# Planung multizentrischer randomisierter klinischer Studien mit kontinuierlichem Endpunkt

Dissertation
zur Erlangung des humanwissenschaftlichen Doktorgrades
in der Medizin
der Georg-August-Universität Göttingen

vorgelegt von
Markus Harden
aus Wilhelmshaven

Göttingen, 2020

**Betreuungsausschuss**

**Erstbetreuer:**

Professor Dr. Tim Friede, (Gutachter)
Institut für Medizinische Statistik, Universitätsmedizin Göttingen

**Weitere Betreuer/innen:**

Professor Dr. Heike Bickeböller, (Gutachterin)
Institut für Genetische Epidemiologie, Universitätsmedizin Göttingen

Professor Dr. Thomas Kneib,
Professur für Statistik und Ökonometrie, Georg-August-Universität Göttingen

Professor Dr. Jürgen Brockmöller,
Institut für Klinische Pharmakologie, Universitätsmedizin Göttingen

**Weitere Mitglieder der Prüfungskommission:**

Professor Dr. Markus Zabel,
Klinik für Kardiologie und Pneumologie, Universitätsmedizin Göttingen

Professor Dr. Thomas Meyer,
Klinik für Psychosomatische Medizin und Psychotherapie, Universitätsmedizin Göttingen

**Tag der mündlichen Prüfung:** 13. März 2020

"To consult a statistician after an experiment is finished is often merely to ask him to conduct a post mortem examination. He can perhaps say what the experiment died of."[1]

RA Fisher (1890 – 1962)

[1] Fisher, R. A. "Presidential Address." Sankhya: The Indian Journal of Statistics (1933-1960), vol. 4, no. 1, 1938, pp. 14–17. JSTOR, www.jstor.org/stable/40383882.

# Erklärung

Hiermit erkläre ich, Markus Harden, die Dissertation mit dem Titel „Planung multizentrischer randomisierter klinischer Studien mit kontinuierlichem Endpunkt“ eigenständig angefertigt und keine anderen als die von mir angegebenen Quellen und Hilfsmittel verwendet zu haben.

Göttingen, den 17. Januar 2020, Markus Harden

# Zusammenfassung

Multizentrische kontrollierte randomisierte klinische Studien sind ein Grundpfeiler der modernen evidenzbasierten Medizin. Die Vorteile der Datenerfassung an mehreren Standorten sind zahlreich, einschließlich einer beschleunigten Rekrutierung und einer besseren Verallgemeinerbarkeit der Ergebnisse. Große konfirmatorische Studien können meist nur im Rahmen eines multizentrischen Studiendesigns realisiert werden, da einzelne Zentren häufig nur kleine Fallzahlen beisteuern können. Trotz erhöhter Kosten und eines großen Bedarfs an Koordination und Standardisierung im Vergleich zu monozentrischen Studien, nimmt die Anzahl an multizentrischen Studien stetig zu. Obwohl sich gemischte lineare Modelle sehr gut eignen, um Cluster-korrelierte Daten auszuwerten, wird die Struktur einer multizentrischen Studie bei der statistischen Planung häufig nicht ausreichend berücksichtigt. Wenn aufgrund einer fehlerhaften Planung eine zu kleine Fallzahl rekrutiert wird, kann dies das Scheitern der Studie zur Folge haben. Im pharmazeutischen Kontext würde dies mit immensen ökonomischen Einbußen einhergehen, zum Beispiel wenn der Zulassungsprozess eines neuen Wirkstoffes von dem Erfolg dieser Studie abhängt. Auch aus ethischer Sicht ist eine zu kleine Fallzahl nicht vertretbar, da Patienten eventuell ein wirksames Medikament vorenthalten wird.

Das Hauptaugenmerk dieser Dissertation liegt auf der Fallzahlplanung von multizentrischen Studien, bei denen zwei Behandlungsgruppen mit einem kontinuierlichen Endpunkt durch gemischte lineare Modelle miteinander verglichen werden. Obwohl in der wissenschaftlichen Literatur bereits Methoden vorgestellt wurden, um eine solche multizentrische Studie zu planen, gehen diese von sehr restriktiven und in der Anwendung unrealistischen Annahmen bezüglich der Randomisierung aus. Das erste Ziel dieser Arbeit war es daher, eine Fallzahlformel zu entwickeln, die weniger strenge Annahmen an das statistische Modell stellt. Ich habe gezeigt, dass man eine Fallzahlplanung für multizentrische Studien mit beliebigen Stichprobengrößen durchführen kann, falls eine Blockrandomisierung für die Allokation der Probanden verwendet wird, was ein sehr gängiges Randomisierungsverfahren ist. Insbesondere habe ich eine untere und obere Schranke für die geschätzte Fallzahl angegeben und in Simulationsstudien gezeigt, dass mit diesem Ansatz die geplante statistische Power erreicht wird. Dadurch wird die Planung von Studien zur Identifizierung von neuen und

wirksamen Therapien verbessert.
Das zweite Ziel dieser Arbeit war die Übertragung der neu entwickelten Fallzahlformel auf Studiendesigns mit interner Pilotstudie zur Fallzahlrekalkulation. Dieses Ziel war motiviert durch die Unsicherheit, die bei der Fallzahlplanung einer multizentrischen Studie, insbesondere durch einen zusätzlichen Varianzparameter, besteht. Ich habe gezeigt, dass man die Fallzahlformel bei adaptiven Studiendesigns mit Fallzahlrekalkulation anwenden kann und dass Fehlannahmen bei der initialen Fallzahlplanung durch eine Rekalkulation der Varianzparameter korrigiert werden können, so dass die geplante statistische Power erreicht wird.

# Abstract

Multicentre controlled randomized clinical trials are a cornerstone of modern evidence-based medicine. The benefits of collecting data from more than one centre are numerous, including accelerated recruitment and better generalizability of results. Large confirmatory trials often rely on multicentre study designs, since most centres are limited to small sample sizes. Despite increasing costs and requirements for coordination and standardization compared to single-centre studies, the number of multicentre trials is steadily increasing. Although linear mixed effects models are very well suited to analyze cluster-correlated data, this structure is often barely accounted for when planning such a multicentre trial. If too few subjects are recruited due to incorrect assumptions at the planning stage, this may lead to the failure of the trial. In the pharmaceutical context, this would be accompanied by immense economic losses, for example if the approval process of a new drug depends on the success of this study. Also from an ethical point of view, a too small sample size is not justifiable, as patients may be deprived of an effective treatment.

This dissertation focuses on the sample size calculation for multicentre trials in which two treatment groups with a continuous endpoint are compared using linear mixed effects models. Although methods to plan such a multicentre trial have already been proposed, they assume very restrictive and unrealistic assumptions regarding treatment randomization. The first objective of this thesis was therefore to develop a sample size formula that makes less strict assumptions about the statistical model. I demonstrated that sample size calculation can be performed for multicentre trials for arbitrary sample sizes, if block-randomization is used for the allocation of subjects, which is a well established randomization technique. In particular, I derived lower and upper boundaries for the calculated sample size and showed in simulation studies that this approach achieves the planned statistical power. This improves the planning of studies to identify new and effective therapies.

The second objective of this work was to apply the newly developed sample size formula to study designs with an internal pilot study for sample size recalculation. This goal was motivated by the uncertainty in sample size planning, especially in multicentre trials which consider an additional nuisance parameter. I have shown that the sample size formula can be applied to adaptive study designs with sample size recalculation. I performed simulation

studies to show that false assumptions regarding the initial sample size calculation can be corrected by recalculating the sample size based on nuisance parameter estimates and that the initially targeted statistical power is achieved.

# Danksagung

Während meiner Arbeit an dieser Dissertation habe ich von vielen Seiten wertvolle Ratschläge, Motivation und Unterstützung erhalten und möchte mich an dieser Stelle dafür bedanken.
Ich danke Herrn Professor Friede für die Überlassung des Themas und die guten Arbeitsbedingungen die ich im Institut für Medizinische Statistik für die Erstellung dieser Arbeit vorgefunden habe. Frau Professor Bickeböller danke ich für die Übernahme des Co-Referates. Herrn Professor Kneib und Herrn Professor Brockmöller danke ich für ihre Unterstützung im Rahmen des Betreuungskomitees.
Meinen Kollegen aus dem Institut für Medizinische Statistik danke ich für das freundliche Umfeld, in dem ich mich die letzten Jahre bewegen und weiterentwickeln durfte. Insbesondere danke ich Christian und Tobias für das Korrekturlesen meiner Arbeit.
Schließlich danke ich meinen Eltern und meiner Schwester für die engelsgleiche Geduld mit der sie mich stets unterstützt und motiviert haben. Susanne danke ich dafür, endlich einen Grund zu haben meinen universitären Kokon zu verlassen.

# Inhaltsverzeichnis

# 1 Einleitung

## 1.1 Klinische Studien in der evidenzbasierten Medizin

Die randomisierte kontrollierte klinische Studie ist ein Grundbaustein der modernen evidenzbasierten Medizin. Ziel einer solchen Studie kann es sein, die Überlegenheit eines Medizinproduktes, die Nichtunterlegenheit einer neuen, günstigereren Behandlungsform oder auch die Bioäquivalenz eines Generikums bezüglich seiner Wirksamkeit und/oder Sicherheit im Vergleich zu bestehenden Standards aufzuzeigen [1].
Gemäß § 4, Absatz 23 im deutschen Arzneimittelgesetz (AMG) versteht man unter einer klinischen Prüfung bei Menschen „jede am Menschen durchgeführte Untersuchung, die dazu bestimmt ist, klinische oder pharmakologische Wirkungen von Arzneimitteln zu erforschen oder nachzuweisen oder Nebenwirkungen festzustellen oder die Resorption, die Verteilung, den Stoffwechsel oder die Ausscheidung zu untersuchen, mit dem Ziel, sich von der Unbedenklichkeit oder Wirksamkeit der Arzneimittel zu überzeugen“. Damit die aus dieser Prüfung gezogenen Schlüsse auch auf andere Personen übertragen werden können, bedarf es einiger Werkzeuge, die sicherstellen sollen, dass die gemessenen Studienergebnisse Unterschiede zwischen den Behandlungen und nicht etwa ungleiche Patientencharakteristika zu Beginn der Studie beschreiben.

### 1.1.1 Grundlegende Prinzipien klinischer Studien

Für aussagekräftige Ergebnisse sollte eine neue Therapie immer wenn möglich gegen eine Referenztherapie im Rahmen einer klinischen Studie verglichen werden, um den zusätzlichen Nutzen der Behandlung herauszustellen [2]. In diesem Fall spricht man von einer kontrollierten Studie. Die erste kontrollierte klinische Studie wird häufig dem Arzt James Lind zugeschrieben, der in seiner Abhandlung von 1757 beschreibt, wie er 1747 während seiner Zeit als Schiffsarzt auf der *HMS Salisbury* zwölf ähnlich schwer an Skorbut erkrankte Seemänner in sechs Gruppen aufteilte, mit verschiedenen Diäten behandelte, und schließlich die Wirksamkeit von Zitrusfrüchten zur Behandlung von Skorbut beobachtete [3].

Als Randomisierung bezeichnet man einen zufälligen Prozess, mit dem Patienten den verschiedenen Behandlungsgruppen zugeordnet werden. Die Randomisierung soll sicherstellen, dass die Behandlungsgruppen gleiche Ausgangsbedingungen zu Beginn der Studie aufweisen, so dass Unterschiede am Studienende auf die Therapie zurückgeführt werden können. Dazu werden die Probanden den Behandlungsgruppen zufällig, und damit unabhängig von Faktoren wie Alter, Erkrankungsgrad oder auch erwartetem Therapieerfolg, zugewiesen. Der Nutzen der Randomisierung wurde insbesondere von Jerzy Neyman und Ronald A. Fisher Anfang der 1920er Jahre beschrieben und rückte fortan verstärkt in den Fokus der Planung von Experimenten [4, 5]. Der wesentliche Vorteil der Randomisierung besteht darin, dass durch eine zufällige Allokation der Probanden eine Strukturgleichheit der Daten zu Beginn der Studie sowohl für beobachtete als auch unbeobachtete Einflussgrößen erreicht wird [6, Kapitel 3]. Die erste randomisierte kontrollierte klinische Studie wurde 1946 in Großbritannien durchgeführt und verfolgte das Ziel, die Wirksamkeit des Antibiotikums Streptomycin bei der Behandlung von Tuberkulose zu untersuchen [7]. Typische Verfahren zur Randomisierung sind die einfache Randomisierung (Münzwurf mit einer fairen Münze), die Blockrandomisierung, die stratifizierte Randomisierung und die adaptive Randomisierung [8]. Am häufigsten wird die Blockrandomisierung verwendet [9], auf die im Folgenden kurz eingegangen werden soll. Als Block oder Blocklänge bezeichnet man eine feste oder zufällige Anzahl an Probanden, für die gleichzeitig die Behandlungszugehörigkeit bestimmt wird, so dass ein vorgegebenes Allokationsverhältnis zwischen den Behandlungsgruppen erreicht wird. Das hat den Vorteil, dass die Behandlungsgruppen trotz Randomisierung gleichmäßig aufgefüllt werden und die Unbalanciertheit der Fallzahlen durch die Blocklänge beschränkt ist [6, Kapitel 3.5]. In der Praxis müssen die Patienten nicht gleichzeitig rekrutiert werden, die Behandlungszugehörigkeit ist allerdings schon im Vorfeld festgelegt und muss daher geheim gehalten werden. Als Nachteil der Blockrandomisierung kann daher angeführt werden, dass eine teilweise Entblindung der Behandlungszugehörigkeiten Rückschlüsse auf andere Probanden desselben Blockes zulässt [1].
Als weiteres Mittel, um das Risiko der bewussten wie unbewussten Ungleichbehandlung der verschiedenen Behandlungsgruppen zu minimieren, werden häufig alle in der Studie involvierten Personen der Studie verblindet. Ziel ist es mit Hilfe der Verblindung sowohl eine Behandlungs- als auch Beobachtungsgleichheit zu erreichen und damit das Risiko von Verzerrungen (englisch *bias*) der Studienergebnisse zu verhindern [10]. Man unterscheidet zwischen einfach- und doppelt- und sogar dreifach-blinden Studien, in denen nur den Probanden oder allen an der Studie beteiligten Personen (Teilnehmer, Behandelnde, Auswertende, Studienleitung) die Gruppenzugehörigkeit der Probanden vorenthalten wird [1]. Eine Entblindung der Daten erfolgt in der Regel für die finale Auswertung.

### 1.1.2 Multizentrische klinische Studien

Eine multizentrische Studie zeichnet sich dadurch aus, dass die Rekrutierung und Behandlung der Probanden an mehreren Zentren erfolgt, während die Behandlung für alle Zentren durch dasselbe Studienprotokoll standardisiert ist. Durch die Rekrutierung an mehreren Standorten kann die Rekrutierungsgeschwindigkeit erhöht beziehungsweise die Rekrutierung ausreichend vieler Probanden z. B. bei einer seltenen Erkrankung womöglich überhaupt erst ermöglicht werden [1, 11]. Außerdem ist die Verallgemeinerung multizentrischer Studienergebnisse im Vergleich zu denen einer monozentrischen Studie erleichtert, da eine Implementierung des Studienprotokolls bereits an mehreren Standorten durchgeführt wurde und die Studienpopulation heterogener wird. Allen Standardisierungen und Vorkehrungen zum Trotz könnten individuelle Zentren unbeabsichtigt systematisch unterschiedliche Daten hervorbringen. Dies kann beispielsweise auf unterschiedliche Patientenpopulationen oder auch Unterschiede in der klinischen Praxis zurückzuführen sein und sollte sowohl in der Planung als auch der Auswertung der Studie berücksichtigt werden [12, 13]. Das Zentrum kann aus statistischer Betrachtungsweise sowohl die Ausgangssituation der Probanden als auch den Behandlungsunterschied zwischen den Interventionen beeinflussen [14].

Als eine der ersten multizentrischen kontrollierten klinischen Studien gilt die Patulin-Studie aus dem Jahr 1944, in der die Wirksamkeit des Penicillium Patulum gegen Placebo bei der Behandlung erkälteter Arbeitnehmer untersucht wurde [15]. Dafür wurden an 14 Standorten $n_1 = 1449$ Probanden rekrutiert, von denen $668$ mit Patulin und $680$ mit einem Placebo behandelt wurden und in die Auswertung eingingen. Die Ergebnisse der Studie sind in Tabelle 1.1 dargestellt.

Tabelle 1.1: Publizierte Ergebnisse der Patulinstudie aus dem Jahr 1944. Für die Behandlungsgruppen sind die Raten an geheilten Patienten oder solche mit Verbesserung dargestellt. Die Differenz dieser Raten ist als Mittelwert $\pm$ Standardabweichung aufgeführt.

| | % geheilt | | | % geheilt oder verbessert | | |
|---|---|---|---|---|---|---|
| Behandlung | 24 Std. | 48 Std. | 1 Woche | 24 Std. | 48 Std. | 1 Woche |
| Patulin ($n = 668$) | 1,6 | 13 | 33 | 59 | 73 | 63 |
| Kontrolle ($n = 680$) | 1,2 | 13 | 37 | 64 | 77 | 69 |
| Differenz | 0,4 | 0 $\pm$ 1,9 | -4 $\pm$ 2,8 | -5 $\pm$ 2,7 | -4 $\pm$ 2,5 | -6 $\pm$ 2,8 |

In dem Artikel wurde ein statistisch signifikanter Gruppenunterschied zu einen Signifikanzniveau von $\alpha = 5\%$ nach einer Woche beschrieben.

Die Anzahl oder zumindest die Sichtbarkeit multizentrischer klinischer Studien ist seit jeher stetig gewachsen, wie man anhand von Abbildung 1.1 sehen kann. Dort ist exemplarisch die Anzahl an Studien aufgetragen, die in der PubMed-Datenbank hinterlegt sind und die Begriffe multizentrisch und Studie in englischer Sprache gemäß der folgenden Suchstrategie im Titel enthalten:
"(Multicentre[title] OR Multi-centre[title] OR Multicenter[title] OR Multi-center[title] ) AND trial[title]"

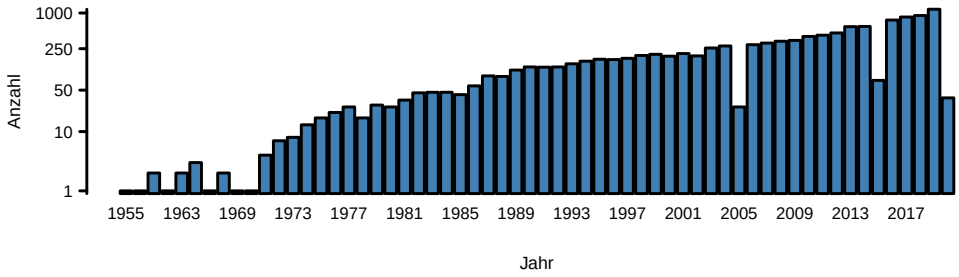


Abbildung 1.1: Anzahl an PubMed-Artikeln zu multizentrischen Studien.

### 1.1.3 Fallzahlplanung in klinischen Studien

Ein wichtiger Aspekt bei der Planung einer klinischen Studie ist die Berechnung der benötigten Fallzahl. Die zu untersuchende Fragestellung wird dazu mit Hilfe eines statistischen Modells in zwei komplementäre Hypothesen übersetzt, die durch Parameter dieses Modells formuliert werden können. Die sogenannte Nullhypothese ($H_0$) repräsentiert dabei den Status Quo, d.h. zum Beispiel die Annahme, dass sich die zu vergleichenden Therapien nicht unterscheiden oder die neue Therapie nicht besser ist als die Kontrolle. Die Alternativhypothese ($H_\mathrm{A}$) beschreibt definitionsgemäß das Gegenteil der Nullhypothese, also jenen Teil des Parameterraumes, in dem eine Ungleichheit der Behandlungsgruppen beziehungsweise eine Überlegenheit der neuen Therapie vorliegt.
Bei der Fallzahlplanung geht es um die Berechnung der benötigten Anzahl an Probanden, um die vorspezifizierte Alternativhypothese der Studie mit einer gewissen Wahrscheinlichkeit aufdecken zu können. Aus ethischer wie ökonomischer Sicht ist es notwendig eine möglichst genaue Einschätzung der Studiengröße vorzunehmen um (a) Patienten vor einer neuen aber unwirksamen Therapie zu schützen und (b) Patienten eine neue und überlegene Therapie nicht unnötig lange vorzuenthalten, beziehungsweise (a) Projekte ohne Erfolgsaussichten rechtzeitig zu beenden und (b) verheißungsvolle Projekte ausreichend zu fördern,

dass sie erfolgreiche sein können [1]. Die Fallzahlplanung basiert im Wesentlichen auf vier zu treffenden Annahmen:

- Wie groß darf die Wahrscheinlichkeit für einen Fehler 1. Art $(\alpha)$ sein?
- Wie groß soll die statistische Power $(1-\beta)$ zum Aufdecken der Alternativhypothese sein?
- Wie groß muss ein Behandlungsunterschied $(\mu^*)$ zwischen den Gruppen mindestens sein um als klinisch relevant betrachtet werden zu können?
- Wie groß ist die Variabilität $(\sigma^2)$ des Endpunkts innerhalb der Gruppen?

Wenn ein Zweigruppenvergleich von Verum $(V)$ gegen Placebo $(P)$ für einen kontinuierlichen Endpunkt mittels eines $t$-Tests angestellt werden soll, wird die Fallzahl $N$ durch die folgende Formel approximativ berechnet [16]

$$N \approx \frac{\sigma^2(k+1)^2}{k} \cdot \left(\frac{q_{1-\alpha/2} + q_{1-\beta}}{\mu^*}\right)^2 . \tag{1.1}$$

Neben den oben definierten Parametern beschreibt $k$ das angestrebte Allokationsverhältnis der Probanden in den beiden Behandlungsgruppen ($n_{\mathrm{V}} = k \cdot n_{\mathrm{P}}$ mit $N = n_{\mathrm{V}} + n_{\mathrm{P}}$) und $q_\gamma$ symbolisiert das $\gamma$-Quantil der Standardnormalverteilung. Abweichungen im Studiendesign können dazu führen, dass Formel (1.1) nicht zu der notwendigen statistischen Power führt, was den Erfolg der Studie gefährden kann. Eine Übersicht zu Fallzahlberechnungen für verschiedene Studiendesigns mit normalverteilten Endpunkten befindet sich in [17]. Auf dieser Grundlage kann nach erfolgreicher Datenerhebung untersucht werden, ob die Daten ausreichend Evidenz gegen $H_0$ und für $H_{\mathrm{A}}$ zeigen.

### 1.1.4 Fallzahlrekalkulation in klinischen Studien

Wie bereits in Kapitel 1.1.1 beschrieben, muss für jede klinische Studie vor der Datenerhebung ein Studienprotokoll erstellt werden, das neben anderen Informationen die benötigte Fallzahl spezifiziert. Da die Annahmen für die Fallzahlplanung auf bereits erfolgten Studien oder anderen, externen Abwägungen basieren, ist diese Fallzahl in der Regel mit Unsicherheit behaftet. Wenn die initial getroffenen Annahmen nicht zutreffen, kann dies das Scheitern der Studie zur Folge haben, auch wenn die neue Therapie der Kontrolle theoretisch überlegen wäre. Da für die Zulassung eines neuen Medikamentes in der Regel zwei positive Zulassungsstudien notwendig sind, kann das Scheitern einer zu klein geplanten Studie einen

substantiellen finanziellen Verlusten bedeuten oder aus der Sicht von Patienten, dass ein möglicherweise wirksames Medikament nicht auf den Markt kommt.
Um trotz eines vorspezifizierten Studienprotokolls eine gewisse Flexibilität des Studiendesigns zu ermöglichen und damit die Wahrscheinlichkeit eines richtig positiven Studienergebnisses zu erhöhen, wurden adaptive Studiendesigns entwickelt. Eine Übersicht zu den verschiedenen Formen adaptiver Studiendesigns findet man beispielsweise in [18, 19, 20, 21, 22]. Eine aus regulatorischer Sicht notwendige Forderung an adaptive Studiendesigns ist die Kontrolle des Fehlerniveaus und die möglichst unverzerrte Schätzung des Behandlungseffektes samt Konfidenzintervallen, wie in [23, 24] beschrieben. In dieser Arbeit beschäftige ich mich ausschließlich mit Fallzahlrekalkulationen basierend auf den Varianzkomponenten und nicht etwa der Neuberechnung des Behandlungseffektes im Rahmen der Fallzahlrekalkulation.
Die Fallzahlrekalkulation basierend auf der Neuberechnung der Varianzkomponenten erfolgt in der Regel in Form einer internen Pilotstudie nach Wittes und Brittain, die aus drei Schritten besteht [25]. Zunächst wird analog zu einem festen Studiendesign die initiale Fallzahl berechnet. Zusätzlich wird ein Zeitpunkt spezifiziert (z.B. wenn die Hälfte der initial geplanten Daten erhoben wurden), zu dem eine Neuberechnung der Fallzahl erfolgen soll. Diese Neuberechnung der Fallzahl basiert dann auf den initial getroffenen Annahmen zu den Fehlerniveaus $\alpha$, $\beta$, dem vermuteten Behandlungsunterschied $\mu^*$ und den aus den neuen Daten geschätzten Varianzparametern. Für die neu berechnete Fallzahl wurden verschiedene Restriktionen wie beispielsweise eine minimale oder maximale Fallzahl vorgeschlagen, um allzu starke Veränderungen durch die Fallzahlrekalkulation abzumildern [25, 26].
Man unterscheidet zwischen der Fallzahlrekalkulation auf Grundlage komparativer Daten, d.h. der Berücksichtigung der Gruppenzugehörigkeit, und nicht-komparativer Daten, was bedeutet, dass die Schätzung, beispielsweise der Varianzkomponenten, ohne Berücksichtigung der Behandlungszugehörigkeit erfolgt [24]. In früheren Artikeln und Guidelines wird häufig von Verfahren auf Grundlage entblindeter und verblindeter Daten gesprochen und beschreibt ebenfalls, ob Berechnungen mit oder ohne Kenntnis der Behandlungszugehörigkeit erfolgt sind. Da Berechnungen auf Grundlage komparativer Daten nicht zwingend eine Entblindung der Probanden alle Beteiligten zur Folge haben, oder dass Ergebnisse einer komparativen Analyse Teilnehmern, Personal oder Monitoren der Studie bekannt gemacht werden, wird diese sprachliche Unterscheidung in manchen Richtlinien vorgenommen [24]. In dieser Arbeit wird die Sicht des Statistikers auf die Daten eingenommen, daher werden die Begriffe nicht-komparativ und verblindet beziehungsweise komparativ und entblindet zum Teil synonym verwendet.

Adaptionen die auf nicht-komparativen Daten basieren, beeinflussen die Wahrscheinlichkeit für einen Fehler 1. Art einer Studie in der Regel nur unwesentlich, daher wird dieser Ansatz insbesondere von regulatorischer Seite favorisiert [23, 24]. Ein weiterer Vorteil nicht-komparativer Methoden besteht darin, dass zum Zeitpunkt der Fallzahlrekalkulation keine Daten entblindet werden müssen.

## 1.2 Motivierendes Beispiel: Die COMPETE II-Studie

Multizentrische Studien können in allen denkbaren Forschungsbereichen auftreten. Wir stellen als Beispiel eine Studie vor, die sich mit dem Krankheitsmanagement von Patienten mit Diabetes befasst.

Holbrook und Kollegen erhoben im Rahmen einer multizentrischen randomisierten Studie Daten, um die Wirksamkeit eines zusätzlichen Managementtools (Verum) bei der Behandlung von erwachsenen Patienten mit Diabetes zu untersuchen [27, 28]. Bei der neuen Behandlung handelte es sich um eine Ergänzung des lokalen Patientenorganisationsprogrammes (Kontrolle), das auf Grundlage der Patientenakte Zusammenfassungen, Empfehlungen und Erinnerungen sowohl für den behandelnden Arzt als auch Patienten erstellt, um die Qualität der fortlaufenden Betreuung zu verbessern. Das Programm war in die elektronischen Patienteninformationssysteme integriert und bot für die Patienten einen Online-Zugriff. Zusätzlich gab es ein automatisiertes telefonisches Erinnerungssystem und vierteljährliche Aufstellungen der Zusammenfassungen. Patienten in der Kontrollgruppe wurden ohne dieses zusätzliche Programm behandelt.

Der primäre Endpunkt der Studie war die Verbesserung eines Gesamtscores gegen den Ausgangswert zu Beginn der Studie. Dieser Score misst die Güte der Behandlung auf einer Skala von 0 bis 10 und basiert auf den folgenden Parametern: Blutdruck, Cholesterin, Hämoglobin, Fußkontrolle, Nierenfunktion, Gewicht, körperliche Aktivität und Rauchverhalten. Vor Beginn der Studie planten die Forscher 508 Patienten zu rekrutieren, um einen Behandlungsunterschied zwischen den Gruppen von einem Punkt mit einer statistischen Power von 80% und einem Signifikanzniveau von 5% mittels t-Test aufzudecken. Bezüglich der angenommenen Variabilität dieses Behandlungseffektes wurden in der Publikation keine Angaben gemacht. Ein Intraklassen-Korrelationskoeffizient von $\rho = 0,08$ wurde in einem frühen Stadium der Rekrutierung, basierend auf einer Teilmenge der behandelnden Zentren, berechnet. Schließlich wurden 511 Patienten in 46 Hausarztpraxen rekrutiert und lokal zufällig den beiden Interventionen zugeordnet. Die Randomisierung basierte auf einer Blockrandomisierung mit Blocklänge $b = 6$ und wurde nach Zentren für ein Allokationsverhältnis von $k = 1$ ($n_{\text{Verum}} = n_{\text{Kontrolle}}$) stratifiziert. Die Anzahl der Patienten je Zen-

trum ist in Abbildung 1.2 dargestellt. Schließlich konnte zwischen den Behandlungsgruppen eine statistisch signifikante Verbesserung des primären Endpunktes zugunsten der neuen Therapie in Höhe von 1,26 Punkten (95% Konfidenzintervall 0,79-1,75; p-Wert$< 0.0001$) beobachtet werden.

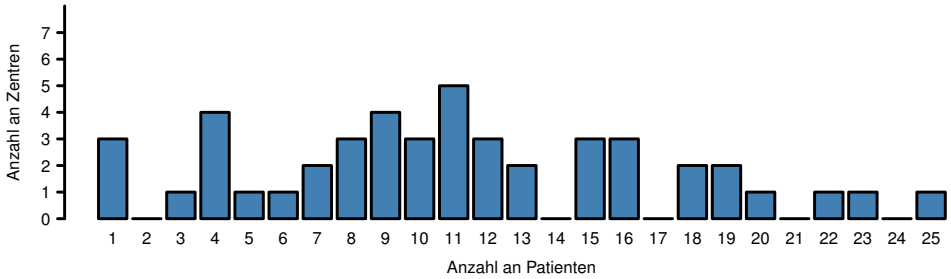


Abbildung 1.2: Anzahl der rekrutierten Patienten der COMPETE II-Studie je Zentrum.

## 1.3 Fragestellungen

Im Rahmen dieser Dissertation habe ich mich mit der Fallzahlplanung multizentrischer randomisierter klinischer Studien befasst. Der Fokus der Arbeit lag dabei auf der Entwicklung einer Fallzahlformel für den Fall, dass die Rekrutierung und Randomisierung der Probanden in vielen kleinen Zentren erfolgt. Dieser Ansatz wurde schließlich auf ein adaptives Studiendesign mit Fallzahlrekalkulation übertragen.

### 1.3.1 Fallzahlplanung in multizentrischen Studien

Die Fallzahlplanung ist ein wesentlicher Bestandteil der Planung einer jeden klinischen Studie. Wenn Probanden im Rahmen einer klinischen Studie an mehreren Standorten rekrutiert werden sollen, muss die Möglichkeit lokaler Unterschiede zwischen den Zentren bei der Planung und Auswertung der Studie berücksichtigt werden [11]. Dieser Einfluss durch die Zentren kann bei der statistischen Modellierung entweder durch einen festen oder zufälligen Effekt beschrieben werden. Für lineare Modelle mit festen Zentrumseffekten wurden von Gallo und Ruvuna zwei Korrekturen für Formel (1.1) beschrieben, die auf eine unbalancierte Zentrumsallokation der Probanden abzielen [29, 30]. Beide Korrekturen basieren auf dem sog. *inefficiency factor*, der auf dem Quotienten der Varianzen für balancierte und unbalancierte Studiendesigns (Typ II oder Typ III Fehler) basiert. Vierron und Giradeau schlugen eine Korrektur der unter (1.1) genannten Fallzahlformel für das Modell

mit zufälligem Zentrumseffekt vor, indem Sie den Intraklassen-Korrelationskoeffizienten $\rho = \sigma^2/(\sigma^2 + \tau^2)$ in Formel (1.1) aufnehmen, wobei $\tau^2$ den zufälligen Effekt der Zentren beschreibt [31, 32]. Van Breukelen und Kollegen wählten einen ähnlichen Ansatz wie Gallo und Ruvuna und ergänzten einen *inefficiency factor*, der allerdings auf der relativen Effizienz von ungleichen versus gleichen Zentrumsgrößen basiert, um die Heterogenität bei Modellen mit zufälligen Effekten zu berücksichtigen [33]. In Vergleichsstudien wurde gezeigt, dass das Modell mit zufälligen Effekten dem Modell mit festen Zentrumseffekten in vielen Situationen überlegen ist, vor allem dann, wenn die Anzahl an Probanden je Zentrum klein ist [13, 34].
Alle oben beschriebenen Verfahren treffen starke Annahmen an das statistische Modell, die in realen Daten nicht zwingend vorliegen müssen. So verlangen einige Verfahren balancierte Studiendesigns, d.h., es wird vorausgesetzt, dass die Behandlungsgruppen je Zentrum identisch groß sind oder dass die Rekrutierungsgeschwindigkeit und Stichproben je Behandlungsgruppe schon zur Planung feststehen oder bekannt sind. Diese Annahmen sind häufig unrealistisch und vereinfachen die Fallzahlformel zu sehr, was schließlich zu schlechten Resultaten, d.h. einer zu geringen Power der Studie führen kann, wenn diese Annahmen nicht zutreffen. Ein Ziel meiner Forschung war es daher, eine Fallzahlformel für multizentrische Studien zu entwickeln, die weniger strikte Annahmen an das statistische Modell stellt als bisher beschriebene Methoden, und schließlich zu untersuchen, welche Fallzahl in welchen Situationen angebracht ist.

### 1.3.2 Fallzahlrekalkulation in multizentrischen Studien

Ein wiederkehrendes Problem bei der Planung klinischer Studien ist die Vorspezifikation der Effektgröße $(\mu^*)$ und Variabilität des Endpunktes $(\sigma^2)$ für die Fallzahlplanung. In manchen Fällen gibt es externe Pilotstudien, auf deren Grundlage man ein erstes Verständnis für die erwarteten Ergebnisse in der gewünschten Population entwickeln konnte, doch meist basieren dieser Erfahrungen auf kleineren Stichproben oder die bisher untersuchten Populationen sind nicht ohne weiteres mit denen der neuen Studie vergleichbar. Daher herrscht zum Planungszeitpunkt der Studie immer eine gewisse Unsicherheit bezüglich der Wahl der Parameter zur Fallzahlplanung. Um das Risiko falsch negativer Studienergebnisse, beispielsweise durch eine zu geringe Fallzahl, zu verringern, wurden adaptive Studiendesigns entwickelt, um solche initialen Fehler im Verlauf der Studie zu korrigieren. Aus regulatorischer Sicht ist bei adaptiven Studiendesigns insbesondere darauf zu achten, dass alle geplanten Veränderungen im Vorfeld der Studie beschrieben, wenn auch noch nicht spezifiziert, werden und das Signifikanzniveau der Studie durch dieses Eingreifen nicht erhöht

wird [1, 23, 35, 24]. Für die Fallzahlrekalkulation ist es demnach zulässig, die Varianzparameter durch in der Studie erhobene Daten neu zu schätzen, solange das Fehlerniveau davon unberührt bleibt.
Da multizentrische Studien weitere Parameter aufweisen, die die Fallzahl beeinflussen und im Vorfeld der Studie mitunter nicht zu schätzen sind, werden bestehende Techniken zur Fallzahlrekalkulation mit der in Kapitel 1.1.3 vorgestellten Fallzahlformel kombiniert. Es muss untersucht werden, welche Parameter die Fallzahl beeinflussen, ob diese in Einklang mit den regulatorischen Anforderungen während einer Zwischenauswertung berechnet werden können und ob die geplante statistische Power erreicht wird.

## 1.4 Aufbau der Arbeit

In dieser Arbeit bespreche ich Lösungsansätze für die Fragestellungen, die in Kapiteln 1.3.1 und 1.3.2 vorgestellt wurden. Die Ergebnisse meiner Forschung wurden als Originalarbeiten in wissenschaftlichen Zeitschriften publiziert, die in einem Peer-Review Verfahren begutachtet wurden [36, 37].
In Kapitel 2 stelle ich eine Zusammenfassung meiner Ergebnisse dar, die sich analog zu den zuvor beschriebenen Fragestellungen und zwei Abschnitte gliedern. In Kapitel 3 diskutiere ich die Ergebnisse und getroffene Annahmen hinsichtlich des Studiendesigns und statistischer Methoden.

# 2 Methoden zur Fallzahlplanung und Fallzahlrekalkulation in multizentrischen Studien

## 2.1 Fallzahlplanung in multizentrischen Studien

Die Ergebnisse meiner Forschung zur Fallzahlplanung multizentrischer Studien wurden in [36] publiziert. Im Folgenden sind die Ergebnisse dieses Artikels zusammengefasst.
Man betrachte ein gemischtes lineares Modell für den Vergleich zweier Therapien $i = 1, 2$ an mehreren Zentren $j = 1, \ldots, c$ hinsichtlich eines kontinuierlichen Endpunktes. Die Anzahl der Probanden je Behandlung und Zentrum wird hier als beliebig angenommen, soll aber aus einer lokalen Blockrandomisierung mit Blocklänge $b$ und einem angestrebten Allokationsverhältnis $n_{1j} = k \cdot n_{2j}$ je Zentrum resultieren, wobei $n_{ij}$ die Anzahl an Probanden mit Behandlung $i$ in Zentrum $j$ bezeichnet. Formal lässt sich das Modell wie folgt beschreiben

$$Y_{ijk} = \mu_0 + u_j + \mu \cdot x_i + \epsilon_{ijk} \tag{2.1}$$

mit paarweise unabhängigen, zufälligen Effekten $u_j$ und Residuen $\epsilon_{ijk}$ mit $\mathrm{E}(u_j) = 0$, $\mathrm{Var}(u_j) = \tau^2 < \infty$, $\mathrm{E}(\epsilon_{ijk}) = 0$, $\mathrm{Var}(\epsilon_{ijk}) = \sigma^2 < \infty$, festem Intercept $\mu_0$, festem Behandlungseffekt $\mu$, Behandlungsindikator $x_i = 1_{\{i=2\}}$ für $i = 1, 2$, Zentren $j = 1, \ldots, c$ und den Probanden $k = 1, \ldots, n_{ij}$ je Zentrum und Behandlung. Die Kovarianz aller Beobachtungen wird durch die Block-Diagonalmatrix $\mathrm{Cov}(Y_{111}, \ldots, Y_{2cn_{2c}}) = \bigoplus_{j=1}^{c} \left[\sigma^2 \mathbf{I}_{n_j} + \tau^2 \mathbf{J}_{n_j}\right]$ für alle $N$ Probanden beschrieben, wobei $N = \sum_{i=1}^{2} \sum_{j=1}^{c} n_{ij}$ und $\bigoplus$ die direkte Summe der Kovarianzmatrizen je Zentrum bezeichnet. In dieser Arbeit bezeichnet $\mathbf{I}_{n_j}$ die $n_j$-dimensionale Identitätsmatrix und $\mathbf{J}_{n_j}$ die $n_j$-dimensionale Matrix bestehend aus Einsen und $n_j = n_{1j} + n_{2j}$.
Der Behandlungsunterschied zwischen den beiden Behandlungsgruppen wird durch die Nullhypothese $H_0 : \mu = 0$ gegen die zweiseitige Alternativhypothese $H_A : \mu \neq 0$ getestet.

Da die Schätzung von $\mu$ auf Mittelwerten von kontinuierlichen Zufallsvariablen beruht, kann die Teststatistik auf Grundlage des zentralen Grenzwertsatzes gegen ein Normalverteilungsquantil $q_{1-\alpha/2}$ zu gewähltem Signifikanzniveau $\alpha$ verglichen werden.
Basierend auf dem statistischen Modell (2.1) kann $\widehat{\mu} = \overline{Y}_{2\cdot\cdot} - \overline{Y}_{1\cdot\cdot}$ als erwartungstreuer und konsistenter Schätzer für den Behandlungsunterschied $\mu$ verwendet werden, wobei $\overline{Y}_{i\cdot\cdot} = 1/N_i \sum_{j=1}^{c} \sum_{k=1}^{n_{ij}} Y_{ijk}$. Die Varianz dieses Schätzers ist

$$\mathrm{Var}\,(\widehat{\mu}) = \sigma^2 \frac{N}{N_1 N_2} + \tau^2 \sum_{j=1}^{c} \left( \frac{n_{1j}}{N_1} - \frac{n_{2j}}{N_2} \right)^2 \tag{2.2}$$

und kann unter der Annahme $N_1 = k \cdot N_2$ und mit $\Delta_j^2 := \left(\frac{n_{1j}}{k} - n_{2j}\right)^2 \in [0, m^*]$ wie folgt dargestellt werden

$$\mathrm{Var}\,(\widehat{\mu}) = \sigma^2 \frac{(k+1)^2}{kN} + \tau^2 \frac{(k+1)^2}{N^2} \sum_{j=1}^{c} \Delta_j^2. \tag{2.3}$$

Ersetzt man in (2.3) die unbekannten Varianzparameter $\sigma^2$ und $\tau^2$ durch die folgenden konsistenten Schätzer

$$\widehat{\sigma}^2 = \frac{1}{2c} \sum_{i=1}^{2} \sum_{j=1}^{c} \frac{1}{n_{ij}-1} \sum_{k=1}^{n_{ij}} \left(Y_{ijk} - \overline{Y}_{ij\cdot}\right)^2 \text{ und}$$

$$\widehat{\tau}^2 = \frac{1}{2} \sum_{i=1}^{2} \frac{1}{c-1} \sum_{j=1}^{c} \left(\overline{Y}_{ij\cdot} - \overline{Y}_{i\cdot\cdot}\right)^2$$

gelangt man zu einer unter $H_0$ asymptotisch normalverteilten Teststatistik

$$T = \frac{\widehat{\mu}}{\sqrt{\widehat{\mathrm{Var}\,(\widehat{\mu})}}} \overset{H_0}{\sim} \mathrm{N}(0,1) \tag{2.4}$$

und damit zu der Fallzahlformel

$$N_{\mathrm{MC}}^k(\Delta_j^2) = \left( \frac{q_{1-\alpha/2} + q_{1-\beta}}{\mu} \right)^2 \cdot \left( \frac{\sigma^2 (k+1)^2}{2k} + \sqrt{\frac{\sigma^4 (k+1)^4}{4k^2} + \frac{\tau^2 (k+1)^2 \mu^2 \sum_{j=1}^{c} \Delta_j^2}{\left(q_{1-\alpha/2} + q_{1-\beta}\right)^2}} \right). \tag{2.5}$$

Der Ausdruck $\Delta_j^2$ beschreibt die Abweichung der Fallzahlen je Zentrum von dem vorge-

gebenen Allokationsverhältnis $k$ und hat als obere Grenze $m^* = b^2/(k+1)^2$, da eine Blockrandomisierung je Zentrum angenommen wird. Die Verteilung von $\Delta_j^2$ hängt damit von der Menge aller Randomisierungstupel $\Pi_b^k$ ab, wobei

$$\Pi_b^k = \left\{ (x_1, \ldots, x_b) \in \Pi_b \middle| \sum_{\ell=1}^{b} 1_{\{x_\ell = 1\}} = \frac{kb}{k+1} = b - \sum_{\ell=1}^{b} 1_{\{x_\ell = 2\}} \right\} \tag{2.6}$$

mit $\Pi_b := \{(x_1, \ldots, x_b) | x_\ell \in \{1, 2\}\}$. Sie kann durch die Blocklänge $b$, den Allokationsparameter $k$ und die Anzahl der Probanden im letzten Randomisierungsblock je Zentrum $r_j = n_j \bmod b$ beschrieben werden (s. Fig. 1 in [36]). Der Erwartungswert von $\Delta_j^2 | r_j$ ist ohne Kenntnis der eigentlichen Fallzahlen berechenbar und bietet sich daher als Substitut in Formel (2.5) an. Exemplarisch ist dieser Erwartungswert in Abbildung 2.1 für eine $1:1$-Allokation dargestellt.

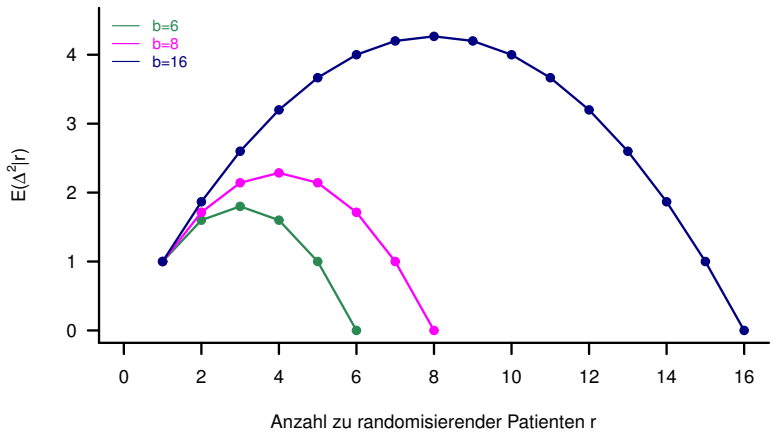


Abbildung 2.1: **Erwartungswert von $\Delta^2|r$.** Bedingter Erwartungswert der Abweichung der Fallzahlen vom Allokationsverhältnis $k = 1$ für variierende Anzahl an Probanden $r = 1 \ldots, b$ und Blocklänge $b$.

Mit Hilfe der getroffenen Annahmen erhält man die Fallzahlformel

$$N_{\mathrm{MC}}^k = \left( \frac{q_{1-\alpha/2} + q_{1-\beta}}{\mu} \right)^2 \cdot \left( \frac{\sigma^2 (k+1)^2}{2k} + \sqrt{\frac{\sigma^4 (k+1)^4}{4k^2} + \frac{\tau^2 (k+1)^2 \mu^2 \sum_{j=1}^{c} \mathrm{E}\left(\Delta_1^2 | r_j\right)}{\left(q_{1-\alpha/2} + q_{1-\beta}\right)^2}} \right). \tag{2.7}$$

Diese allgemeine Form kann weiter konkretisiert werden, wenn genauere Informationen

über die einzelnen Zentren vorliegen. Als untere Grenze der benötigten Fallzahl dient Formel (1.1), wenn man den Einfluss einer womöglich vorhandenen Heterogenität der Zentren ignoriert [16]. Eine obere Grenze der benötigten Fallzahl erhält man, wenn man davon ausgeht, dass in jedem unvollständigen Randomisierungsblock $\sqrt{m^*} = b/(k+1)$ Probanden dieselbe Therapie erhalten und $\mathrm{E}\left(\Delta_1^2 | b/(k+1)\right)$ in Formel (2.7) einsetzt. Zusätzlich kann man die Formel konkretisieren, wenn davon ausgegangen werden kann, dass alle Zentren identisch viele Patienten im letzten Randomisierungsblock rekrutieren werden. Als robuste Alternative schlagen ich vor, in jedem Zentrum den Ausdruck $\mathrm{E}\left(\Delta_1^2 | r_j\right)$ durch den Mittelwert

$$\frac{1}{b}\sum_{\ell=1}^{b} \mathrm{E}\left(\Delta_1^2 | \ell\right) =: \overline{\mathrm{E}\left(\Delta_1^2 | \cdot\right)}. \tag{2.8}$$

zu ersetzen. Der Einfluss von Zentrumsheterogenität, Blocklänge und Anzahl der rekrutierenden Zentren auf die Fallzahl ist exemplarisch in Abbildung 2.2 dargestellt. Die gewählten Parameter beruhen exemplarisch teilweise auf publizierten Studienergebnissen der COMPETE II-Studie, die die Wirksamkeit eines unterstützenden Management Tools bei der Behandlung von Diabetespatienten durch den Hausarzt untersucht hat [27, 28].

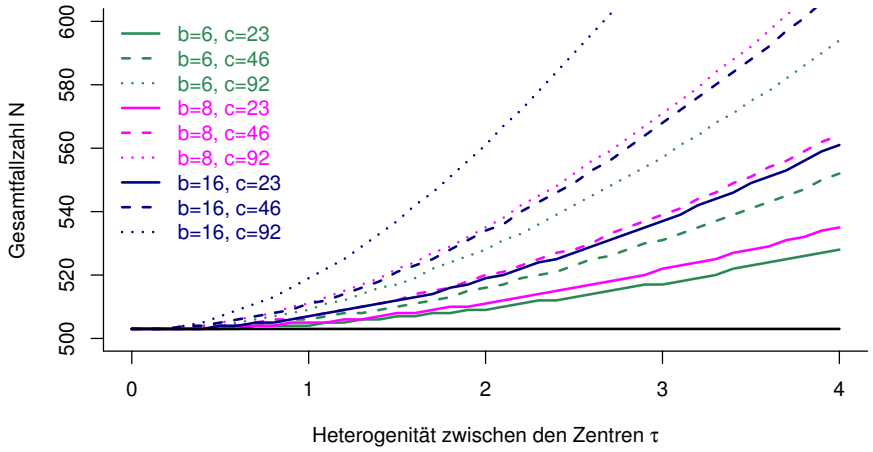


Abbildung 2.2: **Exemplarische Fallzahlberechnung basierend auf $N_{\mathrm{MC}}^k$.** Basiert auf $\mu = 1$, $\sigma = 4$, variierender Blocklänge $b$, Anzahl an Zentren $c$ und Heterogenität der Zentren $\tau$. Die schwarze durchgezogene Linie beschreibt die untere Grenze der Fallzahl für die gewählten Parameter ($N = 503$).

Die Eigenschaften dieser robusten Fallzahlformel habe ich mit Hilfe von Simulationsstudien

mit der freien Statistiksoftware R untersucht [38]. Der für die Simulationen verwendete Programmcode wurde im Rahmen des Artikels veröffentlicht. Die gewählten Parameter basieren ebenfalls zum Teil auf den Ergebnissen der COMPETE II-Studie [27, 28].
In Abbildung 2.3 ist die statistische Power in Abhängigkeit von der Anzahl an rekrutierenden Zentren und variierender Blocklänge dargestellt (zugehörige Fallzahlen in Table 2, [36]). Jeder Punkt ist das Resultat von $n_{\text{sim}} = 10\,000$ Simulationsdurchläufen. Für die dargestellten Simulationsergebnisse wurde angenommen, dass die Fallzahl gemäß einer Multinomialverteilung mit zufälligen Wahrscheinlichkeiten auf die Zentren aufgeteilt wird, d.h.

$$(n_1, \ldots, n_c)' \sim \text{Multi}_c\,(N, p_1^*, \ldots, p_c^*) \text{ mit } p_j^* = \frac{p_j}{\sum_{k=1}^{c} p_k} \text{ und } p_j \overset{iid}{\sim} U[0;1].$$

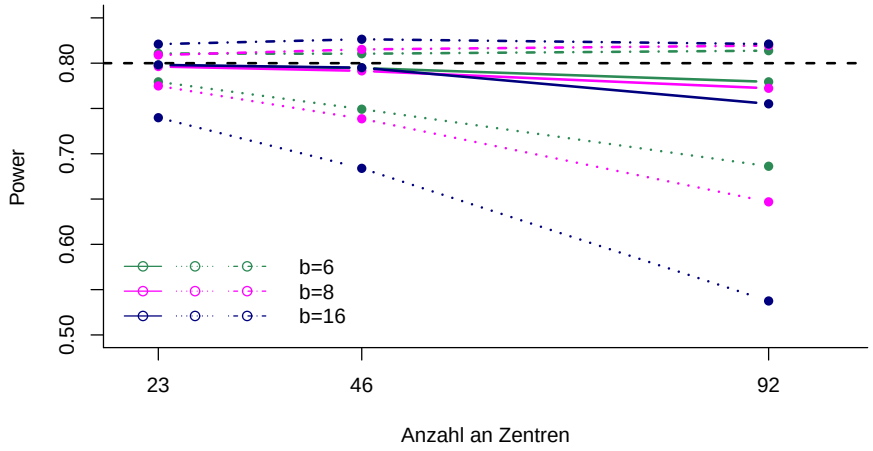


Abbildung 2.3: **Powersimulationen** $N_{\text{MC}}^k$. Simulierte Power für einen klinisch relevanten Effekt $\mu^* = 1$, $\sigma = \tau = 4$ und variierende Blocklänge $b$ und Anzahl an Zentren $c$. Die schwarz-gestrichelte Linie markiert die angestrebte Power von 0.8. Durchgezogene Linien markieren Ergebnisse basierend auf robuster Fallzahlformel mit $\overline{\text{E}\left(\Delta_1^2|\cdot\right)}$, während die gestrichelt-gepunktete Linie die obere Grenze gemäß $\text{E}\left(\Delta_1^2|m^*\right)$ und die gepunktete Linie die Ergebnisse bei Planung mit Formel (1.1) markiert.

Man erkennt, dass die Berücksichtigung der Heterogenität der Studienzentren wichtig ist, insbesondere dann, wenn viele Zentren bei der Rekrutierung der Probanden mitwirken. Die obere Grenze liegt etwas oberhalb der angestrebten Power von $0.80$, die vorgeschlagene robuste Variante der Fallzahlformel erreicht die statistische Power für eine kleine bis mittlere

Anzahl an Zentren und unterschreitet die angestrebte Power etwas im Falle vieler Zentren. Zusammenfassend habe ich eine Fallzahlformel für multizentrische Studien entwickelt, die eine mögliche Heterogenität der Zentren berücksichtigt und auch für ungleich große Zentren verwendet werden kann. Mit Hilfe von Monte-Carlo Simulationen habe ich gezeigt, dass die Planung durch die Fallzahlformel zu der geplanten statistischen Power führt.

## 2.2 Fallzahlrekalkulation in multizentrischen Studien

Da die Planung einer klinischen Studie in der Regel basierend auf bereits erhobenen Daten erfolgt, kann die Fallzahlplanung der Studie auf falschen Annahmen beruhen. Um diese Annahmen unter Einhaltung statistischer Prinzipien korrigieren zu können, wurden verschiedene adaptive Studiendesigns entwickelt, die die Anpassung gewisser Designoptionen während der Laufzeit der Studie zulassen. Für multizentrische Studien habe ich die in Kapitel 2.1 vorgestellte Fallzahlformel in ein adaptives Studiendesign mit Fallzahlrekalkulation implementiert. Die Ergebnisse wurden in [37] publiziert und werden in diesem Kapitel zusammengefasst.

Das zugrunde liegende statistische Modell ist für dieses Verfahren identisch zu dem in Kapitel 2.1. Ebenso gehe ich an dieser Stelle von einer Blockrandomisierung mit fester Blocklänge und ungleich großen Studienzentren aus. Im Gegensatz zu der bisherigen Betrachtung eines festen Studiendesigns ist nun aber vorgesehen, dass zu einem festen Zeitpunkt während der Rekrutierung der Patienten eine Teilauswertung durchgeführt wird, um die getroffenen Annahmen an die Daten bezüglich der Fallzahlplanung zu adjustieren.

In Kapitel 2.1 wurde gezeigt, dass für die Berechnung der Fallzahl einer multizentrischen Studie der angenommene Behandlungseffekt $\mu^*$, $\alpha$- und $\beta$-Fehler, die Variabilität der Beobachtungen $\sigma^2$ und $\tau^2$, die Blocklänge $b$, das Allokationsverhältnis $k$, die Anzahl der Zentren $c$ und die Abweichung der Fallzahlen im Zentrum vom angestrebten Allokationsverhältnis $\Delta_j^2$ spezifiziert werden müssen. Um eine Kontrolle des Fehlerniveaus zu bewahren, habe ich die Rekalkulation von $\sigma^2$, $\tau^2$ und $\Delta_j^2$ auf Grundlage nicht-komparativer Daten genauer betrachtet. Ich habe untersucht, wie sich die empirische Verteilung der $\Delta_j^2$ verändert, wenn man den Zeitpunkt der Fallzahlrekalkulation verschiebt, und ob man mit dieser Information die finalen Blocklängen in den Zentren schätzen kann (Figure 2, [37]). Da ich für die Schätzung der finalen $\Delta_j^2$ zum Zeitpunkt der Fallzahlrekalkulation keine überzeugenden Ergebnisse erhalten habe, beschränke ich mich im Folgenden auf die Ergebnisse zur Schätzung der Varianzparameter $\sigma^2$ und $\tau^2$.

Diese können ohne Kenntnis der Behandlungszugehörigkeit, aber verzerrt, wie folgt aus den Daten geschätzt werden

$$\widehat{\sigma}_b^2 = \frac{1}{N-c} \sum_{i=1}^{2} \sum_{j=1}^{c} \sum_{k=1}^{n_{ij}} \left(Y_{ijk} - \overline{Y}_{\cdot j \cdot}\right)^2 \text{ und} \tag{2.9}$$

$$\widehat{\tau}_b^2 = \frac{1}{c-1} \sum_{j=1}^{c} \left(\overline{Y}_{\cdot j \cdot} - \overline{Y}_{\cdots}\right)^2 . \tag{2.10}$$

Die Betrachtung anderer Varianzschätzer auf Grundlage nicht-komparativer Daten wie in [39, 40] habe ich verworfen, da analytisch gezeigt wurde, dass diese Schätzer in typischen Situationen klinischer Studien eine mitunter deutlich größere Varianz haben [41]. Ich habe eine Korrektur für den Schätzer von $\tau_b^2$ betrachtet, um zu untersuchen inwiefern die Verzerrung von $\widehat{\tau}_b^2$ die Fallzahlrekalkulation beeinflusst, beziehungsweise ob diese Verzerrung effizient reduziert werden kann. Der Algorithmus einer Fallzahlanpassung auf Grundlage nicht-komparativer Daten (BSSR) gestaltet sich wie folgt:

Fallzahlrekalkulation auf Grundlage nicht-komparativer Daten (BSSR):

1. Berechnung der initialen Fallzahl $N_{\text{init}}$ basierend auf prä-spezifizierten Parametern für Formel (2.7) und Festlegung der Anzahl an Probanden $\lambda \cdot N_{\text{init}}$ mit denen die Rekalkulation der Varianzparameter erfolgen wird, $\lambda \in (0; 1)$.
2. Berechnung von $\widehat{\sigma}_b^2$ und $\widehat{\tau}_b^2$ basierend auf $\rho \cdot N_{\text{init}}$ Probanden.
3. Neuberechnung der Fallzahl $N_1$ mit neu geschätzten Werten für $\sigma^2$ und $\tau^2$.
4. Rekrutierung weiterer Probanden bis $N_{\text{final}} = \max(N_1; \lambda \cdot N_{\text{init}})$ erreicht ist. Wenn im Vorfeld eine obere Grenze $N_{\max}$ für die Anzahl der Probanden festgelegt wurde, dann rekrutiere

   $$N_{\text{final}} = \min\{\max(N_1; \lambda \cdot N_{\text{init}}); N_{\max}\}.$$

   Probanden. Ebenso kann eine untere Schranke $N_{\min}$ spezifiziert werden.
5. Finale Analyse aller $N_{\text{final}}$ Probanden.

Bei der Fallzahlrekalkulation ist es von grundlegender Bedeutung, dass das Fehlerniveau durch die Zwischenauswertung der Daten nicht erhöht wird [23, 24] . Daher habe ich den vorgestellten Ansatz zur Fallzahlrekalkulation mit Hilfe von Monte-Carlo Simulationsstudien bezüglich der Wahrscheinlichkeit eines Fehlers 1. und 2. Art untersucht. Der verwendete Programmcode wurde zusammen mit dem Artikel publiziert [37]. Die gewählten Parameter wurden analog zu Kapitel 2.1 durch die COMPETE II-Studie motiviert [27, 28].
Da der korrigierte Schätzer für $\tau^2$ bei kleinen Fallzahlen zu einer liberalen Teststatistik bezüglich der Wahrscheinlichkeit eines Fehlers 1. Art führen kann (Figure 4, [37]), beschränke ich mich an dieser Stelle auf die Ergebnisse für die Varianzschätzer aus (2.9) und (2.10) und eine Fallzahlrekalkulation auf Grundlage der halben initialen Fallzahl ($\lambda = 0.5$),

wie in Abbildung 2.4 dargestellt.

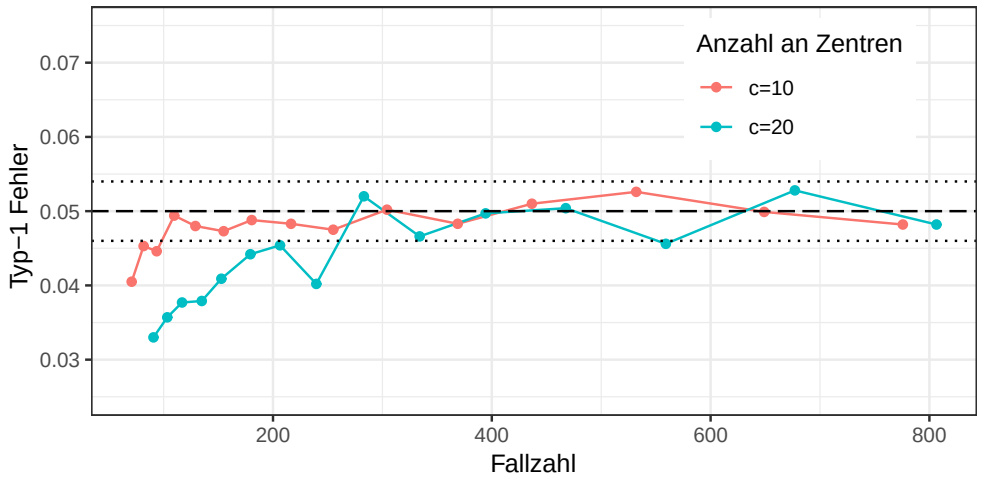


Abbildung 2.4: **Fehler 1. Art Fallzahlrekalkulation.** Simulierter Fehler 1. Art für $\mu = 0$, $c = 10$ und $c = 20$ Zentren, $\sigma^2 = \tau^2 = 16$ eine Blocklänge von $b = 16$ und variierende Fallzahl. Die Fallzahlrekalkulation findet auf Grundlage der halben initialen Fallzahl statt. Die gestrichelte Linie zeigt das angestrebte Signifikanzniveau von $\alpha = 0.05$, die gepunkteten Linien beschreiben den Simulationsfehler.

Man erkennt, dass das Testverfahren mit den Schätzern (2.9) und (2.10) die Wahrscheinlichkeit eines Fehlers 1. Art für alle gezeigten Parameterkonstellationen zum Signifikanzniveau $\alpha$ kontrolliert. Für große benötigte Fallzahlen, d.h. kleine Behandlungseffekte liegt der simulierte Typ-1 Fehler fast immer im erwarteten Zufallsstreubereich, bei kleinen Fallzahlen, und damit vielen kleinen Zentren, zeigt der Test ein leicht konservatives Verhalten. Bei einem Vergleich der simulierten Wahrscheinlichkeit eines Typ-1 Fehlers zwischen der Schätzung der Varianzparameter auf Grundlage komparativer beziehungsweise nicht-komparativer Daten konnte für die gewählten Parameter kein wesentlicher Unterschied beobachtet werden (Figure 3 und 4, [37]).
Die statistische Power wurde ebenfalls mit Hilfe von Monte-Carlo Simulationen untersucht. Dafür wurde ein Vergleich von einem festen Studiendesign ohne Fallzahladjustierung mit einem adaptiven Design mit Fallzahlrekalkulation auf Grundlage der halben initialen Fallzahl ($\lambda = 0.5$) betrachtet. Die Ergebnisse sind in Abbildung 2.5 dargestellt.
Die statistische Power wird für alle Parameterkonstellationen erreicht, wenn eine Fallzahlrekalkulation durchgeführt wird. Das gilt insbesondere dann, wenn die initiale Fallzahlpla-

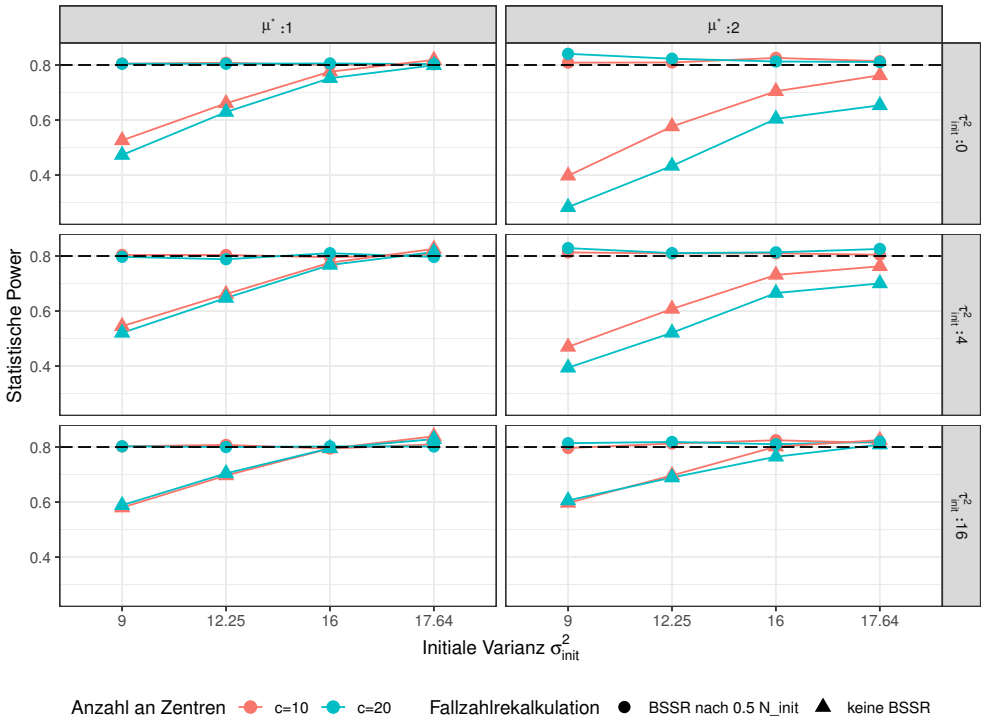


Abbildung 2.5: **Power Fallzahlrekalkulation.** Simulierte Power für $c = 10$ und $c = 20$ Zentren, einen Behandlungseffekt von $\mu^* = 1$ und $\mu^* = 2$ und variierende initiale Werte für $\sigma^2_{init}$ und $\tau^2_{init}$, wobei $\sigma^2 = \tau^2 = 16$. Die gestrichelte Linie zeigt die angestrebte statistische Power von $1 - \beta = 0.8$.

nung auf falschen Annahmen bezüglich der Varianzkomponenten $\sigma^2$ und $\tau^2$ beruht. Für die klassische Planung ohne Fallzahlrekalkulation sieht man, dass falsch getroffene Annahmen starke Auswirkungen auf die erreichte statistische Power haben. Dass das Verfahren mit Fallzahlrekalkulation im Falle korrekter initialer Varianzparameter zu einer etwas höheren statistischen Power führt ($\sigma^2_{\text{init}} = \tau^2_{\text{init}} = 16$) liegt daran, dass die bei der Fallzahlrekalkulation verwendeten Schätzer (2.9) und (2.10) verzerrt sind und die wahren Werte immer etwas überschätzen (Figure 7, [37]). Im Falle kleinerer Studien, d.h. $\mu^* = 2$ bei $c = 20$ rekrutierenden Zentren, tritt dieser Effekt verstärkt auf.
In weiteren Simulationsszenarien konnte kein Unterschied bei der Fallzahlrekalkulation basierend auf komparativen und nicht-komparativen Daten beobachtet werden (Figure 5, [37]). Dies ist vermutlich damit zu erklären, dass die betrachteten Szenarien zu hinreichend großen Fallzahlen führen ($N > 100$). Zusammenfassend habe ich gezeigt, dass die in Kapitel 2.1 vorgestellte Fallzahlformel zur Fallzahlrekalkulation verwendet werden kann und dass sowohl der Fehler 1. Art als auch die geplante statistische Power erreiche werden.

# 3 Diskussion

In dieser Arbeit habe ich eine neue Fallzahlformel für die Planung multizentrischer randomisierter klinischer Studien entwickelt und gezeigt, wie man diese auch im Rahmen eines adaptiven Studiendesigns für die Fallzahlrekalkulation verwenden kann.
Der Vorteil der entwickelten Fallzahlformel im Vergleich zu bereits beschriebenen Verfahren besteht darin, dass nur schwache Annahmen an das statistische Modell gestellt werden und eine beliebige – sofern auf eine Blockrandomisierung zurückzuführende – Behandlungsallokation in den Zentren zulässig ist. Die Blocklänge sollte nach Aussage der ICH E9 Guideline [1] dabei ausreichend kurz gewählt werden, um stark ungleiche Stichprobenumfänge zu vermeiden, aber auch so groß, dass die Vorhersagbarkeit der Behandlungszugehörigkeit der letzten Probanden in einem Randomisierungsblock gewahrt bleibt. Dass die von mir neu entwickelte Fallzahlformel insbesondere bei großen Blocklängen verwendet werden kann, habe ich in den Simulationsstudien gezeigt. Falls eine minimale Blocklänge verwendet wird ($b = 2$ für zwei Behandlungsgruppen) oder keine Heterogenität der Zentren vorliegt ($\tau^2 = 0$), dann vereinfacht sich die vorgestellte Fallzahlformel zum klassischen Ergebnis (1.1). Die Fallzahlformel lässt sich auch im Falle einer Blockrandomisierung mit zufälliger Blocklänge anwenden. Dabei ist jedoch zu beachten, dass die Bestimmung der erwarteten Abweichungen $\mathrm{E}\left(\Delta_j^2 | r_j\right)$ von der Spannweite der Blocklängen abhängt.
Eine Limitierung der vorgestellten Fallzahlformel besteht in der Notwendigkeit, die Heterogenität der Zentren im Vorfeld der Studie zu spezifizieren. Dies ist eine Information, die selten im Vorhinein der Studie bekannt ist. Es gibt zwar einige Artikel, in denen Schätzer für den Intraklassen-Korrelationskoeffizienten $\rho = \sigma^2/(\sigma^2 + \tau^2)$ angegeben sind, allerdings basieren diese Werte meist auf Cluster-randomisierten Studien, und es bleibt zu klären, inwieweit diese Schätzwerte die Heterogenität in einer neuen Studie widerspiegeln. Eine weitere Einschränkung des angenommenen Modells ist die Annahme, dass der Behandlungseffekt in allen Zentren identisch ist. Diese Annahme deckt sich allerdings mit der Forderung der ICH E9 Guideline, keine Zentrum-Behandlungs-Interaktion in der primären Analyse zu modellieren [1]. Man könnte untersuchen, inwiefern eine Fallzahlplanung bei dem Vorliegen einer solchen Interaktion zwischen Zentrum und Behandlung erfolgen kann, allerdings wäre dann ein weiterer Heterogenitätsparameter zu bestimmen, der noch schwie-

riger im Vorfeld zu spezifizieren ist.
Grundsätzlich sollten sich die vorgestellten Ideen auch auf andere Endpunkte (Binomial- oder Poissonverteilung) übertragen lassen. Da die Fallzahlformel auf der Varianz des Behandlungseffektschätzers beruht, verändert sich diese mit dem gewählten Endpunkt und Schätzer. Zusätzlich muss beachtet werden, dass die modellierte Heterogenität bei anderen Endpunkten womöglich auch den angenommen Behandlungseffekt beeinflusst, da Erwartungswert und Varianz in diesen Verteilungen von denselben Parametern beeinflusst werden.
Die Ungewissheit bezüglich der Zentrumsheterogenität zu Beginn einer Studie war ein Grund die Fallzahlformel mit einer adaptiven Fallzahlrekalkulation zu kombinieren. Der Vorteil der verblindeten Fallzahlrekalkulation im Vergleich zu einem festen Studiendesign besteht darin, dass die Studie mit der festgelegten statistischen Power durchgeführt werden kann, auch wenn die initialen Annahmen an die Varianzparameter falsch sind. Bei den Simulationsergebnissen der Fallzahlrekalkulation konnte kein Unterschied bezüglich der statistischen Power zwischen der verblindeten und unverblindeten Fallzahlrekalkulation beobachtet werden und in allen betrachteten Szenarien wurde das Fehlerniveau eingehalten. In anderen Arbeiten wurde gezeigt, dass die Verwendung der entblindeten Varianzschätzer zu einer Inflation des Typ-1 Fehlers führen kann [25, 26, 42]. Dass ein solcher Effekt in dieser Arbeit nicht beobachtet wurde, liegt vermutlich daran, dass ich von Studien mit vielen Zentren ausgegangen bin, was eine moderate Gesamtfallzahl erfordert und der erwartete Effekt dann noch nicht zum Tragen kommt.
Für Studien mit Blockrandomsierung wurde ein alternativer Varianzschätzer vorgeschlagen, der den Vorteil hat, dass er unverzerrt ist, obwohl er auf Grundlage nicht-komparativer Daten berechnet wird [39, 40]. Dieser Ansatz der Schätzung wurde sowohl auf Crossover- als auch Clusterrandomisierte Studien übertragen [43, 44], wird in dieser Arbeit aber nicht näher betrachtet, da er (a) auf balancierten Daten in den Randomisierungsblöcken beruht und (b) gezeigt wurde, dass die Varianz des Schätzers in für klinische Studien üblichen Situationen deutlich größer ist als die des in dieser Arbeit betrachteten Schätzers [41].
Die ideale Größe einer internen Pilotstudie hängt von Faktoren wie der Rekrutierungsgeschwindigkeit, Dauer der Studie und der Unsicherheit bezüglich der initial gewählten Parameter ab [45]. Da zum Zeitpunkt der Fallzahlrekalkulation womöglich noch nicht alle Patienten die Beobachtungsdauer für den primären Endpunkt erreicht haben, wurden Methoden entwickelt, die einen Surrogatendpunkt für kurze Beobachtungsdauern verwenden, um die Genauigkeit der Varianzschätzung zu erhöhen [46, 42]. Diese Methode könnte auf multizentrische Studien übertragen werden, war aber nicht Gegenstand dieser Arbeit.

# Literaturverzeichnis


[1] ICH E9 Statistical Principles for Clinical Trials. International Conference on Harmonisation, 1998. URL abgerufen am 15.01.2020: https://www.ema.europa.eu/en/documents/scientific-guideline/ich-e-9-statistical-principles-clinical-trials-step-5_en.pdf.

[2] ICH E10 Choice of control group in clinical trials. International Conference on Harmonisation, 2001. URL abgerufen am 15.01.2020: https://www.ema.europa.eu/en/documents/scientific-guideline/ich-e-10-choice-control-group-clinical-trials-step-5_en.pdf.

[3] James Lind. *A treatise of the Scurvy in three parts. Containing an inquiry into the nature, causes and cure of that disease, together with a critical and chronological view of what has been published on the subject.* Edinburgh: Printed by Sands, Murray and Cochran for A Kincaid and A Donaldson, 1753.

[4] Jerzy Splawa-Neyman, Dorota M Dabrowska, and TP Speed. On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Statistical Science*, pages 465–472, 1990.

[5] Ronald Aylmer Fisher. *Statistical Methods for Research Workers*. Oliver and Boyd, Edinburgh, 1925.

[6] William F Rosenberger and John M Lachin. *Randomization in clinical trials: theory and practice*. John Wiley & Sons, 2015.

[7] John Crofton and DA Mitchison. Streptomycin resistance in pulmonary tuberculosis. *British Medical Journal*, 2(4588):1009, 1948.

[8] Elaine M Beller, Val Gebski, and Anthony C Keech. Randomisation in clinical trials. *Medical Journal of Australia*, 177(10):565–567, 2002.

[9] Lisa N Yelland, Brennan C Kahan, Elsa Dent, Katherine J Lee, Merryn Voysey, Andrew B Forbes, and Jonathan A Cook. Prevalence and reporting of recruitment,

randomisation and treatment errors in clinical trials: a systematic review. *Clinical Trials*, 15(3):278–285, 2018.

[10] Simon J Day and Douglas G Altman. Blinding in clinical trials and other studies. *BMJ*, 321(7259):504, 2000.

[11] ICH guideline E17 on general principles for planning and design of multi-regional clinical trials. International Conference on Harmonisation, 2016. URL abgerufen am 15.01.2020: `https://www.ema.europa.eu/en/documents/scientific-guideline/ich-guideline-e17-general-principles-planning-design-multi-regional-clinical-trials-step-5-first_en.pdf`.

[12] David G Weiss, William O Williford, Joseph F Collins, and Stephen F Bingham. Planning multicenter clinical trials: a biostatistician's perspective. *Controlled Clinical Trials*, 4(1-2):53–64, 1983.

[13] Brennan C Kahan and Tim P Morris. Analysis of multicentre trials with continuous outcomes: when and how should we account for centre effects? *Statistics in Medicine*, 32(7):1136–1149, 2013.

[14] Stephen Senn. Some controversies in planning and analysing multi-centre trials. *Statistics in Medicine*, 17(15-16):1753–1765, 1998.

[15] Medical Research Council Patulin Trials Committee et al. Clinical trial of patulin in the common cold. *The Lancet*, 2:373–5, 1944.

[16] Steven A Julious and Roger J Owen. Sample size calculations for clinical studies allowing for uncertainty about the variance. *Pharmaceutical Statistics*, 5(1):29–37, 2006.

[17] Steven A Julious. Sample sizes for clinical trials with normal data. *Statistics in Medicine*, 23(12):1921–1986, 2004.

[18] Peter Bauer, Frank Bretz, Vladimir Dragalin, Franz König, and Gernot Wassmer. Twenty-five years of confirmatory adaptive designs: opportunities and pitfalls. *Statistics in Medicine*, 35(3):325–347, 2016.

[19] Frank Bretz, Franz Koenig, Werner Brannath, Ekkehard Glimm, and Martin Posch. Adaptive designs for confirmatory clinical trials. *Statistics in Medicine*, 28(8):1181–1217, 2009.

[20] Shein-Chung Chow. Adaptive clinical trial design. *Annual Review of Medicine*, 65:405–415, 2014.

[21] Laura E Bothwell, Jerry Avorn, Nazleen F Khan, and Aaron S Kesselheim. Adaptive design clinical trials: a review of the literature and clinicaltrials. gov. *BMJ Open*, 8(2):e018320, 2018.

[22] Gernot Wassmer and Werner Brannath. *Group sequential and confirmatory adaptive designs in clinical trials*. Springer, 2016.

[23] Reflection paper on methodological issues in confirmatory clinical trials planned with an adaptive design. European Medicines Agency, 2007. URL abgerufen am 15.01.2020: `https://www.ema.europa.eu/en/documents/scientific-guideline/reflection-paper-methodological-issues-confirmatory-clinical-trials-planned-adaptive-design_en.pdf`.

[24] Adaptive Design Clinical Trials for Drugs and Biologics Guidance for Industry. Food and Drug Administration, 2019. URL abgerufen am 15.01.2020: `https://www.fda.gov/media/78495/download`.

[25] Janet Wittes and Erica Brittain. The role of internal pilot studies in increasing the efficiency of clinical trials. *Statistics in Medicine*, 9(1-2):65–72, 1990.

[26] Martin A Birkett and Simon J Day. Internal pilot studies for estimating sample size. *Statistics in Medicine*, 13(23-24):2455–2463, 1994.

[27] Anne Holbrook, Lehana Thabane, Karim Keshavjee, Lisa Dolovich, Bob Bernstein, David Chan, Sue Troyan, Gary Foster, Hertzel Gerstein, COMPETE II Investigators, et al. Individualized electronic decision support and reminders to improve diabetes care in the community: Compete ii randomized trial. *Canadian Medical Association Journal*, 181(1-2):37–44, 2009.

[28] Rong Chu, Lehana Thabane, Jinhui Ma, Anne Holbrook, Eleanor Pullenayegum, and Philip James Devereaux. Comparing methods to estimate treatment effects on a continuous outcome in multicentre randomized controlled trials: a simulation study. *BMC Medical Research Methodology*, 11(1):21, 2011.

[29] Paul P Gallo. Practical issues in linear models analyses in multicenter clinical trials. *Biopharmaceutical Report*, 6:2–9, 1998.

[30] Francis Ruvuna. Unequal center sizes, sample size, and power in multicenter clinical trials. *Drug Information Journal*, 38(4):387–394, 2004.

[31] Emilie Vierron and Bruno Giraudeau. Sample size calculation for multicenter randomized trial: taking the center effect into account. *Contemporary Clinical Trials*, 28(4):451–458, 2007.

[32] Emilie Vierron and Bruno Giraudeau. Design effect in multicenter studies: gain or loss of power? *BMC Medical Research Methodology*, 9(1):39, 2009.

[33] Gerard JP van Breukelen, Math JJM Candel, and Martijn PF Berger. Relative efficiency of unequal versus equal cluster sizes in cluster randomized and multicentre trials. *Statistics in Medicine*, 26(13):2589–2603, 2007.

[34] Ruth M Pickering and Mark Weatherall. The analysis of continuous outcomes in multi-centre trials with small centre sizes. *Statistics in Medicine*, 26(30):5445–5456, 2007.

[35] Reflection paper on methodological issues in confirmatory clinical trials planned with an adaptive design. Food and Drug Administration, 2016. URL abgerufen am 15.01.2020: `https://www.fda.gov/media/92671/download`.

[36] Markus Harden and Tim Friede. Sample size calculation in multi-centre clinical trials. *BMC Medical Research Methodology*, 18(1):156, 2018.

[37] Markus Harden and Tim Friede. Sample size reestimation in muti-centre randomized controlled clinical trials based on non-comparative data. *Biometrical Journal*, pages 1–16, 2020.

[38] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2008. ISBN 3-900051-07-0.

[39] Jitendra Ganju and Biao Xing. Re-estimating the sample size of an on-going blinded trial based on the method of randomization block sums. *Statistics in Medicine*, 28(1):24–38, 2009.

[40] Biao Xing and Jitendra Ganju. A method to estimate the variance of an endpoint from an on-going blinded trial. *Statistics in Medicine*, 24(12):1807–1814, 2005.

[41] Tim Friede and Meinhard Kieser. Blinded sample size re-estimation in superiority and noninferiority trials: bias versus variance in variance estimation. *Pharmaceutical Statistics*, 12(3):141–146, 2013.

[42] Tim Friede and Meinhard Kieser. Sample size recalculation in internal pilot study designs: a review. *Biometrical Journal*, 48(4):537–555, 2006.

[43] Michael J Grayling, Adrian P Mander, and James MS Wason. Blinded and unblinded sample size reestimation procedures for stepped-wedge cluster randomized trials. *Biometrical Journal*, 60(5):903–916, 2018.

[44] Michael J Grayling, Adrian P Mander, and James MS Wason. Blinded and unblinded sample size reestimation in crossover trials balanced for period. *Biometrical Journal*, 60(5):917–933, 2018.

[45] Christy Chuang-Stein, Keaven Anderson, Paul Gallo, and Sylva Collins. Sample size reestimation: a review and recommendations. *Drug Information Journal*, 40(4):475–484, 2006.

[46] Thomas Asendorf, Robin Henderson, Heinz Schmidli, and Tim Friede. Sample size re-estimation for clinical trials with longitudinal negative binomial counts including time trends. *Statistics in Medicine*, 38(9):1503–1528, 2019.

# A Appendix

# A.1 Veröffentlichungen

## Veröffentlichungen

### Methodische Publikationen

2018 Harden, M., Friede, T. Sample size calculation in multi-centre clinical trials. BMC Med Res Methodol 18, 156 (2018). https://doi.org/10.1186/s12874-018-0602-y

2020 Harden, M, Friede, T. Sample size recalculation in multicenter randomized controlled clinical trials based on noncomparative data. Biometrical Journal. 2020; 1– 16. https://doi.org/10.1002/bimj.201900138

### Klinische Publikationen

2015 Krasnianski A., Bohling G., Harden M., Zerr I. (2015). Psychiatric symptoms in patients with sporadic Creutzfeldt-Jakob disease in Germany. The Journal of Clinical Psychiatry, 76(9), 1209-1215.

2016 Straube S., Harden M., Schröder H., Arendacka B., Fan X., Moore R.A., Friede T. (2016). Back schools for the treatment of chronic low back pain: possibility of benefit but no convincing evidence after 47 years of research – systematic review and meta-analysis. Pain, Oct;157(10):2160-72.

Wetz A.J., Perl T., Brandes I.F., Harden M., Bauer M., Bräuer A. (2016). Unexpectedly high incidence of hypothermia before induction of anesthesia in elective surgical patients. Journal of Clinical Anesthesia, Nov;34:282-9.

2017 Balcarek P., Rehn S., Howells N.R., Eldridge J.D., Kita K., Dejour D., Nelitz M., Banke I.J., Lambrecht D., Harden M. and Friede T. (2017). Results of medial patellofemoral ligament reconstruction compared with trochleoplasty plus individual extensor apparatus balancing in patellar instability caused by severe trochlear dysplasia: a systematic review and meta-analysis. Knee Surgery, Sports Traumatology, Arthroscopy, Dec;25(12):3869-3877.

Schütz E., Fischer A., Beck J., Harden M., Koch M., Wuensch T., Stockmann M., Nashan B., Kollmar O., Matthaei J., Kanzow P., Walson P.D., Brockmöller J., Oellerich M. (2017). Graft-derived cell-free DNA, a noninvasive early rejection and graft damage marker in liver transplantation: A prospective, observational, multicenter cohort study. PLoS Medicine, Apr 25;14(4):e1002286.

2018 Grimmsmann T., Harden M., Fiß T., Himmel W. (2018). The influence of hospitalisation on the initiation, continuation and discontinuation of benzodiazepines and Z-drugs–an observational study. Swiss Medical Weekly, Feb 14;148:w14590.

2019 Ott M., Avendaño-Guzmán E., Ullrich E., Dreyer C., Strauss J., Harden M., Schön M., Schön M.P., Bernhardt G., Stadelmann C., Wegner C., Brück W., Nessler S. (2019). Laquinimod, a prototypic quinoline-3-carboxamide and aryl hydrocarbon receptor agonist, utilizes a CD155-mediated natural killer/dendritic cell interaction to suppress CNS autoimmunity. Journal of Neuroinflammation, Feb 26;16(1):49.

Bauer A., Klemm M., Rizas K.D., Hamm W., von Stülpnagel L., Dommasch M., Steger A., Lubinski A., Flevari P., Harden M., Friede T., Kääb S., Merkely B., Sticherling C., Willems R., Huikuri H.V., Malik M., Schmidt G., Zabel M., & EU-CERT-ICD investigators (2019). Prediction of mortality benefit based on periodic repolarisation dynamics in patients undergoing prophylactic implantation of a defibrillator: a prospective, controlled, multicentre cohort study. The Lancet, Oct 12;394(10206):1344-1351.

Zabel M., Schlögl S., Lubinski A., Svendsen J.H., Bauer A., Arbelo E., Brusich S., Conen D., Cygankiewicz I., Dommasch M., Flevari P., Galuszka J., Hansen J., Hasenfuß G., Hatala R., Huikuri H.V., Kenttä T., Kucejko T., Haarmann H., Harden M., Iovev S., Kääb S., Kaliska G., Katsimardos A., Kasprzak J.D., Qavoq D., Lüthje L., Malik M., Novotný T., Pavlović N., Perge P., Röver C., Schmidt G., Shalganov T., Sritharan R., Svetlosak M., Sallo Z., Szavits-Nossan J., Traykov V., Vandenberk B., Velchev V., Vos M.A., Willich S.N., Friede T., Willems R., Merkely B., Sticherling C., & EU-CERT-ICD investiga-

tors. (2019). Present criteria for prophylactic ICD implantation: Insights from the EU-CERT-ICD (Comparative Effectiveness Research to Assess the Use of Primary ProphylacTic Implantable Cardioverter Defibrillators in EUrope) project. Journal of Electrocardiology, Volume 57, Supplement, November–December, Pages S34-S39

2020 Junttila J., Pelli A., Kenttä T.V., Friede T., Willems R., Bergau L., Malik M., Vandenberk B., Vos M.A., Schmidt G., Merkely B., Lubinski A., Svetlosak M., Braunschweig F., Harden M., Zabel M., Huikuri H.V., Sticherling C. for the EU-CERT-ICD Investigators (2020). Appropriate shocks and mortality in patients with versus without diabetes with prophylactic implantable cardioverter defibrillators. Diabetes Care; Jan; 43(1): 196-200.

Pelli A., Kentta T., Junttila J., Bergaus L., Zabel M., Malik M., Reichlin T., Willems R., Vos M., Harden M., Friede T., Sticherling C. (2020). Electrocardiogram as a predictor of survival without appropriate shocks in primary prophylactic ICD patients: A retrospective multi-center study. International Journal of Cardiology, (in press).

Gross O., Tönshoff B., Weber L. T., Pape L., Latta K., Fehrenbach H., Lange-Sperandio B., Zappel H., Hoyer P., Staude H., König S., John U., Gellermann J., Hoppe B., Galiano M., Hoecker B., Ehren R., Lerch C., Kasthan C. E., Harden M., Boeckhaus J., Friede T., for the GPN Study Group and EARLY PRO-TECT Alport investigators. (2020). A multicenter, randomized, placebo-controlled, double-blind phase 3 trial with open-arm comparison indicates safety and efficacy of nephroprotective therapy with ramipril in children with Alport's syndrome. Kidney International.

Zabel M., Willems R., Lubinski A., Bauer A., Brugada J., Conen D., Flevari11 P., Hasenfuß G., Hatala R., Huikuri H. V., Malik M., Pavlović N., Schmidt G., Sritharan1 R., Schlögl1 S., Szavits-Nossan J., Traykov V., Tuinenburg A. E., Willich S. N., Harden M., Friede T., Svendsen J. H., Sticherling C., Merkely B., & the EU-CERT-ICD Study Investigators . (2020). Clinical effectiveness of primary prevention implantable cardioverter defibrillators: results of the EU-CERT ICD controlled multicentre cohort study. European Heart Journal, (in press).

Harden and Friede *BMC Medical Research Methodology* (2018) 18:156
https://doi.org/10.1186/s12874-018-0602-y

BMC Medical Research Methodology

RESEARCH ARTICLE Open Access

# Sample size calculation in multi-centre clinical trials

Markus Harden* and Tim Friede


## Abstract

**Background:** Multi-centre randomized controlled clinical trials play an important role in modern evidence-based medicine. Advantages of collecting data from more than one site are numerous, including accelerated recruitment and increased generalisability of results. Mixed models can be applied to account for potential clustering in the data, in particular when many small centres contribute patients to the study. Previously proposed methods on sample size calculation for mixed models only considered balanced treatment allocations which is an unlikely outcome in practice if block randomisation with reasonable choices of block length is used.

**Methods:** We propose a sample size determination procedure for multi-centre trials comparing two treatment groups for a continuous outcome, modelling centre differences using random effects and allowing for arbitrary sample sizes. It is assumed that block randomisation with fixed block length is used at each study site for subject allocation. Simulations are used to assess operation characteristics such as power of the sample size approach. The proposed method is illustrated by an example in disease management systems.

**Results:** A sample size formula as well as a lower and upper boundary for the required overall sample size are given. We demonstrate the superiority of the new sample size formula over the conventional approach of ignoring the multi-centre structure and show the influence of parameters such as block length or centre heterogeneity. The application of the procedure on the example data shows that large blocks require larger sample sizes, if centre heterogeneity is present.

**Conclusion:** Unbalanced treatment allocation can result in substantial power loss when centre heterogeneity is present but not considered at the planning stage. When only few patients by centre will be recruited, one has to weigh the risk of imbalance between treatment groups due to large blocks and the risk of unblinding due to small blocks. The proposed approach should be considered when planning multi-centre trials.


**Keywords:** Block randomisation, Linear mixed model, Random effects

## Background

When planning a randomized controlled clinical trial, sample size considerations are necessary to assess how many subjects are needed, e.g. to demonstrate a beneficial effect of a new treatment. These considerations usually are based on initial assumptions regarding a clinically meaningful treatment effect, the variability in the data and prespecified type I and type II error rates. Patients are often recruited in more than one centre, for example to account for a low incidence of the disease [1]. Since these centres might differ in some ways, between-centre heterogeneity at baseline needs to be accounted for in the analysis and therefore sample size planning.

When analysing continuous outcomes, baseline differences between centres can be accounted for using either a linear fixed-effects or a linear mixed model. Due to the central limit theorem, sample size calculation can often be based on the normal approximation

$$N = \frac{\sigma^2(k+1)^2}{k}\left(\frac{q_{1-\alpha/2}+q_{1-\beta}}{\mu^*}\right)^2 \quad (1)$$

where $N$ denotes the total sample size, $\mu^*$ the assumed treatment effect, $\sigma^2$ the variance of the observations, $k$ the allocation ratio between treatment groups and $q_\gamma$ the $\gamma$-quantile of the standard normal distribution [2].

*Correspondence: markus.harden@med.uni-goettingen.de
Department of Medical Statistics, University Medical Centre Göttingen, Humboldtallee 32, 37073 Göttingen, Germany

There have been various attempts to extend Formula (1) to multi-centre trials, e.g. by including a multiplicative factor to account for deviations from the standard design. Gallo as well as Ruvuna suggested such an inefficiency factor to account for centre size imbalances in the fixed effects setting [3, 4]. Both methods rely on the proportion of the treatment effects' variances between balanced and imbalanced centre sizes, where the balanced case gives optimal power.

Van Breukelen and colleagues introduced an inefficiency factor for the mixed model [5]. This factor is based on the relative efficiency of unequal versus equal cluster sizes for the weighted least squares estimator, assuming a linear mixed effects model with an interaction between study site and treatment effect. Fedorov and Jones consider sample size formulas for balanced multi-centre designs and suggest simulations for more complex situations [6]. Vierron and Giradeau suggested a *design-effect* to adjust Eq. (1) for different study designs [7, 8].

All of the approaches mentioned above assume balanced treatment allocation by centre. Randomisation techniques such as block randomisation do not guarantee equal group sizes in all centres, especially if centres are small and block lengths are large. The normal approximation in (1) gives a lower boundary of the necessary sample size, but underpowered trials could occur, especially when between centre heterogeneity is large. We believe that this assumption is too strict for real trials and therefore suggest a sample size formula that accounts for unequal sample sizes.

It has been demonstrated that mixed models tend to yield better results compared to fixed effects models, especially when the number of patients per centre is small [9, 10]. For a small number of centres, however, the fixed effects design might result in better results, because the between-centre variation is likely to be estimated with bias in mixed models in that situation. We therefore aim to construct a sample size formula that accounts for baseline heterogeneity between study centres, assuming a linear mixed model for multi-centre designs.

## Methods

### Statistical model and estimators

We assume a linear mixed-effects model with a fixed intercept $\mu_0$, random effects $u_j$, $j = 1, \ldots, c$ to account for centre heterogeneity at baseline, and a fixed treatment effect $\mu$. The data are assumed to follow some continuous distribution allowing for unequal sample sizes. The statistical model is given by

$$Y_{ijk} = \mu_0 + u_j + \mu \cdot x_i + \epsilon_{ijk} \tag{2}$$

for pairwise independent $u_j, \epsilon_{ijk}$ with $E(u_j) = 0$, $Var(u_j) = \tau^2 < \infty$, $E(\epsilon_{ijk}) = 0$, $Var(\epsilon_{ijk}) = \sigma^2 < \infty$, treatment indicator $x_i = 1_{\{i=2\}}$ for treatment groups $i = 1, 2$, centres $j = 1, \ldots, c$ and individuals $k = 1, \ldots, n_{ij}$ for each treatment-centre combination. The shared random effect $u_j$ within centres induces the following covariance matrix $Cov(Y_{111}, \ldots, Y_{2cn_{2c}}) = \bigoplus_{j=1}^{c} \left[\sigma^2 \mathbf{I}_{n_j} + \tau^2 \mathbf{J}_{n_j}\right]$, including all $N$ observations with $N = \sum_{i=1}^{2} \sum_{j=1}^{c} n_{ij}$. Here, $\mathbf{I}_{n_j}$ denotes the $n_j$-dimensional identity matrix and $\mathbf{J}_{n_j}$ the $n_j$-dimensional matrix consisting of ones only with $n_j = n_{1j} + n_{2j}$. We assume zero risk of contamination of the control group.

We are interested in differences between treatment groups and test the null hypothesis $H_0 : \mu = 0$ against the two-sided alternative $H_A : \mu \neq 0$. The distribution of the estimated treatment effect $\widehat{\mu}$ can be approximated by a normal distribution, if the sample size is sufficiently large (say sample sizes larger 30). It follows

$$T = \frac{\widehat{\mu}}{\sqrt{\widehat{\mathrm{Var}}\,(\widehat{\mu})}} \overset{H_0}{\sim} \mathrm{N}(0, 1). \tag{3}$$

The null hypothesis can be rejected if the test statistic $|T|$ exceeds the quantile $q_{1-\alpha/2}$ of the reference distribution for some fixed type I error rate $\alpha \in (0, 1)$. In order to apply the statistical test, suitable estimators for the unknown parameters have to be chosen.

We choose $\widehat{\mu} = \overline{Y}_{2\cdot\cdot} - \overline{Y}_{1\cdot\cdot}$ to measure treatment group differences, where $\overline{Y}_{i\cdot\cdot} = \frac{1}{N_i} \sum_{j=1}^{c} \sum_{k=1}^{n_{ij}} Y_{ijk}$ denotes the group mean in treatment group $i$. This estimator is unbiased, even if centres recruited patients for one treatment group only. The variance of $\widehat{\mu}$ can be written as

$$\mathrm{Var}\,(\widehat{\mu}) = \sigma^2 \frac{N}{N_1 N_2} + \tau^2 \sum_{j=1}^{c} \left( \frac{n_{1j}}{N_1} - \frac{n_{2j}}{N_2} \right)^2. \tag{4}$$

Details on the derivation can be found elsewhere [8].

$\mathrm{Var}\,(\widehat{\mu})$ depends on the overall sample size $N$, the variances $\sigma^2$ and $\tau^2$, and additionally the sample sizes by treatment group ($N_1$, $N_2$), number of study centres ($c$), and the sample sizes within study centres ($n_{1j}, n_{2j}$). In case of a perfectly balanced randomisation, the differences between centres cancel out and the treatment effect's variance only depends on sample sizes $N, N_1, N_2$ and variance $\sigma^2$, resulting in a sample size formula similar to (1).

The unknown variance parameters $\tau^2$ and $\sigma^2$ can be estimated using the following quadratic forms

$$\widehat{\sigma}^2 = \frac{1}{2c} \sum_{i=1}^{2} \sum_{j=1}^{c} \frac{1}{n_{ij} - 1} \sum_{k=1}^{n_{ij}} \left( Y_{ijk} - \overline{Y}_{ij\cdot} \right)^2$$

$$\widehat{\tau}^2 = \frac{1}{2} \sum_{i=1}^{2} \frac{1}{c-1} \sum_{j=1}^{c} \left( \overline{Y}_{ij\cdot} - \overline{Y}_{i\cdot\cdot} \right)^2.$$

### Additional assumptions for sample size calculation

$N_i$ and $n_{ij}$ are determined by recruitment and treatment allocation. We aim to replace all $N_i$, $n_{ij}$ in (4) by their expectations that can be calculated based on the randomisation procedure and planned allocation proportion between treatment groups.

In the following, we want to calculate the overall sample size $N$ and assume

1. a block randomisation stratified by centre with fixed block length $b$,
2. $k:1$ allocation ratio at each study site for $k \in \mathbb{N}$,
3. proportion of overall sample sizes between treatment groups according to allocation: $N_1 = kN_2$.

### Block randomisation

Since randomisation will not always result exactly in the planned allocation, we take a closer look at the randomisation process. Block randomisation with fixed block length $b$ is a procedure where every $b$ subjects get randomised between treatment groups at a time [11]. Complete blocks do always fulfil the planned $k:1$ allocation ratio. The block size should be unknown to investigators to strengthen the blinding in the trial.

In this article, we assume patients to be assigned to treatment groups $i = 1,2$ within centres, for a fixed $k:1$ allocation ratio. This means that in each randomisation block $b$ patients are randomized between treatment groups 1 and 2 in a way that for each patient receiving treatment 2, $k$ patients will receive treatment 1. The set of randomisation tuples $\Pi_b^k$ depends on block length $b$ and allocation parameter $k$. It is defined as follows

$$\Pi_b^k = \left\{ (x_1,\ldots,x_b) \in \Pi_b \,\middle|\, \sum_{\ell=1}^{b} 1_{\{x_\ell=1\}} = \frac{kb}{k+1} = b - \sum_{\ell=1}^{b} 1_{\{x_\ell=2\}} \right\} \tag{5}$$

where $\Pi_b := \{(x_1,\ldots,x_b) | x_\ell \in \{1,2\}\}$.

Treatment allocation imbalances can only occur in incomplete blocks with an upper boundary of $kb/(k+1)$. The choice of $k$ can be based on several assumptions as ethics, costs and other factors and will not be discussed further in this article. This topic is covered in more detail in a review by Dumville and colleagues [12]. It is available in many software packages and is therefore easy to apply [13, 14]. Further advantages and disadvantages of block randomisation are considered in the Discussion.

### Derivation of the sample size formula

The underlying idea of sample size calculation is to find the overall sample size $N$, such that the quantile $q_{1-\alpha/2}$ of the reference distribution under the null hypothesis equals the quantile $q_\beta^*$ of the reference distribution under a fixed alternative for type I and II error rates $\alpha$ and $\beta$.

Since we do assume a normally distributed test statistic, $q_\beta^*$ can be approximated by a shifted N(0, 1)-quantile $q_\beta^* \approx q_\beta + \frac{\mu}{\mathrm{Var}(\mu)}$ resulting in the following equation to construct a sample size formula

$$(q_{1-\alpha/2} + q_{1-\beta})^2 = \frac{\mu^2}{\mathrm{Var}(\mu)}. \tag{6}$$

By isolating the sample size $N$, which is part of $\mathrm{Var}(\mu)$, one can derive a sample size formula. In case of an ideal allocation, i. e., $n_{1j} = kn_{2j}$ for all centres, (6) is equal to (1). Since this is unlikely to be observed, unbalanced designs are taken into account by incorporating expectations with respect to the randomisation procedure.

We derive the sample size formula for the general case of a $k:1$ allocation ratio and assume the overall treatment group sample sizes to fulfil $N_1 = kN_2$ and therefore $N = (k+1)N_2$. This leads to the set of randomisation tuples in (5). By taking assumptions 1-3, the variance of $\widehat{\mu}$ given in (4) simplifies to

$$\mathrm{Var}\,(\widehat{\mu}) = \frac{\sigma^2(k+1)^2}{kN} + \frac{\tau^2(k+1)^2}{N^2}\sum_{j=1}^{c}\Delta_j^2 \tag{7}$$

where $\Delta_j^2 := \left(\frac{n_{1j}}{k} - n_{2j}\right)^2 \in [0, m^*]$ describe each centre's imbalance that will result from incomplete blocks with $m^* = b^2/(k+1)^2$. The (discrete) probability distribution of $\Delta_j^2$ depends on $\Pi_b^k$ and the number of patients in the last block which equals the remainder of the Euclidean division $r_j = n_j \bmod b$. An example of $\Delta_j^2|r_j$ for a single centre is illustrated in Fig. 1.

The probability distribution of $\Delta_j^2|r_j$ is fully described by block length $b$, allocation parameter $k$ and $r_j$. For planning purposes it therefore seems reasonable to replace $\Delta_j^2|r_j$ by its expectation $\mathrm{E}\left(\Delta_j^2|r_j\right)$ to eliminate sample sizes $n_{1j}$ and $n_{2j}$ from (7). The expectation of the probability distribution can easily be derived as

$$\mathrm{E}\left(\Delta_j^2|r_j\right) = \frac{1}{m^*}\sum_{\ell=0}^{m^*} p(\ell|r_j)\cdot\ell \tag{8}$$

where $p(\cdot|r_j)$ denotes the conditional density function of $\Delta_j^2|r_j$. These expectations are shown in Fig. 2 for a single randomization block, $k = 1,2,3$ and various block lengths $b$. Since the expected imbalance is the largest for $k = 1$ we will restrict simulations to this case.

The expected imbalance between treatment groups $\mathrm{E}\left(\Delta_j^2|r_j\right)$ increases with block length. This happens due to the fact that the probability to receive an incomplete randomisation block increases with increasing block length $b$. It is maximised when the last randomisation

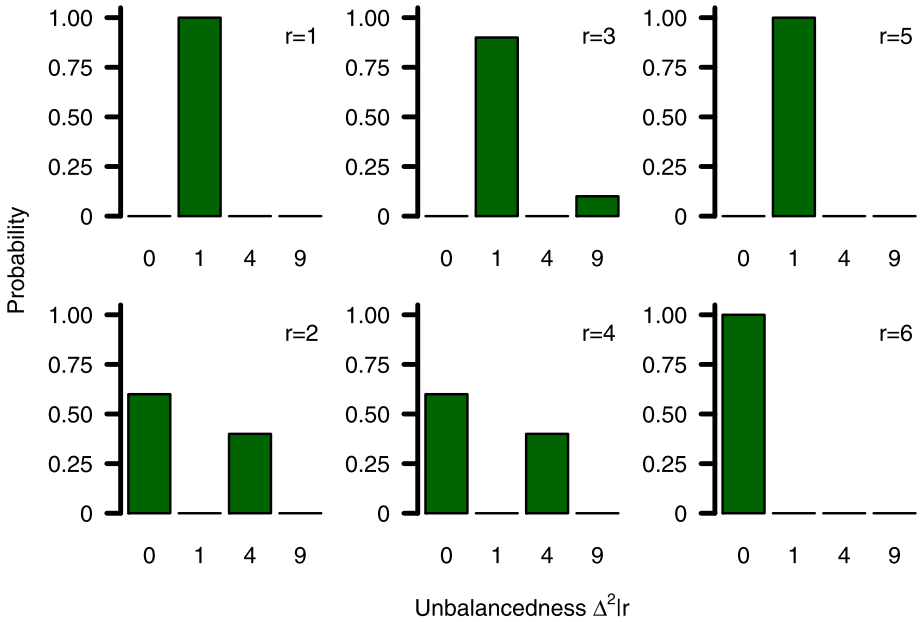


**Fig. 1** Probability distribution. Conditional probability distributions of $\Delta^2|r$ for varying numbers of randomized subjects $r = 1, \ldots, b = 6$

block only consists of $\frac{bk}{k+1}$ patients receiving treatment 1 or $\frac{b}{k+1}$ patients receiving treatment 2, respectively. If we replace $\Delta_j^2|r_j$ by $\mathrm{E}\left(\Delta_j^2|r_j\right)$, we can transform (6) into the following sample size formula for multi-centre trials (derivation is given in the appendix, see Additional file 1)

$$N_{\mathrm{MC}}^k = \frac{\sigma^2(k+1)^2}{2k}\left(\frac{q_{1-\alpha/2}+q_{1-\beta}}{\mu}\right)^2 + \sqrt{\frac{\sigma^4(k+1)^4}{4k^2} + \frac{\tau^2(k+1)^2\mu^2\sum_{j=1}^{c}\mathrm{E}\left(\Delta_j^2|r_j\right)}{\left(q_{1-\alpha/2}+q_{1-\beta}\right)^2}} \cdot \left(\frac{q_{1-\alpha/2}+q_{1-\beta}}{\mu}\right)^2. \quad (9)$$

## Simulations

### *General settings*

We perform a simulation study to assess the accuracy of the sample size formulas in terms of statistical power using R, version 3.3.1 [15]. For each scenario, we repeat $n_{\mathrm{sim}} = 10,000$ independent simulation runs. The R package `blockrand` is used for block randomisation [13]. All data are generated based on the statistical model described in (2), assuming a normal distribution for $u_j$ and $\epsilon_{ijk}$. The test statistic $T$ given in (3) is used for all power simulations and it is approximated by a standard normal distribution under the null hypothesis. The effect of block randomisation is strongest for $k = 1$, we therefore present simulation results for this setting only. The assumed effect size $\mu^*$ and values for variance components $\sigma^2$ and $\tau^2$ are based on an example trial described in the next section.

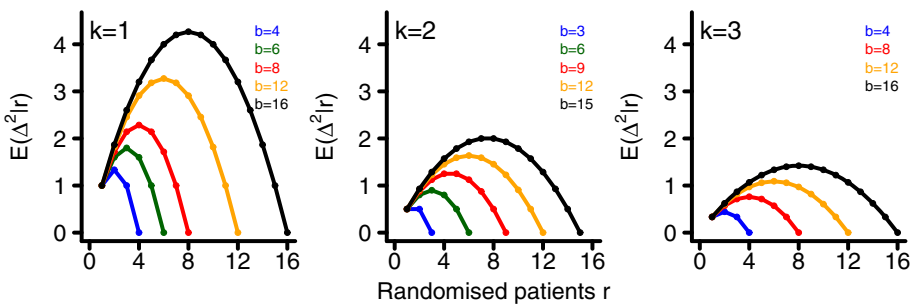


**Fig. 2** Expectation of $\Delta^2|r$. Conditional expected imbalance between treatment groups for allocation parameter $k = 1, 2, 3$ and various numbers of subjects $r = 1 \ldots, b$ and block lengths $b$

All code used for simulations is available in the Appendix, see Additional files 2, 3, 4, 5, 6, 7 and 8.

#### *Subject-to-centre allocation*

We consider equally as well as unequally sized study centres based on the following methods.

1. *Equally sized study centres:* Only in this situation, we can predict the sample size very precisely, since we can specify $r_j$ correctly prior to recruitment. Here, study centres are assumed to be equal. Since this assumption is limited to the fixed overall sample size $N$, we distribute the overall sample size to centres as follows

$$\text{Equal: } n_j = \left\lfloor \frac{N}{c} \right\rfloor + 1_{\{j \leq m\}}$$

for $m = N \bmod c$ and $j = 1, \ldots, c$.

2. *Unequally sized study centres:* In most experiments, the true allocation of patients cannot be foreseen. Recruitment rates can be estimated beforehand, but since only the number of patients in the last randomisation block affects the presented sample size formulas, no precise estimation of unbalanced treatment allocation can be made. To model this situation, study centres are assumed to be unequal but limited to the fixed overall sample size $N$. We use a multinomial distribution to generate unequal sample sizes by centre, assuming the following scenarios

$$\text{Unequal 1: } (n_1, \ldots, n_c)' \sim \text{Multi}_c\,(N, p_1, \ldots, p_c) \quad \text{with } p_j = \frac{1}{c}$$

$$\text{Unequal 2: } (n_1, \ldots, n_c)' \sim \text{Multi}_c\left(N, p_1^*, \ldots, p_c^*\right) \quad \text{with } p_j^* = \frac{p_j}{\sum_{k=1}^{c} p_k} \text{ and } p_j \overset{iid}{\sim} U[0;1]\,.$$

#### Example: The COMPETE II trial

Multi-centre trials are applied in many different disease areas. We present an example in the setting of disease management systems and use this trial to illustrate the sample size approach proposed.

Holbrook and colleagues conducted a randomized, multi-centre trial to investigate the benefit of an individualized electronic decision support system in adult patients diagnosed with type 2 diabetes [16]. This new intervention provided patient specific summaries and recommendations based on electronic medical records, aiming to improve the quality of diabetes management between patients and general practitioners. The tool was integrated into the practice work flow and offered web-based access by patients. In addition, an automated telephone reminder system was provided and all patients received a colour-coded printout quarterly. The control treatment consisted of usual care without use of this tool.

Primary outcome was a composite score difference compared to baseline. The composite score measured process quality on a scale from 0 to 10, based on the following parameters: blood pressure, cholesterol, glycated haemoglobin, foot check, kidney function, weight, physical activity, and smoking behaviour. The clinical targets are described in the original article. It was assessed at baseline and 6 months after randomization.

For this trial, 511 patients from 46 primary care providers were randomly assigned to intervention or control. At planning stage, the investigators aimed to recruit 508 patients to achieve 80% power to detect a difference of 1 for the primary outcome between treatment groups using a two-sided t-test with a type-1 error rate of $\alpha = 0.05$. No information on the assumed standard deviation is given in the article. Block-randomisation was stratified by study site in blocks of six, following a 1:1 allocation scheme.

The absolute measured improvement of composite scores between treatment groups was 1.26 (95% confidence interval (CI) 0.79-1.75; $p$-value $< 0.001$) favouring the new intervention.

## Results

### Approaches to sample size calculation

As long as no values for $r_j$ are assumed, Formula (9) cannot be used for sample size calculation. We consider a setting, where each centre will have at most one incomplete randomisation block. In the following, we present different ways to specify values of $\mathrm{E}\left(\Delta_j^2 | r_j\right)$ for each centre prior to recruitment.

The resulting sample size formulas are listed in Table 1. More detailed explanations are provided in the following subsections.

#### *Lower boundary*

Given an ideal treatment allocation ($n_{1j} = kn_{2j}$ for all centres), potential centre differences cancel out and do not affect the statistical power of the trial. In this case, sample size formula (1) would be suitable to plan the trial. However, the trial still had to be analysed with a mixed effects model to get an unbiased estimate for $\mathrm{Var}(\mu)$, since $\mathrm{Var}(Y_{ijk}) = \sigma^2 + \tau^2$. This formula is the standard approach to sample size calculation and is often used to plan multicentre trials. It therefore serves as one reference for sample size calculation and power simulations.

#### *Equal centre sizes*

The exact number of patients by study centre will be unknown at the planning stage, but in some situations, it might be reasonable to assume centres to be equally large.

**Table 1** Overview of sample size formulas

| | |
|---|---|
| Lower boundary | $N^k_{\text{lower}} = \left(\frac{q_{1-\alpha/2}+q_{1-\beta}}{\mu}\right)^2 \left(\frac{\sigma^2(k+1)^2}{k}\right)$ |
| Equal centres | $N^k_{\text{MC,E}} = \left(\frac{q_{1-\alpha/2}+q_{1-\beta}}{\mu}\right)^2 \left(\frac{\sigma^2(k+1)^2}{2k} + \sqrt{\frac{\sigma^4(k+1)^4}{4k^2} + \frac{\tau^2(k+1)^2\mu^2 c\,\mathrm{E}\left(\Delta_1^2\vert r_1\right)}{\left(q_{1-\alpha/2}+q_{1-\beta}\right)^2}}\right)$ |
| Unequal centres | $N^k_{\text{MC,U}} = \left(\frac{q_{1-\alpha/2}+q_{1-\beta}}{\mu}\right)^2 \left(\frac{\sigma^2(k+1)^2}{2k} + \sqrt{\frac{\sigma^4(k+1)^4}{4k^2} + \frac{\tau^2(k+1)^2\mu^2 c\overline{\mathrm{E}\left(\Delta_1^2\vert\cdot\right)}}{\left(q_{1-\alpha/2}+q_{1-\beta}\right)^2}}\right)$ |
| Upper boundary | $N^k_{\text{upper}} = \left(\frac{q_{1-\alpha/2}+q_{1-\beta}}{\mu}\right)^2 \left(\frac{\sigma^2(k+1)^2}{2k} + \sqrt{\frac{\sigma^4(k+1)^4}{4k^2} + \frac{\tau^2(k+1)^2\mu^2 c\,\mathrm{E}\left(\Delta_1^2\middle\vert\frac{b}{k+1}\right)}{\left(q_{1-\alpha/2}+q_{1-\beta}\right)^2}}\right)$ |

Lower and upper boundaries as well as sample size formulas for equal and unequal centre sizes

In this scenario, the sum of expected differences simplifies to a single quantity

$$\sum_{j=1}^{c} \mathrm{E}\left(\Delta_j^2|r_j\right) \approx c \cdot \mathrm{E}\left(\Delta_1^2|r_1\right) \tag{10}$$

which needs to be specified for sample size estimation.

Since the overall sample size $N$ and the number of subjects in the last randomisation block depend on each other, $r_1$ still has to be specified. The unspecified value of $\mathrm{E}\left(\Delta_1^2|r_1\right)$ can be determined by calculating the sample size $N^k_{\text{MC,E}}$ for each $r_1 \in \{1,\ldots,b\}$. Keep $N^k_{\text{MC,E}}\left(r_1^*\right)$ with

$$r_1^* = \underset{r_1}{\operatorname{argmin}} \left| \left( \frac{N^k_{\text{MC,E}}(r_1)}{c} \bmod b \right) - r_1 \right|. \tag{11}$$

***Unequal centre sizes***

For the general case, we suggest using the average of the expected imbalance for each centre

$$\mathrm{E}\left(\Delta_j^2|r_j\right) \approx \frac{1}{b}\sum_{\ell=1}^{b} \mathrm{E}\left(\Delta_1^2|\ell\right) =: \overline{\mathrm{E}\left(\Delta_1^2|\cdot\right)}. \tag{12}$$

This basically assumes a univariate distribution of $r_j$ on $[1,\ldots,b]$.

***Upper boundary***

We can identify an upper boundary of the sample size, given that all parameters are specified correctly at planning stage. The maximal imbalance between treatments would occur, if each centre recruited and allocated $r_j = b/(k+1)$ or $r_j = kb/(k+1)$ subjects to a single treatment group resulting in $\Delta_j^2 = m^*$. Therefore, the most conservative sample size calculation will be performed with the following approximation

$$\sum_{j=1}^{c} \mathrm{E}\left(\Delta_j^2|r_j\right) \approx c \cdot \mathrm{E}\left(\Delta_1^2\middle|\frac{b}{k+1}\right). \tag{13}$$

**Example: Sample size calculation**

To give an example of the application of the proposed sample size formula, we demonstrate the effects of between-centre heterogeneity combined with incomplete block randomisation based on the COMPETE II trial.

The number of patients per study site is reported in [17]. Based on those numbers, we know the completeness of all randomisation blocks by centre as shown in Fig. 3.

We use these numbers to illustrate the influence of unbalanced treatment group allocation on sample size and statistical power based on the assumed model. In total, 40 incomplete randomisation blocks ($r < 6$) out of 46 study sites occurred. Based on the assumed model, statistical power of the analysis might be reduced due to those 40 incomplete randomisation blocks, compared to a trial, where the same amount of patients would have been recruited at a single centre.

If we were to plan a new trial with similar features ($\mu = 1$, $\sigma = 4$, type 1 and type 2 error rates $\alpha = 0.05$ and $\beta = 0.2$, respectively) we could plug these values into the sample size formula $N^1_{\text{MC,U}}$ for unequal centre sizes. The assumption of uniformly distributed values of $r_j$ on $[1,\ldots,b]$ could be underpinned by a $\chi^2$-test for goodness-of-fit ($p = 0.4159$). The influence of intra-class correlation $\rho \in [0, 0.5]$, number of centres $c \in \{23, 46, 92\}$ and block lengths $b \in \{6, 8, 16\}$ on the sample size is shown in Fig. 4.

For a block length of $b = 6$, as chosen in the trial, no substantial influence of the intraclass correlation $\rho$ on the overall sample size can be observed. The reported value of $\rho = 0.08$ in the trial would not require a sample size adjustment compared to the standard approach ($N^k_{\text{lower}}$). For larger block lengths, however, a strong increase of the estimated sample size can be seen, especially for $\rho > 0.2$ and an increasing number of centres.

**Power simulations**

In addition to sample size calculations, we present some power simulation results for parameter settings based on the COMPETE II trial. We specify a treatment effect of $\mu = 1$, standard deviation $\sigma = 4$ and intraclass-correlation coefficient $\rho = 0.5$. Data was generated for various block lengths, numbers of centres and subject-to-centre allocation schemes. Resulting sample sizes and associated statistical power are given in Table 2 and Fig. 5.

Analyses based on the lower boundary formula do not achieve the planned power of 0.8. The deviance from the nominal level increases with block length and number of

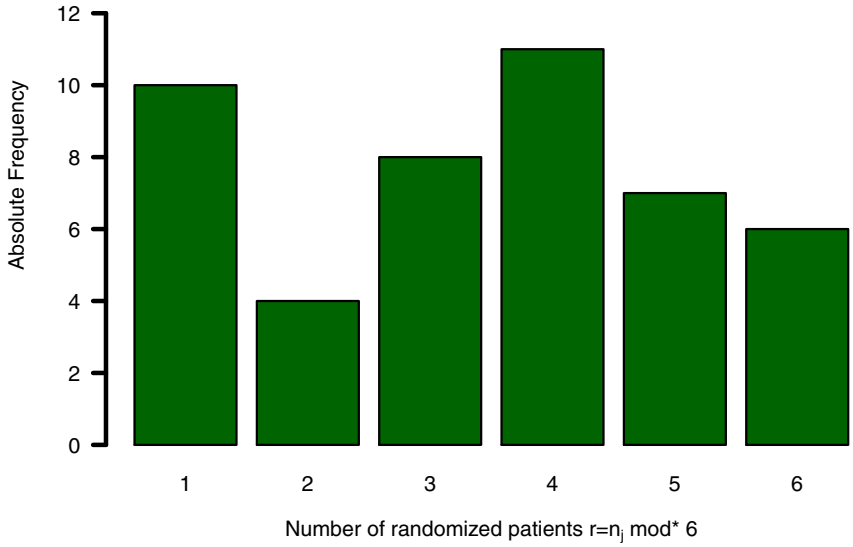


**Fig. 3** Example: Block randomisation. Number of final randomisation blocks by centre with block length $b = 6$ based on the COMPETE II trial [17]

centres. The upper boundary formula results in power levels that exceed the nominal level slightly.

The new sample size formulas for equal and unequal centre sizes lead to reasonable power simulation results. Even if subject allocation is performed randomly in each simulation run (Unequal 2), adequate statistical power can be obtained.

Subject allocation might, by chance, result in more complete randomisation blocks for $N^1_{\mathrm{MC,U}}$ than $N^1_{\mathrm{upper}}$. This leads to some situations where the formula with smaller sample size $\left(N^1_{\mathrm{MC,U}}\right)$ achieves higher (estimated) power. This observation underlines the necessity to take block length and patient recruitment into consideration when planning large multi-centre trials.

## Discussion

Patient enrolment and treatment allocation are key elements of every successful clinical trial. Randomisation techniques are used to achieve comparable treatment groups minimizing the risk of selection bias. Unfortunately, these randomisation procedures can result in unequal treatment group sizes and therefore a power loss compared to a balanced trial. Such imbalances cannot be determined prior to the trial, but we presented a way to estimate these values based on expectations.

The ICH E9 Guideline encourages the use of block randomisation and states the following on the choice of block sizes [1]: *"Care should be taken to choose block lengths that are sufficiently short to limit possible imbalance, but that*

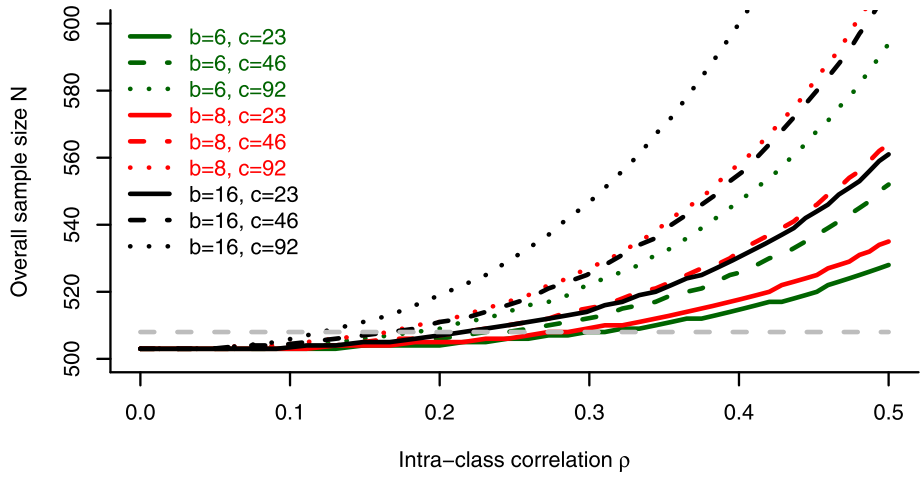


**Fig. 4** Example: Sample size based on $N^1_{\mathrm{MC,U}}$. Derived for $\mu = 1, \sigma = 4$, varying block length $b$, number of centres $c$ and intra-class correlation $\rho$. Dashed grey line represents the planned sample size of the trial ($N = 508$)

**Table 2** Example: Comparison of sample size formulas

| Block length | Formula | Number of centres | | |
|---|---|---|---|---|
| | | c=23 | c=46 | c=92 |
| 6 | $N^k_{\text{lower}}$ | 503 | 503 | 503 |
| | $N^k_{\text{MC,E}}$ | 525 | 524 | 569 |
| | $N^k_{\text{MC,U}}$ | 528 | 552 | 594 |
| | $N^k_{\text{upper}}$ | 541 | 575 | 634 |
| 8 | $N^k_{\text{lower}}$ | 503 | 503 | 503 |
| | $N^k_{\text{MC,E}}$ | 525 | 587 | 606 |
| | $N^k_{\text{MC,U}}$ | 535 | 564 | 616 |
| | $N^k_{\text{upper}}$ | 551 | 592 | 662 |
| 16 | $N^k_{\text{lower}}$ | 503 | 503 | 503 |
| | $N^k_{\text{MC,E}}$ | 586 | 603 | 762 |
| | $N^k_{\text{MC,U}}$ | 561 | 610 | 692 |
| | $N^k_{\text{upper}}$ | 587 | 654 | 762 |

Derived for $\mu = 1, \sigma = 4, \rho = 0.5$, varying block lengths and varying numbers of centres

*are long enough to avoid predictability towards the end of the sequence in a block. Investigators and other relevant staff should generally be blind to the block length; the use of two or more block lengths, randomly selected for each block, can achieve the same purpose."*

The results shown remain valid when using variable block sizes, since incomplete blocks can occur using either method. Only the determination of expected values $\mathrm{E}\left(\Delta^2|r\right)$ is more complicated for variable blocks, because it depends on the range of block sizes used.

In a recent systematic review on prevalence and reporting of recruitment, randomisation and treatment errors in phase III randomized, controlled trials, stratified block randomisation was identified as randomisation technique of choice in 50% of 82 included studies published in New England Journal of Medicine, Lancet, Journal of the American Medical Association, Annals of Internal

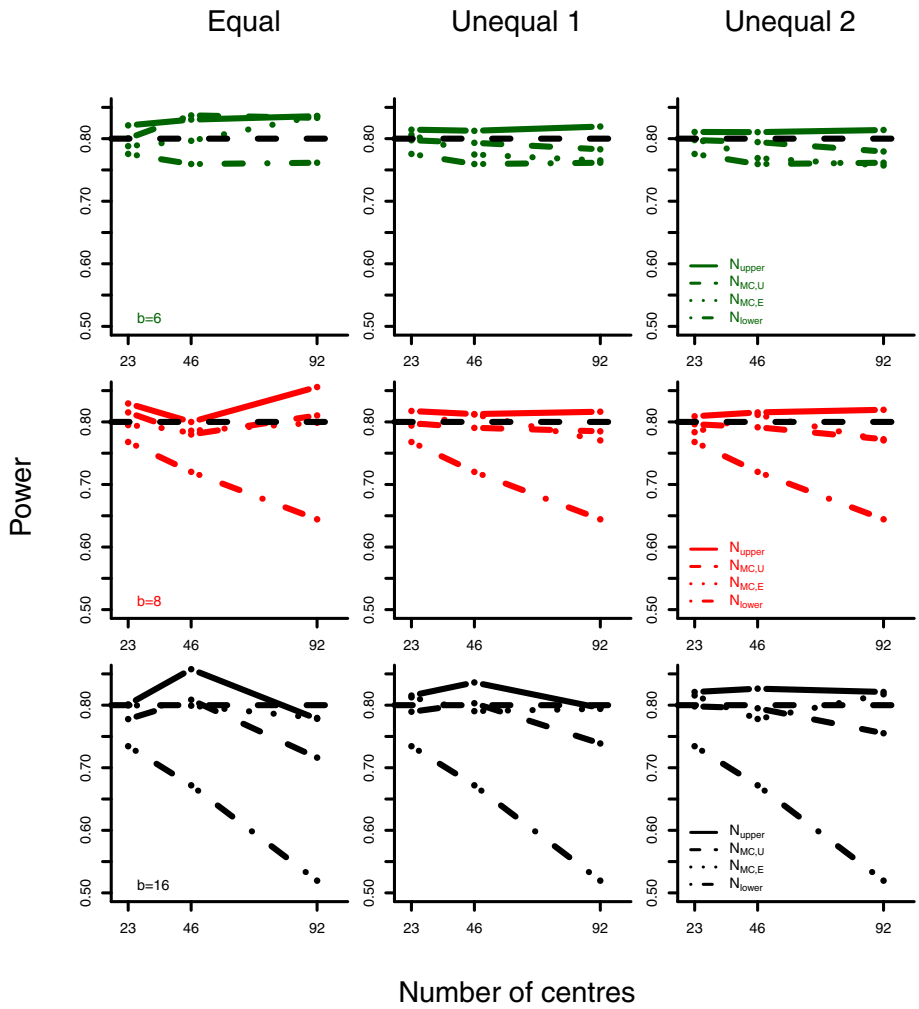


**Fig. 5** Example: Power simulations. Simulated power based on the planned sample sizes for $\mu = 1, \sigma = 4, \rho = 0.5$, varying numbers of centres and varying block lengths. Dashed black line represents the targeted power of 0.8

Medicine, or British Medical Journal between January and March 2015 [18]. The median number of participants per trial was 650 (range 40-84,496) and the number of centres varied between 1 and 1,161 with a median of 24. There are a number of trials that used fixed block lengths greater than 10 to allocate subjects between two treatment groups [19, 20]. Trials using random block randomisation almost never report the underlying block sizes, therefore we can not compare fixed versus random blocks any further. Overall, these observations support our idea to account for incomplete blocks to plan a multi-centre trial using either method for randomisation.

One limitation of our approach is a lack of knowledge on centre heterogeneity at the planning stage. There have been various articles with estimates of intraclass-correlation coefficients (ICC) derived from cluster-randomized trials. These estimates can be used to get an initial guess for centre heterogeneity in multicentre trials. A nice overview is given in the following text book [21]. Also, the implementation of an adaptive sample size reestimation procedure could account for this problem as it has been applied for cluster randomized trials and the fixed effects multi-centre trial design [22–24]. The development of sample size reestimation strategies based on the approach proposed here is subject to ongoing research. When planning an individually-randomized multi-centre trial there is a risk of control group contamination. This can partly be handled in placebo-controlled pharmacological trials or when proper blinding of patients and researchers is implemented [25]. Alternatively a cluster-randomized trial could be used to prevent contamination of treatment groups. This would, however, be associated with a higher sample size compared to the multi-centre design [26].

Here, we assumed a constant treatment effect across the centres. This is in line with the ICH E9 Guideline, which demands to avoid treatment-by-centre interactions in the primary analysis [1]. Therefore, this is at least for the planning of a trial an adequate assumption. Nevertheless, sensitivity analyses might explore treatment-by-centre interactions. Extending the sample size approach to a model including treatment-by-centre interactions is subject to future research.

## Conclusion

Imbalances in treatment allocation will lead to a power loss in multi-centre trials, if baseline heterogeneity is present. This risk can be accounted for when using appropriate methods for sample size calculation. To reduce uncertainty of sample size calculation, we recommend to calculate lower and upper boundaries in addition to the sample size.

## Additional files

**Additional file 1:** This file contains the calculation of the sample size formula. (PDF 24 kb)

**Additional file 2:** This file contains all source code used for simulations. (R 13 kb)

**Additional file 3:** This file contains the code used for Fig. 1. (R 3 kb)

**Additional file 4:** This file contains the code used for Fig. 2. (R 13 kb)

**Additional file 5:** This file contains the code used for Fig. 3. (R 1 kb)

**Additional file 6:** This file contains the code used for Fig. 4. (R 4 kb)

**Additional file 7:** This file contains the code used for Fig. 5. (R 12 kb)

**Additional file 8:** This file contains the code used for Table 2. (R 1 kb)

### Abbreviations

CI: Confidence interval; ICC: Intraclass-correlation coefficient

### Acknowledgements

The authors are grateful to Christian Röver for his helpful comments.

### Funding

This work was partly funded by the the German Federal Ministry of Education and Research project "Biostatistical Methods for efficient Evaluation of Individualised Therapies (BIMIT)", BMBF-05M13MGE.

### Availability of data and materials

All data generated or analysed during this study are included in this published article and its supplementary information files.

### Authors' contributions

MH and TF conceived the concept of this study, MH carried out the simulations and drafted the manuscript. TF critically reviewed and made substantial contributions to the manuscript. All authors commented on and approved the final manuscript.

### Ethics approval and consent to participate

Not applicable.

### Consent for publication

Not applicable.

### Competing interests

The authors declare that they have no competing interests.

## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Received: 9 May 2018 Accepted: 1 November 2018
Published online: 29 November 2018

### References

1. Ich: Ich harmonized tripartite guideline e9: Statistical principles for clinical trials. Stat Med. 1999;18:1905–42.
2. Julious SA, Owen RJ. Sample size calculations for clinical studies allowing for uncertainty about the variance. Pharm Stat. 2006;5(1):29–37.
3. Gallo PP. Practical issues in linear models analyses in multicenter clinical trials. Biopharm Rep. 1998;6:2–9.
4. Ruvuna F. Unequal center sizes, sample size, and power in multicenter clinical trials. Drug Inf J. 2004;38(4):387–94.
5. van Breukelen GJ, Candel MJ, Berger MP. Relative efficiency of unequal versus equal cluster sizes in cluster randomized and multicentre trials. Stat Med. 2007;26(13):2589–603.
6. Fedorov V, Jones B. The design of multicentre trials. Stat Methods Med Res. 2005;14(3):205–48.
7. Vierron E, Giraudeau B. Sample size calculation for multicenter randomized trial: taking the center effect into account. Contemp Clin Trials. 2007;28(4):451–8.

8. Vierron E, Giraudeau B. Design effect in multicenter studies: gain or loss of power? BMC Med Res Methodol. 2009;9(1):39.
9. Pickering RM, Weatherall M. The analysis of continuous outcomes in multi-centre trials with small centre sizes. Stat Med. 2007;26(30):5445–56.
10. Kahan BC, Morris TP. Analysis of multicentre trials with continuous outcomes: when and how should we account for centre effects? Stat Med. 2013;32(7):1136–49.
11. Matts JP, Lachin JM. Properties of permuted-block randomization in clinical trials. Control Clin Trials. 1988;9(4):327–44.
12. Dumville J, Hahn S, Miles J, Torgerson D. The use of unequal randomisation ratios in clinical trials: a review. Contemp Clin Trials. 2006;27(1):1–12.
13. Snow G. Blockrand: Randomization for Block Random Clinical Trials. 2013. R package version 1.3. https://CRAN.R-project.org/package=blockrand.
14. Asghari-Jafarabadi M, Sadeghi-Bazargani H. Randomization: techniques and software-aided implementation in medical studies. J Clin Res & Governance. 2015;4(2).
15. R Development Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2008. R Foundation for Statistical Computing. ISBN 3-900051-07-0. http://www.R-project.org.
16. Holbrook A, Thabane L, Keshavjee K, Dolovich L, Bernstein B, Chan D, Troyan S, Foster G, Gerstein H, Investigators CI, et al. Individualized electronic decision support and reminders to improve diabetes care in the community: Compete ii randomized trial. Can Med Assoc J. 2009;181(1-2):37–44.
17. Chu R, Thabane L, Ma J., Holbrook A, Pullenayegum E, Devereaux PJ. Comparing methods to estimate treatment effects on a continuous outcome in multicentre randomized controlled trials: a simulation study. BMC Med Res Methodol. 2011;11(1):21.
18. Yelland LN, Kahan BC, Dent E, Lee KJ, Voysey M, Forbes AB, Cook JA. Prevalence and reporting of recruitment, randomisation and treatment errors in clinical trials: a systematic review. Clin Trials. 2018;15(3):278. https://journals.sagepub.com/doi/10.1177/1740774518761627.
19. Yaxley JW, Coughlin GD, Chambers SK, Occhipinti S, Samaratunga H, Zajdlewicz L, Dunglison N, Carter R, Williams S, Payton DJ, et al. Robot-assisted laparoscopic prostatectomy versus open radical retropubic prostatectomy: early outcomes from a randomised controlled phase 3 study. The Lancet. 2016;388(10049):1057–66.
20. Döring G, Meisner C, Stern M, Group FVTS, et al. A double-blind randomized placebo-controlled phase iii study of a pseudomonas aeruginosa flagella vaccine in cystic fibrosis patients. Proc Natl Acad Sci. 2007;104(26):11020–25.
21. Teerenstra S, Moerbeek M. Power Analysis of Trials with Multilevel Data. Boca Raton: Chapman and Hall/CRC; 2015.
22. Lake S, Kammann E, Klar N, Betensky R. Sample size re-estimation in cluster randomization trials. Stat Med. 2002;21(10):1337–50.
23. Jensen K, Kieser M. Blinded sample size recalculation in multicentre trials with normally distributed outcome. Biom J. 2010;52(3):377–99.
24. van Schie S, Moerbeek M. Re-estimating sample size in cluster randomised trials with active recruitment within clusters. Stat Med. 2014;33(19):3253–68.
25. Moerbeek M. Randomization of clusters versus randomization of persons within clusters: which is preferable? Am Stat. 2005;59(2):173–9.
26. Torgerson DJ. Contamination in trials: is cluster randomisation the answer? Bmj. 2001;322(7282):355–7.

Received: 8 May 2019 | Revised: 8 October 2019 | Accepted: 4 January 2020

DOI: 10.1002/bimj.201900138

RESEARCH PAPER

Biometrical Journal

# Sample size recalculation in multicenter randomized controlled clinical trials based on noncomparative data

**Markus Harden**[1] | **Tim Friede**[1,2]

[1]Department of Medical Statistics, University Medical Centre Göttingen, Göttingen, Germany

[2]DZHK (German Center for Cardiovascular Research), Partner Site Göttingen, Göttingen, Germany

**Correspondence**
Markus Harden, Department of Medical Statistics, University Medical Centre Göttingen, Humboldtallee 32, 37073 Göttingen, Germany.
Email: markus.harden@med.uni-goettingen.de

This article has earned an open data badge "**Reproducible Research**" for making publicly available the code necessary to reproduce the reported results. The results reported in this article could fully be reproduced.

**Abstract**

Many late-phase clinical trials recruit subjects at multiple study sites. This introduces a hierarchical structure into the data that can result in a power-loss compared to a more homogeneous single-center trial. Building on a recently proposed approach to sample size determination, we suggest a sample size recalculation procedure for multicenter trials with continuous endpoints. The procedure estimates nuisance parameters at interim from noncomparative data and recalculates the sample size required based on these estimates. In contrast to other sample size calculation methods for multicenter trials, our approach assumes a mixed effects model and does not rely on balanced data within centers. It is therefore advantageous, especially for sample size recalculation at interim. We illustrate the proposed methodology by a study evaluating a diabetes management system. Monte Carlo simulations are carried out to evaluate operation characteristics of the sample size recalculation procedure using comparative as well as noncomparative data, assessing their dependence on parameters such as between-center heterogeneity, residual variance of observations, treatment effect size and number of centers. We compare two different estimators for between-center heterogeneity, an unadjusted and a bias-adjusted estimator, both based on quadratic forms. The type 1 error probability as well as statistical power are close to their nominal levels for all parameter combinations considered in our simulation study for the proposed unadjusted estimator, whereas the adjusted estimator exhibits some type 1 error rate inflation. Overall, the sample size recalculation procedure can be recommended to mitigate risks arising from misspecified nuisance parameters at the planning stage.

**KEYWORDS**

adaptive design, hierarchical model, internal pilot study, linear mixed model

## 1 | INTRODUCTION

Multicenter randomized controlled clinical trials are a key element of modern evidence-based medicine and the number of publications of such trials is constantly increasing (Danielsen, Okholm, Pommergaard, Burcharth, & Rosenberg, 2014). The need to recruit subjects at more than one location often arises from the need to reach a sufficient sample size in limited time.

Also, conducting a multicenter trial with multiple sites offers some form of replication within the trial, which is advantageous compared to monocenter trials to "provide a better basis for the subsequent generalization of its findings," see ICH (1998), Section 3.2. Compared to single-center trials, multicenter trials should account for potential heterogeneity between study sites. This heterogeneity can be considered as an additional nuisance parameter for planning and analysis purposes. The statistical model of such a trial can be implemented in a Gaussian linear fixed-effects or random-effects model, if a continuous outcome is considered. An advantage of a random-effects model is that only a single parameter has to be estimated compared to a fixed-effects model or a stratified analysis that will include at least one additional model parameter for each center. A fixed effects model should be used when only few centers are included, to avoid bias in estimation of between-center variation (Kahan & Morris, 2013).

Sample size formulas for multicenter trials have been described for both fixed and random-effects models. An overview of existing approaches and a new sample size formula is given in Harden and Friede (2018). In contrast to other approaches, this formula does not assume balanced numbers of observations across the treatment groups within centers. Since exact balance is unlikely to be observed in multicenter trials in practice, we believe that this is a useful approach to sizing multicenter trials.

Sample size calculations are typically informed by results from previous trials. Therefore, fixed study designs, for example, studies with an a priori fixed sample size are at risk to be planned with incorrect values when differences between previous trials and the new one occur. Adaptive study designs have been developed to reduce the risk of false negative study results, due to initial misspecification and have been adopted to various trial designs (Wassmer & Brannath, 2016, e.g., Chapters 2, 6, 9). One option is the implementation of an internal pilot study (Wittes & Brittain, 1990). This means that assumed values of nuisance parameters for the initial sample size calculation will be replaced by estimates from the accruing data during the course of the trial. They can be calculated in a blinded fashion using an estimator of the overall variance based on the data lumped across treatment groups, as proposed by Gould (1992), Kieser and Friede (2003), Zucker, Wittes, Schabenberger, and Brittain (1999), or based on unblinded data, for example, Wittes and Brittain (1990), Denne and Jennison (1999), Coffey and Muller (1999), Miller (2005). An overview to this kind of study design is given in Friede and Kieser (2006) and Proschan (2009). Since treatments cannot always be blinded in clinical trials but adaptations can still be based on the data pooled across the treatment groups, the terms "blinded" and "unblinded" are nowadays often replaced by "noncomparative" and "comparative," respectively, to indicate that adaptive designs are not limited to blinded trials. In the following, we will call these procedures comparative or noncomparative.

There are several guidance documents on the use of adaptive designs and sample size recalculation in particular. "Whenever possible, methods for blinded sample size reassessment that properly control the type 1 error rate should be used" as stated in EMA (2007), Section 4.2.2. It is mentioned further that "sufficient justification should be made," whenever unblinded data need to be reassessed. In a recent draft guidance on "Adaptive designs for clinical trials of drugs and biologics" by the U.S. Food and Drug Administration, it is said that "adequately prespecified adaptations based on noncomparative data have a negligible effect in the type 1 error probability. This makes them an attractive choice in many settings, particularly when uncertainty about event probabilities or endpoint variability is high," see FDA (2018), Section IV. Regarding comparative data, it is stated that "adaptations based on comparative data generally do directly increase the type 1 error probability and induce bias in treatment effect estimates. Therefore, statistical methods should take into account the adaptive trial design" (Section V, lines 414–416). The adaptation of sample sizes based on comparative data is suggested, when "there is considerable uncertainty about the true treatment effect size" (Section V.B). The use of adaptive designs is of course not limited to pharmacological interventions and specific guidance was released for example for device trials (FDA, 2016). Following these recommendations, we will present an approach for sample size recalculation based on noncomparative data to deal with uncertainties in nuisance parameters.

To the best of our knowledge, only few approaches for sample size recalculation in multicenter trials have been suggested so far. Shih and Long (1998) consider multicenter trials with unequal variances between treatment groups but equal group sizes within centers. Jensen and Kieser (2010) presented a sample size recalculation procedure based on a sample size formula by Ruvuna (2004) for the fixed effects model. In this article, we aim to apply the fixed sample size formula for the mixed effects model described in Harden and Friede (2018) to an internal pilot study approach to allow for sample size recalculation during the trial.

The manuscript is organized as follows. We introduce the COMPETE II trial as the motivating example in Section 2. Section 3 defines the statistical model and describes the sample size formula for the fixed study design. We extend the sample size formula to a sample size recalculation procedure in Section 4. For the sake of completeness, we will look at sample size recalculation based on comparative, as well as noncomparative data. In Section 5, we assess operation characteristics of the new approach based on a simulation study with parameters similar to the COMPETE II trial. We discuss the results and close with a conclusion in Section 6.

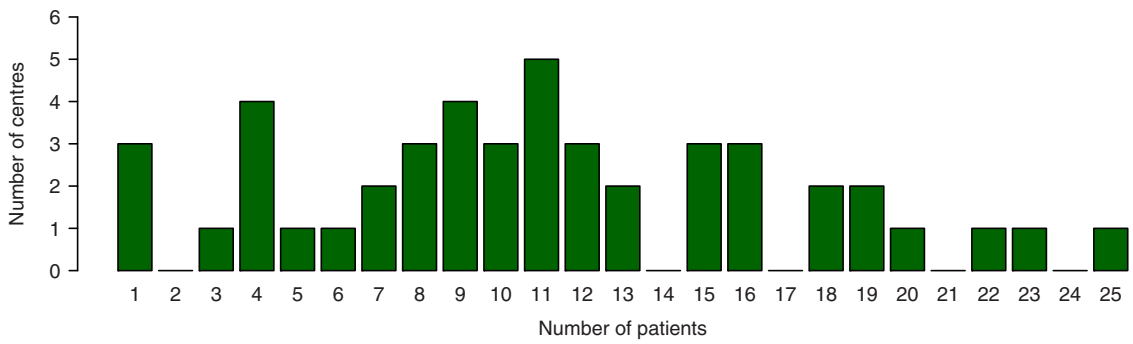


**FIGURE 1** Number of patients by center recruited for the COMPETE II trial

## 2 | MOTIVATING EXAMPLE: THE COMPETE II TRIAL

Holbrook et al. (2009) investigated the benefit of an individualized electronic decision support system in adult patients diagnosed with type 2 diabetes. Primary outcome was a composite score difference compared to baseline measuring process quality on a scale from 0 to 10, based on the following parameters: blood pressure, cholesterol, glycated haemoglobin, foot check, kidney function, weight, physical activity, and smoking behavior. The clinical targets are described in the original article. It was assessed twice, at baseline and 6 months after randomization.

For this multicenter trial, 511 patients from 46 primary care providers (here referred to as centers) were randomly assigned to intervention or control. Block-randomization was stratified by study site in blocks of six, following a 1:1 allocation scheme. The number of patients by study center are published in another article by Chu et al. (2011) and displayed in Figure 1. At the planning stage, the investigators aimed to recruit 508 patients to achieve 80% power assuming a difference of 1 for the primary outcome between treatment groups using a two-sided $t$-test with a significance level of $\alpha = 0.05$.

We will illustrate ideas and simulations using this example. Obvious parameters that can be reestimated from interim data are the within- and the between-group variance. Additionally, information on subject recruitment can be retrieved at the interim stage. We want to explore whether this information can be used to improve sample size recalculation.

## 3 | STATISTICAL MODEL AND FIXED SAMPLE SIZE CALCULATION

The statistical model is based on a linear mixed-effects model described as follows

$$Y_{ijk} = \mu_0 + u_j + \mu \cdot x_i + \epsilon_{ijk}$$

with fixed intercept $\mu_0$, treatment effect $\mu$, treatment indicator $x_1 = 0$ and $x_2 = 1$, and pairwise independent $u_j$, $\epsilon_{ijk}$ satisfying $\mathrm{E}(u_j) = 0, \mathrm{Var}(u_j) = \tau^2 < \infty, \mathrm{E}(\epsilon_{ijk}) = 0, \mathrm{Var}(\epsilon_{ijk}) = \sigma^2 < \infty$ and $\mathrm{Cov}(u_j, \epsilon_{ij'k}) = 0$. Indices refer to treatment groups $i = 1, 2$, centers $j, j' = 1, \dots, c$ and subjects within centers and treatment groups $k = 1, \dots, n_{ij}$, sample sizes are defined as $n_j = n_{1j} + n_{2j}$, $N_i = \sum_{j=1}^{c} n_{ij}$ and $N = N_1 + N_2$. Subjects within centers share a common random effect that creates a block-diagonal covariance matrix given by $\mathrm{Cov}(Y_{111}, \dots, Y_{2cn_{2c}}) = \bigoplus_{j=1}^{c} (\sigma^2 \mathbf{I}_{n_j} + \tau^2 \mathbf{J}_{n_j})$ with the $n_j$-dimensional identity matrix $\mathbf{I}_{n_j}$ and the $n_j$-dimensional matrix $\mathbf{J}_{n_j}$ consisting of ones only. The structure of the study design is shown in Table 1.

In addition to the statistical model, the following assumptions are necessary to use the closed sample size formula as presented in Harden and Friede (2018):

1. Block randomization with fixed block length $b$, locally applied at each center for treatment allocation,
2. fixed $R : 1$ allocation ratio for each randomization block with $R \in \mathbb{N}$,
3. the proportion of overall sample sizes between treatment groups fulfils the assumed allocation ratio $N_1 = RN_2$.

The unknown parameter of interest $\mu$ can be estimated in an unbiased fashion based on overall treatment group means $\widehat{\mu} = \overline{Y}_{2\cdot\cdot} - \overline{Y}_{1\cdot\cdot}$, with $\overline{Y}_{i\cdot\cdot} = \sum_{j,k} Y_{ijk}/N_i$ for $i = 1, 2$. The variance of $\widehat{\mu}$ has been calculated elsewhere, see Vierron and Giraudeau

**TABLE 1** Study design of a multicenter trial for a two treatment comparison

| Center | Treatment | Observations | | | | Sample sizes by center |
|---|---|---|---|---|---|---|
| 1 | 1 | $Y_{111}$ | $Y_{112}$ | … | $Y_{11n_{11}}$ | $n_{11}$ |
| | 2 | $Y_{211}$ | $Y_{212}$ | … | $Y_{21n_{21}}$ | $n_{21}$ $\Big\} =: n_1$ |
| ⋮ | | | | | | |
| $j$ | 1 | $Y_{1j1}$ | … | $Y_{1jn_{1j}}$ | | $n_{1j}$ |
| | 2 | $Y_{2j1}$ | … | | $Y_{2jn_{2j}}$ | $n_{2j}$ $\Big\} =: n_j$ |
| ⋮ | | | | | | |
| $c$ | 1 | $Y_{1c1}$ | … | | $Y_{1cn_{1c}}$ | $n_{1c}$ |
| | 2 | $Y_{2c1}$ | … | $Y_{2cn_{2c}}$ | | $n_{2c}$ $\Big\} =: n_c$ |

(2009), Appendix I, and is given by

$$\mathrm{Var}\left(\widehat{\mu}\right) = \sigma^2 \frac{N}{N_1 N_2} + \tau^2 \sum_{j=1}^{c} \left( \frac{n_{1j}}{N_1} - \frac{n_{2j}}{N_2} \right)^2 .$$

The unknown nuisance parameters $\sigma^2$ and $\tau^2$ can be estimated by

$$\widehat{\sigma}^2 = \frac{1}{N-2c} \sum_{i=1}^{2} \sum_{j=1}^{c} \sum_{k=1}^{n_{ij}} \left( Y_{ijk} - \overline{Y}_{ij.} \right)^2 \text{ and} \tag{1}$$

$$\widehat{\tau}^2 = \frac{1}{2(c-1)} \sum_{i=1}^{2} \sum_{j=1}^{c} \left( \overline{Y}_{ij.} - \overline{Y}_{i..} \right)^2 , \tag{2}$$

resulting in the following test statistic, which follows asymptotically a normal distribution

$$T = \frac{\widehat{\mu}}{\sqrt{\widehat{\sigma}^2 \frac{N}{N_1 N_2} + \widehat{\tau}^2 \sum_{j=1}^{c} \left( \frac{n_{1j}}{N_1} - \frac{n_{2j}}{N_2} \right)^2}} \approx N(\delta, 1). \tag{3}$$

Under the null hypothesis of no treatment effect $\mathrm{H}_0 : \mu = 0$, $T$ follows asymptotically a standard normal distribution, while $\delta = \sqrt{N}\mu/\sigma$ under the two-sided alternative $\mathrm{H}_\mathrm{A} : \mu \neq 0$. The null hypothesis will be rejected, if $|T| > q_{1-\alpha/2}$ for $\alpha \in (0, 1)$, where $q_\gamma$ denotes the $\gamma$-quantile of the standard normal distribution. We determined a sample size formula for $\widehat{\mu}$ in a recent article (Harden & Friede, 2018) given by

$$N_{\mathrm{MC}} = \left( \frac{q_{1-\alpha/2} + q_{1-\beta}}{\mu^*} \right)^2 \left( \frac{\sigma^2 (R+1)^2}{2R} + \sqrt{\frac{\sigma^4 (R+1)^4}{4R^2} + \frac{[\tau(R+1)\mu^*]^2 \sum_{j=1}^{c} \Delta_j^2}{\left( q_{1-\alpha/2} + q_{1-\beta} \right)^2}} \right), \tag{4}$$

where $\mu^*$ describes the assumed treatment effect and $\Delta_j^2 := (n_{1j}/R - n_{2j})^2$ the deviation between planned and observed allocation ratio within centers. Nuisance parameters $\sigma^2$ and $\tau^2$ as well as deviations $\Delta_j^2$ are unknown at the planning stage and must be replaced by reasonable guesses prior to the trial. For instance, $\Delta_j^2$ could be replaced by expected values as shown in Harden and Friede (2018). If sample sizes within centers match the planned allocation ratio, that is, $\Delta_j^2 = 0$ for all $j$, between-center heterogeneity will not decrease the statistical power of the trial and (4) reduces to

$$N_{\mathrm{lower}} = \left( \frac{q_{1-\alpha/2} + q_{1-\beta}}{\mu^*} \right)^2 \frac{\sigma^2 (R+1)^2}{R} .$$

An upper boundary of the required sample size is given, if every center would contribute an incomplete randomization block with $b/(R+1)$ subjects receiving the same treatment. Here, we assume that every center can at most contribute one incomplete block.

# 4 | SAMPLE SIZE RECALCULATION

We use Formula (4) to construct a sample size recalculation procedure for multicenter trials with continuous outcomes and arbitrary sample sizes per treatment arm and center. The goal of an internal pilot study is to gain information on nuisance parameters of the trial during recruitment to improve sample size calculations.

## 4.1 | What can be learned from noncomparative interim data?

In order to achieve the planned statistical power, we will analyze what can be learned based on interim data to improve nuisance parameter estimation. In the considered multicenter design, we assume values for the variability between observations $\sigma^2$ and the dependency of observations within study sites $\tau^2$. The influence of $\tau^2$ on the statistical power is influenced by sample size deviations within centers $\Delta_j^2$. These values require knowledge of the treatment group allocation, that is, unblinding of the data. In a sample size review based on noncomparative data, we can, however, estimate the distribution of $\Delta_j^2$ at the interim stage and impute $\mathrm{E}(\Delta_j^2)$ into Formula (4) for sample size recalculation. In this article, the number of centers $c$ is assumed to be fixed.

## 4.2 | Nuisance parameter estimation

We suggest the following noncomparative quadratic forms to estimate unknown nuisance parameters $\sigma^2$ and $\tau^2$ at interim

$$\widehat{\sigma}_b^2 = \frac{1}{N-c}\sum_{i=1}^{2}\sum_{j=1}^{c}\sum_{k=1}^{n_{ij}}\left(Y_{ijk} - \overline{Y}_{\cdot j\cdot}\right)^2 \text{ and} \tag{5}$$

$$\widehat{\tau}_b^2 = \frac{1}{c-1}\sum_{j=1}^{c}\left(\overline{Y}_{\cdot j\cdot} - \overline{Y}_{\cdots}\right)^2. \tag{6}$$

When a comparative estimation of nuisance parameters seems appropriate, estimators described in Formulas (1) and (2) can be applied. All these estimators do not assume any distribution function and can be calculated as long as $n_j > 2\ \forall j$ ($n_{ij} > 2\ \forall j$) based on noncomparative (comparative) data. Some of these estimators are biased with the following expected values

$$\mathrm{E}\left(\widehat{\sigma}^2\right) = \sigma^2,$$

$$\mathrm{E}\left(\widehat{\sigma}_b^2\right) = \sigma^2 + \frac{\mu^2}{N-c}\sum_{j=1}^{c}\frac{n_{1j}n_{2j}}{n_j},$$

$$\mathrm{E}\left(\widehat{\tau}^2\right) = \tau^2 + \frac{\sigma^2}{2c}\sum_{i=1}^{2}\sum_{j=1}^{c}\frac{1}{n_{ij}} \text{ and}$$

$$\mathrm{E}\left(\widehat{\tau}_b^2\right) = \tau^2 + \frac{\sigma^2}{c}\sum_{j=1}^{c}\frac{1}{n_j} + \frac{\mu^2}{c-1}\sum_{j=1}^{c}\left(\frac{n_{2j}}{n_j} - \frac{1}{c}\sum_{\ell=1}^{c}\frac{n_{2\ell}}{n_\ell}\right)^2.$$

Detailed derivations are given in the Appendix. Friede and Kieser (2001) and others showed that the bias of the noncomparative variance estimator $\widehat{\sigma}_b^2$ is neglectable in typical situations for clinical trials. The bias of estimators for $\tau^2$ can partly be adjusted using estimators

$$\tilde{\tau}^2 = \max\left\{\widehat{\tau}^2 - \frac{\widehat{\sigma}^2}{2c}\sum_{i=1}^{2}\sum_{j=1}^{c}\frac{1}{n_{ij}}; 0\right\} \tag{7}$$

$$\tilde{\tau}_b^2 = \max\left\{\widehat{\tau}_b^2 - \frac{\widehat{\sigma}_b^2}{c}\sum_{j=1}^{c}\frac{1}{n_j}; 0\right\}. \tag{8}$$

In the following, we will refer to estimators $\hat{\tau}^2$ and $\hat{\tau}_b^2$ as *unadjusted* estimators, while referring to $\tilde{\tau}^2$ and $\tilde{\tau}_b^2$ as *adjusted* estimators of $\tau^2$. We will assess how these biases affect the sample size determination process and if corrections shown in Formulas (7) and (8) improve results using simulation studies.

### 4.3 | Incomplete randomization blocks within centers

In addition to nuisance parameters, the unknown dispersion by center $\Delta_j^2$ has to be be determined during sample size review. Since the quantity is defined by sample sizes $n_{ij}$ which are unknown at interim, ideas similar to Harden and Friede (2018) can be used to calculate the expectation of the distribution of $\Delta_j^2$ which only depends on block length $b$, allocation ratio $R$ and the number of patients in the last randomization block $r_j$, all of which are available from noncomparative data. In order to investigate how well the distribution of $\Delta_j^2$ can be estimated at an interim stage, we use the COMPETE II trial-based center sizes to compare the distribution of block lengths at interim stages round$\{\rho \cdot (n_1, \dots, n_c)\}$, for $\rho \in (0.2, 0.3, 0.4, 0.5)$ to the ones displayed in Figure 1. The results are shown in Figure 2.

It can be seen that the distribution of block lengths at any interim stage does not look alike the final distribution after including all 511 subjects. We therefore recommend to assume a uniform distribution of $r_j$ on values $\{1, \dots, b\}$ for sample size calculation in general, leading to the following approximation

$$\sum_{j=1}^{c} \Delta_j^2 \approx \frac{c}{b} \sum_{k=1}^{b} \mathrm{E}\left(\Delta_j^2 \middle| r_k = k\right).$$

### 4.4 | Recalculation procedure

All parameters in Formula (4) are either known or can be estimated as described above. For sample size recalculation noncomparative, as well as comparative nuisance parameter estimators as listed in Formulas (1), (2), (5), (6), (7), and (8) can be applied. The sample size recalculation procedure is executed as follows:

(i) Calculate initial sample size $N_{\text{init}}$ using formula $N_{\text{MC}}$ based on initial assumptions on $\alpha$, $\beta$, $\mu^*$, $\sigma^2$, $\tau^2$, and $\Delta_j^2$ and specify a number of subjects $N_{\text{BSSR}} = \rho \cdot N_{\text{init}}$ which will be used to recalculate nuisance parameters with $\rho \in (0, 1)$.

(ii) Estimate nuisance parameters $\sigma^2$, $\tau^2$ based on the first $N_{\text{BSSR}}$ subjects recruited.

(iii) Calculate sample size $N_1$ using formula $N_{\text{MC}}$ based on initial $\alpha$, $\beta$, and $\mu$ and estimated $\sigma^2$, $\tau^2$.

(iv) Recruit additional subjects into the study until $N_{\text{final}} = \max(N_1; N_{\text{BSSR}})$ is reached. If an upper limit $N_{\max}$ is given for recruitment, $N_{\text{final}} = \min\{\max(N_1; N_{\text{BSSR}}); N_{\max}\}$.

(v) Perform final analysis based on $N_{\text{final}}$ subjects.

## 5 | SIMULATION STUDY

### 5.1 | General settings

A simulation study is carried out to assess the operation characteristics of the sample size recalculation procedure. All analyses are performed using `R`, version 3.6.1 (R Development Core Team, 2008). Each simulation scenario consists of $n_{\text{sim}} = 10\,000$ simulation runs to estimate the type 1 and 2 error rates. For data generation, the R-package `blockrand` is used for block randomization (Snow, 2013). A normal distribution is assumed for both the random effects $u_j$ and observation errors $\epsilon_{ijk}$. Source code to reproduce the results is available as Supporting Information online at https://onlinelibrary.wiley.com/doi/10.1002/bimj.201900138 at the end of the article.

The parameter settings of the generated data are motivated by the COMPETE II trial. Block length is chosen to be large to show the benefit of the new approach, which comes into play for unbalanced data. If not stated otherwise, we set the treatment effect to $\mu = 1$, nuisance parameters $\sigma^2 = \tau^2 = 16$, block length $b = 16$ and the allocation ratio $R = 1$. The number of centers $c$, timing of the sample size recalculation $\rho$ and the assumed treatment effect under the alternative for type 1 error simulations will vary. We generate unequally sized study centers distributing a fixed overall sample size $N$ to $c$ centers based on a multinomial

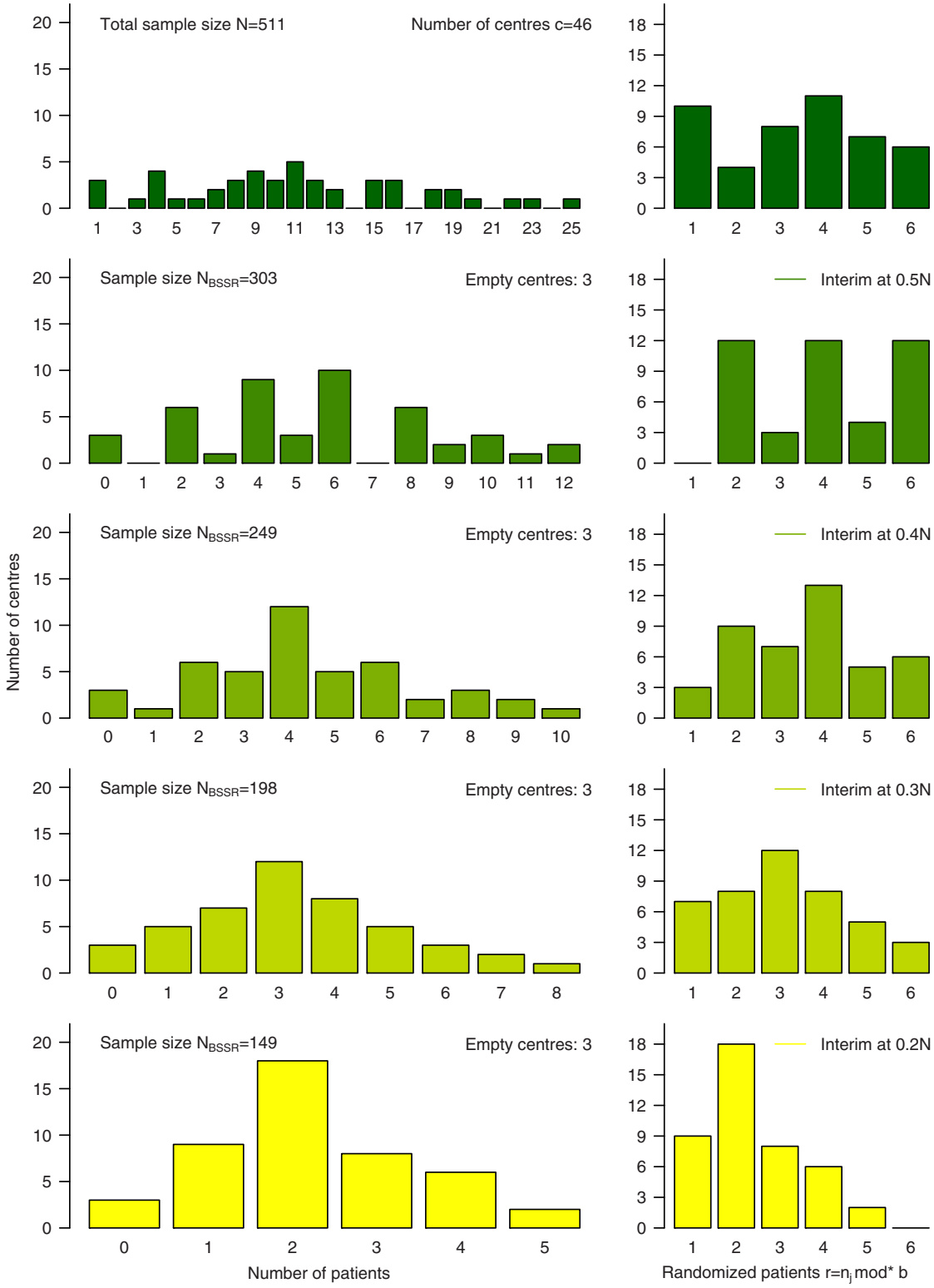


**FIGURE 2** COMPETE II trial-based center sizes for different time points of interim analysis. We define $b$ mod* $b = b$, since zero subjects refer to a previous block with $b$ subjects, assuming all centers are not empty

**TABLE 2** Overview of nuisance parameter estimates used for sample size recalculation

| | | Estimator for | |
|---|---|---|---|
| **Noncomparative** | **Adjusted** | $\sigma^2$ | $\tau^2$ |
| No | No | $\widehat{\sigma}^2$ | $\widehat{\tau}^2$ |
| Yes | No | $\widehat{\sigma}_b^2$ | $\widehat{\tau}_b^2$ |
| No | Yes | $\widehat{\sigma}^2$ | $\tilde{\tau}^2$ |
| Yes | Yes | $\widehat{\sigma}_b^2$ | $\tilde{\tau}_b^2$ |

**TABLE 3** Sample sizes without sample size recalculation for nuisance parameters $\sigma^2 = \tau^2 = 16$ and varying treatment effects $\mu^*$ and number of centers $c$

| | Treatment effect $\mu^*$ | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Centers** | **0.82** | **0.9** | **1** | **1.11** | **1.22** | **1.35** | **1.49** | **1.65** |
| 1 | 750 | 624 | 506 | 410 | 340 | 278 | 230 | 188 |
| 10 | 775 | 640 | 530 | 435 | 364 | 302 | 252 | 210 |
| 20 | 800 | 673 | 554 | 459 | 387 | 324 | 274 | 230 |

| | Treatment effect $\mu^*$ (continued) | | | | | | |
|---|---|---|---|---|---|---|---|
| **Centers** | **1.82** | **2.01** | **2.23** | **2.46** | **2.72** | **3** | **3.32** |
| 1 | 154 | 128 | 104 | 86 | 70 | 58 | 48 |
| 10 | 177 | 149 | 125 | 106 | 90 | 77 | 66 |
| 20 | 196 | 167 | 142 | 122 | 105 | 91 | 79 |

distribution with random center sizes as described in Jensen and Kieser (2010), that is,

$$\left(n_1, \ldots, n_c\right)' \sim \text{Multi}_c\left(N, p_1^*, \ldots, p_c^*\right) \text{ with } p_j^* = \frac{p_j}{\sum_{k=1}^{c} p_k} \text{ and } p_j \overset{iid}{\sim} U[0;1],$$

but force $n_j > 0$ for every center $j$. The different nuisance parameter estimators given in Section 4.2 are compared and we will present the difference between sample size recalculation based on noncomparative and comparative data. We consider different estimators for nuisance parameters as listed in Table 2. Also, the test statistic described in (3) will be calculated using the adjusted as well as the unadjusted estimator for $\tau^2$. We will refer to the test statistic in accordance to the nuisance estimator of $\tau^2$.

## 5.2 | Type 1 error rate

In this section, we present simulation results for the estimated type 1 error rate. We simulate data from $c = 10$ and $c = 20$ centers with varying sample sizes determined by $\mu^* \in \{0.82, \ldots, 3.32\}$ and a varying proportion of data used for sample size recalculation $\rho \in \{0.3, \ldots, 0.8\}$. Initial sample sizes before sample size recalculation for varying center sizes are shown in Table 3. For the purpose of comparison, we included sample sizes for a balanced t-test in that table (centers = 1). The results for a prespecified two-sided type 1 error rate of $\alpha = 0.05$ are shown in Figures 3 and 4. Each subfigure contains four panels that represent results for noncomparative and comparative, as well as adjusted and unadjusted estimators for the nuisance parameters $\sigma^2$ and $\tau^2$ described in (1) – (8), which were applied for sample size recalculation.

The unadjusted test statistic controls the type 1 error rate and shows some conservative behavior for larger treatment effects and an increasing number of centers. The use of noncomparative or comparative nuisance estimators does not seem to affect the estimated type 1 error rate in situations considered for this simulation study. The adjusted estimator of $\tau^2$ at interim does not affect the estimated type 1 error rate either.

When using the adjusted test statistic for the final analysis, some inflation of the estimated type 1 error rate can be observed for fewer centers and larger treatment effects. Again, no influence of noncomparative/comparative or adjusted/unadjusted nuisance parameters used for sample size recalculation can be seen for the simulated type 1 error. Due to the possible inflation of the type 1 error rate of the adjusted test statistic, we do not consider it for power analyses.

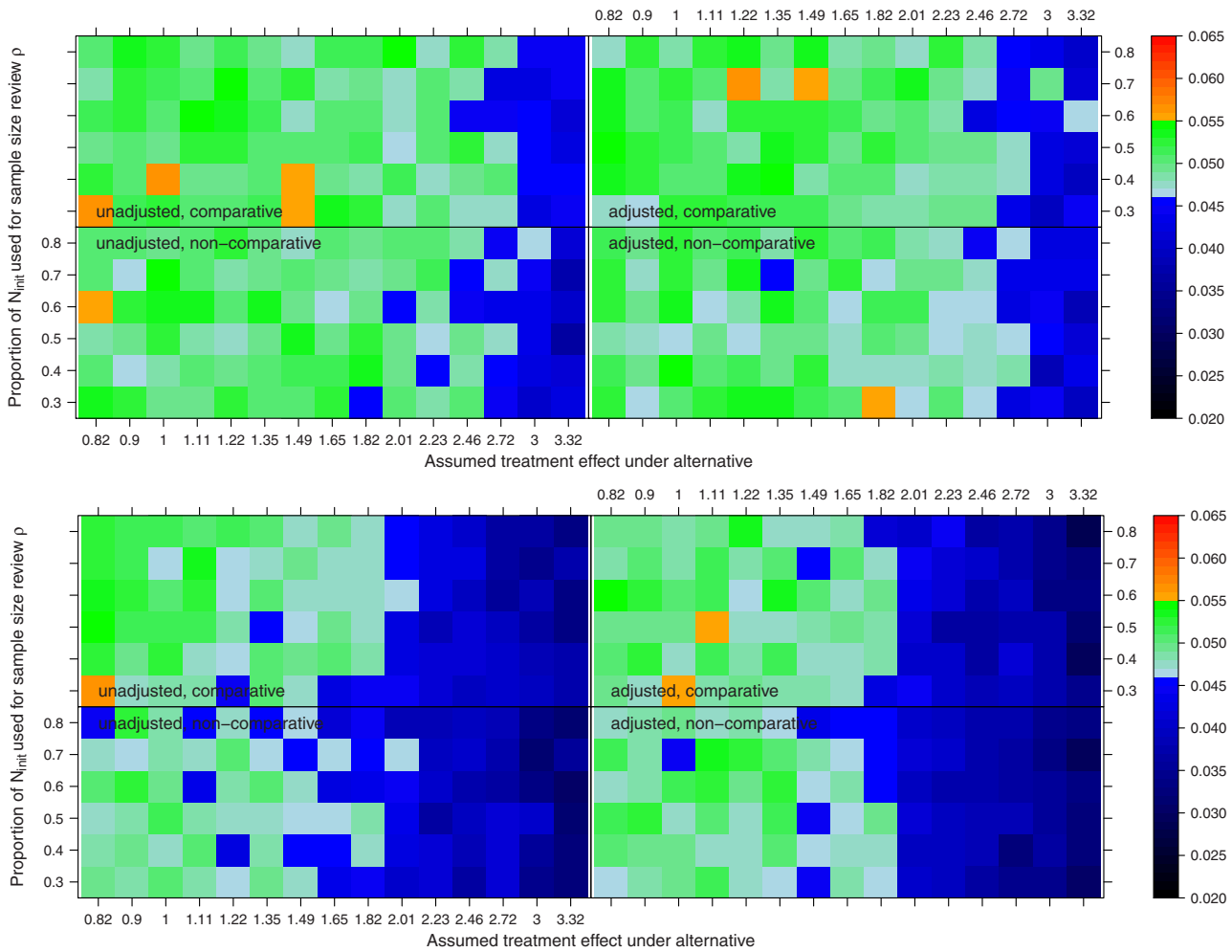


**FIGURE 3** Simulated type 1 error rate for the nonadjusted test statistic. Light-colored regions describe the 99% confidence band of the simulation error. Left (right)-panel figures show results for unadjusted (adjusted) estimators of $\tau^2$ at interim. Comparative (noncomparative) nuisance parameters for $\sigma^2$ and $\tau^2$ are used at interim in the top (bottom) panels, respectively

## 5.3 | Power

In this section, we explore, whether the sample size recalculation procedure achieves the pre-specified statistical power for varying parameter settings. The simulation results refer to the parameter settings described earlier. The desired statistical power is set to 0.8. Simulation results are presented in Figure 5, which shows the behavior of the unadjusted test statistic for varying values of initially (mis)-specified nuisance parameters, varying number of centers and treatment effects.

The power simulation results show in general that an initial misspecification of nuisance parameters will be corrected by the sample size recalculation. For a larger treatment effects, we can see, however, that the adjusted estimator for $\tau^2$ can lead to underpowered trials. The unadjusted estimator for $\tau^2$ will at least lead to the planned power level. No difference between noncomparative and comparative estimators can be observed in the simulation settings considered here.

The distribution of nuisance parameters estimated at interim and final sample sizes is shown in Figures 6 and 7. Here we only presents results for one scenario, where both nuisance parameters are correctly assumed at the initial planning stage of the trial.

For a small treatment effect of $\mu = 1$ we observe only little differences regarding estimated nuisance parameters and resulting sample sizes, as seen Figure 5. For a larger treatment effect ($\mu = 2$) we can see, however, that resulting sample sizes are substantially increased for unadjusted nuisance parameter estimated or $\tau^2$. This increase seems to be influenced by the number of centers and overall number of subjects. Noncomparative variances estimated are slightly larger than comparative estimators, but this difference does not seem to affect sample size recalculation for the parameters considered here.

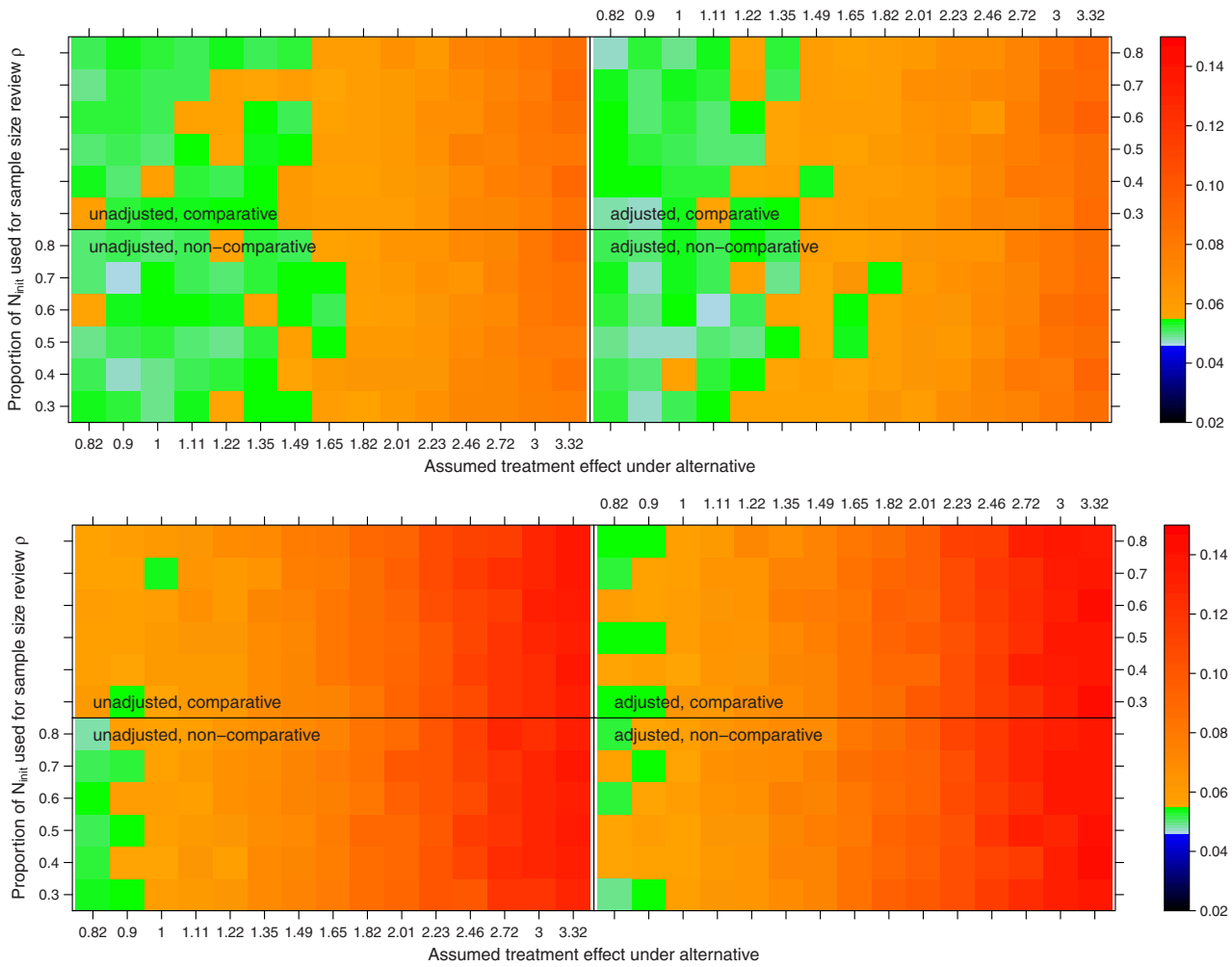


**FIGURE 4** Simulated type 1 error rate for the adjusted test statistic. Light-colored regions describe the 99% confidence band of the simulation error. Left (right)-panel figures show results for unadjusted (adjusted) estimators of $\tau^2$ at interim. Comparative (noncomparative) nuisance parameters for $\sigma^2$ and $\tau^2$ are used at interim in the top (bottom) panels, respectively

## 6 | DISCUSSION AND CONCLUSIONS

The specification of nuisance parameters is crucial for sample size calculation. Today, it is standard practice to use sample size recalculation procedures based on noncomparative data to correct inappropriate initial guesses of such parameters, generally without any practically relevant inflation of the type 1 error rate.

In this article, we presented a sample size recalculation procedure based on noncomparative data for a random-effects model of multicenter trials. The underlying sample size formula was described previously and depends, in addition to treatment effect, variance, and type 1 and 2 error rates, on the number of centers and heterogeneity between study sites (Harden & Friede, 2018). Whereas other approaches assume balanced data within centers, we relax this assumption here. However, we consider block randomization to approximate the imbalance.

Based in simulation results, we suggest to use the unadjusted nuisance parameter of $\tau^2$ for sample size recalculation as well as the final analysis to control the type 1 error rate and achieve the desired power. Power simulations confirm that a recalculation of the sample size based on noncomparative data is an adequate method to correct initial misspecification of nuisance parameters. This is especially helpful in multicenter trials, since the heterogeneity between centers is often unknown at the planning stage and since it is barely reported in publications of clinical trials. In terms of type 1 error rate and power, we cannot detect any differences between sample size recalculation based on noncomparative or comparative data. We think that this observation is mainly due to the rather large sample sizes typical for multicenter trials and therefore considered in the presented simulations.

An alternative variance estimator specifically for block-randomized trials has been suggested (Ganju & Xing, 2009; Xing & Ganju, 2005). This estimator is calculated for each randomization block, therefore blindness of the data can be preserved.

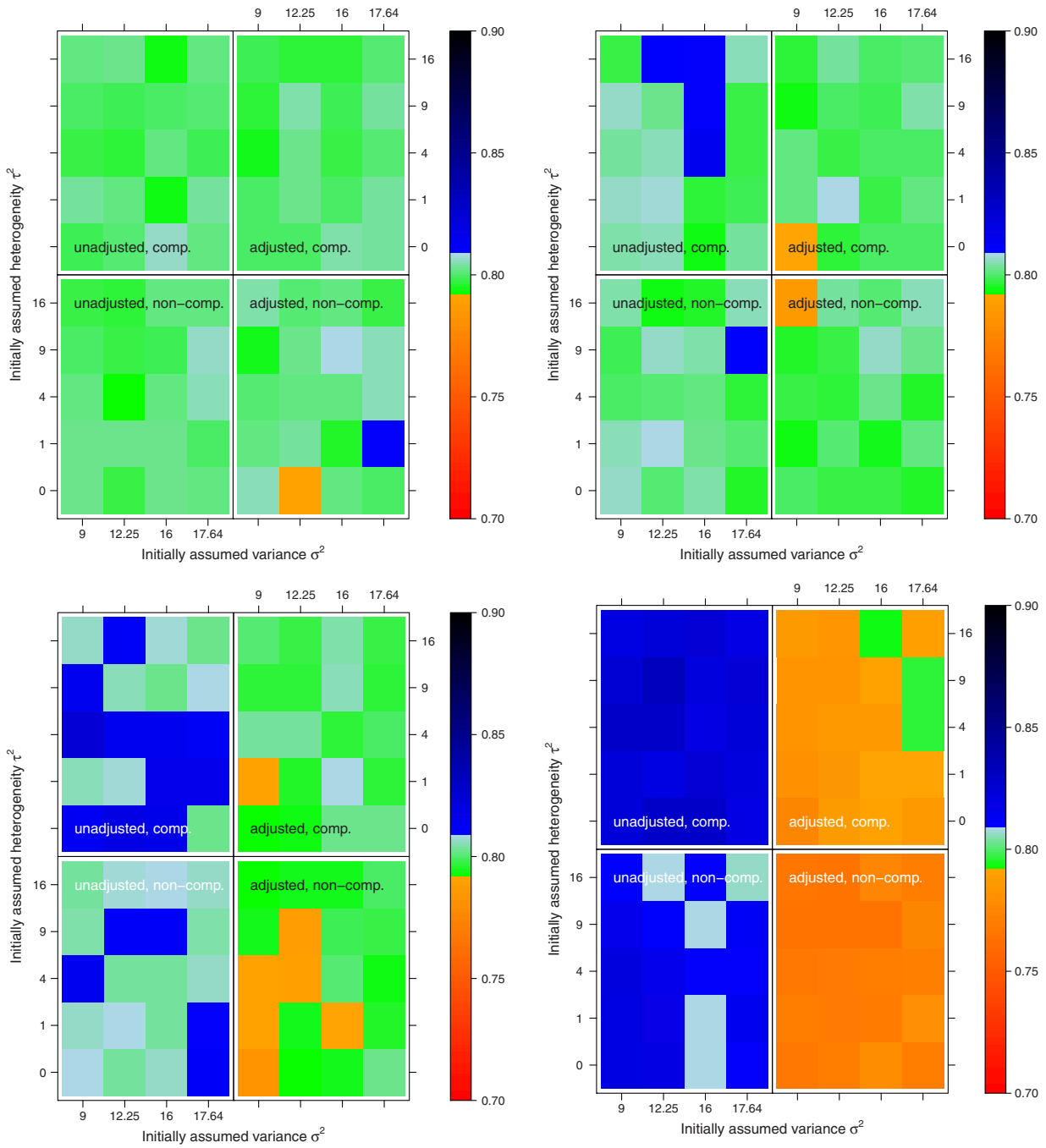


**FIGURE 5** Simulated power for the nonadjusted test statistic. Light-colored regions describe the 99% confidence band of the simulation error. Left (right)-panel figures show results for unadjusted (adjusted) estimators of $\tau^2$ at interim. Comparative (noncomparative) nuisance parameters for $\sigma^2$ and $\tau^2$ are used at interim in the top (bottom) panels, respectively. Comparative is abbreviated as comp

The estimator is unbiased, if all randomization blocks are complete. This estimation technique has also been extended to $\tau^2$ in cross-over as well as cluster randomized trials (Grayling, Mander, & Wason, 2018a, 2018b). Since this estimator relies on balanced data, we did not consider it here. Also, it has been demonstrated that the estimator's variance is larger compared to the one-sample variance estimator in situations that are common in clinical trials (Friede & Kieser, 2013).

In our motivating example, the COMPETE II trial, no sample size recalculation method was applied. However, we do believe that multicenter trials such as the COMPETE II trial can benefit from sample size recalculation methods. In practice, their implementation is not too different from previously proposed approaches. Data of multicenter trials are usually stored in a central trial database and all calculations can be based on a data export at interim. An ideal time point for the sample size

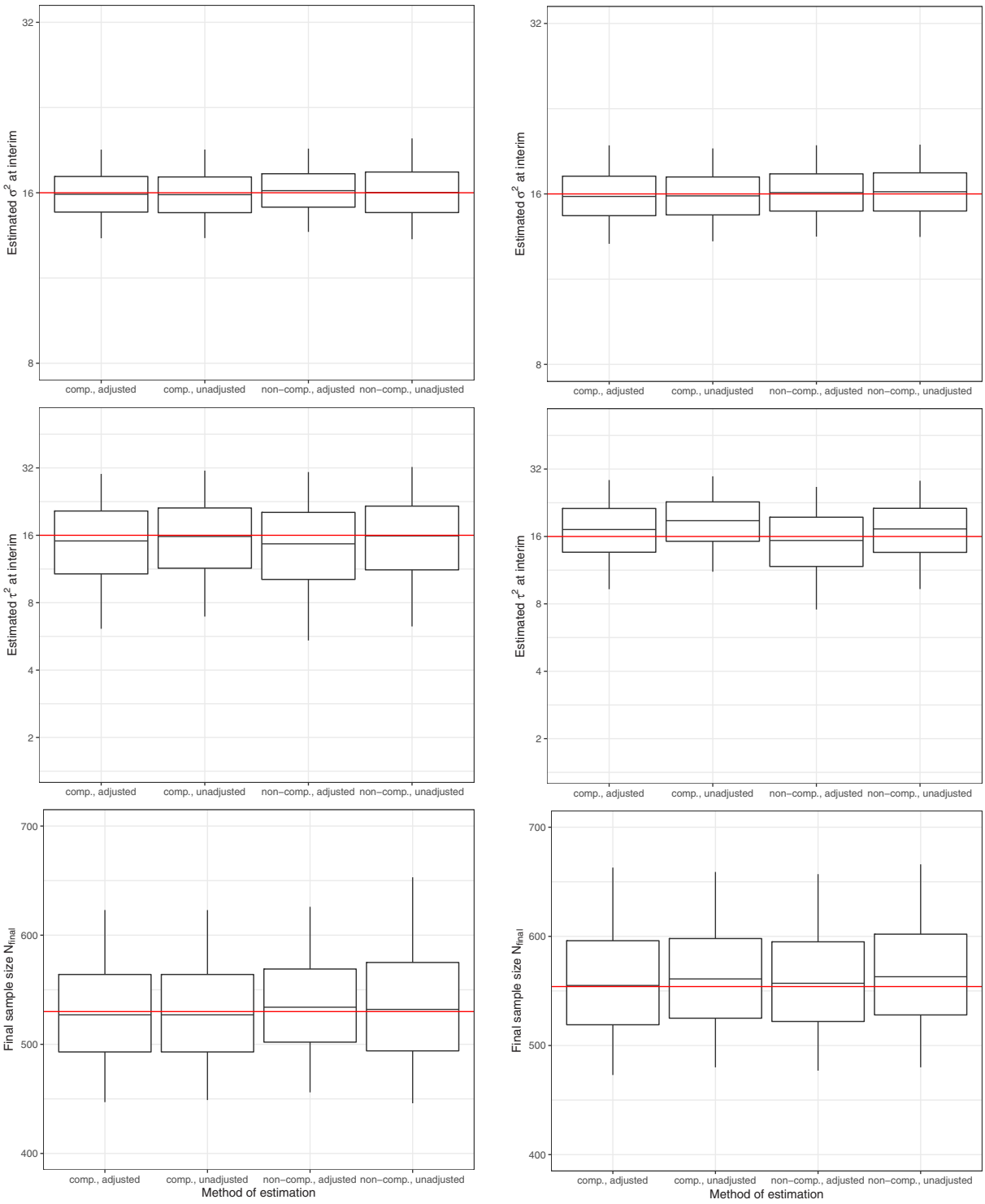


**FIGURE 6** Distribution of reestimated nuisance parameters and sample sizes based on $\mu = 1$ and varying center sizes. Results based on a block length of $b = 16$ and nuisance parameters $\sigma^2 = \tau^2 = 16$ and a sample size recalculation after half of the initial sample size is recruited. Red line represents true nuisance parameters and sample size of 530 (554) subjects for $c = 10$ ($c = 20$) centers based on true values. Comparative is abbreviated as comp

recalculation depends on several factors, such as pace of recruitment, trial duration, as well as uncertainty on initial choices of nuisance parameters (Chuang-Stein, Anderson, Gallo, & Collins, 2006). Given that not all patients enrolled at interim might have completed follow-up, approaches combining short- and long-term data have been proposed to increase precision of the estimated variance components, which might be transferred to multicenter trials, see, for example, Asendorf, Henderson, Schmidli, and

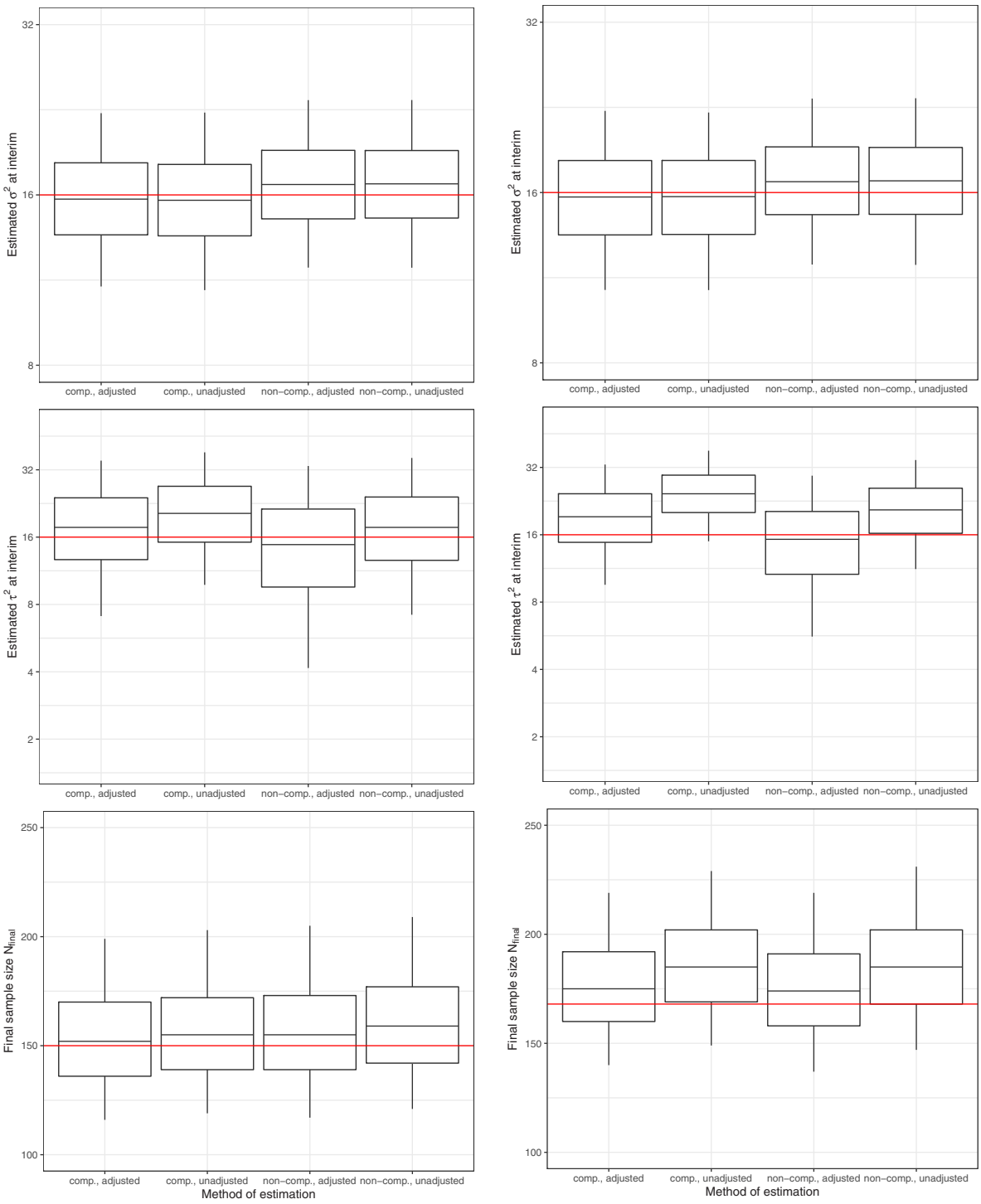


**FIGURE 7** Distribution of reestimated nuisance parameters and sample sizes based on $\mu = 2$ and varying center sizes. Results based on a block length of $b = 16$ and nuisance parameters $\sigma^2 = \tau^2 = 16$ and a sample size recalculation after half of the initial sample size is recruited. Red line represents true nuisance parameters and sample size of 150 (168) subjects for $c = 10$ ($c = 20$) centers based on true values. Comparative is abbreviated as comp

Friede (2019) for an application to longitudinal counts and Friede and Kieser (2006) for an overview. For further reading on practical guidance, we refer to the summary by Pritchett et al. (2015).

The integration of between-center heterogeneity into sample size considerations is a necessary step to control statistical power in multicenter trials. The use of a sample size recalculation procedure based on noncomparative data is a helpful tool to account for uncertainty in nuisance parameters at the planning stage of a trial.

## CONFLICT OF INTEREST

The authors have declared no conflict of interest.

## OPEN RESEARCH BADGES

This article has earned an Open Data badge for making publicly available the digitally-shareable data necessary to reproduce the reported results. The data is available in the Supporting Information section.

This article has earned an open data badge "**Reproducible Research**" for making publicly available the code necessary to reproduce the reported results. The results reported in this article could fully be reproduced.

## ORCID

*Markus Harden* https://orcid.org/0000-0003-2533-9396
*Tim Friede* https://orcid.org/0000-0001-5347-7441

## REFERENCES

Asendorf, T., Henderson, R., Schmidli, H., & Friede, T. (2019). Sample size re-estimation for clinical trials with longitudinal negative binomial counts including time trends. *Statistics in Medicine*, *38*(9), 1503–1528.

Chu, R., Thabane, L., Ma, J., Holbrook, A., Pullenayegum, E., & Devereaux, P. J. (2011). Comparing methods to estimate treatment effects on a continuous outcome in multicentre randomized controlled trials: A simulation study. *BMC Medical Research Methodology*, *11*(1), 21.

Chuang-Stein, C., Anderson, K., Gallo, P., & Collins, S. (2006). Sample size reestimation: A review and recommendations. *Drug Information Journal*, *40*(4), 475–484.

Coffey, C. S., & Muller, K. E. (1999). Exact test size and power of a Gaussian error linear model for an internal pilot study. *Statistics in Medicine*, *18*(10), 1199–1214.

Danielsen, A. K., Okholm, C., Pommergaard, H.-C., Burcharth, J., & Rosenberg, J. (2014). Number of published randomized controlled multi center trials testing pharmacological interventions or devices is increasing in both medical and surgical specialties. *PloS One*, *9*(7), e101383.

Denne, J. S., & Jennison, C. (1999). Estimating the sample size for at-test using an internal pilot. *Statistics in Medicine*, *18*(13), 1575–1585.

EMA (2007). *Reflection paper on methodological issues in confirmatory clinical trials planned with an adaptive design*. London: EMEA.

FDA (2016). *Adaptive designs for medical device clinical studies*. Washington DC: Food and Drug Administration.

FDA (2018). *Guidance for industry: Adaptive design clinical trials for drugs and biologics*. Washington DC: Food and Drug Administration.

Friede, T., & Kieser, M. (2001). A comparison of methods for adaptive sample size adjustment. *Statistics in Medicine*, *20*(24), 3861–3873.

Friede, T., & Kieser, M. (2006). Sample size recalculation in internal pilot study designs: A review. *Biometrical Journal*, *48*(4), 537–555.

Friede, T., & Kieser, M. (2013). Blinded sample size re-estimation in superiority and noninferiority trials: Bias versus variance in variance estimation. *Pharmaceutical Statistics*, *12*(3), 141–146.

Ganju, J., & Xing, B. (2009). Re-estimating the sample size of an on-going blinded trial based on the method of randomization block sums. *Statistics in Medicine*, *28*(1), 24–38.

Gould, A. L. (1992). Interim analyses for monitoring clinical trials that do not materially affect the type I error rate. *Statistics in Medicine*, *11*(1), 55–66.

Grayling, M. J., Mander, A. P., & Wason, J. M. (2018a). Blinded and unblinded sample size reestimation in crossover trials balanced for period. *Biometrical Journal*, *60*(5), 917–933.

Grayling, M. J., Mander, A. P., & Wason, J. M. (2018b). Blinded and unblinded sample size reestimation procedures for stepped-wedge cluster randomized trials. *Biometrical Journal*, *60*(5), 903–916.

Harden, M., & Friede, T. (2018). Sample size calculation in multi-centre clinical trials. *BMC Medical Research Methodology*, *18*(1), 156.

Holbrook, A., Thabane, L., Keshavjee, K., Dolovich, L., Bernstein, B., Chan, D.... COMPETE II Investigators (2009). Individualized electronic decision support and reminders to improve diabetes care in the community: COMPETE II randomized trial. *Canadian Medical Association Journal*, *181*(1-2), 37–44.

ICH (1998). *ICH Topic E9 Statistical principles for clinical trials* (Report No. CPMP/ICH/363/96). London: European Medicines Agency.

Jensen, K., & Kieser, M. (2010). Blinded sample size recalculation in multicentre trials with normally distributed outcome. *Biometrical Journal*, *52*(3), 377–399.

Kahan, B. C., & Morris, T. P. (2013). Analysis of multicentre trials with continuous outcomes: When and how should we account for centre effects? *Statistics in Medicine*, *32*(7), 1136–1149.

Kieser, M., & Friede, T. (2003). Simple procedures for blinded sample size adjustment that do not affect the type I error rate. *Statistics in Medicine*, *22*(23), 3571–3581.

Miller, F. (2005). Variance estimation in clinical studies with interim sample size reestimation. *Biometrics*, *61*(2), 355–361.

Pritchett, Y. L., Menon, S., Marchenko, O., Antonijevic, Z., Miller, E., Sanchez-Kam, M.... Prucka, W. R. (2015). Sample size re-estimation designs in confirmatory clinical trials–current state, statistical considerations, and practical guidance. *Statistics in Biopharmaceutical Research*, *7*(4), 309–321.

Proschan, M. A. (2009). Sample size re-estimation in clinical trials. *Biometrical Journal*, *51*(2), 348–357.
R Development Core Team (2008). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.
Ruvuna, F. (2004). Unequal center sizes, sample size, and power in multicenter clinical trials. *Drug Information Journal*, *38*(4), 387–394.
Shih, W. J., & Long, J. (1998). Blinded sample size re-estimation with unequal variances and center effects in clinical trials. *Communications in Statistics*, *27*(2), 395–408.
Snow, G. (2013). *blockrand: Randomization for block random clinical trials*. R package version 1.3. Vienna, Austria: R Foundation for Statistical Computing.
Vierron, E., & Giraudeau, B. (2009). Design effect in multicenter studies: gain or loss of power? *BMC Medical Research Methodology*, *9*(1), 39.
Wassmer, G., & Brannath, W. (2016). *Group sequential and confirmatory adaptive designs in clinical trials*. Berlin, Germany: Springer.
Wittes, J., & Brittain, E. (1990). The role of internal pilot studies in increasing the efficiency of clinical trials. *Statistics in Medicine*, *9*(1-2), 65–72.
Xing, B., & Ganju, J. (2005). A method to estimate the variance of an endpoint from an on-going blinded trial. *Statistics in Medicine*, *24*(12), 1807–1814.
Zucker, D. M., Wittes, J. T., Schabenberger, O., & Brittain, E. (1999). Internal pilot studies II: Comparison of various procedures. *Statistics in Medicine*, *18*(24), 3493–3509.

## SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

**How to cite this article:** Harden M, Friede T. Sample size recalculation in multicenter randomized controlled clinical trials based on noncomparative data. *Biometrical Journal*. 2020;1–16. https://doi.org/10.1002/bimj.201900138

## APPENDIX: EXPECTED VALUES OF THE NUISANCE PARAMETER ESTIMATORS

Given the model in Section 3 and formulating every estimator as a quadratic form, the expectations of variance components given in (1), (2), (5), and (8) can be calculated as shown below.

Let $\mathbf{Y}_{n_{ij}} = (Y_{ij1}, \dots, Y_{ijn_{ij}})'$ denote the vector of all observations in treatment group $i$ and center $j$ and $\mathbf{P}_{n_{ij}} = \mathbf{I}_{n_{ij}} - n_{ij}^{-1}\mathbf{J}_{n_{ij}}$ the $n_{ij}$-dimensional centering matrix with the $n_{ij}$-dimensional identity matrix $\mathbf{I}_{n_{ij}}$ and the $n_{ij} \times n_{ij}$-dimensional matrix $\mathbf{J}_{n_{ij}}$ consisting of ones only.

$$\begin{aligned}
\mathrm{E}\left(\widehat{\sigma}^2\right) &= \frac{1}{2c}\sum_{i=1}^{2}\sum_{j=1}^{c}\frac{1}{n_{ij}-1}\mathrm{E}\left(\sum_{k=1}^{n_{ij}}\left(Y_{ijk}-\overline{Y}_{ij\cdot}\right)^2\right)\\
&= \frac{1}{2c}\sum_{i=1}^{2}\sum_{j=1}^{c}\frac{1}{n_{ij}-1}\mathrm{E}\left(\mathbf{Y}_{n_{ij}}'\mathbf{P}_{n_{ij}}\mathbf{Y}_{n_{ij}}\right)\\
&= \frac{1}{2c}\sum_{i=1}^{2}\sum_{j=1}^{c}\frac{1}{n_{ij}-1}\operatorname{trace}\left[\mathbf{P}_{n_{ij}}\cdot\left(\sigma^2\cdot\mathbf{I}_{n_{ij}}+\sigma_c^2\cdot\mathbf{J}_{n_{ij}}\right)\right]\\
&= \sigma^2
\end{aligned}$$

A similar calculation can be done for $\widehat{\tau}^2$ when denoting $\overline{\mathbf{Y}}_i = (\overline{Y_{i1\cdot}}, \dots, \overline{Y_{i1\cdot}})'$ the vector of group means for treatment group $i$ at centers $j = 1, \dots, c$ and $\mathbf{P}_c$ die $c$-dimensional centering matrix.

$$\begin{aligned}
\mathrm{E}\left(\widehat{\tau}^2\right) &= \frac{1}{2(c-1)}\sum_{i=1}^{2}\mathrm{E}\left(\sum_{j=1}^{c}\left(\overline{Y}_{ij\cdot}-\overline{Y}_{i\cdot\cdot}\right)^2\right)\\
&= \frac{1}{2(c-1)}\sum_{i=1}^{2}\mathrm{E}\left(\overline{\mathbf{Y}}_i'\mathbf{P}_c\overline{\mathbf{Y}}_i\right)
\end{aligned}$$

$$= \frac{1}{2(c-1)} \sum_{i=1}^{2} \text{trace} \left[ \mathbf{P}_c \cdot \text{Diag} \left( \frac{\sigma^2}{n_{ij}} + \tau^2 \right)_{\{j=1,\ldots,c\}} \right]$$

$$= \frac{1}{2(c-1)} \sum_{i=1}^{2} \sum_{j=1}^{c} \frac{c-1}{c} \left( \frac{\sigma^2}{n_{ij}} + \tau^2 \right)$$

$$= \tau^2 + \frac{\sigma^2}{2c} \sum_{i=1}^{2} \sum_{j=1}^{c} \frac{1}{n_{ij}}$$

For the noncomparative variance estimator denote $\mathbf{Y}_{n_j} = (\mathbf{Y}'_{n_{1j}}, \mathbf{Y}'_{n_{2j}})'$ the vector of observations at center $j$, $\boldsymbol{\mu_j} = \mathrm{E}(\mathbf{Y}_{n_j}) = (\mu_0 \mathbf{1}'_{n_{1j}}, (\mu_0 + \mu)\mathbf{1}'_{n_{2j}})'$ the vector of expectations and $\mathbf{P}_{n_j}$ the $n_j$-dimensional centering matrix.

$$\mathrm{E}\left(\hat{\sigma}_b^2\right) = \frac{1}{c} \sum_{j=1}^{c} \frac{1}{n_j - 1} \mathrm{E}\left( \sum_{i=1}^{2} \sum_{k=1}^{n_{ij}} \left( Y_{ijk} - \overline{Y}_{\cdot j \cdot} \right)^2 \right)$$

$$= \frac{1}{c} \sum_{j=1}^{c} \frac{1}{n_j - 1} \mathrm{E}\left( \mathbf{Y}'_{n_j} \mathbf{P}_{n_j} \mathbf{Y}_{n_j} \right)$$

$$= \frac{1}{c} \sum_{j=1}^{c} \frac{1}{n_j - 1} \left( \text{trace} \left[ \mathbf{P}_{n_j} \cdot \left( \sigma^2 \cdot \mathbf{I}_{n_j} + \sigma_c^2 \cdot \mathbf{J}_{n_j} \right) \right] + \boldsymbol{\mu}'_j \mathbf{P}_{n_j} \boldsymbol{\mu_j} \right)$$

$$= \sigma^2 + \frac{1}{c} \sum_{j=1}^{c} \frac{1}{n_j - 1} \cdot \boldsymbol{\mu}'_j \mathbf{P}_{n_j} \boldsymbol{\mu_j}$$

$$= \sigma^2 + \frac{\mu^2}{N - c} \sum_{j=1}^{c} \frac{n_{1j} n_{2j}}{n_j}$$

And with a similar calculation

$$\mathrm{E}\left(\hat{\tau}_b^2\right) = \frac{1}{c-1} \mathrm{E}\left( \sum_{j=1}^{c} \left( \overline{Y}_{\cdot j \cdot} - \overline{Y}_{\cdots} \right)^2 \right)$$

$$= \frac{1}{c-1} \mathrm{E}\left( \overline{\mathbf{Y}}_{\cdot}' \mathbf{P}_c \overline{\mathbf{Y}}_{\cdot}' \right)$$

$$= \frac{1}{c-1} \left( \text{trace} \left[ \mathbf{P}_c \cdot \text{Diag} \left( \frac{\sigma^2}{n_j} + \sigma_c^2 \right)_{\{j=1,\ldots,c\}} \right] + \boldsymbol{\mu}'_c \mathbf{P}_c \boldsymbol{\mu}_c \right)$$

$$= \frac{1}{c-1} \left( \sum_{j=1}^{c} \frac{c-1}{c} \left( \frac{\sigma^2}{n_j} + \tau^2 \right) + \boldsymbol{\mu}'_c \mathbf{P}_c \boldsymbol{\mu}_c \right)$$

$$= \tau^2 + \frac{\sigma^2}{c} \sum_{j=1}^{c} \frac{1}{n_j} + \frac{1}{c-1} \cdot \boldsymbol{\mu}'_c \mathbf{P}_c \boldsymbol{\mu}_c$$

$$= \tau^2 + \frac{\sigma^2}{c} \sum_{j=1}^{c} \frac{1}{n_j} + \frac{\mu^2}{c-1} \sum_{j=1}^{c} \left( \frac{n_{2j}}{n_j} - \frac{1}{c} \sum_{\ell=1}^{c} \frac{n_{2\ell}}{n_\ell} \right)^2$$