Constant Effect in Randomized Clinical Trials with Quantitative Outcome.

A Methodological Review

PhD student:	Jordi Cortés Martínez
Advisors:	Erik Cobo Valeri
	José Antonio González Alastrué
Institution:	Universitat Politècnica de Catalunya

Acknowledgments

First of all, I want to thank my thesis supervisors, **Erik Cobo** and **José Antonio González**, for their determination and perseverance with me. My dedication to the thesis during these last years has been intermittent and, on occasion, I have not been able to invest all the time that I had wanted due to the work derived from other projects in which I was immersed. My supervisors have always encouraged me to move on and find the time, even when it seemed none was available. Without them this thesis would not have been possible.

There are many other people whom I want to thank. I will begin with **Lupe Gómez**. In 2017, she made it possible for me to return to work at the University after a 3-year journey working in companies outside the UPC and the university environment. Since she hired me on her staff within a research project she has given me support to be able to finish my thesis and has shown understanding when I needed to devote more time to this work and prioritize it over the research I was carrying out with her.

In the article derived from my thesis, numerous authors have contributed. To all of them, I show my gratitude. To **Stephen Senn** and **Mike Campbell** for having invested hours in their last phase of their career (currently, both are retired) to provide brilliant ideas to be implemented in this work. As an example, some of their contributions are the use of a random effects model and the correction in the estimation of the standard errors of the model for small samples. **Nuncia Medina** was the precursor of the investigation carried out and without her there would have been no further work. I would like to thank **Markus Vogler** and **Marta Vilaró** for the time spent reviewing and collecting part of the data used in this investigation. Finally, **Matt Elmore** has been the person in charge of improving the level of English in this document, articles, answers to the referees and even emails that I sent to foreign researchers.

Together with the authors, the article referees have had a relevant role in the quality of the article derived from this work, published in the journal F1000. This journal has the peculiarity of publishing open reviews and, therefore, it is possible to observe the excellent work they have done. The first three referees of the paper, **Ian White, Saskia le Cessie** and **Erica Moodie** were the most critical in the initial phase but they are also the ones who have contributed the most to any improvements, despite the fact that on many occasions we have strongly disagreed on the essential message of this research. **Richard Stevens, David Nunan, Vance W. Berger** and **Dennis W. Lendrem** were the referees who showed more understanding for our work and the derived conclusions. Thanks to them, we have achieved a final version of the manuscript whose acceptance led to it being indexed in the main databases. There are other researchers who have provided us

with help and advice during this investigation. Among them, I would like to highlight **Miquel Salicrú** and **Joan Bigorra**, whose ideas have contributed to the improvement of this work. Overall, many of the members of the **GRBIO** research group have given me various forms of support on numerous circumstances.

Many people from the university and the department have provided me with help on bureaucratic issues. **Jordi Castro** and **Jan Graffelman** have been the directors of the doctoral program during the period I have completed the thesis. They have provided me with timely advice on how to deal with certain formal aspects of the thesis presentation. **Carme Macias** and **Celia Martín** are the secretaries of the department that have helped me with the procedures of the doctorate process and other related matters. **Toni Font**, the computer scientist of my department, has always had an excellent predisposition for helping me with the uncountable computer problems I have had during this last period. Finally, I am grateful for the help received by the library staff, **Gemma Flaquer** and **Luz Arbeloa** who have guided me through my doubts, especially regarding the bibliographic review.

I also want to especially mention the other doctoral and master students with whom I have shared many seminars and a multitude of conversations around our projects: **Marta Bofill, David Blanco, Cecilia Superchi, Yovi Alarcón, Guillermo Villacampa, Lore Zumeta** and **Pol Castellano**. They have made this journey through the desert more enjoyable.

There are some professors from the department who, despite not having directly participated in the realization of this thesis, they have been indirectly involved. For me, my relationship with them has been special. **Pilar Muñoz** and **Jaume Barceló** were the persons who allowed me to work at the university as a researcher in a previous stage. Surely, if I had not started that phase with them, I would not be currently writing these lines. **Roser Rius** and **Klaus Langohr** are two endearing colleagues for me. With them, I have shared meetings, seminars and even teaching during these years, and they have made my stay and work within the university much more gratifying.

Last but not least, I want to thank **Sonia Navarro**, my partner, and **Iker Cortés**, my son for all the time I have stolen from them during these years. I hope to return part of this time in the coming years. With them, I have learned to value the really important things in life. To my sister, **Marisol Cortés,** who has always advised me on the important decisions of my life. To my parents, **Benito Cortés** and **Ignacia Martínez,** I owe everything and because of them I am here.

Preface

- Good Morning Doctor

- Good Morning, I have bad news. You have high blood pressure.

- Oh! Is this serious?

- No, you do not have to worry. Taking some pills, you can control your hypertension.

- Great! But I'm a bit skeptical with medications, you know? Surely these pills will lower my pressure right?

- Of course. If you want, I can give you the article of the study that demonstrates this medication decreased by an average of 5 mmHG the systolic blood pressure compared to those who did not take such medication.

- Mmmm! I see. But does that mean that all patients lost 5 mmHg of systolic blood pressure?

- No. This is an average difference of the pressures between the two groups at the end of the study. For each patient, it is not possible to know the exact effect on their systolic blood pressure, due to the fundamental problem of causal inference

- The fundamental problem of what?

- I want to say that you cannot know it because only one measurement was taken per patient and it depends on if the pill was taken or not.

- But then, it could be that for some patients the medication is not effective or it's even harmful, right?

- Well, this could be one possibility, but it is not the most likely.

- How do you know?

[Continue reading to know more]

Table of contents

List of figures

List of tables

Abbreviations

Acronym	Meaning
ACE	Average Causal Effect
B	Baseline outcome (measured at the beginning of the study)
BC	Baseline outcome (measured at the beginning of the study) in the Control or reference arm
BT	Baseline outcome (measured at the beginning of the study) in the Treated or experimental arm
C	Control or reference arm
CI	Confidence Interval
DM	Delta Method
df	Degrees of freedom
MLE	Maximum Likelihood estimation/estimator
NI	No Information
NS	Non-Significant
O	Outcome
OC	Outcome (measured at the end of the study) in the Control or reference arm
OT	Baseline outcome (measured at the beginning of the study) in the treated or experimental arm
RCT	Randomized Controlled Trial
S	Significant
SD	Standard deviation
SE	Standard Error
T	Treated or experimental arm
WoS	Web of Science

Abstract

The past decade has seen continuous growth in so-called precision medicine, due especially to great advances in the genetics. While applying it presently goes unquestioned in certain fields like oncology, it is more controversial in other medical specialties that usually practice it. Precision medicine is justified under two assumptions. First, it must be more cost-effective than the universal standard of care, as a world with limited resources requires that an individual treatment's benefits be inversely related to the number of people on whom it is effective. Second, and most importantly, the intervention under study should actually show different responses among patients or subgroups of them, which this work focuses on.

Strictly speaking, the fundamental problem of causal inference makes the latter requirement impossible to prove, because a conventional trial observes patient outcome only under a single treatment. However, the variability of a continuous outcome provides important information about the presence (or absence) of a constant treatment effect, of which a direct consequence is that outcome variance remains unchanged under different treatment regimens. Thus, homoscedasticity may be a useful tool for testing the hypothesis of a homogeneous effect.

Our work here conducts a methodological review of randomized clinical trials (RCT) with two treatment arms and a quantitative primary endpoint. Among other variables, we collected the outcome and baseline variances for each treatment group with two purposes: to quantify the outcome variance ratio between the experimental and reference groups; and to estimate the proportion of studies with variance discrepancies large enough to be attributed to a heterogeneous treatment effect among participants. This variance comparison was carried out between treatment arms (independent by randomization) and over time, contrasting the end-of-study and baseline outcomes.

The Medline database provided us 208 randomized clinical trials fulfilling the eligibility criteria and published in the years 2004, 2007, 2010 and 2013. A random effects model was used to estimate the variance ratios (experimental to reference), of which the mean was 0.89, 95% CI from 0.81 to 0.97. Thus, contrary to popular belief, the point estimate indicate that the experimental treatments reduce the variability of patient response by 11%. The experimental group's variance ratio (final to baseline) in the comparison over time was 0.86, 95% CI from 0.76 to 0.98, meaning lower variability at the end of the study.

This analysis provides no statistical evidence to justify ruling out a constant intervention effect on our target population in four out of five studies (80.3%, 95% CI from 74.1 to 85.3%). This percentage barely changed in four sensitivity analyses with percentage point estimates ranging

from 79.8 to 90.0%. Among the studies that we found evidence of a non-constant intervention effect, the experimental group showed 7.2% and 12.5%, respectively, greater and lower outcome variance than the reference arm. The high number of studies with lower variability in the experimental group can be explained by the *ceiling* and *floor* effects of some measurement scales, which generally group patients at one of the scale boundaries in cases of highly effective interventions.

This work aims to show that comparing variances provides evidence on whether or not precision medicine is a sensible choice for a specific treatment. When both arms have equal variances, a simple interpretation is that the treatment effect is constant. If true, searching for any predictors of a differential response is futile. This means that the average treatment effect can be viewed as an individual treatment effect, which justifies using a single clinical guideline for all patients fulfilling the eligibility criteria. This in turn supports using parallel controlled trials to guide decision-making in these circumstances.

Abstract (spanish)

La medicina de precisión ha experimentado un auge continuo en la última década debido sobre todo a grandes avances en la genética. Aunque su uso es actualmente incuestionable en campos como la oncología, es más controvertido en otras especialidades médicas. La medicina de precisión queda justificada bajo dos supuestos. Por una parte, debe ser más rentable que el tratamiento estándar en el sentido que los beneficios individuales de un tratamiento deben relacionarse inversamente con el número de personas en las que es realmente eficaz. En segundo lugar, y más importante, la intervención debe actuar realmente de forma diferencial entre los pacientes.

Formalmente, el problema fundamental de la inferencia causal establece que este último requisito es indemostrable debido a que los ensayos convencionales muestran la respuesta de cada paciente bajo un único tratamiento. Sin embargo, la variabilidad de una respuesta continua proporciona información valiosa sobre la presencia de un efecto constante, siendo una consecuencia directa que la variabilidad permanece inalterable bajo diferentes intervenciones. Por tanto, el estudio de la homoscedasticidad de la respuesta puede ser una herramienta útil para probar la hipótesis de homogeneidad del efecto.

Se realizó una revisión metodológica de ensayos clínicos aleatorizados paralelos con una variable respuesta principal cuantitativa. Se recogió información referente a las varianzas de dicha variable respuesta al final y al inicio del estudio para cada grupo de tratamiento con dos propósitos: estimar la razón de varianzas y estimar la proporción de estudios con discrepancias en la varianza lo

suficientemente grandes como para ser atribuidas a un efecto heterogéneo. Se compararon las varianzas entre brazos de tratamiento (independiente por la asignación aleatoria) y a lo largo del tiempo, comparando las varianzas de las respuestas al final y al inicio del estudio.

Se obtuvieron 208 ensayos clínicos publicados en los años 2004, 2007, 2010 y 2013 de la base de datos *Medline* que cumplían los criterios de elegibilidad. El análisis principal se basó en un modelo de efectos aleatorios que estimó la media de las razones de varianzas (experimental vs. control) en 0,89, IC95% de 0,81 a 0,97. Contrariamente a la creencia popular, los tratamientos experimentales redujeron en media la variabilidad de la respuesta del paciente en un 11%. La razón de varianzas dentro del grupo experimental a lo largo del tiempo (final vs. basal) fue de 0,86, IC del 95% de 0,76 a 0,98, implicando una variabilidad menor al final del estudio.

Nuestro análisis principal no proporcionó evidencia estadística para descartar un efecto constante del tratamiento en nuestra población objetivo en cuatro de cada cinco estudios (80,3%, IC95% de 74,1 a 85,3%). Este porcentaje apenas cambió en cuatro análisis de sensibilidad que arrojaron estimaciones puntuales entre 79,8 y 90,0%. Entre los estudios en los que se halló evidencia de un efecto no constante, un 7.2% y un 12,5% presentaron una mayor y menor variabilidad en el grupo experimental, respectivamente. Este resultado podría explicarse por los efectos *techo* y *suelo* característicos de algunas escalas, que tienden a agrupar a los pacientes en alguno de sus extremos cuando las intervenciones son altamente eficaces.

El objetivo de este trabajo es mostrar que la comparación de varianzas proporciona evidencia sobre si la medicina de precisión es una opción razonable para un tratamiento específico ya que una interpretación simple de la presencia de homoscedasticidad es que el efecto del tratamiento es constante. En caso de ser constante, la búsqueda futura de predictores de una respuesta diferencial es inútil y el efecto promedio del tratamiento puede asimilarse como un efecto individual. Esto justificaría el uso de una única guía clínica para todos los pacientes que cumplan los criterios de elegibilidad y respalda el uso de ensayos paralelos para guiar la toma de decisiones.

1 Introduction. State of the art and objectives

1.1 Synopsis

Precision medicine is the Holy Grail of interventions that are tailored to a patient's individual characteristics. As a patient's response can vary with a new treatment, clinical trials try to estimate individual treatment effects and therefore outcome variability may be greater among treated participants than reference patients. However, the conventional design and analysis of randomized trials might gain an advantage by assuming that each individual benefits by the same amount. We have reviewed parallel group trials with quantitative outcomes to examine this assumption. The results of the present work controvert popular belief, indicating that variability is reduced in some treatments, which thus obviates the pursuit of precision medicine in those cases. Conversely, some increase in treatment effect has been observed in a few interventions, which could provide more tailored treatments through finer selection criteria. This study demonstrates that homoscedasticity might be used to assess whether the eligibility criteria in clinical trials needs to be refined once precision medicine provides new effect modifiers.

1.2 Precision medicine

The idea of precision medicine is to develop prevention and treatment strategies that take individual characteristics into account. The prospect of applying this concept broadly has been dramatically improved by recent developments in large-scale biological databases (such as the human genome sequence), new 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]. Most of the emphasis underlying this initiative was made in the field of oncology, even though the incursion of precision medicine has recently increased in other areas of disease study, such as psychiatry. Nevertheless, it is uncertain whether these fields could fit well into this new paradigm, because their benefits have not been as well established as in oncology[3,4].

There are several terms to conceptualize the idea behind a tailored intervention. The distinction between ***precision*** and ***personalized medicine*** is fuzzy, and there is no consensus in the literature, neither in the definition nor in the difference between both terms. Each one focuses on identifying which treatments will be effective for which patients based on genetic, environmental and lifestyle factors. In both cases, preventative and therapeutic interventions are concentrated on those who will benefit while avoiding expenses and side effects for those who will not. For instance, the

National Research Council does not make any distinction between these terms beyond the particular concern about the word *personalized*, which could be misinterpreted in the sense that it can be understood that treatments are being developed uniquely for each individual when the word can – and often does – refer to interventions applied to subgroups of patients[5]. Other authors have delved further into the idiosyncrasies of each term and find subtle differences between both terms[6]. The term *personalized* (much older and prior to current advances in the field of genetics) is associated with integrated medicine in which the patient is viewed as a whole (patient-centered), thus considering all her/his possible components, from biochemistry to behavior to subjective well-being to environmental exposure. In contrast, *precision medicine* is seen as a new all-integrating endeavor of rational, data-driven, mechanism-based health care that focuses more on the disease[7]. In some way, the term precision could be literally interpreted in its original linguistic meaning as a deterministic (not probabilistic) knowledge of all relevant details of a system that would allow discovering the cause of any malfunction and deduce possible successful remedies.

On the other hand, ***individualized medicine*** combines standardization with individualization, i.e., the target is fixed on individuals[8]. Whereas the former two terms pertain to prevention, diagnosis, and treatment, *individualized medicine* is directly related to the ***N-of-1*** studies and, therefore, the definition is more close to a specific clinical research strategy. Table 1 tries to summarize the main peculiarities of each term.

Table 1. Characteristics of the different terms related to tailored interventions.

Terms	Origin	Application	Design	Focus	GS entries*
Personalized	c. XIX	Subgroups	Any	Patient-centered	406,000
Precision	2010	Subgroups	Any	Disease-centered	784,000
Individualized	1986	Individuals	N-of-1	Patient-centered	19,500

**GS: Google Scholar entries (10th April 2020) for the searches: i) "Personalized medicine"; ii) "Precision medicine"; and iii) "Individualized medicine" OR "N of 1" OR "N of one" OR "N-of-1"*

Despite the fact that technological advances have allowed for the development of *personalized/precision* medicine in recent times, the idea of treating the patient in the most individualized way is very old. As Simeon-Denis Poisson et al.[9] pointed out in reference to a previous article by Double[10],

> "*In the field of statistics, that is to say in the various attempts at numerical assessment of facts, the first task is to lose sight of the individual seen in isolation, to consider him only as a fraction of the species. He must be stripped of his individuality so as to eliminate anything accidental that this individuality might introduce into the issue in hand.*

> *In applied medicine, on the contrary, the problem is always individual, facts to which a solution must be found only present themselves one by one; it is always the patient's individual personality that is in question, and in the end it is always a single man with all his idiosyncrasies that the physician must treat. For us, the masses are quite irrelevant to the issue*".

In recent times, many supporters and detractors have generated much controversy about the utility of precision medicine, but the reality is that precision medicine is gaining more and more ground: in the last decade, projects related to genetics have received 50% more funding than those ones with the goal of disease prevention, and there has been a 90% decrease in the terms "*public*" and "*population*" in the titles of published articles[11].The defenders warn about the presence of a treatment–patient interaction in many interventions, arguing that medicine requires a different type of clinical trial that focuses on individual (not average) responses to therapy[12,13]. Conversely, most of the industry guidelines rely on the assumption of the constant effect among patients and they warn that the interpretation of heterogeneity treatment effects among patients is controversial[14]. Another criticism lies in the fact that precision medicine amplifies the gap between social classes and moves away from the global objective of achieving a healthier population:[11] although they can be cost-effective in some cases, the development of expensive medications diverts resources from the R&D of more effective global therapies[15]. One last current weakness concerns the application of precision medicine being limited to some medical areas. For example, the SHIVA trial[16] compared a molecularly targeted therapy based on tumor molecular profiling (precision medicine) versus the conventional therapy for advanced cancer; no statistically significant difference was found between groups in regard to the primary endpoint progression-free survival (hazard ratio 0·88, 95% CI 0·65–1·19).

1.3 The constant effect assumption is a misnomer

Variability in a clinical trial outcome measure should interest researchers, because it conveys important information about the treatment effect and whether or not there is a need for precision medicine. Does variability come only from unpredictable sources of patient variability? Or should it also be attributed to a different treatment effect that requires more precise prescription rules?[17–19] Usually, researchers assess treatment effect modifications among subgroups based on relevant variables. The term *interaction* refers to divergent effects in each subgroup of patients (*stratum*), and its determination requires large samples in all stratums sharing the same treatment effect. This often results in underpowered interaction analyses[20]. The main problem is, however, that those stratification factors should be known in advance and be measurable. This in turn implies that

when new variables are discovered and introduced into the causal path, these can only be analyzed in an exploratory manner and an additional clinical trial will be needed to obtain confirmatory results. Fortunately, one observable consequence of a constant effect is that the treatment will not affect variability, and therefore the outcome variances in both groups should be equal.

1.3.1 The fundamental problem of causal inference

The fundamental problem of causal inference is that for each patient in a parallel group trial we can only know the response for one of the interventions. That is, we observe the response to either the reference treatment or the new treatment, but not both. By experimentally controlling unknown confounders through randomization, a clinical trial may estimate the average causal effect (ACE). However, in order to translate this population estimate into effects for individual patients, additional assumptions are needed, with the premise of a constant effect being the simplest one. In this work, we try to elucidate whether the comparison of observed variances may shed some light on the non-observable individual treatment effect.

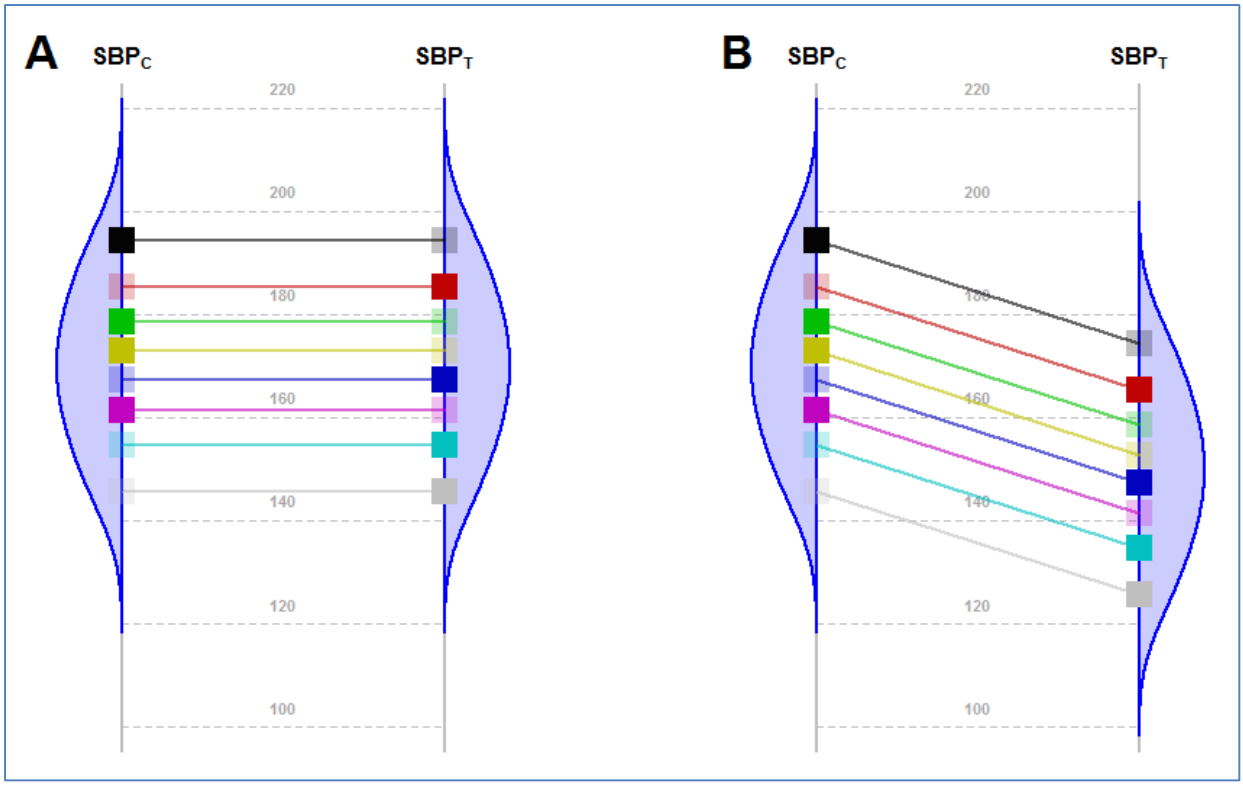
1.3.2 Scenarios for the non-observable individual effect.

Let us begin with a description of what can be observed in different hypothetical scenarios. First, imagine a fictional trial with eight participants (four in each arm) and systolic blood pressure (SBP) as the primary endpoint. Figure 1A shows the potential outcome values that we could obtain from each patient if the treatment effect was null. As explained above, a participant in a parallel RCT is allocated to a single arm and, thus, we observe only one outcome, which is represented in the figure by squares with fully saturated colors. Opposite, transparent squares represent the missing potential outcomes that would have been observed if the volunteer were to have been allocated to the other group. As the intervention has no effect at all, both groups have the same distribution (i.e., mean and variance). Having equal variances across treatment groups is referred to as homoscedasticity. Figure 1B shows the scenario of a constant effect, meaning that the effect is exactly the same in every patient: the intervention lowers the SBP by a single, constant value in each participant and, again, variability remains unalterable.

For instance, the study from Duran-Cantolla et al.[21] compared the 24-hour SBP among 340 patients randomized to either continuous positive airway pressure (CPAP) or sham–CPAP, and they found 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 agree with the trial design's assumption of a constant effect and nothing contradicts the inference that each patient exhibits a constant reduction

of 2.1 mmHg, although the uncertainty derived from random allocation makes the results compatible with a constant effect that lies anywhere between 0.4 and 3.7.

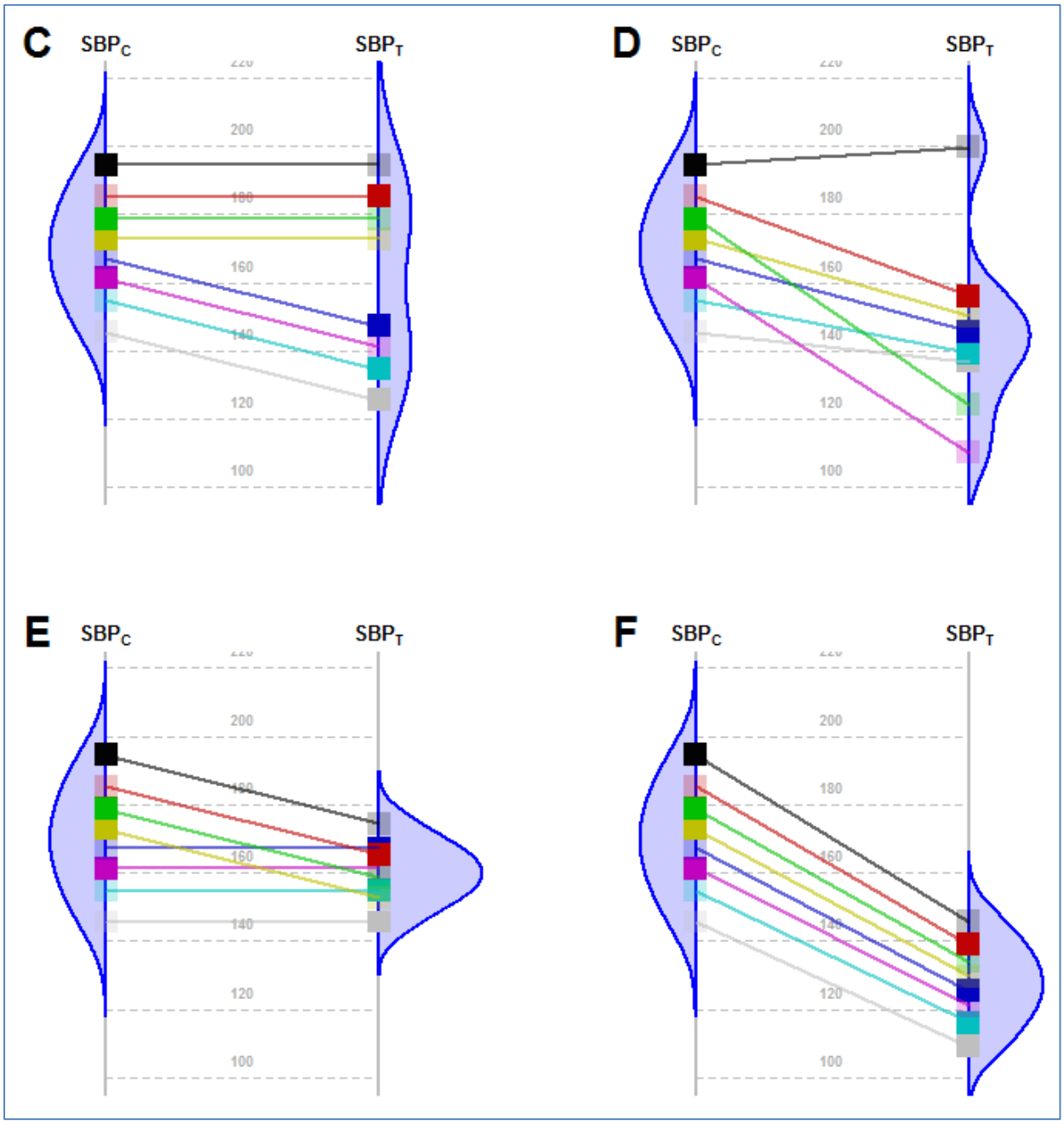
Figure 1. SBP values of each patient in both groups (C: Control; T: Treated) under a constant treatment effect. Fully saturated colored squares represent observed values, and transparent squares are missing outcomes (not observed) values. The line slope indicates the individual and non-observable effect for each patient. Densities of the potential outcome distributions are represented assuming normality.



On the other hand, if the treatment effect varies among patients, the variance may be greater (or lower) in treated patients. An example of this situation can be found in the study of Kojima et al.[22], in which the primary outcome measure is the 3-hour postprandial area under the curve (AUC) of apolipoprotein (Apo) B48. Outcome SDs were 0.78 and 0.16 in the experimental and reference group, respectively, with a variance ratio equal to 23.77. This is consistent with an intervention that increases variability and with diverse individual treatment effects that need further refinements with the help of precision medicine.

Let us imagine two completely different scenarios that may apply to this increased variability. Panel C of Figure 2 represents a hypothetical scenario with a couple of different effects in two subpopulations (*subgroup interaction*). Although the effects are identical within any subgroup, the overall observable distribution in the treated arm would have higher variability. Here, we need precision medicine in order to find more restrictive eligibility criteria and to identify any criterion that classifies patients in those subpopulations, as well as to be able to assume a constant effect.

Figure 2. SBP values of each patient in both groups (C: Control; T: Treated) for several heterogeneous treatment effect scenarios. Fully saturated colored squares represent observed values, and transparent squares represent missing outcomes. The line slope indicates the individual non-observable effect for each patient. Densities of the potential outcome distributions are represented assuming normality.



However, in panel D, the treatment has a variable (i.e., *random*) effect in each patient, resulting also in greater variability within the treated arm. Unfortunately, there is no longer any subgroup of patients sharing a common effect, and results from previous patients are poorly predictive about the outcomes of future ones. When taken to this extreme case, no group of patients shares a common effect, and precision medicine should become individualized medicine by means of observing the outcome in a specific patient receiving both treatments. This is only feasible for

chronic, stable conditions without carry-over effects, meaning that we can run specific *N-of-1* trials, as previously suggested[13,23–26]. This "treatment by patient interaction" was already highlighted by W. S. Gosset in the data of his 1908 paper proposing the Student t-distribution[27]. As we have already stated, the presence of greater variance in the treated arm leads to a hard interpretation of the main treatment effect,[14] and 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).

Alternatively, interactions can result in smaller variances in the treated arm. In panel E, there is a different effect in two subgroups, but the variability is now reduced. Again, the best solution would be to identify the subpopulations in order to refine the selection criteria and to provide a better estimate of the effect size. In panel F, the treatment has a stabilizing effect, with higher blood pressure falling more in severe patients, and resulting in lower variability in the treated arm. Notice that in the last scenario, although heteroscedasticity is present, there is no problem at all: the intervention has improved the conditions of every patient and, although the individualized treatment effect cannot be provided, the outcome distribution represents a more stable situation that can be described in order to summarize treatment effects. Our most extreme example of a study with the smallest variance ratio comes from Kim et al.[28], in which the outcome is 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). The 11 control patients had an average score of around 42 at baseline and post-baseline, while the mean score of 11 treated participants decreased from 42 to 24, which results in a statistically significant reduction. The respective SDs at the end of the trial were 3 and 16 for the treated and control arms, respectively (variance ratio equal to 0.035), implying that variance was reduced by a factor of approximately 28.

1.3.3 Sample size depends on the nature of the treatment effect

Whether an intervention has a constant effect or not is a relevant issue in itself to determine the scope of a given intervention, but it is also important for the design of a trial that aims to find a mean difference in the primary endpoint. The sample size calculation in RCTs with this purpose is performed on the basis of a single parameter to specify the treatment effect (commonly called Δ). This fact implicitly implies the assumption of a constant effect; otherwise, at least one additional parameter would be required to specify the variability of this effect.

Appendix A shows two tables containing sample size calculations drawn from 20 articles in the journals Trials and NEJM. None of them explicitly mentions that the treatment effect could be

variable. In addition to being reasonable, the constant effect assumption simplifies the problem, since the presence of a heterogeneous treatment effect has consequences on the needed sample size. By means of examples, this Appendix A introduces four different scenarios: constant treatment effect, random treatment effect and two types of interactions. The required sample sizes for these situations differ from each other, ranging from 32 to 75 patients, even though the parameters used in the traditional formula are the same.

1.4 Objectives and structure

After describing the challenges derived from the variability of the treatment effect, we introduce the objectives and structure of this PhD thesis. The data used in this study came from a methodological review which is extensively described in Chapter 2. The two main objectives of this work are:

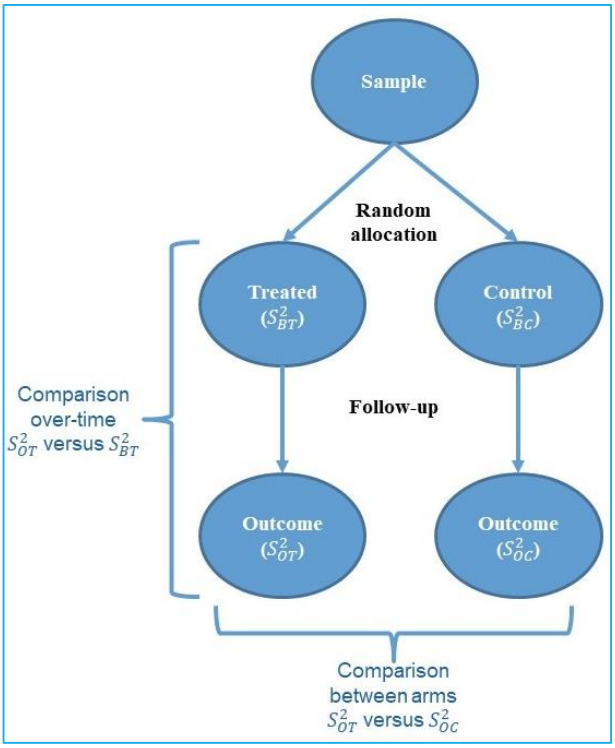
- To compare the variance of the main outcome between arms (at the end of the study) and over time (from baseline to the end of the study in the experimental group) in RCTs published in medical journals and to provide an estimation of the variance ratio.
- To estimate the proportion of studies with a plausible constant effect based on this variance comparison.

We compared the variances in a set of parallel RCTs with quantitative outcome in two types of comparisons (Figure 3):

- **Between arms at the end of the study**: Treated (T) vs. Control (C)
- **Between final and baseline outcome in the experimental arm**: Outcome (O) vs. Baseline (B)

This distinction is pertinent because of the nature of the relationship between the data; while in the comparison between arms the data of each group are independent. In the comparison over time the data are related through the different measures in the same patient.

Figure 3. Schema of the two studied comparisons in a parallel RCT.



We have implemented five different methodologies to carry out the objectives for both comparisons. Three of them are based on the theory behind the random effects model (REM) and two are based on the usual tests of variance comparison.

- **Based on REM:**
  - ***Random effects model***. Main Analysis.
  - ***Heuristic procedure***. Sensitivity Analysis I.
  - ***Simulation study***. Sensitivity Analysis II.
- **Based on variance comparison tests:**
  - ***Usual tests for variance comparison***. Sensitivity Analysis III.
  - ***Mixture distributions for the p-values***. Sensitivity Analysis IV.

These are explained in detail in the following chapters. Their common goal is to determine which proportion of the population studies could take advantage of the precision medicine setup due to the absence of a constant effect. In addition, the first main methodology also provides an estimate of the variance ratio between arms and between final and baseline outcome. See Table 2 for an overview of the complete structure of this work.

Table 2.Schema of how the objectives, the type of analysis and the type of comparisons are related to the chapters/sections of this document.

Chapter/ Section	Analysis		Objective		Type of comparison	
	Type	Specific analysis	Variance ratio estimation	Proportion of constant-effect trials	Between Arms	Over time
Chapter 3	Random effects model	***MA***. Random effects model	✓	✓	✓	✓
Section 4.1		***SA1.*** Heuristic procedure		✓	✓	✓
Section 4.2		***SA2.*** Simulation study		✓	✓	
Section 4.3	Variance test	***SA3.*** Variance tests		✓	✓	✓
Section 4.4		***SA4.*** Mixture distributions		✓	✓	✓

MA: Main Analysis; SA: Sensitivity Analysis.

2 Data source. Methodological review

The data for our work was gathered through a methodological review, which differs from systematic or bibliographic reviews in that it is not limited to a specific topic but instead focuses on peculiar characteristics of a design or analysis. In our case, we wanted to collect parallel clinical trials (see Section 2.1.1) through a search strategy that is detailed in Section 2.1.2. The list of collected variables in the dataset is specified in Section 2.1.3 and some issues related to the data and the software used in the analysis are present in Section 2.1.4. Finally, Sections 2.2.1 and 2.2.2 show the flow diagram for selecting the data and a preliminary descriptive analysis, respectively.

2.1 Methods

2.1.1 Population

Our target population was parallel randomized clinical trials with quantitative outcomes (continuous or not); so, studies with categorical or time-to-event primary endpoints were not included. Trials needed to provide enough information to assess both homoscedasticity assumptions in the primary endpoint: between arms and over time. Therefore, baseline and final standard deviations for the main outcome were necessary or, failing that, at least one measure (variances, standard errors or mean confidence intervals) that would allow us to calculate them (see Appendix B for some examples).

2.1.2 Search strategy

Articles on parallel clinical trials from the years 2004, 2007, 2010 and 2013 were selected from the Medline database via EBSCO with the following criteria:

Years 2004 and 2007
AB(clinical trial AND random*) AND AB(change OR evolution OR (difference AND baseline)*

Years 2010 and 2013
AB((trial AND random*) AND ((change OR evolution) AND (difference AND baseline)))*

For the years 2004 and 2007, we selected all papers that met our inclusion criteria; while for the years 2010 and 2013, we chose a random sample of 300 papers from the articles retrieved from the search.

One limitation of the search strategy was the usage of terms "clinical trial" and "trial" instead of setting a filter for this type of design. Although this limitation works against specificity (studies

that were not clinical trials would be returned), it does not limit sensitivity (very few clinical trials omit these terms in the abstract). Therefore, the only drawback was the need for more effort to screen the articles.

The initial purpose of the data collection was not to assess the homoscedasticity of the treatment effect, but primarily to estimate the correlation between the baseline and final measurements of the primary endpoint; under this rationale, the word "difference" was paired with "baseline" in an alternative option that is included in the abovementioned criteria.

In June 2017, we tried to reproduce this search strategy via PubMed (the original searches were run on the EBSCO platform) by adapting the syntax in a pertinent way in order to assess the variation of the returned articles some time later. A noteworthy different number of articles was observed from those previously retrieved. Two explanations would be that some factors may vary over time, such as the indexing of articles or that the platform for accessing the Medline database was distinct (EBSCO versus PubMed) in the two periods. Table 3 shows the specific search strategy conducted and the number of articles returned for each year.

Table 3. Summary of the original methodological review and the validation.

	Original Strategy (EBSCO)			**Validation strategy (PubMed)**		
Year	**no. papers**	**Search date**	**Search**	**no. papers**	**Search date**	**Search**
2004	**266**	January 2005	AB(clinical trial* AND random*) AND AB (change OR evolution OR (difference AND baseline))	**326**	June 2017	(clinical trial*[Title/Abstract] AND random*[Title/Abstract]) AND (change[Title/Abstract] OR evolution[Title/Abstract] OR (difference[Title/Abstract] AND baseline[Title/Abstract])
2007	**348**	January 2008		**503**	June 2017	
2010	**478**	January 2015	AB((trial* AND random*) AND ((change OR evolution) AND (difference AND baseline)))	**319**	June 2017	(trial*[Title/Abstract] AND random*[Title/Abstract]) AND (change[Title/Abstract] OR evolution[Title/Abstract] AND (difference[Title/Abstract] AND baseline[Title/Abstract]))
2013	**657**	January 2015		**442**	June 2017	

Data were collected by two different researchers in two phases: 2004/2007 and 2010/2013. After merging the two databases, two statisticians checked the correctness of the data by reviewing all articles from the first two years (2004, 2007) and a random sample (10% of the total) from the last two years (2010, 2013).

2.1.3 Variables

Various pieces of information regarding the clinical trials of the selected articles were collected and they were reflected in the following variables: baseline and outcome standard deviations; experimental and reference interventions; initial and final sample size in each group; medical field labeled according to *Web of Science* (WoS) classification; primary endpoint; disease under study; kind of disease (chronic or acute); outcome type (measured or scored); intervention type (pharmacological or not); improvement direction (upwards or downwards) and whether or not the main analysis of the trial proved the intervention efficacy. The last five variables were involved in the subgroup analyses and they are widely explained in Section 3.1.6.

For studies with more than one quantitative outcome, primary endpoint was determined according to the following hierarchical criteria: (1) variable appearing in either the objective or hypothesis; (2) employed in sample size determination; (3) response in main statistical analysis; or (4) first quantitative variable reported in results. In the same way, the choice of the "experimental" or "treated" group was determined depending on their role in the following sections of the article: (1) objective or hypothesis; (2) sample size determination; or (3) study rationale. In trials with more than two treatment arms, the experimental intervention was the one involved in the first comparison reported in the results.

2.1.4 Software and Data

All analyses were performed with the R statistical package version 3.6.3 or higher. We used the *rma* function from *metafor* package to fit the random effects models.

Our data is made available through two sources:

- A shiny app (see Section 6.1) that allows the user to interact with the data by means of several visualization tools: http://shiny-eio.upc.edu/pubs/F1000_precision_medicine/[29]
- The *Figshare* repository: https://doi.org/10.6084/m9.figshare.5552656[30]

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

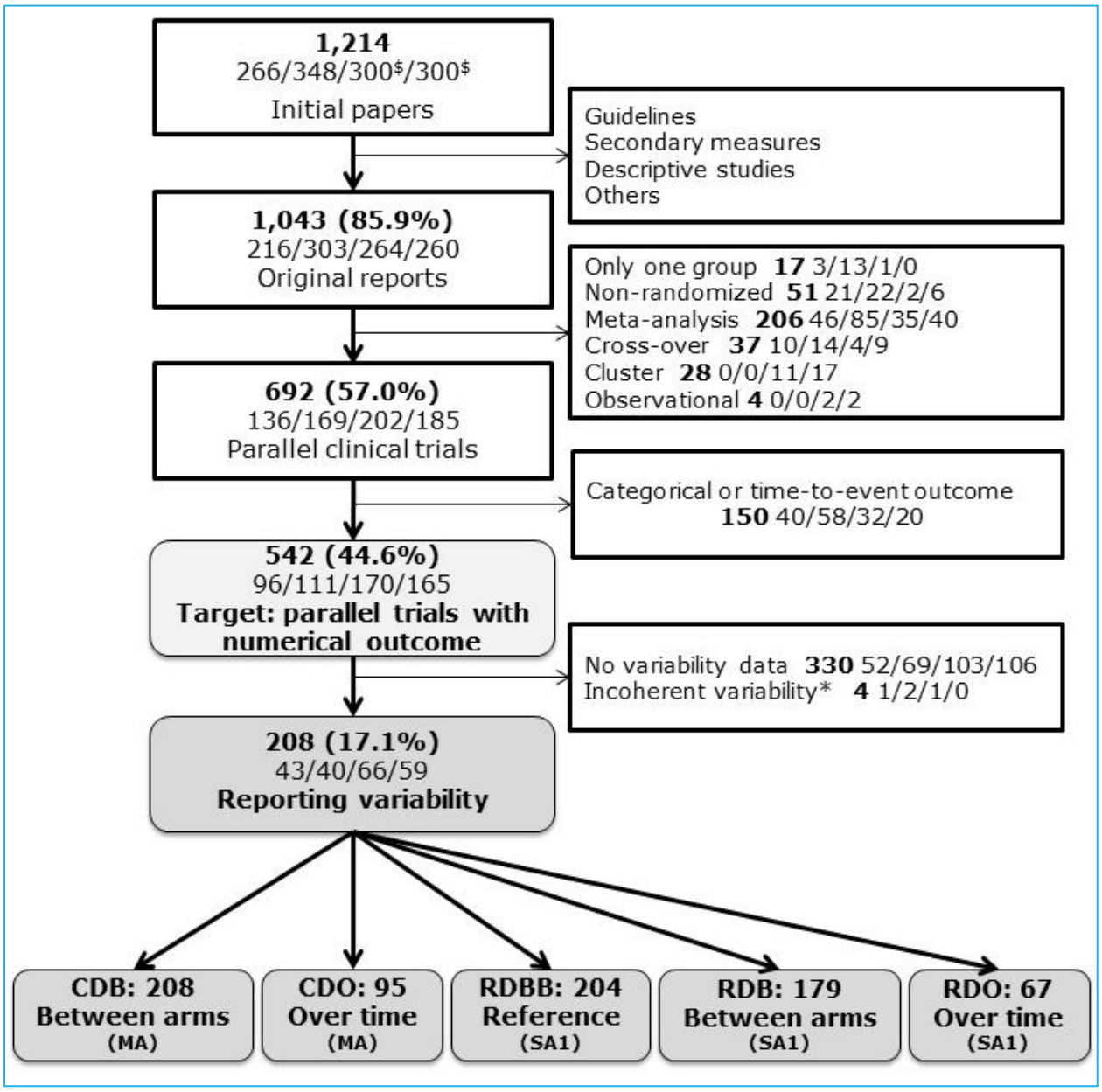
2.2 Results

2.2.1 Flowchart

Exactly 1,749 articles were retrieved from the search, but a random selection of 300 articles in 2014 and 2017 involved a total of 1,214 initial papers, with 542 of them pertaining to the target population. Of those, 212 contained enough information to conduct the analysis, but four papers were excluded because the change over time in the reported variance of the outcome (final minus

baseline) was inconsistent with both the baseline and final variances, which would lead to infeasible absolute correlation estimates above 1. Ultimately, 208 (38.4%) were included in the analysis (Figure 4).

Figure 4. Percentages represent the number of papers with respect to the ones retrieved from the methodological review. The number of articles for each year (2004/2007/2010/2013) is specified in each box separated by slashes. $300 papers were randomly selected for years 2010 and 2013.



MA: Main analysis; SA1: First Sensitivity analysis; CDB: Complete Dataset for Between-arm comparison; CDO: Complete Dataset for comparison Over time; RDBB: Reduced Dataset for Between-arm comparison at Baseline; RDB; Reduced Dataset for Between-arm comparison; RDO: Reduced Dataset for comparison Over time.

The end of the flow chart is split into several datasets that were used, depending on the type of analysis or comparison. The complete datasets for the main analysis (MA) and all the sensitivity analyses except the first one (SA2, SA3, SA4) were called CDB (Complete Dataset for Between-arm comparison) and CDO (Complete Dataset for comparison Over time). The datasets used in the first sensitivity analysis (SA1), which are subsets of the previous ones, were named RDBB (Reduced Dataset for Between-arm comparison at Baseline), RDB (Reduced Dataset for Between-arm comparison) and RDO (Reduced Dataset for comparison Over time).

2.2.2 Descriptive results

Following the WoS criteria, 156 articles (75%) pertained to exactly one medical field; 47 (22.6%) to more than one and 5 (2.4%) were not classified. The main areas of study were: General & Internal Medicine ($n = 31$, 14.9%), Nutrition & Dietetics ($n = 21$, 10.1%), Endocrinology & Metabolism ($n = 19$, 9.1%), and Cardiovascular System & Cardiology ($n = 16$, 7.7%). See Appendix C for more detailed information and the complete list of the medical fields.

Mainly, collected studies were non-pharmacological ($n = 122$, 58.6%); referred to chronic conditions ($n = 101$, 57.4%); had an outcome measured with units ($n = 132$, 63.8%) instead of a constructed scale; were measured ($n = 125$, 60.1%) rather than assessed; and had lower values associated with an improvement in health status ($n = 141$, 67.8%). Regarding the primary objective of the trials, the authors proved the intervention efficacy in 83 (39.9%) studies.

Table 4 shows the main statistics for the logarithm of the sample variance (S^2) at the end and at the beginning of the trial in both arms, as well as their differences. The mean of the $\log(S^2)$ in the experimental arm at the end of the study (2.76) is lower than in the control group (2.81), and it is also lower than the outcome measure in the experimental group at baseline (2.88).

Table 4. Descriptive statistics for the logarithm of the S^2 (rows 1-4) and for their differences (rows 5-7).

		n	mean	sd	min.	Q1	median	Q3	max.
Baseline	Reference or Control arm (BC)	208	**2.77**	3.22	-7.82	0.35	**3.26**	4.94	11.28
	Experimental or Treated arm (BT)	208	**2.88**	3.29	-7.01	0.28	**3.60**	5.19	13.65
Outcome	Reference or Control arm (OC)	208	**2.81**	3.26	-7.82	0.36	**3.20**	5.08	11.56
	Experimental or Treated arm (OT)	208	**2.76**	3.26	-7.01	0.19	**3.02**	5.02	13.47
Differences	Between arms at the end (OT-OC)	208	**-0.05**	0.82	-4.29	-0.44	**0.00**	0.34	3.92
	Over time in Experimental arm (OT-BT)	208	**-0.12**	0.76	-3.19	-0.34	**-0.05**	0.21	3.17
	Over time between arms [(OT-BT) - (OC-BC)]	208	**-0.16**	0.77	-2.94	-0.44	**-0.10**	0.12	5.11

B: Baseline, O: Outcome, T: Treated or experimental, C: Control or reference.

Figure 5 and Figure 6 highlight these differences in each single study for both comparisons: 113 (54.3%) studies revealed lower variability in the outcome in the experimental arm; 12 (5.8%) trials reported exactly the same variability in both arms at the end of the study; and 83 (39.9%) provided a higher variability in the reference arm. Arrow lengths represent the variance difference between arms. Although it is subtle, the magnitude of the differences seems slightly higher in trials with less variability in the experimental group.

Figure 5. Arrows go from the outcome variance in the control group to the outcome variance in the treated group for each trial, and it may be increasing (blue) or decreasing (red). Y-axis is log-scaled.

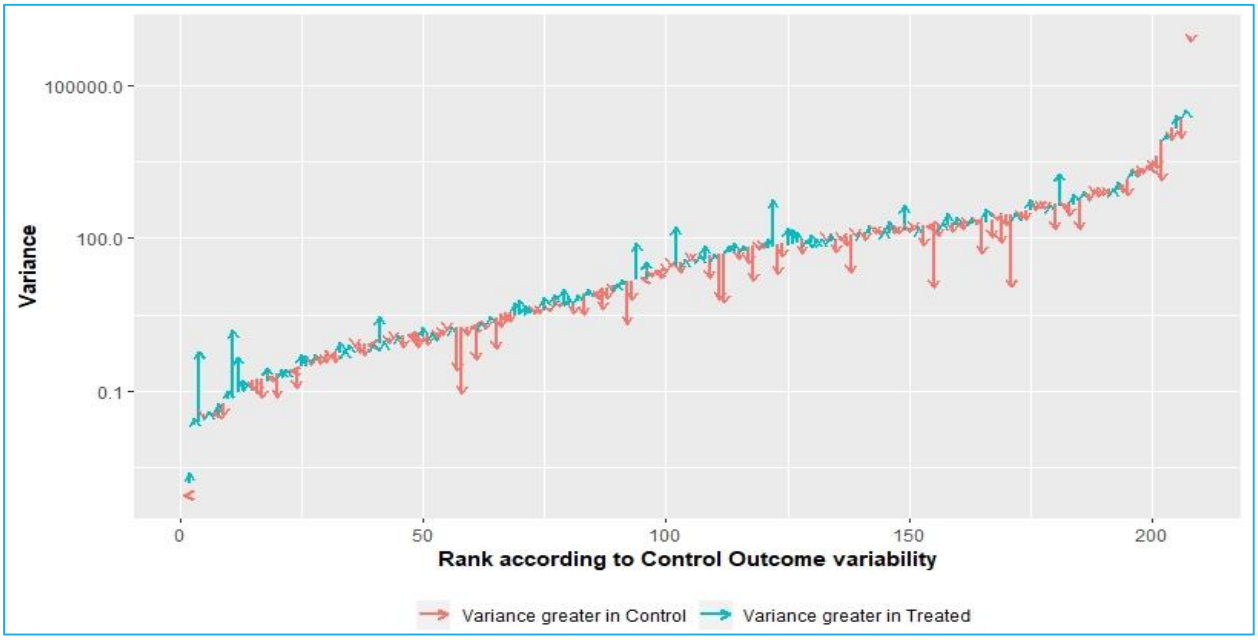
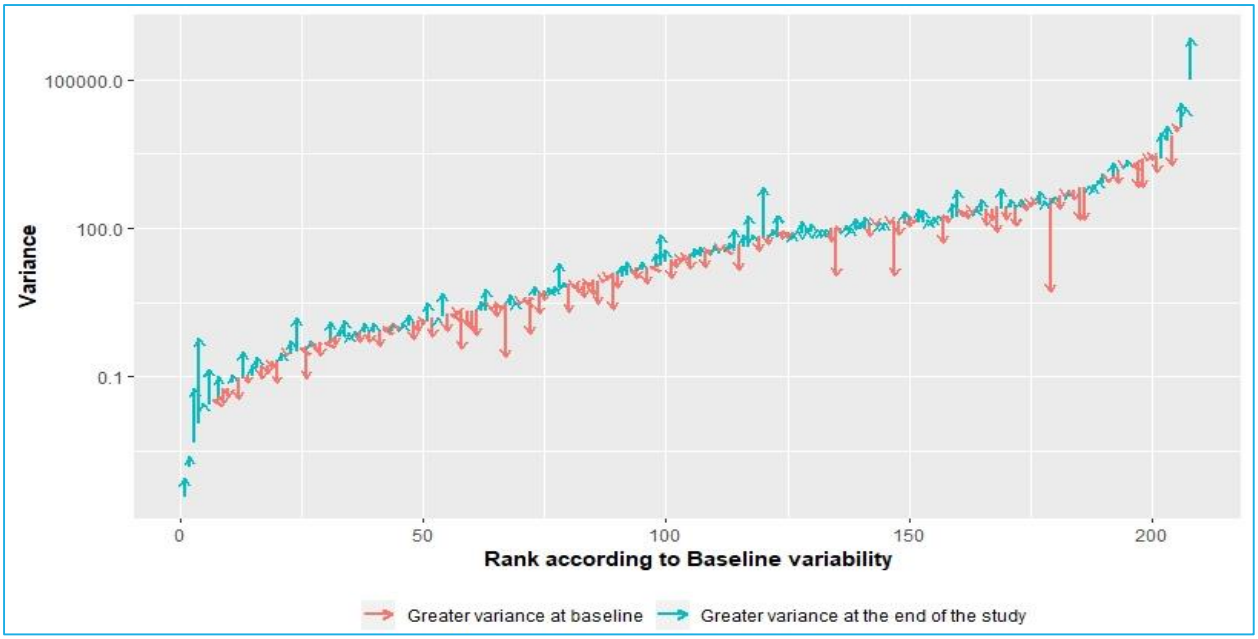


Figure 6. Arrows go from the baseline variance to the outcome variance in the treated group for each trial, and it may be increasing (blue) or decreasing (red) over time. Y-axis is log-scaled.



For the comparison over time in the treated arm, 97 (46.6%) interventions increased the variability; 101 (48.6%) decreased it; and the remaining 11 (4.8%) did not affect the dispersion of the outcome.

3 Random effects models analysis

Summary key points

- The estimate of the variance ratio both between arms and over time can be modeled by means of random effects models.
- The variance of this statistic in the model can be split into two sources of variability: the random sampling and heterogeneity among studies.
- What the models provided:
  - An estimate of the variance ratio.
  - The proportion of studies whose variability discrepancies are not simply due to random sampling, but also to real heterogeneity.
- The main results:
  - The point estimate of the variance ratio was 0.86 (treated versus control) and 0.89 (final versus baseline in treated arm), implying that experimental interventions decrease the variance on average.
  - The percentages of studies with discrepancies in variability that can only be attributed to the random sampling were 86% and 66% for the comparison between arms and over time, respectively. Precision medicine in these trials is not justified.

We applied the theory that underlies random effects models to the framework of a meta-analysis study, though we did not focus on the differences in the outcome means but instead on the ratio of the outcome variances.

Chapter 3 is divided into two large blocks that present the methods and the results of the random effects model analysis. As regards the methods (Section 3.1), we begin by providing an overview of the methodology (Section 3.1.1). Subsequently, the models for comparison between arms (Section 3.1.2) and over time (Section 3.1.3) are specified. The most challenging issue when fitting these models was distinguishing between the random variability and the heterogeneity among studies. The methodology used to estimate the former is explained in Sections 3.1.2.1 and 3.1.3.1 for both comparisons. Due to its complexity, some simulations were performed to confirm the previous theoretical results. The way to create and interpret the funnel plots resulting from the models is addressed in Section 3.1.4, and the potential bias of the estimators of the model is assessed in Section 3.1.5. The last section on the methods (Section 3.1.6) describes the factors considered in the subgroup analysis and provides examples. Section 3.2 basically contains the results for our two objectives: estimating the variance ratios (Section 3.2.1) and the proportion of constant-effect studies (Section 3.2.2); and validating the model (Section 3.2.3). It also includes the main findings of the subgroup analysis (Section 3.2.4). In addition, Appendix D shows a collateral analysis that weighs the concordance between both comparisons.

3.1 Methods

3.1.1 Explanation

A random effects meta-analysis model assumes that the observed estimates of treatment effect can vary across studies either because of real differences in the effect in each study (*heterogeneity*) or due to the sampling process (*random variability*). Thus, even if all the involved studies had an infinite sample size, the observed study effects would still vary because of the real differences in "treatment" effects. Such heterogeneity is caused by differences in study populations, interventions received, follow-up length, and other factors[31]. For us, the need for this kind of model is clear since all of these settings are completely diverse in collected studies: the only common link among them was that they had a quantitative outcome, but there was no other commonality beyond this.

In our context, the model response to be estimated was:

$$\hat{y}_i = log\left(\frac{S_{i1}^2}{S_{i0}^2}\right)$$

where S_{i1}^2 is the sample variance of the outcome of the experimental arm for the *i-th* study and S_{i0}^2 is either the variance of the outcome in the reference arm (comparison between-arms) or the variance of the outcome at baseline in the treated arm (comparison over time). From now on, we will refer to both variance ratios and their logarithms as measures of "*variability discrepancy*" or "*variance discrepancy*". Later, it will be explained why this response is log-transformed.

We dealt with the problem of combining the estimated variance discrepancies from a series of *k* RCTs, with sample sizes n_{Ti} and n_{Ci} in treated and control groups, respectively. A random effects model for meta-analysis stipulates that the observed $\hat{y}_i$, from the *i*-th study is made up of two additive components: the true difference for the specific study (θ_i) and the sampling error (e_i).

$$y_i = \theta_i + e_i \qquad e_i \sim N(0, v_i^2) \qquad i = 1, \ldots, k$$

The variance v_i^2 of e_i is the sampling variance reflecting within-study variability, which especially depends on the sample size. This *within-study variance* or *sampling variance* is usually unknown, and conventional meta-analysis studies estimate this using the data of the *i-th* observed trial. However, we did not have access to the raw data and had to make some assumptions in order to estimate it. In addition to the sampling error associated with each study, the random effects model includes a component for a true variability discrepancy among trials influenced by several other factors. That is to say, the model explicitly accounts for possible heterogeneity and stipulates that $\theta_i = \mu + T_i$, where θ_i is the true variance discrepancy in the *i-th* study; μ is the expected variance

discrepancy for the entire population; and $T_i = \theta_i - \mu$ is the deviation of the *i*-th study's effect from μ. The variance (τ^2) of T_i is the *inter-study variance* or *heterogeneity* and represents the degree to which true variance discrepancy vary across studies. The special case of $\tau^2 = 0$ would represent lack of heterogeneity, i.e., the variance discrepancies are all equal among studies and their common value is μ.[32]

In Subsections 3.1.2 (between arms) and 3.1.3 (over time), the models for both comparisons are specified.

3.1.2 Between-arm comparison

For measuring the variance discrepancy in between-arm comparison, we fitted a random effects model for the logarithm of the outcome variance ratio (at the end of the study) as the response, and this included the study as a random effect, T_i while the logarithm of the variance ratio at baseline was a fixed effect.

Between arms model
$$\log\left(\frac{S_{OT}^2}{S_{OC}^2}\right)_i = \mu + T_i + \beta \cdot \log\left(\frac{S_{BT}^2}{S_{BC}^2}\right)_i + e_i \quad with \quad e_i \sim N(0, v_i^2) \quad and \quad T_i \sim N(0, \tau^2)$$

S_{OT}^2, S_{OC}^2 represent the sample variances of the outcome in each arm at the end of the study and S_{BT}^2, S_{BC}^2 serves as the counterparts at baseline. The μ parameter is the expected value of the response across all the studies; T_i is the deviation of the *i*-th study's effect from μ, which is assumed to be Normally distributed with variance τ^2 that represents the heterogeneity associated with the study population; β is the coefficient for the model response measured at baseline; and e_i represents the sampling error associated with each trial, which is also Normally distributed with variance v_i^2.

As there was only one available measure (primary endpoint) for each study, the data did not allow us to differentiate both sources of variability: (i) the random sampling (also called within-study) variability (v_i^2); and (ii) heterogeneity (τ^2). To isolate the latter, the former was theoretically estimated using a method explained in Section 3.1.2.1 which provides the following estimate:

$$\hat{v}_i^2 = \frac{2}{n_{OT_i} - 2} + \frac{2}{n_{OC_i} - 2}$$

with n_{OT_i} and n_{OC_i} being the final sample sizes in the experimental and reference arms in the *i*-th study, respectively.

In addition, we explored the models for forced values of β:

- $\beta = 0$. The model that does not take into account the baseline heteroscedasticity.
- $\beta = 1$. The model equivalent to the analysis of change.

Table 5 shows a complete definition of the models involved in the between-arm comparison. In any case, it is well known that the most efficient analysis is the one that does not place any restriction on the estimate of the β parameter[33].

Table 5. Model descriptions for between-arm comparison.

Response	β	ID	Model
$\log\left(\frac{S_{OT}^2}{S_{OC}^2}\right)$ Between arms discrepancy	Unadjusted. $\beta = 0$	1	$\text{Log}(S_{OT}^2/S_{OC}^2)_i = \mu + T_i + e_i$
	Adjusted. Estimated β	2	$\text{Log}(S_{OT}^2/S_{OC}^2)_i = \mu + T_i + \beta_1 \cdot \log(S_{BT}^2/S_{BC}^2)_i + e_i$
	Adjusted (offset). $\beta = 1$	3	$\text{Log}(S_{OT}^2/S_{OC}^2)_i = \mu + T_i + 1 \cdot \log(S_{BT}^2/S_{BC}^2)_i + e_i$

ID: model identifier for later use

3.1.2.1 Standard error of $\log(S_{OT}^2/S_{OC}^2)$ in independent samples (between arms)

The aim of this section is to prove that the standard error of the statistic logarithm of the variance ratio is:

$$\sqrt{\frac{2}{n_{OT_i} - 2} + \frac{2}{n_{OC_i} - 2}}$$

Proof

First step. We need to prove the next convergence in distribution (D):

$$\xi = \sqrt{n-1} \cdot (S_n^2 - \sigma^2) \xrightarrow{D} N(0, 2\sigma^4)$$

Let S_n^2 be the sample variance. The statistic expressed below follows a χ^2 distribution with $(n-1)$ degrees of freedom provided that the sample comes from a Normal distribution:

$$\frac{(n-1) \cdot S_n^2}{\sigma^2} \sim \chi_{n-1}^2$$

Knowing the expected value $[(n-1)]$ and the variance $[2(n-1)]$ of a random variable with a χ^2 distribution, the expected value and variance of the sample variance are:

$$E\left[\frac{(n-1) \cdot S_n^2}{\sigma^2}\right] = \frac{n-1}{\sigma^2} E[S_n^2] = (n-1) \rightarrow E[S_n^2] = \sigma^2$$

$$V\left[\frac{(n-1) \cdot S_n^2}{\sigma^2}\right] = \frac{(n-1)^2}{\sigma^4} V[S_n^2] = 2(n-1) \rightarrow V[S_n^2] = \frac{2\sigma^4}{n-1}$$

And the expected value and the variance of ξ can be easily deducted:

$$E\left[\sqrt{n-1}\cdot(S_n^2-\sigma^2)\right]=\sqrt{n-1}\cdot E[S_n^2-\sigma^2]=\sqrt{n-1}\cdot(E[S_n^2]-\sigma^2)=0$$

$$V\left[\sqrt{n-1}\cdot(S_n^2-\sigma^2)\right]=(n-1)\cdot V[S_n^2-\sigma^2]=(n-1)\cdot V[S_n^2]=(n-1)\frac{2\sigma^4}{n-1}=2\sigma^4$$

The convergence to a Normal distribution is explained by the convergence of the *Maximum Likelihood Estimators* to this distribution.

Second step. Deducting the variance of the logarithm of the sample variance. The univariate version of the *Delta* method[34] (*DM*) or equivalently, the use of the approximation by the Taylor series allows us to estimate this variance. Let θ and σ^2 be constant values and let X_n be a sequence of random variables satisfying:

$$\sqrt{n}\cdot(X_n-\theta)\xrightarrow{D}N(0,\sigma^2)$$

The first order approximation of the Taylor series for X_n of any function g satisfying that $g'(\theta)$ exists and is non-zero can be expressed as:

$$g(X_n)\approx g(\theta)+g'(\theta)\cdot(X_n-\theta)$$

In our specific context, $\theta=\sigma^2$ and $g(x)=\log(x)$ imply:

$$\log(S_n^2)\approx\log(\sigma^2)+\frac{1}{\sigma^2}\cdot(S_n^2-\sigma^2)$$

We can now calculate the target variance:

$$V[\log(S_n^2)]\approx V\left[\log(\sigma^2)+\frac{1}{\sigma^2}\cdot(S_n^2-\sigma^2)\right]=V[\log(\sigma^2)]+V\left[\frac{1}{\sigma^2}\cdot(S_n^2-\sigma^2)\right]=0+\frac{1}{\sigma^4}V[(S_n^2-\sigma^2)]$$

$$=\frac{1}{\sigma^4}V[S_n^2]=\frac{1}{\sigma^4}\cdot\frac{2\sigma^4}{n-1}=\frac{2}{n-1}$$

Therefore, the standard error (SE) of the log-sample variance is approximated by $\sqrt{2/(n-1)}$ and the within-variance for our statistic in the model can be deducted assuming independence between groups:

$$V\left[\log\left(\frac{S_{OT}^2}{S_{OC}^2}\right)\right]=V[\log(S_{OT}^2)-\log(S_{OC}^2)\,]=V[\log(S_{OT}^2)\,]+V[\log(S_{OC}^2)\,]-$$

$$2Cov[\log(S_{OT}^2),\log(S_{OC}^2)]\overset{independent}{=}V[\log(S_{OT}^2)\,]+V[\log(S_{OC}^2)]=\frac{2}{n_{OT_i}-1}+\frac{2}{n_{OC_i}-1}$$

The approximated SE with this methodology is the square root of this variance:

$$SE_{DM}=\sqrt{\frac{2}{n_{OT_i}-1}+\frac{2}{n_{OC_i}-1}}$$

Although it may seem anti-intuitive, the variance of this statistic does not depend on the magnitude of the variability of the random variable but only on the sample size of each group.

As the sample sizes of the collected trials were not very large and this result relies on asymptotic properties, we tried to correct the obtained expression in a way to what is done in the estimation

of GEE models,[35] in which a degrees-of-freedom-adjusted strategy is proposed to correct the bias in small sample size scenarios.

$$V\left[\log\left(\frac{S_{OT}^2}{S_{OC}^2}\right)\right] \approx \frac{2}{n_{OT_i}-2}+\frac{2}{n_{OC_i}-2} \rightarrow SE_{DM}^{C} \approx \sqrt{\frac{2}{n_{OT_i}-2}+\frac{2}{n_{OC_i}-2}}$$

We verified by simulation that this approach provides a better fit than the first approximation, which underestimates the actual variance in small-sample scenarios.

Comparing SE_{DM} and SE_{DM}^{C} by simulation

A simulation was carried out in order to assess the performance of both estimators. We compared empirical estimates (SE_E) of the standard error of the log-variance coming from the simulation with the previous analytically deducted estimators (SE_{DM} and SE_{DM}^{C}). Two different scenarios were considered to carry out this comparison:

- **Setting A**. A wide range of sample sizes and equal variances in both arms.
- **Setting B**. Simulation parameters based on collected sample sizes and sample variances.

Setting A

The next box contains the parameter settings to perform this simulation.

Parameters of the simulation

- 100,000 replications
- (Constant) variance of the first group (σ_1^2): 1
- (Constant) variance of the second group (σ_2^2): 1
- Total Sample size (n): 12, 24, 48, 96, 192, 384, 768, 1536
- Sample size in control group (n_2): $n/2, n/3, n/4$

The empirical estimator SE_E was obtained according to the procedure explained in the box below.

Estimator SE_E

For each iteration:

- Simulate two samples with sizes n_1 and n_2 from the Standard Normal distribution ($\mu_1 = \mu_2 = 0$; σ_1^2=σ_2^2=1).
- Calculate the sample variance (S_1^2 and S_2^2) for each sample.
- Calculate the statistic $\log(S_1^2/S_2^2)$

The empirical SE_E is the **standard deviation** of the last statistic through all the iterations.

Figure 7 shows the ratios of these two estimators with SE_E. The different color lines are function of the sample size (x-axis) and the allocation ratio (panels). With small sample sizes, it can be seen

that the variance obtained from the delta method (SE_{DM}) infra-estimates the real variance, especially with unbalanced groups. This could be a problem, since 39% of our studies have sample sizes below 50 (Table 6 shows the mean and some quantiles). In contrast, the correction of the delta method estimator (SE_{DM}^{C}) lessens this divergence.

Figure 7. Ratios between theoretical and empirical standard errors as function of the total sample size (x-axis) and allocation ratio (panels) between arms: SE_{DM}/SE_E (red lines) and SE_{DM}^{C}/SE_E (blue lines). Y-axes are log-scaled.

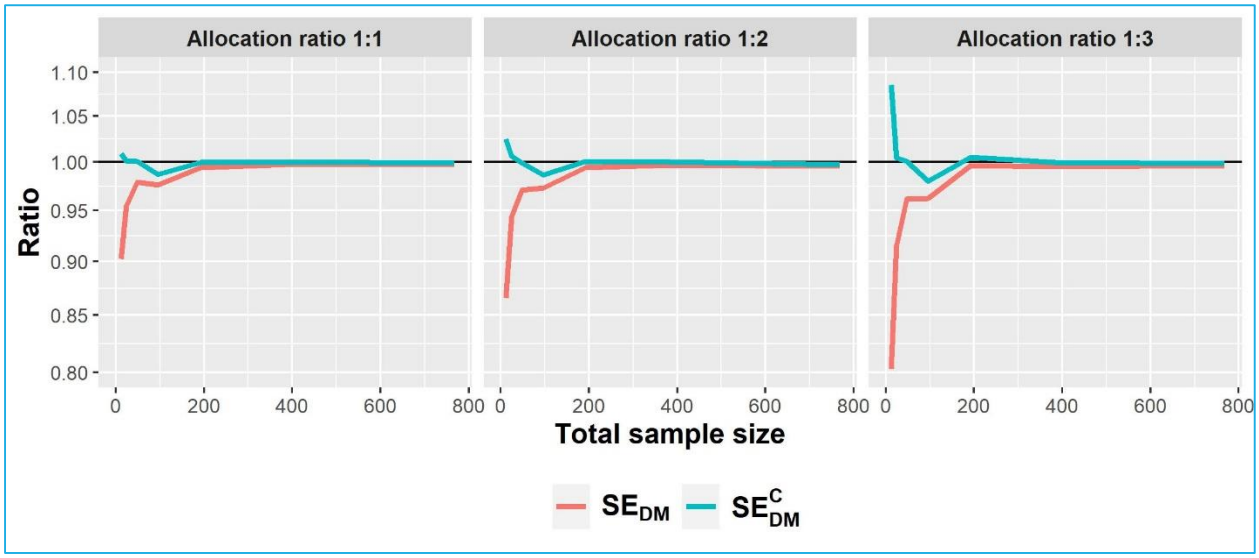


Table 6. Descriptive statistics for the overall/total sample sizes of collected studies.

Mean	Min.	Q1	Median	Q3	Max.
119.40	12.00	37.75	66.00	161.20	852.00

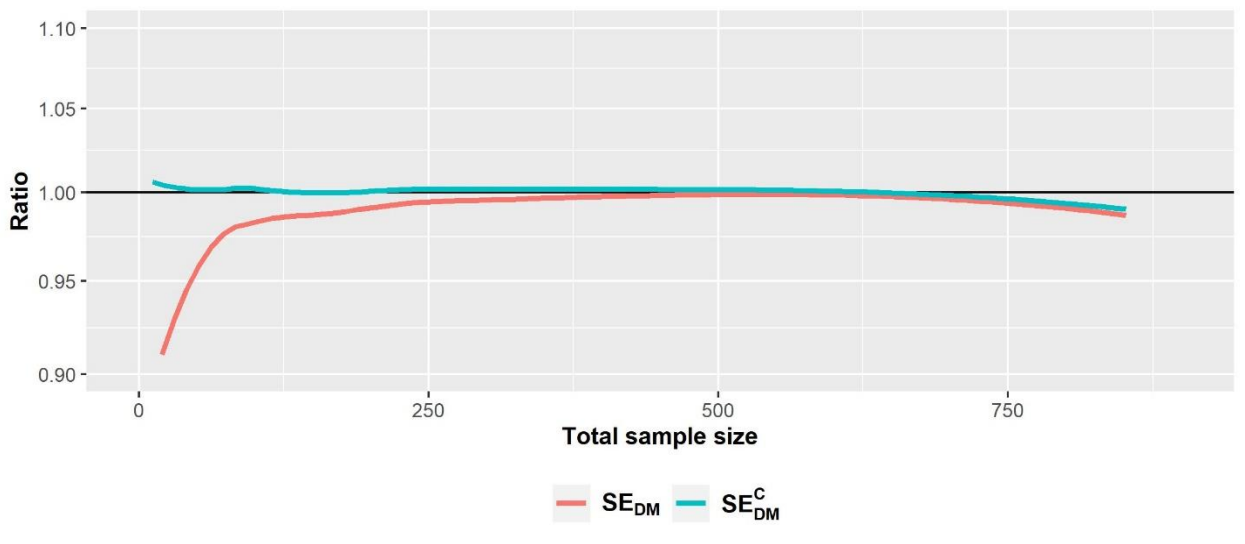
Setting B

Once we saw that SE_{DM}^{C} outperforms the same statistic without correction in a wide range of sample size scenarios, we decided to explore if the conclusions hold by conducting another simulation using settings that were closer to the observed data, i.e., with the sample sizes and variances obtained from the 208 collected trials.

Parameters of the simulation	
❖ 10,000 replications ❖ Real variance of the first group (σ_{1i}^2): $S_{OT_i}^2$ ❖ Real variance of the second group (σ_{2i}^2): $S_{OC_i}^2$ ❖ Sample size in first group (n_{1i}): n_{OT_i} ❖ Sample size in second group (n_{2i}): n_{OC_i}	$S_{OT_i}^2$: sample outcome variance in treated arm in the *i*-th study $S_{OC_i}^2$: sample outcome variance in control arm in the *i*-th study n_{OT_i}: sample size in the experimental arm in the *i*-th study n_{OC_i}: sample size in the control arm in the *i*-th study

In Figure 8, discrepancies among both estimators SE_{DM}^{C} and SE_E were pointless (<1%) and did not depend on sample size or the unbalanced sample variances. Larger discrepancies between SE_E and SE_{DM} were present for very small sample sizes. For instance, SE_{DM} underestimated the actual SE by more than 8% in those studies with less than 30 patients and by up to 5% in trials with around 50 participants.

Figure 8. Ratios SE_{DM}/SE_E (red line) and SE_{DM}^{C}/SE_E (blue line) vs. the total sample size (x-axis). Y-axis is log-scaled.



3.1.3 Comparison over time

An analogous model was employed to assess the homoscedasticity over time.

Over time model

$$\log\left(\frac{S_{OT}^2}{S_{BT}^2}\right)_i = \mu + T_i + \beta \cdot \log\left(\frac{S_{OC}^2}{S_{BC}^2}\right)_i + e_i \quad with \quad e_i \sim N(0, v_i^2) \ and \ T_i \sim N(0, \tau^2)$$

Note that the only difference from the between-arm model is that the terms S_{OC}^2 and S_{BT}^2 are interchanged. Again, we needed to distinguish between the two sources of variability. The random sampling variability deducted in Section 3.1.3.1 providing the following estimated within-study variance:

$$\hat{v}_i^2 = \frac{4}{n_{OT_i} - 1} - 2log\left(1 + \frac{2 \cdot r_i(Y_{OT}, Y_{BT})^2}{n_{OT_i}^2/(n_{OT_i} - 1)}\right)$$

where $r_i(Y_{OT}, Y_{BT})$ is the sample correlation between outcomes measured at baseline and at the end of the *i*-th study in the experimental arm. Therefore, for this model involving paired samples, we

required the covariance or the correlation (or some information that allows us to deduct them, such as the variance or the standard error of the difference between baseline and final outcomes) between the logarithm of the variances (final and initial). Only 95 studies had reported this information and were used in this analysis.

As in the between-arm model, we explored the models for forced values of β ($\beta = 0$ and $\beta = 1$). See Table 7 for the expression of the models.

Table 7. Model descriptions for comparison over time.

Response	β	ID	Model
$\log\left(\frac{S_{OT}^2}{S_{BT}^2}\right)$ Over time discrepancy	Unadjusted. $\beta = 0$	4	$\text{Log}(S_{OT}^2/S_{BT}^2)_i = \mu + T_i + e_i$
	Adjusted. Estimated β	5	$\text{Log}(S_{OT}^2/S_{BT}^2)_i = \mu + T_i + \beta_1 \cdot \log(S_{OC}^2/S_{BC}^2)_i + e_i$
	Adjusted (offset). $\beta = 1$	6	$\text{Log}(S_{OT}^2/S_{BT}^2)_i = \mu + T_i + 1 \cdot \log(S_{OC}^2/S_{BC}^2)_i + e_i$

ID: model identifier for later use

3.1.3.1 Standard error of $\log(S_{OT}^2/S_{BT}^2)$ in paired samples (over time)

In order to estimate the variance of the logarithm of the sample variances ratios for the paired case (over time), we need to estimate the covariance between the logarithms of the sample variances.

$$V\left[log\left(\frac{S_{OT}^2}{S_{BT}^2}\right)\right] = V[\log(S_{OT}^2) - log(S_{BT}^2)] = V[\log(S_{OT}^2)] + V[\log(S_{BT}^2)] - 2Cov[\log(S_{OT}^2), \log(S_{BT}^2)]$$

The variance of the log-variance can be estimated as in Section 3.1.2.1:

$$V[\log(S_{OT}^2)] = \frac{2}{n_{OT_i} - 2} \qquad V[\log(S_{BT}^2)] = \frac{2}{n_{BT_i} - 2}$$

The next boxes show how to estimate the covariance by dividing the procedure into three steps. The reason for splitting the proof into three parts is to later assess by simulation the approximations made after each step.

Proof – First Step

Demonstrate: $\boldsymbol{Cov\left[\log(S_{OT}^2), \log(S_{BT}^2)\right] \approx \log\left[1 + \frac{Cov[S_{OT}^2, S_{BT}^2]}{E[S_{OT}^2] \cdot E[S_{BT}^2]}\right]}$

We are going to define the following random variables:

$U = \log(S_{OT}^2)$ $V = \log(S_{BT}^2)$ $W = U + V$

$X = \exp(U) = S_{OT}^2$ $Y = \exp(V) = S_{BT}^2$

We will use the following two properties:

- Relationship between covariance and expected values for X, Y and their product:

$$Cov(X,Y) = E(XY) - E(X)E(Y)$$

- Taylor expansion of *X:*

$$E[X] = E[\exp(U)] \overset{(1)}{\approx} \mathrm{E}\left[\exp(\mu_U) + \frac{\exp(\mu_U)}{1}(U-\mu_U) + \frac{\exp(\mu_U)}{2}(U-\mu_U)^2\right] =$$

$$= \exp(\mu_U) + \exp(\mu_U)\cdot E[U-\mu_U] + \frac{\exp(\mu_U)}{2}E[(U-\mu_U)^2] =$$

$$= \exp(\mu_U) + 0 + \frac{\exp(\mu_U)}{2}\sigma_U^2 = \exp(\mu_U)\cdot\left[1+\frac{\sigma_U^2}{2}\right] \overset{(2)}{\approx} \exp(\mu_U)\cdot\exp\left(\frac{\sigma_U^2}{2}\right) = \exp\left(\mu_U + \frac{\sigma_U^2}{2}\right)$$

(1) *Second order Taylor expansion around* μ_U: $exp(U) \approx \exp(\mu_U) + \frac{\exp(\mu_U)}{1}(U-\mu_U) + \frac{\exp(\mu_U)}{2}(U-\mu_U)^2$

(2) *First order Taylor expansion around 0*: $\exp\left(\frac{\sigma_U^2}{2}\right) \approx 1 + \frac{\sigma_U^2}{2}$

[This approximation is exact for *U* Normal and good enough for not too large departures from Normality]

We can express the expected value for the X and Y product as follows:

$$E[XY] = E[\exp(U)\cdot\exp(V)] = E[exp(W)] \approx \exp\left(\mu_W + \frac{\sigma_W^2}{2}\right)$$

The variance of W is:

$$Var(W) = Var(U) + Var(V) + 2Cov(U,V)$$

Using the last two expressions, we can obtain another one for the exponential of the covariance:

$$1 + \frac{Cov(X,Y)}{E(X)E(Y)} = \frac{E(XY)}{E(X)E(Y)} = \frac{\exp\left(\mu_W + \frac{\sigma_W^2}{2}\right)}{\exp\left(\mu_U + \frac{\sigma_U^2}{2}\right)\cdot\exp\left(\mu_V + \frac{\sigma_V^2}{2}\right)}$$

$$= \frac{\exp\left(\mu_U + \mu_V + \frac{1}{2}\cdot(\sigma_U^2 + \sigma_V^2 + 2Cov(U,V))\right)}{\exp\left(\mu_U + \frac{\sigma_U^2}{2}\right)\cdot\exp\left(\mu_V + \frac{\sigma_V^2}{2}\right)} \approx exp[Cov(U,V)]$$

Therefore:

$$Cov(U,V) \approx log\left(1 + \frac{Cov(X,Y)}{E(X)\cdot E(Y)}\right) \rightarrow Cov[\log(S_{OT}^2), \log(S_{BT}^2)] = \log\left[1 + \frac{Cov[S_{OT}^2, S_{BT}^2]}{E[S_{OT}^2]\cdot E[S_{BT}^2]}\right]$$

$$\approx \log\left[1 + \frac{Corr[S_{OT}^2, S_{BT}^2]\cdot\sqrt{V[S_{OT}^2]\cdot V[S_{BT}^2]}}{S_{OT}^2\cdot S_{BT}^2}\right]$$

Proof – Second step

Demonstrate: $\boldsymbol{Corr[S_{OT}^2, S_{BT}^2] \approx Corr[Y_{OT}, Y_{BT}]^2}$

Some authors have proposed methods for estimating the covariance of the two sample variances using sample cumulants and k-statistics[36], but information about the moments of order higher than 2 was either not available or

these methods were not suitable in our setting. As an alternative, an approximated relationship between the abovementioned correlations is proposed by Muirhead et al.[37]: the squared correlation of the outcomes is a good enough approximation of the correlation between the sample variances of the outcomes. Therefore, we can reformulate the expression obtained in step 1.

$$\log\left[1+\frac{Corr[S_{OT}^2,S_{BT}^2]\cdot\sqrt{\mathrm{V}[S_{OT}^2]\cdot V[S_{BT}^2]}}{S_{OT}^2\cdot S_{BT}^2}\right]=\log\left[1+\frac{Corr[Y_{OT},Y_{BT}]^2\cdot\sqrt{\mathrm{V}[S_{OT}^2]\cdot V[S_{BT}^2]}}{S_{OT}^2\cdot S_{BT}^2}\right]$$

Proof – Third step

Demonstrate: $\frac{2s_{XT}^4}{df_{XT}}$ **is an estimator of** $V\left[S_{XT}^2\right]$

Let Y be a random variable defined as:

$$Y=\frac{(n-1)\cdot S^2}{\sigma^2}\sim\chi_{n-1}^2$$

As we have already mentioned:

$$V\left[\frac{(n-1)\cdot S^2}{\sigma^2}\right]=V[\mathrm{Y}]\rightarrow V[S^2]=\frac{\sigma^4}{(n-1)^2}\cdot V[\mathrm{Y}]=\frac{\sigma^4}{(n-1)^2}\cdot 2\cdot(n-1)=\frac{2\sigma^4}{n-1}$$

So:

$$\widehat{V[S^2]}=\frac{2s^4}{n-1}=\frac{2s^4}{df}$$

where lowercase *s* is the estimate of the parameter σ. The expression obtained in step 2 can be reformulated:

$$Cov(U,V)=Cov[\log(S_{OT}^2),\log(S_{BT}^2)]=\log\left[1+\frac{Corr[S_{OT}^2,S_{BT}^2]\cdot\sqrt{\mathrm{V}[S_{OT}^2]\cdot V[S_{BT}^2]}}{S_{OT}^2\cdot S_{BT}^2}\right]$$

This is estimated by:

$$\widehat{Cov(U,V)}=\log\left[1+\frac{Corr[Y_{OT},Y_{BT}]^2\cdot\sqrt{\frac{2s_{OT}^4}{df}\cdot\frac{2s_{BT}^4}{df}}}{s_{OT}^2\cdot s_{BT}^2}\right]\approx\log\left[1+\frac{Corr[Y_{OT},Y_{BT}]^2\cdot\frac{2s_{OT}^2\cdot s_{BT}^2}{df}}{s_{OT}^2\cdot s_{BT}^2}\right]$$

$$\approx\log\left[1+\frac{2\cdot Corr[Y_{OT},Y_{BT}]^2}{df}\right]$$

We have achieved a complete expression for the variance of the logarithm of the variance ratio:

$$V\left[log\left(\frac{S_{OT}^2}{S_{BT}^2}\right)\right]=\frac{2}{df_{OT}-1}+\frac{2}{df_{OB}-1}-2\cdot\log\left[1+\frac{2\cdot Corr[Y_{OT},Y_{BT}]^2}{df_{OT}}\right]$$

The appropriate degrees of freedom to calculate the covariance should be df_{OT} (not df_{OB}) since this is based on those patients with complete follow-up information.

Validation by simulation

A simulation was performed in order to assess if this approximation was good enough to be used in the random effects model.

Parameters of the simulation

- Replications ($nsim$) for each fixed σ^2: 1,000
- SD (σ) for the change over time from baseline to the end of the study in the treated arm:
  - Minimum σ (σ_{min}): $0.1 \rightarrow Equivalent\ to\ \rho \approx 0.99$
  - Maximum σ (σ_{max}): $10 \rightarrow Equivalent\ to\ \rho \approx 0.10$
  - Values: $\sigma_{min} = 0.1, 0.1e^{0.02}, 0.1e^{0.04}, \ldots, 0.1e^{4.6}, 10 = \sigma_{max}$
- Sample size by group (n): 100

Remark: σ_{min} and σ_{max} values were chosen to cover a wide spectrum of correlations*. The sequence of values (232 in total) was chosen based on a geometric series, thus making the values of the correlations more equidistant.

*Knowing how the baseline values (Z) and final outcomes (Y) were generated, the relationship between σ and ρ can be deducted:

$$\left.\begin{matrix} Z \sim N(0,1) \\ C \sim N(0,\sigma) \end{matrix}\right\} \rightarrow Y = Z + C \sim N\left(0, \sqrt{1+\sigma^2}\right)$$

$$Cov(Z,Y) = Cov(Z, Z+C) = Cov(Z,Z) + Cov(Z,C) = V(Z) = 1$$

$$\rho_{Z,Y} = \frac{Cov(Z,Y)}{\sigma_Z \cdot \sigma_Y} = \frac{1}{1 \cdot \sqrt{1+\sigma^2}} = \frac{1}{\sqrt{1+\sigma^2}}$$

Regarding to the simulation parameters, several scenarios with different treatment effects were tested, but this parameter did not lead to any influence at all in the results.

Simulation procedure

- For each sigma(σ^2) value and for each iteration ($nsim$):
  a. Generate a Standard Normal sample $Z \sim N(0,1)$ representing the baseline values.
  b. Let $C \sim N(0,\sigma)$ be the change over time, so the response at the end of the study is generated as $Y=Z+C$.
  c. Estimate the correlation and covariance between Y and Z.
  d. Estimate the variances for Y and Z.
- Calculate the covariance of the sample variances for the $nsim$ iterations.
- Calculate the means of the correlation and covariance between Y and Z for the $nsim$ iterations, respectively.
- Store the different indicators before/after each step of the proof:

a. Left-hand term: $Cov(S_{OT}^2, S_{BT}^2)$

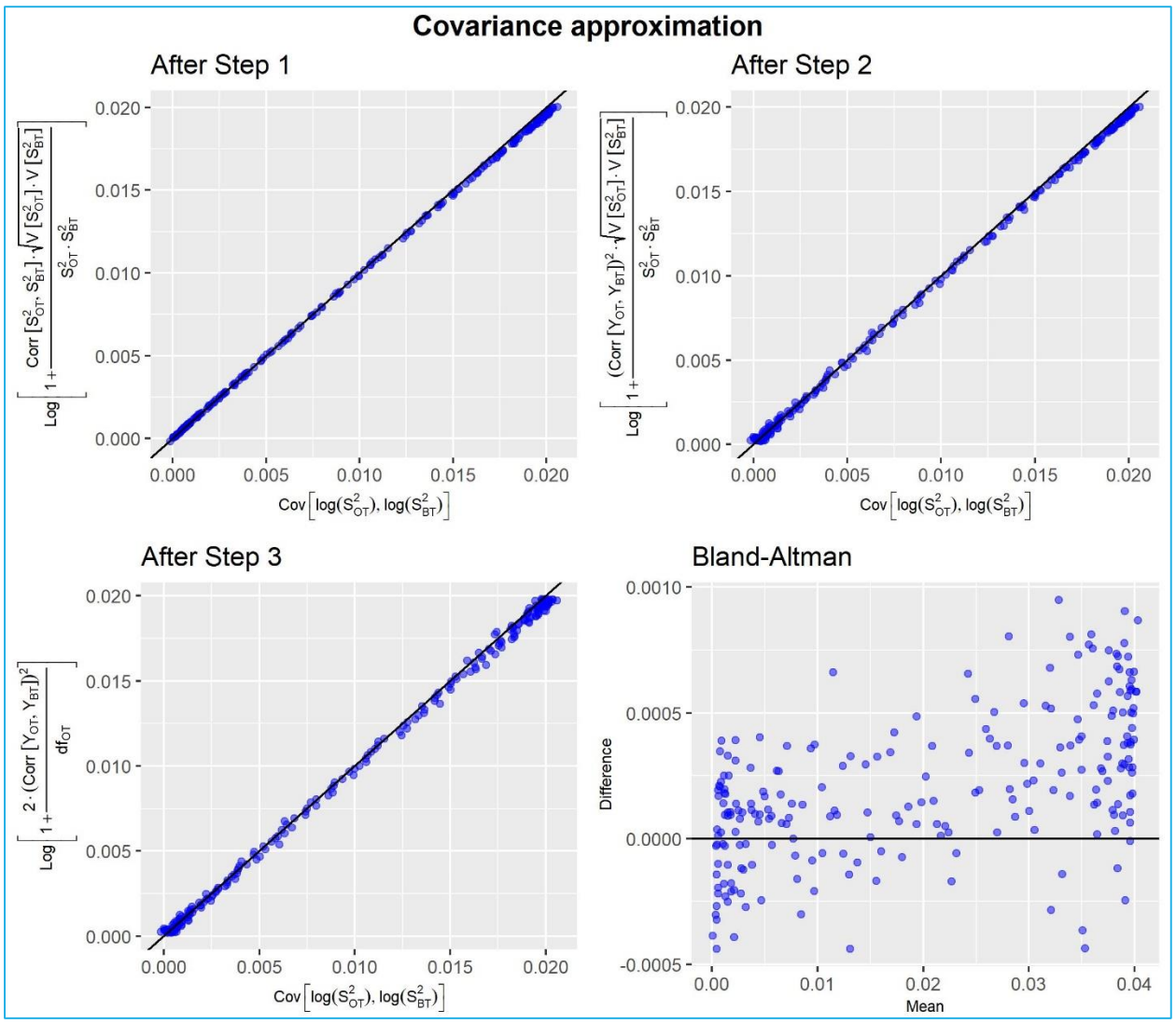
b. Right-hand term after the first step: $log\left[1 + \frac{Corr[S_{OT}^2, S_{BT}^2] \cdot \sqrt{V[S_{OT}^2] \cdot V[S_{BT}^2]}}{S_{OT}^2 \cdot S_{BT}^2}\right]$

c. Right-hand term after the second step: $log\left[1 + \frac{Corr[Y_{OT}, Y_{BT}]^2 \cdot \sqrt{V[S_{OT}^2] \cdot V[S_{BT}^2]}}{S_{OT}^2 \cdot S_{BT}^2}\right]$

d. Right-hand term after the third step: $log\left[1 + \frac{2 \cdot Corr[Y_{OT}, Y_{BT}]^2}{df_{OT}}\right]$

The quality of the approximations obtained is shown in Figure 9, which compares the left-hand side of the formula with the different right-hand sides after each step.

Figure 9. The three first panels: scatterplot between the covariance of the log-variances (x-axis) and the right-hand expression achieved after each step of the previous proof. The last bottom-right panel: Bland-Altman plot of the concordance between the left- and right-hand sides of the final expression.



The approximation is almost perfect after the first step (absolute difference in mean between real value and the approximated value, $0.17 \cdot 10^{-3}$) and good enough after second ($0.26 \cdot 10^{-3}$) and third ($0.28 \cdot 10^{-3}$) ones. The last panel represents the Bland-Altman plot, which assesses the concordance between the left-hand side of the formula and the last expression; the growing trend of the points on this plot indicates a slightly poorer performance of the approximation as the covariance between the log-variance increases: a difference between the approximate and real values equal to $0.5 \cdot 10^{-3}$ is equivalent to a relative increase of 1.25% in the covariance values for larger magnitudes around 0.04.

3.1.4 Funnel plots

Centered at zero, pseudo-funnel plots for the measurement of interest (variance discrepancy) as a function of its standard error are reported in order to help investigate asymmetries. We use the term *pseudo* because they are centered on the 0 value instead of the point estimate. The aim of representing these plots is to emphasize potential asymmetries in the variance ratios. As a complement to the visual inspection, a non-parametric rank correlation test based on Kendall's τ statistic —similar to those used to assess publication bias in the meta-analysis context— was performed to determine such asymmetries[38]. Confidence intervals were obtained by bootstrap.

A confidence interval region within the funnel plot is drawn around zero, with bounds equal to ±1.96 SE. The usual way of representing the confidence regions in the presence of heterogeneity among studies is to consider that the SE takes into account the two sources of variability:

$$SE_i = \sqrt{v_i^2 + \tau^2}$$

However, since we were interested in discerning which studies have a variance discrepancy that is not explained by random sampling, we will only include the first term of the previous expression to represent the confidence regions:

$$SE_i = \sqrt{v_i^2}$$

This analysis related the points outside the triangle with those studies having different variability either between groups (comparison between arms) or between baseline and final variability (comparison over time).

3.1.5 Validation of the random effects model

We assessed if the results obtained using the random effects model matched our expectations. Specifically, we studied the behavior of the three main statistics of the model: $\hat{\mu}, \hat{\tau}$ and I^2. First, a simulation study was conducted, in which we used the REM of the estimated $\hat{\mu}$ and $\hat{\tau}$ to compare them with the ones in the specific parameter setting. Second, the *Jackknife* estimator was used to evaluate the bias of these three estimators.

3.1.5.1 Assessing $\hat{\mu}$, $\hat{\tau}$, I^2 in random effects model by simulation

The appropriateness of the random effects model was evaluated by simulation taking into account the results obtained after applying the corrected Delta method. We wanted to discern if the model was able to separate the within (random) and between (studies) variabilities. The next boxes show the characteristics of the conducted simulation.

Parameters of the simulation

- 100 replications
- Real variance ratio (σ_1^2/σ_2^2): 0.1, 0.2, 0.3,…,2
- Real heterogeneity (τ): 0, 0.1 ,0.2 ,…1

Simulation procedure

For each iteration and each combination of simulation parameters:

- The response is generated according to the next structure:

$$y_i = Real\ Variance\ Ratio + Heterogeneity + Random\ Error$$

$$y_i = \mu + T_i + \epsilon_i \qquad T_i \sim N(0,\tau) \quad \epsilon_i \sim N(0, SE_{DM}^C)$$

$$y_i = log\left(\frac{\sigma_1^2}{\sigma_2^2}\right) + N(0,\tau) + N\left(0, \sqrt{\frac{2}{n_{T_i}-2} + \frac{2}{n_{C_i}-2}}\right)$$

- Each model is fitted with the *rma* function from the *metafor* R package.
- The estimated μ and τ in the model were stored.

All estimated μ and τ were averaged for all iterations corresponding to the same parameter setting.

3.1.5.2 *Jackknife* estimator

Jackknife is a resampling technique that is especially useful for variance and bias estimation. The Jackknife estimator of a parameter is found by systematically leaving out each observation from a dataset and calculating the *i*-th estimate for each subsample. The final estimate is found by averaging all these estimates[39,40].

Bias estimation

Let $\hat{\theta}$ be the estimator to assess and $\hat{\theta}_{(i)}$ the same estimator applied over the whole sample without the *i*-th observation. The $\hat{\theta}_{(\cdot)}$ is the mean of all these estimators which allows us to obtain the estimated bias of an estimator calculated over the entire sample:

$$\hat{\theta}_{(\cdot)} = \frac{1}{n} \cdot \sum \hat{\theta}_{(i)} \rightarrow \widehat{Bias} = (n-1) \cdot (\hat{\theta}_{(\cdot)} - \hat{\theta})$$

3.1.6 Subgroup analysis

With the subgroup analysis, we want to know whether homoscedasticity is indeed modified or not by measurable factors such as the positive result of the trial's main objective. Conditions studied together with an example for each of their levels are explained in the box below.

Factors considered in the subgroup analysis

- **Intervention efficacy in the primary endpoint** (*No/Yes*). This refers to whether the authors rejected the null hypothesis of no effect for the primary endpoint according to their pre-specified significance level (α).
  - **Example *No*:** *Hormone treatment provided no overall protection against functional decline in nondisabled postmenopausal women 65 years or older in 6 years of follow-up*[41]
  - **Example *Yes*:** *This study shows that self-management of hypertension, consisting of regular self-measurements of blood pressure and a simple predetermined titration plan for antihypertensive drugs, is more effective in lowering systolic blood pressure than the usual care during 1 year*[42]

- **Intervention type** (*Non-pharmacological/Pharmacological*). Basically, this criterion distinguishes between studies that tested drugs or other kinds of interventions.
  - **Example *Non-pharmacological*:** Physical exercise vs. Cognitive-behavioral therapy[43]
  - **Example *Pharmacological*:** Metformin vs. Placebo[44]

- **Outcome type** (*Measured/Scored*). This variable classifies the outcomes into those that were the result of a measurement of a physical parameter and those that were the result of a rating scale.
  - **Example *Measured*:** Body weight[45]
  - **Example *Scored*:** Montgomery-Asberg depression rating scale (MADRS)[46]

- **Condition type** (*Acute/Chronic*). If the main disease considered in the eligibility criteria was acute or chronic. This classification was made with the help of a researcher with medical training. In some trials, the participants were healthy patients without any relevant pathology.
  - **Example *Acute*:** Individuals with first mild stroke[47]
  - **Example *Chronic*:** Active Crohn's disease[48]
  - **Example *Healthy*:** Late postmenopausal women[49]

- **Measurement type** (*Assessed/Automatic*). This should not be confused with the type of variable outcome. This criterion refers to whether the measurement is carried out subjectively or objectively.
  - **Example *Assessed*:** Plaque index (example of assessed and measured outcome)[50]
  - **Example *Automatic*:** Star excursion balance test (example of automatic and scored outcome)[51]

- **Improvement** (*Downwards/Upwards*). If the improvement of a patient is associated with a higher (upwards) or smaller (downwards) value of the outcome.
  - **Example *Downwards*:** Intraocular Pressure[52]
  - **Example *Upwards*:** Visual acuity[53]

3.2 Results

3.2.1 Estimate of variance ratios

The point-estimate coefficients of the REM for both comparisons are shown in the next box, and Table 8 shows the estimates for all the fitted models including 95% confidence intervals.

Estimated coefficients

Between arms $\rightarrow \log\left(\frac{S^2_{OT}}{S^2_{OC}}\right)_i = \mathbf{-0.12} + T_i + \mathbf{0.47} \cdot \log\left(\frac{S^2_{BT}}{S^2_{BC}}\right)_i + e_i \quad T_i \sim N(0, \hat{\tau}^2 = \mathbf{0.30})$

Over time $\rightarrow \log\left(\frac{S^2_{OT}}{S^2_{BT}}\right)_i = \mathbf{-0.15} + T_i + \mathbf{0.62} \cdot \log\left(\frac{S^2_{OC}}{S^2_{BC}}\right)_i + e_i \quad T_i \sim N(0, \hat{\tau}^2 = \mathbf{0.35})$

Table 8. Estimated coefficients from the random effects models.

Response in the model	Type of model	$\hat{\mu}$ (95%CI)	$\hat{\beta}$ (95%CI)	$\hat{\tau}$ (95%CI)
Outcome variance ratio	**CDB (n = 208)**			
	Unadjusted[1]	-0.11 (-0.20, -0.01)	0	0.60 (0.54, 0.71)
	Adjusted[2]	-0.12 (-0.21, -0.03)	0.48 (0.30, 0.65)	0.55 (0.49, 0.65)
	Adjusted (offset)[3]	-0.14 (-0.24, -0.04)	1	0.60 (0.55, 0.72)
Outcome vs. baseline ratio in treated group	**CDO (n = 95)**			
	Unadjusted[4]	-0.14 (-0.30, 0.01)	0	0.71 (0.61, 0.85)
	Adjusted[5]	-0.15 (-0.28, -0.02)	0.63 (0.42, 0.84)	0.59 (0.51, 0.73)
	Adjusted (offset)[6]	-0.16 (-0.29, -0.02)	1	0.62 (0.55, 0.78)

$\hat{\mu}$: Average of the model response. $\hat{\beta}$: Coefficient for the model response measured at baseline (comparison between groups) or in control arm (comparison over time). $\hat{\tau}$: standard deviation of the study's random effect (heterogeneity). Superscripts represent the fitted model identifier (see Table 5 and Table 7).

Regarding the comparison between arms, the adjusted point estimate of the ratio (experimental to reference group) of the outcome variances was 0.89 ($e^{-0.12}$). This indicates that, contrary to popular belief, treatments seem to reduce the variability of the patient's response. The comparison over time provided another interesting and similar result: the average variability at the end of the study was 14% ($e^{-0.15} = 0.86$) lower than that at baseline. This may be due to some measurements having *ceiling* or *floor* effects, whereby the top or bottom range of a variable is truncated. An example of a *ceiling* effect arises when one is unable to score better than 100% on

a quality of life measure. The previous mentioned article about the PCL-C scale based on the sum of 17 Likert symptoms[28] is an example of a *floor* effect: if the intervention reduced the number of affected symptoms, then variance, as well as average, would also be decreased.

The measures of heterogeneity could be considered as quite high keeping in mind that we are on a logarithmic scale. The value of $\tau^2 = 0.30$ in the comparison between arms implies that expected typical variations could increase typically by 73% $\left(e^{\sqrt{0.30}} = 1.73\right)$ or decrease by 42% $\left(e^{-\sqrt{0.30}} = 0.58\right)$ the variance discrepancies between one study and another. For the comparison over time, these fluctuations were even greater going from an increase of 81% $\left(e^{\sqrt{0.35}} = 1.81\right)$ to a decrease of 45% $\left(e^{-\sqrt{0.35}} = 0.55\right)$.

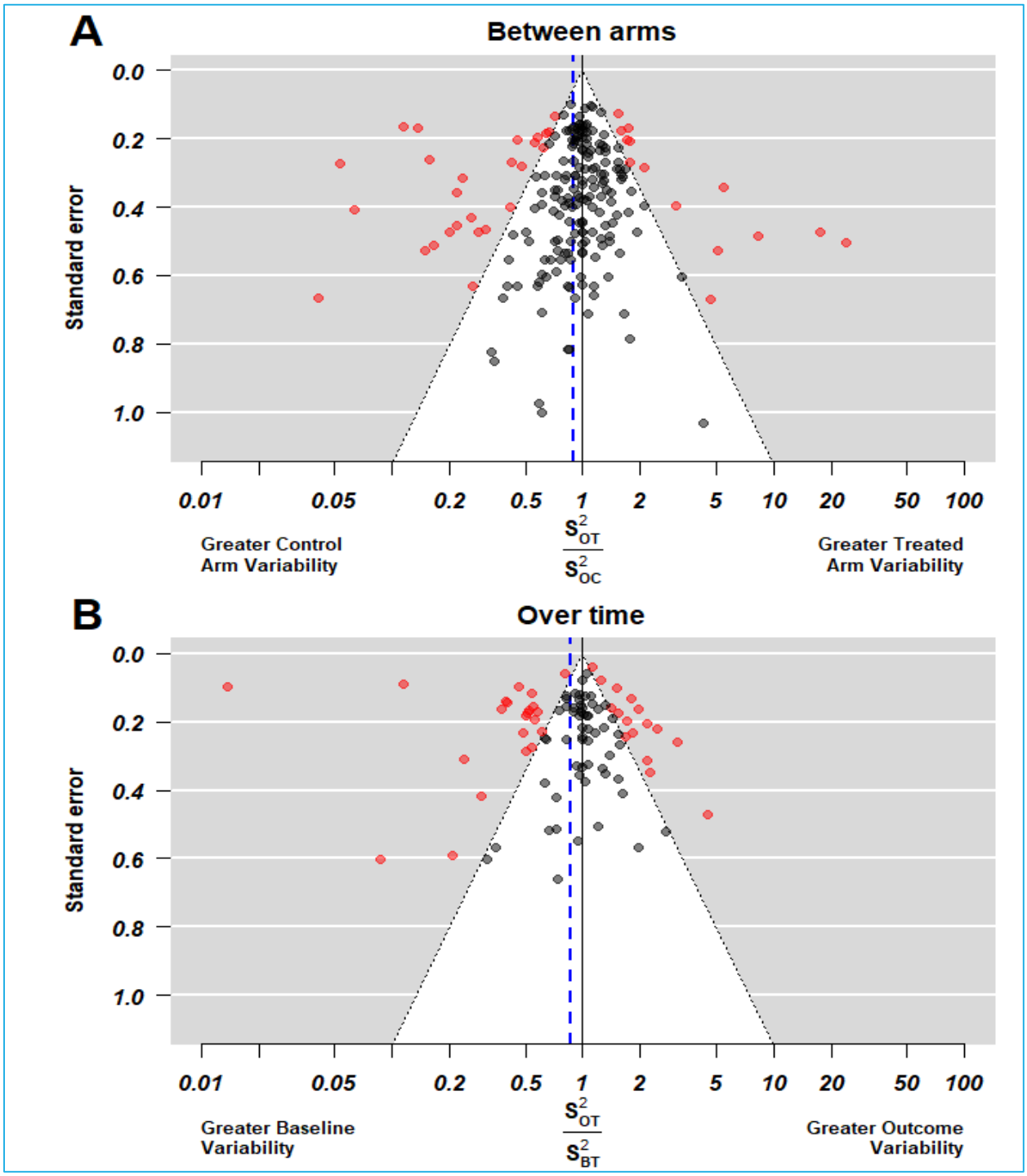
In the between-arm comparison, the value of $\hat{\beta}$ being lower than 1 ($\hat{\beta} = 0.47$) could be interpreted as a regression to the mean effect. In other words, in those trials whose randomization process produced greater variance discrepancies by chance (or due to methodological failures, as we will explain in Section 4.1.1), these discrepancies could be mitigated during the trial follow-up. In the second model, the same value of $\hat{\beta}$ being below 1 ($\hat{\beta} = 0.62$) indicated that the variations in the dispersion of the response over time were more moderate in the experimental arm.

For the between-arm comparison, the point estimate $\hat{\mu}$ ranged from -0.14 ($\beta = 1$) to -0.11 ($\beta = 0$). The analysis without forced β gave a point estimate equal to -0.12 (95% CI from -0.21 to -0.03), making the average variance ratios (treated to control) between 0.81 and 0.97 compatible with the data. For the comparison over time, the point estimate $\hat{\mu}$ ranged from -0.16 ($\beta = 1$) to -0.14 ($\beta = 0$). Without constraining the β value, the point estimate was equal to -0.15 (95% CI from -0.28 to -0.02), causing average variance ratios (final to baseline) between 0.76 and 0.98, which are compatible with our data.

3.2.2 Estimated proportion of studies with heteroscedasticity

The funnel plots of both comparisons were used to assess the percentage of studies having more extreme values of the variance ratio that those expected by chance. In Panel A of Figure 10, points on the right and on the left indicate higher and lower outcome variability for the treated patients, respectively. Points on the right in Panel B indicate higher variability in the experimental arm at the end of the study, as expected in a scenario of heterogeneous treatment effect, and the 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 reduced variability.

Figure 10. Funnel plots of variance ratio between arms with 208 studies (Panel A) and for comparison over time with the 95 studies for which the variance of the difference between the basal and final response was available (Panel B). Vertical axes represent the SE derived from the model. X axis is log-scaled.



For the between-arm comparison, the REM found 15 studies (7.2%) with greater variability and 26 (12.5%) with lower variability in the experimental arm, respectively; in the remaining 167 studies (80.3%), there was no evidence of differences in variability between groups. For the comparison over time, among the 95 studies with enough information to conduct the analysis, 16 (16.8%) of them showed greater variability at the end of the study, while 22 (23.2%) trials reduced

the outcome variability over time. In the remaining 57 studies (60.0%), there was no evidence of differences in variability regarding this comparison.

Although the visual perception of asymmetry was more marked in the funnel plot of panel B, the test to assess it in the comparison over time provided a Kendall's τ statistic equal to -0.084 (95% CI from -0.224 to 0.059), which include the null value and, thus, there was no evidence to state such asymmetry. In contrast, Kendall's τ coefficient was -0.122 (95% CI from -0.202 to -0.034) when comparing treatment groups, thus manifesting the existence of asymmetry. It is likely that the divergent conclusions in both comparisons are also partially conditioned by the number of studies involved in each analysis. Regardless of this formal analysis, both funnels plots clearly show a trend towards having less variability in the outcome of the treated arm.

3.2.3 Validation of the random effect models

3.2.3.1 Assessing $\hat{\mu},\ \hat{\tau},\ I^2$ in random effects model by simulation

We have inferred and drawn some conclusions based on the random effects models. We have assumed that the estimates provided by the fitted models using maximum likelihood estimation (MLE) are unbiased and precise. The theory about MLE already guarantees good properties based on reasonable premises. We checked the model estimates for the parameters $\hat{\mu}, \hat{\tau}$ and I^2 from the model 2 (adjusted model for between-arms comparison, see Table 5) with the values resulting from a simulation study under the scenarios explained in Section 3.1.5.1.

Figure 11 compares the real with the averaged estimated parameters throughout the 100 replications. The estimators were not biased except for very small heterogeneities where the estimated heterogeneity ($\hat{\tau}$) overestimates the actual parameter (τ).

3.2.3.2 *Jackknife* estimator

Another well-known methodology for evaluating the bias of an estimator is the *Jackknife* method. This technique allows obtaining an alternative estimate whose proximity to the initial estimate is an indicator of absence of bias. Figure 12 shows that the three estimators of μ, τ and $I^2$are not biased since the *Jackknife* estimators (horizontal dashed lines) match the estimates obtained from the model (red crosses).

Figure 11. Plots comparing $\hat{\mu}$ (left y-axis) and $\hat{\tau}$ (right y-axis) with the real values fixed in the simulation (x-axes respectively). Axes in the left plot are log-scaled.

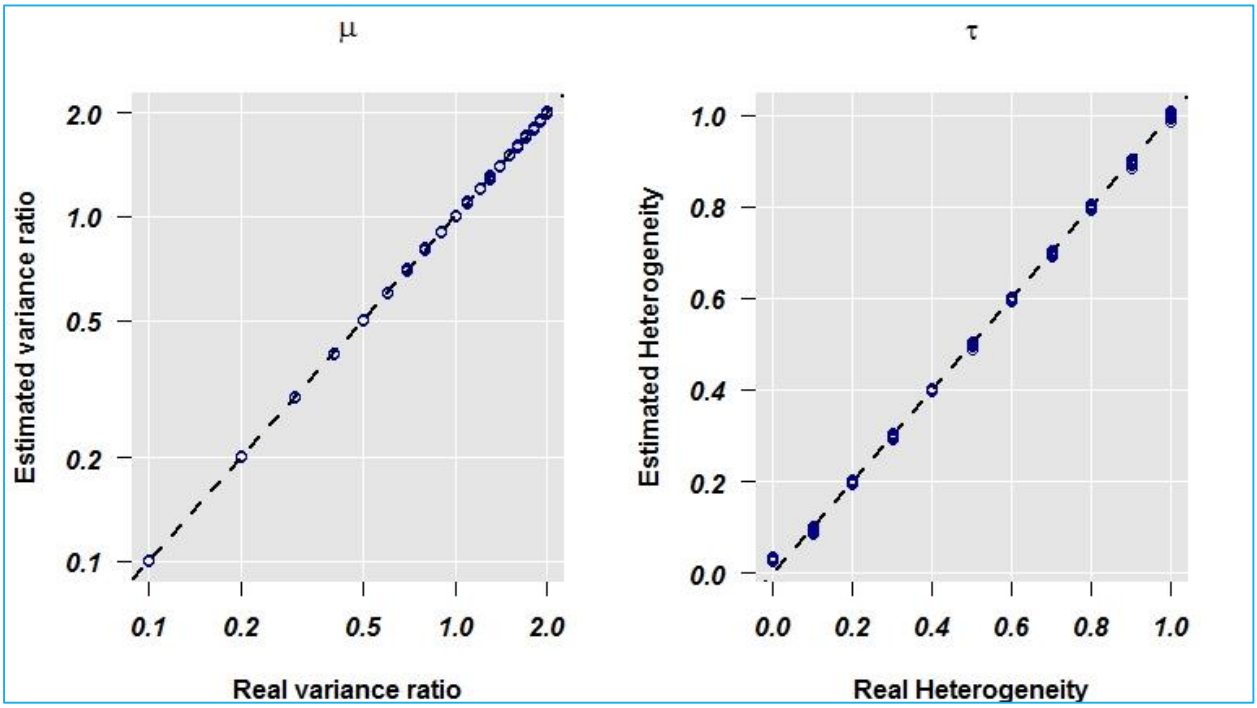
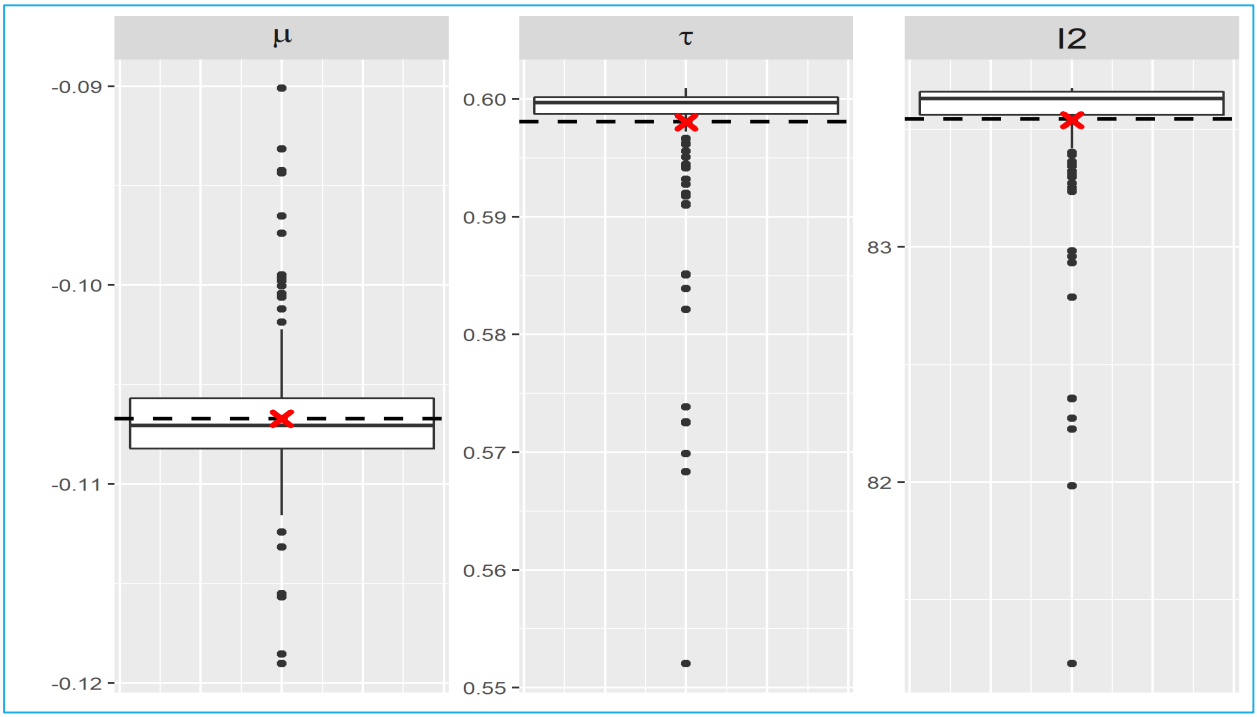


Figure 12. Distribution of the $\hat{\theta}_{(i)}$ *Jackknife* estimators for the main statistics involved in the REM. Dashed lines represent the estimates from the model ($\hat{\theta}$). Red crosses represent the estimates obtained using the *Jackknife* method ($\hat{\theta}_{(.)}$)

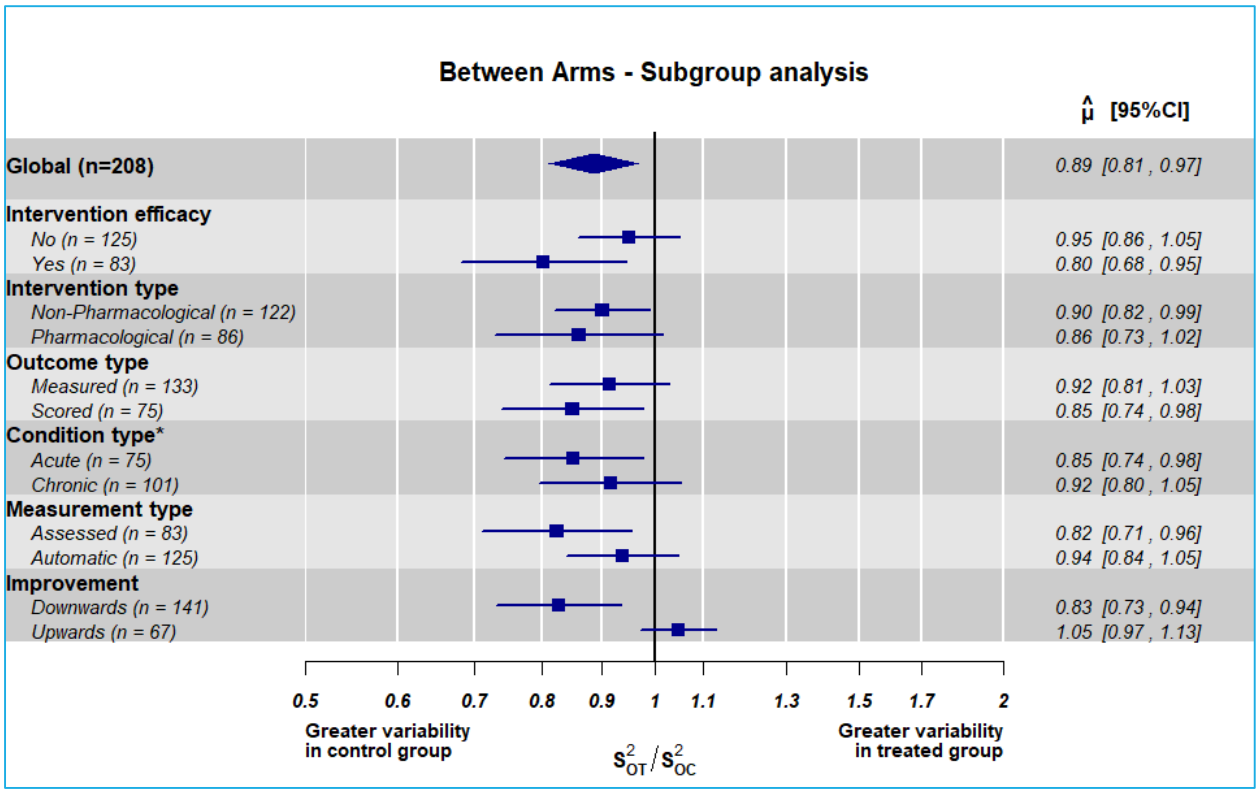


3.2.4 Subgroup analyses

We collected some features of the studies: whether or not the comparison for the primary endpoint yielded "positive" results; the experimental intervention type; the outcome type; the nature of the illness; the measurement type; and the improvement direction.

For each subgroup variable and the overall sample, Figure 13 shows the 95% CI of the outcome variance ratio between arms. Interventions of trials with "positive" results also had an effect on reducing variability, which had already been observed in other studies[54,55]. Not only this, but it has been previously shown that there is a positive correlation between the effect size and its heterogeneity[56,57], that is, if the intervention has an effect on the mean, it also has an effect on diminishing the variance, which is in accordance with some kind of stabilizing effect. For this reason, this factor was hypothesized a priori to be the most relevant. On the other hand, studies that failed to demonstrate the intervention's efficacy were almost centered around the point that represents homoscedasticity: a treatment effect that does not affect the centrality parameter will rarely affect the dispersion.

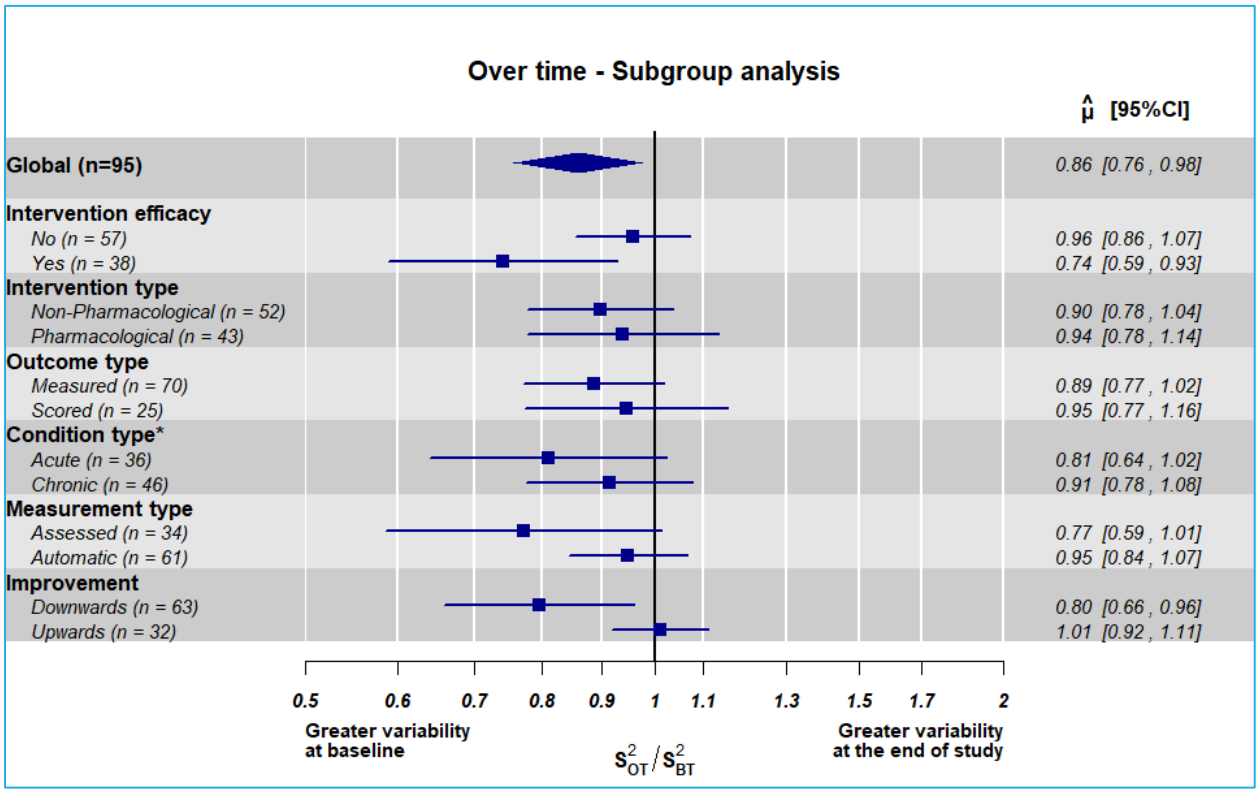
Figure 13. For whole data and each subgroup, the point estimate and the 95% confidence intervals for the outcome variance ratio between Treated (T) and Controls (C), after adjusting by baseline discrepancies through the *rma* function (Model 3 of Table 11). X-axis is log-scaled. *32 studies were performed with healthy participants, i.e., without any particular disease.



Regarding other factors, trials with assessed measurement, showed a trend of having less variability in the experimental group. However, the most marked difference was observed in relation to the direction of improvement. For those trials with an endpoint with lower values associated with a health improvement the variance was clearly lower in the treated arm, while in those studies with outcomes whose values positively correlate with health status, there was a trend in the experimental group of having greater variances. Estimates of differences according to the remaining subgroups did not raise major concerns.

Figure 14 shows the 95% CI for the over-time variance ratio in the experimental arm. No major additional conclusions arise beyond those already mentioned above.

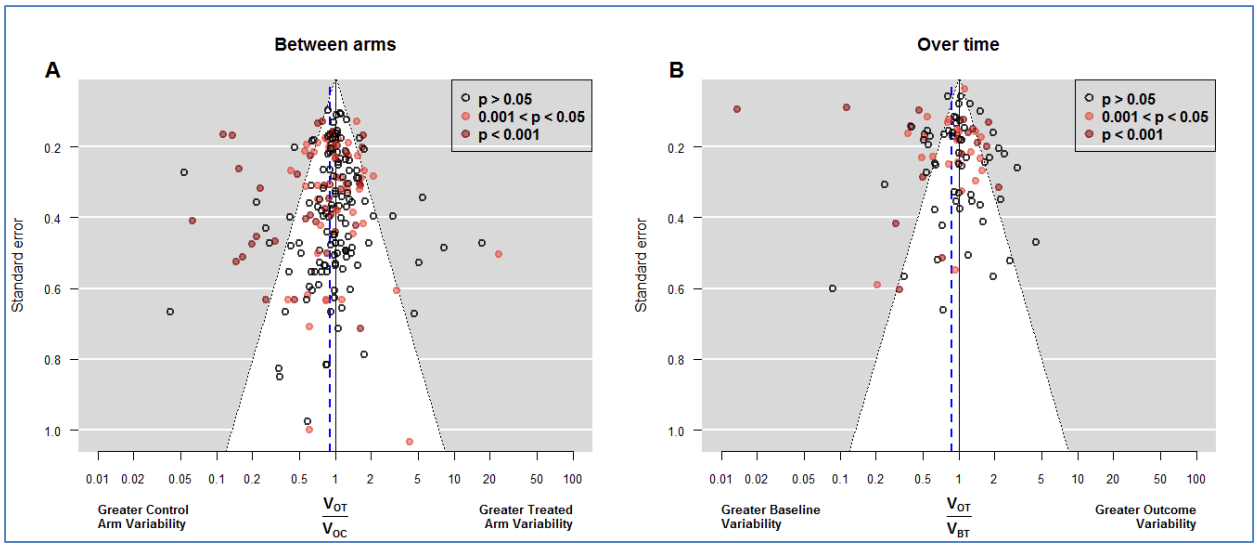
Figure 14. For whole data and each subgroup, the point estimate and the 95% confidence intervals for the estimated variance ratio between Outcome (O) and Baseline (B), after adjusting by the change over time ratio in the control group through the *rma* function (Model 6 of Table 11). X-axis is in log scale.*13 studies were performed with healthy participants.



The most interesting factor involved in the subgroups analyses was whether or not the trials to demonstrated the treatment efficacy. The pseudo funnels plot presented in Section 3.2.2 were replicated by stratifying the statistical significance achieved in the main analysis of the trial measure, using the corresponding p-values. Figure 15 highlights the studies according to the level of such statistical significance. We expected to observe more positive studies (red points) with

markedly different variances (outside the triangles). In Panel A, more red points on the left side indicates that changes in the average, came with reductions in the variance. In Panel B, the largest number of points also on the left side indicates a majority of experimental interventions that reduced variability, of which several of them yielded significant results in the primary endpoint. Nevertheless, the dependence of the statistical significance on the sample size must be taken into account and, consequently, this figure must be interpreted with caution.

Figure 15. Funnel plots of variance ratio between arms (Panel A) and over time (Panel B). Vertical axis indicates precision for the comparison of variances with points outside the triangle being statistically significant. Red points indicate significant differences between means (main trial objective).



The first two sections of appendix E introduce some ancillary analyses regarding the subgroup analyses, which were not included in the main body because they did not provide relevant findings.

4 Sensitivity analyses

This chapter collects a series of sensitivity analyses that were carried out to assess the robustness of the analysis proposed in Chapter 3. Four alternative methodologies were implemented. The first two methods are based on the random effects model. The first (Section 4.1) aims to infer which studies have discrepancies in variability by removing those trials that would contribute more to heterogeneity. The second (Section 4.2) simulates scenarios to inspect which of them would provide better concordance between simulation parameters and the estimates coming from the random effect models analysis. The other two methods are based on the usual tests for variance comparison: either on their direct application (Section 4.3) or by means of a more exhaustive analysis of the p-values resulting from these tests (Section 4.4). Each section referring to sensitivity analyses starts with a summary box that is intended as a guide for further reading.

4.1 Sensitivity Analysis I: Heuristic procedure

Summary key points

- By simulation, we were able to know the expected heterogeneity in a random effects model in a situation of perfect randomness.
- *Baseline* variance discrepancies between treatment arms observed in some studies cannot be derived from a randomization procedure. These initial discrepancies might cause greater outcome variances variability.
- Removing the studies one by one allowed us to identify which trials had *outcome* variability discrepancies.
- Thirty out of 208 and 32 out of 95 studies had to be eliminated in the comparisons between arms and over time, respectively, to achieve a scenario of non-heterogeneity (or constant effect).

This methodology pursues two goals. On the one hand, we wanted to identify those studies whose discrepancy in baseline variability cannot be due solely to chance and, on the other hand, we aimed to determine those trials whose differences in variances at the end of the study are so large that they cannot be attributed to a non-constant treatment effect. This methodology provides a way to determine the number of studies that are promoting the presence of heterogeneity in variance discrepancies and, therefore, we can highlight trials that are suspect of a non-constant effect.

The *heuristic* term is applied to this methodology on the grounds that, the result does not rely on a formal statistical theory but on a logical strategy that nevertheless does not guarantee optimal results.

4.1.1 Methods

4.1.1.1 Explanation

Because of random allocation, baseline sample variances were expected to be identical on average, with some minor variation among studies due to random sampling. In order to obtain a measure of the baseline variability discrepancy between arms, we fitted a *reference* random effects model for the ratio of baseline variances.

Reference model
$$log\left(\frac{S_{BT}^2}{S_{BC}^2}\right)_i = \mu + T_i + e_i \qquad with \quad e_i \sim N(0, v_i^2) \quad and \quad T_i \sim N(0, \tau^2)$$

As mentioned, the expected heterogeneity of this model should be equal to 0, due to randomization. If some heterogeneity is present after randomization, it can be explained by methodological flaws similar to those presented by Carlisle[58], which are from honest errors to data fabrication. For example, the study by Hsieh et al.[59] used an analysis adjusted for baseline values, though these values were obtained one month after the intervention was administered. The reference model allowed knowing the proportion of studies that could have additional heterogeneity due to these methodological impurities and, therefore, the random allocation variability could be separated from the undesired heterogeneity.

At the end of any trial, a possible explanation for discrepancies in the variances could be a direct consequence of the randomization method. Allocation is often performed by permuted blocks or minimization to ensure a balance in prognostic factors between treatment groups, but the block effect or the minimization variables involved in the allocation are rarely fitted in the analysis of clinical trials. This has an undesired consequence: to the extent that the block or covariate is predictive of the outcome, it will increase the variance estimate because the whole purpose of minimizing is to balance for predictive factors and reduce their effect on the true outcome variation; but if the effects are not separately fitted (i.e., adding further predictors in the ANOVA), this additional sum of squares increases the residual term[54].

Simulating studies without heterogeneity will yield a set of heterogeneity measure values under the scenario of a constant effect. Therefore, a criterion is required to decide how much heterogeneity could come from a population of randomized studies in normal circumstances. In other words, what cut-off value for the heterogeneity (τ) determines that the initial discrepancy was too high to occur only by chance? Once again, we rely on a simulation described in Section 4.1.2.3 to verify that such discrepancies were highly unlikely to occur after randomization and, on

the other, determine this boundary. In summary, the application of this methodology was carried out in three steps:

1. **Simulation in order to know the τ_{Target}.** By simulation, we obtained an estimate of the distribution of the τ statistic under a non-heterogeneity scenario. We chose the arbitrary 0.90 quantile of this distribution as a cut-off (τ_{Target}) in order to determine that greater heterogeneity than this value was unlikely in randomized studies.
2. **Baseline variance discrepancies**. Using the reference model, we have removed from the sample all those studies whose variance discrepancy cannot be due solely to chance, up until the heterogeneity achieved the value τ_{Target}. This helped us to detect studies with an excessive and irrational variance discrepancy at baseline.
3. **Outcome variance discrepancies.** Trials with more extreme variance discrepancies between arms or over time have been removed one by one until the random effects model provided heterogeneity as close as possible to the τ_{Target}. These deleted studies were considered to be those that had different variances, because the experimental treatment either increased or decreased the outcome variance.

Two questions arise regarding whether this heuristic methodology is appropriate. First of all, did any studies in our sample have suspicious high variance discrepancies at baseline? Section 4.1.2.1 lists the studies with the most divergent values; at first glance, one can see that the discrepancies of these studies are clearly anomalous when keeping in mind that they come from a randomization process. The second question concerns whether there was a relationship between the baseline and final variance discrepancies; albeit fuzzy, a weak association was present (see Section 4.1.2.2).

4.1.2 Results

4.1.2.1 Studies with unexpected baseline heteroscedasticity

Table 9 lists the studies with more extreme baseline variance ratios. They were suspected of having methodological impurities. All of them except one ($ID = 5$) were studies with relative small sample size.

From Figure 16 to Figure 22, more details about these studies are provided. The outcome variance in the treated arm in the first study was exactly the squared value of the one in the control arm. This fact could lead to think that it was a typo due to confusion between variances and standard deviations, but this explanation was discarded because the information was consistent throughout the whole article. The second study was the only phase II trial in all collected trials and the one with lower sample size. The fifth trial on the list surprisingly showed higher variance discrepancy in those patients with complete follow-up (per protocol analysis) than in the allocated patients (intention to treat analysis).

Table 9. Descriptive statistics of baseline variances in studies with $S_{BT}^2/S_{BC}^2 > 4$ or $S_{BT}^2/S_{BC}^2 < 0.25$.

ID	S_{BT}^2	S_{BC}^2	S_{BT}^2/S_{BC}^2	n_{BT}	n_{BC}
1	256	16	**16**	15	18
2	7,007	675	**10.4**	6	5
3	0.476	0.102	**4.65**	12	10
4	44.6	9.73	**4.58**	31	27
5	5.29	23.1	**0.23**	145	143
6	17.4	99.6	**0.18**	20	20
7	0.012	0.084	**0.14**	20	16

Figure 16. Study 1 (*Passive versus Active Stretching of Hip Flexor Muscles in Subjects with Limited Hip Extension: A Randomized Clinical Trial*.).

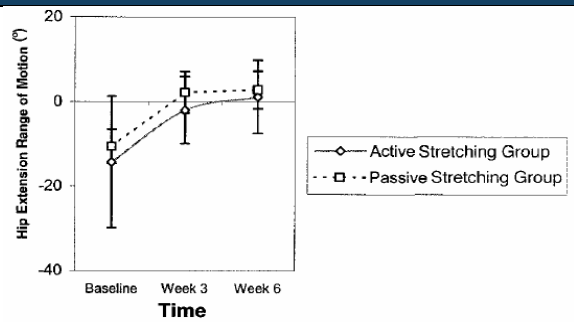


Figure 5 of this paper shows that 95% CIs of the baseline means are quite different. The width of the interval in the active group (rhombus) is four-fold the width of the interval in the passive group (square).

Figure 17. Study 2 (*The COPE Healthy Lifestyles TEEN Program: Feasibility, Preliminary Efficacy, & Lessons Learned from an After School Group Intervention with Overweight Adolescents*).

	Baseline		
	N	*M*	*SD*
Weight			
COPE	7	212.36	83.71
Control	5	189.90	25.99

Table 4 of this paper reports baseline standard deviations with a ratio equal to 3.22, which is equivalent to a variance ratio of 10.4.

Figure 18. Study 3 (*Additive beneficial effects of lactotripeptides intake with regular exercise on endothelium-dependent dilatation in postmenopausal women*).

	Placebo		LTP
	Before	After	Before
BA diameter (mm)	3.8 ± 0.1	3.8 ± 0.1	3.6 ± 0.2

Table 2 of this paper reports the baseline SE of the mean. Variances (0.102 and 0.476) can be deduced from SE and sample sizes, and their ratio is 4.65. The precision (number of significant figures) when reporting these values might have played a relevant role in this discrepancy.

Figure 19. Study 4 (*Consumption of yogurts fortified in vitamin D and calcium reduces serum parathyroid hormone and markers of bone resorption: a double-blind randomized controlled trial in institutionalized elderly women*).

	CY Group	FY Group
	d 0, n = 27	d 0, n = 31
25OHD, nmol/L	16.2 (0.6)	19.2 (1.2)

Table 2 of this paper reports the baseline mean standard errors. Deducted variances are 9.73 and 44.6, leading to a variance ratio equal to 4.58.

Figure 20. Study 5 (*Randomized, double-blind, controlled study of losartan in children with proteinuria*).

Twelve weeks of treatment with losartan significantly reduced proteinuria compared with amlodipine/placebo. There was a 35.8% (95% CI: 27.6% to 43.1%) reduction in mean (SD) UPr/Cr from 1.27 (2.30) to 0.83 (2.48) in losartan-treated patients compared with a 1.4% (−10.3% to 14.5%) increase with amlodipine/placebo from 1.55 (4.81) to 1.60 (6.73), $P \leq 0.001$.

Baseline standard deviations reported in the text are 2.30 and 4.81, with a variance ratio equivalent to 0.23.

Figure 21. Study 6 (*Riluzole as an adjunctive therapy to risperidone for the treatment of irritability in children with autistic disorder: a double-blind, placebo-controlled, randomized trial*).

ABC-C subscale	Week	Patient scores [mean (SD)]	
		Riluzole + risperidone	Placebo + risperidone
Irritability	0	21.40 (4.18)	22.10 (9.98)

Table 2 provides the standard deviations for the primary endpoint (irritability). Their ratio is 0.42 and the variance ratio is 0.18.

Figure 22. Study 7 (*Comparison between sitagliptin and nateglinide on postprandial lipid levels: The STANDARD study*).

Table 1 Baseline patient profile and post-treatment data

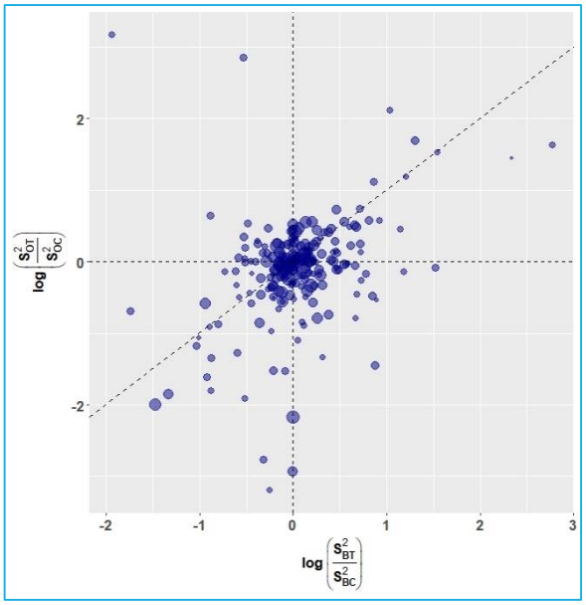
	Sitagliptin group		Nateglinide group	
	Before treatment	After 3 mo of treatment	Before treatment	After 3 mo of treatment
Age (yr)	63.85 ± 12.92	-	66.44 ± 9.02	-
Sex: M/F (*n*)	15/5	15/5	12/4	12/4
Body weight (kg)	68.79 ± 12.89[1]	69.30 ± 13.18	58.36 ± 8.54	58.75 ± 8.52
Systolic blood pressure (mmHg)	134.3 ± 19.3	126.5 ± 11.1	129.1 ± 13.9	122.7 ± 14.8
Diastolic blood pressure (mmHg)	81.5 ± 12.9	73.3 ± 7.4[b]	76.9 ± 8.8	74.3 ± 10.2
HbA_{1c} (%)	7.2 ± 0.7	7.0 ± 0.8[b]	7.2 ± 0.4	6.8 ± 0.5[b]
Glycated albumin (%)	19.3 ± 3.6	17.7 ± 2.6[b]	19.9 ± 3.4	18.1 ± 2.8[b]
1.5-AG (μmol/L)	65.5 ± 45.8	81.6 ± 46.2[b]	51.0 ± 22.7	69.5 ± 29.3[b]
Fasting blood glucose (mmol/L)	7.8 ± 1.7	7.3 ± 1.6	8.1 ± 1.0	7.6 ± 1.5
Blood glucose 1 h postprandially (mmol/L)	12.0 ± 3.0	10.6 ± 3.1[b]	13.6 ± 2.3	11.0 ± 2.8[b]
Blood glucose 3 h postprandially (mmol/L)	7.6 ± 2.5	6.6 ± 2.0[b]	7.6 ± 2.5	6.5 ± 1.9[a]
Apo-B48 AUC (g/L per hour)	2.48 ± 0.11	1.94 ± 0.78[a]	3.14 ± 0.29	2.29 ± 0.16

Table 1 shows the baseline characteristics. The ratio for the primary endpoint Apo-B48 AUC iss equal to 0.14, which is the smallest variance ratio of the whole sample in our study.

4.1.2.2 Relationship between baseline and outcome variance discrepancies.

Figure 23 shows the relationship between outcome and baseline variance discrepancies. The Pearson correlation coefficient between both discrepancies was 0.35 indicating that the differences in the variability between-arms at the end of the study might be slightly influenced by imbalances at baseline, although most of the studies are concentrated in the center of the plot without any apparent relationship. Note the presence of more studies in the first and third quadrant pointing out to a positive correlation between the two represented magnitudes.

Figure 23. Between-arm variance discrepancy at the end of the study as a function of baseline discrepancy. The point size is proportional to the square root of the study sample size.



4.1.2.3 Expected heterogeneity under randomization by simulation. Estimation of τ_{Target}

The feasibility of the observed baseline heterogeneity in our data was assessed by simulating scenarios with identical variances in both arms. The variances were retrieved from the treated arms and sample sizes were also the same that in the 208 collected studies.

Parameters of the simulation	
❖ Total iterations ($nsim$): 10,000 ❖ Variance in treated and control arms (σ_i^2): $S_{T_i}^2$ ❖ Sample size in treated group (n_{i1}): n_{T_i} ❖ Sample size in control group (n_{i2}): n_{C_i}	$S_{T_i}^2$: sample outcome variance in treated arm in the *i*-th study n_{T_i}: sample size in the treated arm in the *i*-th study n_{C_i}: sample size in the control arm in the *i*-th study

Simulation procedure
❖ For each iteration: ○ For each study, generate n_{T_i} and n_C baseline values for treated and control groups, respectively, from a Normal distribution $N(\mu = 0, \sigma = S_{T_i})$ ○ A random effect models is fitted to the generated data and the heterogeneity parameter (τ) , and the random sampling variability is stored The quantile 0.90 of all the simulated τ's will be the τ_{Target}

As an illustration, the forest-plot in Figure 24 is taken to compare the 95% CI of the baseline variance ratio from a randomly chosen instance of simulated data with the 95% CI obtained for the real data at baseline and at the end of the study. In the simulated data, the point estimates vary from one iteration to another, but the interval width for each study remains constant because it depends only on the sample size in each group. The number of studies with confidence intervals for real baseline variance discrepancy that do not include the value 1 (n=26, 12.5%), which represents perfect homoscedasticity, is higher that the number of intervals not including the value 1 provided by simulated data (n=14, 6.7% in the represented plot; and 10.5 studies, 5.1% on average throughout the complete simulation)

Figure 24. Forest plot of 95% CI for the estimated variance ratio for each study: single instance of simulated data with no effect and no heterogeneity (left); the reference model (middle); and outcome model (right). Red intervals do not contain the value 1. Studies are sorted according their point estimate and, thus, the order of the studies in the two last plots can vary.

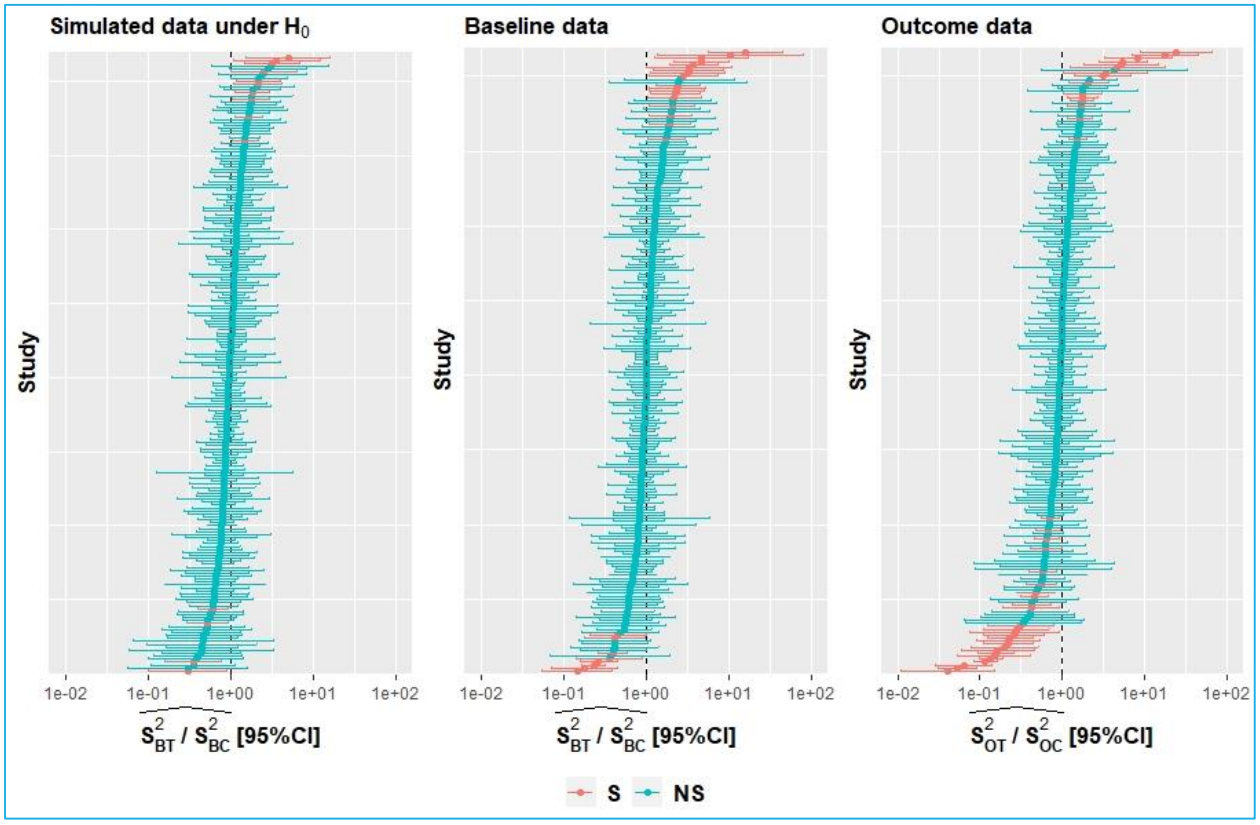
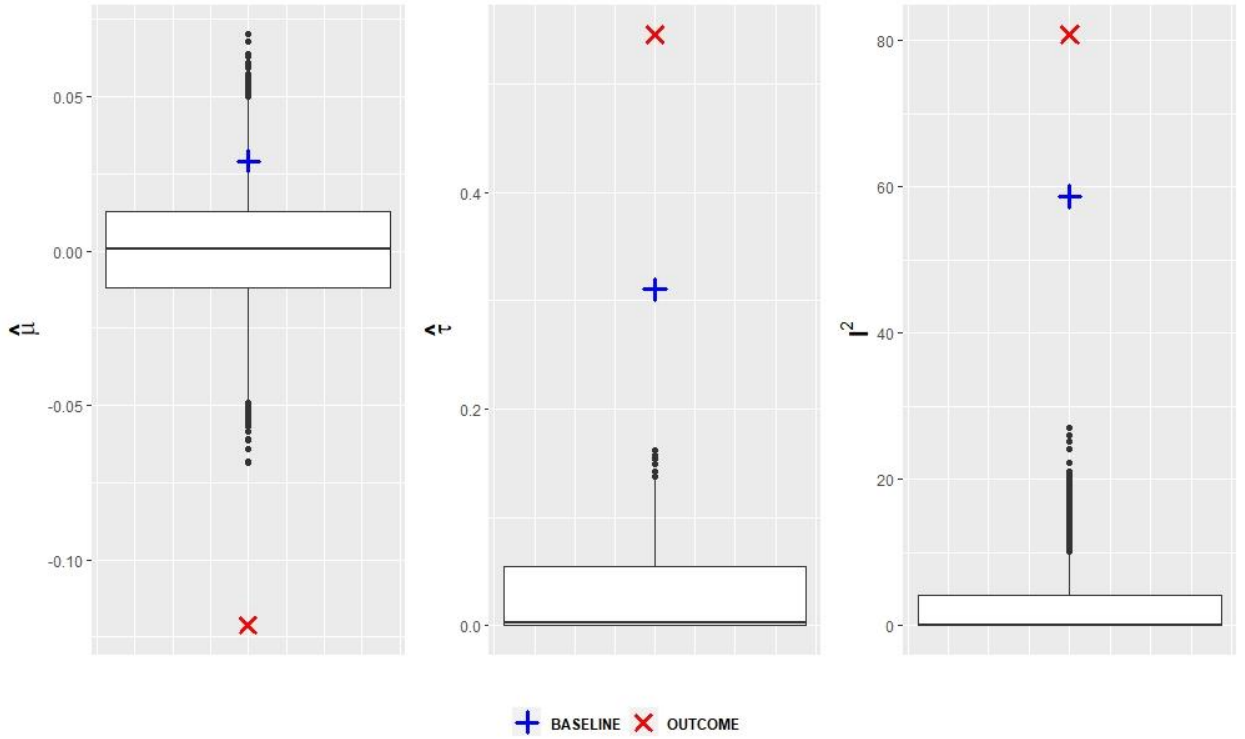


Figure 25 shows the distribution of the three estimators for the main model parameters (μ, τ *and* I^2) provided by the complete simulation procedure. The 2nd and 3rd plots of this figure show that indicators of heterogeneity in the *baseline* data (blue crosses) were clearly higher than expected compared with the simulated distributions when no heterogeneity was present. This causes suspicion regarding whether the random allocation of some trials was correct, and this fact could imply more heterogeneity at the end of the study (red crosses).

Figure 25. Boxplots of the 3 parameters of interest along 10,000 simulations based on a random effects model applied to data with no heterogeneity and equal variances in both arms. Blue and red crosses represent the estimated parameters using the random effects model with baseline variances (reference model) and outcomes variances (outcome model) as response, respectively.



An average $\tau = 0.03$ resulted from the simulation, which is a plausible value under a non-heterogeneity scenario. Table 10 shows the quantiles for the three parameters under study throughout the 10,000 iterations. The quantile 0.90 of the τ values was defined as the τ_{Target}. Studies were removed from the three random effects models (reference, between-arms and over-time) until the estimated heterogeneity was below (or very close to) this τ_{Target}, which was equal to 0.08 in the simulated data.

Table 10. Estimated values for the three main parameters in the *reference* model and the adjusted between arms model (model 2 of Table 5), and the quantiles of the three parameters in the simulated data.

Parameter	Models		Quantiles in simulated data										
	Reference	Between arms	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
μ	0.03	-0.11	-0.07	-0.02	-0.01	-0.01	0.00	0.00	0.01	0.01	0.02	0.02	0.07
τ	0.31	0.60			0.00			0.00	0.03	0.05	0.06	**0.08**	0.16
I^2	58.73	83.55			0.00			0.01	1.27	3.07	5.21	8.18	27.08

4.1.2.4 Proportion of studies causing heterogeneity

We first modeled the baseline variance ratio as the response in the complete database (CDB, $n = 208$), which was expected to have a null heterogeneity. The reference model provided an estimate of the logarithm of the variance ratio equal to 0.03, 95% CI from -0.03 to 0.09 and, surprisingly, the estimated baseline heterogeneity was $\hat{\tau} = 0.31$ with an $I^2 = 58.7$ (Table 10), a very high value if we bear in mind the randomization procedure. However, the exclusion of the four most extreme studies lessened the value of $\hat{\tau}$ to 0.066 at baseline (RDBB, $n = 204$), slightly lower than the $\tau_{\text{Target}} = 0.08$. When the outcome discrepancy rather than the baseline variances was modeled as the response in the CDB, heterogeneity was almost doubled (0.55), and we needed to exclude up to 29 studies so that we could lower this heterogeneity to 0.08 in the reduced database for the between-arm comparison (RDB, $n = 179$), with 10 out of 208 (4.8%) studies increasing their variance in the treated group and 19 (9.1%) doing just the opposite. Finally, in the comparison over time, 28 studies had to be eliminated to achieve a $\tau = 0.079$ in the reduced database for the comparison over time (RDO, $n = 67$). Specifically, we had to remove 10 (10.5%) and 18 (18.9%) studies with, respectively, greater and lower variance at the end of the study. The comparisons among all the model heterogeneities (τ estimates) are presented in Table 11. Unadjusted and offset models showed greater heterogeneity in all situations since these models have less flexibility due to the forced value of β.

Table 11. Estimated heterogeneities ($\hat{\tau}$) from the random effects models. Superscripts represent the fitted model identifier (see Table 5 & Table 7).

Response in the model	**Type of model**	$\hat{\tau}$	
		Full Data (main analysis)	Reduced data (heuristic analysis)
		CDB (n=208)	**RDBB (n=204)**
Baseline variance ratio	Reference model	0.31	0.066
		CDB (n=208)	**RDB (n=179)**
Outcome variance ratio	Unadjusted[1]	0.60	0.085
	Adjusted[2]	0.55	0.080
	Adjusted (offset)[3]	0.60	0.188
		CDO (n=95)	**RDO (n=67)**
Outcome vs. baseline ratio in treated group	Unadjusted[4]	0.71	0.137
	Adjusted[5]	0.59	0.079
	Adjusted (offset)[6]	0.62	0.277

4.2 Sensitivity analysis II: Simulation study

Summary key points

- Misclassification errors were not controlled in the analysis based on random effects models.
- A simulation study was performed to explore under which simulated scenarios we would attain the parameters we have estimated in the principal analysis.
- A variable proportion of studies with random effect were generated under several settings.
- The models in a scenario where 10% of trials have random treatment effects produced coefficients that were similar to those from the random effects model analysis.

4.2.1 Methods

4.2.1.1 Explanation

Several assumptions were made about the conclusions derived from the random effects model in the Chapter 3. The most important was that studies that fell outside the region of natural variability in the funnel plot came from studies with heteroscedasticity, and vice versa. In the absence of a design for controlling probabilities of the type I and type II errors, an unknown proportion of these studies could be misclassified. The purpose of this sensitivity analysis was, on the one hand, to simulate realistic situations with specified proportions of studies having a random treatment effect (and, therefore, heteroscedasticity) and, on the other, to explore under which of these scenarios the model parameter estimates for between-arm comparison match those previously obtained. Hence, these scenarios provide us the proportion of trials with random effects. The simulation may be conducted in several ways, and it also requires some assumptions:

- **All the treatment effects are additive**. Usually, the main analysis of the collected trials rests on a comparison of the outcome mean, which assumes an additive treatment effect. If trialists were to have expected a multiplicative treatment effect, then they would have log-transformed the main outcome to achieve an additive effect.
- **There is a proportion π_R of studies with a random treatment effect**. It is sensible to assume that some interventions produce a different treatment effect, depending on the patient. We simulated a proportion π_R of interventions with a variable (random) treatment effect and another proportion (1-π_R) with a constant effect.
- **The random treatment effects have different variability among studies**. Some interventions may produce more heterogeneous effects than others. The variability of the random treatment effect was generated using a uniform distribution between 0 and θ_M, which was one of the simulation parameters.
- **On average, the variability in the control arm is greater than in the treated arm**. Despite the fact that we also simulated scenarios with higher variability in the treated arm,

the results obtained under the settings with greater variability in the control arm fitted better to the main results shown in Chapter 3.

The simulation was performed only for the between-arm comparison because of two main reasons. First, the sample size was quite small for the comparison over time: only 95 studies do not provide enough precision to this specific goal. Second, and most importantly, we had the constraint of a fixed correlation between outcome and baseline values for each collected study. This restriction made the simulation procedure much more complex and, in turn, would result in obtaining estimates that lack the necessary reliability.

4.2.1.2 Simulation models

The next box describes the models for simulating the data. Two separate models for the interventions with constant and random treatment effects were set.

Simulation model

Constant effect

$$Y_B \sim N(0,1)$$
$$Y_{OT} = Y_B + \gamma_T \qquad \gamma_T = k$$
$$Y_{OC} = Y_B + \gamma_C \qquad \gamma_C = k'$$

Random effect

$$Y_B \sim N(0,1)$$
$$Y_{OT} = Y_B + \gamma_T \qquad \gamma_T \sim N(0,1)$$
$$Y_{OC} = Y_B + \gamma_C \qquad \gamma_C \sim N(0,\sigma_C) \quad \sigma_C \sim U(0,\theta_M)$$

Y_B: Baseline values; Y_{OT}: Potential outcome in treated arm; Y_{OC}: Potential outcome in control arm; γ_T: Change from baseline to the end of the study in treated arm; γ_C: Change from baseline to the end of the study in control arm; k, k': constant values; σ_C: standard deviation of γ_C; θ_M: maximum value of σ_C.

The box below provides the selected simulation parameters.

Parameters of the simulation

- 100 replications (*nsim*)
- Maximum standard deviations for the treatment effect on the control arm (θ_M): 1, 3, 5, 7
- Expected treatment effect $(E[\gamma_T - \gamma_C])$: 0
- Proportion of studies with random effect (π_R): 0, 0.05, 0.10, 0.15, … , 0.5
- Sample size in treatment and control group: n_{T_i}, n_{C_i}

Some remarks about the choice of the simulation parameters should be highlighted. The number of replications (100) was enough to draw conclusions because, as the number of iterations increased, the differences in the simulated model parameters showed that the results remained quite stable above 50 replications. In order to obtain a more accurate estimation of π_R, once the most plausible value for θ_M among the predefined subset of values (1, 3, 5 and 7) was obtained, other values for θ_M were tested around the value that provided a closest estimates to the target coefficients. In the same way as with θ_M, we tested values in the full range from 0 to 1 for π_R, but only values up to 0.5 are shown. The expected treatment effect had no impact at all on the simulation results; hence, for simplicity, we assigned it a 0 value, which is equivalent to an absence of treatment effect. Note that the model is flexible enough to allow studies with random treatment effects that provide either lower or greater variation in the experimental arm, since variance for the change over time in the reference group is bounded within a range (0, θ_M) that covers values above and below 1, which is the standard deviation in the treated group.

4.2.2 Results

Figure 26 shows that the simulated parameters closest to the target estimates were reached in a setup that combines a maximum random treatment effect value of 7 ($\theta_M = 7$, last panel) with a proportion of studies whose random treatment effect equals 0.10 ($\pi_R = 0.10$). Under this scenario, the solid lines (simulation estimates) cross the dashed lines (random effects model estimates) at approximately the same x-value. Additional simulation scenarios around the optimal setting of $\theta_M = 7$ and $\pi_R = 0.10$ were tested: we explored the grid from $\theta_M = 6.5$ to $\theta_M = 7.5$ and from $\pi_R = 0.05$ to $\pi_R = 0.15$, in both cases, with evenly spaced intervals of 0.01. For each combination, we calculated the mean square error (MSE) between the simulated data's mean estimates and the estimates from the random effects model analysis. The parameters for which a lower MSE was achieved were $\theta_M = 7.2$ and $\pi_R = 0.1$. Table 12 shows the values of the estimates coming from both sources: real and simulated data.

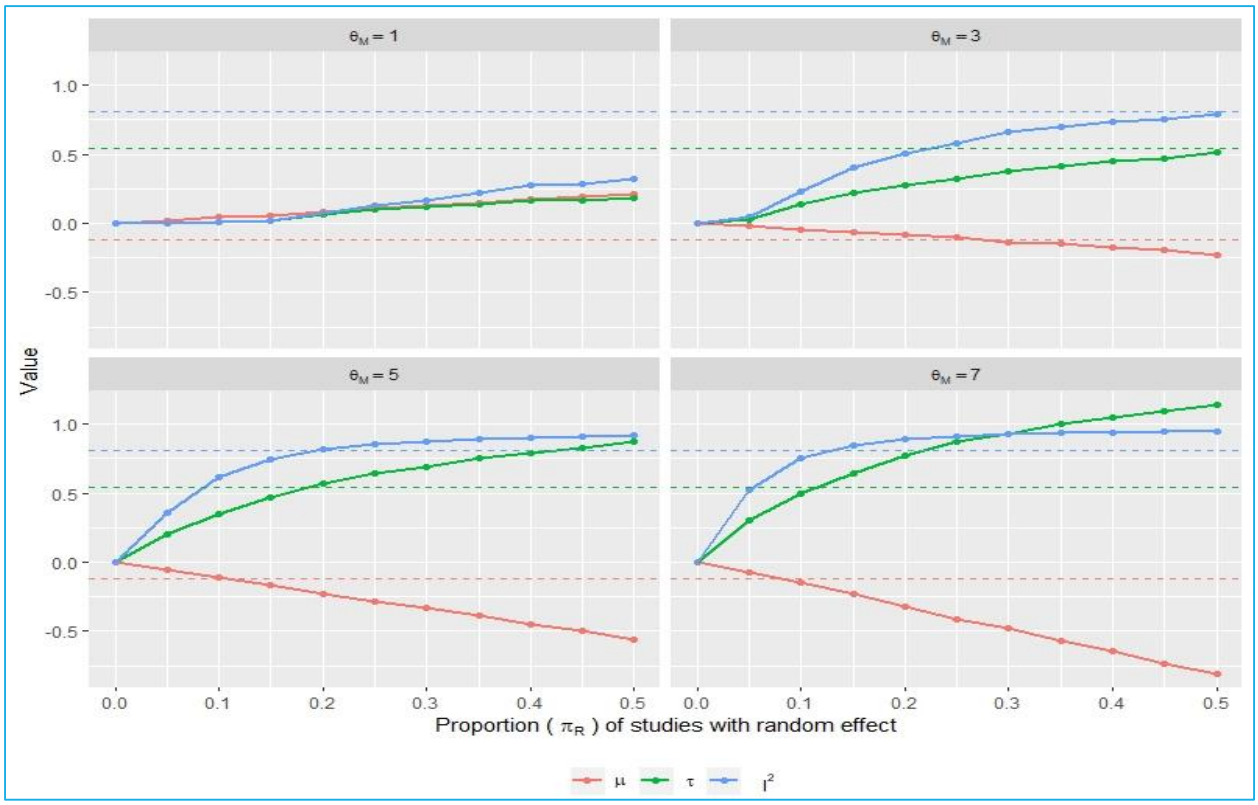
Table 12. Parameter estimates from the real and data from the closest simulated scenario.

Data source	Estimated parameter			MSE
	μ	τ	I^2	
Collected data	-0.121	0.546	0.809	0.00080
Simulated data in the optimal setting ($\theta_M = 7.2$; π_R =0.1)	-0.162	0.528	0.788	

The value $\theta_M = 7.2$ represents the maximum standard deviation of the control group in the studies with random effects. It has been interpreted as if the standard deviation in the reference group is,

on average, approximately three and a half times greater than in the experimental group. This value may seem very high, but it may be due to studies where the patients in the experimental group are either subjected to severe monitoring or to really efficient interventions that place outcomes of treated patients in a narrower range of normality.

Figure 26. Solid lines represent the average estimates of the parameters μ, τ, I^2 in an additive treatment effect setup, depending on the maximum variability of the random effect (θ_M) in each panel and on the proportion of studies with random effects (π_R) on the x-axis. Dashed lines are the estimated values from the analysis using the real data for the between-arm comparison.



In the fitted models with simulated data under this scenario ($\theta_M = 7.2$), 1.4% and 8.6% of studies were expected to show greater variability in, respectively, the control and treated arms:

$$P(\sigma_C < \sigma_T) = P(\sigma_C < \sigma_T|\gamma_C = k') \cdot P(\gamma_C = k') + P(\sigma_C < \sigma_T|\gamma_C \neq k') \cdot P(\gamma_C \neq k')$$

$$= 0 \cdot (1 - \pi_R) + P[U(0, \theta_M) < 1] \cdot \pi_R = \frac{1}{7.2} \cdot 0.1 = 0.014$$

$$P(\sigma_C > \sigma_T) = P(\sigma_C > \sigma_T|\gamma_C = k') \cdot P(\gamma_C = k') + P(\sigma_C > \sigma_T|\gamma_C \neq k') \cdot P(\gamma_C \neq k')$$

$$= 0 \cdot (1 - \pi_R) + P[U(0, \theta_M) > 1] \cdot \pi_R = \frac{6}{7.2} \cdot 0.1 = 0.086$$

Note that in the above expression ($\gamma_C = k'$) and ($\gamma_C \neq k'$) equates to the presence of a constant and random effect, respectively.

4.3 Sensitivity Analysis III: Comparison using common tests for comparing variances

Summary key points
❖ Formal tests for variance comparison can be applied to each single collected study. ❖ An independent sample test was performed for between-arm comparison, and a paired test was implemented for comparison over time. ❖ Almost 20% and 40% of the studies showed statistically significant differences in the between-arm and over-time comparisons, respectively.

4.3.1 Methods

4.3.1.1 Explanation

This method aimed to assess the homoscedasticity in each single study through the usual tests of variance comparisons:

- ❖ Between outcomes in both arms with the usual F-test for independent samples
- ❖ Between baseline and outcome in the treated arm with a specific test for paired samples[60].

Both tests were two-sided with a significance level of $\alpha = 5\%$. The formal hypotheses for both comparisons are expressed below.

Independent samples	Paired samples
$\begin{cases} H_0: \sigma_{OT}^2 = \sigma_{OC}^2 \\ H_1: \sigma_{OT}^2 \neq \sigma_{OC}^2 \end{cases}$	$\begin{cases} H_0: \sigma_{OT}^2 = \sigma_{BT}^2 \\ H_1: \sigma_{OT}^2 \neq \sigma_{BT}^2 \end{cases}$

4.3.1.2 Independent samples

Although there are other alternatives, for the independent samples situation, an F-test was implemented for comparing variances. The statistic for this test is:

$$F = \frac{S_{OT}^2}{S_{OC}^2} \sim F_{n_{OT}-1, n_{OC}-1}$$

This statistic follows an F distribution with $(n_T - 1)$ and $(n_C - 1)$ degrees of freedom under the null hypothesis of homoscedasticity given the next premises: 1) the population outcome is normally distributed (this is reasonable because collected trials were designed to compare means); and 2) the samples are independent, which is guaranteed by design.

4.3.1.3 Paired samples

For the paired comparison, the *Lothar Sachs test*[60] was implemented for comparing variances. The statistic for this test is:

$$t = \frac{(Q_{OT} - Q_{BT}) \cdot \sqrt{n-2}}{\sqrt{Q_{OT}Q_{BT} - Q_{OT,BT}^2}} \sim t_{n-2}$$

where: $Q_{OT} = (n-1) \cdot S_{OT}^2$; $Q_{BT} = (n-1) \cdot S_{BT}^2$; $Q_{OT,BT} = (n-1) \cdot Cov(Y_{OT}, Y_{BT})$

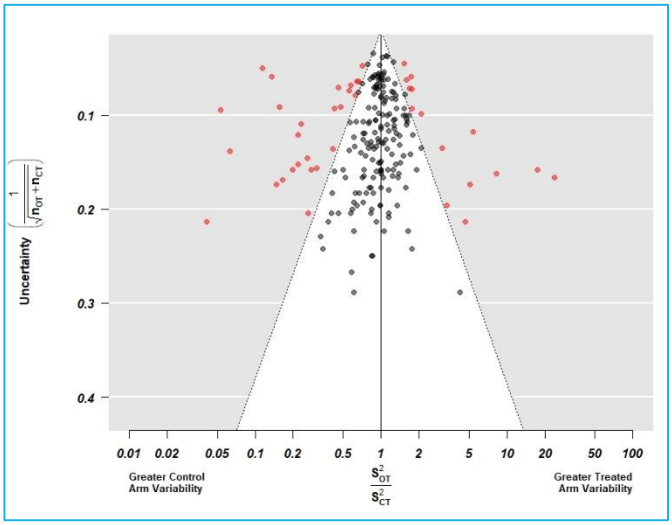
The statistic follows a t-distribution with $(n-2)$ degrees of freedom under the null hypothesis of equal variances, keeping in mind two assumptions: 1) the population outcome is normally distributed; and 2) the samples are paired. As this test was applied to a before-after comparison, this last premise is met in any case.

4.3.2 Results

4.3.2.1 Independent samples: Between arms

Using the F-test, we found that 41 trials (19.7%) showed heteroscedasticity in the outcome over the 208 collected studies. Among them, 26 (63.4%) presented lower variability in the treated arm, and 15 (36.6%) had significantly greater variability in the experimental group. We set the inverse of the square root of the total sample size at the end of the study as a measure of uncertainty in Figure 27, which tries to emulate the funnel plot obtained from the random effects model. This figure is a little bit tricky in the sense that the limits of the non-rejection region are not a linear function of this uncertainty measure, especially in small trials. For instance, both the studies of Moe et al.[61] and Wöhrle et al.[62] had 40 patients (uncertainty= $1/\sqrt{40} = 0.16$) with comple-te follow-up at the end of the study. However, in the former, the groups were balanced and the upper limit of the non-rejection region is 2.53 ($F_{0.975,19,19}$); while in the latter, with unbalanced arms (12 vs. 28), the upper limit is 3.14 ($F_{0.975,27,11}$). Consequently, it would have been possible for a study with different variances (red point) to have fallen inside the triangle, although did not occur in any case. The triangle limits, as a function of the uncertainty, are the result of fitting a linear regression to the boundaries of the non-rejection region of the test.

Figure 27. Funnel plot for the variance discrepancy between arms. Red points mean heteroscedastic trials according to the F-test. The white triangle represents the non-rejection region. Uncertainty was measured as the inverse of the square root of the total sample size.



4.3.2.2 Paired samples: Over time

Using the variance test for paired samples[60], 38 (40%) trials presented heteroscedasticity over time among the 95 studies with available covariance between pre- and post-outcome. Among them, 22 (57.9%) show lower variability at the end of the study, and 16 (42.1%) studies had significant greater variability at baseline.

Figure 28. Funnel plot for the variance discrepancy over time. Red points represent studies with heteroscedasticity according to the paired test based on the Q-statistic. The white triangle is the non-rejection region. Uncertainty was measured as the inverse of the square root of the baseline sample size.

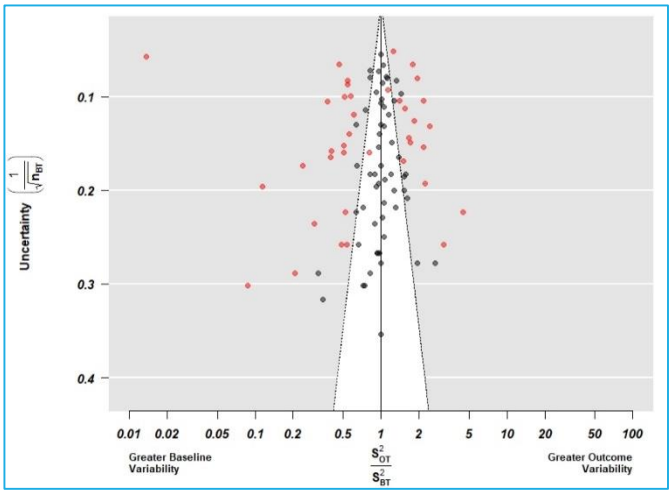


Figure 29. *Q* statistic versus a measure of uncertainty. The dotted lines that limit the white shaded region represent the upper and lower limits according to a *t*-distribution with *n-2* degrees of freedom for each study.

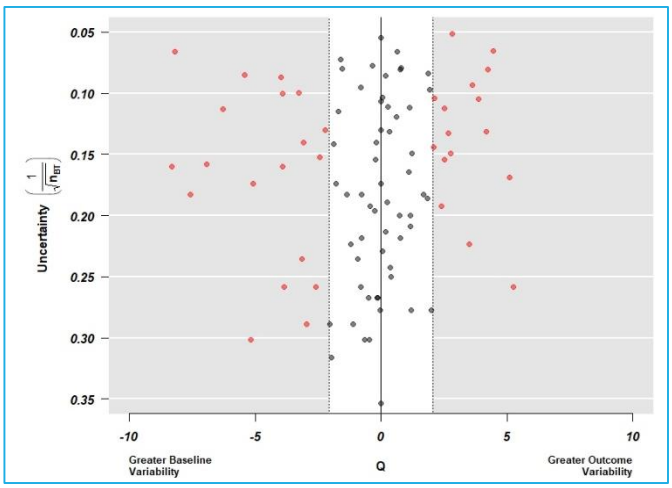


Figure 28 shows the funnel plot for the over-time variance ratio. The horizontal axis represents the ratio for the final (O) and baseline (B) variances in the treated arm. The vertical axis is the inverse of the square root of the sample size at baseline, which represents a possible measure of uncertainty. The limits of the triangle are not directly related to this measure, because the covariance also plays an important role in their configuration. Given S_{BT}^2 and covariance of each single study, we calculated the limit value of S_{OT}^2, from which the null hypothesis of equal variances would be rejected according to the test, either because S_{OT}^2 is too large (red points on the right) or because it is too small (red points on the left). Once these limit values were calculated for each study, the boundaries of the non-rejection region were the result of applying a linear regression to these limits for individual studies as a function of the uncertainty ($R^2 = 0.985$ and $R^2 = 0.991$ for the lower and upper limits, respectively).

Since the limits of the triangle depend not only on *n* but also on the correlation, there can be studies with significant differences within the triangle and vice versa. For example, the leftmost black

point is a study[63] located outside the triangle with a moderate correlation (r = 0.30), while the significant (red point) study[64] with a ratio of variances closest to 1 is the one with a higher correlation (r = 0.99). To avoid this problem, another plot including the *Q* statistic on one of the axes has been represented in Figure 29. As it follows a Student's t-distribution under the null hypothesis, the boundaries of the non-rejection region are around two, regardless of the studies' sample sizes.

4.4 Sensitivity Analysis IV: Comparison through distribution mixture

Summary key points
❖ The p-values obtained in the tests for variance comparison come from two distributions: a uniform distribution under equal population variances and an unknown distribution, otherwise. ❖ This distribution mixture can be modeled to elucidate the proportion of studies under each hypothesis. ❖ The estimated proportion of studies with no evidence of heteroscedasticity at the end of the study was 0.80 (95% CI from 0.72 to 0.88) in the between-arm comparison and 0.57 (95% CI from 0.44 to 0.71) in the comparison over time

The American Statistical Association (ASA) published a statement in 2016 with a set of recommendations regarding the use and interpretation of the p-values[65]. Among the various suggestions, they advised about that scientific conclusions and policy decisions should not be based only on whether a p-value passes a specific threshold. We fully agree with this point and believe that the results of the previous sensitivity analysis (which are based on hypothesis tests that lack power) should be complemented by a more sophisticated analysis, which is presented in this Section 4.4.

4.4.1 Methods

4.4.1.1 Explanation

Since the collected studies were not designed to find differences in variances, but in means, the previous tests in Section 4.3 could lack statistical power. For this reason, we devised our new and final method for estimating the proportion of trials that have a variance discrepancy. The method is based on fitting a mixture of distributions to the p-values that resulted from the previous variance comparison tests, and it is similar to the methodology applied by Pounds et al.[66]

Under the null hypothesis of equal variances, it is well known that the p-values follow a uniform distribution (U) in the interval [0,1]. In contrast, under the alternative hypothesis, studies with different population outcome variances provide p-values coming from an unknown distribution, which might depend on the magnitude of the variance difference. Therefore, these p-values come from two different populations and they lead to a distribution mixture. One of the distributions that compound the mixture should be uniform. For the other distribution, we first explored the empirical distribution of the p-values (Figure 30 and Figure 31 for both comparisons). We used the one-sided p-values (under the alternative hypothesis that the treatment arm outcome has greater variance) in order to distinguish those studies with lower or greater variance in the treated group's outcome. That is, small p-values will correspond to trials with less variability in the outcome of the treated group, while p-values very close to 1 will come from studies with lower variability in the outcome of this arm. Both empirical distributions show two peaks (close to 0 and 1, respectively) caused by those trials with real variance discrepancies between groups and over time. Q–Q plots show how closely the empirical distributions form a theoretical uniform distribution. Both comparisons (especially the comparison over time) indicate the presence of trials with different variability and, therefore, with a non-constant effect. The aim of this methodology is to quantify the percentage of these studies.

Figure 30. Distribution of one-sided p-values for between-arm comparison: histogram (left) and Q-Q plot for a theoretical uniform distribution (right).

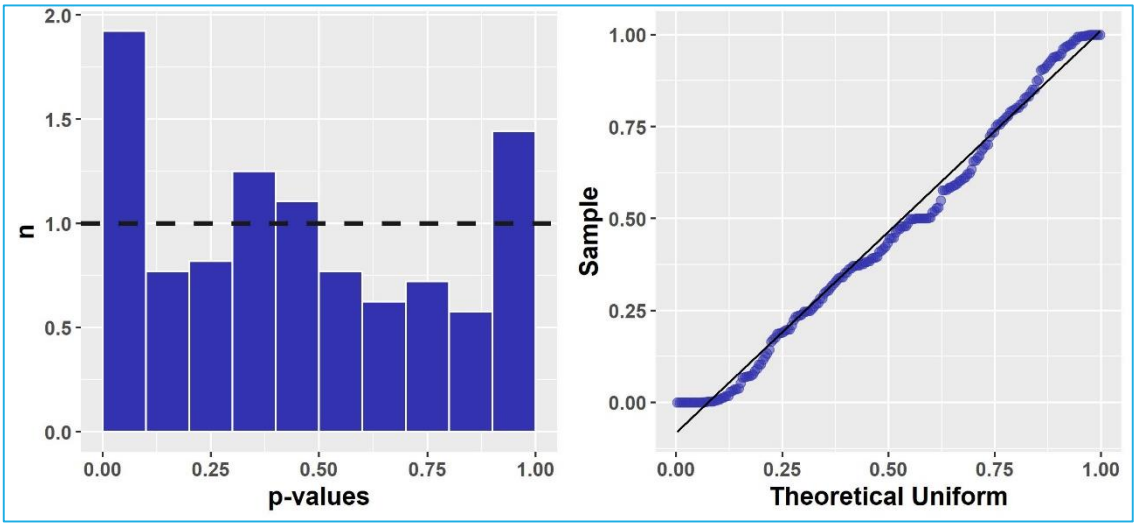
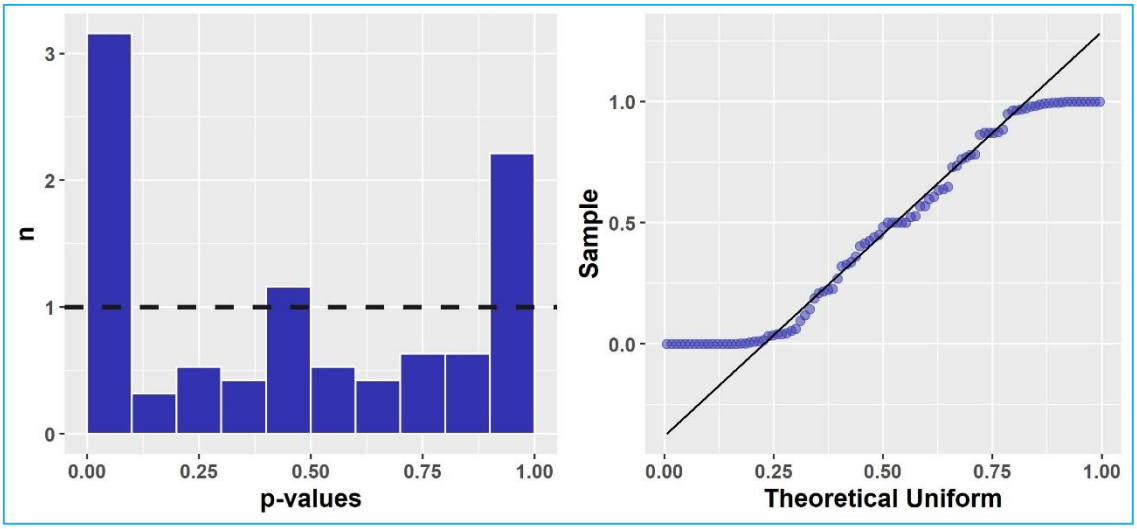


Figure 31. Distribution of one-sided p-values for comparison over time: histogram (left) and Q-Q plot for a theoretical uniform distribution (right).



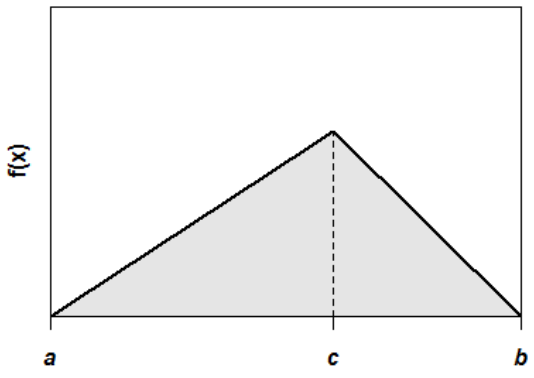
To fit the distribution mixture, we should choose one or more distributions to model the p-values coming from the studies with variance discrepancies while taking into account the observed peaks in both histograms. We fitted four distribution mixtures to these data with the goal of estimating the proportion of studies derived from trials with different final outcome variability or variance discrepancy over time.

4.4.1.2 Distributions

This section provides an overview of the distributions that were used to model the p-values coming from heteroscedastic trials. Four different distributions are presented.

- ***Triangular (T)***. This distribution (Figure 32) has three parameters (a, b and c) and a piecewise density probability function (f_T).

Figure 32. Triangular distribution.



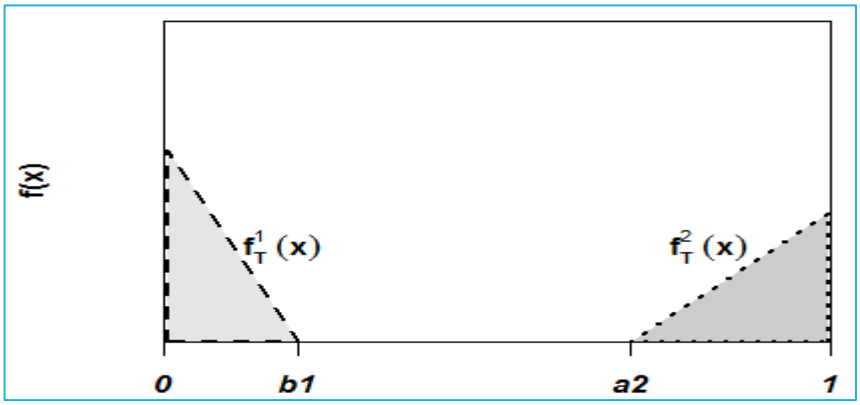
$$f_T(x;a,b,c) = \begin{cases} 0 & if \quad x < a \\ \dfrac{2(x-a)}{(b-a)(c-a)} & if \quad a \le x \le c \\ \dfrac{2(b-x)}{(b-a)(b-c)} & if \quad c < x \le b \\ 0 & if \quad b < x \end{cases}$$

In our specific case, we need two triangular distributions to cover the two empirical peaks, in which case, only one parameter is enough to determine each triangle's distribution on the condition of two additional restrictions:

- They should be right triangles with $a = c$ (or $c = b$) in order to accommodate the peaks in the lower (or higher) p-values.
- The vertical side of the triangles should be located at $x = 0$ (or $x = 1$). This constraint implies that $a = c = 0$ for the triangle at the left (or $c = b = 1$ for the triangle at the right).

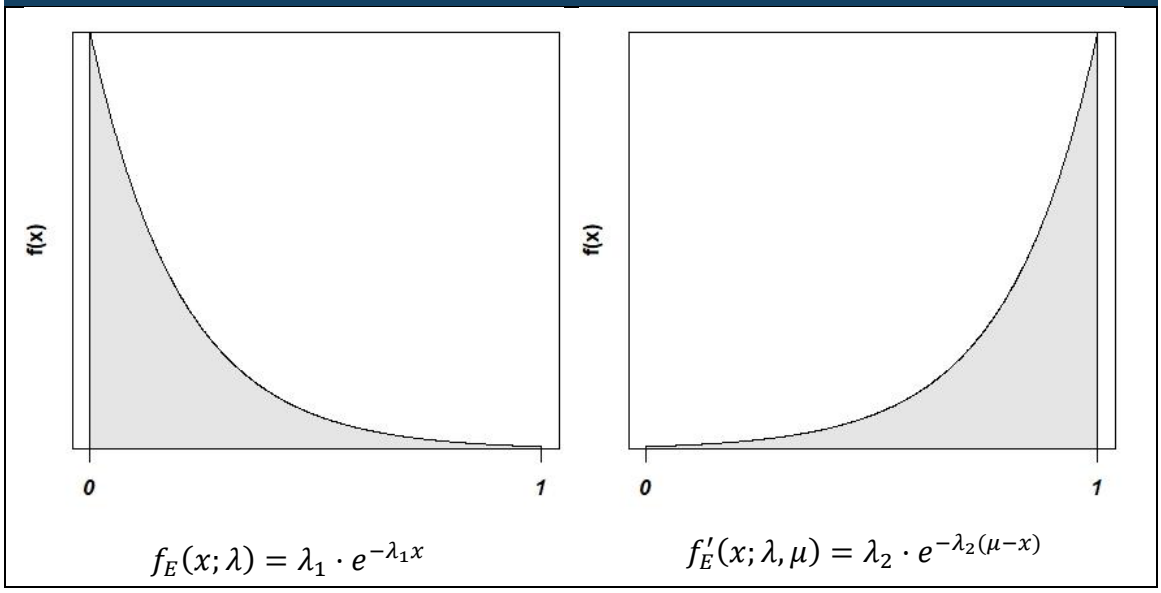
Consequently, the only parameters that should be estimated are parameter b (called b_1) for the left-hand triangular distribution (f_T^1) and parameter a (called a_2) for the right-hand triangular distribution (f_T^2) (Figure 33).

Figure 33. Triangular distributions constrained to additional restrictions: right triangles located at the extremes.



❖ ***Exponential (E)***. We used a usual exponential with event rate λ_1 to fit the left peak and a translated and inverted exponential distribution with event rate λ_2 and fixed location parameter $\mu = 1$ to fit the right peak. Their densities (f_E, f_E') are below Figure 34.

Figure 34. Exponential distribution at left and translated exponential distribution at right.

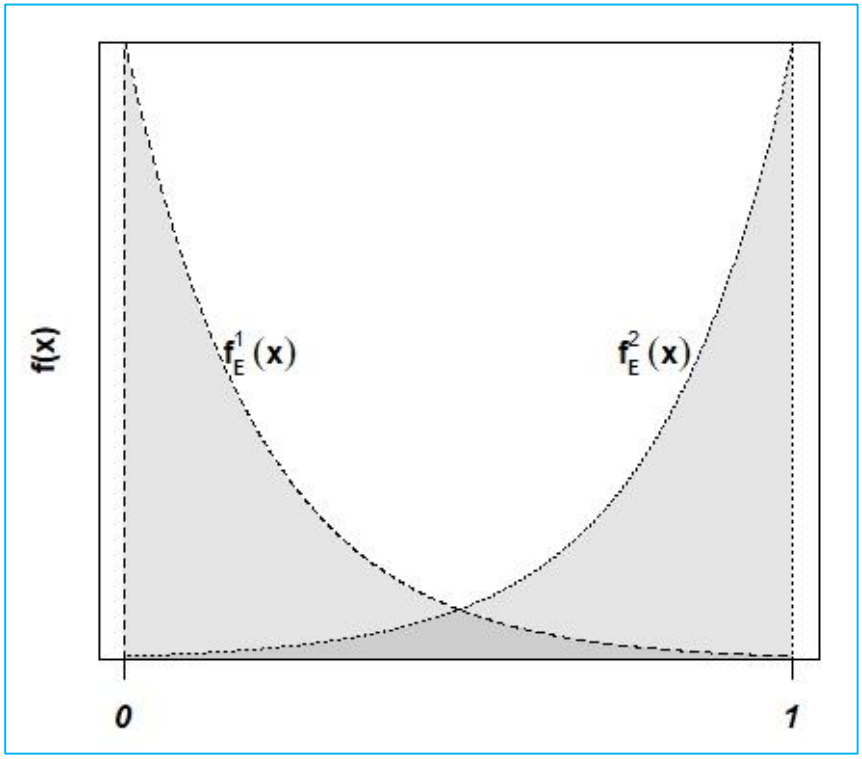


$$f_E(x;\lambda) = \lambda_1 \cdot e^{-\lambda_1 x} \qquad f_E'(x;\lambda,\mu) = \lambda_2 \cdot e^{-\lambda_2(\mu - x)}$$

As this distribution is not bounded, we standardized it to the domain of the interval (0,1) and both exponentials are included in the distribution mixture (Figure 35):

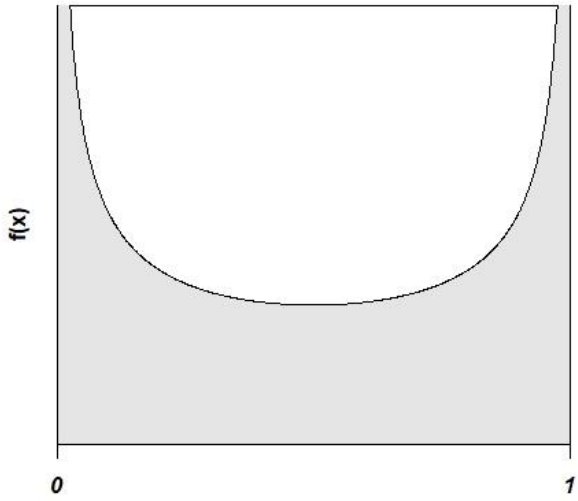
$$f_E^1(x;\lambda) = \lambda_1 \cdot e^{-\lambda_1 x}/(1-e^{-\lambda_1}) \qquad f_E^2(x;\lambda,\mu) = \lambda_2 \cdot e^{-\lambda_2(\mu-x)}/(1-e^{-\lambda_2})$$

Figure 35. Combining two exponential distributions.



❖ ***Beta (B)***. This distribution has two parameters of shape (α) and scale (β). It is especially appropriate for this kind of data because its domain is the interval (0,1) and is enough flexible to fit both peaks. Its density (f_β) is represented in Figure 36.

Figure 36. Beta distribution

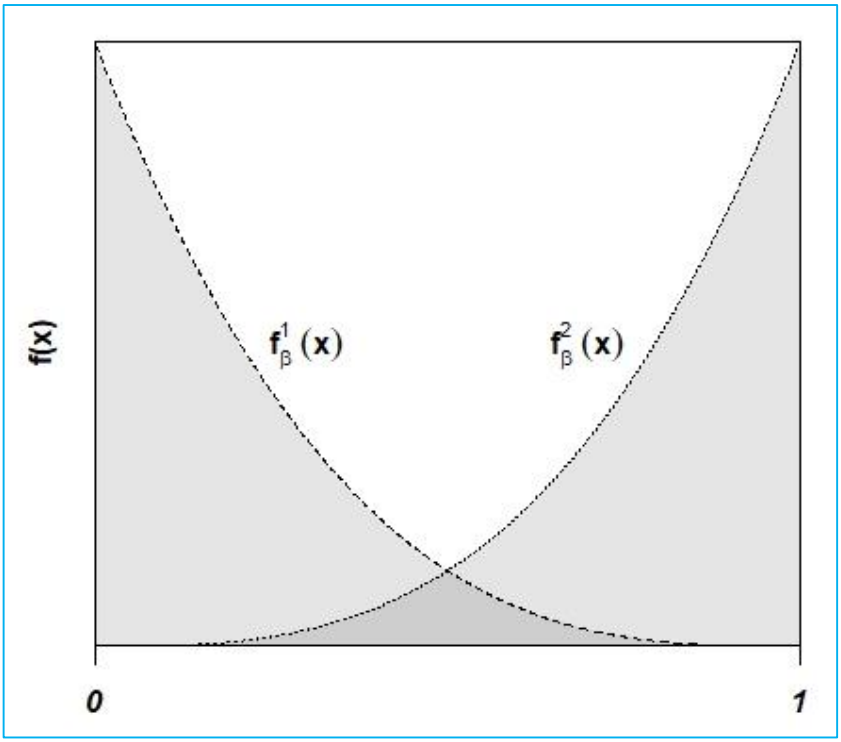


$$f_B(x;\alpha,\beta) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}$$

$$B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

As with the triangular and exponential distributions, two beta distributions (f_β^1, f_β^2) are combined in order to obtain more flexibility (Figure 37).

Figure 37. Combining two beta distributions.



To sum up, four different distribution mixtures were tested, and they are expressed in the next box.

Simulation model
❖ U(0,1) + $T(a_1 = 0, b_1, c_1 = a_1 = 0) + T(a_2, b_2 = 1, c_2 = b_2 = 1)$.
❖ U(0,1) + $E(\lambda_1) + E'(\lambda_2, \mu = 1)$
❖ U(0,1) + $B(\alpha, \beta)$
❖ U(0,1) + $B(\alpha_1, \beta_1) + B(\alpha_2, \beta_2)$

4.4.1.3 Optimization formulae

The parameters of the mixture were estimated by maximizing the log-likelihood using the *Augmented Lagrangian Minimization Algorithm* method[67]. After this, the best distribution mixture was selected by means of two goodness-of-fit measures: the Akaike Information Criterion (AIC) and the Kolmogorov-Smirnov statistic. The four Log-Likelihood functions from the four mixture distributions, with their corresponding constraints are detailed in the following boxes.

Uniform + 2 Triangle distributions

Max $$l = \sum_{i=1}^{i=n} \log\left[\pi_0 + \pi_{11} \cdot \left(\frac{2}{b_1} - \frac{2x_i}{b_1^2}\right) + (1 - \pi_0 - \pi_{11}) \cdot \left(\frac{2}{(1-a_2)^2} \cdot (x_i - a_2)\right)\right]$$

Subject to
$$\begin{array}{rcl} \pi_0 & \geq & 0 \\ \pi_{11} & \geq & 0 \\ \pi_0 + \pi_{11} & \leq & 1 \end{array} \qquad \begin{array}{rcl} b_1 & \geq & 0 \\ a_2 & \geq & 0 \end{array} \qquad \begin{array}{rcl} b_1 & \leq & 1 \\ a_2 & \leq & 1 \end{array}$$

where π_0, π_{11} represent the proportion of studies coming from, respectively, the (uniform) distribution under the null hypothesis and the left triangular distribution (part of the alternative hypothesis).

Uniform + 2 Exponential distributions

Max $$l = \sum_{i=1}^{i=n} \log\left[\pi_0 + \pi_{11} \cdot \frac{\lambda_1 \cdot e^{-\lambda_1 x_i}}{1 - e^{-\lambda_1}} + (1 - \pi_0 - \pi_{11}) \cdot \frac{\lambda_2 \cdot e^{-\lambda_2 (1-x_i)}}{1 - e^{-\lambda_2}}\right]$$

Subject to
$$\begin{array}{rcl} \pi_0 & \geq & 0 \\ \pi_{11} & \geq & 0 \\ \pi_0 + \pi_{11} & \leq & 1 \end{array} \qquad \begin{array}{rcl} \lambda_1 & \geq & 0 \\ \lambda_2 & \geq & 0 \end{array}$$

where π_0, π_{11} represent the proportion of studies coming from, respectively, the (uniform) distribution under the null hypothesis and the left exponential distribution (part of the alternative hypothesis).

Uniform + 1 Beta distribution

Max $$l = \sum_{i=1}^{i=n} \log\left[\pi_0 + (1 - \pi_0) \cdot \frac{x_i^{\alpha_1 - 1}(1-x_i)^{\beta_1 - 1}}{B(\alpha_1, \beta_1)}\right]$$

Subject to
$$\begin{array}{rcl} \pi_0 & \geq & 0 \\ \alpha_1 & \geq & 0 \\ \beta_1 & \geq & 0 \\ \pi_0 & \leq & 1 \end{array}$$

where π_0 is the proportion of events coming from the (uniform) distribution under the null hypothesis.

Uniform + 2 Beta distributions

Max $$l = \sum_{i=1}^{i=n} \log\left[\pi_0 + \pi_{11} \cdot \frac{x_i^{\alpha_1 - 1}(1-x_i)^{\beta_1 - 1}}{B(\alpha_1, \beta_1)} + (1 - \pi_0 - \pi_{11}) \cdot \frac{x_i^{\alpha_2 - 1}(1-x_i)^{\beta_2 - 1}}{B(\alpha_2, \beta_2)}\right]$$

Subject to
$$\begin{array}{rcl} \pi_0 & \geq & 0 \\ \pi_{11} & \geq & 0 \\ \pi_0 + \pi_{11} & \leq & 1 \end{array} \qquad \begin{array}{rcl} \beta_1 & \geq & 0 \\ \alpha_2 & \geq & 0 \\ \beta_2 & \geq & 0 \end{array} \qquad \begin{array}{rcl} \alpha_1 & \geq & 0 \\ \alpha_1 - \beta_1 & \geq & 0 \\ \alpha_2 - \beta_2 & \leq & 0 \end{array}$$

where π_0, π_{11} represent the proportion of studies coming from, respectively, the (uniform) distribution under the null hypothesis and the left beta distribution (part of the alternative hypothesis). The last two restrictions are imposed to guarantee that π_{11} is the proportion associated with the left-hand (and not to the right-hand) beta distribution.

Once the best model was found, the proportion of studies with different population outcome variance was estimated as $1 - \hat{\pi}_0$, being $\hat{\pi}_0$ the estimated proportion of p-values coming from the uniform distribution.

4.4.2 Results

4.4.2.1 Between-arm comparison

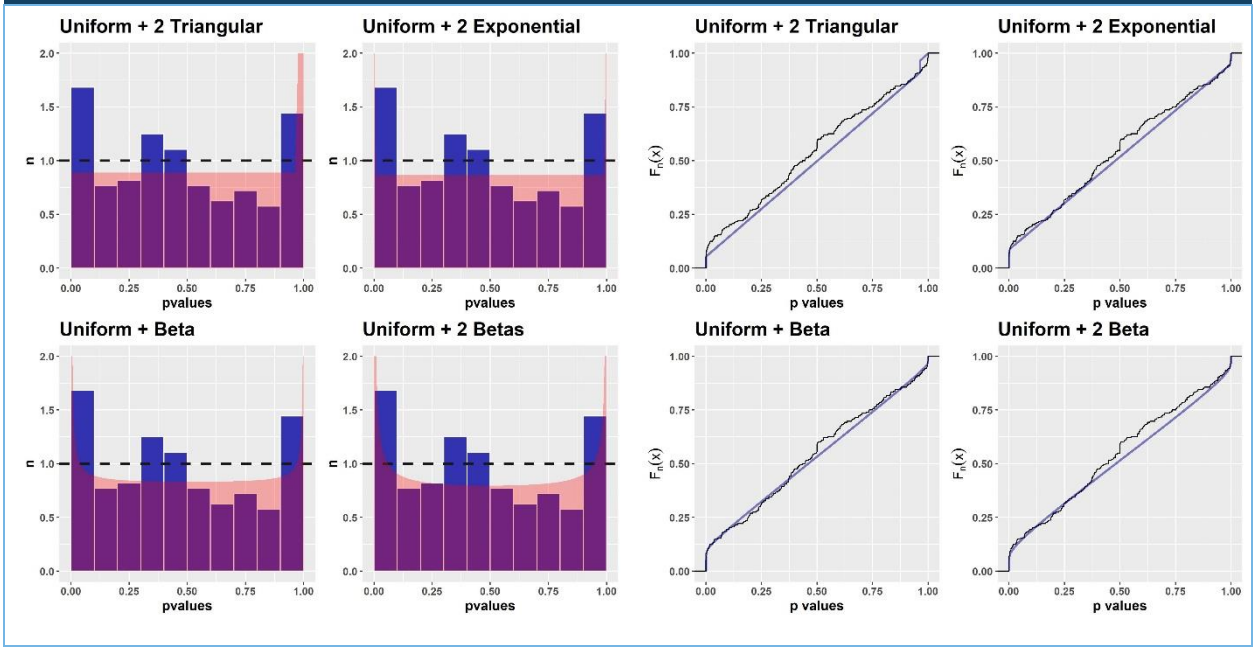
For the between-arm comparison, we used MLE to fit our four models to the empirical p-values, and several goodness-of-fit measures were calculated. Table 13 provides a summary of the results derived from each distribution mixture, and Figure 38 shows the model and empirical data comparison of the density and distribution functions.

Table 13. Goodness-of-fit measures and estimated proportion of constant-effect trials in between-arm comparison.

Mixture Distribution	no. parameters	AIC	KS	$\hat{\pi}_0$ (95% *CI*)
Uniform + 2 triangular	4	-126.0	0.10	0.88 (0.84, 0.94)
Uniform + 2 exponential	4	-139.6	0.08	0.86 (0.81, 0.91)
Uniform + 1 beta	3	-562.9	0.06	0.80 (0.72, 0.88)
Uniform + 2 betas	6	-568.8	0.08	0.71 (0.58, 0.84)

KS: Kolmogorov-Smirnov statistic; $\hat{\pi}_0$: estimated proportion of p-values coming from a uniform distribution; CI: Confidence interval

Figure 38. Distribution mixtures in between-arm comparison: histogram of the empirical data with overlapped theoretical density (left) and comparison of theoretical (blue) versus empirical (black) cumulative density functions.



The fits provided by mixtures formed by exponential and triangular distributions were clearly worse than those obtained using the beta distribution while taking into account the AIC statistic. This may be due in part to these distributions having poor flexibility in presenting different curvature patterns that capture the shape of the empirical peaks. Regarding the models that involve

the beta distribution, the two goodness-of-fit indicators gave discordant results: while the model with two beta distributions presented a better AIC value, the mixture distribution with a single beta distribution derived a better value of the Kolmogorov-Smirnov (KS) statistic. The KS statistic value is easily perceived at the right hand of Figure 38: a smaller maximum distance between the empirical and theoretical distributions is present for the mixture with a single beta distribution. The bottom Q-Q plots hint at the reason for this mismatch. The model that includes two beta distributions fitted all the quantiles almost perfectly, except for those located in the central range (near 0.5). The presence of a less marked peak in the central part of the empirical distribution of the p-values penalizes this mixture, resulting in a higher KS statistic value than, the distribution with a single beta.

Applying the principle of parsimony, we chose the distribution with the fewest number of parameters, which is the one that uses a single distribution to model the two empirical peaks. With this distribution mixture, the estimated proportion of studies coming from a uniform distribution and, therefore, from a population of constant effects was 0.80 (95% CI from 0.72 to 0.88). Unfortunately, as this mixture distribution fits a unique distribution to all the studies under the alternative hypothesis, we could not distinguish from among these trials which one provided less (p-values close to 0) or greater (p-values close to 1) variance in the treated arm.

4.4.2.2 Comparison over time

We applied the same methodology to the 95 p-values that resulted from the paired tests for over-time comparisons of the variances. We used the same four abovementioned mixture distributions. In this case, Table 14 shows that the model using two beta distributions provided a better fit in both the AIC and the KS statistics, and Figure 39 validates this finding. Thus, we have chosen the mixture that comprises two beta distributions.

Table 14. Goodness-of-fit measures and estimated proportion of constant effects trials in comparison over time.

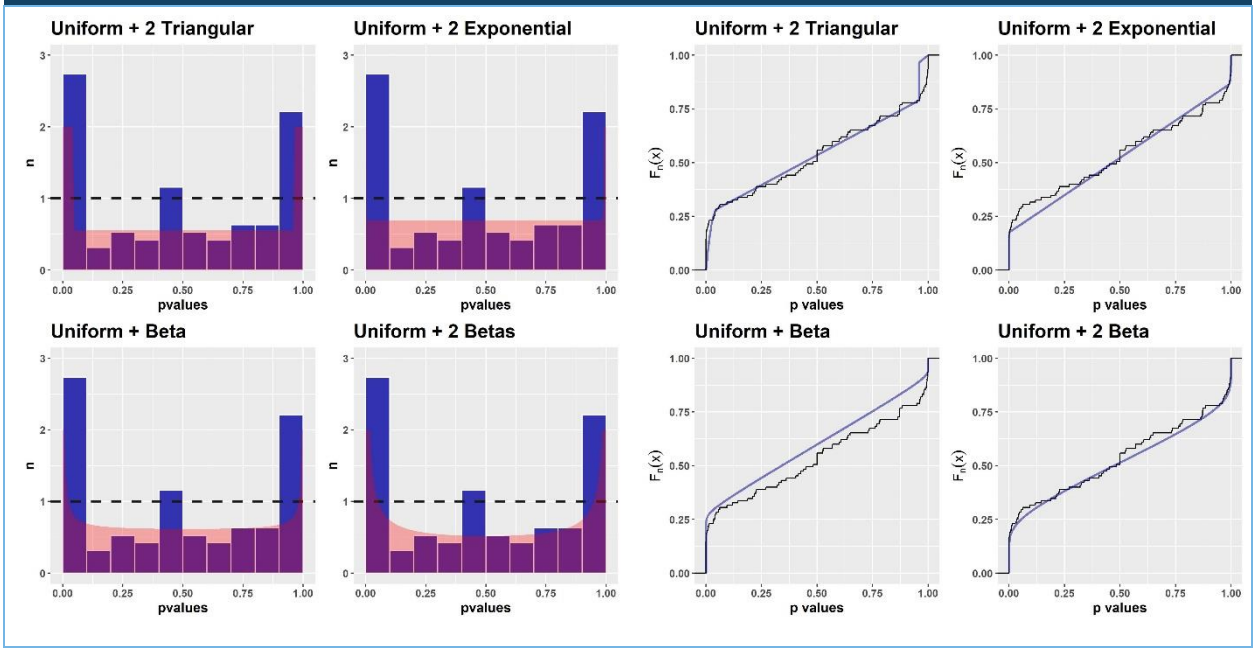
Mixture Distribution	no. parameters	AIC	KS	$\hat{\pi}_0$ (95% CI)
Uniform + 2 triangulars	4	-122.1	0.18	0.55 (0.44, 0.66)
Uniform + 2 exponentials	4	-217.7	0.09	0.70 (0.60, 0.79)
Uniform + 1 beta	3	-1356.4	0.11	0.57 (0.44, 0.71)
Uniform + 2 betas	6	-1393.4	0.06	0.35 (0.15, 0.55)

KS: Kolmogorov-Smirnov statistic; $\hat{\pi}_0$: estimated proportion of p-values coming from a uniform distribution; CI: Confidence interval.

According to the selected model, the proportion of trials in which the intervention did not affect the dispersion of the outcome was 0.35 (95% CI from 0.15 to 0.55). This estimate is clearly inferior to the one for the comparison between arms; but in some way it is reasonable, since the "time effect" makes a contribution by modifying the variability of some primary endpoints. Furthermore, this estimate agrees neither with that of the analysis from Chapter 3 nor with the estimates of the other sensitivity analyses over time.

This model uses two different distributions to model each of the empirical distribution peaks, and it allows us to estimate the proportion of tests that point towards an increase (0.03) or a decrease (0.62) of the variance over time.

Figure 39. Distribution mixtures in comparison over time: histogram of the empirical data with overlapped theoretical density (left) and comparison of theoretical (blue) versus empirical (black) cumulative density functions.



The last ancillary analysis in appendix E contains a simulation study for assessing the appropriateness of the distribution mixture approach by simulation.

5 Discussion

5.1 Summary findings and explanations

The aim of this work is primarily twofold. First, it studies whether the comparison of variances can provide any evidence about the constant effect assumption. Second, it seeks to provide preliminary evidence on how frequently this assumption behind precision medicine holds. To answer these questions, we use published RCTs with numerical outcomes to estimate, on the one hand, the outcome variance ratio between two randomized groups and, on the other, an analogous ratio over time between outcome and baseline variances in the treated arm.

Regarding the first objective, all analyses point toward an unexpected result: the outcome variability in the experimental group is lower than both that of the reference group and the variability at the beginning of the study. The point estimates of all the fitted models (see Table 8) show that the variance in the experimental group is between 10.4% and 13.1% lower than in the control group and between 13.1% and 14.8% lower than the baseline. Specifically, the variance of the treated group in the adjusted models is 11.3% (95% CI from 3.0 to 18.9%) lower than that of the control group and 13.9% (95% CI from 2.0 to 24.4%) lower than that of the baseline.

Concerning the second objective, we have provided a rough estimate of the proportion of interventions with different variability in both arms, which might benefit from more precise medicine. Considering the most extreme result of the between-arm comparison indicated in Table 15, one out of 14 interventions (n=15, 7.2%) had greater variance in the treated arm while one out of eight interventions (n=26, 12.5%) had lower variance. Overall, we have found evidence of effect variation in only one out of five trials (n=41, 19.7%), suggesting a limited role for tailored interventions. These might be pursued either through finer selection criteria (common effect within specific subgroups) or with *n-of-1* trials (no subgroups of patients with a common effect). The most important finding is that the remaining 80.3% of interventions were compatible with a constant treatment effect.

The analyses of the change over time in the treated arm generally agree with the findings of the between-arm comparison, 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 increasing naturally over time — contrary to our results.

Table 15. Classification of the studies according to their variability discrepancies with different methods (main and sensitivity analyses). ¥Over-time comparison was performed only on studies reporting enough information to obtain the variability of the change from baseline to outcome.

Comparing variances	N	Method	After treatment, variability …		
			Increased n (%)	Decreased n (%)	Did not changed (n %)
Outcome between treatment arms	208	**Random effects model**	**15 (7.2%)**	**26 (12.5%)**	**167 (80.3%)**
		Heuristic method	10 (4.8%)	19 (9.1%)	179 (86.0%)
		Simulation study	3 (1.4%)	18 (8.7%)	187 (90.0%)
		F test	14 (6.7%)	26 (12.5%)	168 (80.8%)
		Mixture distribution	-	-	166 (79.8%)
Outcome versus baseline in treated arm	95¥	**Random effects model**	**16 (16.8%)**	**22 (23.2%)**	**57 (60.0%)**
		Heuristic method	10 (10.5%)	18 (18.9%)	67 (70.5%)
		Paired test	16 (16.8%)	22 (23.2%)	57 (60.0%)
		Mixture distribution	3 (3.2%)	59 (62.1%)	33 (34.7%)

Regarding the subgroup analyses, we found that variability seems to decrease for effective treatments; 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; and 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 thus lack the scope to respond very much), which subsequently incurs savings in side effects and costs.

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 implying greater severity, e.g., pain), greater variance (20.4%) is obtained in the treated arm (see Figure 13). However, in

the studies with outcomes that measure the degree of healthiness (high values implying better condition, e.g., mobility), the average variances match between arms, which does not suggest a ceiling effect. As mentioned previously, another reason might be that the treatment effect is not additive on the scale used for analysis, suggesting that further exploration of other metrics and transformations would be suitable. For instance, if the treatment acts proportionally rather than linearly, the logarithm of the outcome would be a better scale.

5.2 Limitations

The results of this work rely on published articles, which raises some relevant issues. First, some of our analyses are based on Normality assumptions that are unverifiable without access to raw data.

Second, 330 out of 542 (61.6%) manuscripts act contrary to CONSORT guideline[68] advice in that they do not report variability (Figure 4). Thus, the included studies may not be representative. The selected articles are from the years 2004 to 2013, and it is probable that nowadays the percentage of trials that do not report variability would be much lower.

Third, trials are usually powered to test constant effects and thus the presence of greater variability would lead to an underpowered design. In other words, if the control group variance is used to plan the trial, increased treatment group variance would reduce power (perhaps leading to non-publication).

Fourth, and in relation to the aforementioned limitation, the sample size of the collected studies does not provide enough power to detect differences between variances. For this reason, additional sensitivity analyses have been proposed (e.g., sensitivity analyses I, II and IV) that take this concern into account, reaching similar results.

Fifth, the heterogeneity observed in the random effects model may be the result of methodological inaccuracies arising from typographical errors in data translation, inadequate follow-up, insufficient reporting, or even data fabrication[58].

A sixth limitation is that many clinical trials are not completely randomized. For example, multicenter trials often use a 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[54].

The seventh limitation refers to our selection criteria for collecting articles. Our work focuses only on parallel clinical trials and, despite these being the most frequent, we are leaving out other designs, such as crossover or cluster trials. Furthermore, trials with quantitative outcome are not

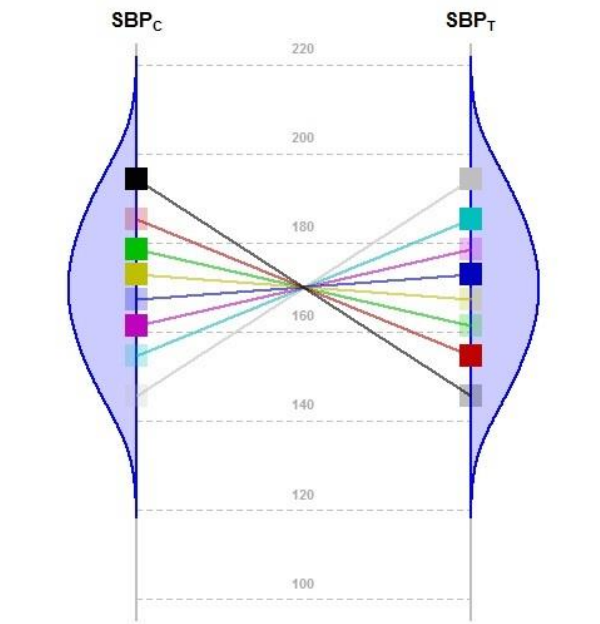
comparable to those with binary or time-to-event endpoints. The former, in general, requires a lower sample size to detect clinically relevant treatment effects, and it would not be convenient to relate a small number of patients with a low-quality study.

Eighth, the choice of the experimental and reference groups may be controversial in some cases where both interventions include active treatments. However, the objectives of the studies in most cases dispel this doubt, and conflicts were usually resolved through the consensus of two evaluators.

The ninth limitation refers to the fact that an alternative to subgroup analysis would have been to include the subgroup variables as covariates in the models. Even though this is feasible, this option could cause some instability in models fitted with so many covariates and, further, the variable selection procedure would have exponentially increased the number of models to analyze.

Finally, the most important limitation of our study arises from the fact that, although a constant effect always implies homoscedasticity on the chosen scale, the reverse is not true; i.e., homoscedasticity does not necessarily imply a constant effect. For example, the highly specific and non-parsimonious situation reflected in Figure 40 reveals homoscedasticity but without a constant effect. Nevertheless, a constant effect is the simplest explanation for homoscedasticity. This point is further discussed in the next subsection.

Figure 40. Example of homoscedasticity without constant effect.



5.2.1 Conditions for homoscedasticity without constant effect

Under an additive model, the relationship between the two potential (Y^0, Y^1) outcomes in each group is the following:

$$Y^1 = Y^0 + E \qquad E \sim N(\mu_E, \sigma_E)$$

With *E* being the treatment effect. We can devise four situations, which are detailed in the box below.

Relationship between variances according to different scenarios

- **Fixed effect model**. If the treatment effect is constant [$V(E) = 0$] and independent of the potential outcomes, the homoscedasticity holds.

$$V(Y^1) = V(Y^0 + E) = V(Y^0) + V(E) = V(Y^0)$$

- **Random effects model with <u>no correlation</u> between treatment effect and the potential outcome in the control group**. As σ_E^2 is positive, $V(Y^1)$ should be larger than $V(Y^0)$, implying heteroscedasticity.

$$V(E) = \sigma_E^2 > 0 \rightarrow V(Y^1) = V(Y^0 + E) = V(Y^0) + V(E) > V(Y^0)$$

- **Random effects model with <u>positive correlation</u> between treatment effect and the potential outcome in the control group**. This situation also leads to heteroscedasticity, with $V(Y^1) > V(Y^0)$

$$V(E) = \sigma_E^2 > 0, \rho_{Y^0,E} > 0 \rightarrow V(Y^1) = V(Y^0 + E) = V(Y^0) + V(E) + 2Cov(Y^0, E) > V(Y^0)$$

- **Random effects model with <u>negative correlation</u> between treatment effect and the potential outcome in the control group.** Let *K* be equal to $-\frac{\sigma_E}{2\sigma_{Y_0}}$; then, the following situations can occur:
  - *Homoscedasticity*. $V(Y^1) = V(Y^0)$ if $\rho_{Y^0,E} = K$
  - *Heteroscedasticity* with $V(Y^1) > V(Y^0)$ if $\rho_{Y^0,E} > K$
  - *Heteroscedasticity* with $V(Y^1) < V(Y^0)$ if $\rho_{Y^0,E} < K$

Proof:

$$V(Y^1) = V(Y^0 + E) = V(Y^0) + V(E) + 2Cov(Y^0, E) = V(Y^0) + V(E) + 2 \cdot \rho_{Y^0,E} \cdot \sigma_{Y^0} \cdot \sigma_E$$

$$V(Y^1) = V(Y^0) \leftrightarrow \sigma_E^2 + 2\rho_{Y^0,E}\sigma_{Y^0}\sigma_E = 0 \leftrightarrow \sigma_E + 2\rho_{Y^0,E}\sigma_{Y^0} = 0 \leftrightarrow \boldsymbol{\rho_{Y^0,E} = -\frac{\sigma_E}{2\sigma_{Y_0}}}$$

Thus, under the standard parameterization of an additive effect, the only situation without a constant effect but with homoscedasticity has a specific negative correlation ($-\frac{\sigma_E}{2\sigma_{Y_0}}$). For example, if the variances of the effect and the outcome in the control group are identical, then this negative correlation should be -0.5 in order for the assumption to hold. We have proved that we need a very specific correlation in order to obtain homoscedasticity without a constant effect.

5.3 Main conclusions and impact

We have aimed to show that comparing variances provides evidence on whether or not precision medicine is a sensible choice for a specific study. When both arms have equal variances, then a simple interpretation is that the treatment effect is constant, which, if correct, would render futile any search for predictors of differential response. This means that the average treatment effect can

be seen as an individual treatment effect (not directly observable), which supports the use of a unique clinical guideline for all patients within the eligibility criteria. This in turn also supports using parallel controlled trials to guide decision-making in these circumstances. Otherwise, heteroscedasticity may suggest a need to further specify the eligibility criteria or search for an additive scale[54,69]. Because interaction analyses cannot include unknown variables, there might be value in repeating trials once any new potential interaction variable emerges (e.g., a new biomarker) as a candidate for a new subgroup analysis. However, we should highlight that past attempts to corroborate statistically significant subgroup differences have failed because the initially observed interactions could not be reproduced[19].

We have described how homoscedasticity can be assessed when reporting trials with numerical outcomes, regardless of whether every potential effect modifier is known. As far as we know, no studies prior to ours[70], compare variances in order to assess the suitability of precision medicine. To this date, February 5, 2021, we were aware of 20 publications that had cited our work, most of them coming from psychology journals (n=14). Next, we will briefly summarize our contribution to these studies, identifying those that use our methodology (n=10) and those that cite us for other reasons (n=10).

Among the studies that emulate our analysis based on the random effects model applied on the logarithm of the variance ratio, a recent 2019 meta-analysis[71] published in JAMA Psychiatry reproduced our method and achieved results similar to ours on 52 RCTs to assess the efficacy of antipsychotic drugs in patients with schizophrenia: the estimated average variance ratio (treated versus controls) was 0.97 (95% CI 0.95 to 0.99). Therefore, no evidence was found to indicate that antipsychotic drugs increased the outcome variance, thus suggesting no personal response to treatment but instead indicating that the variance was slightly lower in the treatment group than in the control group. Even though it cannot be ruled out that subsets of patients respond differently to treatment, it remains possible that the average treatment effect is a reasonable assumption for the individual patient. Another meta-analysis performed by Munkhölm et al.[72] included 222 RCTs assessing antidepressants in patients with major depressive disorder, with at least one group of patients receiving placebo. Again, they found no evidence for greater variance overall in the antidepressant group compared with placebo (variance ratio=1.00, 95% CI: from 0.98 to 1.01), and the heterogeneity among studies was null ($I^2 = 0\%$). Again, their findings did not provide empirical support for individual differences in response to antidepressants. A similar study conducted by Plöderl et al.[73] explored the need to personalize antidepressant prescriptions in patients with depression. Applying the same technique, they found an estimated average variance ratio between outcomes equal to 1.01, with a 95% CI from 0.99 to 1.02. Senior et al.[74] applied a

version of our method using two alternative measures: the *log-transformed variability ratio* (the ratio of two standard deviations) and the *log-transformed CV ratio* (the ratio of two coefficients of variation) under the setting of dependent responses. They found a heterogeneous treatment effect for lifestyle interventions on gestational weight gain in obese women and a homogenous treatment effect for low-GI diets on glycemic control in diabetics. Mills et al.[75] show several methods to assess the heterogeneity based on variance comparison both for a single trial when the raw data is available and for meta-analyses. They present our methodology as a feasible one in the absence of pre-specified effect modifiers. Radua et al.[76] performed a meta-analysis to estimate the variance ratio at 6 (95% CI 0.89–1.12) and at 12 months (95% CI 0.94–1.25) in treatments for individuals at clinical high risk for psychosis. They also found no evidence of greater variance on average in any group. In a systematic review of RCTs including participants with psychiatric disorders conducted by Winkelbeiner et al.[77], an average greater variability in the active stimulation group than in the reference arm (variability ratio=1.05; 95% CI, 1.01-1.11) was found. This result might indicate that there is a component of variation in the treatment effect due to patient-by-treatment or subgroup-by-treatment interaction. In another systematic review, Watson et al.[78] studied the effect of a pain neuroscience education intervention in order to quantify the inter-individual variation in pain, disability and psychosocial outcomes using a random effects meta-analysis. Treatment effect heterogeneity could not be proved. Neumeier et al.[79] assessed the heterogeneity of antipsychotic treatment effects in patients with schizophrenia disorders. In this case, they found some variance increase in treated patients with respect to weight gain (1.08; 95% CI: 1.02-1.14) and prolactin levels (1.38; 95% CI: 1.17-1.62) outcomes. Finally, the last study to use our methods to this date was conducted by Smith et al.[80] They did not find significant differences in the pre-post body weight change variance between the study arms when comparing Low-Carbohydrate and Low-Fat Diets. The application of the comparison of variances in all these studies to evaluate the heterogeneity of the treatment effect, leads to postulate the proposed methodology as a valid procedure to discern whether or not to personalize the treatments.

Other studies have cited our work to strengthen some of their assertions, especially, in the discussion section. In a letter published in *Jama Psychiatry*, Winkelbeiner et al.[81] cited our study to argue that different treatment effects might lead to different variances between arms both at individual or at subgroup level. Atkinson et al.[82] shows that the dichotomization of continuous variables in psychological research implies biased conclusions. Based on our findings, they claim that any substantial treatment effect heterogeneity that is larger than the heterogeneity in the data owing to random within subject variability over time would be revealed if the SD of changes in the treatment arm is larger than that in the comparator group. In an editorial of *Acta Psychiatrica*

Scandinavica, Homan et al.[83] weighed the benefits of clozapine and regretted that the treatment effect heterogeneity is very rarely evaluated or tested, although they know that the presence of heterogeneity is scarce. In another article of the same first author[84], the investigators (wrongly) cited us to justify the lack of a placebo group in their study based on the fact that the outcome variability in the treated group would have been greater than the (not observed) variance in a placebo group. Feczko et al.[85] present several methods for assessing the heterogeneity together with their limitations. They urge for the need to evaluate the homogeneity assumption in all the clinical studies and they mention the variance comparison as an option to achieve this goal. Hieronymus et al.[86] in a way, criticize our method by reasoning that similar variabilities do not exclude a heterogeneous treatment effect. We guess that this statement is motivated by Figure 40 (also present in the article derived from this work). As we have already explained, the purpose of this illustration is to show how odd this situation is and make it clear to the reader that the most plausible cause of homoscedasticity is a constant treatment effect. There are three articles about the role of instrumental variables on causal inference that cited our work. Sandu et al.[87] propose a new two-step randomization analysis to enhance the value of feasibility studies. Among the list of limitations, they remark that some assumption to apply the instrumental variables theory might not be met since the observed intervention effect could only apply to a specific subgroup of patients referring to some scenarios drawn in Figure 1 and Figure 2. Pires et al.[88] comment on that these assumptions could be empirically checked and give the example that the instrument effect heterogeneity would generally imply that treatment is heteroscedastic with respect to the instrument. Bowden et al.[89] emphasize one limitation of our study, which is that the low number of trials with statistically different variances reflect a lack of power to detect these discrepancies, which generally require larger sample sizes. Finally, Aron[90] in a chapter of the book "*Precision Medicine and Complexity*" stands out that the recent success of personalized medicine might be due more to the hype than to scientific evidence.

There are several reasons why the findings of this work do not invalidate precision medicine. First, some studies indicate glaringly different variability in the response, thus indicating the presence of a non-constant effect. This heterogeneity might be the result of relevant undetected factors interacting with the treatment, which would indeed justify the suitability of precision medicine. Second, the outcomes of some types of interventions, such as surgeries, are greatly influenced by the skills and training of those administering the intervention. Such situations could have some effect on increasing variability. Third, this study focuses on numerical endpoints, for which time-to-event and categorical outcomes are out of scope.

We do not intend to discourage researchers from pursuing precision medicine, but to instead encourage them to get a better sense of its potential at the outset. As Senn[91] stated, there is room for improvement in this field, and any a posteriori inference that subgroups have a better response to a specific treatment is a bad idea. On the other hand, the *N* of 1 trials is somehow a good choice for assessing differences between administering the same intervention at various times on the same individual, specifically when seeking to compare the data with differences from other drugs that are administered the same way.

Critics of medical protocols developed through classical evidence-based medicine may argue that ATE does not imply a constant effect. However, we propose that some evidence is needed to indicate that the effect is not constant before advocating for precision medicine (rather than the opposite). In summary, medical researchers must be transparent about their premise; and statisticians should develop methods to study these premises.

The contribution of this work to the scientific community is twofold. On the one hand, it is intended that researchers do not evade the question of whether the constant effect premise is reasonable in the interventions or treatments in their studies. On the other hand, a new methodology is provided that has been proven feasible in view of the studies that have cited our article derived from this PhD doctoral thesis.

5.4 Future work

Future lines of work are divided into 3 main objectives:

1. To further explore the implications of sample size having a heterogeneous treatment effect in studies with continuous response.
2. To provide evidence by a literature review to indicate the degree of heterogeneity in treatments effects from other types of studies, such as crossover RCTs and time-to-event studies.
3. To propose new techniques for the comparison of variances to test the heterogeneity of the treatment effect in a single specific study.

The goal of the first objective is to broaden and formally address the topic discussed in Section 1.3.3. We have already seen that assuming a constant treatment effect when there is none could have important implications for the sample size, since this can lead to different variabilities between arms. The problem is even greater if the outcome distribution in the experimental group is bimodal, where the distribution of the sample mean statistic no longer has the usual variance of $\sigma/\sqrt{n}$, which is derived from the central limit theorem. Therefore, the classic sample calculations are invalid. If trialists can anticipate different behaviors among individuals, it is necessary to make

assumptions in the study design and take them into account in the sample size calculations in order to achieve adequate power. Appendix A contains a first attempt to enumerate several examples under different scenarios in which some issues could arise due to a non-constant treatment effect.

Regarding the second objective, one of the medical fields where personalized medicine plays an important role is oncology. Studies seeking evidence in this area often have a time-to-event outcome (usually to disease progression or death). It would be interesting to review how the variability of the response behaves in both groups and what the implications would be in trials that use hazard ratio as a treatment effect measure and that are based on the proportional hazard assumption. This would incur additional difficulties, such as the presence of censored times or variability not only in the treatment effect among patients but also over time. Dealing with these issues is a challenge we want to overcome.

Finally, because trialists are interested in whether their interventions produce a reasonably constant effect on the entire target population, they would thus prefer to test this hypothesis in a single study. Being able to verify the heterogeneity of the treatment effect would be a useful tool for health administrations to authorize drugs. Currently, aside from subgroup analyses based on pre-specified variables, no restrictions on heterogeneity are required by drug regulatory agencies. As a consequence, it is just as feasible to put on the market a drug with a constant effect as it is to authorize a drug with the same average treatment effect but also huge variability and that may even be harmful to some patients. Precision medicine is becoming increasingly fashionable due to the belief that one specific intervention can produce different responses in each patient. For this reason, our aim will be to measure this heterogeneity in a single RCT by following the methodology developed in this work, which is based on the comparison of outcome variances. This is not a new pursuit, as the extension of the CONSORT reporting guidelines for non-pharmacological interventions[92] requests that the performance of interventionists be reported. Rather than describing individual performance or ranking the surgeons, the correct challenge is to ascertain how much heterogeneity is added by interventionists. If not us, we would be happy to see other researchers explore the clinical acceptability of this alternative way of reporting heterogeneity. Some work has already been done in this regard. Caughey et al.[93] defined a method for finding one-sided confidence intervals based on the idea of the permutations test that Fisher[94] proposed in the last century. Nevertheless, the drawback of these intervals is that they are very inaccurate, and adding the information provided by the observed variances could be interesting for testing heterogeneity and, subsequently, for prioritizing interventions and treatments that are closer to a constant treatment effect.

6 Shiny app and R code

Until the eighteenth century, disciplines based on deductive reasoning adhered to self-contained treaties that needed no additional information for self-validation. With the growth of statistics at the beginning of the twentieth century, researchers published their results based on data that—although indispensable for confirming their hypotheses—were not published for different reasons: space, privacy, budgetary considerations or perhaps egocentrism.

In studies with patient data, strict measures should be established to maintain privacy. Regulatory agencies have always demanded that clinical trial data be made available, but it was only in the past decade that its publication began to be promoted[95–97] , although the intended purposes for this varied: for comparing results; studying deviations from the protocol (http://compare-trials.org/); or allowing others to re-study them with new objectives.

Given the material and human cost of data collection, greater access can lead to less waste of resources[98]. In this study, we have built an application with the R *Shiny* package in order to share data interactively (see Section 6.1) . In addition, our data has also been made available through the *Figshare* repository:

https://figshare.com/articles/review_homoscedasticity_clinical_trials_csv/5552656

6.1 Shiny app

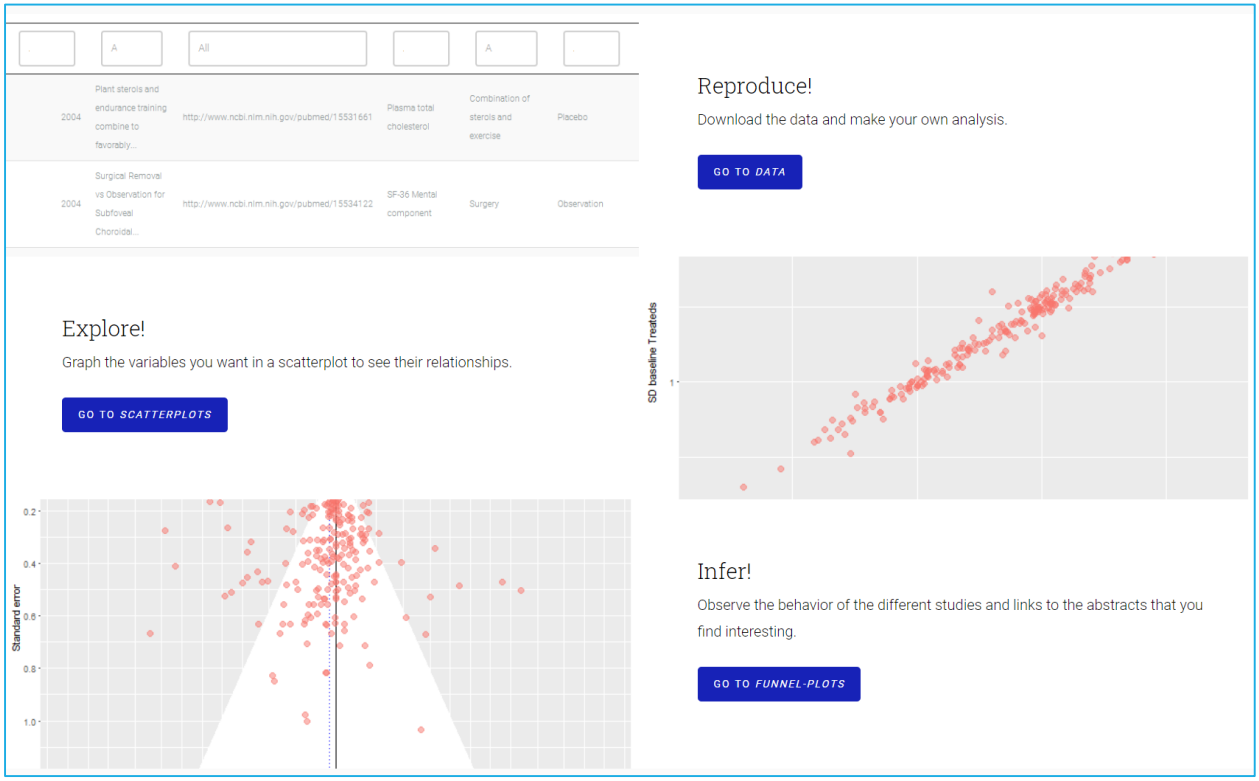
The main features of the Shiny app we developed for making the data more accessible are described in this section. The app is accessible via the following link:

http://shiny-eio.upc.edu/pubs/F1000_precision_medicine/

Shiny is an R package that allows sharing and interacting with data, thus opening up new opportunities for statisticians in the Open Data era. This kind of web app allows researchers to explore the data in a very simple way while sometimes arriving at new findings, detecting errors or discovering abnormalities in the analyses.

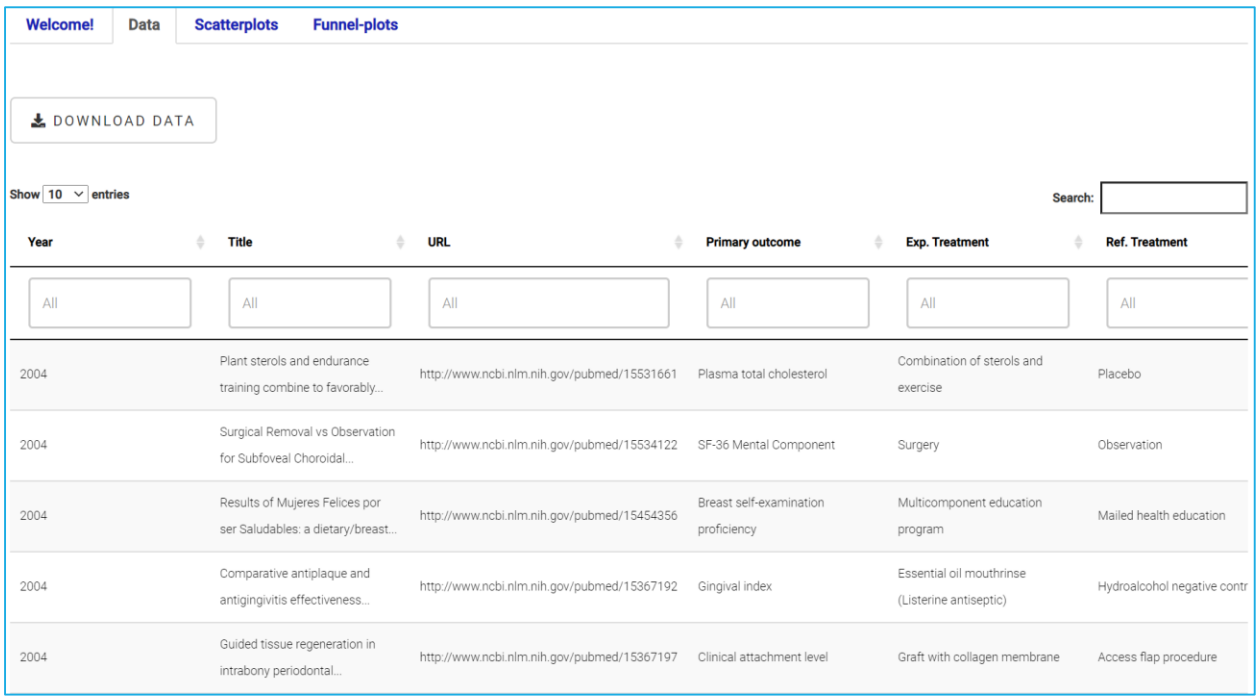
Figure 41 shows a screenshot of the web app's homepage, which contains a menu for accessing the different features.

Figure 41. Screenshot of the Shiny app homepage. It allows directly accessing the main features.



The *Data* tab gives access to the data in a comma-separated value (CSV) format by means of a download button or direct visualization of it on the web interface. The latter way also allows filtering any of the subgroup variables and performing global or partial searches for any term or figure in the entire dataset (Figure 42)

Figure 42. Screenshot of the *Data* tab of the Shiny app. Data can be filtered and/or downloaded.



Year	Title	URL	Primary outcome	Exp. Treatment	Ref. Treatment
All	All	All	All	All	All
2004	Plant sterols and endurance training combine to favorably...	http://www.ncbi.nlm.nih.gov/pubmed/15531661	Plasma total cholesterol	Combination of sterols and exercise	Placebo
2004	Surgical Removal vs Observation for Subfoveal Choroidal...	http://www.ncbi.nlm.nih.gov/pubmed/15534122	SF-36 Mental Component	Surgery	Observation
2004	Results of Mujeres Felices por ser Saludables: a dietary/breast...	http://www.ncbi.nlm.nih.gov/pubmed/15454356	Breast self-examination proficiency	Multicomponent education program	Mailed health education
2004	Comparative antiplaque and antigingivitis effectiveness...	http://www.ncbi.nlm.nih.gov/pubmed/15367192	Gingival index	Essential oil mouthrinse (Listerine antiseptic)	Hydroalcohol negative contr
2004	Guided tissue regeneration in intrabony periodontal...	http://www.ncbi.nlm.nih.gov/pubmed/15367197	Clinical attachment level	Graft with collagen membrane	Access flap procedure

The *Scatterplots* tab (Figure 43) supports building customized scatterplots with the following features:

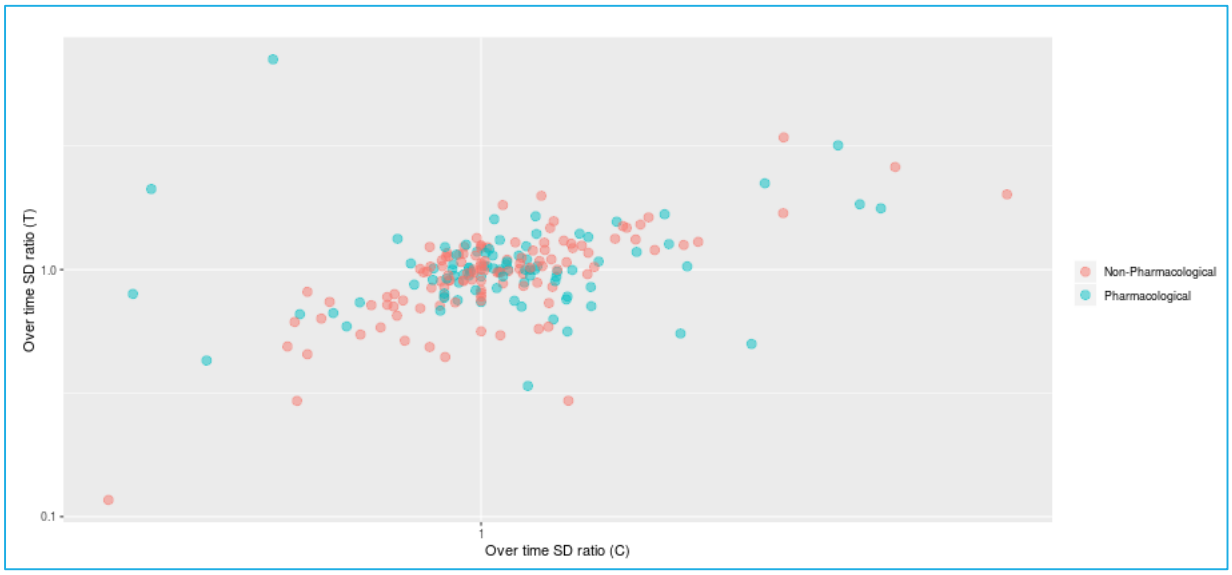
- Choice of variables represented on each axis (e.g., standard deviations, variances or sample sizes).
- Choice of the axis scale (additive or logarithmic).
- Choice of a subgroup factor in order to color-code points according to their categories.
- Dynamic filter for selecting studies according to specific characteristics.

In addition, the user can select a certain area on the graphic at the left to zoom in on the right side of the screen. If a point is clicked in this last graphic, the corresponding study details (such as the article title and other information) are shown at the bottom of the screen. Furthermore, clicking the title links the user directly to the published article's abstract on the PubMed interface.

Figure 43. Screenshot of the *Scatterplots* tab of the Shiny app.

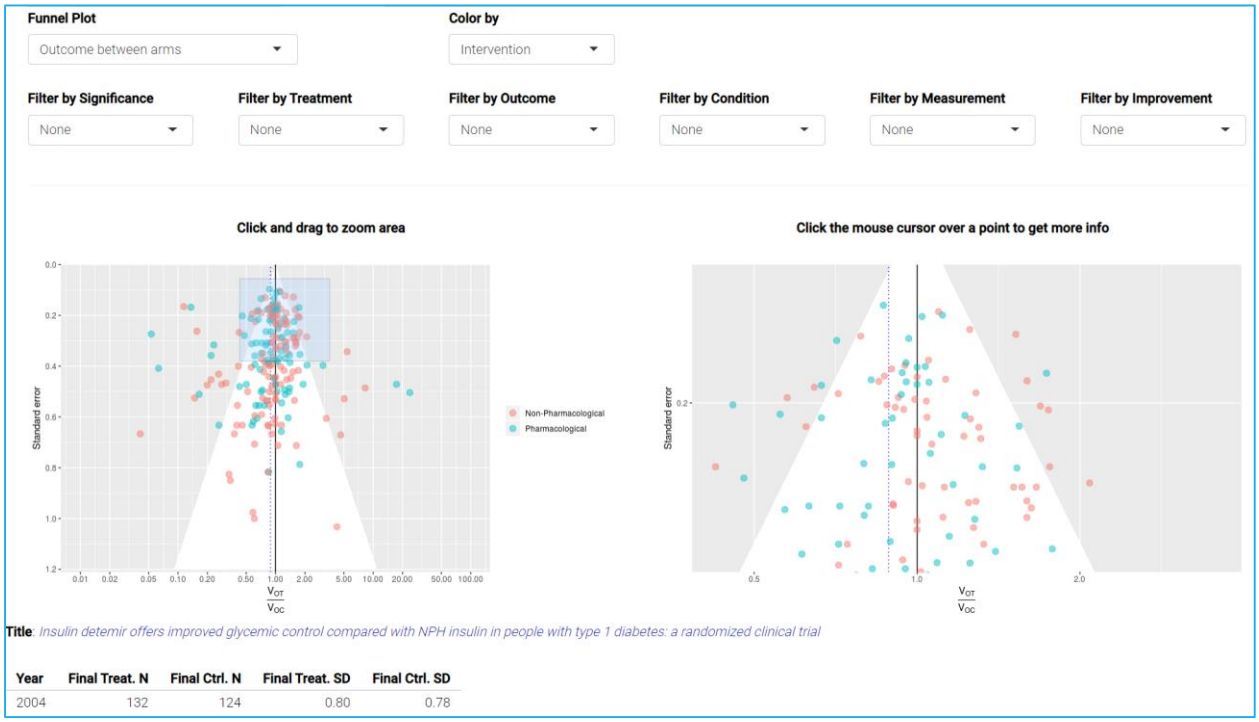
For an example of a completely ad hoc exploratory analysis not conducted in this work, but which can be carried out with the Shiny application, we can consider a comparison of variance discrepancies over time between both treatment groups. Figure 45 highlights a great correlation between these two measures, where the most discordant studies are located in the second and fourth quadrants and correspond to pharmacological interventions. Although these findings should be taken with caution because they are not pre-specified in advance, they demonstrate how this app can serve to generate new hypotheses for future research.

Figure 44. Screenshot of the scatterplot representing variance discrepancies over time in treated arm versus control arm.



Finally, as its name suggests, the *Funnel plots* tab (Figure 45), contains different funnel plot graphics. It brings together practically all the characteristics mentioned for the *Scatterplots* tab and can be useful for inquiring about studies with very specific characteristics in a remarkably easy way. For example, one can filter by pharmacological trials on chronic diseases.

Figure 45. Screenshot of the Shiny app's *Funnel plots* tab for building customized funnel plots.



6.2 R code

The R code used for the random effects model analysis of the published paper related to this work can be found on the *Zenodo* platform by following this link: https://doi.org/10.5281/zenodo.1239539[99]. The complete R code for conducting all the analyses and generating the figures shown in this work is available at the GitHub repository: https://github.com/jordicortes40/constant-effect-RCT

The GitHub platform is designed to share the code while researchers simultaneously develop it. This repository contains four main folders:

- ***data***. This contains the data collected from the 208 studies involved in this work.
- ***code***. The scripts with the R code for conducting the analyses.
- ***results_figures***. Some of the figures presented in this document
- ***results_tables***. Some of the main tables resulting from this work, in .txt format.

Sections 6.2.1 and 6.2.2 briefly explain the role of each R script contained in the *code* folder. Scripts were run using R version 4.0.1.

6.2.1 Main scripts

Each of the main R scripts can be run independently. For example, the script *summary_table.R* produces the table with information on the proportion of studies with a constant effect for each analysis, and it runs each of the previous scripts in order to provide this table. Below is a brief explanation of each script:

- ***read_data***. This installs and loads libraries. It also reads and cleans the data.
- ***descriptive***. Descriptive of the dataset containing information on our 208 randomized clinical trials.
- ***MA_main_analysis_rma***. Main analysis based on random effects models.
- ***SA_I_heuristic***. Sensitivity Analysis I. Heuristic procedure based on removing studies one by one until achieving negligible heterogeneity.
- ***SA_II_simulation***. Sensitivity Analysis II. Simulation study for determining the conditions under which the estimations of the random effects model will be obtained.
- ***SA_III_usual_tests***. Sensitivity Analysis III. Based on classic tests for comparing variances.
- ***SA_IV_mixture_distribution***. Sensitivity Analysis IV. Based on fitting a mixture distribution to the p-values derived from the previous Sensitivity Analysis III.
- ***summary_table***. Main results of all previous analyses.

6.2.2 Ancillary scripts

These scripts are called on by the main scripts and contain pieces of code that are placed in separate files to facilitate readability, due to their complexity. The names of these scripts along with short descriptions are listed below:

- ***functions***. Ancillary functions used in the main scripts.
- ***rma_models***. Fit of all models for the main analysis.
- ***subgroups***. Subgroup analyses.
- ***rma_models_reduced_data***. Fit of all models for Sensitivity Analysis II.
- ***SE_validation_BA***. Validation of standard errors for the between-arm comparison.
- ***SE_validation_OT***. Validation of standard errors for the comparison over time.

6.2.3 R libraries

Several R packages were used to carry out the analyses. They are listed below, together with their main purpose.

- ***data.table***. This library allows working with a class of objects (*data.table*) similar to data.frames, but with some properties that streamline data handling.
- ***weights***. This is for weighted t-tests, of which we performed some (not shown in this work) with the aim of comparing their results with those obtained for the main and sensitivity analyses.
- ***catspec***. For building good tables.
- ***alabama***. For optimizing the log-likelihood in Sensitivity Analysis IV.
- ***metaphor***. For fitting the random effects model.
- ***epitools***. For calculating the odds ratios.
- ***ggplot2***. For plotting good visualizations.
- ***ggpubr***. For arranging several *ggplots* in a single window with *ggarrange*.
- ***gridExtra***. For arranging several *ggplots* in a single window with *grid.arrange*.
- ***bootstrap***. For estimating the confidence interval for Kendall's τ statistic.

Funding

This research was partially funded by the Ministerio de Ciencia e Innovación (Spain) (PID2019-104830RB-I00), the Ministerio de Economía y Competitividad (Spain) (MTM2015-64465-C2-1-R (MINECO/FEDER)), the Ministerio de Ciencia e Innovación y Universidades (PID2019-104830RB-I00) and the Departament d'Economia i Coneixement de la Generalitat de Catalunya (Spain) (2017 SGR 622 (GRBIO)).

Appendix A: Sample size in studies with quantitative outcome: scenarios

Special attention should go to sample size in when designing RCTs. If the number of patients exceeds that which is required, the trial will be unnecessarily expensive and prolonged, due to exposing an unnecessary number of patients to a less efficacious drug. In contrast, if this number is too small, the study will lack power for detecting real treatment effects by statistical methods. In that case, the voluntary participation in any trial will have been in vain. A researcher should strike a balance between enrolling enough participants for detecting relevant differences and not having too many patients that resources are wasted unnecessarily. The consideration of whether the treatment effect is constant or not should be borne in mind. The usual sample size rationale specified by statisticians in RCTs includes a single parameter (without specifying its variability) that quantifies the desirable treatment effect. Consequently, there is a premise of a constant and unique effect that agrees with the clinical and legal interpretations that the benefit is the same for all the patients fulfilling the eligibility criteria – or at least similar enough to be considered homogeneous. Thus, as only a single effect is specified in most RCTs, the assumption that the ACE equals the single unit effect underlies the rationale behind the sample size calculation.

As an example, 10 protocols of RCTs published in the journal Trials in October 2017[100–109] (Table S - 1) and 10 RCTs published in *NEJM* in June/July 2017[110–119] (Table S - 2) provide a unique effect in the sample size calculation. Although many of these studies do not deal with quantitative response variables, the examples could be extrapolated to our context. For example, many of these trials refer to "*a reduction of some adverse event of X%*" in the treated group or to a "*change of X points in some measurement scale*"; but in no case do they mention whether a different effect should be considered among different patients.

Table S - 1. Ten RCT protocols published in October 2017 in the journal Trials. The last column (*Sample size explanation*) includes the paragraph in the statistical analysis section, which specified the sample size calculation. Specific sentences that denote a constant treatment effect are highlighted in bold.

Date	Title	Sample size explanation
Oct 23	*A multi-centre randomised trial to compare the effectiveness of geriatrician-led admission avoidance hospital at home versus inpatient admission*	Our proposed study effect estimate is based on a control group (inpatient admission) event rate at 12 months of 50% living in a residential setting, **with a 10% reduction** to 40% in the admission avoidance hospital at home group, equal to a relative risk of 0.8

Oct 30	*SCORE: Shared care of Colorectal cancer survivors: protocol for a randomised controlled trial*	Evidence of a substantial detriment associated with shared care as **measured by a reduction of 0.6** (or worse) on a patient-reported outcome measure would be detected with 80% power (at the 2.5% one-sided level of significance)
Oct 23	*Efficacy of inhaled HYdrogen on neurological outcome following Brain Ischemia During post-cardiac arrest care (HYBRID II trial): study protocol for a randomized controlled trial*	On the basis of published data […] and **assumed that the absolute risk reduction by HI is 15%**; that is, the favourable neurological outcome rate improves from 50% to 65% with HI. A sample size of 167 patients in each group will provide 80% power to detect a 15% change in the proportion of good neurological outcomes (CPCs of 1 and 2), from 50% to 65%,
Oct 25	*Neoadjuvant everolimus plus letrozole versus fluorouracil, epirubicin and cyclophosphamide for ER-positive, HER2-negative breast cancer: study protocol for a randomized pilot trial*	Because of the exploratory nature of this study, statistical power was not calculated to assess specific study outcomes […]. **This sample size calculation was developed in accordance with the recommendations by previous reports on sample size determinatio**n for pilot studies.
Oct 27	*Long-term Effects of high-doSe pitavaStatin on Diabetogenicity in comparison with atorvastatin in patients with Metabolic syndrome (LESS-DM): study protocol for a randomized controlled trial*	The sample size was calculated by **assuming an expected difference of 0.2%** in hemoglobin A1c change between the groups and a population variance of 0.5%, based on previous studies.
Oct 26	*Safety of tubal ligation by minilaparotomy provided by clinical officers versus assistant medical officers: study protocol for a noninferiority randomized controlled trial in Tanzanian women*	Assuming a 3% major AE rate in the control group (AMOs), we will demonstrate **noninferiority within the margin of 2%** at a one-sided significance level of $\alpha = 0.05$ and a power of 80% (calculated when AE rates in both arms are the same) with a sample size of 895 per arm (1790 women in total).
Oct 27	*Improving oxygen therapy for children and neonates in secondary hospitals in Nigeria: study protocol for a stepped-wedge cluster randomised trial*	Based on an alpha of 5%, we calculated that we would have approximately 80% power **to detect a 35% reduction** in pneumonia case fatality rate (6.1% to 4%), and 90% power **to detect a 20% reduction** in neonatal case fatality rate (8.2% to 6.6%)
Oct 27	*The effects of exercise on the quality of life of patients with breast cancer (the UMBRELLA Fit study): study protocol for a randomized controlled trial*	[…] we assume a 6-point increase in QoL in the control group and a 16-point increase in the intervention group […]. As a result, **we estimate a mean improvement of 13 points in the intervention group** ((70*16 + 30*6)/100 = 13) instead of 16 points **and a mean improvement of 6 points in the control group**.

Oct 30	*Targeting low- or high-normal Carbon dioxide, Oxygen, and Mean arterial pressure After Cardiac Arrest and REsuscitation: study protocol for a randomized pilot trial*	In a previous, small RCT, the use of 30% FiO2, compared with 100% FiO2, resulted in approximately 50% increase of NSE values at 48 h in the subset of patients treated with hypothermia. Assuming this previous finding, a study with 39 patients in each arm would have a power of 80%, with the significance set at 0.05, **to detect a 50% increase in NSE.**
Oct 24	*A randomised controlled trial assessing the severity and duration of depressive symptoms associated with a clinically significant response to sertraline versus placebo, in people presenting to primary care with depression (PANDA trial): study protocol for a randomised controlled trial*	The results from previous metaanalyses suggest that the effect size of SSRIs versus placebo is **about an 11% reduction in the Hamilton Rating Scale** for Depression (HAM-D) score […] Our best estimate of the minimal clinically important difference (MCID) from the PANDA cohort study is that this corresponds to a **14 percentage points (95% CI 10 to 17 percentage points) reduction in score on the PHQ-9** […] Therefore giving the power for effect sizes of 11 and 14 percentage points is reasonable and conservative in the light of the confidence limits and the previous results from the systematic review.

Table S - 2. Calculation of the sample size assuming a constant effect in 10 RCTs published in the NEJM during June/July 2017.

Date	Title	Sample size explanation
Jun 22	Efficacy of Recombinant Influenza Vaccine in Adults 50 Years of Age or Older	The sample size required to provide 80% power to show the noninferiority of relative vaccine efficacy was 4257 participants per treatment group, assuming influenza **attack rates of 1.6% in the RIV4 group and 2.0%** in the IIV4 group
Jun 22	Cluster-Randomized, Crossover Trial of Head Positioning in Acute Stroke	We estimated that at least 100 patients with acute ischemic stroke would need to be assigned to a head position at each hospital (i.e., 50 patients per intervention phase [or "period"]) across 120 centers (a total of 12,000 patients) for the study to have 90% power to detect a **16% or greater relative shift** in levels of disability outcome between intervention groups at 90 days in the ordinal logistic-regression analysis, at an alpha level of 0.05
Jun 22	Oral Glucocorticoid–Sparing Effect of Benralizumab in Severe Asthma	We estimated that 70 patients per group would be required for the trial to detect a difference in the primary end point between each benralizumab group and the placebo group with 86% power by means of a Wilcoxon rank-sum test with a two-sided level of 5%. Our estimation was based on simulations that used data from the Steroid Reduction with Mepolizumab Study (SIRIUS), which yielded a median percentage **reduction from baseline of 50%** in the glucocorticoid dose in the active-treatment group, as compared with no reduction in the placebo group.
Jun 22	First-Line Nivolumab in Stage IV or Recurrent Non–Small-Cell Lung Cancer	The sample-size estimation for the primary efficacy analysis population (patients with PDL1 expression level of ≥5%) was based on expected median progression-free survival of 7 months in the chemotherapy group and an **overall HR for disease progression or death of 0.71** favoring nivolumab.

Jun 29	Thyroid Hormone Therapy for Older Adults with Subclinical Hypothyroidism	[…] provided the trial with 80% power to detect a change with levothyroxine treatment (vs. placebo) of **3.0 points on the Hypothyroid Symptoms score and 4.1 points on the Tiredness score** with our revised maximum expected number of recruited participants of 750, with changes of **3.5 points and 4.9 points**, respectively, with our minimum expected number of 540 participants.
Jun 29	Trial of Electrical Direct-Current Therapy versus Escitalopram for Depression	The sample size was estimated on the basis of results from our previous study, the Sertraline versus Electrical Current Therapy for Treating Depression Clinical Study (SELECT-TDCS), with the use of an attrition rate of 13% and a noninferiority margin of 50% of the comparative efficacy of placebo versus escitalopram. The noninferiority margin was based on our hypothesis that tDCS would be associated with **at least 50% of the difference in efficacy** of escitalopram as compared with placebo
Jun 29	A Placebo-Controlled Trial of Antibiotics for Smaller Skin Abscesses	The trial was designed as a superiority trial with 80% power **to detect a 10-percentage-point absolute difference in cure rates** (e.g., 85% vs. 95%) among the three study groups in the population that could be evaluated.
Jul 6	Treatment of Endometriosis-Associated Pain with Elagolix, an Oral GnRH Antagonist	We calculated that the enrollment of 875 women in Elaris EM-I and 788 in Elaris EM-II (with the latter adjusted according to the withdrawal rate in Elaris EM-I) would provide a power of more than 90% to determine the two primary end points in each trial, **assuming response rates of 55% in each elagolix group and 29% in the placebo group**, at a two-sided alpha level of 0.025.
Jul 13	Adjuvant Pertuzumab and Trastuzumab in Early HER2-Positive Breast Cancer	The trial was designed to have 80% power to detect **a hazard ratio of 0.75** at a 5%, two-sided significance level
Jul 13	Follow-up of Prostatectomy versus Observation for Early Prostate Cancer	We revised our sample on the basis of estimates that 740 men enrolled over a period of 7 years, with an additional 8 years of follow-up, would provide 91% power **to detect a 25% relative reduction** in all-cause mortality, assuming a median survival of 10 years [*Paragraph from the original study*]

In a study that expects a non-constant treatment effect among patients, an additional parameter considering this variability would be necessary, since the distribution of the sample mean would be affected. The omission of such information indicates that such a treatment effect is expected to be homogeneous. An example of the wording in a sample size calculation that assumes certain variability would be: "*N patients are needed to detect a mean change of X in the primary endpoint, with a standard deviation of the treatment effect equal to Y*"

The following parts of this appendix are a modest attempt to exemplify the underlying problem of how variability in the intervention effect could condition sample size. No complete formal development is included, and we intend only to illustrate the fact that different casuistic errors occur when beginning with different situations with common features (for example, the Average

Treatment Effect [ATE] equals 1). Ignoring them leads to other errors of several magnitudes. We considered four possible scenarios:

A. **Constant treatment effect**. Equivalent to panel B of Figure 1.
B. **Constant treatment effect with interaction**. Equivalent to panels *C* and *E* of Figure 2.
C. **Constant treatment effect with interaction and different baseline characteristics**. This creates a new scenario with no equivalence, as seen in Figure 1 and Figure 2.
D. **Random treatment effect**. Equivalent to panel *D* in Figure 2.

Figure S - 1. Four hypothetical scenarios for calculating sample size.

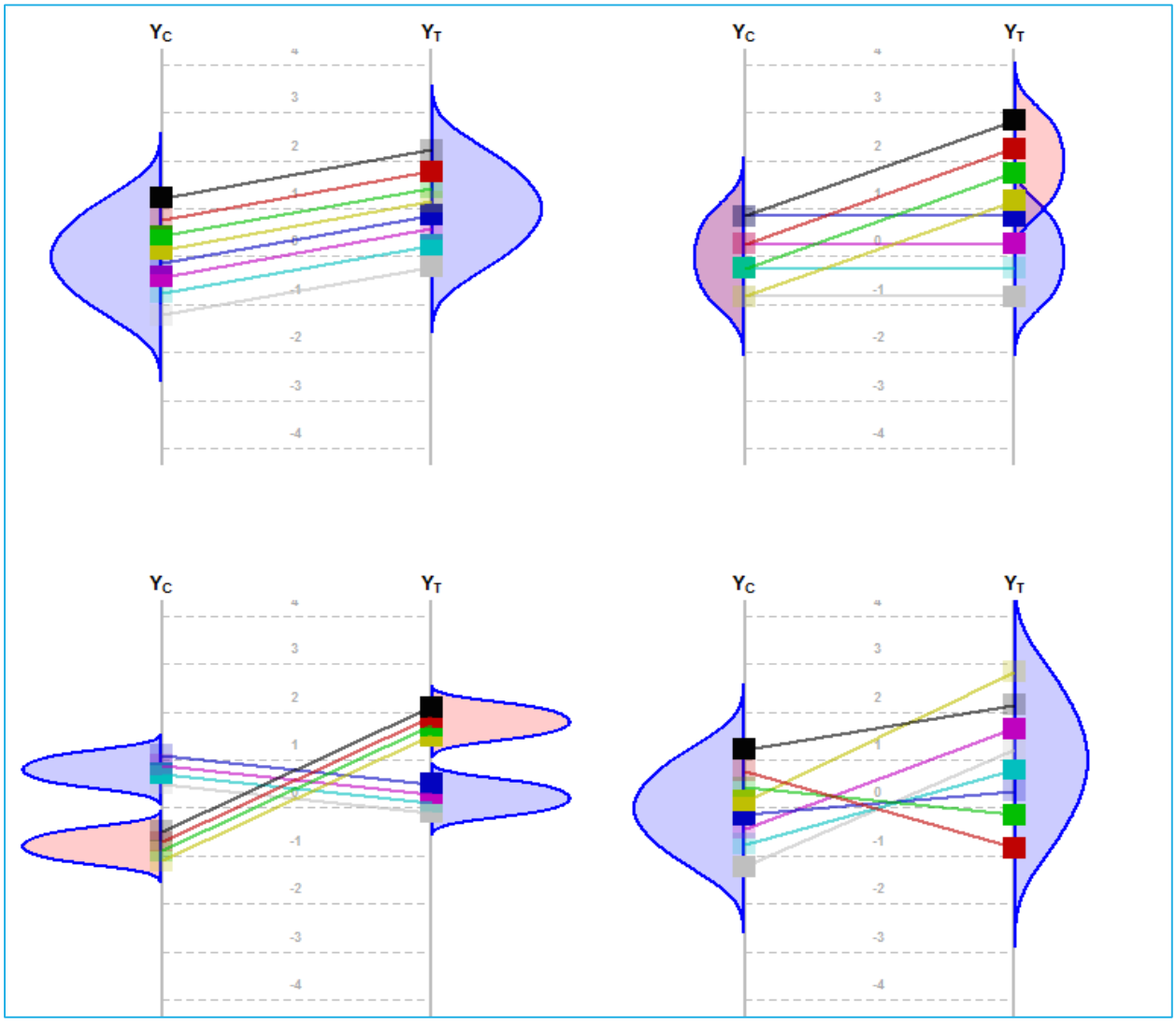


Figure S - 1 represents these situations. In all of them, we consider that the distribution of the potential outcome in the control arm (Y_C, at left in the four graphs) has an expected null value and an overall variance equal to 1. In addition, the Average Treatment effect (ATE) equals 1 in all scenarios. In the following pages, the required sample size for detecting a difference in means between arms (balanced) of one unit (population difference) with a probability of a one-side type

I error of α=0.05 and power of 80% was calculated for each of the scenarios. We will discuss what considerations should be taken into account in each case and what would occur by ignoring them.

Scenario A: Constant effect

The usual sample size formula for a mean comparison between two populations is[120]:

$$n = \frac{2\sigma^2 \cdot \left(Z_{1-\alpha/2} + Z_{1-\beta}\right)^2}{\Delta^2}$$

where *n* is the sample size in each group; σ^2 is the outcome variance; α and β are the probabilities of Type I and Type II errors, respectively; and Δ is the average (constant in this case) population treatment effect for which we want a specific power $(1-\beta)$ to detect it. As we will see, the key point in that formula is the determination of the **outcome variance (σ^2)**

In order to simplify the problem, we have chosen very simple values for the indicators of the distributions of the potential responses. The following expected values and variances for the outcome are assumed in the treated (T) and control (C) arms:

$$E[Y_C] = 0 \;;\; V[Y_C] = 1$$

$$E[Y_T] = 1 \;;\; V[Y_T] = 1$$

Under this setting, the derivation of the needed sample size for each group (n) does not present any complication and can be directly obtained by applying the above formula:

$$n = \frac{2 \cdot 1^2 \cdot (1.96 + 0.84)^2}{1^2} \approx 16$$

Scenario B: Effect Interaction with same baseline characteristics

In this setup, there are two subpopulations (e.g., men (M) and women (W)), each one with different treatment effects (Δ) of the intervention (constant treatment effect within the population). Again, we have chosen plain values that meet the constraint that the average treatment effect is 1:

$$\Delta^M = 0 \;;\; \Delta^W = 2$$

Assuming Normality for both subpopulations in the control (C) arm, the distributions of the outcome in the treated (T) arm are also Normal with the same variance but different expected values:

$$Y_C^M \sim N(\mu = 0, \sigma^2 = 1) \;\;;\;\; Y_T^M \sim N(\mu = 0, \sigma^2 = 1)$$

$$Y_C^W \sim N(\mu = 0, \sigma^2 = 1) \;\;;\;\; Y_T^W \sim N(\mu = 2, \sigma^2 = 1)$$

The overall variance of the outcome in the control arm is lower than in the experimental arm. That is, assuming the same proportion (0.5) of men and women, the overall variance in the treated arm doubled the one in the control arm:

Proof

$$E[Y_T] = 0.5 \cdot E[Y_T^M] + 0.5 \cdot E[Y_T^W] = 0.5 \cdot 0 + 0.5 \cdot 2 = 1$$

$$\left.\begin{array}{l} E[(Y_T^M)^2] = V[Y_T^M] + E[Y_T^M]^2 = 1 + 0^2 = 1 \\ E[(Y_T^W)^2] = V[Y_T^W] + E[Y_T^W]^2 = 1 + 2^2 = 5 \end{array}\right\} \rightarrow E[Y_T^2] = 0.5E[(Y_T^M)^2] + 0.5E[(Y_T^W)^2] = 0.5 \cdot 1 + 0.5 \cdot 5 = 3$$

$$V[Y_T] = E[Y_T^2] - E[Y_T]^2 = 3 - 1^2 = 2$$

Although the real variance doubled the expected one in the presence of this interaction, the variance of the sample mean random variable is not $2/n$, as anyone could expect if the Central Limit theorem is applied, but it is the same as in the constant effect scenario:

$$V[\bar{Y}_T] = V\left[\frac{Y_{T_1}^M + Y_{T_2}^M + \cdots + Y_{T_{n/2}}^M + Y_{T_1}^W + Y_{T_2}^W + \cdots + Y_{T_{n/2}}^W}{n}\right] = \frac{1}{n^2} \cdot [n/2 \cdot V[Y^M] + n/2 \cdot V[Y^W]]$$

$$= \frac{1}{n^2} \cdot \frac{n}{2} \cdot \left[V[Y^M] + V[Y^W]\right] = \frac{1}{2n} \cdot (1 + 1) = \frac{1}{n} = V[\bar{Y}_C]$$

Thus, provided that the averaged effect remains constant in respect to the constant effect scenario, the required sample size would be the same independently of the level of interaction and its magnitude.

Table S - 3 shows different situations according to researcher behavior. Basically, the researcher should make two decisions: 1) how much variability to consider in the sample size calculation and 2) whether to include an interaction term in the model for the analysis. From what we have seen, the correct choices would be 1) to choose the overall variability derived from the control group and 2) to consider the interaction term within the model as it actually exists. Therefore, the best decision (second row of the table completely in green) under this setting consists of 1) estimating the outcome variance from the control group (usual approach) and 2) conducting the analysis while taking into account the interaction. If the researcher takes into account the variability of the control group but does not include the interaction (first row), then the excess variability of the experimental group produced by the interaction will lead to a loss of power in the analysis. On the other hand, if trialists take into account the overall variances of both treatment arms (3rd and 4th row) based on previous studies, it can have several consequences, all of them undesirable. Not considering the interaction term (3rd row) would provide a well-powered study but with an unrealistic estimated treatment effect for any of the sex subgroups. Furthermore, the sample size would be larger than under the correct choice scenario. Finally, if the interaction in the model (4th row) is considered, the variability within each sex subgroup would be artificially inflated and lead to a study with a power that is higher than desired.

Table S - 3. Power analysis according to researcher behavior under the interaction setting. *T: Treatment S: Sex covariate*

Researcher behavior		Power
σ^2 used in sample size formula	**Model Analysis**	
Based on the overall variance of the control group	$Y \sim T + S$	Underpowered
	$Y \sim T * S$	Well powered
Based on the overall variances of the control and experimental groups	$Y \sim T + S$	Well powered
	$Y \sim T * S$	Overpowered

Scenario C: Effect Interaction with different baseline characteristics

We deal with the situation that the expected values and variances in both arms are known but the baseline characteristics differ between two subpopulations (e.g, M and W). Therefore, the outcome distribution in the control arm also differs between them. We will assume that the distributions of the potential outcomes in both strata of the control (C) group are:

$$Y_C^M \sim N(\mu = 0.8, \sigma^2 = 0.36) \;\; ; \;\; Y_C^W \sim N(\mu = -0.8, \sigma^2 = 0.36)$$

It is easy to demonstrate that the expected value and variance for the overall population is identical to the previous scenarios if men and women are distributed 50/50 in both groups.

$$E[Y_C] = 0 \;\; ; \;\; V[Y_C] = 1$$

Proof

$$E[Y_C] = 0.5 \cdot E[Y_C^M] + 0.5 \cdot E[Y_C^W] = 0.5 \cdot 0.8 + 0.5 \cdot (-0.8) = 0$$

$$\left.\begin{matrix} E[(Y_C^M)^2] = V[Y_C^M] + E[Y_C^M]^2 = 0.36 + 0.8^2 = 1 \\ E[(Y_C^W)^2] = V[Y_C^W] + E[Y_C^W]^2 = 0.36 + 0.8^2 = 1 \end{matrix}\right\} \rightarrow E[Y_C^2] = 0.5(E[(Y_C^M)^2] + E[(Y_C^W)^2]) = 0.5(1 + 1) = 1$$

$$V[Y_C] = E[Y_C^2] - E[Y_C]^2 = 1 - 0^2 = 1$$

The values for the distributions in the control group and the effects are chosen so that the averaged effect would be 1 and the outcome variance remains constant $(V[Y_T] = 1)$ in the overall experimental arm. Keeping this in mind, the following constant effects for each subpopulation are considered:

$$\Delta^M = -0.6 \;\; ; \;\; \Delta^W = 2.6$$

Thus, the potential outcomes in the Treated (T) group are:

$$Y_T^M \sim N(\mu = 0.2, \sigma^2 = 0.36) \;\; ; \;\; Y_T^W \sim N(\mu = 1.8, \sigma^2 = 0.36)$$

If someone tries to find statistical differences in means between the treated and control arms without taking into account the two subpopulations, the sample size required should be based on

the variance of the mixture distributions (i.e., $\sigma^2 = 1$). Then, the sample size (n) required for each group in this scenario should be based on the sample mean random variables and the differences between them with following distributions:

$$\bar{Y}_T \sim N(\mu = 1, \sigma^2 = 1/n)$$
$$\bar{Y}_C \sim N(\mu = 1, \sigma^2 = 1/n)$$

However, the real variance of the sample mean random variable in the control arm is derived as follows:

$$V(\bar{Y}_C) = V\left[\frac{Y_{C_1}^M + Y_{C_2}^M + \cdots + Y_{C_{n/2}}^M + Y_{C_1}^W + Y_{C_2}^W + \cdots + Y_{C_{n/2}}^W}{n}\right] = \frac{1}{n^2} \cdot [n/2 \cdot V[Y^M] + n/2 \cdot V[Y^W]]$$

$$= \frac{1}{n^2} \cdot n/2 \cdot [V[Y^M] + V[Y^W]] = \frac{1}{2n} \cdot 2 \cdot 0.36 = \frac{0.36}{n}$$

In a similar way, it can be seen that the variance for the sample mean random variable in the treated arm is the same, and, therefore the variance of the mean difference and its distribution is:

$$V[\bar{Y}_T - \bar{Y}_C] = \frac{2 \cdot 0.36}{n}$$

$$\bar{Y}_T - \bar{Y}_C \sim N(\mu = 1, \sigma^2 = 0.72/n)$$

If the researcher is able to anticipate this interaction, the required sample size would be:

$$n = \frac{2 \cdot \sqrt{0.72} \cdot (1.96 + 0.84)^2}{1^2} \approx 14$$

Table S - 4 shows the consequences for power and sample size based on researcher's decisions. In this setting, the usual approach represented in the second row of the table entails a non-desirable overpowered design with a larger sample size than the one actually needed. The variability of the statistic of the sample mean required in the third and fourth rows of the table should be calculated by taking into account the a priori subgroups involved in the interaction.

Table S - 4. Power analysis according to researcher behavior under interaction setting according to baseline characteristics. *T: Treatment effect; S: Sex covariate.*

Researcher behavior		**Power**
σ^2 used in sample size formula	**Analysis**	
Based on the overall variance of the control group	$Y \sim T + S$	Well powered
	$Y \sim T * S$	Overpowered
Based on the real variance of the sample mean statistic of the control group	$Y \sim T + S$	Underpowered
	$Y \sim T * S$	Well powered

Scenario D: Random effect

We deal here with the case that there are no subpopulations that differ from each other, that the treatment effect (Δ) is specific for each individual, and that Δ is distributed according to a Normal distribution with the following expected value and variance:

$$Y_C \sim N(\mu = 0, \sigma^2 = 1) \ ; \ \Delta \sim N(\mu = 1, \sigma^2 = 1)$$

The potential outcome in the Treated (T) group is directly deducted: $Y_T \sim N(\mu = 1, \sigma^2 = 2)$

Both the outcome variance and the sample mean variance in the treated group are inflated, and this should be contemplated in the sample size. With unequal variances between arms, the sample size is the minimum *n* that meets the inequality[121]:

$$n_C \geq \frac{\left(\frac{s_T^2}{A} + s_C^2\right) \cdot \left(t_{df,1-\frac{\alpha}{2}} + t_{df,1-\beta}\right)^2}{\Delta^2}$$

where A is the ratio between sample sizes in both groups (n_T/n_C) and is the optimal sample size ratio equal to the ratio of standard deviations. Therefore:

$$n_C \geq \frac{\left(\frac{s_T^2}{\frac{s_T}{s_C}} + s_2^2\right) \cdot \left(t_{df,1-\frac{\alpha}{2}} + t_{df,1-\beta}\right)^2}{\Delta^2} = \frac{\left(\frac{s_T}{s_C} + s_C^2\right) \cdot \left(t_{df,1-\frac{\alpha}{2}} + t_{df,1-\beta}\right)^2}{\Delta^2}$$

$$= \frac{(2+1) \cdot \left(t_{df,1-\frac{\alpha}{2}} + t_{df,1-\beta}\right)^2}{1^2} = 3 \cdot \left(t_{df,1-\frac{\alpha}{2}} + t_{df,1-\beta}\right)^2$$

The required sample size is $n_C = 25$ and $n_T = 50$. Table S - 5 contains the results of several strategies in the design and analysis phases of the trial. Again, the most common decision (second row of Table S - 5) entails a non-desirable underpowered design with a sample size lower than the one that is actually needed.

Table S - 5. Power analysis according to researcher behavior in a random treatment effect setting. *T: Treatment effect.*

Researcher decision		**Power**
σ^2 used in sample size formula	**Analysis**	
Based on the variance of the control group	$Y \sim T \quad \sigma_T = 0 \ (fixed\ effect)$	Under or overpowered
	$Y \sim T \quad \sigma_T > 0 \ (random\ effect)$	Underpowered
Based on the variance of the control and the random effect	$Y \sim T \quad \sigma_T = 0 \ (fixed\ effect)$	Overpowered
	$Y \sim T \quad \sigma_T > 0 \ (random\ effect)$	Well powered

Sample size calculations in all scenarios

Table S - 6 summarizes all the abovementioned scenarios. The purpose of these illustrations through numerical examples was not to make a formal study of the entire spectrum of situations that may arise, but to illustrate a non-trivial issue: the presence of a non-constant treatment effect has relevant implications for the sample size calculation and should not be neglected. The fact that RCTs are looking for differences in the population means does not remove this problem.

Table S - 6. Summary results of power analysis under each scenario.

Scenario	Sample size				Power assuming constant effect
	Correct calculation		Calculation assuming constant effect		
	Control group	Treated group	Control group	Treated group	
Constant effect	16	16	16	16	Well powered
Interaction	16	16			Well powered
Interaction with different characteristics	14	14			Overpowered
Random effect	25	50			Underpowered

Appendix B: Reporting uncertainty

Some collected studies did not report the standard deviation (S) or the variance (S^2) of the primary endpoint. Instead, they reported standard errors or confidence intervals. This appendix explains how the standard deviations were obtained in these cases.

Studies that report standard errors

The estimated standard error (SE) of the mean is defined as:

$$\widehat{SE} = \frac{S}{\sqrt{n}} \rightarrow S = \widehat{SE} \cdot \sqrt{n}$$

Example: In the study by Yancy et al.[122], they report the SE for the primary endpoint (pH) at baseline (week 0) and at the end of the follow-up (week 24).

Table S - 7 summarizes how we transform the standard errors in standard deviations using the formula above.

Table S - 7. Example of calculating S from SE.

Moment	Group	N	SE	S
Baseline (week 0)	Control (LFD)	12	0.006	$S = 0.006 \cdot \sqrt{12} = 0.02$
	Intervention (LCKD)	27	0.004	$S = 0.004 \cdot \sqrt{27} = 0.02$
End of follow-up (week 24)	Control (LFD)	12	0.009	$S = 0.009 \cdot \sqrt{12} = 0.03$
	Intervention (LCKD)	27	0.006	$S = 0.006 \cdot \sqrt{27} = 0.03$

Studies that report confidence intervals

The (1-α)% confidence interval of the population mean from one sample is defined as:

$$IC(1-\alpha, \mu) = \bar{x} \mp t_{0.975,n-1} \frac{S}{\sqrt{n}} = \bar{x} \mp \delta \rightarrow S = \frac{\delta \cdot \sqrt{n}}{t_{0.975,n-1}}$$

Example: In the study by McManus[123], they report the 95% confidence interval for the primary endpoint (systolic blood pressure) at baseline (month 0) and at the end of the follow-up (month 12). Table S - 8 summarizes how we transform the 95% CIs in standard deviations using the formula above.

Table S - 8. Example of calculating S from 95% confidence interval.

Moment	Group	N	95% CI	S
Baseline (month 0)	Control (self-management)	264	(150.3 to 153.3)	$S = \frac{1.5 \cdot \sqrt{264}}{1.97} = 12.38$
	Intervention (tele-monitoring)	263	(150.6 to 153.6)	$S = \frac{1.5 \cdot \sqrt{263}}{1.97} = 12.35$
End of follow-up (month 12)	Control (self-management)	246	(138.0 to 142.2)	$S = \frac{2.1 \cdot \sqrt{246}}{1.97} = 16.72$
	Intervention (tele-monitoring)	234	(132.6 to 137.1)	$S = \frac{2.25 \cdot \sqrt{234}}{1.97} = 17.47$

Appendix C: Medical specialties

Table S - 9 shows the frequency and percentage of the medical fields pertaining to the collected articles, according to WoS.

Table S - 9. Medical fields according to WoS criteria for the collected studies.

Medical field	N	%	Medical field	N	%
General & Internal Medicine	31	14.9	Sport Sciences	5	2.4
Nutrition & Dietetics	21	10.1	Surgery	5	2.4
Endocrinology & Metabolism	19	9.1	Anesthesiology	3	1.4
Cardiovascular System & Cardiology	16	7.7	Gastroenterology & Hepatology	3	1.4
Dentistry, Oral Surgery & Medicine	14	6.7	Integrative & Complementary Medicine	3	1.4
Neurosciences & Neurology	14	6.7	Nursing	3	1.4
Pharmacology & Pharmacy	13	6.3	Public, Environmental & Occupational Health	3	1.4
Psychiatry	12	5.8	Reproductive Biology	3	1.4
Ophthalmology	11	5.3	Dermatology	2	1.0
Health Care Sciences & Services	9	4.3	Science & Technology - Other Topics	2	1.0
Orthopedics	9	4.3	Allergy	1	0.5
Pediatrics	9	4.3	Biomedical Social Sciences	1	0.5
Obstetrics & Gynecology	7	3.4	Engineering	1	0.5
Psychology	7	3.4	Immunology	1	0.5
Rehabilitation	7	3.4	Mathematics	1	0.5
Respiratory System	6	2.9	Microbiology	1	0.5
Rheumatology	6	2.9	Otorhinolaryngology	1	0.5
Urology & Nephrology	6	2.9	Physiology	1	0.5
Geriatrics & Gerontology	5	2.4	Research & Experimental Medicine	1	0.5
Oncology	5	2.4	Transplantation	1	0.5

Prior to this classification, Erik Cobo (EC), José Antonio González (JAG) (the supervisors of this work), and Jordi Cortés (JC) classified these articles. In a first round, EC and JC classified while blinded to each other; they matched 127 out of 208 articles to a specific medical field. In the remaining 81 articles, JAG decided between two discordant medical categories. The concordance between the classifications that we and WoS ultimately assigned is shown in Table S - 10.

Table S - 10. Concordance of medical field between researchers' allocation and WoS classification.

Independent evaluators (JAG, EC, JC)	WoS: Cardiovascular System & Cardiology	Dermatology	Endocrinology & Metabolism	Gastroenterology & Hepatology	General & Internal Medicine	Obstetrics & Gynecology	Urology & Nephrology	Dentistry, Oral Surgery & Medicine	Oncology	Ophthalmology	Pediatrics	Respiratory System	Neurosciences & Neurology	Psychiatry	Psychology	Rheumatology	Rehabilitation	Geriatrics & Gerontology	Health Care Sciences & Services	Integrative & Complementary Med.	Nutrition & Dietetics	Orthopedics	Pharmacology & Pharmacy	Research & Experimental Medicine	Science & Technology-Other Topics	Sport Sciences	Surgery	Subtotal	More than one category	Without category	Grand Total
Cardiovascular	15		2		6	2	1										1				7		3		1			38	3	1	***42***
Dermatology		2			2											1			1									6			***6***
Endocrinology			14		4	1					1									1	5						1	27	1	1	***29***
Gastroenterology				2	1																1							4			***4***
General medicine																			1		1							2	3		***5***
Gynecology					1																							1	4		***5***
Hematology																					1							1			***1***
Nephrology					1		3																					4	2		***6***
Odontology								13																				13	1		***14***
Oncology									1						1				1									3	4		***7***
Ophthalmology					1					10													2					13	1		***14***
Otolaryngology																												0	1		***1***
Pediatric					1						1																	2	3		***5***
Pneumology					1						1	2																4	5		***9***
Psychology/ Psychiatry/Neurology			1		2						1		5	7	1			3			1							21	8	1	***30***
Rheumatology/ Traumatology			1		4											3		1		1		1	1	1	1	2		16	13	1	***30***
Grand Total	***15***	***2***	***18***	***2***	***24***	***3***	***4***	***13***	***1***	***10***	***4***	***2***	***5***	***7***	***2***	***4***	***1***	***4***	***3***	***2***	***16***	***1***	***6***	***1***	***2***	***2***	***1***	***155***	***48***	***4***	***208***

Appendix D: Concordance between both comparisons

We wanted to analyze the concordance between both criteria to detect an increase in the outcome variance of the comparisons between arms (treated vs. control) and over time in the treated group (outcome vs. baseline). The hypothesis behind this analysis was that a true increase in the variability of the treated group's outcome variance should be detected in both comparisons. Concordant studies would be more suspect of having a non-constant effect, and discordant studies deserve a deeper analysis to explore the existence of some methodological impurities, which will be commented on Section 4.1. However, this latest and more qualitative scrutiny was outside the scope of this work and, thus, we did not formally address it. Figure S - 2 combines the information from both comparisons. Although most of the studies were distributed in a random way around the point (1,1) of perfect concordance, the Pearson correlation coefficient is 0.74. However, Spearman's correlation coefficient is more appropriate due to the presence of outliers, and it was equal to 0.36, which can be considered a mild correlation. Since the regression line passes through the origin of the logarithmic scale (1,1), the usual correlation coefficients can serve the purpose of measuring agreement, provided that the variances of the two involved measures have similar magnitude.

Figure S - 2. Variance discrepancy between arms vs. over time. Colors indicate the availability of the information from the over-time model. Point size is proportional to the precision of each study. The dashed blue line is the linear regression fitted to all the points.

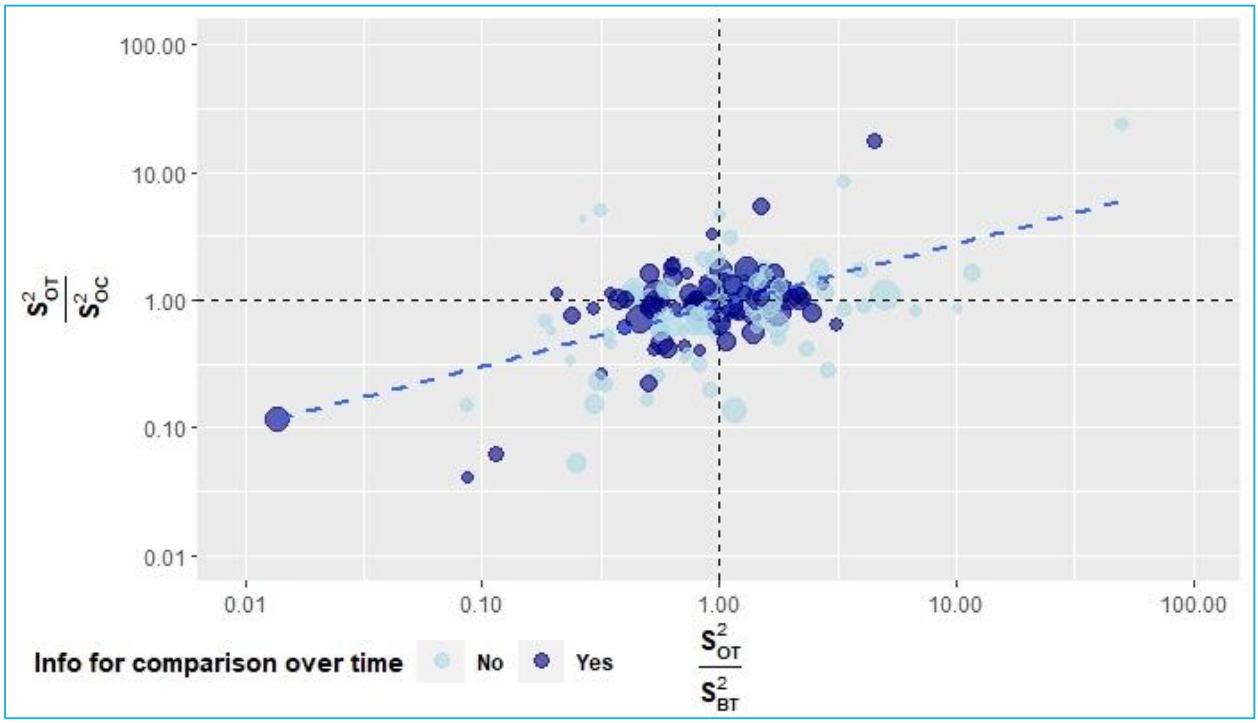


Table S - 11 shows a summary of the concordance between both comparisons dichotomized according to their statistical significance in the random effects models.

Table S - 11. Concordance between both comparisons.

Comparison over time		Significant between-arm comparison		Total
Complete reporting	Significant	No	Yes	
No	-	87	26	113
Yes	No	52	5	57
	Yes	28	10	38
Total		167	41	208

The concordance between both comparisons may be assessed only in those studies (n = 95) where the significance of the results is available for both comparisons. A positive association was found for these studies, with variance discrepancies between both comparisons (OR=3.61, 95% CI from 1.14 to 12.9). Among the 10 studies for which both comparisons showed statistically significant differences in variances, two and seven had respectively greater and lower variances in the treated arm at the end of the follow-up. Surprisingly, one single study presented significantly greater variance at the end of the study in the over-time comparison, but this was lower than the variance in the control arm at the end. In other words, despite the fact that the treated patients increased the outcome variability over time, it increased even more in the reference group.

These results should be cautiously interpreted, because the significance is highly influenced by the trial's sample size, which was determined to achieve a specific power in order to compare outcome means (not variances).

We will now explore more detailed information regarding the concordance in order to formulate tentative interpretations of each kind of (dis)concordance. Table S - 12 shows such interpretations for the specific situations given in Figure S - 2. It should be borne in mind that these are plausible interpretations that can help a researcher understand their results; but in no case have they been verified. Concordance may be evaluated and examined only in those studies (n = 95) where the significance of the results is available for both comparisons.

Table S - 12. Interpretive summary table for the concordance between both comparisons.

Between Arms	Over Time	*n/N (%)*	Interpretation
NS	**NI**	*87/208 (42%)*	**Studies without information on the concordance between the two variance discrepancies**. No conclusion available.
	NS	*53/95 (56%)*	**Ineffective treatments**. Generally, not modifying the centrality parameter implies that the dispersion parameter is not affected either.
	S	*28/95 (29%)*	**Homogeneous patients at the start of the study**. In the 13 studies with this pattern, the increasing variance over time in the experimental group did not trigger differences in variability between arms. This indicates that the included sample was too homogeneous in the trial and the outcome would have increased its variability in a natural way. **Mild diseases**. In contrast, these are the 15 studies with decreasing variability in the experimental group over time and without differences between groups at the end of the study. They probably deal with a condition under study that naturally fades over time.
S	**NI**	*26/208 (12%)*	**Studies without information on the concordance between the two variance discrepancies**. No conclusion available.
	NS	*4/95 (4%)*	**Control group not properly followed**. Here, the experimental group has no increase in variability over time but there are differences at the end of the study, which could be due to poor monitoring of patients in the control group.
	S	*10/95 (11%)*	**The treatment modifies the variability**. There are two and seven studies with, respectively, increasing and decreasing variance in the experimental arm in both comparisons. In these studies, the change in variability is directly due to treatment, as patients in the experimental group increase or decrease their variability over time. This change does not occur equally in the reference group and thereby causes a difference in the variances at the end of the study between treatment arms. There is a single study with increasing variability over time and decreasing variability between-arms, both with no plausible explanation.

S: significant; NS: Not significant; NI: No information

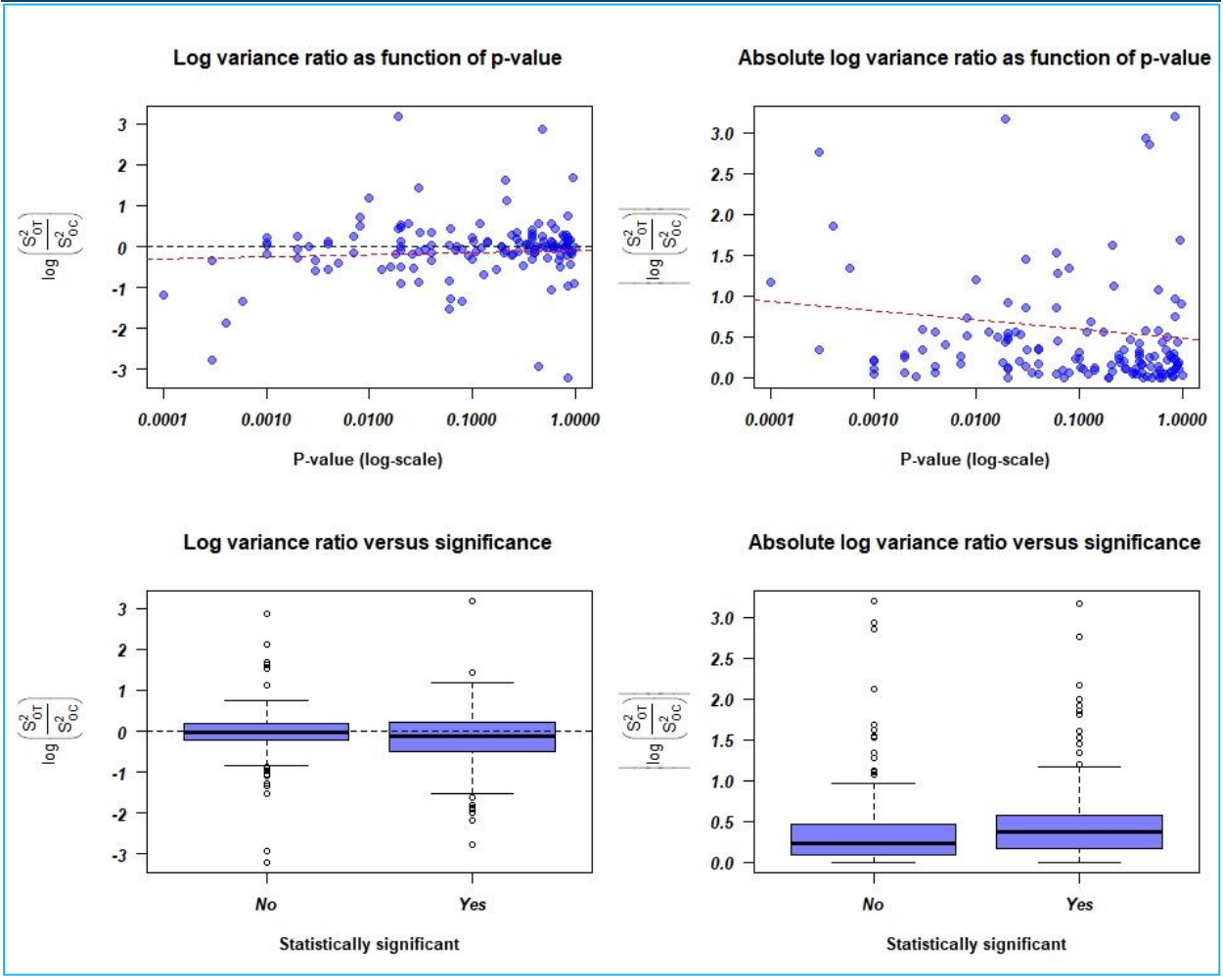
Appendix E: Ancillary analyses

Subgroup analysis according to the significance

One subgroup analysis was performed according to the significance of the main analysis of each collected trial. Thus, the variable significance was dichotomized. We wanted to explore if some relationship existed between the variance discrepancy between arms and the significance of the main analysis of each study, but by measuring it through the uncategorized p-values.

Figure S - 3 shows the between-arm variance as functions of the main study result's p-value and of statistical significance.

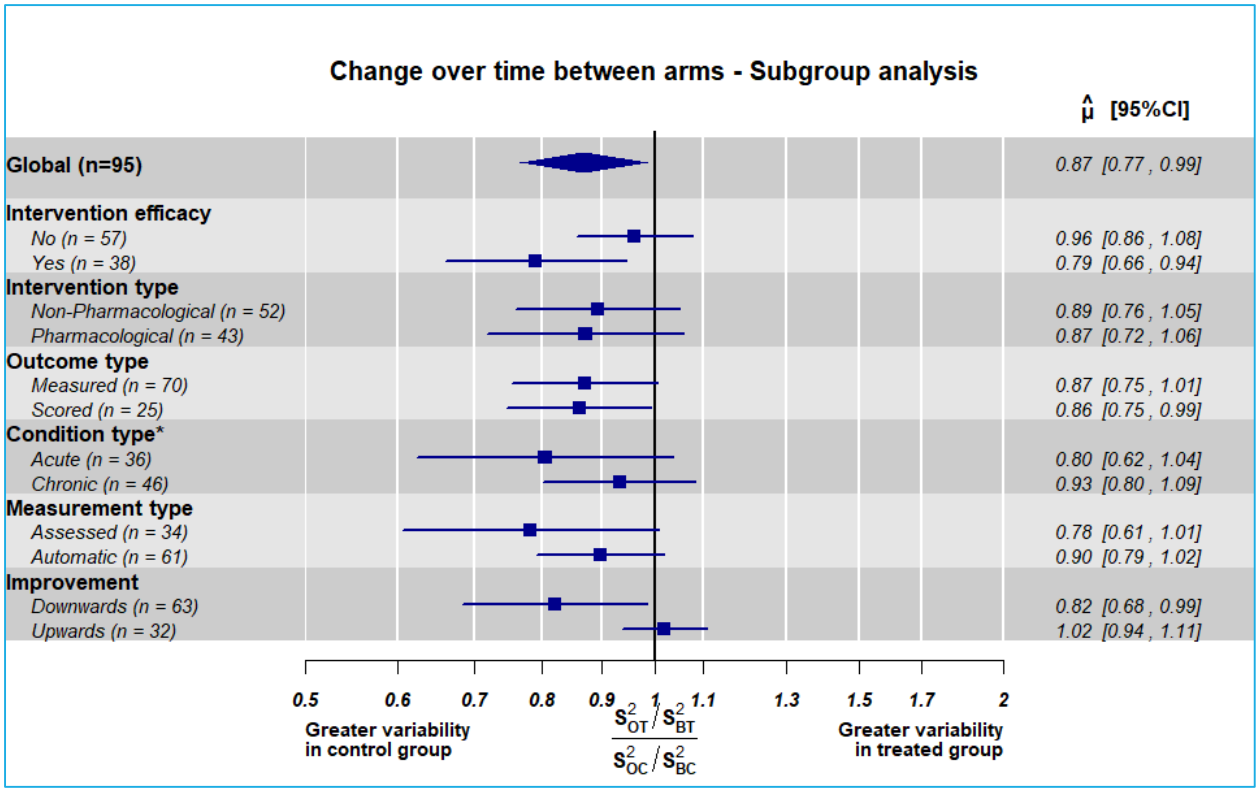
Figure S - 3. Relationship between statistically significant differences and the variance discrepancy in the between-arm comparison.



Subgroup analysis for change in the comparison over time between arms

Figure S - 4 shows the difference in the change in variability (from baseline to the end of the study) between arms. The results were quite similar to those observed in the subgroup analysis regarding the comparison over time, and they do not yield any conclusions beyond to those already mentioned in Section 3.2.4.

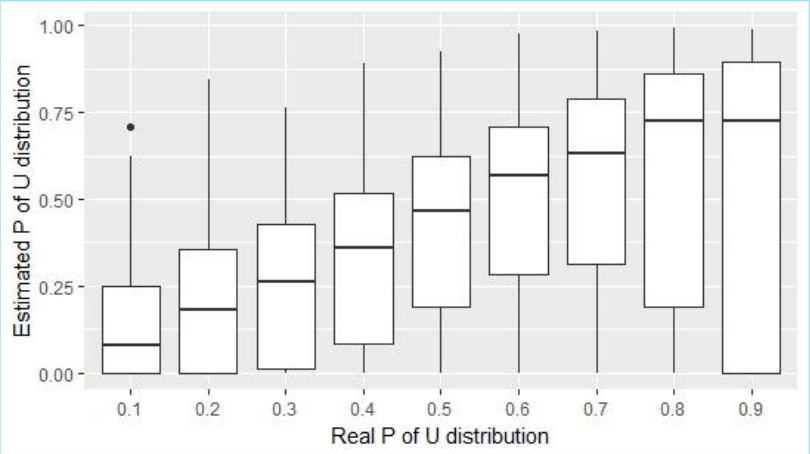
Figure S - 4. For whole data and each subgroup, point estimate and the 95% confidence intervals for the estimated outcome (O) variance ratio between Treated (T) and Controls (C) (Model 7 of Table 11). X-axis is in log scale.*13 studies were performed with healthy participants.



Simulation study to check the distribution mixture approach

We conducted a simple simulation to check the validity of the results. In the sensitivity analysis based on the mixture distribution, we included each p-value in all the components of the likelihood function. In other words, each p-value contributes in the same way to all distributions. An alternative approach would be to assign each p-value to a single distribution. We performed the simulation with a specific proportion (0.1, 0.2…, 0.9) of p-values coming from a uniform distribution and with a complementary proportion of p-values coming from a beta distribution with both parameters (shape and scale) equal to 0.5. For each proportion, we conducted 1000 runs and collected the estimates of the proportion ($P = \hat{\pi}_0$) of studies under the null hypothesis. The distributions of the estimated proportions under each scenario are shown in Figure S - 5. Each boxplot represents the estimated proportions of p-values coming from a uniform distribution according to the actual proportion simulated, as indicated by the label on the horizontal axis. In all cases, estimates were slightly biased towards lower proportions, especially those with a real large proportion. However, the most worrisome problem is that these estimates present a large amount of variability. For example, under the scenario with an actual proportion (P) of studies with a constant effect equal to 0.90, the estimates ranged from 0 to 1, almost uniformly.

Figure S - 5. Distribution of the estimates according to the real proportion of p-values coming from a uniform distribution.



In the paper by Pounds et al.[66], the authors did not have to deal with this problem because their distribution was unimodal with a single peak of around 0, unlike our data. Furthermore, the same paper warns that it is unfeasible to obtain a precise estimate of the proportion of p-values coming from a specific distribution, although it is nevertheless possible to find an upper bound for this proportion. This could be partially explained by our simulation results: the upper bound for the estimates increases as the real proportion of homoscedastic studies increases; but the lower bound remains almost constant, independently of the actual proportion.

Appendix F: References of collected studies

References 2004

1. Aller, Rocio, Daniel Antonio De Luis, Olatz Izaola, Fernando La Calle, Lourdes Del Olmo, Luis Fernandez, Teresa Arranz, and J. M. Gonzalez Hernandez. 2004. "Effect of soluble fiber intake in lipid and glucose leves in healthy subjects: A randomized clinical trial." Diabetes Research and Clinical Practice 65 (1): 7-11.
2. Bleie, Øyvind, Helga Refsum, Per Magne Ueland, Stein Emil Vollset, Anne Berit Guttormsen, Ebba Nexo, Jørn Schneede, Jan Erik Nordrehaug, and Ottar Nygård. 2004. "Changes in basal and postmethionine load concentrations of total homocysteine and cystathionine after B vitamin intervention." American Journal of Clinical Nutrition 80 (3): 641-648.
3. Booth, Sarah L., Jennifer M. Sacheck, Ronenn Roubenoff, Gerard E. Dallal, Jeffrey B. Blumberg, Ines Golly, and Koichiro Hamada. 2004. "Effect of vitamin E supplementation on vitamin K status in adults with normal coagulation status." American Journal of Clinical Nutrition 80 (1): 143-148.
4. Breneman, Debra, Ronald Savin, Christine VandePol, George Vamvakias, Sharon Levy, and James Leyden. 2004. "Double-blind, randomized, vehicle-controlled clinical trial of once-daily benzoyl peroxide/clindamycin topical gel in the treatment of patients with moderate to severe rosacea." International Journal of Dermatology 43 (5): 381-387.
5. Caballero, A. Enrique, Adriano Delgado, Carlos A. Aguilar-Salinas, Alberto Naranjo Herrera, Jose Luis Castillo, Tatiana Cabrera, Francisco J. Gomez-Perez, and Juan A. Rull. 2004. "The differential effects of metformin on markers of endothelial activation and inflammation in subjects with impaired glucose tolerance: A placebo-controlled, randomized clinical trial." Journal of Clinical Endocrinology and Metabolism 89 (8): 3943-3948.
6. Charles, C. H., K. M. Mostler, L. L. Bartels, and S. M. Mankodi. 2004. "Comparative antiplaque and antigingivitis effectiveness of a chlorhexidine and an essential oil mouthrinse: 6-month clinical trial." Journal of Clinical Periodontology 31 (10): 878-884.
7. Chow, Song-Nan, Ting-Chen Chang, Yih-Ron Lien, Ming Chen, Shao-Pei Cheng, and Ruey-Jien Chen. 2004. "Effect of conjugated equine estrogen in combination with two different progestogens on the risk factors of coronary heart disease in postmenopausal Chinese women in Taiwan: A randomized one-year study." Acta Obstetricia et Gynecologica Scandinavica 83 (7): 661-666.
8. Dimeglio, Linda A., Tina M. Pottorff, Sheryl R. Boyd, Lisa France, Naomi Fineberg, and Erica A. Eugster. 2004. "A randomized, controlled study of insulin pump therapy in diabetic preschoolers." Journal of Pediatrics 145 (3): 380-384.
9. Dodson, Thomas B. 2004. "Management of mandibular third molar extraction sites to prevent periodontal defects." Journal of Oral and Maxillofacial Surgery 62 (10): 1213-1224.
10. Eickholz, Peter, Diana-Maria Krigar, Bernadette Pretzl, Harald Steinbrenner, Christof Dörfer, and Ti-Sun Kim. 2004. "Guided Tissue Regeneration with Bioabsorbable Barriers. II. Long-Term Results in Infrabony Defects." Journal of Periodontology 75 (7): 957-965.

11. Fitzgthhon, Marian L., Susan M. Gapstur, and Sara J. Knight. 2004. "Results of mujeres felices por ser saludables: a dietary/breast heafth randomized chnicail trial for latino women." Annals of Behavioral Medicine 28 (2): 95-104.
12. Fox, Norma Lynn, Byron J. Hoogwerf, Susan Czajkowski, Ruth Lindquist, Gilles Dupuis, J. Alan Herd, Lucien Campeau, Ann Hickey, Franca B. Barton, and Michael L. Terrin. 2004. "Quality of life after coronary artery bypass graft: Results from the POST CABG trial." Chest (The American College of Chest Physicians) 126 (2): 487-495.
13. Francetti, Luca, Massimo Del Fabbro, Matteo Basso, Tiziano Testori, and Roberto Weinstein. 2004. "Enamel matrix proteins in the treatment of intra-bony defects. A prospective 24-month clinical trial." Journal of clinical periodontology 31 (1): 52-9.
14. Francis, B. A., L. T. Du, S. Berke, M. Ehrenhaus, D. S. Minckler, and Cosopt Study Group. 2004. "Comparing the fixed combination dorzolamide-timolol (Cosopt) to concomitant administration of 2% dorzolamide (Trusopt) and 0.5% timolol -- a randomized controlled trial and a replacement study." Journal of clinical pharmacy and therapeutics 29 (4): 375-80.
15. Halterman, Jill S, Peter G Szilagyi, H Lorrie Yoos, Kelly M Conn, Jeffrey M Kaczorowski, Robert J Holzhauer, Sherri C Lauver, Tia L Neely, Patrick M Callahan, and Kenneth M McConnochie. 2004. "Benefits of a school-based asthma treatment program in the absence of secondhand smoke exposure: results of a randomized clinical trial." Archives of pediatrics and adolescent medicine 158 (5): 460-7.
16. Hawkins, Barbara S., Paivi H Miskala, Eric B Bass, Neil M Bressler, Ashley L Childs, Carol M Mangione, Marta J Marsh, and Submacular Surgery Trials Research Group. 2004. "Surgical removal vs observation for subfoveal choroidal neovascularization, either associated with the ocular histoplasmosis syndrome or idiopathic: II. Quality-of-life findings from a randomized clinical trial: SST Group H Trial: SST Report No. 10." Archives of ophthalmology 122 (11): 1616-28.
17. Home, Philip, Paul Bartley, David Russell-Jones, Hélène Hanaire-Broutin, Jan Evert Heeg, Pascale Abrams, Mona Landin-Olsson, Birgitte Hylleberg, Hanne Lang, and Eberhard Draeger. 2004. "Insulin Detemir Offers Improved Glycemic Control Compared with NPH Insulin in People with Type 1 Diabetes: A randomized clinical trial." Diabetes Care 27 (5): 1081-1087.
18. Hsieh, Lisa Li-Chen, Chung-Hung Kuo, Ming-Fang Yen, and Tony Hsiu-Hsi Chen. 2004. "A randomized controlled clinical trial for low back pain treated by acupressure and physical therapy." Preventive Medicine 39 (1): 168-176.
19. Iwasaki, Koh, Seiichi Kobayashi, Yuri Chimura, Mayumi Taguchi, Kazumi Inoue, Shigehumi Cho, Tetsuo Akiba, Hiroyuki Arai, Jong-Chol Cyong, and Hidetada Sasaki. 2004. "A randomized, double-blind, placebo-controlled clinical trial of the Chinese herbal medicine "ba wei di huang wan" in the treatment of dementia." Journal of the American Geriatrics Society 52 (9): 1518-21.
20. Jirecek, Stefan, Andreas Lee, Imre Pavo, Gerald Crans, Wolfgang Eppel, and Rene Wenzl. 2004. "Raloxifene prevents the growth of uterine leiomyomas in premenopausal women." Fertility and Sterility 81 (1): 132-136.

21. Joos, Stefanie, Benno Brinkhaus, Christa Maluche, Nathalie Maupai, Ralf Kohnen, Nils Kraehmer, Eckhart G. Hahn, and Detlef Schuppan. 2004. "Acupuncture and moxibustion in the treatment of active Crohn's disease: A randomized controlled study." Digestion 69 (3): 131-139.
22. Juniper, Elizabeth F., Klas Svensson, Arm Christin Mörk, and Elisabeth Ståhl. 2004. "Measuring Health-Related Quality of Life in Adults during an Acute Asthma Exacerbation." Chest 125 (1): 93-97.
23. Kaltwasser, J. Peter, Peter Nash, Dafna Gladman, Cheryl F. Rosen, Frank Behrens, Peter Jones, Jürgen Wollenhaupt, Franziska G. Falk, and Philip Mease. 2004. "Efficacy and safety of leflunomide in the treatment of psoriatic arthritis and psoriasis: A multinational, double-blind, randomized, placebo-controlled clinical trial." Arthritis and Rheumatism 50 (6): 1939-1950.
24. Karlson, Elizabeth W., Matthew H. Liang, Holley Eaton, Jie Huang, Lisa Fitzgerald, Malcolm P. Rogers, and Lawren H. Daltroy. 2004. "A randomized clinical trial of a psychoeducational intervention to improve outcomes in systemic lupus erythematosus." Arthritis and Rheumatism 50 (6): 1832-1841.
25. Kwon, Hyuk-Sang, Jae-Hyoung Cho, Hee-Soo Kim, Bok-Re Song, Seung-Hyun Ko, Jung-Min Lee, Sung-Rae Kim, et al. 2004. "Establishment of blood glucose monitoring system using the internet." Diabetes care 27 (2): 478-83.
26. Lissin, Lynette Wroblewski, Roberta Oka, Subbu Lakshmi, and John P Cooke. 2004. "Isoflavones improve vascular reactivity in post-menopausal women with hypercholesterolemia." Vascular medicine (London, England) 9 (1): 26-30.
27. Mathus-Vliegen, E. M.H., M. L. Van Ierland-Van Leeuwen, and A. Terpstra. 2004. "Lipase inhibition by orlistat: Effects on gall-bladder kinetics and cholecystokinin release in obesity." Alimentary Pharmacology and Therapeutics 19 (5): 601-611.
28. McCracken, G. I., L. Heasman, F. Stacey, N. Steen, M. DeJager, and P. A. Heasman. 2004. "A clinical comparison of an oscillating/rotating powered toothbrush and a manual toothbrush in patients with chronic periodontitis." Journal of Clinical Periodontology 31 (9): 805-812.
29. Molnár, Zsolt, András Mikor, Tamás Leiner, and Tamás Szakmány. 2004. "Fluid resuscitation with colloids of different molecular weight in septic shock." Intensive Care Medicine 30 (7): 1356-1360.
30. Mundorf, Thomas K., Takahiro Ogawa, Hiroaki Naka, Gary D. Novack, and R. Stephens Crockett. 2004. "A 12-month, multicenter, randomized, double-masked, parallel-group comparison of timolol-LA once daily and timolol maleate ophthalmic solution twice daily in the treatment of adults with glaucoma or ocular hypertension." Clinical Therapeutics 26 (4): 541-551.
31. Orringer, Jeffrey S., Sewon Kang, Ted Hamilton, Wendy Schumacher, Soyun Cho, Craig Hammerberg, Gary J. Fisher, Darius J. Karimipour, Timothy M. Johnson, and John J. Voorhees. 2004. "Treatment of Acne Vulgaris With a Pulsed Dye Laser." Jama 291 (23): 2834.

32. Osman, Abdulfatah, Javier Otero, Alberto Brizolara, Sergio Waxman, George Stouffer, Peter Fitzgerald, and Barry F. Uretsky. 2004. "Effect of rosiglitazone on restenosis after coronary stenting in patients with type 2 diabetes." American Heart Journal 147 (5): e21-e25.
33. Pearson, P. J.K., S. A. Lewis, J. Britton, and A. Fogarty. 2004. "Vitamin E supplements in asthma: A parallel group randomised placebo controlled trial." Thorax 59 (8): 652-656.
34. Redondo, Javier Rivera, Carmen Moratalla Justo, Francisca Valdepeñas Moraleda, Yolanda García Velayos, José Juan Osés Puche, Julio Ruiz Zubero, Teresa González Hernández, Loreto Carmona Ortells, and Miguel Ángel Vallejo Pareja. 2004. "Long-term efficacy of therapy in patients with fibromyalgia: A physical exercise-based program and a cognitive-behavioral approach." Arthritis Care & Research 51 (2): 184-192.
35. Saavedra, Miguel A, Leobardo Tera, Federico Galva, Juan M Miranda, Luis J Jara, Leonor Barile, and The Grecia Group. 2004. “A Randomized Double-Blind, Multicenter, Controlled Clinical Trial of Cyclosporine Plus Chloroquine vs. Cyclosporine Plus Placebo in Early-Onset Rheumatoid Arthritis." Archives of Medical Research 35 (1): 36-42.
36. Silva, Robert Carvalho Da, Julio César Joly, Antonio Fernando Martorelli de Lima, and Dimitris N. Tatakis. 2004. "Root Coverage Using the Coronally Positioned Flap With or Without a Subepithelial Connective Tissue Graft." Journal of Periodontology 75 (3): 413-419.
37. Spinhoven, Philip, Moniek Ter Kuile, Ank M.J. Kole-Snijders, Menno Hutten Mansfeld, Dirk Jan Den Ouden, and Johan W.S. Vlaeyen. 2004. "Catastrophizing and internal pain control as mediators of outcome in the multidisciplinary treatment of chronic low back pain." European Journal of Pain 8 (3): 211-219.
38. Tulloch, J. F.Camilla, William R. Proffit, and Ceib Phillips. 2004. "Outcomes in a 2-phase randomized clinical trial of early class II treatment." American Journal of Orthodontics and Dentofacial Orthopedics 125 (6): 657-667.
39. Varady, K.A., N. Ebine, C.A. Vanstone, W.E. Parsons, and P.J.H. Jones. 2004. "Plant sterols and endurance training combine to favorably alter plasma lipid profiles in previously sedentary hypercholesterolemic adults after 8 wk." American Journal of Clinical Nutrition 80 (5): 1159-1166.
40. Vouros, Ioannis, Elena Aristodimou, and Antonis Konstantinidis. 2004. "Guided tissue regeneration in intrabony periodontal defects following treatment with two bioabsorbable membranes in combination with bovine bone mineral graft. A clinical and radiographic study." Journal of clinical periodontology 31 (10): 908-17.
41. Winters, Michael V, Charles G Blake, Jennifer S Trost, Toni B Marcello-brinker, Lynne Lowe, Matthew B Garber, and Robert S Wainner. 2004. "Passive Versus Active Stretching of Hip Flexor Muscles in Subjects With Limited Hip Extension: A Randomized Clinical Trial." Physical Therapy 84 (9): 800-807.
42. Wollert, Kai C., Gerd P. Meyer, Joachim Lotz, Stefanie Ringes-Lichtenberg, Peter Lippolt, Christiane Breidenbach, Stephanie Fichtner, et al. 2004. "Intracoronary autologous bone-marrow cell transfer after myocardial infarction: The BOOST randomised controlled clinical trial." Lancet 364 (9429): 141-148.

43. Zinman, Lorne H, Mylan, Eduardo T, Khin T, Sven Gogov, and Vera Bril. 2004. "Low-Intensity Laser Therapy for Painful Symptoms of Diabetic Sensorimotor." Emerging Treatments and Technologies 27 (4): 921-924.

References 2007

1. Albert-Kiszely, A., B. E. Pjetursson, G. E. Salvi, J. Witt, A. Hamilton, G. R. Persson, and N. P. Lang. 2007. "Comparison of the effects of cetylpyridinium chloride with an essential oil mouth rinse on dental plaque and gingivitis - a six-month randomized controlled clinical trial." Journal of Clinical Periodontology 34 (8): 658-667.
2. Allen, Jerilyn K., Diane M. Becker, Peter O. Kwiterovich, Kathleen A. Lindenstruth, and Carol Curtis. 2007. "Effect of soy protein-containing isoflavones on lipoproteins in postmenopausal women." Menopause 14 (1): 106-114.
3. Bagheri, Abbas, Maryam Aletaha, Hossein Saloor, and Shahin Yazdani. 2007. "A randomized clinical trial of two methods of fascia lata suspension in congenital ptosis." Ophthalmic Plastic and Reconstructive Surgery 23 (3): 217-221.
4. Brennan, Gerard P., Steve D. Moffit, Weston Lindsay, Aaron Swalberg, Brian Rodriquez, Julie M. Fritz, James W. Matheson, and Stephen J. Hunter. 2007. "Is There a Subgroup of Patients With Low Back Pain Likely to Benefit From Mechanical Traction?" Spine 32 (26): E793-E800.
5. Christodoulakos, George E., Irene V. Lambrinoudaki, Emmanuel V. Economou, Constantinos Papadias, Nikolaos Vitoratos, Constantinos P. Panoulis, Evangelia E. Kouskouni, Sofia A. Vlachou, and George C. Creatsas. 2007. "Circulating chemoattractants RANTES, negatively related to endogenous androgens, and MCP-1 are differentially suppressed by hormone therapy and raloxifene." Atherosclerosis 193 (1): 142-150.
6. Claudino, AM, IR De Oliveira, JC Appolinario, TA Cordas, M Duchesne, and R Sichieri. 2007. "Double-blind, randomized, placebo-controlled trial of topiramate plus cognitive-behavior therapy in binge-eating disorder." Journal of Clinical Psychiatry 68 (9): 1324-1332.
7. Dans, Antonio Miguel Limcaco, Maria Vanessa C Villarruz, Cecilia A. Jimeno, Mark Anthony U Javelosa, Joel Chua, Rhida Bautista, and Gwyneth Giselle B Velez. 2007. "The effect of Momordica charantia capsule preparation on glycemic control in Type 2 Diabetes Mellitus needs further studies." Journal of Clinical Epidemiology 60 (6): 554-559.
8. Davenport, Marsha L., Brenda J. Crowe, Sharon H. Travers, Karen Rubin, Judith L. Ross, Patricia Y. Fechner, Daniel F. Gunther, et al. 2007. "Growth hormone treatment of early growth failure in toddlers with turner syndrome: A randomized, controlled, multicenter trial." Journal of Clinical Endocrinology and Metabolism 92 (9): 3406-3416.
9. Drager, Luciano F., Luiz A. Bortolotto, Adelaide C. Figueiredo, Eduardo M. Krieger, and Geraldo Lorenzi-Filho. 2007. "Effects of continuous positive airway pressure on early signs of atherosclerosis in obstructive sleep apnea." American Journal of Respiratory and Critical Care Medicine 176 (7): 706-712.
10. Feldman, Robert M., Angelo P. Tanna, Ronald L. Gross, Alice Z. Chuang, Laura Baker, Adam Reynolds, and Thomas C. Prager. 2007. "Comparison of the Ocular Hypotensive Efficacy of Adjunctive Brimonidine 0.15% or Brinzolamide 1% in Combination with Travoprost 0.004%." Ophthalmology 114 (7).

11. Glaros, A. G., Z. Owais, and L. Lausten. 2007. "Reduction in parafunctional activity: A potential mechanism for the effectiveness of splint therapy." Journal of Oral Rehabilitation 34 (2): 97-104.
12. H.S., Feng, Pinheiro I.C.M., Grande S.R., Pannuti C.M., Barros F.J.N., Lotufo R.F.M., Hsu Shao Feng, et al. 2007. "Effectiveness of a triclosan/copolymer dentifrice on dental plaque and gingivitis in Brazilian individuals with cerebral palsy." Special Care in Dentistry 27 (4): 144-148.
13. Heimann, Heinrich, Karl Ulrich Bartz-Schmidt, Norbert Bornfeld, Claudia Weiss, Ralf Dieter Hilgers, and Michael H. Foerster. 2007. "Scleral Buckling versus Primary Vitrectomy in Rhegmatogenous Retinal Detachment. A Prospective Randomized Multicenter Clinical Study." Ophthalmology 114 (12).
14. Kratochvil, C.J., D. Faries, B. Vaughan, A. Perwien, J. Busner, K. Saylor, S. Kaplan, C. Buermeyer, and R. Swindle. 2007. "Emotional Expression During Attention-Deficit/Hyperactivity Disorders Treatment: Initial Assessment of Treatment Effects." Journal of Child and Adolescent Psychopharmacology 17 (1): 51-62.
15. Kurlan, R., J. Cummings, R. Raman, and L. Thal. 2007. "Quetiapine for agitation or psychosis in patients with dementia and parkinsonism." Neurology 68 (17): 1356-1363.
16. Lam, Dennis S.C., Carmen K.M. Chan, Shaheeda Mohamed, Timothy Y.Y. Lai, Vincent Y.W. Lee, David T.L. Liu, Kenneth K.W. Li, Patrick S.H. Li, and Mahesh P. Shanmugam. 2007. "Intravitreal Triamcinolone plus Sequential Grid Laser versus Triamcinolone or Laser Alone for Treating Diabetic Macular Edema. Six-Month Outcomes." Ophthalmology 114 (12): 2162-2168.
17. Lee, Jonghyun, William R. Freeman, Stanley P. Azen, Eun Jee Chung, and Hyoung Jun Koh. 2007. "Prospective, randomized clinical trial of intravitreal triamcinolone treatment of neovascular age-related macular degeneration: One-year results." Retina 27 (9): 1205-1213.
18. Lee, Yu Mi, Thomas Skurk, Michael Hennig, and Hans Hauner. 2007. "Effect of a milk drink supplemented with whey peptides on blood pressure in patients with mild hypertension." European Journal of Nutrition 46 (1): 21-27.
19. Leone, Antonio Maria, Leonarda Galiuto, Barbara Garramone, Sergio Rutella, Maria Benedetta Giannico, Salvatore Brugaletta, Matteo Perfetti, et al. 2007. "Usefulness of Granulocyte Colony-Stimulating Factor in Patients With a Large Anterior Wall Acute Myocardial Infarction to Prevent Left Ventricular Remodeling (The Rigenera Study)." American Journal of Cardiology 100 (3): 397-403.
20. Li, J., H. Tian, Q. Li, N. Wang, T. Wu, Y. Liu, Z. Ni, et al. 2007. "Improvement of insulin sensitivity and β-cell function by nateglinide and repaglinide in type 2 diabetic patients - A randomized controlled double-blind and double-dummy multicentre clinical trial." Diabetes, Obesity and Metabolism 9 (4): 558-565.
21. Lin-Tan, Dan Tzu, Ja Liang Lin, Tzung Hai Yen, Kuan Hsing Chen, and Yen Lin Huang. 2007. "Long-term outcome of repeated lead chelation therapy in progressive non-diabetic chronic kidney diseases." Nephrology Dialysis Transplantation 22 (10): 2924-2931.

22. Luty, Suzanne E, Janet D Carter, Janice M McKenzie, Alma M Rae, Christopher M. A. Frampton, Roger T Mulder, and Peter R Joyce. 2007. "Randomised controlled trial of interpersonal psychotherapy and cognitive–behavioural therapy for depression." British Journal of Psychiatry 190 (6): 496-502.
23. Maiti, Rituparna, N. K. Agrawal, D. Dash, and B. L. Pandey. 2007. "Effect of Pentoxifylline on inflammatory burden, oxidative stress and platelet aggregability in hypertensive type 2 diabetes mellitus patients." Vascular Pharmacology 47 (2-3): 118-124.
24. Martínez-Abundis, Esperanza, Carlos Alejandro Molina-Villa, Manuel González-Ortiz, José Antonio Robles-Cervantes, and José Antonio Saucedo-Ortiz. 2007. "Effect of surgically removing subcutaneous fat by abdominoplasty on leptin concentrations and insulin sensitivity." Annals of Plastic Surgery 58 (4): 416-419.
25. McTiernan, Anne, Bess Sorensen, Melinda L. Irwin, Angela Morgan, Yutaka Yasui, Rebecca E. Rudolph, Christina Surawicz, et al. 2007. "Exercise effect on weight and body fat in men and women." Obesity 15 (6): 1496-1512.
26. Melnyk, Bernadette Mazurek, Leigh Small, Dianne Morrison-Beedy, Anne Strasser, Lisa Spath, Richard Kreipe, Hugh Crean, Diana Jacobson, Stephanie Kelly, and Judith O'Haver. 2007. "The COPE Healthy Lifestyles TEEN Program: Feasibility, Preliminary Efficacy, & Lessons Learned from an After School Group Intervention with Overweight Adolescents." Journal of Pediatric Health Care 21 (5): 315-322.
27. Milrod, Barbara, Andrew C Leon, Fredric Busch, Marie Rudden, Michael Schwalberg, John Clarkin, Andrew Aronson, et al. 2007. "A Randomized Controlled Clinical Trial of Psychoanalytic Psychotherapy for Panic Disorder." American Journal of Psychiatry 164 (2): 265-272.
28. Pfützner, A., M. Hanefeld, G. Lübben, M. M. Weber, E. Karaglannis, C. Köhler, C. Hohberg, and T. Forst. 2007. "Visfatin: A putative biomarker for metabolic syndrome is not influenced by pioglitazone or simvastatin treatment in nondiabetic patients at cardiovascular risk - Results from the PIOSTAT study." Hormone and Metabolic Research 39 (10): 764-768.
29. Piccirillo, Jay F., Joshua Finnell, Anna Vlahiotis, Richard A. Chole, and Edward Spitznagel. 2007. "Relief of Idiopathic Subjective Tinnitus." Archives of Otolaryngology–Head & Neck Surgery 133 (4): 390.
30. Rosie, Juliet, and Denise Taylor. 2007. "Sit-to-stand as home exercise for mobility-limited adults over 80 years of age - GrandStand System™ may keep you standing?" Age and Ageing 36 (5): 555-562.
31. Rossetti, Luca, Costas H. Karabatsas, Fotis Topouzis, Michele Vetrugno, Marco Centofanti, Andreas Boehm, Ananth Viswanathan, Christian Vorwerk, and David Goldblum. 2007. "Comparison of the Effects of Bimatoprost and a Fixed Combination of Latanoprost and Timolol on Circadian Intraocular Pressure." Ophthalmology 114 (12): 2244-2252.
32. Soheilian, Masoud, Alireza Ramezani, Bijan Bijanzadeh, Mehdi Yaseri, Hamid Ahmadieh, Mohammad H. Dehghan, Mohsen Azarmina, Siamak Moradian, Homa Tabatabaei, and Gholam A. Peyman. 2007. "Intravitreal bevacizumab (Avastin) injection alone or combined

with triamcinolone versus macular photocoagulation as primary treatment of diabetic macular edema." Retina 27 (9): 1187-1195.
33. Stewart, Catherine E., David A. Stephens, Alistair R. Fielder, and Merrick J. Moseley. 2007. "Objectively monitored patching regimens for treatment of amblyopia: Randomised trial." British Medical Journal 335 (7622): 707-711.
34. Tugay, Nazan, Türkan Akbayrak, Funda Demirtürk, Ilkim Çitak Karakaya, Özge Kocaacar, Umut Tugay, Mehmet Gürhan Karakaya, and Fazli Demirtürk. 2007. "Effectiveness of transcutaneous electrical nerve stimulation and interferential current in primary dysmenorrhea." Pain Medicine 8 (4): 295-300.
35. Vallance, Jeffrey K.H., Kerry S. Courneya, Ronald C. Plotnikoff, Yutaka Yasui, and John R. Mackey. 2007. "Randomized controlled trial of the effects of print materials and step pedometers on physical activity and quality of life in breast cancer survivors." Journal of Clinical Oncology 25 (17): 2352-2359.
36. Wessels, Tina, Thomas Ewert, Heribert Limm, Berid Rackwitz, and Gerold Stucki. 2007. "Change factors explaining reductions of "interference" in a multidisciplinary and an exercise prevention program for low back pain." Clinical Journal of Pain 23 (7): 629-634.
37. Wilson, Andrew M., Randall Harada, Nandini Nair, Naras Balasubramanian, and John P. Cooke. 2007. "L-arginine supplementation in peripheral arterial disease: No benefit and possible harm." Circulation 116 (2): 188-195.
38. Wister, Andrew V., Nadine Loewen, Holly Kennedy-Symonds, Brian McGowan, Bonnie McCoy, and Joel Singer. 2007. "One-year follow-up of a therapeutic lifestyle intervention targeting cardiovascular disease risk." Cmaj 177 (8): 859-865.
39. Yancy, W. S., M. K. Olsen, T. Dudley, and E. C. Westman. 2007. "Acid-base analysis of individuals following two weight loss diets." European Journal of Clinical Nutrition 61 (12): 1416-1422.
40. Zhang, Ge, Ling Qin, and Yinyu Shi. 2007. "Epimedium-derived phytoestrogen flavonoids exert beneficial effect on preventing bone loss in late postmenopausal women: A 24-month randomized, double-blind and placebo-controlled trial." Journal of Bone and Mineral Research 22 (7): 1072-1079.

References 2010

1. Beerens, M. W., M. H. Van Der Veen, H. Van Beek, and J. M. Ten Cate. 2010. "Effects of casein phosphopeptide amorphous calcium fluoride phosphate paste on white spot lesions and dental plaque after orthodontic treatment: A 3-month follow-up." European Journal of Oral Sciences 118 (6): 610-617.
2. Bennell, K. L., M. A. Hunt, T. V. Wrigley, D. J. Hunter, F. J. McManus, P. W. Hodges, L. Li, and R. S. Hinman. 2010. "Hip strengthening reduces symptoms but not knee load in people with medial knee osteoarthritis and varus malalignment: A randomised controlled trial." Osteoarthritis and Cartilage (Elsevier Ltd) 18 (5): 621-628.
3. Bennell, Kim L, Bernadette Matthews, Alison Greig, Andrew Briggs, Anne Kelly, Margaret Sherburn, Judy Larsen, and John Wark. 2010. "Effects of an exercise and manual therapy program on physical impairments, function and quality-of-life in people with osteoporotic vertebral fracture: A randomised, single-blind controlled pilot trial." BMC musculoskeletal disorders 11: 36.
4. Bergenstal, Richard M., Carol Wysham, Leigh MacConell, Jaret Malloy, Brandon Walsh, Ping Yan, Ken Wilhelm, Jim Malone, and Lisa E. Porter. 2010. "Efficacy and safety of exenatide once weekly versus sitagliptin or pioglitazone as an adjunct to metformin for treatment of type 2 diabetes (DURATION-2): A randomised trial." The Lancet (Elsevier Ltd) 376 (9739): 431-439.
5. Bergman, Gert J., Jan C. Winters, Klaas H. Groenier, Betty Meyboom-de Jong, Klaas Postema, and Geert J. van der Heijden. 2010. "Manipulative Therapy in Addition to Usual Care for Patients With Shoulder Complaints: Results of Physical Examination Outcomes in a Randomized Controlled Trial." Journal of Manipulative and Physiological Therapeutics (National University of Health Sciences) 33 (2): 96-101.
6. Bot, M., F. Pouwer, J. Assies, E. H.J.M. Jansen, M. Diamant, F. J. Snoek, A. T.F. Beekman, and P. De Jonge. 2010. "Eicosapentaenoic acid as an add-on to antidepressant medication for co-morbid major depression in patients with diabetes mellitus: A randomized, double-blind placebo-controlled study." Journal of Affective Disorders (Elsevier B.V.) 126 (1-2): 282-286.
7. Buckley, Jonathan D., Rebecca L. Thomson, Alison M. Coates, Peter R.C. Howe, Mark O. DeNichilo, and Michelle K. Rowney. 2010. "Supplementation with a whey protein hydrolysate enhances recovery of muscle force-generating capacity following eccentric exercise." Journal of Science and Medicine in Sport 13 (1): 178-181.
8. Bxarone, Paolo, Werner Poewe, Stefan Albrecht, Catherine Debieuvre, Dan Massey, Olivier Rascol, Eduardo Tolosa, and Daniel Weintraub. 2010. "Pramipexole for the treatment of depressive symptoms in patients with Parkinson's disease: a randomised, double-blind, placebo-controlled trial." The Lancet Neurology (Elsevier Ltd) 9 (6): 573-580.
9. Castro, Mario, Adalberto S. Rubin, Michel Laviolette, Jussara Fiterman, Marina De Andrade Lima, Pallav L. Shah, Elie Fiss, et al. 2010. "Effectiveness and safety of bronchial thermoplasty in the treatment of severe asthma: A multicenter, randomized, double-blind, sham-controlled clinical trial." American Journal of Respiratory and Critical Care Medicine 181 (2): 116-124.

10. Dangour, Alan D, Elizabeth Allen, Diana Elbourne, Nicky Fasey, Astrid E Fletcher, Pollyanna Hardy, Graham E Holder, et al. 2010. "Effect of 2-y n-3 long-chain polyunsaturated fatty acid supplementation on cognitive function in older people: a randomized, double-blind, controlled trial." The American journal of clinical nutrition 91 (6): 1725-32.
11. Defoor, Jgm. 2010. "Effect of creatine supplementation as a potential adjuvant therapy to exercise training in cardiac patients: A randomized controlled trial." Clinical Rehabilitation 24 (11): 988-999.
12. Desch, Steffen, Daniela Kobler, Johanna Schmidt, Melanie Sonnabend, Volker Adams, Mahdi Sareban, Ingo Eitel, Matthias Blüher, Gerhard Schuler, and Holger Thiele. 2010. "Low vs. Higher-dose dark chocolate and blood pressure in cardiovascular high-risk patients." American Journal of Hypertension (Nature Publishing Group) 23 (6): 694-700.
13. Diamant, Michaela, Luc Van Gaal, Stephen Stranks, Justin Northrup, Dachuang Cao, Kristin Taylor, and Michael Trautmann. 2010. "Once weekly exenatide compared with insulin glargine titrated to target in patients with type 2 diabetes (DURATION-3): an open-label randomised trial." The Lancet (Elsevier Ltd) 375 (9733): 2234-2243.
14. Durán-Cantolla, Joaquín, Felipe Aizpuru, Jose María Montserrat, Eugeni Ballester, Joaquín Terán-Santos, Jose Ignacio Aguirregomoscorta, Mónica Gonzalez, et al. 2010. "Continuous positive airway pressure as treatment for systemic hypertension in people with obstructive sleep apnoea: Randomised controlled trial." BMJ (Online) 341 (7783): 1142.
15. Elliott, Amanda F., Louis D. Burgio, and Jamie Decoster. 2010. "Enhancing caregiver health: Findings from the resources for enhancing Alzheimer's Caregiver Health II intervention." Journal of the American Geriatrics Society 58 (1): 30-37.
16. Engelmann, Markus G., Hans D. Theiss, Christine Theiss, Volkmar Henschel, Armin Huber, Bernd J. Wintersperger, Stefan O. Schoenberg, Gerhard Steinbeck, and Wolfgang M. Franz. 2010. "G-CSF in patients suffering from late revascularised ST elevation myocardial infarction: Final 1-year-results of the G-CSF-STEMI Trial." International Journal of Cardiology (Elsevier Ireland Ltd) 144 (3): 399-404.
17. Esty, Mary Lee, C. C. Stuart Donaldson, Len Ochs, David V. Nelson, Robert M. Bennett, Gary J. Sexton, Kim D. Jones, and Andre Barkhuizen. 2010. "Neurotherapy of Fibromyalgia?" Pain Medicine 11 (6): 912-919.
18. Fitzgerald, Diarmaid, Nanthana Trakarnratanakul, Barry Smyth, and Brian Caulfield. 2010. "Effects of a Wobble Board-Based Therapeutic Exergaming System for Balance Training on Dynamic Postural Stability and Intrinsic Motivation Levels." Journal of Orthopaedic & Sports Physical Therapy 40 (1): 11-19.
19. Ford, Anna L., Cecilia Bergh, Per Södersten, Matthew A. Sabin, Sandra Hollinghurst, Linda P. Hunt, and Julian P.H. Shield. 2010. "Treatment of childhood obesity by retraining eating behaviour: Randomised controlled trial." BMJ (Online) 340 (7740): 250.
20. Fowler, E. G., L. M. Knutson, S. K. DeMuth, K. L. Siebert, V. D. Simms, M. H. Sugi, R. B. Souza, R. Karim, and S. P. Azen. 2010. "Pediatric Endurance and Limb Strengthening (PEDALS) for Children With Cerebral Palsy Using Stationary Cycling: A Randomized Controlled Trial." Physical Therapy 90 (3): 367-381.

21. Fritsche, A., M. Larbig, D. Owens, and H. U. Häring. 2010. "Comparison between a basal-bolus and a premixed insulin regimen in individuals with type 2 diabetes-results of the GINGER study." Diabetes, Obesity and Metabolism 12 (2): 115-123.
22. Gerstein, Hertzel C., Robert E. Ratner, Christopher P. Cannon, Patrick W. Serruys, Héctor M. García-García, Gerrit-Anne van Es, Nikheel S. Kolatkar, et al. 2010. "Effect of Rosiglitazone on Progression of Coronary Atherosclerosis in Patients With Type 2 Diabetes Mellitus and Coronary Artery Disease." Circulation 121 (10): 1176-1187.
23. Gibbons, Christopher E., Brian G. Pietrosimone, Joseph M. Hart, A. Saliba Susan, and Christopher D. Ingersoll. 2010. "Transcranial magnetic stimulation and volitional quadriceps activation." Journal of Athletic Training 45 (6): 570-579.
24. Goldman, Rose H., William B. Stason, Sung Kyun Park, Rokho Kim, Sharmila Mudgal, Roger B Davis, and Ted J Kaptchuk. 2010. "Low-dose amitriptyline for treatment of persistent arm pain due to repetitive use." Pain 149 (1): 117-123.
25. Heisler, Michele, Sandeep Vijan, Fatima Makki, and John D. Piette. 2010. "Diabetes Control With Reciprocal Peer Support Versus Nurse Care Management." Annals of Internal Medicine 153 (8): 507.
26. Ho, S. G.Y., C. K. Yeung, and H. H.L. Chan. 2010. "Methotrexate versus traditional Chinese medicine in psoriasis: A randomized, placebo-controlled trial to determine efficacy, safety and quality of life." Clinical and Experimental Dermatology 35 (7): 717-722.
27. Hong, K. S., D. W. Kang, H. J. Bae, Y. K. Kim, M. K. Han, J. M. Park, J. H. Rha, et al. 2010. "Effect of cilnidipine vs losartan on cerebral blood flow in hypertensive patients with a history of ischemic stroke: A randomized controlled trial." Acta Neurologica Scandinavica 121 (1): 51-57.
28. Ishii, Hideki, Masami Nishio, Hiroshi Takahashi, Toru Aoyama, Miho Tanaka, Takanobu Toriyama, Tsuneo Tamaki, et al. 2010. "Comparison of Atorvastatin 5 and 20 mg/d for Reducing F-18 Fluorodeoxyglucose Uptake in Atherosclerotic Plaques on Positron Emission Tomography/Computed Tomography: A Randomized, Investigator-Blinded, Open-Label, 6-Month Study in Japanese Adults Scheduled." Clinical Therapeutics (Elsevier Inc.) 32 (14): 2337-2347.
29. Jia, Guo, Mao Bin Meng, Zong Wen Huang, Xia Qing, Wang Lei, Xiao Nan Yang, Song Shan Liu, et al. 2010. "Treatment of functional constipation with the Yun-chang capsule: A double-blind, randomized, placebo-controlled, dose-escalation trial." Journal of Gastroenterology and Hepatology (Australia) 25 (3): 487-493.
30. Jorge, Ricardo E, Laura Acion, David Moser, Harold P. Adams, and Robert G Robinson. 2010. "Escitalopram and Enhancement of Cognitive Recovery Following Stroke." Archives of General Psychiatry 67 (2): 187.
31. Kabra, S. K., R. Pawaiya, Rakesh Lodha, Arti Kapil, Madhulika Kabra, A. Satya Vani, G. Agarwal, and S. S. Shastri. 2010. "Long-term daily high and low doses of azithromycin in children with cystic fibrosis: A randomized controlled trial." Journal of Cystic Fibrosis (European Cystic Fibrosis Society) 9 (1): 17-23.

32. Kattelmann, Kendra K., Kibbe Conti, and Cuirong Ren. 2010. "The Medicine Wheel Nutrition Intervention: A Diabetes Education Study with the Cheyenne River Sioux Tribe." Journal of the American Dietetic Association (Elsevier Inc.) 109 (9): 1532-1539.
33. Kemmler, Wolfgang. 2010. "Exercise Effects on Bone Mineral Density, Falls, Coronary Risk Factors, and Health Care Costs in Older Women." Archives of Internal Medicine 170 (2): 179.
34. Kemp, Sue, Ian Roberts, Carrol Gamble, Stuart Wilkinson, Joyce E. Davidson, Eileen M. Baildam, Andrew Gavin Cleary, Liza J. McCann, and Michael W. Beresford. 2010. "A randomized comparative trial of generalized vs targeted physiotherapy in the management of childhood hypermobility." Rheumatology 49 (2): 315-325.
35. Kiritsi, Olga, Konstantinos Tsitas, Nikolaos Malliaropoulos, and Grogorios Mikroulis. 2010. "Ultrasonographic evaluation of plantar fasciitis after low-level laser therapy: Results of a double-blind, randomized, placebo-controlled trial." Lasers in Medical Science 25 (2): 275-281.
36. Klusmann, Verena, Andrea Evers, Ralf Schwarzer, Peter Schlattmann, Friedel M. Reischies, Isabella Heuser, and Fernando C. Dimeo. 2010. "Complex mental and physical activity in older women and cognitive performance: A 6-month randomized controlled trial." Journals of Gerontology - Series A Biological Sciences and Medical Sciences 65 A (6): 680-688.
37. Koyasu, Masayoshi, Hideki Ishii, Masato Watarai, Kenji Takemoto, Yasuya Inden, Kyosuke Takeshita, Tetsuya Amano, Daiji Yoshikawa, Tatsuaki Matsubara, and Toyoaki Murohara. 2010. "Impact of acarbose on carotid intima-media thickness in patients with newly diagnosed impaired glucose tolerance or mild type 2 diabetes mellitus: A one-year, prospective, randomized, open-label, parallel-group study in Japanese adults with established co." Clinical Therapeutics 32 (9): 1610-1617.
38. Lawrence, Jean M., Lori Laffel, Tim Wysocki, Dongyuan Xing, Roy W. Beck, Elbert S. Huang, Brett Ives, et al. 2010. "Quality-of-life measures in children and adults with type 1 diabetes: Juvenile diabetes research foundation continuous glucose monitoring randomized trial." Diabetes Care 33 (10): 2175-2177.
39. Maher, Carol A., Marie T. Williams, Tim Olds, and Alison E. Lane. 2010. "An internet-based physical activity intervention for adolescents with cerebral palsy: A randomized controlled trial." Developmental Medicine and Child Neurology 52 (5): 448-455.
40. Manske, Robert C., Matt Meschke, Andrew Porter, Barbara Smith, and Michael Reiman. 2010. "A randomized controlled single-blinded comparison of stretching versus stretching and joint mobilization for posterior shoulder tightness measured by internal rotation motion loss." Sports Health 2 (2): 94-100.
41. McManus, Richard J., Jonathan Mant, Emma P. Bray, Roger Holder, Miren I. Jones, Sheila Greenfield, Billingsley Kaambwa, et al. 2010. "Telemonitoring and self-management in the control of hypertension (TASMINH2): A randomised controlled trial." The Lancet 376 (9736): 163-172.
42. Meyer, Elisabeth, Roseli G Shavitt, Carl Leukefeld, Elizeth Heldt, Fernanda P Souza, Paulo Knapp, and Aristides V Cordioli. 2010. "Adding motivational interviewing and thought mapping to cognitive-behavioral group therapy: results from a randomized clinical trial TT -

Adicionando a entrevista motivacional e o mapeamento cognitivo à terapia cognitivo-comportamental em grupo: resultad." Brazilian Journal of Psychiatry 32 (1): 20-29.

43. Michael, Yvonne L., Rachel Gold, JoAnn E. Manson, Erin M. Keast, Barbara B. Cochrane, Nancy F. Woods, Robert G. Brzyski, S. Gene McNeeley, and Robert B. Wallace. 2010. "Hormone therapy and physical function change among older women in the Women's Health Initiative." Menopause 17 (2): 295-302.
44. Modrego, P. J., N. Fayed, J. M. Errea, C. Rios, M. A. Pina, and M. Sarasa. 2010. "Memantine versus donepezil in mild to moderate Alzheimer's disease: A randomized trial with magnetic resonance spectroscopy." European Journal of Neurology 17 (3): 405-412.
45. Moe, Sharon M., Akber Saifullah, Robert E. LaClair, Sohail A. Usman, and Zhangsheng Yu. 2010. "A randomized trial of cholecalciferol versus doxercalciferol for lowering parathyroid hormone in chronic kidney disease." Clinical Journal of the American Society of Nephrology 5 (2): 299-306.
46. Nerla, Roberto, Dario Pitocco, Francesco Zaccardi, Giancarla Scalone, Ilaria Coviello, Roberto Mollo, Giovanni Ghirlanda, Gaetano A. Lanza, and Filippo Crea. 2010. "Effect of pioglitazone on systemic inflammation is independent of metabolic control and cardiac autonomic function in patients with type 2 diabetes." Acta Diabetologica 47 (SUPPL. 1): 117-122.
47. Parle, J., L. Roberts, S. Wilson, H. Pattison, A. Roalfe, M. S. Haque, C. Heath, M. Sheppard, J. Franklyn, and F. D.R. Hobbs. 2010. "A randomized controlled trial of the effect of thyroxine replacement on cognitive function in community-living elderly subjects with subclinical hypothyroidism: The Birmingham elderly thyroid study." Journal of Clinical Endocrinology and Metabolism 95 (8): 3623-3632.
48. Poole, Chris, Brandon Bushey, Cliffa Foster, Bill Campbell, Darryn Willoughby, Richard Kreider, Lem Taylor, and Colin Wilborn. 2010. "The effects of a commercially available botanical supplement on strength, body composition, power output, and hormonal profiles in resistance-trained males." Journal of the International Society of Sports Nutrition 7: 1-9.
49. Rijal, Raju, Bikram Prasad Shrestha, Girish Kumar Singh, Mahipal Singh, Pravin Nepal, Guru Prasad Khanal, and Pramila Rai. 2010. "Comparison of Ponseti and Kite's method of treatment for idiopathic clubfoot." Indian journal of orthopaedics 44 (2): 202-7.
50. Rosado, Jorge L, Karla E González, María Del C Caamaño, Olga P García, Roxana Preciado, and Mauricio Odio. 2010. "Efficacy of different strategies to treat anemia in children: a randomized clinical trial." Nutrition journal 9 (1): 40.
51. Sadrzadeh-Yeganeh, Haleh, Ibrahim Elmadfa, Abolghasem Djazayery, Mahmoud Jalali, Ramin Heshmat, and Maryam Chamary. 2010. "The effects of probiotic and conventional yoghurt on lipid profile in women." British Journal of Nutrition 103 (12): 1778-1783.
52. Saiman, Lisa, Michael Anstead, Nicole Mayer-Hamblett, Larry C. Lands, Margaret Kloster, Jasna Hocevar-Trnka, Christopher H. Goss, et al. 2010. "Effect of azithromycin on pulmonary function in patients with cystic fibrosis uninfected with Pseudomonas aeruginosa: A randomized controlled trial." JAMA - Journal of the American Medical Association 303 (17): 1707-1715.

53. Salehi, Bahman, Reza Imani, Mohammad Reza Mohammadi, Jalil Fallah, Mohammad Mohammadi, Ahmad Ghanizadeh, Ali Akbar Tasviechi, Ardalan Vossoughi, Shams Ali Rezazadeh, and Shahin Akhondzadeh. 2010. "Ginkgo biloba for Attention-Deficit/Hyperactivity Disorder in children and adolescents: A double blind, randomized controlled trial." Progress in Neuro-Psychopharmacology and Biological Psychiatry (Elsevier Inc.) 34 (1): 76-80.
54. Scheen, André J., Guillaume Charpentier, Carl Johan Östgren, Åsa Hellqvist, and Ingrid Gause-Nilsson. 2010. "Efficacy and safety of saxagliptin in combination with metformin compared with sitagliptin in combination with metformin in adult patients with type 2 diabetes mellitus." Diabetes/Metabolism Research and Reviews 26 (7): 540-549.
55. Swoboda, Kathryn J., Charles B. Scott, Thomas O. Crawford, Louise R. Simard, Sandra P. Reyna, Kristin J. Krosschell, Gyula Acsadi, et al. 2010. "SMA CARNI-VAL trial part I: Double-blind, randomized, placebo-controlled trial of L-carnitine and valproic acid in spinal muscular atrophy." PLoS ONE 5 (8).
56. Tabibi, Hadi, Hossein Imani, Mehdi Hedayati, Shahnaz Atabak, and Leila Rahmani. 2010. "Effects of soy consumption on serum lipids and apoproteins in peritoneal dialysis patients: A randomized controlled trial." Peritoneal Dialysis International 30 (6): 611-618.
57. Thackeray, Anne, Julie M. Fritz, Gerard P. Brennan, Faisel M. Zaman, and Stuart E. Willick. 2010. "A Pilot Study Examining the Effectiveness of Physical Therapy as an Adjunct to Selective Nerve Root Block in the Treatment of Lumbar Radicular Pain From Disk Herniation: A Randomized Controlled Trial." Physical Therapy 90 (12): 1717-1729.
58. Tiwari, Agnes, Daniel Yee Tak Fong, Kwan Hok Yuen, Helina Yuk, Polly Pang, Janice Humphreys, and Linda Bullock. 2010. "Effect of an advocacy intervention on mental health in Chinese women survivors of intimate partner violence: A randomized controlled trial." JAMA - Journal of the American Medical Association 304 (5): 536-543.
59. Webb, Nicholas J.A., Chun Lam, Tom Loeys, Shahnaz Shahinfar, Juergen Strehlau, Thomas G. Wells, Emanuela Santoro, Denise Manas, and Gilbert W. Gleim. 2010. "Randomized, double-blind, controlled study of losartan in children with proteinuria." Clinical Journal of the American Society of Nephrology 5 (3): 417-424.
60. Weie, Kristin, Corinna Brandsch, Bianca Zernsdorf, Germaine S. Nkengfack Nembongwe, Kathleen Hofmann, Klaus Eder, and Gabriele I. Stangl. 2010. "Lupin protein compared to casein lowers the LDL cholesterol:HDL cholesterol-ratio of hypercholesterolemic adults." European Journal of Nutrition 49 (2): 65-71.
61. Wilkens, Philip, Inger B. Scheel, Oliver Grundnes, Christian Hellum, and Kjersti Storheim. 2010. "Effect of glucosamine on pain-related disability in patients with chronic low back pain and degenerative lumbar osteoarthritis: A randomized controlled trial." JAMA - Journal of the American Medical Association 304 (1): 45-52.
62. Witt, Petra M., Jeppe H. Christensen, Marianne Ewertz, Inge V. Aardestrup, and Erik B. Schmidt. 2010. "The incorporation of marine n-3 PUFA into platelets and adipose tissue in pre- and postmenopausal women: A randomised, double-blind, placebo-controlled trial." British Journal of Nutrition 104 (3): 318-325.

63. Wöhrle, Jochen, Nico Merkle, Volker Mailänder, Thorsten Nusser, Peter Schauwecker, Fabian von Scheidt, Klaus Schwarz, et al. 2010. "Results of Intracoronary Stem Cell Therapy After Acute Myocardial Infarction." American Journal of Cardiology (Elsevier Inc.) 105 (6): 804-812.
64. Yoshizawa, Mutsuko, Seiji Maeda, Asako Miyaki, Maiko Misono, Youngju Choi, Nobutake Shimojo, Ryuichi Ajisaka, and Hirofumi Tanaka. 2010. "Additive beneficial effects of lactotripeptides intake with regular exercise on endothelium-dependent dilatation in postmenopausal women." American Journal of Hypertension (Nature Publishing Group) 23 (4): 368-372.
65. Zimmermann, Michael B, Christophe Chassard, Fabian Rohner, Eliézer K N'goran, Charlemagne Nindjin, Alexandra Dostal, Jürg Utzinger, Hala Ghattas, Christophe Lacroix, and Richard F Hurrell. 2010. "The effects of iron fortification on the gut microbiota in African children: a randomized controlled trial in Cote d'Ivoire." The American journal of clinical nutrition 92 (6): 1406-15.
66. Zimmern, Philippe, Heather J. Litman, Elizabeth Mueller, Peggy Norton, and Patricia Goode. 2010. "Effect of fluid management on fluid intake and urge incontinence in a trial for overactive bladder in women." BJU International 105 (12): 1680-1685.

References 2013

1. Agnoletti, Davide, Yi Zhang, Claudio Borghi, Jacques Blacher, and Michel E. Safar. 2013. "Effects of antihypertensive drugs on central blood pressure in humans: A preliminary observation." American Journal of Hypertension 26 (8): 1045-1052.
2. Andreyev, H. Jervoise N., Barbara E. Benton, Amyn Lalji, Christine Norton, Kabir Mohammed, Heather Gage, Kjell Pennert, and James O. Lindsay. 2013. "Algorithm-based management of patients with gastrointestinal symptoms in patients after pelvic radiation treatment (ORBIT): A randomised controlled trial." The Lancet (Elsevier Ltd) 382 (9910): 2084-2092.
3. Asemi, Z., M. Samimi, Z. Tabassi, M. Naghibi Rad, A. Rahimi Foroushani, H. Khorammian, and A. Esmaillzadeh. 2013. "Effect of daily consumption of probiotic yoghurt on insulin resistance in pregnant women: A randomized controlled trial." European Journal of Clinical Nutrition (Nature Publishing Group) 67 (1): 71-74.
4. Baek, Hey Sung, Juhwan Cho, Joo Hwa Kim, Jae Won Oh, and Ha Baik Lee. 2013. "Ratio of leukotriene E4 to exhaled nitric oxide and the therapeutic response in children with exercise-induced bronchoconstriction." Allergy, Asthma and Immunology Research 5 (1): 26-33.
5. Baudic, Sophie, Nadine Attal, Alaa Mhalla, Daniel Ciampi de Andrade, Serge Perrot, and Didier Bouhassira. 2013. "Unilateral repetitive transcranial magnetic stimulation of the motor cortex does not affect cognition in patients with fibromyalgia." Journal of Psychiatric Research (Elsevier Ltd) 47 (1): 72-77.
6. Belgacem, Bénédicte, Candy Auclair, Marie Christine Fedor, David Brugnon, Marie Blanquet, Olivier Tournilhac, and Laurent Gerbaud. 2013. "A caregiver educational program improves quality of life and burden for cancer patients and their caregivers: A randomised clinical trial." European Journal of Oncology Nursing (Elsevier Ltd) 17 (6): 870-876.
7. Bonjour, Jean Philippe, Valérie Benoit, Flore Payen, and Marius Kraenzlin. 2013. "Consumption of yogurts fortified in vitamin D and calcium reduces serum parathyroid hormone and markers of bone resorption: A double-blind randomized controlled trial in institutionalized elderly women." Journal of Clinical Endocrinology and Metabolism 98 (7): 2915-2921.
8. Carroll, Matthew W., Doosoo Jeon, James M. Mountz, Jong Doo Lee, Yeon Joo Jeong, Nadeem Zia, Myungsun Lee, et al. 2013. "Efficacy and safety of metronidazole for pulmonary multidrug-resistant tuberculosis." Antimicrobial Agents and Chemotherapy 57 (8): 3903-3909.
9. Chen, Chun Hsin, Ming Chyi Huang, Chung Feng Kao, Shih Ku Lin, Po Hsiu Kuo, Chih Chiang Chiu, and Mong Liang Lu. 2013. "Effects of adjunctive metformin on metabolic traits in nondiabetic clozapine-treated patients with schizophrenia and the effect of metformin discontinuation on body weight: A 24-week, randomized, double-blind, placebo-controlled study." Journal of Clinical Psychiatry 74 (5).
10. Cho, Yu-Jeong, Yun-Kyung Song, Yun-Yeop Cha, Byung-Cheul Shin, Im-Hee Shin, Hi-Joon Park, Hyang-Sook Lee, et al. 2013. "Acupuncture for chronic low back pain: a multicenter, randomized, patient-assessor blind, sham-controlled clinical trial." Spine 38 (7): 549-57.

11. Cormie, P., R. U. Newton, N. Spry, D. Joseph, D. R. Taaffe, and D. A. Galvão. 2013. "Safety and efficacy of resistance exercise in prostate cancer patients with bone metastases." Prostate Cancer and Prostatic Diseases 16 (4): 328-335.
12. Duberg, Anna, Lars Hagberg, Helena Sunvisson, and Margareta Möller. 2013. "Influencing self-rated health among adolescent girls with dance intervention: A randomized controlled trial." Archives of Pediatrics and Adolescent Medicine 167 (1): 27-31.
13. Faghihi, Gita, Amir Hossein Siadat, and Fariba Iraji. 2013. "The efficacy of acyclovir in treatment of the pemphigus vulgaris." Journal of Research in Medical Sciences 18 (11): 976-978.
14. Findley, Austin D, Matthew T Siedhoff, Kumari A Hobbs, John F. Steege, Erin T Carey, Christina A. McCall, and Anne Z Steiner. 2013. "Short-term effects of salpingectomy during laparoscopic hysterectomy on ovarian reserve: a pilot randomized controlled trial." Fertility and Sterility 100 (6): 1704-1708.
15. Fjorback, Lone Overby, Mikkel Arendt, Eva Ørnbøl, Harald Walach, Emma Rehfeld, Andreas Schröder, and Per Fink. 2013. "Mindfulness therapy for somatization disorder and functional somatic syndromes - Randomized trial with one-year follow-up." Journal of Psychosomatic Research (Elsevier Inc.) 74 (1): 31-40.
16. Ghaleiha, Ali, Effat Mohammadi, Mohammad Reza Mohammadi, Mehdi Farokhnia, Amirhossein Modabbernia, Habibeh Yekehtaz, Mandana Ashrafi, Elmira Hassanzadeh, and Shahin Akhondzadeh. 2013. "Riluzole as an adjunctive therapy to risperidone for the treatment of irritability in children with autistic disorder: A double-blind, placebo-controlled, randomized trial." Pediatric Drugs 15 (6): 505-514.
17. Ghassemi, Fariba, Nazanin Ebrahimiadib, Ramak Roohipoor, Sasan Moghimi, and Fateme Alipour. 2013. "Nerve fiber layer thickness in eyes treated with red versus green laser in proliferative diabetic retinopathy: Short-term results." Ophthalmologica 230 (4): 195-200.
18. Gitlin, Laura N., Lynn Fields Harris, Megan C. Mccoy, Nancy L. Chernett, Laura T. Pizzi, Eric Jutkowitz, Edward Hess, and Walter W. Hauck. 2013. "A home-based intervention to reduce depressive symptoms and improve quality of life in older African Americans: A randomized trial." Annals of Internal Medicine 159 (4): 243-252.
19. Grant, Suzanne J, Dennis Hsu-Tung Chang, Jianxun Liu, Vincent Wong, Hosen Kiat, and Alan Bensoussan. 2013. "Chinese herbal medicine for impaired glucose tolerance: a randomized placebo controlled trial." BMC complementary and alternative medicine 13 (1): 104.
20. Guinan, E., J. Hussey, J. M. Broderick, F. E. Lithander, D. O'Donnell, M. J. Kennedy, and E. M. Connolly. 2013. "The effect of aerobic exercise on metabolic and inflammatory markers in breast cancer survivors - A pilot study." Supportive Care in Cancer 21 (7): 1983-1992.
21. Gupta, Samir K., Deming Mi, Michael P. Dubé, Chandan K. Saha, Raymond M. Johnson, James H. Stein, Matthias A. Clauss, Kieren J. Mather, Zeruesenay Desta, and Ziyue Liu. 2013. "Pentoxifylline, Inflammation, and Endothelial Function in HIV-Infected Persons: A Randomized, Placebo-Controlled Trial." PLoS ONE 8 (4).
22. Hassanzadeh Delui, Mahdy, Hedayatollah Fatehi, Morteza Manavifar, Maral Amini, Majid Ghayour-Mobarhan, Mahdi Zahedi, and Gordon Ferns. 2013. "The effects of Panax ginseng

on lipid profile, pro-oxidant: Antioxidant status and high-sensitivity C reactive protein levels in hyperlipidemic patients in Iran." International Journal of Preventive Medicine 4 (9): 1045-1051.

23. Howe, Samuel T., Phillip M. Bellinger, Matthew W. Driller, Cecilia M. Shing, and James W. Fell. 2013. "The effect of beta-alanine supplementation on isokinetic force and cycling performance in highly trained cyclists." International Journal of Sport Nutrition and Exercise Metabolism 23 (6): 562-570.
24. Husebye, Trygve, Jan Eritsland, Carl Müller, Leiv Sandvik, Harald Arnesen, Ingebjørg Seljeflot, Arild Mangschau, Reidar Bjørnerheim, and Geir Øystein Andersen. 2013. "Levosimendan in acute heart failure following primary percutaneous coronary intervention-treated acute ST-elevation myocardial infarction. Results from the LEAF trial: A randomized, placebo-controlled study." European Journal of Heart Failure 15 (5): 565-572.
25. Kerry, Sally M, Hugh S Markus, Teck K Khong, Geoffrey C Cloud, Jenny Tulloch, Denise Coster, Judith Ibison, and Pippa Oakeshott. 2013. "Home blood pressure monitoring with nurse-led telephone support among patients with hypertension and a history of stroke: a community-based randomized controlled trial." CMAJ : Canadian Medical Association journal = journal de l'Association medicale canadienne 185 (1): 23-31.
26. Kim, Sang Hwan, Suzanne M. Schneider, Margaret Bevans, Len Kravitz, Christine Mermier, Clifford Qualls, and Mark R. Burge. 2013. "PTSD symptom reduction with mindfulness-based stretching and deep breathing exercise: Randomized controlled clinical trial of efficacy." Journal of Clinical Endocrinology and Metabolism 98 (7): 2984-2992.
27. Kojima, Yuichi, Hideyoshi Kaga, Shinu Hayashi, Toru Kitazawa, Yuko Iimura, Makoto Ohno, Michiyasu Yoshitsugu, Mutsunori Fujiwara, and Toru Hiyoshi. 2013. "Comparison between sitagliptin and nateglinide on postprandial lipid levels: The STANDARD study." World Journal of Diabetes 4 (1): 8.
28. Kosmala, Wojciech, David J. Holland, Aleksandra Rojek, Leah Wright, Monika Przewlocka-Kosmala, and Thomas H. Marwick. 2013. "Effect of If-channel inhibition on hemodynamic status and exercise tolerance in heart failure with preserved ejection fraction: A randomized trial." Journal of the American College of Cardiology 62 (15): 1330-1338.
29. Lakerveld, Jeroen, Sandra D. Bot, Mai J. Chinapaw, Maurits W. van Tulder, Piet J. Kostense, Jacqueline M. Dekker, and Giel Nijpels. 2013. "Motivational interviewing and problem solving treatment to reduce type 2 diabetes and cardiovascular disease risk in real life: A randomized controlled trial." International Journal of Behavioral Nutrition and Physical Activity 10: 1-9.
30. Lavender, Tina, Carol Bedwell, Stephen A. Roberts, Anna Hart, Mark A. Turner, Lesley Anne Carter, and Michael J. Cork. 2013. "Randomized, Controlled Trial Evaluating a Baby Wash Product on Skin Barrier Function in Healthy, Term Neonates." JOGNN - Journal of Obstetric, Gynecologic, and Neonatal Nursing (Elsevier Masson SAS) 42 (2): 203-214.
31. Ledderer, Loni, Karen La Cour, Ole Mogensen, Erik Jakobsen, Rene DePont Christensen, Jakob Kragstrup, and Helle Ploug Hansen. 2013. "Feasibility of a psychosocial rehabilitation

intervention to enhance the involvement of relatives in cancer rehabilitation: Pilot study for a randomized controlled trial." Patient 6 (3): 201-212.

32. Lushchyk, Tanya, Sankha Amarakoon, José P. Martinez-Ciriano, L. Ingeborgh Van Den Born, G. Seerp Baarsma, and Tom Missotten. 2013. "Bevacizumab in age-related macular degeneration: A randomized controlled trial on the effect of injections every 4 weeks, 6 weeks and 8 weeks." Acta Ophthalmologica 91 (6): 456-461.
33. McLean LM, Walton T, Rodin G, Esplen MJ, and Jones JM. 2013. "A couple-based intervention for patients and caregivers facing end-stage cancer: outcomes." Psycho-Oncology 22 (1): 28-38.
34. Mousavi, Ateke, Mohammadreza Vafa, Tirang Neyestani, Mohammadebrahim Khamseh, and Fatemeh Hoseini. 2013. "The effects of green tea consumption on metabolic and anthropometric indices in patients with type 2 diabetes." Journal of Research in Medical Sciences 18 (12): 1080-1086.
35. Paramba, Firjeeth C., Vamanjore A. Naushad, Nishan Purayil, Osama H. Mohammed, and Prem Chandra. 2013. "Randomized controlled study of the antipyretic efficacy of oral paracetamol, intravenous paracetamol, and intramuscular diclofenac in patients presenting with fever to the emergency department." Therapeutics and Clinical Risk Management 9 (1): 371-376.
36. Prezio, Elizabeth A., Dunlei Cheng, Bijal A. Balasubramanian, Kerem Shuval, Darla E. Kendzor, and Dan Culica. 2013. "Community Diabetes Education (CoDE) for uninsured Mexican Americans: A randomized controlled trial of a culturally tailored diabetes education and management program led by a community health worker." Diabetes Research and Clinical Practice (Elsevier Ireland Ltd) 100 (1): 19-28.
37. Putman, Melissa S., Sarah A B Pitts, Carly E. Milliren, Henry A. Feldman, Kristina Reinold, and Catherine M. Gordon. 2013. "A randomized clinical trial of vitamin D supplementation in healthy adolescents." Journal of Adolescent Health (Elsevier Ltd) 52 (5): 592-598.
38. Ribeiro, Fernanda V., Renato C.V. Casarin, Maria A.G. Palma, Francisco H.N. Júnior, Enilson A. Sallum, and Márcio Z. Casati. 2013. "Clinical and microbiological changes after minimally invasive therapeutic approaches in intrabony defects: A 12-month follow-up." Clinical Oral Investigations 17 (7): 1635-1644.
39. Rochette, Annie, Nicol Korner-Bitensky, Duane Bishop, Robert Teasell, Carole L. White, Gina Bravo, Robert Côté, et al. 2013. "The YOU CALL–WE CALL Randomized Clinical Trial." Circulation: Cardiovascular Quality and Outcomes 6 (6): 674-679.
40. Rosenquist, Peter B., Andrew Krystal, Karen L. Heart, Mark A. Demitrack, and William Vaughn McCall. 2013. "Left dorsolateral prefrontal transcranial magnetic stimulation (TMS): Sleep factor changes during treatment in patients with pharmacoresistant major depressive disorder." Psychiatry Research (Elsevier) 205 (1-2): 67-73.
41. Rouhani, Mohammad Hossein, Roya Kelishadi, Mahin Hashemipour, Ahmad Esmaillzadeh, and Leila Azadbakht. 2013. "The effect of an energy restricted low glycemic index diet on blood lipids, apolipoproteins and lipoprotein (a) among adolescent girls with excess weight: A randomized clinical trial." Lipids 48 (12): 1197-1205.

42. Sabariego, Carla, Andrea E. Barrera, Silvia Neubert, Marita Stier-Jarmer, Cristina Bostan, and Alarcos Cieza. 2013. "Evaluation of an ICF-based patient education programme for stroke patients: A randomized, single-blinded, controlled, multicentre trial of the effects on self-efficacy, life satisfaction and functioning." British Journal of Health Psychology 18 (4): 707-728.
43. Santos, Vanessa R., Jadson A. Lima, Tamires S. Miranda, Tiago E.D. Gonçalves, Luciene C. Figueiredo, Marcelo Faveri, and Poliana M. Duarte. 2013. "Full-mouth disinfection as a therapeutic protocol for type-2 diabetic subjects with chronic periodontitis: Twelve-month clinical outcomes. A randomized controlled clinical trial." Journal of Clinical Periodontology 40 (2): 155-162.
44. Sathi, P., S. Kalyan, C. L. Hitchcock, M. Pudek, and J. C. Prior. 2013. "Progesterone therapy increases free thyroxine levels-data from a randomized placebo-controlled 12-week hot flush trial." Clinical Endocrinology 79 (2): 282-287.
45. Scholten, L., A. M. Willemen, B. F. Last, H. Maurice-Stam, E. M. van Dijk, E. Ensink, N. Zandbelt, A. van der Hoop-Mooij, C. Schuengel, and M. A. Grootenhuis. 2013. "Efficacy of Psychosocial Group Intervention for Children With Chronic Illness and Their Parents." Pediatrics 131 (4): e1196-e1203.
46. Sihvonen, Raine, Mika Paavola, Antti Malmivaara, Ari Itälä, Antti Joukainen, Heikki Nurmi, Juha Kalske, and Teppo L.N. Järvinen. 2013. "Arthroscopic Partial Meniscectomy versus Sham Surgery for a Degenerative Meniscal Tear." New England Journal of Medicine 369 (26): 2515-2524.
47. Skjeie, Holgeir, Trygve Skonnord, Arne Fetveit, and Mette Brekke. 2013. "Acupuncture for infantile colic: A blinding-validated, randomized controlled multicentre trial in general practice." Scandinavian Journal of Primary Health Care 31 (4): 190-196.
48. Steg, P. G., E. Lopez-de-Sà, F. Schiele, M. Hamon, T. Meinertz, J. Goicolea, K. Werdan, and J. L. Lopez-Sendon. 2013. "Safety of intravenous ivabradine in acute ST-segment elevation myocardial infarction patients treated with primary percutaneous coronary intervention: A randomized, placebo-controlled, double-blind, pilot study." European Heart Journal: Acute Cardiovascular Care 2 (3): 270-279.
49. Stern, J., M. Larsson, P. Kristiansson, and T. Tydén. 2013. "Introducing reproductive life plan-based information in contraceptive counselling: An RCT." Human Reproduction 28 (9): 2450-2461.
50. Strand, Vibeke, David Fiorentino, ChiaChi Hu, Robert M Day, Randall M Stevens, and Kim A Papp. 2013. "Improvements in patient-reported outcomes with apremilast, an oral phosphodiesterase 4 inhibitor, in the treatment of moderate to severe psoriasis: results from a phase IIb randomized, controlled study." Health and quality of life outcomes 11 (1): 82.
51. Subramanian, Sharath, Hamed Emami, Esad Vucic, Parmanand Singh, Jayanthi Vijayakumar, Kenneth M. Fifer, Achilles Alon, et al. 2013. "High-dose atorvastatin reduces periodontal inflammation: A novel pleiotropic effect of statins." Journal of the American College of Cardiology 62 (25): 2382-2391.

52. Udani, Jay K., and David W. Bloom. 2013. "Effects of kivia powder on Gut health in patients with occasional constipation: A randomized, double-blind, placebo-controlled study." Nutrition Journal (Nutrition Journal) 12 (1): 1.
53. Van der Marck, Marjolein A., Bastiaan R. Bloem, George F. Borm, Sebastiaan Overeem, Marten Munneke, and Mark Guttman. 2013. "Effectiveness of multidisciplinary care for Parkinson's disease: A randomized, controlled trial." Movement Disorders 28 (5): 605-611.
54. Weber, Michel, Laurent Kodjikian, Friedrich E. Kruse, Zbigniew Zagorski, and Catherine M. Allaire. 2013. "Efficacy and safety of indomethacin 0.1% eye drops compared with ketorolac 0.5% eye drops in the management of ocular inflammation after cataract surgery." Acta Ophthalmologica 91 (1): 15-21.
55. Weiner, Debra K., Charity G. Moore, Natalia E. Morone, Edward S. Lee, and C. Kent Kwoh. 2013. "Efficacy of periosteal stimulation for chronic pain associated with advanced knee osteoarthritis: A randomized, controlled clinical trial." Clinical Therapeutics (Elsevier) 35 (11): 1703-1720.e5.
56. Wiltheiss, Gina A., Cheryl A. Lovelady, Deborah G. West, Rebecca J.N. Brouwer, Katrina M. Krause, and Truls Østbye. 2013. "Diet Quality and Weight Change among Overweight and Obese Postpartum Women Enrolled in a Behavioral Intervention Program." Journal of the Academy of Nutrition and Dietetics (Elsevier) 113 (1): 54-62.
57. Wolak, Talya, Elizaveta Aliev, Boris Rogachev, Yael Baumfeld, Carlos Cafri, Mahmoud Abu-Shakra, and Victor Novack. 2013. "Renal safety and angiotensin II blockade medications in patients undergoing non-emergent coronary angiography: A randomized controlled study." Israel Medical Association Journal 15 (11): 682-687.
58. Yuan, Wei An, Shi Rong Huang, Kai Guo, Wu Quan Sun, Xiao Bing Xi, Ming Cai Zhang, Ling Jun Kong, Hua Lu, Hong Sheng Zhan, and Ying Wu Cheng. 2013. "Integrative TCM Conservative Therapy for Low Back Pain due to Lumbar Disc Herniation: A Randomized Controlled Clinical Trial." Evidence-Based Complementary and Alternative Medicine 2013: 1-8.
59. Ziaaddini, Hassan, Batoul Ebrahim-Nejad, and Nouzar Nakhaee. 2013. "The Effectiveness of Group Therapy on the Family Functioning of Individuals under Methadone Treatment: A Clinical Trial." Addiction & health 5 (1-2): 1-6.

References

1. Collins, F. S. & Varmus, H. A New Initiative on Precision Medicine. *N. Engl. J. Med.* **372**, 793–795 (2015).
2. Kohane, I. S. HEALTH CARE POLICY. Ten things we have to do to achieve precision medicine. *Science* **349**, 37–8 (2015).
3. Dearment, A. How best to describe precision medicine beyond oncology? It's complicated. *MedCityNews* (2019). Available at: https://medcitynews.com/2019/07/how-best-to-describe-precision-medicine-beyond-oncology-its-complicated/?rf=1. (Accessed: 6th December 2020)
4. Gameiro, G., Sinkunas, V., Liguori, G. & Auler-Júnior, J. Precision Medicine: Changing the way we think about healthcare. *Clinics* **73**, e723 (2018).
5. National Research Council. *Toward Precision Medicine*. (National Academies Press, 2011). doi:10.17226/13284
6. Huang, S. & Hood, L. Personalized, Precision, and N-of-One Medicine: A Clarification of Terminology and Concepts. *Perspect. Biol. Med.* **62**, 617–639 (2019).
7. Hawgood, S., Hook-Barnard, I. G., O'Brien, T. C. & Yamamoto, K. R. Precision medicine: Beyond the inflection point. *Sci. Transl. Med.* **7**, 300ps17-300ps17 (2015).
8. Wang, Z.-G., Zhang, L. & Zhao, W.-J. Definition and application of precision medicine. *Chinese J. Traumatol.* **19**, 249–250 (2016).
9. Poisson, M., Dulong, Larrey & Double. Statistical research on conditions caused by calculi by Doctor Civiale. *Int. J. Epidemiol.* **30**, 1246–1249 (2001).
10. Double, F. J. Recherches de statistiques sur l'affection calculeuse par M. le docteur Civiale. *Comtes Rendus l'Acad,mie des Sci.* 167–167 (1835).
11. Bayer, R. & Galea, S. Public Health in the Precision-Medicine Era. *N. Engl. J. Med.* **373**, 499–501 (2015).
12. Schork, N. J. Personalized medicine: Time for one-person trials. *Nature* **520**, 609–611 (2015).
13. Shamseer, L. *et al.* CONSORT extension for reporting N-of-1 trials (CENT) 2015: Explanation and elaboration. *BMJ* **350**, h1793 (2015).
14. ICH Expert Working Group. ICH HARMONISED TRIPARTITE GUIDELINE STATISTICAL PRINCIPLES FOR CLINICAL TRIALS E9 STATISTICAL PRINCIPLES FOR CLINICAL TRIALS. (1998).
15. Tannock, I. F. & Hickman, J. A. Limits to Personalized Cancer Medicine. *N. Engl. J. Med.* **375**, 1289–1294 (2016).
16. Le Tourneau, C. *et al.* Molecularly targeted therapy based on tumour molecular profiling versus conventional therapy for advanced cancer (SHIVA): a multicentre, open-label, proof-of-concept, randomised, controlled phase 2 trial. *Lancet Oncol.* **16**, 1324–1334 (2015).
17. Schork, N. J. Personalized medicine: Time for one-person trials. *Nature* **520**, 609–611 (2015).

18. Willis, J. C. D. & Lord, G. M. Immune biomarkers: the promises and pitfalls of personalized medicine. *Nat. Rev. Immunol.* **15**, 323–329 (2015).

19. Wallach, J. D. *et al.* Evaluation of Evidence of Statistical Support and Corroboration of Subgroup Claims in Randomized Clinical Trials. *JAMA Intern. Med.* **177**, 554–560 (2017).

20. Sathi, P., Kalyan, S., Hitchcock, C. L., Pudek, M. & Prior, J. C. Progesterone therapy increases free thyroxine levels-data from a randomized placebo-controlled 12-week hot flush trial. *Clin. Endocrinol. (Oxf).* **79**, 282–287 (2013).

21. Durán-Cantolla, J. *et al.* Continuous positive airway pressure as treatment for systemic hypertension in people with obstructive sleep apnoea: randomised controlled trial. *BMJ* **341**, c5991 (2010).

22. Kojima, Y. *et al.* Comparison between sitagliptin and nateglinide on postprandial lipid levels: The STANDARD study. *World J. Diabetes* **4**, 8 (2013).

23. Araujo, A., Julious, S. & Senn, S. Understanding Variation in Sets of N-of-1 Trials. *PLoS One* **11**, e0167167 (2016).

24. Senn, S. Individual response to treatment: is it a valid assumption? *BMJ* **329**, 966–8 (2004).

25. Senn, S. Mastering variation: variance components and personalised medicine. *Stat. Med.* **35**, 966–977 (2016).

26. Wang, R., Lagakos, S. W., Ware, J. H., Hunter, D. J. & Drazen, J. M. Statistics in Medicine — Reporting of Subgroup Analyses in Clinical Trials. *N. Engl. J. Med.* **357**, 2189–2194 (2007).

27. Senn, S. & Richardson, W. The first t-test. *Stat. Med.* **13**, 785–803 (1994).

28. Kim, S. H. *et al.* PTSD Symptom Reduction With Mindfulness-Based Stretching and Deep Breathing Exercise: Randomized Controlled Clinical Trial of Efficacy. *J. Clin. Endocrinol. Metab.* **98**, 2984–2992 (2013).

29. Cortés, J. Shiny-app: Does evidence support the high expectations placed in precision medicine? A review. (2017). Available at: http://shiny-eio.upc.edu/pubs/F1000_precision_medicine/. (Accessed: 30th January 2020)

30. Cortés, J. Dataset: review_homoscedasticity_clinical_trials.csv. (2017). doi:10.6084/m9.figshare.5552656

31. Riley, R. D., Higgins, J. P. T. & Deeks, J. J. Interpretation of random effects meta-analyses. *Bmj-British Med. J.* **342**, 964–967 (2011).

32. DerSimonian, R. & Kacker, R. Random-effects model for meta-analysis of clinical trials: An update. *Contemp. Clin. Trials* **28**, 105–114 (2007).

33. Senn, S. *Statistical issues in drug development*. (Wiley, 1997).

34. Bartlett, M. S. & Kendall, D. G. The Statistical Analysis of Variance-Heterogeneity and the Logarithmic Transformation. *Suppl. to J. R. Stat. Soc.* **8**, 128 (1946).

35. Mancl, L. A. & DeRouen, T. A. A covariance estimator for GEE with improved small-sample properties. *Biometrics* **57**, 126–34 (2001).

36. Kaplan, E. L. Tensor Notation and the Sampling Cumulants of k-Statistics. *Biometrika* **39**, 319 (1952).

37. Muirhead, R. J. *Aspects of Multivariate Statistical Theory*. (John Wiley & Sons, Ltd., 1982). doi:10.2307/2987858

38. Begg, C. B. & Mazumdar, M. Operating Characteristics of a Rank Correlation Test for Publication Bias. *Biometrics* **50**, 1088 (1994).

39. Burnham, K. P. & Efron, B. The Jackknife, the Bootstrap and Other Resampling Plans. *Biometrics* **39**, 816 (1983).

40. Wikipedia. *Jacknife resampling*. *Wikipedia* doi:10.1002/9780470057339.vaj001

41. Michael, Y. L. *et al.* Hormone therapy and physical function change among older women in the Women's Health Initiative. *Menopause* **17**, 295–302 (2010).

42. McManus, R. J. *et al.* Telemonitoring and self-management in the control of hypertension (TASMINH2): a randomised controlled trial. *Lancet* **376**, 163–172 (2010).

43. Redondo, J. R. *et al.* Long-term efficacy of therapy in patients with fibromyalgia: A physical exercise-based program and a cognitive-behavioral approach. *Arthritis Care Res. (Hoboken).* **51**, 184–192 (2004).

44. Caballero, A. E. *et al.* The differential effects of metformin on markers of endothelial activation and inflammation in subjects with impaired glucose tolerance: A placebo-controlled, randomized clinical trial. *J. Clin. Endocrinol. Metab.* **89**, 3943–3948 (2004).

45. Chen, C.-H. *et al.* Effects of Adjunctive Metformin on Metabolic Traits in Nondiabetic Clozapine-Treated Patients With Schizophrenia and the Effect of Metformin Discontinuation on Body Weight. *J. Clin. Psychiatry* **74**, e424–e430 (2013).

46. Luty, S. E. *et al.* Randomised controlled trial of interpersonal psychotherapy and cognitive–behavioural therapy for depression. *Br. J. Psychiatry* **190**, 496–502 (2007).

47. Rochette, A. *et al.* The YOU CALL-WE CALL randomized clinical trial impact of a multimodal support intervention after a mild stroke. *Circ. Cardiovasc. Qual. Outcomes* **6**, 674–679 (2013).

48. Joos, S. *et al.* Acupuncture and Moxibustion in the Treatment of Active Crohn's Disease: A Randomized Controlled Study. *Digestion* **69**, 131–139 (2004).

49. Zhang, G., Qin, L. & Shi, Y. Epimedium-Derived Phytoestrogen Flavonoids Exert Beneficial Effect on Preventing Bone Loss in Late Postmenopausal Women: A 24-Month Randomized, Double-Blind and Placebo-Controlled Trial. *J. Bone Miner. Res.* **22**, 1072–1079 (2007).

50. Albert-Kiszely, A. *et al.* Comparison of the effects of cetylpyridinium chloride with an essential oil mouth rinse on dental plaque and gingivitis ? a six-month randomized controlled clinical trial. *J. Clin. Periodontol.* **34**, 658–667 (2007).

51. Fitzgerald, D., Trakarnratanakul, N., Smyth, B. & Caulfield, B. Effects of a Wobble Board-Based Therapeutic Exergaming System for Balance Training on Dynamic Postural Stability and Intrinsic Motivation Levels. *J. Orthop. Sport. Phys. Ther.* **40**, 11–19 (2010).

52. Francis, B. A., Du, L. T., Berke, S., Ehrenhaus, M. & Minckler, D. S. Comparing the fixed combination dorzolamide-timolol (CosoptR) to concomitant administration of 2% dorzolamide (TrusoptR) and 0.5% timolol - a randomized controlled trial and a replacement study. *J. Clin. Pharm. Ther.* **29**, 375–380 (2004).

53. Heimann, H. *et al.* Scleral Buckling versus Primary Vitrectomy in Rhegmatogenous Retinal Detachment. *Ophthalmology* **114**, 2142-2154.e4 (2007).

54. Senn, S. Controversies concerning randomization and additivity in clinical trials. *Stat. Med.* **23**, 3729–3753 (2004).

55. Jamieson, J. Measurement of Change and the Law of Initial Values: A Computer Simulation Study. *Educ. Psychol. Meas.* **55**, 38–46 (1995).

56. Senn, S. Trying to be precise about vagueness. *Stat. Med.* **26**, 1417–1430 (2007).

57. Greenlaw & Nicola. Constructing Appropriate Models for Meta-Analyses. (University of Glasgow, 2009).

58. Carlisle, J. B. Data fabrication and other reasons for non-random sampling in 5087 randomised, controlled trials in anaesthetic and general medical journals. *Anaesthesia* **72**, 944–952 (2017).

59. Hsieh, L. L.-C., Kuo, C.-H., Yen, M.-F. & Chen, T. H.-H. A randomized controlled clinical trial for low back pain treated by acupressure and physical therapy. *Prev. Med. (Baltim).* **39**, 168–176 (2004).

60. Sachs, L. *Applied statistics : a handbook of techniques*. (Springer-Verlag, 1982).

61. Moe, S. M., Saifullah, A., LaClair, R. E., Usman, S. A. & Yu, Z. A Randomized Trial of Cholecalciferol versus Doxercalciferol for Lowering Parathyroid Hormone in Chronic Kidney Disease. *Clin. J. Am. Soc. Nephrol.* **5**, 299–306 (2010).

62. Wöhrle, J. *et al.* Results of Intracoronary Stem Cell Therapy After Acute Myocardial Infarction. *Am. J. Cardiol.* **105**, 804–812 (2010).

63. Francetti, L., Del Fabbro, M., Basso, M., Testori, T. & Weinstein, R. Enamel matrix proteins in the treatment of intra-bony defects. A prospective 24-month clinical trial. *J. Clin. Periodontol.* **31**, 52–9 (2004).

64. Bennell, K. L. *et al.* Hip strengthening reduces symptoms but not knee load in people with medial knee osteoarthritis and varus malalignment: A randomised controlled trial. *Osteoarthr. Cartil.* **18**, 621–628 (2010).

65. Wasserstein, R. L. & Lazar, N. A. The ASA Statement on p -Values: Context, Process, and Purpose. *Am. Stat.* **70**, 129–133 (2016).

66. Pounds, S. & Morris, S. W. Estimating the occurrence of false positives and false negatives in microarray studies by approximating and partitioning the empirical distribution of p-values. *Bioinformatics* **19**, 1236–42 (2003).

67. Madsen, K., Nielsen, H. & Tingleff, O. *Optimization with constraints*. *IMM, Technical University of …* (2004).

68. Moher, D. *et al.* CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. *BMJ* **340**, (2010).

69. ROTHMAN, K. J., GREENLAND, S. & WALKER, A. M. CONCEPTS OF INTERACTION. *Am. J. Epidemiol.* **112**, 467–470 (1980).

70. Cobo, E. *et al.* Effect of using reporting guidelines during peer review on quality of final manuscripts submitted to a biomedical journal : masked randomised trial. *Bmj* **6783**, 1–11 (2011).

71. Winkelbeiner, S., Leucht, S., Kane, J. M. & Homan, P. Evaluation of Differences in Individual Treatment Response in Schizophrenia Spectrum Disorders. *JAMA Psychiatry* **76**, 1063 (2019).

72. Munkholm, K., Winkelbeiner, S. & Homan, P. Individual response to antidepressants for depression in adults – a simulation study and meta-analysis. *Psyarxiv* (2019). doi:10.31234/osf.io/m4aqc

73. Plöderl, M. & Hengartner, M. P. What are the chances for personalised treatment with antidepressants? Detection of patient-by-treatment interaction with a variance ratio meta-analysis. *BMJ Open* **9**, e034816 (2019).

74. Senior, A. M., Viechtbauer, W. & Nakagawa, S. Revisiting and expanding the meta-analysis of variation: The log coefficient of variation ratio. *Res. Synth. Methods* **11**, 553–567 (2020).

75. Mills, H. L. *et al.* Detecting heterogeneity of intervention effects using analysis and meta-analysis of differences in variance between arms of a trial. *medRxiv* 1–29 (2020). doi:10.1101/2020.03.07.20032516

76. Radua, J., Davies, C. & Fusar-Poli, P. Evaluation of variability in individual response to treatments in the clinical high-risk state for psychosis: A meta-analysis. *Schizophr. Res.* **227**, 20–27 (2021).

77. Winkelbeiner, S. *et al.* Treatment effect variation in brain stimulation across psychiatric disorders. *medRxiv* 1–12 (2020). doi:10.1101/2020.05.02.20088831

78. Watson, J. A. *et al.* Inter-Individual Differences in the Responses to Pain Neuroscience Education in Adults With Chronic Musculoskeletal Pain: A Systematic Review and Meta-Analysis of Randomized Controlled Trials. *J. Pain* **22**, 9–20 (2021).

79. Neumeier, M. S. *et al.* Examining side effect variation of antipsychotic treatment in schizophrenia spectrum disorders. *medRxiv* (2020). doi:10.1101/2020.07.27.20162727

80. Smith, E. S., Smith, H. A., Betts, J. A., Gonzalez, J. T. & Atkinson, G. A Systematic Review and Meta-Analysis Comparing Heterogeneity in Body Mass Responses Between Low-Carbohydrate and Low-Fat Diets. *Obesity* **28**, 1833–1842 (2020).

81. Winkelbeiner, S. & Homan, P. Is Variance Ratio a Valid Indicator of Heterogeneous Treatment Effect?—Reply. *JAMA Psychiatry* **77**, 217–218 (2019).

82. Atkinson, G., Williamson, P. & Batterham, A. M. Issues in the determination of 'responders' and 'non-responders' in physiological research. *Exp. Physiol.* **104**, 1215–1225 (2019).

83. Homan, P. & Kane, J. M. Clozapine as an early-stage treatment. *Acta Psychiatr. Scand.* **138**, 279–280 (2018).

84. Homan, P. *et al.* Structural similarity networks predict clinical outcome in early-phase psychosis. *Neuropsychopharmacology* **44**, 915–922 (2019).

85. Feczko, E. & Fair, D. A. Methods and Challenges for Assessing Heterogeneity. *Biol. Psychiatry* **88**, 9–17 (2020).

86. Hieronymus, F., Hieronymus, M., Nilsson, S., Eriksson, E. & Østergaard, S. D. Individual variability in treatment response to antidepressants in major depression: comparing trial-level and patient-level analyses. *Acta Psychiatr. Scand.* **142**, 443–445 (2020).

87. Sandu, M. R. *et al.* A novel application of two-step Mendelian randomization: applying the results of small feasibility studies of interventions to infer causal effects on clinical endpoints. *Br. J. cancer. Conf. 2018 Natl. cancer Res. Inst. cancer Conf. NCRI 2018. United kingdom* **119**, 34-35 (2018).

88. Hartwig, F. P., Bowden, J., Wang, L., Smith, G. D. & Davies, N. M. Average causal effect estimation via instrumental variables: the no simultaneous heterogeneity assumption. *arXiv* (2020).

89. Bowden, J., Bornkamp, B., Glimm, E. & Bretz, F. Connecting Instrumental Variable methods for causal inference to the Estimand Framework. *arXiv* 1–28 (2020).

90. Aron, D. C. *Complex Systems in Medicine*. *Complex Systems in Medicine* (2020). doi:10.1007/978-3-030-24593-1

91. Senn, S. Statistical pitfalls of personalized medicine. *Nature* **563**, 619–621 (2018).

92. Boutron, I., Altman, D. G., Moher, D., Schulz, K. F. & Ravaud, P. CONSORT Statement for Randomized Trials of Nonpharmacologic Treatments: A 2017 Update and a CONSORT Extension for Nonpharmacologic Trial Abstracts. *Ann. Intern. Med.* **167**, 40 (2017).

93. Caughey, D., Dafoe, A. & Miratrix, L. Beyond the Sharp Null: Randomization Inference, Bounded Null Hypotheses, and Confidence Intervals for Maximum Effects. 1–37 (2017).

94. Fisher, R. A. *The Design of Experiments*. (1935).

95. Sharing, T. I. C. of I. for F. in T. D. Toward Fairness in Data Sharing. *N. Engl. J. Med.* **375**, 405–407 (2016).

96. Taichman, D. B. *et al.* Sharing Clinical Trial Data — A Proposal from the International Committee of Medical Journal Editors. *N. Engl. J. Med.* **374**, 384–386 (2016).

97. Krumholz, H. M. & Waldstreicher, J. The Yale Open Data Access (YODA) Project — A Mechanism for Data Sharing. *N. Engl. J. Med.* **375**, 403–405 (2016).

98. Ioannidis, J. P. A. *et al.* Increasing value and reducing waste in research design, conduct, and analysis. *Lancet (London, England)* **383**, 166–75 (2014).

99. Cortés, J. R code for analysis of homoscedasticity in clinical trials. (2017). doi:10.5281/zenodo.1239539

100. Barone, M. A. *et al.* Safety of tubal ligation by minilaparotomy provided by clinical officers versus assistant medical officers: study protocol for a noninferiority randomized controlled trial in Tanzanian women. *Trials* **18**, 499 (2017).

101. Gal, R. *et al.* The effects of exercise on the quality of life of patients with breast cancer (the UMBRELLA Fit study): study protocol for a randomized controlled trial. *Trials* **18**, 504 (2017).

102. Graham, H. R. *et al.* Improving oxygen therapy for children and neonates in secondary hospitals in Nigeria: study protocol for a stepped-wedge cluster randomised trial. *Trials* **18**, 502 (2017).

103. Jakkula, P. *et al.* Targeting low- or high-normal Carbon dioxide, Oxygen, and Mean arterial pressure After Cardiac Arrest and REsuscitation: study protocol for a randomized pilot trial. *Trials* **18**, 507 (2017).

104. Jefford, M. *et al.* SCORE: Shared care of Colorectal cancer survivors: protocol for a

randomised controlled trial. *Trials* **18**, 506 (2017).

105. Park, J.-B. *et al.* Long-term Effects of high-doSe pitavaStatin on Diabetogenicity in comparison with atorvastatin in patients with Metabolic syndrome (LESS-DM): study protocol for a randomized controlled trial. *Trials* **18**, 501 (2017).

106. Salaminios, G. *et al.* A randomised controlled trial assessing the severity and duration of depressive symptoms associated with a clinically significant response to sertraline versus placebo, in people presenting to primary care with depression (PANDA trial): study protocol for. *Trials* **18**, 496 (2017).

107. Shepperd, S. *et al.* A multi-centre randomised trial to compare the effectiveness of geriatrician-led admission avoidance hospital at home versus inpatient admission. *Trials* **18**, 491 (2017).

108. Tamura, T., Hayashida, K., Sano, M., Onuki, S. & Suzuki, M. Efficacy of inhaled HYdrogen on neurological outcome following BRain Ischemia During post-cardiac arrest care (HYBRID II trial): study protocol for a randomized controlled trial. *Trials* **18**, 488 (2017).

109. Wu, W. *et al.* Neoadjuvant everolimus plus letrozole versus fluorouracil, epirubicin and cyclophosphamide for ER-positive, HER2-negative breast cancer: study protocol for a randomized pilot trial. *Trials* **18**, 497 (2017).

110. Anderson, C. S. *et al.* Cluster-Randomized, Crossover Trial of Head Positioning in Acute Stroke. *N. Engl. J. Med.* **376**, 2437–2447 (2017).

111. Brunoni, A. R. *et al.* Trial of Electrical Direct-Current Therapy versus Escitalopram for Depression. *N. Engl. J. Med.* **376**, 2523–2533 (2017).

112. Carbone, D. P. *et al.* First-Line Nivolumab in Stage IV or Recurrent Non–Small-Cell Lung Cancer. *N. Engl. J. Med.* **376**, 2415–2426 (2017).

113. Daum, R. S. *et al.* A Placebo-Controlled Trial of Antibiotics for Smaller Skin Abscesses. *N. Engl. J. Med.* **376**, 2545–2555 (2017).

114. Dunkle, L. M. *et al.* Efficacy of Recombinant Influenza Vaccine in Adults 50 Years of Age or Older. *N. Engl. J. Med.* **376**, 2427–2436 (2017).

115. Nair, P. *et al.* Oral Glucocorticoid–Sparing Effect of Benralizumab in Severe Asthma. *N. Engl. J. Med.* **376**, 2448–2458 (2017).

116. Stott, D. J. *et al.* Thyroid Hormone Therapy for Older Adults with Subclinical Hypothyroidism. *N. Engl. J. Med.* **376**, 2534–2544 (2017).

117. Taylor, H. S. *et al.* Treatment of Endometriosis-Associated Pain with Elagolix, an Oral GnRH Antagonist. *N. Engl. J. Med.* **377**, 28–40 (2017).

118. von Minckwitz, G. *et al.* Adjuvant Pertuzumab and Trastuzumab in Early HER2-Positive Breast Cancer. *N. Engl. J. Med.* **377**, 122–131 (2017).

119. Wilt, T. J. *et al.* Follow-up of Prostatectomy versus Observation for Early Prostate Cancer. *N. Engl. J. Med.* **377**, 132–142 (2017).

120. Chow, S.-C., Wang, H. & Shao, J. *Sample size calculations in clinical research*. (Chapman & Hall_CRC, 2008).

121. Singer, J. A simple procedure to compute the sample size needed to compare two

independent groups when the population variances are unequal. *Stat. Med.* **20**, 1089–1095 (2001).

122. Yancy, W. S., Olsen, M. K., Dudley, T. & Westman, E. C. Acid-base analysis of individuals following two weight loss diets. *Eur. J. Clin. Nutr.* **61**, 1416–1422 (2007).

123. McManus, R. J. *et al.* Telemonitoring and self-management in the control of hypertension (TASMINH2): A randomised controlled trial. *Lancet* **376**, 163–172 (2010).