Does evidence support the high expectations placed in precision medicine? A bibliographic review [version 1; referees: awaiting peer review]

Jordi Cortés¹, 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

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.
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.

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; referees: awaiting peer review] F1000Research 2018, 7:30 (doi: 10.12688/f1000research.13490.1)

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.

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.

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 Figsshare repository16.

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.\textsuperscript{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.\textsuperscript{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". 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)\textsuperscript{10-13}. This "treatment by patient interaction" was already highlighted by W. S. Gosset in the data of his 1908 paper proposing the Student t-distribution\textsuperscript{14}. 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.\textsuperscript{10}, 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 effect. 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 samples when the variance of the paired difference was available. All tests were two-sided (α=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)

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).

![Flow-chart of the articles in the study](image)

**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). 500 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 Carlisle. Fortunately, the exclusion of the four most extreme papers reduced it to 0.07; one of them was the study by Hsieh et al., 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 and in the line of other works that had found a positive correlation between the effect size and its heterogeneity: 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.

<table>
<thead>
<tr>
<th>Comparing variances</th>
<th>N</th>
<th>Method</th>
<th>Increased n (%)</th>
<th>Decreased n (%)</th>
<th>Not changed n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Outcome between treatment arms</td>
<td>208</td>
<td>F test</td>
<td>15 (7.2%)</td>
<td>26 (12.5%)</td>
<td>167 (80.3%)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Random model</td>
<td>11 (5.3%)</td>
<td>19 (9.1%)</td>
<td>178 (85.6%)</td>
</tr>
<tr>
<td>Outcome versus baseline in treated arm</td>
<td>95</td>
<td>Paired test</td>
<td>16 (16.8%)</td>
<td>22 (23.2%)</td>
<td>57 (60.0%)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Random model</td>
<td>13 (13.7%)</td>
<td>19 (20.0%)</td>
<td>63 (66.3%)</td>
</tr>
</tbody>
</table>
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 CONSORT 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 inaccuracies 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 true.

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. 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](https://example.com/figure4.png)  
**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”1. 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/](http://shiny-eio.upc.edu/pubs/F1000_precision_medicine/)

The Figshare repository: [https://doi.org/10.6084/m9.figshare.5552656](https://doi.org/10.6084/m9.figshare.5552656)

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](https://doi.org/10.5281/zenodo.1133609)

**Competing interests**

No competing interests were disclosed.

**Grant information**

Partially supported by Methods in Research on Research (MiRoR, Marie Sklodowska-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:

- 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


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