F show only possible individual treatment effects on variances but not on means.

We are deeply grateful to Ian White and Saskia le Cessie for highlighting the need to clarify this crucial issue.

Ian White

From here, we'll answer specific issues

This paper considers randomised trials (RCTs) of treatment versus control with a quantitative outcome. It observes that *if* treatment effects are homogeneous (the same for all trial participants) *then* the outcome variance will be the same in both trial arms. It therefore reviews the extent to which the outcome variance is the same across trial arms in 208 published RCTs. It finds 41 RCTs with significant differences in outcome variance, and that it is more common for the outcome variance to be smaller in the treatment arm rather than larger.

My overall comment is that the analysis results are useful, but they need to be made clearer, and the interpretation should be much more cautious. Points marked * must be addressed to make the article scientifically sound (with ** the top priority).

We are grateful to Prof. Ian White for his suggestions, which will definitively help us to improve our manuscript.

Background

1.*Abstract, background: "The conventional design of randomized trials assumes that each

individual benefits by the same amount.” This is also asserted elsewhere in the paper, but it is not true. From a causal inference perspective, a RCT estimates the average causal effect, which is well defined in the presence of treatment effect heterogeneity. This is why the trials community worries so much about external generalisability: for example, if a trial treated 60% women and 40% men and showed a benefit of treatment, then a clinician treating women and men in the same ratio can be confident of giving a benefit overall, but a clinician treating women and men in a different ratio cannot be so confident. This point (repeated elsewhere) is not essential to the paper’s argument, so should be removed.

2. *Similarly, the argument “The assumption that the average effect equals the single unit effect underlies the rationale behind the usual sample size calculation, where only a single effect is specified” (Introduction) is false. Sample size calculations relate only to comparisons of group averages.

Thanks again for highlighting this hugely important issue. We have addressed these two comments in the previous common answer.

Methods

The methods used appear entirely appropriate. However they are not well described.

1. *Terminology must be improved. For example, the key outcome in this study is the ratio of variances between treatment and control arms, and this (or its opposite) is variously called “homoscedasticity”, “heterogeneity”, even “concordance”. The authors should choose a term and stick with it. Similarly for the “random mixed effects model” which later becomes the “random model”. (I’m going to use “homoscedasticity” and “random-effects model”).

Thanks. To simplify the notation, we have deleted the term “concordance”. We also reserved the term heterogeneity for the τ^2 statistic resulting from the mixed-effects model (see next answer). Furthermore, we have homogenized the terms for referring to the “random-effects model” throughout the text.

2. The authors are doing a meta-analysis, even though they don’t call it that, so the term “heterogeneity” should be reserved for “variation between studies”, i.e. τ^2 in the random effects models.

We appreciate this insightful observation. In the random-effects model, we measured homoscedasticity with the μ parameter, and heterogeneity between studies, through τ^2 . In order to clarify this as much as possible, we have specified in the Methods section that τ^2 is used for measuring heterogeneity; and this has also been included between brackets in the Results section:

“The estimated value of τ^2 provides a measure of heterogeneity, that is, to what extent the value of μ is applicable to all studies. The larger τ^2 is, the less the homogeneity”

3. *It’s not clear to me what the “random model” results in Table 1 are. Since this is a model across studies, how can it count individual studies? If empirical Bayes estimates of study-specific effects

are being tested, this must be explained.

We used the Delta method to estimate the within study variability (specifically, the variance of the logarithm of the outcome variance ratio). We have included this explanation in the Methods section: “As there is only one available measure for each study, both sources of variability cannot be empirically differentiated: (i) within study or random or that one related to sample size; and (ii) heterogeneity. In order to isolate the second, the first was theoretically estimated using the Delta method –as explained in Sections V and VI of Supplementary material“

4. Trials that are “significant” are combined - “Subgroup analyses suggest that only significant interventions had an effect on reducing variability” - but interventions that increase the mean should be separated from those that decrease the mean. The later conclusion that “The variability seems to decrease for treatments that perform significantly better than the reference” suggests a different distinction (better/worse is not the same as larger/smaller because outcomes may be positive or negative) and is not supported by the results presented.

Thanks for this great contribution. Following your suggestion, we have sought in each primary endpoint for whether improvements in the response correspond to higher (e.g., mobility) or lower (e.g., pain) values. This new factor has been included in the subgroup analysis (see new figures S5-S7 clicking [here](#) or in the Supplementary Material), thus providing an argument for the existence of a “floor” effect in those studies where a lower value corresponds to a better condition. We have added an interpretation of this finding in the Discussion:

“This reduced variability could also be due to methodological reasons. One is that some measurements may have a “ceiling” or “floor” effect (e.g., in the extreme case, if a treatment heals someone, no further improvement is possible). In fact, according to the subgroup analysis of the studies with outcomes that indicate the degree of disease (high values imply greater severity; e.g., pain), a greater variance (25%) is obtained in the experimental arm (see Figure S5). However, in the studies with outcomes that measure the degree of healthiness (high values imply better condition; e.g., mobility), the average variances match between arms and do not suggest a ceiling effect.”

In addition, we have included this new factor (direction of the improvement) in the [Shiny app](#).

On the other hand, all the significant studies were in favor of the experimental group; therefore, in our context, “statistically significant” is equivalent to “better response in the experimental group”. We have specified this statement in the manuscript and we have kept the sentence: “the authors found statistically significant differences between the arms (all of them in favor of the experimental group) in 83 (39.9%) studies”

5. Abstract, Results: “The adjusted point estimate of the mean ratio (treated to control group) of the outcome variances” is not clear without reading the whole text. Again, defining a term (“outcome variance ratio”?) will help.

Thanks. Corrected both in the Abstract and the main text:

Before [Abstract]: We assessed homoscedasticity by comparing the outcome variability between treated and control arms

After [Abstract]: We assessed homoscedasticity by comparing the variance of the primary endpoint between arms through the outcome variance ratio (treated to control group).

Before [Abstract]: The adjusted point estimate of the mean ratio (treated to control group)

After [Abstract]: The adjusted point estimate of the mean outcome variance ratio (treated to control group) ...

Before [Methods]: ... we fitted a random-mixed effects model using the logarithm of the variance ratio at the end of the trial...

After [Methods]: ... we fitted a random-effects model using the logarithm of the outcome variance ratio at the end of the trial ...

6. *Table 1, “variability is... increased”: from the text, this means “significantly increased”, which should be clarified.

Thanks. We have corrected it:

Before [Table 1]: increased/decreased

After [Table 1]: significantly increased / significantly decreased

Interpretation

The results may be interpreted in many ways, which are sensibly discussed by the authors. Most importantly, treatment effect homogeneity implies homoscedasticity, but the converse (“homoscedasticity implies treatment effect homogeneity”) is not true: this is demonstrated very nicely in Figure 4. Homoscedasticity is scale-dependent: for example, it may be removed (or created) by a log transformation (mentioned in the Discussion).

1. *The authors omit one alternative explanation of homoscedasticity over time: clinical trial populations have eligibility criteria at baseline which may limit baseline variance. For example, a hypertension trial might recruit patients with baseline SBP between 140 and 159 mm Hg. In this case, variance is very likely to naturally increase over time.

Thanks again. We have dealt with this in the Discussion:

“...it has been observed that the variability in the experimental arm also decreases from baseline to the end of the study, *although this comparison is not protected by randomization; for example, the existence of eligibility criteria at baseline may have limited the initial variance (a hypertension trial might recruit patients with baseline SBP between 140 and 159 mm Hg), leading to the variance naturally increasing over time*”

2. **The authors’ conclusions ignore the alternative interpretations noted above. Here are some

examples which are illogical:

- Abstract, Conclusions: “the variance was more often smaller in the intervention group, suggesting, if anything, a reduced role for precision medicine”, and Discussion: “variability tends to be reduced on average after treatment, thus making precision medicine dispensable in most cases”. This is actually false. If a study finds smaller variance in the treated group then we DO have evidence of treatment effect heterogeneity, and indeed the treatment may be doing exactly what medicine should do - making the sickest better while not harming the less sick.

Thanks. We agree. We have addressed this point in the general response above. We provide here further specific comments.

There is heteroscedasticity of effect leading to reduced outcome variability, such as the one shown in examples E and F of Figure 1. Those cases with reduced variability show situations in which the outcome is “under additional control” at the end. The only mathematical model that we can imagine here is the one with an effect correlated with baseline values: higher effects for higher (worse) baseline values. We can imagine this situation for the “ideal” training program: worse participants at the beginning, which further increases or reduces variability. So, although we agree that this is a theoretical heterogeneity, we do not think that it has any practical implication for “individualizing” the treatment: all patients benefit (although to a different degree) from the intervention; and at the end, all patients are “under additional control”.

We have performed some changes in the manuscript in order to clarify this point:

Before [Abstract]: the variance was more often smaller in the intervention group, suggesting, if anything, a reduced role for precision medicine

After [Abstract]: We found that the outcome variance was more often smaller in the intervention group, suggesting that treated patients may end up pertaining more often to reference or “normality” values and thus would not require further precision medicine. However, this result may also be compatible with a reduced effect in some patients, which would require studying whether the effect merits enduring the side effects as well as the economic costs.

Before [Discussion]: variability tends to be reduced on average after treatment, thus making precision medicine dispensable in most cases

After [Discussion]: We found that variability seems to decrease for treatments that perform significantly better than the reference; otherwise, it remains similar. Therefore, the treatment seems to be doing what medicine should do –having larger effects in the most ill patients. Two considerations may be highlighted here: (1) as the outcome range becomes reduced, we may interpret that, following the intervention, this population is under additional control; but also, (2) as subjects are responding differently to treatment, this opens the way for not treating some (e.g. those subjects who are not very ill, and so have no scope to respond very much), with obvious savings in side effects and costs

- Introduction: “If this homoscedasticity holds, there is no need to repeat the clinical trial once a new possible effect modifier becomes measurable” - again, this wrongly assumes the converse stated above.

In this case, we have softened the sentence by changing the term "need" to "evidence".

Before [Introduction]: If this homoscedasticity holds, there is no need to repeat the clinical trial once a new possible effect modifier becomes measurable

After [Introduction]: If this homoscedasticity holds, there is no evidence that the clinical trial should be repeated once a new possible effect modifier becomes measurable

- Discussion: "When both arms have equal variances, then an obvious default explanation is that the treatment is equally effective for all, thus rendering the search for predictors of differential response futile": this is illogical.

We are not sure that we understood why this is illogical. Anyway, we have softened the sentence by changing "an obvious default explanation" to "the simplest explanation".

Before [Discussion]: When both arms have equal variances, then an obvious default explanation is that the treatment is equally effective for all, thus rendering the search for predictors of differential response futile

After [Discussion]: When both arms have equal variances, then the simplest explanation is that the treatment is equally effective for all, thus rendering the search for predictors of differential response futile.

- Discussion: "For most trials, subjects vary little in their response to treatment, which suggests that precision medicine's scope may be less than what is commonly assumed": this is also illogical.

Again, we are not sure that we understood why this is illogical. Nevertheless, we have referred to the limitations derived from Figure 4.

Before [Discussion]: For most trials, subjects vary little in their response to treatment, which suggests that precision medicine's scope may be less than what is commonly assumed

After [Discussion]: For most trials, variability of the response to treatment changes scarcely or even decreases, which suggests that precision medicine's scope may be less than what is commonly assumed – while always taking into account the limitation previously explained in Figure 4.

3. *In the light of the above arguments, I find the statement (Abstract, Conclusions) that "Homoscedasticity is a useful tool for assessing whether or not the premise of constant effect is reasonable" to be highly debatable. Logic suggests it gives a lower bound on the extent of usefulness of precision medicine, and the results of this study do not add any more to this.

We have reduced the ostentatious nature of this phrase, warning the reader that there are limitations to this methodology:

"We have shown that the comparison of variances is a useful but not definitive tool to asses if the design assumption of a constant effect holds."

4. *The objectives in the Discussion should be the same as those stated in the Introduction.

Thanks. We have simplified the objectives in the introduction:

Before: Our objectives were, first, to compare the variability of the main outcome between different arms in clinical trials published in medical journals and, second, to provide a first, rough estimate of the proportion of studies that could potentially benefit from precision medicine. As sensitivity analysis, we explore the changes in the experimental arm's variability over time (from baseline to the end of the study). We also fit a random-effects model to the outcome variance ratio in order to isolate studies with a variance ratio outside their expected random variability values (heterogeneity).

After: Our objectives were, first, to compare the variability of the main outcome between different arms in clinical trials published in medical journals using a random-effects model; and, second, to provide a rough estimate of the proportion of studies that could potentially benefit from precision medicine. Finally, we explore the changes in the experimental arm's variability over time (from baseline to the end of the study).

Also, we have reordered the whole Discussion section according to these objectives:

- 1) Variability comparison between arms and explanation**
- 2) Rough estimate of the studies that potentially benefit from precision medicine (greater variability in experimental arms)**
- 3) Variability comparison between arms and explanation provided in your first suggestion of this section.**

Source data

1. I had trouble opening the source data both in Excel (since the csv file is in fact semi-colon-delimited) and in Stata (which was thrown by line 80). Could it be provided in a more convenient format or with some notes?

We have changed the format (now, columns are comma-delimited) both in the [Shiny app](#) and in the [Figshare](#) repository. We also solved the problem with line 80, which included some unnecessary quotation marks (") in the *Title* field.

Competing Interests: No competing interests were disclosed.

Author Response 06 Nov 2018

Jordi Cortés, Universitat Politècnica de Catalunya, Spain

Following your suggestions together with those of the other reviewers, we have updated the manuscript with a new version that aims to emphasize the fact that researchers' assumption of a constant effect is not clear as long as they do not mention it explicitly. Admittedly, in those studies whose sample size calculation considers some variability in the treatment effect, there is no doubt that this premise has not been considered.

Competing Interests: No competing interests were disclosed.

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com

F1000Research