Historical control arms in oncology clinical trials

Xiaomeng Wang

THESE DE DOCTORAT

NNT : 2023UPASR017

Historical control arms in oncology clinical trials

Contrôles historiques dans les essais cliniques en oncologie

Thèse de doctorat de l'université Paris-Saclay

École doctorale n° 570 : Santé publique (EDSP)
Spécialité de doctorat : Biostatistiques et data sciences
Graduate School : Santé publique
Référent : Université de Versailles-Saint-Quentin-en-Yvelines

Thèse préparée dans l'unité de recherche **U900-Cancer et Génôme : Bioinformatique, Biostatistiques et Epidémiologie des systèmes complexes** (Institut Curie, Université PSL, Inserm),
sous la direction d'**Aurélien LATOUCHE**, PU,
la co-direction de **Roman ROUZIER**, PU-PH,
et la co-supervision de **Ramon HERNANDEZ**,
Global Head of Development Real World Evidence, Sanofi

Thèse soutenue à Paris, le 09 octobre 2023, par

Xiaomeng WANG

Composition du Jury

Membres du jury avec voix délibérative

Jérôme LAMEBRT PU, Université Paris Cité	Président
Carine BELLERA CR, HDR, Institut Bergonié, Université Bordeaux	Rapporteur & Examinatrice
Thomas FILLERON IR, HDR, IUCT Oncopole, Université Toulouse 1	Rapporteur & Examinateur
Emmanuel CHAZARD PU-PH, Université de Lille	Examinateur

université PARIS-SACLAY | ÉCOLE DOCTORALE Santé Publique (EDSP)

Titre : Contrôles historiques dans les essais cliniques en oncologie

Mots clés : contrôle historique, essai clinique, inférence causale, oncologie, données de vie réel

Résumé : L'essai contrôlé randomisé représente la méthode de référence pour établir l'effet causal des traitements expérimentaux par rapport aux traitements de contrôle ou aux placebos. Néanmoins, des problèmes éthiques ou de faisabilité peuvent entraver le processus de randomisation, en particulier dans le développement de médicaments en oncologie. Les autorités réglementaires reconnaissent ces défis et ont accordé des approbations conditionnelles pour des essais à bras unique, avec l'exigence de preuves confirmatoires ultérieures à partir d'études post-approbation.

Dans ce contexte, les bras de contrôle historique ont émergé comme une approche complémentaire aux essais cliniques. En fournissant des informations contextuelles et en améliorant l'interprétation des résultats des essais à bras unique, les bras de contrôle historique visent à réduire le biais dû au manque de randomisation. Bien que diverses méthodes statistiques dans le cadre de l'inférence causale aient été proposées, il existe actuellement un manque de directives régissant leur application dans le processus de développement de médicaments. L'objectif de cette thèse est d'évaluer la faisabilité des bras de contrôle historique, en mettant l'accent sur la disponibilité des données historiques et les méthodes d'analyse statistique appropriées.

Nous avons d'abord mené une revue systématique de l'application des bras de contrôle historique dans le développement de médicaments en oncologie en Europe. Nos résultats indiquent que les contrôles historiques ont été activement soumis aux autorités réglementaires ; cependant, ils ne sont pas systématiquement considérés comme des preuves favorables. Nous avons identifié des limitations significatives et formulé des suggestions correspondantes concernant la conception de l'étude, la sélection des données et l'application des méthodes statistiques.

S'appuyant sur les enseignements tirés de la revue, nous avons mené deux études de cas. La première étude de cas examine un essai à bras unique observationnel qui a évalué l'efficacité du bloc du plan du muscle érecteur du rachis dans la réduction de la douleur post-opératoire lors de la chirurgie du cancer du sein. Nous avons construit un bras de contrôle historique en utilisant des données issues d'essais cliniques précédents et appliqué une analyse de score de propension pour réduire le biais de confusion. La deuxième étude de cas se concentre sur un essai randomisé contrôlé examinant Olaparib plus Bevacizumab en tant que traitement d'entretien de première ligne dans le cancer de l'ovaire. Ici, nous visons à émuler le bras de contrôle en exploitant les données observationnelles de la base de données du monde réel ESME, en utilisant le cadre de l'essai cible.

Cette thèse contribue à faire progresser notre compréhension de la faisabilité et de l'applicabilité des bras de contrôle historique, en mettant en lumière leurs avantages potentiels et leurs applications appropriées dans le domaine du développement de médicaments en oncologie.

université PARIS-SACLAY | ÉCOLE DOCTORALE Santé Publique (EDSP)

Title : Historical control arms in oncology clinical trials

Keywords : historical control arm, clinical trial, causal inference, oncology, real-world data

Abstract : Randomized controlled trials are the gold standard for establishing the causal effect of experimental treatments compared to reference treatments or placebos. Nevertheless, ethical or feasibility issues can hinder the randomization process, especially in oncology drug development. Regulatory bodies acknowledge these challenges and have granted conditional approvals for single-arm trials, with the requirement of further confirmatory evidence from post-approval studies.

Within this context, historical control arms have emerged as an approach to complement clinical trials. By providing contextual information and improving the interpretation of results from single-arm trials, historical control arms aim to reduce the bias due to the lack of randomization. While various statistical methods in the framework of causal inference have been proposed, there is currently a lack of guidelines governing their application in the drug development process. The objective of this thesis is to evaluate the feasibility of historical control arms, focusing on historical data availability and appropriate statistical analysis methods.

We first conducted a systematic review of the application of historical control arms in oncology drug development in Europe. Our findings indicate that the historical controls have been actively submitted to regulatory bodies; however, they are not consistently deemed as supportive evidence. We identified significant limitations and make corresponding suggestions regarding study design, data selection and the application of statistical methods.

Building upon the insights gained from the review, we conducted two case studies. The first case study investigates an observational single-arm trial that evaluated the effectiveness of erector spinae plane block in reducing post-operative pain in breast cancer surgery. We constructed a historical control arm using data from previous clinical trials and applied propensity score analysis to reduce the confounding bias. The second case study focuses on a randomized controlled trial examining Olaparib plus Bevacizumab as first-line maintenance therapy in ovarian cancer. Here, we aim to emulate the control arm by leveraging observational data from the real-world database ESME, employing the target trial framework.

This thesis contributes to advancing our understanding of the feasibility and applicability of historical control arms, shedding light on their potential benefits and appropriate applications in the field of oncology drug development.

There is only one heroism in the world:
to see the world as it is, and to love it.

Il n'ya qu'un héroïsme au monde :
c'est de voir le monde tel qu'il est et de l'aimer.

— Romain Rolland

Acknowledgements

As I step back, I recall the onset of my PhD journey, closely coinciding with the commencement of the Covid-19 pandemic. Now, having crossed the finish line and as life gradually regains its familiar rhythm, I am filled with both relief and gratitude. This journey has been a challenging yet enriching chapter in my life, and I would like to express my deepest gratitude to those who have offered wisdom, support, and companionship.

First and foremost, I am profoundly grateful to my supervisors, Prof. Aurélien Latouche and Prof. Roman Rouzier. Your invaluable advice, consistent support, and infinite patience have shaped my PhD study. I feel privileged to have been mentored by such distinguished individuals, and I will always cherish the knowledge and wisdom you shared.

A special note of gratitude is due to my advisor at Sanofi, Dr. Ramon Hernandez, for leading the CIFRE collaboration between Institut Curie and Sanofi. Your guidance and the opportunity to work with the Development Real World Evidence team has enriched my journey with practical insights. I also wish to thank ANRT for funding this CIFRE program.

My sincere appreciation goes to the jury members, Dr. Carine Bellera, Dr. Thomas Filleron, Prof. Emmanuel Chazard, and Prof. Jérôme Lambert, for kindly reviewing my work. Your time, expertise, and valuable feedback are deeply appreciated.

I had the privilege of collaborating with outstanding researchers and clinicians, Dr. Aline Albi-Feldzer, Dr. Antoine Premachandra, Dr. Mary Saad, and Dr. Nina Oufkir. Your profound medical knowledge and your dedication to helping patients have shown me the purpose and significance of my work.

My gratitude extends to all members of the StaMPM team for your daily support and the enriching exchange of knowledge and experiences. I am thankful to the Inserm U900 team for welcoming me as a doctoral student and to our amiable

assistant, Caroline Dahan Belliere, for her invaluable support. My thanks also go to the Doctoral School of Public Health (EDSP) for providing academic training and guidance along the way.

To my beloved family and friends, your enduring support has been my anchor in tumultuous times. Your love and encouragement have given me the strength to persevere.

Lastly, a special mention goes to Sushi, my one-year-old golden retriever. As my furry companion, your presence has been a source of solace and joy. In your simple routine of eating, playing, sleeping, and loving, I see a life philosophy to aspire to.

With all my heart, I extend my deepest gratitude to each one of you for being an integral part of this special journey. Thank you for making it both memorable and meaningful.

Abstract

Randomized controlled trials are the gold standard for establishing the causal effect of experimental treatments compared to reference treatments or placebos. Nevertheless, ethical or feasibility issues can hinder the randomization process, especially in oncology drug development. Regulatory bodies acknowledge these challenges and have granted conditional approvals for single-arm trials, with the requirement of further confirmatory evidence from post-approval studies.

Within this context, historical control arms have emerged as an approach to complement clinical trials. By providing contextual information and improving the interpretation of results from single-arm trials, historical control arms aim to reduce the bias due to the lack of randomization. While various statistical methods in the framework of causal inference have been proposed, there is currently a lack of guidelines governing their application in the drug development process. The objective of this thesis is to evaluate the feasibility of historical control arms, focusing on historical data availability and appropriate statistical analysis methods.

We first conducted a systematic review of the application of historical control arms in oncology drug development in Europe. Our findings indicate that the historical controls have been actively submitted to regulatory bodies; however, they are not consistently deemed as supportive evidence. We identified significant limitations and make corresponding suggestions regarding study design, data selection and the application of statistical methods.

Building upon the insights gained from the review, we conducted two case studies. The first case study investigates an observational single-arm trial that evaluated the effectiveness of erector spinae plane block in reducing post-operative pain in breast cancer surgery. We constructed a historical control arm using data from previous clinical trials and applied propensity score analysis to reduce the confounding bias. The second case study focuses on a randomized controlled trial examining Olaparib plus Bevacizumab as first-line maintenance therapy in ovarian cancer. Here, we aim to

emulate the control arm by leveraging observational data from the real-world database ESME, employing the target trial framework.

This thesis contributes to advancing our understanding of the feasibility and applicability of historical control arms, shedding light on their potential benefits and appropriate applications in the field of oncology drug development.

Key words: historical control arm, clinical trial, causal inference, oncology, real-world data

Scientific publications

- Wang, X., Dormont, F., Lorenzato, C., Latouche, A., Hernandez, R., & Rouzier, R. (2023). Current Perspectives for External Control Arms in Oncology Clinical Trials: Analysis of EMA approvals 2016-2021. *Journal of Cancer Policy*, 100403.
- Premachandra, A.*, Wang, X.*, Saad, M., Moussawy, S., Rouzier, R., Latouche, A., & Albi-Feldzer, A. (2022). Erector spinae plane block versus thoracic paravertebral block for the prevention of acute postsurgical pain in breast cancer surgery: A prospective observational study compared with a propensity score-matched historical cohort. *Plos one*, 17(12), e0279648.

Table of Contents

List of Tables

List of Figures

List of Main Abbreviations

ATE	Average Treatment Effect
ATT	Average Treatment Effect for the Treated
EF	Estimand Framework
EHRs	Electronic Health Records
EMA	European Medical Agency
ESPB	Erector Spinae Plane Block
FDA	Food and Drug Administration
IDS	Interval Debulking Surgery
IPTW	Inverse Probability of Treatment Weighting
OS	Overall Survival
PDS	Primary Debulking Surgery
PFS	Progression Free Survival
POF	Potential Outcome Framework
PSP	Acute Post-surgical Pain
RCT	Randomized Controlled Trial
RWD	Real-world Data
RWE	Real-world Evidence
SATs	Single-arm Trials
SMD	Standardized Mean Difference
SUTVA	Stable Unit Treatment Value Assumption
TPVB	Thoracic Paravertebral Block
TTF	Target Trial Emulation Framework

Chapter 1

Introduction

Contents

Clinical research lays the groundwork for progress in medicine and serves as the foundation for evidence-based practice. In the evidence hierarchy (Figure 1.1), the top section consists of filtered (secondary) evidence, including systematic reviews, meta-analyses, and critical appraisals. The section below includes unfiltered (primary) evidence, including randomized controlled trials (RCTs), cohort studies, case-controlled studies, case series, and case reports (Burns et al., 2011; Murad et al., 2016). Alongside meta-analyses, high-quality RCTs with a low risk of systematic error provide the highest level of evidence. Randomized controlled trials are considered the gold standard for effectiveness research by demonstrating the superiority of a new treatment over an existing standard treatment or a placebo (Hariton & Locascio, 2018). In clinical research RCTs are used to answer patient-related questions, and in the development of new drugs they form the basis for regulatory authorities' decisions on approval (Kabisch et al., 2011).

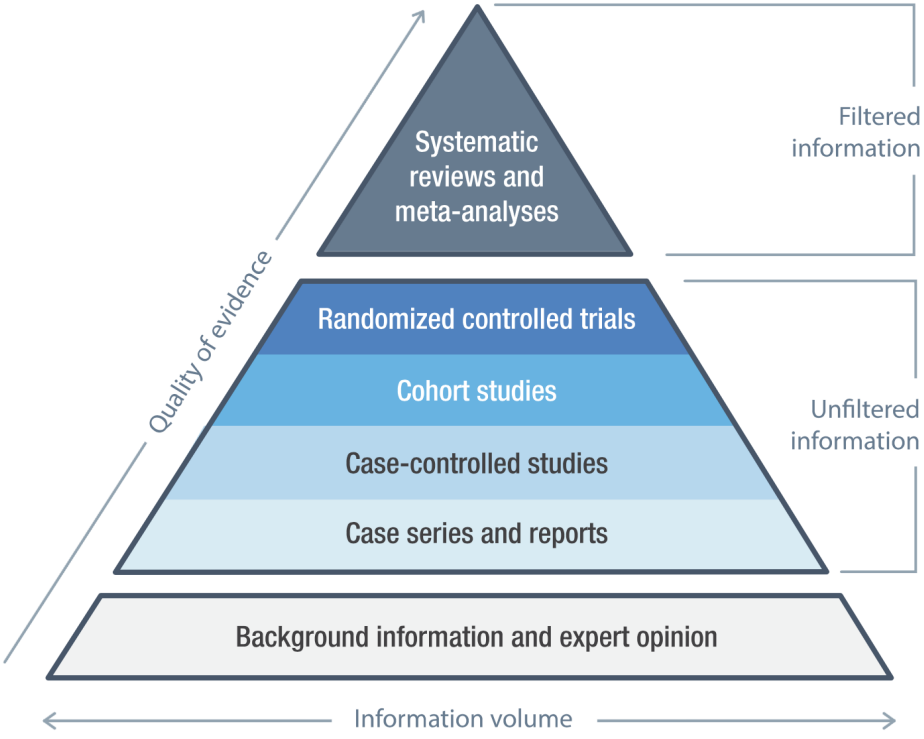


Figure 1.1: Hierarchy of evidence in evidence-based medicine

This chapter serves as an introduction to randomized controlled trials, focusing on the rationale behind using historical control arms when randomization is not feasible. Besides, we provide the definition of historical control arms in various aspects and discuss the statistical issues associated with historical control arms.

- **Section 1.1** provides an overview of randomized controlled trials and their conceptual framework, highlighting their advantages and limitations.
- **Section 1.2** introduces the definition of historical control arms and discusses various aspects of their design.
- **Section 1.3** succinctly presents the context of the thesis and outlines the structure of the manuscript.

1.1 Gold standard: randomized controlled trials

1.1.1 General introduction

A randomized controlled trial is a prospective study that measures the efficacy of an intervention or treatment (Figure 1.2). Subjects are randomly assigned to either an experimental group or a control group. The control group receives a placebo or sham intervention, while the experimental group receives the intervention being studied.

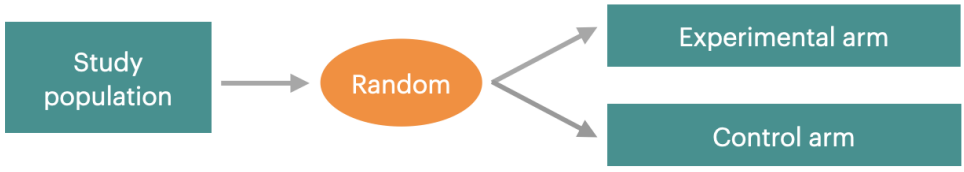


Figure 1.2: Randomized controlled trial

The first published RCT in medicine appeared in the 1948 paper entitled "Streptomycin treatment of pulmonary tuberculosis", which described a Medical Research Council investigation (Geoffrey, Marshall, 1948). By the late 20th century, RCTs had been recognized as the standard method for "rational therapeutics" in medicine (Meldrum, 2000).

In RCTs, the randomization process serves to reduce bias and provides a robust framework for examining cause-effect relationships between interventions and outcomes. By balancing participant characteristics between groups, both observed and unobserved, randomization enables the attribution of any outcome differences to the study intervention. This distinguishing feature is not feasible with other study designs.

The advantages of proper randomization in RCTs include:

- Minimization of selection bias: This type of bias can occur if investigators consciously or unconsciously enroll patients preferentially in certain treatment arms. An effective randomization procedure should be unpredictable, preventing investigators from guessing the group assignment for the next subject based on prior treatment assignments. The risk of selection bias is highest when previous treatment assignments are known (as in unblinded studies) or can be inferred (e.g., if a drug has distinct side effects).
- Minimization of allocation bias (or confounding): Allocation bias arises when covariates that influence the outcome are not evenly distributed between treatment groups, confounding the treatment effect with the effect of the covariates. If the randomization procedure leads to an imbalance in outcome-related covariates across groups, estimates of effect may be biased if not adjusted for these covariates, especially when they are unmeasured and impossible to account for.

When designing an RCT, careful consideration should be given to selecting the target population, interventions to be compared, and the desired outcomes. Power calculations are performed to determine the number of participants needed to reliably detect the existence of a relationship. Subsequently, participants are recruited and randomly assigned to either the intervention or the comparator group.

It is essential to ensure that at the time of recruitment, there is no knowledge of the participant's allocation to a specific group. This is achieved through concealment, often facilitated by automated randomization systems, such as computer-generated methods. Blinding of participants, doctors, nurses, or researchers is also common practice in RCTs to further minimize bias by preventing knowledge of the treatment allocation.

RCTs can be analyzed using various approaches, such as intention-to-treat analysis, where subjects are analyzed in the groups to which they were randomized, or per-protocol analysis, which includes only participants who completed the originally allocated treatment. Intention-to-treat analysis is often regarded as the least biased approach. All RCTs should have pre-specified primary outcomes, be registered in clinical trials databases, and obtain appropriate ethical approvals.

The CONSORT (Consolidated Standards of Reporting Trials) 2010 Statement provides guidelines for reporting parallel two-group RCTs, encompassing four phases:

enrollment, allocation, intervention, follow-up, and data analysis as shown in Figure 1.3 (Schulz et al., 2010).

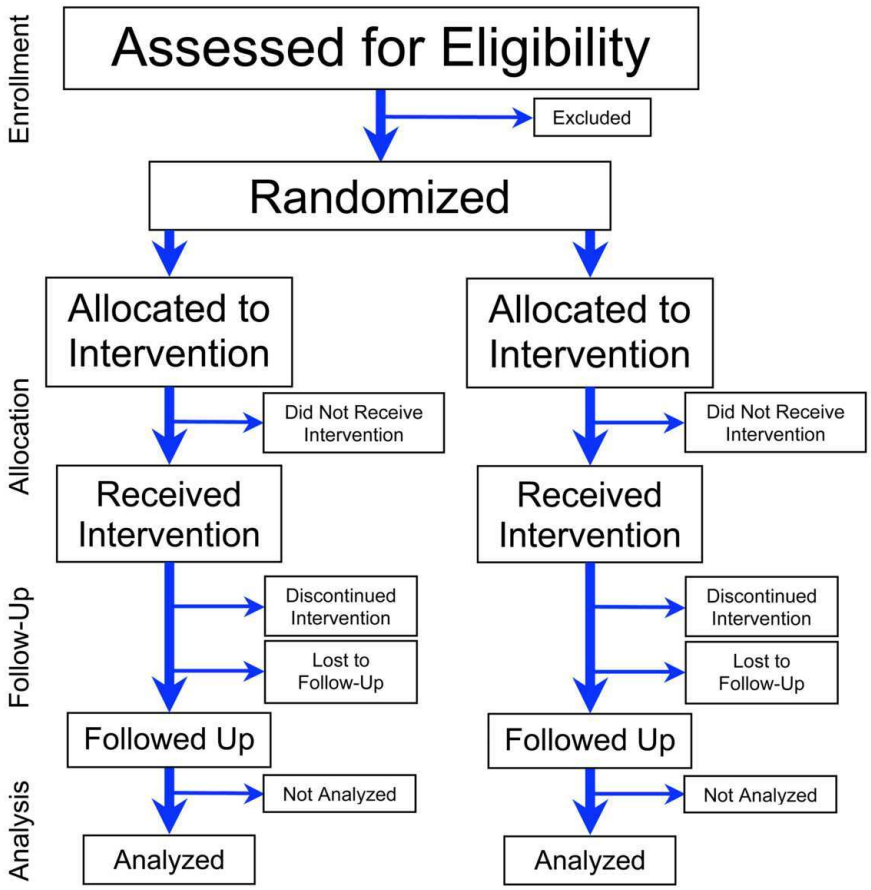


Figure 1.3: Four phases (enrollment, allocation, intervention, follow-up, and data analysis) of a parallel randomized controlled trial

1.1.2 Potential outcome framework

Following a general introduction, we describe a conceptual framework for RCTs.

The fundamental framework to uncover the causal effect of treatment from an RCT is the potential outcome framework (POF), which is also called the Rubin Causal Model (Rubin, 1974). In this framework, an experiment has an intervention or a treatment, and we are interested in its effect on an outcome or multiple outcomes.

In a study with n subjects indexed by $i (i = 1, ..., n)$, considering a treatment with two levels (1 for the active treatment and 0 for the control treatment), each subject i

has a pair of potential outcomes: $Y_i(0)$ and $Y_i(1)$, which are the potential outcomes under the control treatment and the active treatment, respectively.

There are some hidden assumptions of this notation:

- Assumption 1 - No interference: Subject i's potential outcomes do not depend on other subjects' treatments. This is sometimes called the no-interference assumption.
- Assumption 2 - Consistency: There are no other versions of the treatment. The treatment level be well defined, or have no ambiguity at least for the outcome of interest. This is sometimes called the consistency assumption.

Assumption 1 can be violated in infectious diseases. For instance, if one subject's family members receive flu vaccines, the chance of this subject's getting the flu decrease even if he or she does not receive the flu vaccine. Assumption 2 can be violated for treatment with complex components. For instance, when studying the effect of chemotherapy on cancer, the components of chemotherapy need to be specified.

The Assumptions 1 and 2 above together are called the Stable Unit Treatment Value Assumption (SUTVA) (Rubin, 2005).

For each subject, the individual causal effect is defined to be

$$E_i = Y_i(1) - Y_i(0).$$

However, each subject receives only one of the control treatment or the active treatment and we can only observe either $Y_i(0)$ or $Y_i(1)$.

Let $Z_i = \{0, 1\}$ be an indicator variable denoting the treatment actually received ($Z = 0$ for control treatment and $Z = 1$ for active treatment) for subject i. Thus, only one outcome Y_i,

$$Y_i = Z_i Y_i(1) + (1 - Z) Y_i(0),$$

is observed for each subject: the outcome under the actual treatment received, i.e. the experiment only reveals one of subject i's potential outcomes with the other one missing. For this reason, the potential outcomes framework is also called the counterfactual framework.

Identifying individual causal effects is not possible, but we can turn our attention

to an aggregated causal effect: the average causal effect in a population of individuals. The average treatment effect (ATE) is the average effect at the population level, of moving an entire population from untreated to treated (Imbens, 2004). The ATE is defined to be

$$E[Y_i(1) - Y_i(0)].$$

A related measure of treatment effect is the average treatment effect for the treated (ATT) (Imbens, 2004). The ATT is defined as

$$E[Y(1) - Y(0)|Z = 1].$$

The ATT is the average effect of treatment on those subjects who ultimately received the treatment. In an RCT these two measures of treatment effects coincide because the treated population will not differ systematically from the overall population due to randomization. Applied researchers should decide whether the ATE or the ATT is of greater utility or interest in their particular research context. For instance, when the barriers to a patient receiving a particular treatment in a study are substantial, the ATT becomes more relevant and informative than the ATE.

The treatment assignment mechanism, i.e., the probability distribution of Z, plays an important role in inferring causal effects. In RCTs, treatment is assigned by randomization. As a consequence of randomization, an unbiased estimate of the ATE can be directly computed from the study data. An unbiased estimate of the ATE is

$$E[Y_i(1) - Y_i(0)] = E[Y|Z = 1] - E[Y|Z = 0].$$

The aforementioned definition allows one to define the ATE in terms of a difference in means (continuous outcomes) or a difference in proportions or absolute risk reduction (dichotomous outcomes).

1.1.3 Limitations of traditional control arms

While randomized controlled trials offer theoretical advantages, they face certain limitations in practical implementation. These include high costs in terms of time and resources, challenges in generalizability (as volunteer participants may not represent the wider population), and the issue of loss to follow-up. In addition, a recent study

reveals that the failure rate for phase III trials stands at approximately 50 percent (Grignolo & Pretorius, 2016). Consequently, investing substantial resources, time, and costs into running a confirmatory trial might not always yield a viable and beneficial treatment.

The enrollment of a control arm (placebo or standard of care) in clinical trials can present significant ethical and practical challenges. This is especially true when few or no alternative treatments are available, as is often the case in settings such as rare diseases, oncology, and hematology. In such situations, it can be deemed unethical to randomize patients to a placebo or standard-of-care treatment known to have limited efficacy. Even when there is an existing treatment with clinical equipoise, recruitment challenges may become insurmountable if patients and clinicians are reluctant to risk randomization to the standard of care when a potentially more effective treatment is being investigated. These issues are further exacerbated when the population being studied is not large enough to power two treatment arms, or when there is high competition for clinical trial patients.

Besides, the RCTs face challenges in precision medicine due to their design, which is geared towards assessing average treatment effects rather than individual responses. The heterogeneity of treatment effects, ethical concerns in assigning standard care when personalized treatments may be superior, and the need for large, diverse sample sizes for statistical power, are notable limitations (Saad et al., 2017). Additionally, the rapid pace of scientific discovery in biomarkers and targeted therapies can outpace the lengthy timelines of RCTs. The clinical practice needs to adapt to the high specificity of patient selection based on genetic or molecular profiles inherent in precision medicine (Agarwala et al., 2018).

These limitations have contributed to the increasing prevalence of single-arm trials (SATs) in the field of drug development. Between December 1992, and May 2017, the FDA granted accelerated approval to 64 products in hematology or oncology, covering 93 new indications and 53 new molecular entities. The majority of initial indications were supported by single-arm trial designs (72%), relying on clinical experience to interpret the findings (Beaver et al., 2018). In this context, historical control arms have emerged as an alternative approach to complement non-randomized studies (Goring et al., 2019). The primary objective of utilizing historical control arms is to mitigate bias resulting from the absence of randomization in single-arm trials.

1.2 Historical control arms

1.2.1 Definition

In an externally controlled trial, the efficacy of an investigational treatment is evaluated by comparing patients receiving the treatment with a group of subjects external to the study (European Medicines Agency, 2001). This differs from an internal control arm in which patients from the same population are assigned to a control treatment within the same study. The external control group can be composed of patients who were treated prior to the concurrent clinical trial (referred to as the historical control arm) or patients who were treated during the same period but with a different clinical condition (illustrated in Figure 1.4).

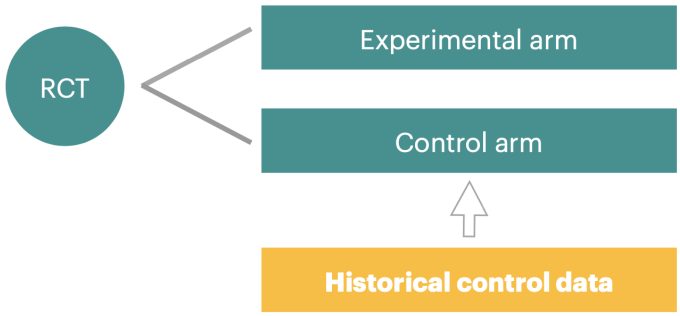


Figure 1.4: Illustration of historical control arm in andomized controlled trials (RCTs)

Historical control arms can be useful in certain situations, particularly when:

- It may not be ethical or feasible to have a control group, such as in studies of cancer or rare diseases where it would be challenging to enroll enough participants for a control group, or in diseases with high mortality rates where withholding treatment could be harmful.

- The effects of the condition without treatment are well-known, so comparison to a control group is unnecessary.

- A new treatment is tested against a well-established standard treatment, and the results of the standard treatment are predictable.

Terminology

Regarding terminology, there is some variation among researchers in the definition of an external control arm. Some researchers consider the external control arm as a broad term encompassing any control that is not a randomized control (Friends of Cancer Research, 2019). Historical control arm, on the other hand, is regarded as a specific subtype of external control, representing a non-concurrent comparator group of patients who received treatment in the past. This can include patient-level data or summary information from medical literature or other sources. Synthetic control is yet another subtype of external control that involves patient-level data from individuals outside of the trial, selected using statistical methods to account for baseline characteristic differences.

However, it should be noted that different studies may use the terms "external control", "historical control", and "synthetic control" interchangeably, employing various data sources and methods for statistical adjustment (Burger et al., 2021; Ghadessi et al., 2020; Hall et al., 2021; Mack et al., 2019; Thorlund et al., 2020; Viele et al., 2014).

In this manuscript, we use the term "historical control arm" to specifically focus on situations where the historical data was collected prior to its inclusion in the final analysis of clinical trials.

Role of historical control arm

In the context of single-arm trials, historical control arms serve as the sole comparator to establish benchmarks and provide contextual references for the investigational treatments, which represents the current focus on the application historical control arms (as illustrated in the upper section of Figure 1.5) (Davi et al., 2020; Ghadessi et al., 2020).

Besides, there is a growing interest in integrating historical control data into randomized controlled trials. An illustrative example is found in randomized controlled trials with unequal randomization, where more patients are randomized into the experimental arm. In such cases, historical control arms can effectively complement the concurrent control group, forming hybrid control arms that combine both internal and external control data to achieve a 1:1 ratio (as illustrated in the lower section of Figure 1.5). This methodological innovation optimizes patient allocation and reduces

the number of patients assigned to the control arm while ensuring robust comparisons (Yuan et al., 2019).

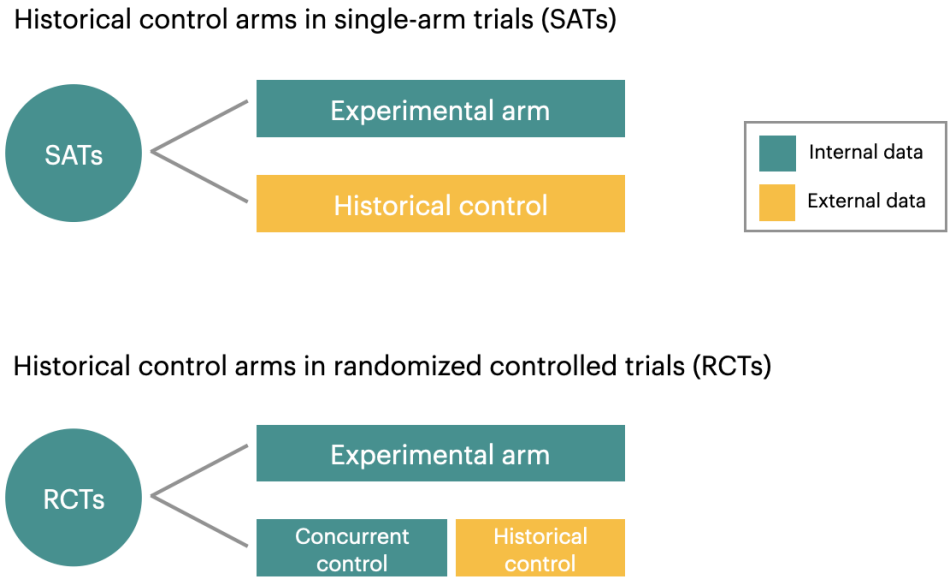


Figure 1.5: Illustration of different roles of historical control arms

Data sources

Historical control data can be drawn from various sources, including:

- **Previous Clinical Trials**: Historical control data from previous clinical trials can be highly valuable, including the control group or the treatment group which has already become the standard of care. Utilizing data from trials conducted for relevant indications and outcomes with similar inclusion and exclusion criteria can strengthen the foundation for new research.
- **Patient Registries**: Patient registries offer uniform data about specific groups of patients sharing a common condition or experience. These registries serve various purposes, such as observing the long-term course of diseases, assessing treatment cost-effectiveness, monitoring safety, evaluating the quality of care, and conducting post-marketing surveillance for drugs or devices. Incorporating data from patient registries into historical control arms can provide valuable insights.
- **Electronic Health Records (EHRs)**: EHRs are digital repositories containing comprehensive patient health information, including medical history, diag-

noses, medications, treatment plans, immunization dates, allergies, radiology images, laboratory and test results, and more. With the increasing prevalence of big data in healthcare, EHRs are becoming an essential source of historical control data. They can be particularly beneficial for indications with a limited number of patients.

- **Disease-specific databases**: Specialized databases that track outcomes for specific diseases can be instrumental in providing historical control data. These databases often contain long-term follow-up data, enabling the assessment of treatment durability and long-term effects.
- **Claims databases**: Claims databases compile information from health insurance claims, encompassing diagnoses, procedures, hospitalizations, prescriptions, and other services covered by insurance. Although they may lack some clinical details, they offer a vast quantity of data from diverse populations, making them valuable sources for historical control arms.
- **Biobanks**: Biobanks store biological samples like blood, tissue, DNA, etc., along with detailed health information. These repositories are particularly advantageous for genetic studies or research focused on biomarkers, as they offer invaluable historical control data for such investigations.

We can categorize the different data sources mentioned above into two primary categories: clinical trial data and observational data, commonly referred to as real-world data (RWD) (Franklin & Schneeweiss, 2017). The selection of the data source depends on the specific research questions under investigation. When the endpoints of interest require evaluation using specialized techniques, such as key biomarker testing, previous clinical trial data may be more appropriate due to the controlled and structured nature of these trials. On the other hand, RWD can be better suited for indications with limited clinical trial experience, as it reflects real-world patient experiences and provides insights into the effectiveness and safety of treatments in diverse populations. By understanding the strengths and limitations of each data source, researchers can make informed decisions to ensure the most relevant and reliable evidence is used for their studies.

1.2.2 Statistical issues

Clinical trials incorporating historical control arms share similarities with observational studies which have two common characteristics:

1. the objective is to elucidate cause-and-effect relationships;
2. it is not feasible to use controlled experimentation.

By this definition, an observational study has the same intent as a randomized experiment: to estimate a causal effect. However, an observational study differs from a randomized experiment in one design issue: the lack of randomization to allocate subjects to treatment and control groups which leads to systematic difference between treated and untreated subjects.

As illustrated in Figure 1.6, if we directly compare difference in outcomes between the treatment and control groups $E[Y|Z = 1] - E[Y|Z = 0]$, our conclusion has no causal interpretation but rather an association relationship. Thus, an unbiased estimate of the average treatment effect cannot be obtained by directly comparing outcomes between the two treatment groups: $E[Y_i(1) - Y_i(0)] \neq E[Y(1) - Y(0)]$ because the use of historical control data can introduce systematic biases due to lack of randomization.

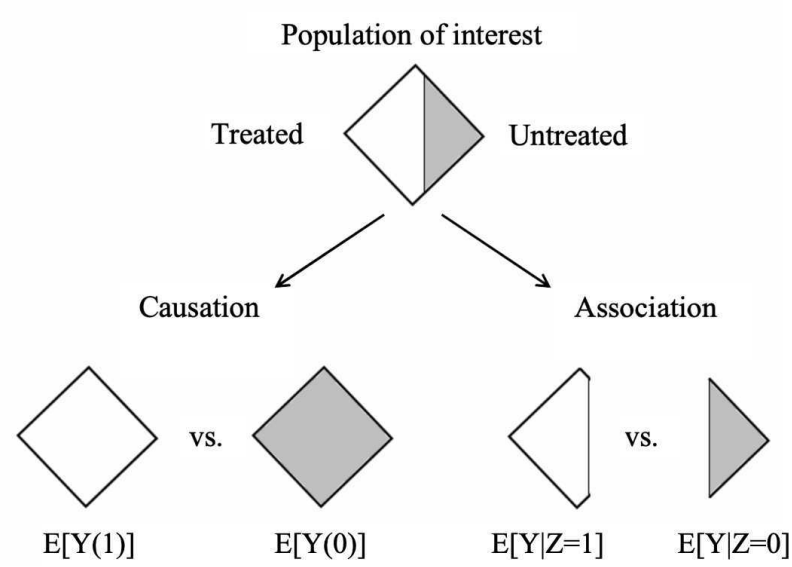


Figure 1.6: Causation versus association

One of the earliest papers to discuss incorporating historical data into both the design and analysis of a new study is a seminal paper by Pocock in 1976 (Pocock, 1976). This paper considers a design where patients are randomized to both treatment and control in the current study, even when "acceptable" historical control data are available. The historical control data are then incorporated into the final analysis of

the current trial.

Pocock proposed six evaluation criteria to assess the exchangeability between new and historical trials. These criteria are relevant regardless of whether multiple historical studies or just one are available. To deem a historical control group acceptable, the following conditions must be met:

- The group must have received a precisely defined standard treatment which must be the same as the treatment for the current trial controls.
- The group must have been part of a recent clinical study which contained the same requirements for patient eligibility.
- The method of treatment evaluation must be the same.
- The distributions of important patient characteristics in the historical group should be comparable with those in the new trial.
- The previous study must have been performed in the same organisation with largely the same clinical investigators.
- There must be no other indications leading one to expect different results between the randomized and historical controls (e.g. differing enrollment rates).

Several of the criteria mentioned are stringent. For instance, insisting that the historical study be conducted within the same organization as the current trial might exclude other literature data and dismiss a wealth of pertinent information. Similarly, mandating identical enrollment rates appears redundant unless significant disparities exist between the trials — a factor often discernible only after the current trial's recruitment concludes. Nevertheless, these criteria serve as a good starting point for identifying relevant historical studies. The FDA has also recognized these criteria and referenced them in the statistical review for drug approvals (FDA, 2015).

1.3 Context and structure of the manuscript

This thesis is financially supported by the Association Nationale Recherche Technologie (ANRT) and Sanofi and is undertaken in collaboration between Institut Curie and Sanofi. Both entities exhibit an interest in the integration of historical control within

clinical trials. The overall objective is to evaluate the feasibility of incorporating historical control data into clinical trials, specifically within the designated case studies of interest.

The manuscript is structured as follows:

- **Chapter 2** introduces the relevant statistical methods to reduce the bias of using historical control data.
- **Chapter 3** reviews current perspectives on historical control arms in the field of cancer drug development, highlighting key concerns associated with their implementation.
- **Chapter 4** presents a case study illustrating the creation of a historical control arm using clinical trial data for a single-arm trial.
- **Chapter 5** presents a case study demonstrating the generation of a historical control arm using observational clinical data, with the application of the target trial framework.
- Finally, **Chapter 6** concludes the thesis, discusses the limitations encountered throughout the research, and proposes potential areas for future work.

Chapter 2

Statistical methodology for historical control arms

Contents

To reduce biases when including external controls in statistical analyses, various methods have been developed including both frequentist and Bayesian approaches.

Bayesian approaches consider outcome heterogeneity and discount the historical control data when incorporating it into the new clinical trial, such as power prior, commensurate prior, or meta-analytic prior (Hobbs et al., 2011; Lewis et al., 2019; Schmidli et al., 2014). As these methods borrow information from the historical data, they are often called "Bayesian borrowing methods".

On the other hand, frequentist approaches involve two primary steps. First, a balance score is estimated using selected covariates that may affect treatment assignment and outcome. Examples of balance scores include propensity score and Mahalanobis distance (Austin, 2011b; De Maesschalck et al., 2000). Second, the balance score is used to create comparable external and internal cohorts through methods like matching, inverse probability weighting, and covariate adjustment.

In this chapter, we assume that the historical control data has been well selected in terms of patient eligibility and characteristics, treatment and outcome evaluation, and we focus on presenting the state-of-the-art statistical methods to adjust for the confounding bias in using historical control arms.

- **Section 2.1** introduces the Bayesian methods in their general forms.
- **Section 2.2** presents the frequentist methods of propensity score analysis.
- **Section 2.3** discusses the considerations of choosing statistical methods when using historical control arms.

2.1 Bayesian methods

To introduce the historical data approaches, we consider a standard trial design comparing one experimental treatment to control, assuming historical data are available for the control arm only. Let D_t, D_c, and D_h denote data from the current treatment group, current control group and historical data, respectively. Let θ_t denote the parameter of interest in the treatment group and θ_c the parameter of interest in the control groups. Where a method assumes that the true underlying parameters are different in the historical and current controls, θ_h denotes the parameter of interest in the historical controls.

2.1.1 Pooling

Incorporating historical data into the analysis of a contemporary trial can begin with the straightforward assumption that the historical control data and the control data from the current trial are exchangeable. In this approach, the historical data serves as the prior information for the control arm of the current study. This prior information is subsequently updated using the control data from the present study, based on Bayes' theorem (Viele et al., 2014). We assume an initial prior distribution, denoted as $\pi_0(\theta_c)$, for the control parameter of interest before considering the historical data, $\pi_0(\theta_c)$ is updated to form a posterior distribution that incorporates the historical data, which forms the prior for the current study control parameter, given by,

$$\pi(\theta_c|D_h) \propto L(\theta_c|D_h)\pi_0(\theta_c) \tag{2.1}$$

where $L(\theta_c|D_h)$ is the likelihood of the historical data. $\pi(\theta_c|D_h)$ is then updated with the control data from the current study using Bayes theorem. This results in a posterior distribution for the control parameter as follows:

$$\pi(\theta_c|D_c, D_h) \propto L(\theta_c|D_c)L(\theta_c|D_h)\pi_0(\theta_c) = L(\theta_c|D_c, D_h)\pi_0(\theta_c),$$

where $L(\theta_c|D_c)$ is the likelihood of the current control data. This is the same as pooling the current and historical control data as if they were from the same study.

2.1.2 Power prior

The power prior assumes that the historical data and current control data are estimating the same underlying parameter of interest θ_c (Chen & Ibrahim, 2000). An initial non-informative prior $\pi_0(\theta_c)$ is assumed for θ_c before the historical data are observed. Then the likelihood of the historical data is raised to a power α_0, where the power quantifies the uncertainty in the similarity between the historical and current studies. The prior for the current study control arm is then,

$$\pi(\theta_c|D_h) \propto \pi_0(\theta_c)L(\theta_c|D_h)^{\alpha_0}$$

where α_0 is a fixed value and lies between zero and one.

α_0 is a weight parameter that controls the degree of borrowing. When α_0 is zero,

there is no borrowing from the historical data which means no historical data are used in the final analysis and the prior reduces to the initial non-informative prior,

$$\pi(\theta_c|D_h) = \pi_0(\theta_c)$$

and when α_0 equals to one, all of the historical data are used in the final analysis which pooling the current and historical controls. In this case, the prior becomes Equation (2.1). The power can be given a value above one, however, in the area of historical data, we consider the most reliable information to be the data from the current randomized controlled trial and are therefore unlikely to want to give the historical data more weight in the final analysis than the current control data. The power α_0 can be interpreted as a relative precision parameter for the historical data.

2.1.3 Modified power prior

When using the power prior, α_0 must be chosen in advance. A natural choice for $\pi(\alpha_0)$ is a beta distribution, given the desirability of a power between zero and one. However, there is no agreement in choosing the value of α_0 (Neuenschwander et al., 2009).

A modified power prior was proposed (Banbeta et al., 2019) to estimate α_0 using available data:

$$\pi(\theta_c, \alpha_0|D_h) \propto C(\alpha_0)\pi_0(\theta_c)L(\theta_c|D_h)^{\alpha_0}\pi(\alpha_0),$$

where $C(\alpha_0)$ is a scaling constant that depends only on α_0:

$$C(\alpha_0) = \frac{1}{\int_{\theta_c} L(\theta_c|D_h)^{\alpha_0}\pi_0(\theta_c)d\theta_c}.$$

2.1.4 Commensurate prior

The commensurate prior approach assumes different underlying parameters for the current and historical controls and the distribution of the parameters of the current data is centered on the corresponding parameters of the historical data (Hobbs et al., 2012).

The location commensurate prior for θ_c is a conditional prior distribution, centered at the historical parameter θ_h with a fixed value τ that controls the cross-study

borrowing. The joint distribution of θ_c and θ_h before the current trial is then given by,

$$\pi(\theta_c, \theta_h | D_h, \tau) \propto \pi(\theta_c | \theta_h, \tau) \pi_0(\theta_h) L(\theta_h | D_h),$$

where $L(\theta_h|D_h)$ is the likelihood of the historical data and $\pi_0(\theta_h)$ is an initial prior for the historical parameter before the historical data are observed. τ controls the degree of borrowing. Lower values of τ indicate increased commensurability (reduced variability) between the current and historical data parameters and induce increased borrowing from the historical data to inform inference on θ_c.

Similar to the modified power prior, there is a single parameter that governs how much historical data are borrowed and incorporated into the final inference on the current study control parameter τ. τ can also be treated as a random variable rather than a fixed value. The choice of prior for τ is similar to the prior for the between study variance parameter in a meta-analysis (Hobbs et al., 2012). It is generally recommended that an informative prior should be used on the between study variance parameter in a meta-analysis since the parameter is not well estimated from the data when there are few studies (Pullenayegum, 2011). An informative prior is required to induce sufficient borrowing from the historical data.

2.1.5 Meta-analytic-predictive (MAP) prior

The meta-analytic-predictive (MAP) prior assumes model parameters are exchangeable and drawn from the same distribution (Schmidli et al., 2014). It is a robust mixture prior composed of a two-component mixture distribution of conjugate priors. The first component of the mixture distribution is an informative component based on the historical data and the second component is a weakly-informative component. The form of the weakly-informative component is dependent on the type of outcome data. The weights given to each component of the mixture distribution in the prior are chosen by the study designer based on how relevant the historical data are thought to be to the current study control data. The prior weight given to the informative component of the robust mixture distribution based on the historical data can be interpreted in a similar way to the power chosen in the power prior, when α_0 is a fixed value. The weakly-informative component of the mixture distribution gives a heavy tailed prior distribution compared to using only the historical data as a prior and adds robustness against prior-data conflict. Using a mixture prior allows added flexibility while maintaining the convenience of using a conjugate prior.

Let $\pi_1(\theta_c), ..., \pi_J(\theta_c)$ be proper probability density functions. Then given weights, $w_1, ..., w_J$, where $w_j > 0$ and $\sum_{j=1}^{J} w_j = 1$, the mixture distribution,

$$\pi(\theta_c) = \sum_{j=1}^{J} w_j \pi_j(\theta_c)$$

is also a proper probability density.

If the individual mixture components are conjugate prior distributions, then the posterior distribution is also a mixture of conjugate distributions with updated parameter values and weights. Assuming the prior density for θ_c of

$$\pi^{(0)}(\theta_c) = \sum_{j=1}^{J} w_j^{(0)} \pi_j^{(0)}(\theta_c),$$

where the superscript (0) denotes a prior distribution or weight, $w_j^{(0)}$ are the prior weights and $\pi_j^{(0)}(\theta_c)$ are the individual conjugate prior mixture distribution components (for the robust mixture prior, $\pi_1^{(0)}(\theta_c|D_h)$ would be the informative component of the mixture distribution based on the historical data and $\pi_2^{(0)}(\theta_c)$ would be a weakly-informative mixture component), the posterior distribution is given by

$$\pi^{(1)}(\theta_c|D_c) = \frac{\sum_{j=1}^{J} w_j^{(0)} \pi_j^{(0)}(\theta_c) L(\theta_c|D_c)}{C} = \sum_{j=1}^{J} w_j^{(1)} \pi_j^{(1)}(\theta_c|D_c)$$

where the superscript (1) denotes a posterior distribution or weight, $L(\theta_c|D_c)$ is the likelihood of the current trial control data and $C = \sum_{j=1}^{J} w_j^{(0)} c_j$,

$$\pi_j^{(1)}(\theta_c|D_c) = \frac{w_j^{(0)}(\theta_c) L(\theta_c|D_c)}{c_j},$$

$$w_j^{(1)} = \frac{w_j^{(0)} c_j}{\sum_{j=1}^{J} w_j^{(0)} c_j},$$

$$c_j = \int_{-\infty}^{\infty} w_j^{(0)}(\theta_c) L(\theta_c|D_c) d\theta_c,$$

where θ_c is either a single parameter or a vector of parameters. The posterior mixture distribution components $\pi_j^{(1)}(\theta_c|D_c)$ are then obtained from standard conjugate Bayesian prior to posterior updates. The updated posterior weights sum to one and

are calculated using the marginal likelihood of the data for each component of the mixture prior distribution.

2.2 Propensity score analysis

Propensity score analysis is a statistical method developed in the early 1980s to address confounding factors and retrieve causal effects in observational (or non-randomized) studies (Rosenbaum & Rubin, 1983). Rosenbaum and Rubin recognized the limitations of traditional regression models in studies studies and proposed the use of propensity scores as a tool to balance observed covariates between treated and control groups, thereby reducing confounding and enabling more reliable causal inference. The propensity score analysis is built in the potential framework presented in Section 1.1.2.

Over the years, propensity score analysis has gained popularity and has become widely used in medical research and other disciplines. Numerous methodological developments and refinements have been made, addressing issues such as missing data, multiple treatments, and non-binary treatments (Austin, 2014; Choi et al., 2019; McCaffrey et al., 2013; Zhao et al., 2020). Various matching algorithms, weighting methods, and sensitivity analyses have been proposed to improve the robustness of propensity score analysis (Austin, 2017; Austin & Stuart, 2015; Rudolph & Stuart, 2018).

Propensity score analysis has been increasingly applied in clinical trials featuring historical control arms to achieve comparability between internal and external patient groups (Gökbuget et al., 2016).

2.2.1 Propensity score

For subject $i(i = 1, ..., n)$, we have p-dimensional observed covariates X_i, a binary treatment indicator $Z_i = \{0, 1\}$, the propensity score is defined to be the probability of treatment assignment conditional on observed baseline covariates:

$$e_i = Pr(Z_i = 1|X_i).$$

We say the propensity score is a balancing score: conditional on the propensity

score, the distribution of measured baseline covariates is similar between treated and untreated subjects:

$$Z \perp\!\!\!\perp X|e.$$

Thus, in a set of subjects all of whom have the same propensity score, the subjects have different values ofobserved baseline covariates X_i, but the distribution of X_i will be the same between the treated and untreated subjects.

The propensity score plays a role in both randomized experiments and observational studies. In randomized experiments, the true propensity score is known and determined by the study design itself. For example, the true propensity score is 0.5 in a 1:1 randomized study. On the other hand, in observational studies, the true propensity score is generally unknown. However, it can be estimated using the available study data. In practice, the propensity score is most often estimated using a logistic regression model, in which treatment status is regressed on observed baseline characteristics. The estimated propensity score is the predicted probability of treatment derived from the fitted regression model. While logistic regression is the most commonly used method for estimating the propensity score, alternative methods have been investigated to address the misspecification issue of the regression model. These include bagging or boosting, recursive partitioning or tree-based methods, random forests, and neural networks (Setoguchi et al., 2008; Westreich et al., 2010).

Treatment assignment can be strongly ignorable if the following two conditions hold: (a) treatment assignment is independent of the potential outcomes conditional on the observed baseline covariates:

$$(Y(1), Y(0)) \perp\!\!\!\perp Z|X$$

(b) every subject has a nonzero probability to receive either treatment:

$$0 < P(Z = 1|X) < 1$$

If treatment assignment is strongly ignorable, conditioning on the propensity score allows one to obtain unbiased estimates of average treatment effects. The aforementioned first condition is also referred to as the "no unmeasured confounders" assumption: the assumption that all variables that affect treatment assignment and outcome have been measured. Given that the assumption of no unmeasured confounders influencing treatment assignment is critical in propensity score analyses, it

is recommended to conduct sensitivity analyses to evaluate the robustness of study findings to this assumption. It is important to recognize that while the assumption of strongly ignorable treatment assignment and the absence of unmeasured confounding is explicitly stated in the context of propensity score analyses, this assumption also underlies regression-based approaches used to estimate treatment effects in observational studies.

Four different propensity score methods are used for reducing the confounding bias when estimating the effects of treatment on outcomes: propensity score matching, stratification (or subclassification) on the propensity score, inverse probability of treatment weighting (IPTW) using the propensity score, and covariate adjustment using the propensity score (Austin, 2011b; Rosenbaum & Rubin, 1983). We present some of the methods in the following subsections.

2.2.2 Variable selection for the propensity score model

The propensity score is defined as the probability of treatment selection conditional on measured baseline covariates. A natural question that arises is what variables should be included in the propensity score model. A logical approach would be to include variables that significantly influence the treatment selection. The primary objective of propensity score analyses is to ensure balance in observed baseline variables across treatment groups. However, achieving balance for every covariate isn't uniformly crucial. Balancing covariates that are prognostically significant is of greater importance than balancing those that solely influence treatment selection without impacting the outcome. In fact, past research indicates that it's more beneficial to incorporate either prognostically significant covariates (those linked to outcomes) or the confounding covariates (those linked to both treatment and outcomes) into the propensity score model than to include variables that only influence the treatment-selection process (Austin et al., 2007).

The set of variables that are either prognostically important or that confound the treatment-outcome relationship can be identified using causal diagrams, supplemented with insights from relevant literature and expert opinions (Hernan & Robins, 2023). Unfortunately, such complete knowledge is often unavailable.

A practical approach was proposed to confounder selection decisions when the somewhat less stringent assumption is made that knowledge is available for each

covariate whether it is a cause of the exposure, and whether it is a cause of the outcome (VanderWeele, 2019). Based on recent theoretically justified developments in the causal inference literature, the following proposal is made for covariate control decisions: control for each covariate that is a cause of the exposure, or of the outcome, or of both; exclude from this set any variable known to be an instrumental variable; and include as a covariate any proxy for an unmeasured variable that is a common cause of both the exposure and the outcome.

It's important to highlight that we don't recommend using statistical hypothesis testing in the analytic sample to identify the necessary variables. This recommendation aligns with the principle of keeping 'design' distinct from 'analysis' and refraining from using outcome data in the propensity score process (Austin & Stuart, 2015).

2.2.3 Propensity score matching

Propensity score matching implies forming matched sets of treated and untreated subjects who share a similar value of the propensity score (Stuart, 2010). Propensity score matching enables the estimation of the ATT (Imbens, 2004). The most common implementation of propensity score matching is one-to-one or pair matching, where treated and untreated individuals are paired based on similar propensity score values.

Once a matched sample has been formed, the treatment effect can be estimated by directly comparing outcomes between the treated and untreated individuals within the matched sample. For continuous outcomes, the treatment effect can be estimated as the difference between the mean outcome for the treated individuals and the mean outcome for the untreated individuals in the matched sample. For dichotomous outcomes, the treatment effect can be estimated as the difference in the proportion of individuals experiencing the event in each group (treated vs. untreated) within the matched sample. The treatment effect for binary outcomes can also be described using the relative risk or the Number Needed to Treat (NNT). Therefore, reporting of treatment effects can be done using the same metrics commonly used in RCTs.

Once the treatment effect has been estimated in the propensity score matched sample, the variance of the estimated treatment effect and its statistical significance can be determined. When using a matched estimator, the variance should be calculated using an appropriate method for paired experiments since the propensity score matched sample does not consist of independent observations. Instead, treated and

untreated individuals within the same matched set share similar propensity score values and their observed baseline covariates are derived from the same multivariate distribution. In the presence of confounding, baseline covariates are associated with outcomes, making matched individuals more likely to have similar outcomes compared to randomly selected individuals. Accounting for the lack of independence in the propensity score matched sample is necessary when estimating the variance of the treatment effect. Recent studies utilizing Monte Carlo simulations have demonstrated that variance estimators incorporating matching more accurately capture the sampling variability of the estimated treatment effect (Austin, 2009a). Therefore, a paired t-test can be utilized to assess the statistical significance of the treatment effect on continuous outcomes, while McNemar's test can be employed for assessing the statistical significance of a difference in proportions for dichotomous outcomes.

The analysis of a propensity score matched sample can mimic that of an RCT, as outcomes can be directly compared between treated and untreated individuals within the propensity score matched sample. In the context of an RCT, it is expected that the distribution of covariates will be similar on average between treatment groups. However, individual RCTs may exhibit residual differences in baseline covariates between treatment groups. Regression adjustment can be employed to mitigate bias arising from residual differences in observed baseline covariates between treatment groups. Regression adjustment enhances precision for continuous outcomes and increases statistical power for continuous, binary, and time-to-event outcomes. Similarly, in propensity score matched samples, achieving covariate balance is a property observed in large samples. Propensity score matching can be combined with additional matching on prognostic factors or regression adjustment (Imbens, 2004).

The choice of methods for forming matched pairs of treated and untreated individuals when matching on the propensity score should be determined based on different methods.

Replacement

First, one must choose between matching without replacement and matching with replacement (Stuart, 2010). When using matching without replacement, once an untreated subject has been selected to be matched to a given treated subject, that untreated subject is no longer available for consideration as a potential match for subsequent treated subjects. As a result, each untreated subject is included in at most

one matched set. In contrast, matching with replacement allows a given untreated subject to be included in more than one matched set. When matching with replacement is used, variance estimation must account for the fact that the same untreated subject may be in multiple matched sets (Hill & Reiter, 2006).

Algorithm

A second choice is between greedy and optimal matching (Stuart, 2010). In greedy matching, a treated subject is first selected at random. The untreated subject whose propensity score is closest to that of this randomly selected treated subject is chosen for matching to this treated subject. This process is then repeated until untreated subjects have been matched to all treated subjects or until one has exhausted the list of treated subjects for whom a matched untreated subject can be found. This process is called greedy because at each step in the process, the nearest untreated subject is selected for matching to the given treated subject, even if that untreated subject would better serve as a match for a subsequent treated subject. An alternative to greedy matching is optimal matching, in which matches are formed so as to minimize the total within-pair difference of the propensity score. Optimal matching did no better than greedy matching in producing balanced matched samples (Gu & Rosenbaum, 1993).

Distance

There are two primary methods for selecting untreated subjects whose propensity score is "close" to that of a treated subject: nearest neighbor matching and nearest neighbor matching within a specified caliper distance. Nearest neighbor matching selects for matching to a given treated subject that untreated subject whose propensity score is closest to that of the treated subject. If multiple untreated subjects have propensity scores that are equally close to that of the treated subject, one of these untreated subjects is selected at random. In this case, no restrictions are placed upon the maximum acceptable difference between the propensity scores of two matched subjects. Nearest neighbor matching within a specified caliper distance is similar to nearest neighbor matching with the further restriction that the absolute difference in the propensity scores of matched subjects must be below some prespecified threshold (the caliper distance). Thus, for a given treated subject, one would identify all the untreated subjects whose propensity score lay within a specified distance of that of

the treated subject. From this restricted set of untreated subjects, the untreated subject whose propensity score was closest to that of the treated subject would be selected for matching to this treated subject. If no untreated subjects had propensity scores that lay within the specified caliper distance of the propensity score of the treated subject, that treated subject would not be matched with any untreated subject. The unmatched treated subject would then be excluded from the resultant matched sample. There are theoretical arguments for matching on the logit of the propensity score, as this quantity is more likely to be normally distributed, and for using a caliper width that is a proportion of the standard deviation of the logit of the propensity score. Recent studies examined optimal caliper widths when estimating risk differences and differences in means (Austin, 2011a). It was suggested that researchers use a caliper of width equal to 0.2 of the standard deviation of the logit of the propensity score as this value (or one close to it) minimized the mean squared error of the estimated treatment effect in several scenarios.

Propensity score matching can be conducted using a variety of packages in R: the Matching (Sekhon, 2011), MatchIt (Ho et al., 2011), and Optmatch (Hansen & Klopfer, 2006) packages allow one to implement a variety of different matching methods.

2.2.4 Inverse probability of treatment weighting using the propensity score

Inverse probability of treatment weighting (IPTW) using the propensity score uses weights based on the propensity score to create a synthetic sample in which the distribution of measured baseline covariates is independent of treatment assignment. As mentioned earlier, let Z_i be an indicator variable denoting whether or not the ith subject was treated; e_i denotes the propensity score for the ith ($i = 1,\ldots,n$) subject. Weights can be defined as

$$w_i = \frac{Z_i}{e_i} + \frac{1 - Z_i}{1 - e_i}.$$

A subject's weight is equal to the inverse of the probability of receiving the treatment that the subject actually received.

Let Y_i denote the outcome variable measured on the ith subject. An estimate of

the ATE is

$$\frac{1}{n}\sum_{i=1}^{n}\frac{Z_i Y_i}{e_i} - \frac{1}{n}\sum_{i=1}^{n}\frac{(1-Z_i)Y_i}{1-e_i}.$$

That variance estimation must account for the weighted nature of the synthetic sample, with robust variance estimation commonly being used to account for the sample weights (Joffe et al., 2004).

The weights may be inaccurate or unstable for subjects with a very low probability of receiving the treatment received. The use of stabilizing weights has been proposed to address this issue (Robins et al., 2000).

Besides, using weights equal to

$$w_{i,ATT} = Z_i + \frac{(1-Z_i)e_i}{1-e_i}$$

allows one to estimate the ATT, whereas the use of weights equal to

$$w_{i,ATC} = \frac{Z_i(1-e_i)}{e_i} + (1-Z_i)$$

allows one to estimate the average effect of treatment in the controls.

2.2.5 Comparison of the different propensity score methods

Propensity score methods for reducing the confounding bias when estimating the effects of treatment on outcomes include propensity score matching, inverse probability of treatment weighting using the propensity score, stratification on the propensity score, and covariate adjustment using the propensity score.

However, several studies indicate that propensity score matching more effectively eliminates systematic differences in baseline characteristics than other methods (Austin et al., 2007; Austin & Mamdani, 2006). n some scenarios, both propensity score matching and IPTW were equally effective at reducing systematic differences between treated and untreated subjects. However, in other cases, propensity score matching was slightly more effective than IPTW (Austin, 2009b). A further distinction among the four propensity score techniques is that both covariate adjustment using the propensity score and IPTW might be particularly sensitive to the accuracy of the propensity score estimation (Rubin, 2004).

Therefore, in this thesis, we focus solely on the methods of propensity score matching and inverse probability of treatment weighting using the propensity score.

2.2.6 Balance diagnostics

The true propensity score is a balancing score: conditional on the true propensity score, the distribution of measured baseline covariates is independent of treatment assignment. Therefore, in strata of subjects that have the same propensity score, the distribution of measured baseline covariates will be the same between treated and untreated subjects. In non-randomized studies, this exact propensity score remains unknown, necessitating estimation from the study data. A crucial aspect of any propensity score analysis is ensuring the model's appropriate specification.

To determine the adequacy of the propensity score model, one should assess if the distribution of observed baseline covariates is consistent between treated and untreated subjects with the same estimated propensity score. If, after conditioning on the propensity score, there remain systematic differences in baseline covariates between treated and untreated subjects, this can be an indication that the propensity score model has not been correctly specified. For propensity score matching, this assessment entails comparing treated and untreated subjects within the matched sample. For IPTW, it involves comparing the two groups in the sample that's weighted by the inverse probability of treatment.

Comparing the similarity of treated and untreated subjects in the matched sample should begin with a comparison of the means or medians of continuous covariates and the distribution of their categorical counterparts between treated and untreated subjects. The standardized difference can be used to compare the mean of continuous and binary variables between treatment groups (multilevel categorical variables can be represented using a set of binary indicator variables.

For a continuous covariate, the standardized difference is defined as

$$d = \frac{(\overline{x}_{treatment} - \overline{x}_{control})}{\sqrt{\frac{s^2_{treatment} - s^2_{treatment}}{2}}}$$

where $\overline{x}_{treatment}$ and $\overline{x}_{control}$ denote the sample mean of the covariate in treated and untreated subjects, respectively, whereas $s^2_{treatment}$ and $s^2_{control}$ denote the sample variance of the covariate in treated and untreated subjects, respectively.

For dichotomous variables, the standardized difference is defined as

$$d = \frac{(\hat{p}_{treatment} - \hat{p}_{control})}{\sqrt{\frac{\hat{p}_{treatment}(1-\hat{p}_{treatment}) - \hat{p}_{control}(1-\hat{p}_{control})}{2}}}$$

where $\hat{p}_{treatment}$ and $\hat{p}_{control}$ denote the prevalence or mean of the dichotomous variable in treated and untreated subjects, respectively.

The standardized difference compares the difference in means in units of the pooled standard deviation. Furthermore, it is not influenced by sample size and allows for the comparison of the relative balance of variables measured in different units. Although there is no universally agreed upon criterion as to what threshold of the standardized difference can be used to indicate important imbalance, a standard difference that is less than 0.1 has been taken to indicate a negligible difference in the mean or prevalence of a covariate between treatment groups (Normand et al., 2001).

The methods described are for use in the context of one-to-one matching on the propensity score. These methods can be adapted to many-to-one matching and IPTW using the propensity score [Joffe et al. (2004); morganDiagnosticRoutineDetection2008].

The standardized difference provides a framework for comparing the mean or prevalence of a baseline covariate between treatment groups in the propensity score matched sample. However, a thorough examination of the comparability of treated and untreated subjects in the propensity score matched sample should not stop with a comparison of means and prevalences. The true propensity score is a balancing score: within strata matched on the true propensity score, the distribution of observed baseline covariates is independent of treatment status. Thus, the entire distribution of baseline covariates, not just means and prevalences, should be similar between treatment groups in the matched sample. Therefore, higher order moments of covariates and interactions between covariates should be compared between treatment groups (Austin, 2009c; Morgan & Todd, 2008). Similarly, graphical methods such as side-by-side boxplots, quantile-quantile plots, cumulative distribution functions, and empirical nonparametric density plots can be used to compare the distribution of continuous baseline covariates between treatment groups in the propensity score matched sample (Austin, 2009c).

Balance diagnostics is a pivotal stage within the framework of propensity score analysis. Approaches to assessing the specification of the propensity score model

are based on comparing the distribution of measured baseline covariates between treated and untreated subjects who have similar propensity score values. Balance diagnostics for assessing the specification of the propensity score are more transparent than validating the precision of an outcome regression model (Austin, 2011b).

Formulating a propensity score model involves an iterative approach. It begins with the specification of an initial propensity score model. Subsequently, the comparability between treated and untreated subjects within the resulting matched sample is evaluated. Should significant residual systematic disparities persist among these groups, adjustments can be made to the initial propensity score model. These adaptations might include the incorporation of supplementary covariates, the introduction of interactions among existing covariates, or the utilization of nonlinear terms to capture the nuanced relationship between continuous covariates and treatment status. This iterative process continues until systematic differences in observed baseline covariates between treated and untreated subjects are either eradicated or minimized to an acceptable degree.

It is important to note that throughout each stage of this iterative procedure, it is not recommended to rely on the statistical significance of the estimated regression coefficients in the propensity score model (assuming a logistic regression model is employed). Instead, the focus should remain directed towards the goal of generating a matched sample where the distribution of observed baseline covariates is similar between treated and untreated subjects.

2.3 Discussion

Choosing between Bayesian methods and propensity score methods when using a historical control arm in clinical trials depends on various factors. Here we discuss the following factors:

Research objective

Propensity score methods are primarily used to balance covariates between treatment groups in observational studies, thus may be preferred when the goal is to control for confounding factors and obtain unbiased estimates of treatment effects. Bayesian methods can offer the flexibility to incorporate prior beliefs regarding treatment effects

if the aim is to harness the synergy between historical and current data to enhance inferences.

Data availability

Implementation of propensity score methods necessitates a comprehensive understanding of the covariates influencing treatment assignment. Adequate data availability, including relevant covariates, is pivotal for effective modeling of treatment assignment probabilities. Bayesian approaches can gracefully handle data with missing values or incomplete historical records. It is possible to specify prior distributions that encapsulate the uncertainty inherent in historical data, providing an advantage in scenarios where data completeness is a challenge.

Regulatory requirements

Considering regulatory requirements is an imperative step in the decision-making process. It is advisable to engage in consultation with regulatory authorities or domain experts to ensure strict adherence to regulatory guidelines. Additionally, assessing whether any specific preferences or recommendations exist regarding the selection of the analytical methodology is crucial.

Chapter 3

Current perspectives on historical control arms

Contents

3.1 Introduction

Well-designed randomized controlled trials (RCTs) are the gold standard for evaluating the efficacy of new treatments (Schulz et al., 2010). RCTs establish causal conclusions by randomly assigning patients to either an investigational or concurrent control treatment that usually consists of a placebo or standard of care. Nevertheless, ethical or feasibility issues can hinder the randomization process, especially in oncology drug development (Agarwala et al., 2018). When a new promising treatment is being studied in early-stage clinical trials for cancer with high unmet needs, enrolling patients into the control arm might be considered unethical. Besides, with the advances in molecular classification and precision medicine in oncology - further dividing patient populations into smaller groups - the number of patients available for a particular clinical trial may be insufficient to produce valid evidence. Regulators acknowledge these challenges and have granted conditional approvals to submissions using single-arm trials requiring further confirmatory evidence from post-approval studies (Goring et al., 2019; Hatswell et al., 2020).

In this context, historical control arms have emerged as an alternative approach to knowledge production that is more pragmatic than RCTs and can be traced back to 1970s (Pocock, 1976; Viele et al., 2014). By providing contextual information and improving the interpretation of single-arm trial results, historical control arms can help reduce the bias due to the lack of randomization (Davi et al., 2020). Furthermore, increasingly available clinical data from historical clinical trials or real-world databases provide the potential to expand the use of historical control data to minimize patient burden and facilitate study conduct in the drug development process (Lim et al., 2018).

Advancements in technology and the evolving policy landscape have created an opportune environment for leveraging real-world data (RWD) to improve clinical evidence generation (Khozin et al., 2017). RWD are qualified by regulators as routinely collected data relating to patient health status and the delivery of health care other than traditional RCTs (Cave et al., 2019; FDA, 2018). These data can be gathered from various sources such as electronic health records (EHRs), claims, registries, or patient-generated data. Clinical evidence regarding the usage and potential benefits or risks of a medical product derived from the analysis of RWD is then considered real-world evidence (RWE).

Regulatory authorities have signaled their support for using RWD to generate

clinical evidence. In 2018, the US Food and Drug Administration (FDA) published the framework for RWE underpinned by three pillars: whether RWD are fit for use, whether the trial or study design can provide adequate evidence, and whether the study conduct meets regulatory requirements (FDA, 2018). In 2019, the European Medicines Agency (EMA) published the Operational, Technical, and Methodological (OPTIMAL) framework for regulatory use of valid RWE in safety, efficacy and benefit-risk monitoring (Cave et al., 2019). The EMA has also outlined its vision that by 2025, the use of real-world evidence will have been enabled, and the value will have been established across the spectrum of regulatory use cases (Arlett et al., 2022).

Recent evidence shows that RCT and RWE findings were not always matched despite attempts to emulate RCT design and confounder adjustment (Franklin et al., 2021). Thus, challenges remain before historical control arms can be an integrated part of decision-making (Eichler et al., 2021; Vanderbeek et al., 2019). On the other hand, researchers proposed the combination of RCTs and RWD for clinical knowledge generation in the era of precision medicine (Agarwala et al., 2018). Historical controls leveraging RWD are of particular interest in oncology drug development to foster patients' access to innovative therapy in the context of segmentation of tumor entities and targeted therapy (Rahman et al., 2021). Learning from previous historical control applications can inform future studies and avoid common pitfalls. However, there have been few systematic discussions about their use in this field and especially the feedback of European regulators on it. Therefore, further research and collaboration are needed to establish frameworks for the use of historical control arms and RWE to improve the efficiency and effectiveness of oncology drug development.

The objective of this study is to perform a comprehensive analysis of historical controls in supporting clinical efficacy in oncology drug development and to gain a deeper understanding of the role of historical controls in regulatory decision-making.

- **Section 3.2** defines the search strategy employed to identify qualifying drug approvals.
- **Section 3.3** presents the outcomes of the identification process for the historical control cases.
- **Section 3.5** examines the statistical features of historical controls, their effect on regulatory decision-making, as well as the principal challenges and solutions for using historical control data to support clinical development.
- Finally, the study is concluded with a concise summary in **Section 3.6**.

3.2 Methodology

3.2.1 Search strategy and findings

To identify regulatory submissions that utilized historical control data to establish clinical efficacy of investigational treatments, we searched European public assessment reports (EPARs) for human medicines (https://www.ema.europa.eu/en/medicines). We included original marketing applications for cancer drugs granted between January 1, 2016, and December 31, 2021, and excluded drugs for diagnostic use only or those withdrawn from the market. The search was conducted in May 2022.

Of the 113 cancer drugs granted marketing authorization by the EMA between 2016 and 2021, we excluded two drugs for diagnostic use only and eight drugs that were withdrawn from the market by the time of searching. After screening the EPARs of the remaining 103 medicines, we identified 18 drug submissions (17%) that utilized historical control data to support clinical efficacy. Figure 3.1 illustrates the selection process we utilized.

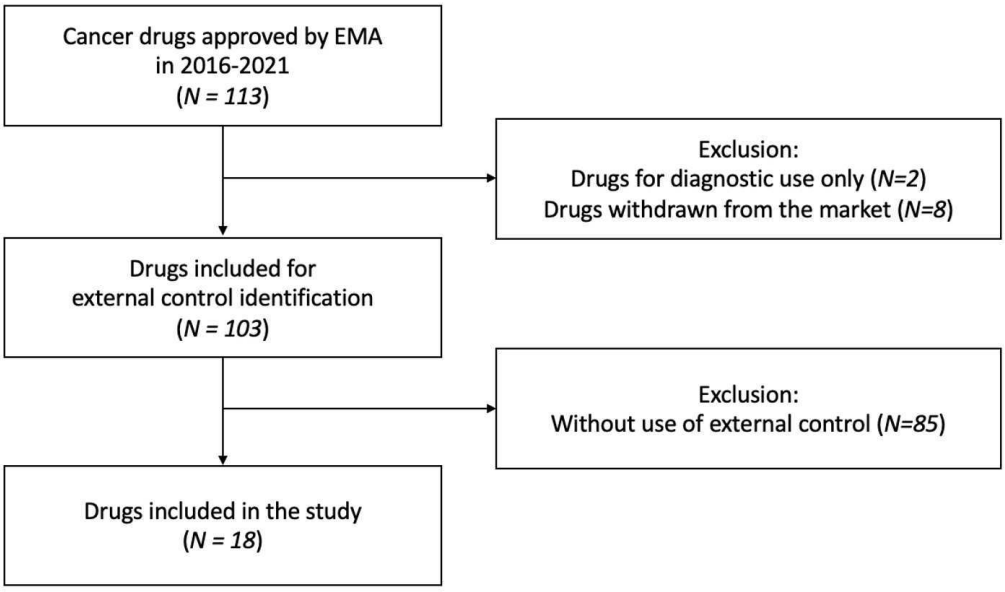


Figure 3.1: Selection of historical control cases from cancer drug approvals

To display a visual representation of the selection results, Figure 3.2 summarizes

the number of screened cancer drugs and eligible historical control cases each year.

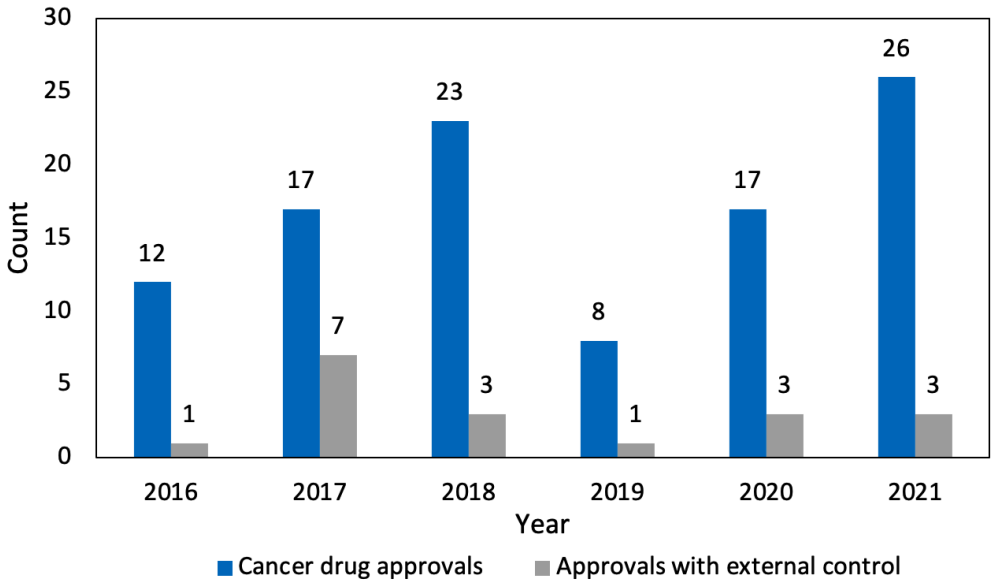


Figure 3.2: Summary of identification results in 2016-2021

Table 3.1 provides a comprehensive summary of 18 drug approvals that have been selected for further analysis. The table contains information including the drug name, drug class, approval year, therapeutic area, indication, and conditional marketing authorization.

Table 3.1: Selected cancer drug approvals using historical controls

Drug	Year	Class of drug	Therapeutic area	Conditional approval
Minjuvi (tafasitamab)	2021	Monoclonal antibody	Lymphoma	Yes
Abecma (idecabtagene vicleucel)	2021	Autologous cellular immunotherapy	Multiple Myeloma	Yes
Enhertu (trastuzumab deruxtecan)	2021	Antibody-drug conjugate	Breast Neoplasms	Yes

Table 3.1: Selected cancer drug approvals using historical controls *(continued)*

Drug	Year	Class of drug	Therapeutic area	Conditional approval
Blenrep (belantamab mafodotin)	2020	Antibody-drug conjugate	Multiple Myeloma	Yes
Rozlytrek (entrectinib)	2020	Kinase inhibitor	Non-Small-Cell Lung Cancer	Yes
Tecartus (brexucabtagene autoleucel)	2020	Autologous cellular immunotherapy	Lymphoma	Yes
Libtayo (cemiplimab)	2019	Monoclonal antibody	Squamous Cell Carcinoma	Yes
Apealea (paclitaxel)	2018	Mitotic inhibitor	Ovarian Neoplasms	No
Verzenios (abemaciclib)	2018	Kinase inhibitor	Breast Neoplasms	No
Bavencio (avelumabl)	2017	Monoclonal antibody	Neuroendocrine Tumors	No
Qarziba (dinutuximab beta)	2017	Monoclonal antibody	Neuroblastoma	No
Rydapt (midostaurin)	2017	Kinase inhibitor	Leukemia, Mastocytosis	No
Tecentriq (atezolizumab)	2017	Monoclonal antibody	Non-Small-Cell Lung Cancer; Urologic Neoplasms	No
Ledaga (chlormethine)	2017	Alkylating agent	Mycosis Fungoides	No
Blitzima (rituximab)	2017	Monoclonal antibody	Lymphoma, Leukemia	No

Table 3.1: Selected cancer drug approvals using historical controls *(continued)*

Drug	Year	Class of drug	Therapeutic area	Conditional approval
Truxima (rituximab)	2017	Monoclonal antibody	Lymphoma, Leukemia	No
Darzalex (daratumumab)	2016	Monoclonal antibody	Multiple Myeloma	No

3.2.2 Data Extraction

We reviewed the clinical efficacy section of the EPARs for the 18 drug approvals and extracted relevant information on pivotal studies and historical controls. Specifically, we analyzed the pivotal study design, use of historical controls, source of historical control data, method of analysis, and the EMA's decision regarding the use of historical controls.

3.3 Results

3.3.1 Characteristics of historical controls

Table 3.2 presents the characteristics of the identified historical controls. It should be noted that some of the drug submissions used multiple historical controls for a single pivotal study or conducted pivotal studies for multiple indications, resulting in a total of 24 historical controls being used across the 18 submissions.

Though all drugs were approved by EMA, some historical controls were not deemed supportive. We evaluated the EMA's decision regarding the use of historical controls and classified it as "accepted" if historical controls were considered as supportive evidence to demonstrate efficacy and "rejected" if they were deemed inadequate for decision-making.

Table 3.2: Historical controls used in cancer durg authorizations

Drug	Pivotal study design	Use of historical control	Source of historical control data	Method of analysis	EMA's decision
Minjuvi	Phase 2, single-arm, open-label, multicentre	Comparative efficacy analysis	RWD	Confounding adjustment	Rejected
Abecma	Phase 2, single-arm, open-label, multicentre	Comparative efficacy analysis	RWD	Confounding adjustment	Rejected
Enhertu - 1	Phase 2, single-arm, open-label, multicenter, 2-part	Comparative efficacy analysis	RWD	Confounding adjustment	Rejected
Enhertu - 2	Phase 2, single-arm, open-label, multicenter, 2-part	Understanding the natural history of disease	Published observational studies	Meta-analysis	Accepted
Blenrep	Phase 2, two-arm, randomized, open-label, multicentre	Historical benchmark	Published observational studies	Descriptive	Accepted
Rozlytrek	Phase 2, single-arm, open-label, multicenter, basket study	Comparative efficacy analysis	RWD	Confounding adjustment	Rejected

Table 3.2: Historical controls used in cancer durg authorizations *(continued)*

Drug	Pivotal study design	Use of historical control	Source of historical control data	Method of analysis	EMA's decision
Tecartus - 1	Phase 2, single-arm, open-label, multicentre	Historical benchmark	Published observational studies	Descriptive	Accepted
Tecartus - 2	Phase 2, single-arm, open-label, multicentre	Historical benchmark	Published observational studies	Meta-analysis	Rejected
Libtayo	Phase 2, single-arm, 3-group, multicenter	Comparative efficacy analysis	RWD	Descriptive	Accepted
Apealea	Phase 3, parallel group, randomised, comparator-controlled, open-label, non-inferiority study	Defining margin of non-interiority	Historical clinical trials	Meta-analysis	Accepted
Verzenios - 1	Phase 2, single-arm, open-label, multicentre	Historical benchmark	Unspecified	Descriptive	Accepted
Verzenios - 2	Phase 2, single-arm, open-label, multicentre	Comparative efficacy analysis	RWD	Confounding adjustment	Rejected

Table 3.2: Historical controls used in cancer durg authorizations *(continued)*

Drug	Pivotal study design	Use of historical control	Source of historical control data	Method of analysis	EMA's decision
Yescarta	Phase 2, single-arm, open-label, multicentre	Comparative efficacy analysis	RWD and historical clinical trials	Meta-analysis	Accepted
Bavencio	Phase 2, single-arm, open-label, multicentre	Understanding the natural history of disease	RWD	Descriptive	Accepted
Qarziba - 1	retrospective data analysis under a compassionate use program	Comparative efficacy analysis	RWD	Descriptive	Accepted
Qarziba - 2	retrospective data analysis under a compassionate use program	Comparative efficacy analysis	Historical clinical trials	Descriptive	Accepted
Rydapt	Phase 2, single-arm, multicentre	Comparative efficacy analysis	Historical clinical trials	Confounding adjustment	Rejected
Tecentriq - 1	Phase 2, single-arm, multicentre	Historical benchmark	Historical clinical trials	Descriptive	Rejected
Tecentriq - 2	Phase 2, single-arm, multicentre, two-cohort	Historical benchmark	RWD	Confounding adjustment	Rejected

Table 3.2: Historical controls used in cancer durg authorizations *(continued)*

Drug	Pivotal study design	Use of historical control	Source of historical control data	Method of analysis	EMA's decision
Ledaga	phase 2, multicenter, randomized, comparator-controlled, third party (observer) blinded, non-inferiority study	Defining margin of non-interiority	Unspecified	Descriptive	Accepted
Blitzima	Phase 1, randomized, controlled, multicentre, 2-arm, parallel-group, double-blind	Comparative efficacy analysis	Historical clinical trials	Descriptive	Accepted
Truxima - 1	Phase 1, randomized, controlled, multicentre, 2-arm, parallel-group, double-	Comparative efficacy analysis	Historical clinical trials	Descriptive	Accepted
Truxima - 2	open-label, single-arm, maintenance study	Defining margin of non-interiority	Historical clinical trials	Descriptive	Accepted

Table 3.2: Historical controls used in cancer durg authorizations *(continued)*

Drug	Pivotal study design	Use of historical control	Source of historical control data	Method of analysis	EMA's decision
Darzalex	Phase 2, open-label, multicentre, 2-arm	Historical benchmark	Published observational studies	Descriptive	Accepted

The characteristics of historical controls in terms of their use for clinical efficacy, data source, analysis method, and EMA's decision are summarized in Table 3.3, Table 3.4, Table 3.5 and Table 3.6 respectively.

Table 3.3: Summary of historical control characteristics (Use of historical control)

Use of historical control	N	%
Comparative efficacy analysis	12	50.0
Defining margin of non-interiority	3	12.5
Historical benchmark	7	29.2
Understanding the natural history of disease	2	8.3

Table 3.4: Summary of historical control characteristics (Source of historical control data)

Source of historical control data	N	%
Historical clinical trials	7	29.2
Published observational studies	5	20.8
RWD	9	37.5
RWD and historical clinical trials	1	4.2
Unspecified	2	8.3

Table 3.5: Summary of historical control characteristics (Method of analysis)

Method of analysis	N	%
Confounding adjustment	7	29.2
Descriptive	13	54.2
Meta-analysis	4	16.7

Table 3.6: Summary of historical control characteristics (EMA's decision)

EMA's decision	N	%
Accepted	15	62.5
Rejected	9	37.5

3.3.2 EMA's decision on historical controls

The decisions made by the EMA regarding the use of historical controls are presented in Table 3.7 and Table 3.8, respectively, categorized based on data sources and analysis methods.

Table 3.7: EMA's decision on historical controls by source of historical control data

Source of external control data	Accepted	Rejected
Real-world data (RWD)	3 (33%)	6 (67%)
Historical clinical trials	5 (71%)	2 (29%)
RWD and historical clinical trials	1 (100%)	0 (0%)
Published observational studies	4 (80%)	1 (20%)
Unspecified	2 (100%)	0 (0%)

Table 3.8: EMA's decision on historical controls by method of analysis

Method of analysis	Accepted	Rejected
Descriptive	12 (92%)	1 (8%)

Table 3.8: EMA's decision on historical controls by method of analysis *(continued)*

Method of analysis	Accepted	Rejected
Confounding adjustment	0 (0%)	7 (100%)
Meta-analysis	3 (75%)	1 (25%)

Table 3.9 displays the limitations recognized by EMA that led to the rejection of historical controls as a valid form of evidence for establishing clinical efficacy.

Table 3.9: Limitations identified by EMA on historical controls

Drug	Source of external control data	Method of analysis	Limitations identified by EMA
Minjuvi	RWD	Confounding adjustment	Heterogeneous patient populations, differences in standard of care received during treatment, suboptimal statistical methodology
Abecma	RWD	Confounding adjustment	Selection bias of the study population, missing data of prognostic factors
Enhertu - 1	RWD	Confounding adjustment	Selection bias of the study population, missing assessment of response, differences in the measurement of endpoint, not optimal statistical methodology
Rozlytrek	RWD	Confounding adjustment	Limitations of the study design, limited data
Tecartus - 2	Published observational studies	Meta-analysis	Heterogeneous patient populations, limited information on the study design
Verzenios - 2	RWD	Confounding adjustment	Heterogeneous patient populations

Table 3.9: Limitations identified by EMA on historical controls *(continued)*

Drug	Source of external control data	Method of analysis	Limitations identified by EMA
Rydapt	Historical clinical trials	Confounding adjustment	Limited information on the baseline characteristics, no correction for the time of initiation of treatment
Tecentriq - 1	Historical clinical trials	Descriptive	Limited information on the determination of historical response rates
Tecentriq - 2	RWD	Confounding adjustment	Heterogeneous patient populations

3.4 Examining selected cases

In order to provide a comprehensive evaluation of historical control arms as assessed by regulatory authorities, we examined two selected cases that involved different data sources, analysis methods, and outcomes during the EMA's evaluation process. The first case study focused on Enhertu and included two historical controls. The EMA rejected the historical control based on real-world data and propensity score matching analysis, while accepting the historical control that utilized meta-analysis based on literature. The second case study involved Minjuvi and included one historical control using observational data and propensity score matching analysis, which was ultimately rejected by the EMA. Each case study provided details on the indication, pivotal study conducted, historical controls submitted, and relevant comments from EMA reviewers. All information was extracted from the European public assessment reports for initial marketing authorization.

3.4.1 Enhertu (trastuzumab deruxtecan)

Enhertu is an antibody-drug conjugate and as monotherapy is indicated for the treatment of adult patients with unresectable or metastatic HER2 positive breast cancer who have received two or more prior anti-HER2 based regimens. The EMA granted Enhertu a conditional marketing authorization (CMA) in 2021.

Pivotal study in clinical efficacy

The assessment of trastuzumab deruxtecan's efficacy primarily relies on the pivotal study U201, which is a phase 2, two-part, open-label, single-arm cohort study. The EMA accepted the single-arm, open-label design of the pivotal study in the context of the conditional marketing authorization. The trial's primary endpoint was the objective response rate (ORR). Secondary endpoints included the duration of response (DoR), progression-free survival (PFS), and overall survival (OS). A total of 184 HER2-positive breast cancer patients were treated with the recommended dose of 5.4 mg/kg. In this treatment group, the ORR was 60.3% (95% CI: 52.9, 67.5), and the median DoR, median PFS, and median OS were not reached.

Historical controls

Two historical controls were included as supportive studies to provide context for the results of the pivotal study: the Unicancer study and a literature-based study.

Historical control 1: Unicancer study

The Unicancer study utilized a French real-world database containing approximately 60,000 patients, of which 19,867 were treated for metastatic breast cancer at 18 cancer centers. From this database, two cohorts were created: the Reference Cohort and the Matched Cohort.

The Reference Cohort consisted of 721 patients with metastatic HER2-positive breast cancer who received treatment between January 2008 and December 2016, with at least one therapy after TDM1 as of September 17, 2018.

To generate the Matched Cohort, patients from the Reference Cohort were propensity-score matched to subjects from Study U201 based on similar baseline characteristics, such as prior treatment with pertuzumab, HR status, presence of

visceral disease, and number of previous treatment lines. The matching was done at a 1:1 ratio without replacement, but due to a limited number of patients within the caliper, not all subjects from Study U201 could be matched. Ultimately, 137 patients from the Reference Cohort were matched with 137 subjects from Study U201.

The primary objective of both the Reference Cohort and Matched Cohort was to describe patient characteristics and clinical features, and a secondary objective was to describe the treatment strategies for these patients. In the Matched Cohort, the ORR was 12.2% (95% CI: 6.2, 18.2), the median PFS for patients was 4.7 months (95% CI: 3.8, 6.0), and the median OS evaluated using the start date of the line of therapy after TDM1 as the reference date was 24.1 months (95% CI: 18.5, 26.4).

Historical control 2: Literature based meta-analysis

A literature-based analysis was conducted to understand the historical context of expected response rates of trastuzumab deruxtecan in the post-trastuzumab setting. The analysis had two main objectives: to review literature on trials involving patients with advanced or metastatic breast cancer who had previously received trastuzumab and chemotherapy, and to perform a model-based meta-analysis to determine the objective response rate, median progression-free survival time, and median overall survival time in this population.

Based on the study's inclusion and exclusion criteria, a total of 8,827 subjects from 37 studies were included in the analysis of ORR, median PFS, and median OS. The results indicated that the overall mean ORR was estimated to be 25.5% (95% prediction range: 17.1, 36.1), and the median PFS was estimated to be 5.8 months (95% prediction range: 3.2, 10.5). In the subgroup of studies that included patients with a median of at least 2 prior chemotherapies or at least 2 prior anti-HER2-based regimens, the median ORR was estimated to be 15% (95% CI: 9, 30), and the median PFS was estimated to be 4.8 months (95% CI: 3.3, 5.5).

EMA review

The analysis of historical data from the real-world database Unicancer and the literature revealed lower efficacy compared to the results observed in the pivotal study U201. However, the EMA acknowledged that the real-world data should be considered exploratory due to several uncertainties. These uncertainties encompass non-optimal matching, which introduces uncertainty in the comparability of patient populations,

as well as the absence of response assessment in 16% of the matched subjects. Furthermore, cohort selection based on post-baseline variables may introduce selection bias, and there are differences in timing and measurement of endpoints (ORR and PFS) between the historical cohort and the pivotal study.

Nevertheless, despite these uncertainties, when comparing the literature data of currently used anti-HER2 regimens in the same setting and target population, trastuzumab deruxtecan still demonstrates a significant therapeutic advantage in terms of efficacy. For instance, the median duration of response (DoR) of 20.8 months is approximately three times longer than the reported median DoR (6.0-8.5 months) or median PFS (4.9-7.8 months). This difference was considered significant enough to outweigh the uncertainties related to the lack of an active comparator arm and allow for indirect comparisons, especially within the context of the requested conditional marketing authorization (European Medicines Agency (EMA), 2020).

3.4.2 Minjuvi (tafasitamab)

Minjuvi is a cancer medication primarily used in combination with lenalidomide (LEN) and subsequently as a monotherapy to treat adults diagnosed with diffuse large B-cell lymphoma (DLBCL) that has relapsed or become unresponsive to other treatments and who are ineligible for autologous stem cell transplantation. The EMA granted Minjuvi a conditional marketing authorization in 2021. Given the rarity of DLBCL, Minjuvi was designated by the EMA an 'orphan medicine' in 2015.

Pivotal study in clinical efficacy

The evaluation of tafasitamab's efficacy primarily relies on the pivotal phase 2 study MOR00208 (L-MIND), which was a multicenter trial with an exploratory hypothesis. It was designed as a single-arm, open-label study. The primary endpoint of the trial was the objective response rate (ORR), while secondary endpoints included the duration of response (DoR), progression-free survival (PFS), and overall survival (OS). A total of 81 patients were enrolled in the trial, with one patient receiving tafasitamab monotherapy. The observed ORR was 56.8% (95% CI: 45.3, 67.8), the median DoR was 34.6 months (95% CI: 26.1, NR), the median PFS was 12.1 months (95% CI: 5.7, NR) and the median OS was 31.6 months (95% CI: 18.3, NR).

Historical control

Among the submitted supportive studies, Study MOR208C206 (RE-MIND) is a retrospective observational study whose objectives were to characterize the effectiveness of LEN monotherapy in the treatment of R/R DLBCL patients and to compare the effectiveness of LEN monotherapy with the efficacy outcomes with tafasitamab-LEN combination therapy.

The RE-MIND study included a multicenter cohort of patients and used similar efficacy endpoints as the pivotal study L-MIND. The eligibility criteria for the observational cohort in the RE-MIND study were the same as those in the L-MIND study, including histological subtypes, number of prior therapy lines, prior therapy types allowed, and ineligibility for autologous stem cell transplantation (ASCT). However, unlike L-MIND study, the RE-MIND study included primary refractory patients and did not require a specific ECOG score at baseline. Additionally, certain laboratory values were not pre-specified in the LEN monotherapy study.

To compare the efficacy between the two treatment cohorts, patients were matched using a propensity score approach based on several variables, including age, Ann Arbor stage, refractoriness to last therapy line, number of prior lines of therapy, history of primary refractoriness, prior ASCT, elevated lactate dehydrogenase (LDH), neutropenia, and anemia.

A total of 76 patients were matched (1:1) to the L-MIND patients. Some baseline characteristics, such as age and refractoriness to the last prior therapy, were similar between the groups. However, there were notable differences in other factors which are of vital importance for the response and prognosis in DLBCL. For estimates of efficacy, patients from the two treatment cohorts were considered as independent sets. As such, analyses for unpaired data were conducted. However, matched data are not independent and analysis methods for paired (correlated) data were also included as sensitivity analyses.

A comparison of endpoints was conducted between the tafasitamab+LEN (L-MIND) and LEN-mono (RE-MIND) groups. The best ORR was 51 patients (95% CI: 55.4, 77.5) for L-MIND and 26 patients (95% CI: 23.7, 46.0) for RE-MIND. The median DoR was 20.5 months (95% CI: 3.3, 13.9) for L-MIND and 4.1 months (95% CI: 1.5, 5.2) for RE-MIND. The median PFS was 12.1 months (95% CI: 5.9, NR) for L-MIND and 4.0 months (95% CI: 3.1, 7.4) for RE-MIND. The median OS estimated

using the Kaplan-Meier method was not reached (95% CI: 15.5, NR) for L-MIND and 9.4 months (95% CI: 5.1, 20.0) for RE-MIND.

EMA review

In the single-arm study of tafasitamab + LEN, it was considered difficult to isolate the treatment effect of tafasitamab. The retrospective, observational study RE-MIND study was submitted as a historical comparator to contextualise the results of the pivotal study. However, due to the heterogeneity in the study populations in the tafasitamab+LEN (L-MIND) and LEN-mono (RE-MIND) groups, uncertainties in the matching analysis, and differences in standard of care received during treatment, the interpretation of the results is limited. Therefore, the findings from the RE-MIND study can only be regarded as exploratory (EMA, 2021).

3.5 Discussion

3.5.1 Use of historical control

We identified that the majority of historical controls (50%) were utilized for comparative efficacy analysis, while 29% served as historical benchmarks in the superiority study design. Additionally, a small percentage of historical controls (8%) provided contextual information on the natural history of disease in rare indications. Furthermore, some historical controls (13%) were used to establish the margin of non-inferiority for the non-inferiority hypothesis. All identified approvals using historical controls were conducted in single-arm trials. No historical controls supplementing the concurrent control arm in RCTs were found. Although there have been discussions about combining external with internal data in RCTs (Gray et al., 2020; Schmidli et al., 2020; Xu et al., 2020), the hybrid control arm has not been used for EMA submissions. Nonetheless, this design enables the assessment of population comparability and may be considered in future historical control studies.

3.5.2 Source of historical control data

The study discovered that historical controls primarily came from individual RWD (38%) and historical clinical trials (29%) for comparable disease settings. In some cases, the historical controls were based on published studies (21%), which included both prospective and retrospective observational studies. A small fraction of cases (8%) did not provide information about the source of the historical control data. Only one case (4%) utilized both individual RWD and historical clinical trial data. These findings indicated that RWD was commonly used in historical controls for cancer drug approvals in recent times.

3.5.3 Method of analysis

Our findings indicate that 54% of the cases relied on descriptive analysis and simple comparison. Furthermore, 17% of the cases utilized meta-analysis with either aggregate or patient-level data. In addition, 29% of the submissions incorporated comparative efficacy analysis, which accounted for confounding covariates by utilizing matching and inverse probability weighting techniques based on propensity score or Mahalanobis distance, using individual patient data from RWD or historical clinical trials.

3.5.4 Regulators' feedback on historical controls

The EMA accepted 63% of the historical controls to shed light on the treatment effect. However, Table 3 shows that only a tiny part of historical controls using RWD (33%) was considered supportive evidence. By contrast, the majority of historical controls using data from historical clinical trials (71%), both RWD and historical clinical trials (100%), published studies (80%), or unknown sources (100%) received positive reviews. Regarding the analysis method, the EMA accepted most historical controls using descriptive analysis (92%) or meta-analysis (75%). In comparison, no case applying methods of confounding adjustment (0%) was considered supportive evidence. These results indicate that the EMA primarily considered the main study results in decision-making, and various data sources or analysis methods did not mitigate the concerns about the systematic bias of historical controls.

The limitations of the historical controls cited by EMA (Table 3.9) primarily con-

cerned the Pocock criteria (Pocock, 1976). Heterogeneity in the external and internal patients was a significant issue. It was caused by differences in standard of care or limited information from the sponsors on the baseline characteristics, especially on significant prognostic factors. The regulators also pointed out that missing or different assessment of endpoints was an essential downside of using RWD in historical controls. Besides, unclear study design and selection bias hindered regulators' understanding and thus their acceptance of the historical controls. Unluckily, these biases were hardly compensated through statistical methodology as the regulators deemed the sponsor applied suboptimal statistical analysis.

3.5.5 Recommendations on historical controls for decision-making

Based on the results of our study, we propose several relevant solutions to enhance the utilization of historical controls in regulatory decision-making.

- Firstly, we recommend a priori trial planning that takes historical controls into account. Sponsors should engage in early discussions with regulatory authorities to align on study design, thus minimizing selection bias and avoiding cherry-picking of data.
- Secondly, we suggest evaluating external data sources using Pocock's criteria to ensure comparability of study populations and reliable assessment of outcomes.
- Thirdly, we advise appropriate statistical analysis with adjustment for confounders to reduce potential bias. we advise applying appropriate statistical analysis techniques that account for confounding factors, thereby reducing potential biases. For frequentist methodology, propensity score analysis can be applied with step-by-step clarification. The doubly robust estimation method can be added for propensity score methods, as propensity score estimation is susceptible to model misspecification (Funk et al., 2011; King & Nielsen, 2019). Additionally, leveraging individual RWD to emulate target trials can ensure the availability of fit-for-purpose data and effective control of confounding factors (Hernán et al., 2022).
- Lastly, we encourage collaboration between sponsors and regulators to develop regulatory guidance on historical controls, covering study design, data selection, and analysis plans.

3.5.6 Limitations of the study

Our study was conducted in a systematic manner with in-depth extractions of historical control characteristics and objective discussions on the issues of historical controls. Despite this, our study was subject to several limitations. First, we limited our scope with a prespecified inclusion period of 2016-2021 under the presumption that historical control using RWD is a recently emerging research interest in oncology. Second, our data extraction was limited to the EMA's public assessment reports rather than the sponsors' original submission documents. As a result, some details on historical controls may have been omitted. Finally, we did not scrutinize the impact of historical control on regulatory pathways, such as conditional marketing authorization. Besides, the prespecified period of 2016-2021 may limit the assessment of post-authorization changes. It can be interesting for future work to study the use of historical controls for diverse regulatory pathways.

3.6 Conclusion

There has been growing research interest in leveraging historical controls to facilitate oncology clinical trials. Medical regulatory agencies have begun to endorse the use of RWD and historical controls and published relevant frameworks. However, there are few systematic discussions about the current applications of historical controls in oncology drug development and the relevant regulatory feedback. In this study, we identified and summarized the characteristics of European oncology drug approvals using historical control data in clinical efficacy. Eighteen eligible submissions leveraging 24 historical controls were included. We discussed the use of historical control, data sources, methods of comparison, and the regulators' feedback.

We found that the historical controls had been actively submitted to the EMA and increasingly used RWD and advanced statistical analysis methods, showing the potential for informing future clinical development. However, some historical controls were not deemed supportive evidence by EMA due to significant limitations regarding heterogeneous patient populations, missing RWD of outcome assessment, and suboptimal study design. This study highlighted that the proper use of historical controls in oncology clinical trials requires carefully assessing data availability and determining the optimal clinical and statistical methodology. For better use of historical controls in oncology clinical development, we suggest a priori study design to avoid selection

bias, sufficient baseline data to ensure the comparability of study populations, consistent endpoint measurements to enable outcome comparison, optimal application of statistical methods for comparative analysis, and collaborative efforts of sponsors and regulators to establish frameworks on historical control arms.

Chapter 4

Historical control arms with clinical data

Contents

4.1 Introduction

The establishment of reliable historical control arms relies heavily on the availability and quality of data (Burger et al., 2021). In this context, two primary sources of data are pivotal: clinical data and observational data. Each data source has unique methodological considerations that must be carefully addressed during the research design process. This chapter will primarily focus on methodological considerations pertaining to clinical data, while the subsequent chapter will delve into observational data.

Using data from previous clinical trials as sources for historical control arms presents several advantages. The most notable advantage is the superior data quality typically associated with rigorous monitoring and compliance to the study protocols. The baseline characteristics of patients, treatment regimens, and outcome measures are defined prospectively, ensuring standardized data collection. Consequently, clinical trials tend to have limited missing data and better data completeness (Franklin & Schneeweiss, 2017).

Nevertheless, clinical trial data can be limited by relatively smaller sample sizes compared to observational databases. Particularly in the realm of diseases with high unmet medical needs, late-stage clinical trials frequently involve a limited cohort size, typically ranging from a few hundred participants to less than a hundred. This could potentially limit the statistical power and generalizability of the derived historical control arm.

Clinical trial data has long been a subject of statistical research, particularly in the context of model-based meta-analysis where aggregated data from multiple trials are analyzed (Chan et al., 2022). When considering historical control arms, aggregated data can serve as an external comparator providing contextual information for the new trial. However, utilizing individual-level data can offer additional insights into the baseline characteristics of patients. Consequently, our study aims to leverage individual-level data from previous clinical trials to construct a historical control arm within a single-arm trial.

The primary objectives of this study are: (1) to build a historical control arm by employing data from previous clinical trials while appropriately addressing between-trial heterogeneity; (2) to assess the effectiveness of propensity score analysis in mitigating bias when utilizing historical control arms.

By achieving these objectives, we aim to study the validity and reliability of utilizing historical control arms in clinical research while accounting for potential confounding factors and sources of variability across trials.

This chapter is structured as follows:

- **Section 4.2** introduces the motivating trial and discusses the selection of historical control data utilized in our case study.
- **Section 4.3** presents the propensity score methodology used to reduce the bias in leveraging historical control arms.
- **Section 4.4** outlines the results derived from our analysis.
- Lastly, in **Section 4.5**, we delve into a detailed discussion of our study.

4.2 Motivating data

4.2.1 Motivating trial: the EPSB study

Acute post-surgical pain (PSP) following breast cancer surgery affects approximately 70% of patients, impacting their quality of life and increasing the risk of developing chronic PSP (Gärtner et al., 2009). Therefore, devising effective pain mitigation strategies is crucial for these patients. Recent data show the effectiveness of regional anesthesia (RA) within multimodal analgesia programs in reducing acute PSP and postoperative opioid usage. Thoracic paravertebral block (TPVB), extensively researched and recognized as the standard for extensive breast cancer surgery, has demonstrated safety under ultrasound guidance (Albi-Feldzer et al., 2021). Alternatively, the erector spinae plane block (ESPB), introduced in 2016, uses a superficial interfacial block in the erector spine muscle plane, distinguishing it from the paravertebral space (Forero et al., 2016). ESPB offers procedural simplicity, speed, and safer injection sites due to more distal anatomical landmarks compared to TPVB. Despite its proven efficiency in preventing acute PSP in spinal and thoracic surgeries, the reliability of this diffusion process remains controversial, especially in the context of breast surgery, where the efficacy of ESPB remains unproven.

In response to this uncertainty, a prospective observational study was conducted to evaluate the safety and efficacy of ESPB with ropivacaine in reducing morphine consumption in the post-anesthesia care unit (PACU) following major breast cancer

surgery. Conducted by the Department of Anesthesiology at Institut Curie, this observational, single-arm trial enrolled 120 patients between December 2018 and August 2019, and included 102 patients in the analysis. The primary outcome was the proportion of patients requiring morphine titration in the PACU. Key secondary outcomes included the total morphine dose in the PACU and the incidence of RA complications. The following data was collected during the study: age, weight and height, the type of surgery, the injected volume and dose of ropivacaine, the occurrence of ESPB-related complications, the rest and mobilization VAS measured during the hospital stay and 24 hours after surgery, the need for rescue morphine, and the overall morphine dose (mg) administered in the PACU.

Existing studies comparing ESPB to TPVB have been limited by their small patient numbers, and their results pooled in multiple meta-analyses have presented conflicting conclusions (Huang et al., 2020; Leong et al., 2021; Xiong et al., 2021). Therefore, the role of ESPB in analgesia strategies for breast surgery is not clearly defined, necessitating a robust comparison of ESPB with TPVB. In this context, we propose to construct a historical control arm for the ESPB trial by leveraging clinical data from previous TPVB trials.

4.2.2 Historical data: the MIRs03 study

The MIRs03 study (NCT02408393) was a large, prospective, multicenter, 1:1 randomized, double-blind, placebo-controlled study that compared the efficacy of TPVB with ropivacaine to a thoracic paravertebral injection of saline in preventing acute and chronic PSP (Albi-Feldzer et al., 2013). Conducted by the Department of Anesthesiology at Institut Curie between March 2015 and June 2018, the study randomized 352 patients undergoing partial or complete mastectomy with or without lymph node dissection to receive a preoperative paravertebral block with either ropivacaine (178 patients in the experimental group) or saline (174 patients in the control group). The procedure of the study is shown in Figure 4.1.

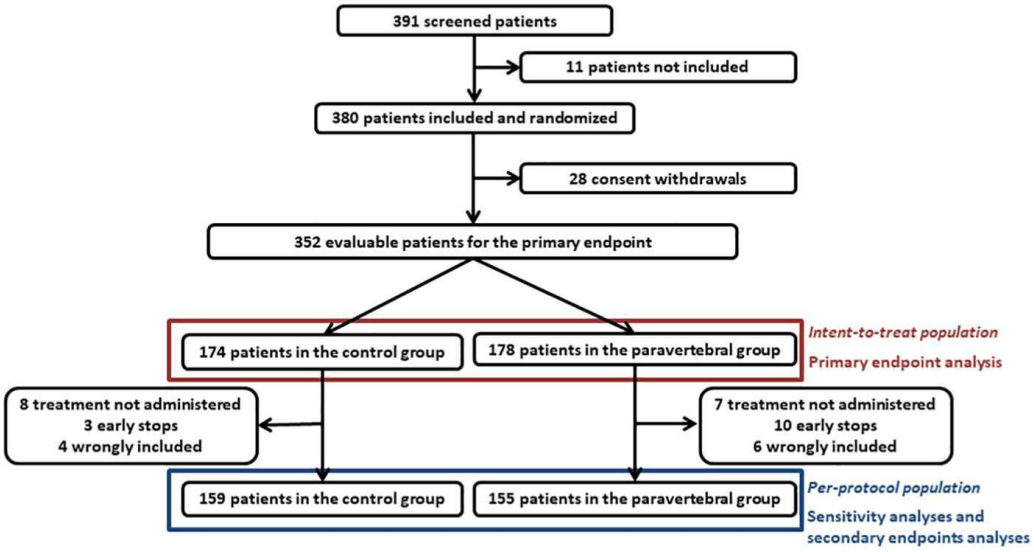


Figure 4.1: Flow chart of the MIRS03 study

The primary outcome was the incidence of chronic pain 3 months after breast cancer surgery, which was reported in 93 of 178 (52.2%) and 83 of 174 (47.7%) patients in the paravertebral and control groups, respectively (odds ratio, 1.20 [95% CI: 0.79-1.82], P=0.394). The secondary outcomes were acute pain, analgesic consumption, nausea and vomiting, chronic pain at 6 and 12 months, neuropathic pain, pain interference, anxiety, and depression.

The general information of the ESPB and MIRs03 studies are summarized in Table 4.1. We plan to use the data from the experimental group of the MIRS03 study to establish a TPVB cohort, serving as the historical comparator for the ESPB study.

Table 4.1: Summary of the motivating trials

Study	ESPB study	MIRs03 study
Design	Prospective observational, single-arm	Placebo-controlled randomized, two-arm
Injection Technique	Erector spinae plane nerve block (ESPB)	Thoracic paravertebral block (TPVB)
Outcome of interest	Incidence of morphine titration after breast cancer surgery	Incidence of morphine titration after breast cancer surgery

Table 4.1: Summary of the motivating trials *(continued)*

Study	ESPB study	MIRs03 study
Primary completion date	2019	2018
Enrollment	120	380

4.3 Methodology

4.3.1 Evaluation of trial comparability

The first step of incorporating historical control arm is to evaluate the comparability of the trials. Differences in patient populations or other trial-specific circumstances can lead to bias when using a historical control arm. Therefore, the possible heterogeneity between the trials should be considered in the analysis.

We apply the Pocock criteria (Pocock, 1976) to evaluate the comparability of the ESPB and MIRs03 studies:

1: *The historical controls must have received a precisely defined standard treatment which must be the same as the treatment for the randomized controls.* In both trials, the experimental arms were administered Ropivacaine, albeit with distinct injection techniques.

2: *The historical controls must have been part of a recent clinical study which contained the same requirements for patient eligibility.* The ESPB and MIRs03 studies were conducted in 2019 and 2018, respectively. Based on the insights of clinical investigators, we assume that there have been no significant advancements in pain management and evaluation within this one-year span.

3: *The methods of treatment evaluation must be the same.* Both studies measured the fraction of patients necessitating morphine titration in the PACU post-breast cancer surgery.

4: *The distributions of important patient characteristics in the historical controls should be comparable with those in the new trial.* We focus on three salient patient attributes, which are prognostic indicators for acute postoperative pain: age, BMI

(Body Mass Index), and the type of breast surgery. The characteristics are compared using statistical tests and calculation of standardized mean differences (SMDs) As shown in Table 4.2, there are significant differences between trials in the distributions of breast surgery type ($p < 0.001$ and $SMD > 0.1$) and BMI ($SMD > 0.1$).

5: *The previous studies must have been performed in the same organization with largely the same clinical investigators.* Both trials were initiated by the same principal investigator and sponsored by Institut Curie.

6: *There must be no other indications leading one to expect differing results between the randomized and historical controls.* As discussed in criterion 2, we maintain that pain management and evaluation methodologies remained static during the trials' execution. To the best of our understanding, there are no other indications for substantial differences between the trials.

Table 4.2: Baseline characteristics of patients between ESPB and MIRs03 trials

Baseline characteristic	ESPB (n=102)	TPVB (n=165)	p-value	SMD
Age (years), mean (SD)	56.5 (12.9)	57.3 (14.1)	0.65	0.057
Weight (kg), mean (SD)	67.20 (13.83)	68.85 (13.87)	0.35	0.119
BMI (kg/m^2), mean (SD)	25.2 (5.0)	25.8 (5.0)	0.34	0.122
Surgery, n (%)			<0.001	0.466
Mastectomy	74 (72.5)	143 (86.7)		
Tumorectomy	20 (19.6)	22 (13.3)		
Axillary lymph node dissection	8 (7.8)	0 (0.0)		

The evaluation of comparability of the ESPB and MIRs03 trials using Pocock criteria is summarized in Table 4.3. While most criteria are satisfied, the one concerning the distribution of pivotal patient characteristics stands as an exception. Given our accessibility to the individual patient data from both trials, we opt to use propensity score analysis to adjust for the heterogeneity observed in the trials' patient characteristics.

Table 4.3: Summary of comparability evaluation of ESPB and MIRs03 trials

Criterion	Evaluation
1. Same treatment	Yes
2. Same eligibility criteria	Yes
3. Same outcome evaluation	Yes
4. Same organization and investigators	Yes
5. Recent trial	Yes
6. Comparable patient characteristics	No

4.3.2 Selection of confounders

The propensity score is defined as the probability of treatment selection conditional on measured baseline covariates. Before estimating the propensity scores, it's essential to select the covariates to incorporate into the estimation. Based on the practical principles of confounder selection (VanderWeele, 2019), we adjust for any covariate that is a cause of the exposure or of the outcome.

In this study, we assume that no confounder is a cause of the exposure to the treatment because the patients in the ESPB and MIRs03 trials were enrolled separately with the same eligibility criteria. Thus, we only focus on the confounders related to the outcome, specifically the significant prognostic factors for acute postoperative pain. According to medical expertise and published evidence on predictors of pain after breast cancer surgery (Wang et al., 2016), we include age, BMI and breast surgery type as listed in Table 4.2. The causal directed acyclic graph of the treatment (ESPB vs. TPVB), the outcome (acute pain after breast cancer surgery) and the confounders is depicted in Figure 4.2.

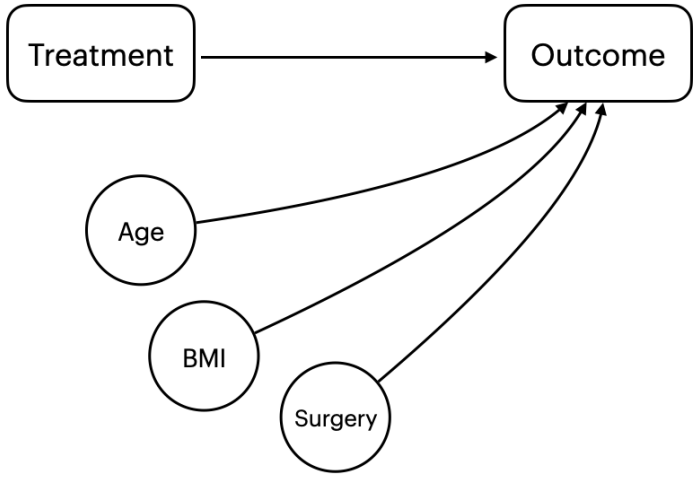


Figure 4.2: Directed acyclic graph of ESPB-MIRs03 study

Note that the selection of confounders should not be based solely on analyzing the correlation between the variables and the outcome using the available data (Austin & Stuart, 2015). During the propensity score estimation, the outcome data ought to remain obscured and should only be accessible at the subsequent stage of outcome analysis.

4.3.3 Adjustment for confounders with propensity score matching

Propensity score estimation

To balance the patient characteristics between the ESPB and TPVB cohorts, we conducted a propensity score matching analysis. The propensity scores were estimated by a multivariable logistic regression model in which the probability of receiving the intervention (ESPB vs. TPVB) was regressed conditional on age, BMI, and breast surgery type.

Matching algorithms

Patients were matched on the logit of the propensity score using a calliper of width equal to 0.2 of the standard deviation of the logit of the estimated propensity scores.

Matching was performed without replacement (i.e., each subject was available for matching only once) in a greedy manner (i.e., at each step in the matching process, the nearest TPVB subject was selected for matching to the given ESPB subject).

Balance diagnostics

Balance diagnostics is a key step for assessing the specification of the propensity score. The balance of covariates between the two arms was checked using standardized mean differences (SMDs) before and after matching. A standardized mean difference of less than 0.1 is considered to indicate a negligible difference in the mean or prevalence of a covariate between groups.

Treatment effect estimation

The risk difference in acute PSP with a 95% confidence interval was estimated as the difference between the probability of receiving morphine titration of TPVB patients and that of ESPB patients in the matched sample. The standard errors were estimated using cluster-robust standard errors to account for pair membership.

Comparisons of patient characteristics of ESPB placement details according to the need for morphine titration were performed with the Mann–Whitney or Student's t test after testing for normality with the Shapiro–Wilk test.

Sensitivity analysis

Estimating the propensity score using logistic regression has the risk of model misspecification. Thus, we built a random forest model in addition to the logistic regression model to assess the sensitivity of the matching result to the propensity score estimation. The random forest model was constructed using the intervention (ESPB vs. TPVB) as the output and the baseline characteristics as inputs. Matching was performed with the same parameters as described above. Note that the true propensity scores are unknown and it is not possible to evaluate directly the model specification. We assess the robustness of the results through balance diagnostics of the confounders.

All tests were two-sided. A P value of less than 0.05 was considered to indicate statistical significance. All analyses were conducted using R statistical software, version

4.0.2 (R Foundation for Statistical Computing, Vienna, Austria).

4.4 Results

4.4.1 Main results

Data collection

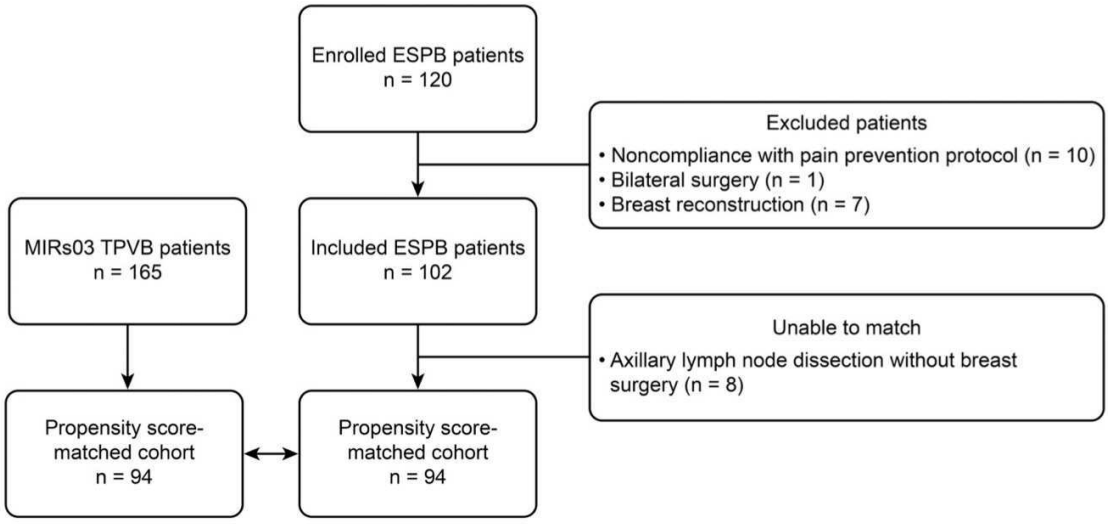


Figure 4.3: Flow chart of the ESPB study

In the ESPB cohort, 120 patients were enrolled between December 2018 and August 2019, and 102 patients were included in the analysis. The reasons for secondary exclusion of the 18 patients are detailed in Figure 4.3. Data from all 178 patients in the experimental arm were employed, of which 13 patients were excluded because of missing data on morphine consumption.

The study sample consisted of 102 ESPB patients and 165 TPVB patients. As shown in Table 4.2, there were statistically significant differences in baseline characteristics regarding breast surgery type (ESPB patients underwent more total mastectomies, $p<0.001$ and SMD > 0.1) and BMI (ESPB patients had lower BMI, SMD > 0.1).

Propensity score matching

The propensity scores were estimated by a multivariable logistic regression model. The standard deviation of the logit of the propensity score is equal to 0.1251. Thus, the caliper is set to be 0.2 of this width and is equal to 0.025.

Figure 4.4 presents the estimated propensity scores using a logistic regression model. The overlapping area of propensity scores of the two groups implies that there are patients who share similar propensity scores and can thus be considered matched pairs.

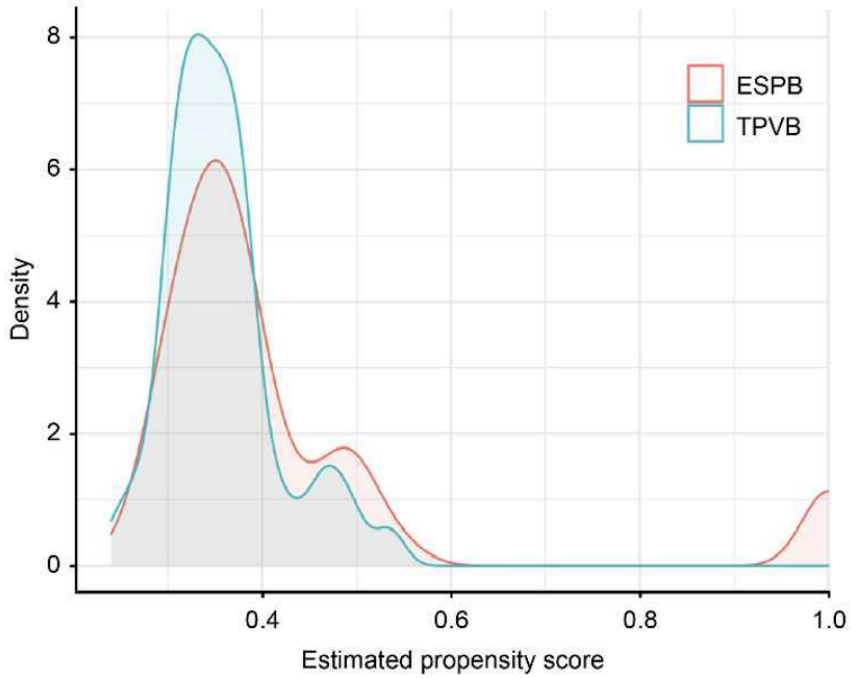


Figure 4.4: Distribution of the estimated propensity scores using a logistic regression model

Propensity score matching formed 94 matched pairs, which means that 94 of 102 ESPB patients were matched with a TPVB patient. Eight ESPB patients who received axillary lymph node dissection were thus excluded from further analysis.

Balance diagnostics

The baseline characteristics of ESPB and TPVB patients in the propensity score-matched sample are described in Table 4.4. The mean and prevalence of continuous

and categorical variables were very similar between the two groups (all SMDs < 0.1).

Table 4.4: Baseline characteristics of patients after propensity score matching

Baseline characteristic	ESPB (n=94)	TPVB (n=94)	p-value	SMD
Age (years), mean (SD)	56.0 (12.7)	55.2 (14.3)	0.72	0.053
Weight (kg), mean (SD)	67.05 (14.08)	67.65 (13.59)	0.77	0.043
BMI (kg/m^2), mean (SD)	25.1 (5.1)	24.9 (4.5)	0.78	0.041
Surgery, n (%)			0.71	0.080
Mastectomy	74 (78.7)	77 (81.9)		
Tumorectomy	20 (21.3)	17 (18.1)		
Axillary lymph node dissection	0 (0.0)	0 (0.0)		

The Figure 4.5 below provides a visualized comparison of covariate balance before and after matching. Propensity scores and all covariates are balanced in the matched sample.

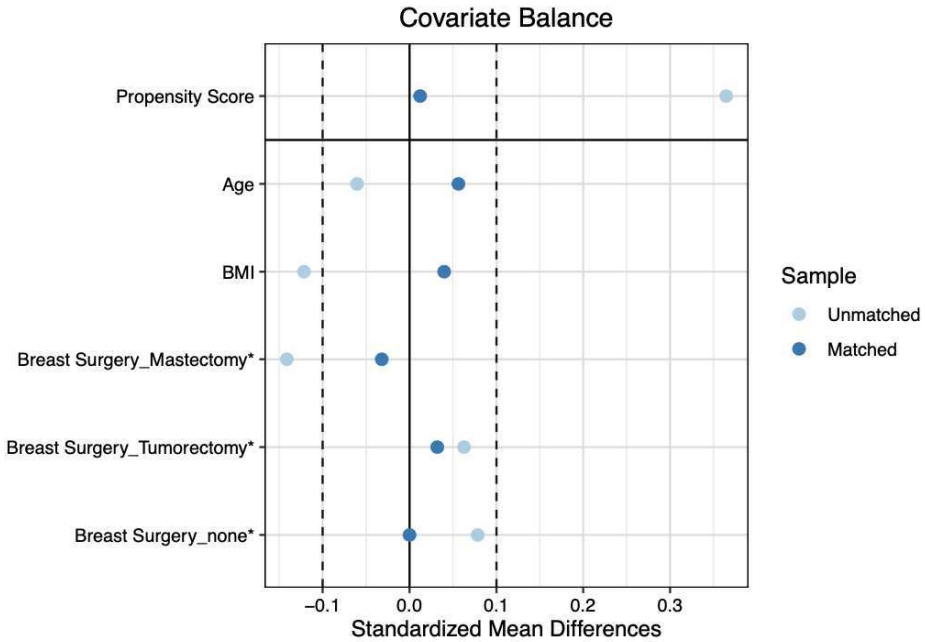


Figure 4.5: Covariate balance before/after propensity score matching

Treatment effect

The primary endpoint of this study was the effect of ESPB on the need for morphine titration after breast surgery in the PACU (indicated in both cohorts when VAS > 3). Table 4.5 shows that in the propensity score-matched sample, the percentage of patients who required morphine titration was significantly higher in the ESPB group than in the TPVB group (74.5% vs. 41.5%, p<0.001). The observed difference between the two groups was 33.0% (95% confidence interval [CI] 19.3%, 46.7%). Regarding secondary outcome, among all the propensity score-matched patients, the overall morphine dose was significantly higher in the ESPB group than in the TPVB group (3.7 mg vs. 2.2 mg, p=0.02).

Table 4.5: Primary and secondary outcomes in propensity score-matched patients

Outcome	ESPB (n=94)	TPVB (n=94)	p-value
Need for morphine titration, n(%)	70 (74.5)	39 (41.5)	< 0.001
Morphine dose (mg), mean (SD)	3.7 (3.3)	2.2 (3.2)	0.02

4.4.2 Sensitivity analysis

Figure 4.6 presents the distributions of the estimated propensity scores using the random forest model, which are similar to those estimated by the logistic regression model in Figure 4.4.

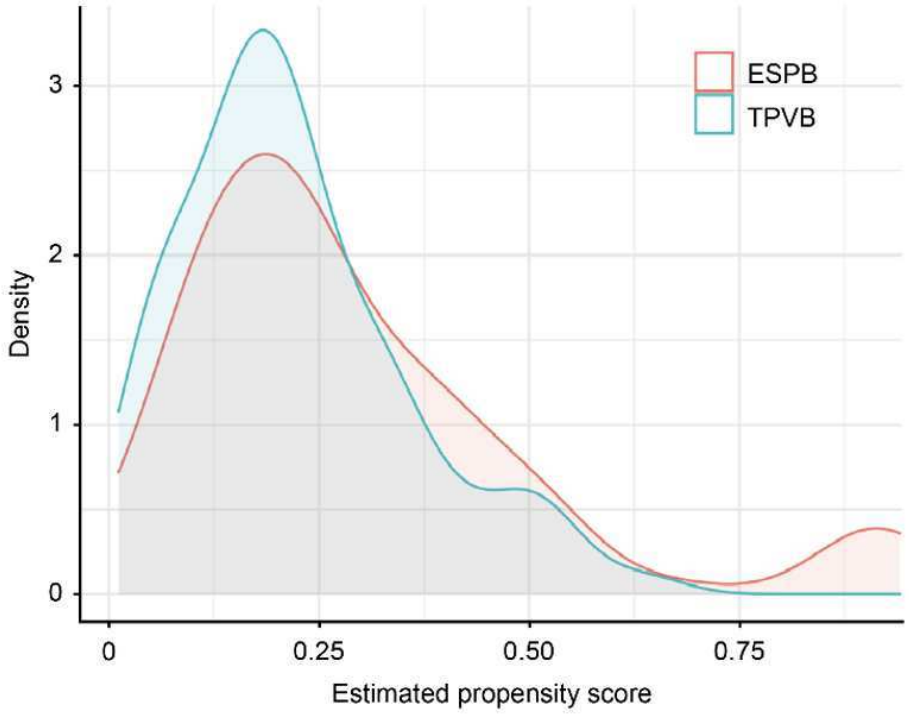


Figure 4.6: Distribution of the estimated propensity scores using a random forest model

Table 4.6 presents the comparisons of the baseline characteristics of patients matched on propensity scores estimated by the random forest model. Ninety-five out of 102 ESPB patients were matched with a TPVB patient. Seven ESPB patients who received axillary lymph node dissection without breast surgery were excluded from matching. Across the baseline covariates, the absolute SMDs of age and BMI were below 0.1, indicating a negligible difference. The SMD of the performed surgery type was 0.169, which slightly exceeded the preset threshold of 0.1 but was lower than the value of 0.466 before matching. The matching process created two groups of patients with more comparable covariates.

Table 4.6: Baseline characteristics patients matched on propensity scores estimated by the random forest model

Baseline characteristic	ESPB (n=95)	TPVB (n=95)	p-value	SMD
Age (years), mean (SD)	55.7 (12.9)	56.3 (14.1)	0.776	0.041
BMI (kg/m^2), mean (SD)	25.1 (5.0)	25.2 (4.5)	0.897	0.019
Surgery, n (%)			0.510	0.169
Mastectomy	74 (77.9)	78 (82.1)		

Table 4.6: Baseline characteristics patients matched on propensity scores estimated by the random forest model *(continued)*

Baseline characteristic	ESPB (n=95)	TPVB (n=95)	p-value	SMD
Tumorectomy	20 (21.1)	17 (17.9)		
Axillary lymph node dissection	1 (1.1)	0 (0.0)		

Table 4.7 presents the comparisons of the outcomes of patients matched on propensity scores estimated by the random forest model. The percentage of patients who required morphine titration was significantly higher in the ESPB group than in the TPVB group (74.7% vs. 38.9%, p<0.001). The observed difference between the two groups was 35.8% (95% CI [22.7%, 48.9%]). Among the patients who received morphine titration, the overall morphine doses were similar between the two groups (5.1 ml vs. 5.8 ml, p=0.14). The results of propensity score matching analysis with the random forest model are consistent with those of the logistic regression model.

Table 4.7: Outcomes of patients matched on propensity scores estimated by the random forest model

Outcome	ESPB (n=95)	TPVB (n=95)	p-value
Need for morphine titration, n(%)	71 (74.7)	37 (38.9)	< 0.001
Overall morphine dose (mg), mean (SD)	5.1 (3.0)	5.8 (2.7)	0.14

4.5 Discussion

In this study, we built a historical control arm for the single-arm observational trial (ESPB study) which assessed the effectiveness and safety of ESPB in preventing acute PSP following major breast cancer surgery. To establish a historical control group, we utilized data from the experimental arm of a randomized clinical trial (MIRs03 study) that evaluated the efficacy and safety of TPVB. Comparatively few studies have directly compared ESPB to TPVB, and the existing results are contradictory. Therefore, our study represents one of the largest clinical investigations comparing the efficacy and safety of ESPB and TPVB.

We discuss several limitations in this study. Firstly, the ESPB study is an obser-

vational cohort study, relying on a historical group for comparison. Using historical data can introduce concerns related to differences between studies. However, we mitigated this issue by ensuring that the ESPB and TPVB trials shared identical designs, including eligibility criteria and perioperative management protocols. This similarity allowed for comparable patient characteristics, intervention effects, and outcome measurements. Furthermore, we performed propensity score matching analysis to balance three significant prognostic factors of acute PSP (age, BMI, and breast surgery type) between the groups. To the best of our knowledge, there were no remaining systematic differences in the baseline covariates that could potentially affect acute PSP outcomes in the propensity score-matched subjects. Since the completion dates of both studies were in 2018 and 2019, respectively, we assumed that there were no substantial changes in clinical practice during this time. Additionally, we obtained consistent results when using propensity scores estimated by both logistic regression and random forest models, demonstrating the robustness of our conclusions.

Nonetheless, it should be noted that unmeasured confounding factors in observational studies may introduce bias. For instance, while the care protocols in the ESPB cohort matched those of the MIRs03 study, the multicenter nature of the latter could theoretically introduce heterogeneity in practice. Table 4.8 shows the number of patients on the experimental arm who received postoperative morphine titration at the five centers in the MIRs03 study. There was no significant difference in the incidences of morphine consumption between centers in the MIRs03 study (p-value = 0.5 of Kruskal-Wallis test). It is worth mentioning that recruitment for the ESPB cohort began in December 2018, following the completion of the MIRs03 study's patient recruitment period, which spanned from March 27, 2015, to June 3, 2018. Therefore, from December 2018 to August 2019, all patients undergoing major breast surgery at Institut Curie were treated with ESPB.

Table 4.8: Incidence of morphine titration among centers in the MIRs03 study

Outcome	Center 01 (n=120)	Center 02 (n=7)	Center 03 (n=25)	Center 04 (n=3)	Center 05 (n=10)	p-value
Need for morphine titration, n (%)	45 (38%)	2 (29%)	12 (48%)	0 (0%)	5 (50%)	p=0.5

In our research, each arm comprises approximately a hundred patients, which is notably fewer than large-scale observational studies that typically involve several hundreds of participants. In the context of small sample sizes, one study found substantial differences in the ATT estimate according to the model selected for propensity score estimation and subsequent matching based on a cohort of 66 children with sickle cell anemia who received either allogeneic bone-marrow transplant or chronic transfusion (Andrillon et al., 2020). This study underlined the importance of thorough sensitivity analyses when using propensity score matching in smaller trial cohorts. We thus conducted sensitivity analyses using a different method for propensity score estimation. We used random forest as an alternative method to logistic regression. We found consistent results of balance diagnostics and treatment effects in our analyses. In both scenarios, the balances were achieved across all covariates. Moreover, the results conclusively indicated that a significantly higher proportion of ESPB patients required morphine titration compared to their TPVB counterparts.

Through this case study, we established a protocol to create a historical control arm for a single-arm trial using data from previous clinical trials. The initial step involves evaluating the comparability of the target trial and historical data based on criteria such as standard treatment, patient eligibility, treatment evaluation, distributions of patient characteristics and completion time. It is imperative to select prognostic patient traits (i.e., the confounders) grounded in medical expertise and causal relationships rather than mere internal data exploration. Once discrepancies between trials are identified, appropriate statistical methods should be employed to account for confounding factors. Individual-level data can be analyzed using propensity score analysis, which helps address potential biases. Additionally, conducting

sensitivity analyses is crucial to assess the robustness of the results, particularly in terms of propensity score estimation and determination of confounding variables.

We acknowledge several advantageous facets inherent to this study's context. The eligibility criteria of patients are consistent in ESPB and MIRs03 studies; and the historical control data from the MIRs03 study are proper and complete. These advantages allowed us to dedicate effort towards appropriate data selection and implementation of propensity score analysis. Based on the insights from this study, we can continue our investigation on historical control arms in clinical trials. The ensuing steps can include:

- Broadening the data source from clinical trial data to real-world observational data;
- Refining the analysis protocol within a causal conceptual framework akin to randomized controlled trials.

Consequently, in the subsequent chapter, we explore the feasibility of historical control arms using larger observational database and target trial framework (Hernán & Robins, 2016).

Acknowledgment

We extend our gratitude to Dr. Aline Albi-Feldzer and Dr. Antoine Premachandra from Institut Curie for providing the data of ESPB and MIRSs03 trials, and for their invaluable collaboration throughout this project.

Chapter 5

Historical control arms with observational data

Contents

5.1 Introduction

With increasing data availability, the observational data which is also referred to real-world data (RWD) provides a big opportunity to substitute for randomized controlled trials (Franklin & Schneeweiss, 2017). RWD are data relating to patient health status and/or the delivery of healthcare routinely collected from different sources (FDA, 2018). These data find utility across numerous domains, such as therapeutic advancement, comparative analysis of effectiveness and safety, reimbursement strategies, regulatory decision-making, and the formulation of clinical guidelines. The reflection of 'diverse real-world practices' enhances the applicability and generalizability of RWD compared to data derived from randomized controlled trials (RCTs). And the abundant availability of RWD positions it as a promising data source for constructing historical control arms in clinical studies. However, these elements also indicate that RWD is less structured and more challenging to analyze. Moreover, the risk of inaccuracies in methodological applications and a shortage of adequate expertise in this area are potential threats to the validity and reliability of RWD studies (Collins et al., 2020).

Defining the research question using a causal inference framework is a crucial step to generate robust evidence for RWD studies (Gokhale et al., 2020). Several causal inference frameworks are employed to aid in defining precise scientific questions. These include the Estimand Framework (EF) and Target Trial Emulation Framework (TTF).

The EF has been progressively used by health authorities and pharmaceutical firms since its introduction in 2017 (ICH, 2021). The framework provides a systematic approach to the definition of the treatment effect under investigation in a clinical trial. An estimand consists of five attributes: treatment, population, variable, population-level summary, and handling of intercurrent events. Each of these attributes is defined in an interdisciplinary discussion during the trial planning phase, based on the clinical question being asked. Intercurrent events are those occurring after treatment initiation and can affect the interpretation or existence of endpoint-associated measurements. In defining the estimand, the primary focus lies on the scientific objective of the trial or study, with due consideration to missing data. Although the EF primarily targets RCTs, its principles can also be applied in estimating treatment effect in single arm trials or observational studies. However, it's worth noting that estimating a causal effect from observational data often presents unique

challenges compared to RCTs. These challenges may include biases resulting from baseline confounding and selection, missing data, and the complexity of determining the comparison's index date.

The TTF represents an additional causal framework capable of refining the specificity of scientific queries within comparative assessments (Bigirumurame et al., 2023; Hernán et al., 2022; Hernán & Robins, 2016). The TTF addresses gaps in the analysis of observational data by applying principles of RCTs to non-randomized comparative assessments. This involves defining a hypothetical randomized trial to address a specific scientific question, followed by specifying how non-randomized data can emulate this trial. Crucial components of a target trial protocol include eligibility criteria, treatment strategies, treatment allocation, initiation and termination of follow-up, outcomes, causal contrasts, and the analytical approach (estimator). This framework can be deployed when combining clinical trial and observational data, as may occur when creating an external comparator to a single-arm trial with observational data.

In this study, we aim to investigate the application of the target trial framework for building a historical control arm using observational data for a target trial in ovarian cancer.

- **Section 5.2** introduces the motivating trial and the source of observational data in this study.
- **Section 5.3** presents the application of the target trial framework in creating historical control arms.
- **Section 5.4** outlines the results derived from our preliminary analysis.
- Finally, **Section 5.5** provides a discussion about the challenges encountered in this study.

5.2 Motivating data

5.2.1 Motivating trial: the PAOLA study

Patients with newly diagnosed, advanced ovarian cancer undergo cytoreductive surgery and platinum-based chemotherapy with curative intent. Olaparib, a PARP inhibitor, has shown significant benefits as maintenance therapy in women with newly diagnosed advanced ovarian cancer who carry a BRCA mutation.

The PAOLA-1 trial (NCT02477644) is a randomized, double-blind, international phase III trial designed to assess the efficacy of olaparib maintenance therapy in conjunction with bevacizumab in patients with newly diagnosed advanced ovarian cancer (Harter et al., 2022; Ray-Coquard et al., 2019). Eligible patients had newly diagnosed advanced high-grade ovarian cancer and were responding to first-line platinum-taxane chemotherapy plus bevacizumab, regardless of surgical outcome or BRCA mutation status. Patients were randomly assigned in a 2:1 ratio to receive either olaparib tablets or a placebo for up to 24 months. The trial enrolled a total of 806 patients, of which 537 received olaparib and 269 received the placebo. After a median follow-up of 22.9 months, the median progression-free survival (PFS) was 22.1 months in the olaparib plus bevacizumab group and 16.6 months in the placebo plus bevacizumab group (hazard ratio for disease progression or death, 0.59; 95% confidence interval [CI], 0.49 to 0.72; P<0.001). The Kaplan–Meier estimates of progression-free survival is shown in Figure 5.1.

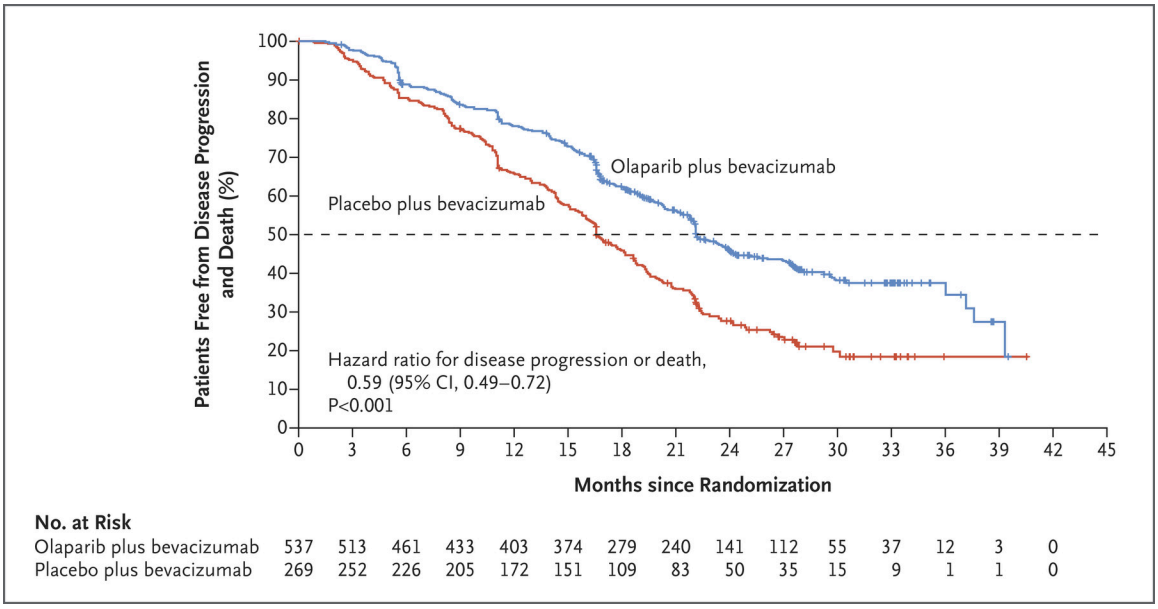


Figure 5.1: Kaplan–Meier estimates of investigator-assessed progression-free survival in the PAOLA-1 study

We noted that the PAOLA-1 study adopted a randomization ratio of 2:1 in favor of patients receiving the innovative treatment. This indicates an underlying motivation of investigating the feasibility of creating a historical control arm for the PAOLA-1 study, as well as other studies operating under similar circumstances. Specifically, the patients assigned to the control arm received the standard of care for ovarian cancer treatment predating the approval of olaparib. Therefore, we propose to use

real-world data in ovarian cancer to create the historical control arm, serving as the comparator for evaluating the outcomes of the experimental treatment.

5.2.2 Observational data: ESME database

Initiated in 2014 by Unicancer R&D and backed by all French Comprehensive Cancer Centres (FCCCs), the Épidémio-Stratégie Médico-Economique (ESME) programme represents an independent academic endeavor aimed at aggregating real-world patient data pertaining to cancer treatment in France. The primary goal of ESME is to chronicle the progress of patient management and modifications in therapeutic strategies over time, framed within an extensive medico-economic perspective. To date, three data platforms have been constructed, encompassing metastatic breast cancer, ovarian cancer, and lung cancer.

The ESME Ovarian Cancer (ESME-OC) database is a comprehensive, real-world, retrospective, multicentric database that centralizes the clinical data of all consecutive patients undergoing treatment for ovarian cancer since 1 January 2011 in any of the 18 French Comprehensive Cancer Centers within the Unicancer network. The ESME-OC database encompasses prospectively collected data from electronic medical records, inpatient hospitalization records, and pharmacy records (Figure 5.2).

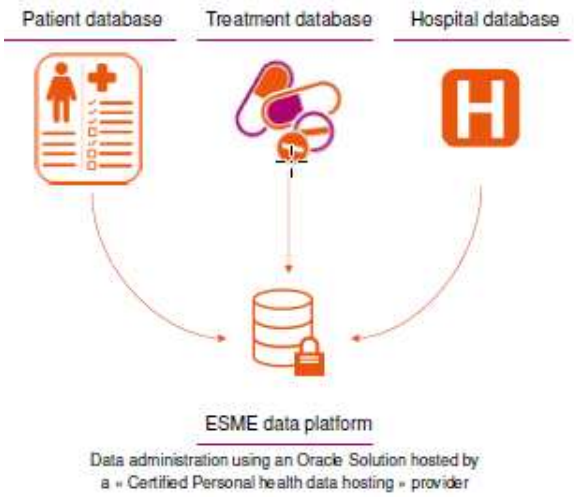


Figure 5.2: ESME data platform

Several recent investigations have used the ESME-OC database to describe the natural disease history of ovarian cancers and to evaluate the optimal timing of de-

bulking surgery (Bini et al., 2022; De Nonneville et al., 2021; Thomas et al., 2022). Therefore, this observational database has the potential to reflect the outcomes of the PAOLA-1 study in routine patient care.

5.3 Methodology

Emulating a target trial involves a systematic approach that aims to replicate, as closely as possible, the conditions and design of a hypothetical randomized controlled trial using non-randomized data. The general steps include:

1. Formulate the research question: Clearly define the clinical question of interest as if we were designing an actual RCT.
2. Define eligibility criteria: Specify inclusion and exclusion criteria for study participants, similar to patient selection in a trial.
3. Specify the interventions of interest: Clearly define the interventions or exposures to be studied; and determine the point in time at which participants would hypothetically be randomized (the so-called "index time").
4. Determine the start of follow-up (time zero): Define the time of starting observing outcomes in relation to the index time.
5. Specify outcomes: Define primary and secondary outcomes of interest; and determine how these outcomes will be measured and verified within the observational data.
6. Determine the end of follow-up: Define when the observation of study participants will end, whether it's after a certain period of time, upon the occurrence of a certain event, or another pre-specified criterion.
7. Plan to address confounding: Identify potential confounders that could bias the relationship between exposure and outcome; and decide on methods to adjust for these confounders, such as stratification, matching, propensity scores, or other statistical techniques.
8. Plan the Analysis: Specify the statistical methods to estimate the effects of the intervention or exposure; and address potential issues like missing data, measurement error, and model assumptions.
9. Execute the study and analyze data: Use the observational data to "mimic" the hypothetical trial, following the pre-specified analysis plan.
10. Interpret results: Discuss findings in the context of the hypothetical RCT de-

sign; and reflect on the limitations of using observational data, any potential sources of bias, and how well the target trial was emulated.

5.3.1 Application of the target trial framework

The objective of the PAOLA-1 study was to evaluate maintenance therapy with a PARP inhibitor (olaparib) as compared with placebo in patients with newly diagnosed advanced ovarian cancer who were receiving chemotherapy plus bevacizumab followed by bevacizumab, regardless of BRCA mutation status.

With the real-world data, we define the scientific research question as "Is there be a difference in progression free survival (PFS) and overall survival (OS) between patients with newly diagnosed advanced ovarian cancer receiving olaparib maintenance therapy in combination with bevacizumab in PAOLA-1 study versus patients with newly diagnosed advanced ovarian cancer who received bevacizumab as part of routine care?"

Then we apply the target trial framework to design a historical control arm for the PAOLA-1 study using ESME data. Table 5.1 illustrates the TTF attributes that define the estimand aligned with the scientific research question. The average treatment effect on the treated (ATT) is the estimand of primary interest. This is the treatment effect difference of using olaparib + bevacizumab in a clinical trial versus using bevacizumab in clinical practice, and hence the target population is defined by the clinical trials population.

Table 5.1: Application of the target trial framework to design a historical control arm for the PAOLA-1 study using ESME data

TTF component	Target trial (PAOLA-1 study)	ESME cohort
Patient population (Eligibility criteria)	Eligibility criteria of PAOLA-1	Same as the target trial
Treatment	Olaparib maintenance therapy in combination with bevacizumab	Bevacizumab as part of routine care

Table 5.1: Application of the target trial framework to design a historical control arm for the PAOLA-1 study using ESME data *(continued)*

TTF component	Target trial (PAOLA-1 study)	ESME cohort
Assignment procedure	Participants were randomly assigned to one of the two treatment settings	Randomization is emulated by mathching or weighting observations for the inverse probability of treatment
Outcomes	Progression free survival (PFS) and overall survival (OS)	Same as the target trial
Start of follow-up	Start of follow-up occurs at the time when the treatment is assigned	Same as the target trial (the real-world start of follow-up occurs at the time when the treatment is initiated in routine care)
Analysis plan	Hazard ratio (HR) with 95% confidence interval (CI)	Same as the target trial

5.3.2 Selection of the real-world cohort

We use the eligibility criteria of the PAOLA-1 study to select the real-world cohort. In the PAOLA-1 study, eligible patients were 18 years of age or older and had newly diagnosed advanced (International Federation of Gynecology and Obstetrics [FIGO] stage III or IV), high-grade serous or endometrioid ovarian cancer, primary peritoneal cancer, or fallopian-tube cancer. Patients with other nonmucinous epithelial ovarian cancers were eligible, provided they had a deleterious germline BRCA1 or BRCA2 mutation. Patients were eligible irrespective of previous surgical outcome (residual macroscopic disease or no residual macroscopic disease after upfront or interval surgery). After first-line treatment with platinumtaxane chemotherapy plus bevacizumab, patients were required to have no evidence of disease or to have had a clinical complete or partial response. Patients had an Eastern Cooperative Oncology Group performance status of 0 or 1 (on a 5-point scale in which higher numbers reflect greater disability), and a tumor sample had to be available for central testing to

determine BRCA mutation status.

To select the historical control cohort from the ESME database with comparable characteristics of patients in the PAOLA-1 study, we first search the variables that correspond with the PAOLA-1 eligibility criteria. In instances where corresponding variables are absent, we resort to using surrogate characteristics guided by expert clinical knowledge.

Besides, ovarian cancer patients follow one of two treatment regimens in clinical practice:

(1) Primary Debulking Surgery (PDS) followed by Adjuvant Chemotherapy (CA), or
(2) Neoadjuvant Chemotherapy (CNA) followed by Interval Debulking Surgery (IDS) and then CA.

To ensure an accurate outcome analysis, it's important to evaluate each treatment regimen separately. For this purpose, we categorize patients based on their treatment regimen using available variables in the ESME database.

5.3.3 Statistical analysis

We implement the inverse probability of treatment weighting (IPTW) on propensity score to balance baseline patient characteristics between the clinical trial arm and the historical control arm. Propensity scores are estimated through a multi-logistic model, defined as probabilities of allocation to the treatment group conditional on selected confounders. Propensity score estimation using other machine learning methods, such as random forest and neural network, is also conducted as sensitivity analysis (Westreich et al., 2010). These confounders are determined a priori based on clinical expert knowledge and the availability of relevant variables in the ESME database.

As our objective focuses on the average treatment effect on the treated, patients from the PAOLA-1 trial are assigned a weight of one, while weights for patients in the ESME database are computed as the ratio of the estimated propensity score to the complement of the estimated propensity score (i.e., odds of being treated in the clinical setting). After IPTW, differences in baseline characteristics are evaluated through standardized mean differences (SMDs) with patient characteristics considered

equilibrated if SMD is less than 0.10. The weighted cohort is then deployed for subsequent outcome analyses.

The treatment effects are estimated by weighted survival analysis. Specifically, we estimate the hazard ratio (HR) using a IPTW weighted Cox proportional hazard model and the 95% CI for the HR using bootstrap approach (Schaubel & Wei, 2011). We also use the IPTW weighted Kaplan–Meier method to calculate survival function estimates and weight log-rank test for intergroup comparisons.

5.4 Results

We have access to the data of 596 patients treated in Institut Curie cancer center, including information on basic characteristics, diagnosis, surgery, received treatment, etc.

Our current work is focused on the selection stage of the real-world cohort. Ovarian cancer patients follow one of two treatment schemas: (1) Primary Debulking Surgery (PDS) + Chemotherapy Adjuvant (CA) or (2) Chemotherapy Neoadjuvant (CNA) + Interval Debulking Surgery (IDS) + CA. Our primary objective is to separate patients based on their respective treatment schema.

Within the ESME database, all surgical information is encapsulated within the SURGERY table. Nonetheless, this table lacks a specific variable to unequivocally identify whether a surgery is classified as debulking or not. As a workaround, we propose to utilize pertinent medical characteristics and select relevant variables from the table. The sequencing of surgical intervention and chemotherapy will also be compared to differentiate between PDS and IDS.

We are currently engaged in the identification of debulking surgery data with the aim of refining our selection of patients who meet the PAOLA-1 eligibility criteria. This work is ongoing, and we will report the analytical findings at a later stage.

5.5 Discussion

The rising interest in historical control arms has led to several studies suggesting frameworks and guidelines for their appropriate use.

One study introduced an evaluation framework for real-world data external comparator studies, allowing for a thorough assessment of existing evidence and associated biases. This four-step process ensures the proper conduct of external comparator studies (Gray et al., 2020). Step 1 includes an assessment of sources of bias and exchangeability. Step 2 adjusts for measured confounders using analytical approaches. Step 3 uses quantitative bias analysis to quantify the impact of potential bias. Step 4, applicable only to augmented randomised controlled trials, combines outcomes between the internal and external comparator groups dependent upon the similarity of the two cohorts.

Another study presented a three-step guideline designed as a robust tool for the critical evaluation of comparisons between external control groups and single-arm trials (Lambert et al., 2022). Step 1 defines an estimand that mirrors a clinical question. This estimand encompasses the treatment effect and the targeted population. Step 2 focuses on the appropriate selection of external controls, whether from previous RCTs or real-world patient data sources such as cohorts, registries, or electronic patient records. Step 3 involves selecting the statistical method that targets the previously defined treatment effect. The chosen method will depend on the nature of the available data, be it individual-level or aggregated external data.

The primary objective of the frameworks and guidelines examined parallels that of the target trial framework: they all aim to ensure the validity of incorporating historical control data in clinical trials. The target trial emulation has a broad scope, aiming to mimic the entirety of a hypothetical RCT in observational data. Thus, we were interested in using this framework as a potential method to evaluate the feasibility of historical control arms.

Target trial emulation has most commonly been used for hypothesis generation and confirmatory studies (Bacic et al., 2020; Hernán & Robins, 2016; Huitfeldt, 2015). In contrast to other observational study paradigms, no standardized guidelines currently exist to enhance research transparency and ensure reproducibility in target trial emulation. Practical guides on how to conduct emulated trials are also not widely available.

In this study, we have defined the methodology to emulate a target trial for the PAOLA-1 study using observational data from the ESME database. During the initial phases, we encountered challenges in selecting patient variables and discerning distinct treatment schemas. To address these challenges, we are collaborating with

clinicians to identify surrogate variables, thereby allowing the creation of an observational cohort based on pre-established eligibility criteria.

We would like to draw attention to lack of initiatives aimed at enhancing the data quality within real-world databases. Our exploration revealed that the granularity of these databases might not adequately represent clinical practices, potentially complicating the extraction of clear answers to research queries. It's imperative to collaborate closely with clinicians to interpret the data, enhance its quality, and streamline real-world data analysis.

Acknowledgment

We extend our sincere appreciation to Dr. Nina Oufkir from Institut Curie for providing the ESME data and offering her medical expertise in ovarian cancer within this project.

Chapter 6

Conclusion

6.1 General discussion

Well-designed randomized controlled trials are the gold standard for evaluating the efficacy of new treatments. RCTs establish causal conclusions by randomly assigning patients to either an investigational or concurrent control treatment that usually consists of a placebo or standard of care. Nevertheless, ethical or feasibility issues can hinder the randomization process, especially in oncology drug development. Historical control arms have emerged as an alternative approach to knowledge production. By providing contextual information and improving the interpretation of single-arm trial results, historical control arms can help reduce the bias due to the lack of randomization (Lim et al., 2018). Furthermore, increasingly available clinical data from historical clinical trials or real-world databases provide the potential to expand the use of historical control data to minimize patient burden and facilitate study conduct in the drug development process (Khozin et al., 2017). Regulatory authorities have signaled their support for using real-world data to generate clinical evidence (Cave et al., 2019; FDA, 2018). However, recent evidence shows that RCT and RWE findings were not always matched despite attempts to emulate RCT design and confounder adjustment (Franklin et al., 2021). Thus, challenges remain before historical control arms can be an integrated part of decision-making.

In this context, this thesis aims to evaluate the feasibility of incorporating historical control arms into clinical trials, drawing from combined perspectives of different stakeholders in drug development.

We realized that historical control arms in clinical trials have increasingly garnered the attention of various stakeholders in drug development and evaluation. Their relevance and application have often been scrutinized from a multitude of perspectives. Within this ambit, there are four primary roles that shape the discourse around historical controls: sponsors, regulators, clinical investigators, and biostatisticians. The varied interpretations from these groups underscored the necessity to initiate our study with a comprehensive summary and understanding of the historical control concept. Thus, in Chapter 1, we provided an overview of randomized controlled trials and their conceptual framework, highlighting their advantages and limitations. And we introduced the definition of historical control arms with various aspects of their design.

The main concern of using historical control arms revolves around the potential for biases, given that past patients might differ in various ways from current patients. There might also be differences in diagnostic methods, care standards, or other factors that could influence outcomes. Properly accounting for these potential differences is critical to ensure the validity of study findings using historical controls. Therefore, we reviewed the developed statistical methods related to historical control arms in Chapter 2. There are two major frameworks: Bayesian and frequentist methods. Bayesian borrowing methods consider outcome heterogeneity and discount the historical control data when incorporating it into the new clinical trial, such as power prior, commensurate prior, or meta-analytic prior. Frequentist approaches involve two primary steps. First, a balance score is estimated using selected covariates that may affect treatment assignment and outcome. Second, the balance score is used to create comparable external and internal cohorts through methods like matching, inverse probability weighting, stratification and covariate adjustment. Propensity score is one balance score that is being widely used in medical research. We thus presented the concept and main methods of propensity score analysis. Note that Bayesian methods are suitable for aggregate outcome data from one or several previous trials without directly considering the information on patient characteristics. On the other hand, propensity score analysis methods can be applied to individual data with patient-level characteristics. The choice of statistical methods should be corresponding to the available data type.

One of the pivotal stakeholders in the drug development process is the regulator. Despite numerous guidelines that regulators have published over the years, the feedback they provide on specific applications is of utmost importance to many in the

industry. Our investigation in Chapter 3 focused on examining real drug approval cases in recent years wherein historical controls were used. In this review, we identified and summarized the characteristics of European oncology drug approvals using historical control data in clinical efficacy. Eighteen eligible submissions leveraging 24 historical controls were included. We discussed the use of historical control, data sources, methods of comparison, and the regulators' feedback.

It emerged that sponsors are increasingly relying on historical control data as supplementary evidence in their drug approval submissions, reflecting the currently emerging research interests in this field. In their applications, sponsors often incorporated individual data from clinical trials as well as RWD, and they have applied the statistical methods we have discussed in the outset of the thesis.

However, we surprisingly found that most of the historical controls were not deemed supportive evidence by EMA due to significant limitations regarding heterogeneous patient populations, missing RWD of outcome assessment, and suboptimal study design. The regulators have maintained stringent views concerning historical control arms in the face of emerging research trends. Two primary reasons underscored this cautious approach. Firstly, the historical control data submitted by sponsors frequently failed to demonstrate the comparability of internal and external patients, making it challenging to assure consistent and reliable outcomes. Secondly, even though many statistical methods have been developed to reduce bias, the application of these methods was often found to be imprecise, falling short in addressing and mitigating biases. This study highlighted that the proper use of historical controls in oncology clinical trials requires carefully assessing data availability and determining the optimal clinical and statistical methodology. The insights gained from the review directed our following studies to focus on evaluation of comparability of internal and external populations, reasonably selection of confounders and appropriate implementation of statistical analysis when incorporating historical control arms into clinical trials.

Recognizing the inherent challenges associated with historical control arms, we embarked on two case studies. These were primarily grounded in individual data sourced from clinical trials and RWD, respectively. These case studies further elucidated the nuances of using historical control arms in different contexts and added a practical dimension to our theoretical discussions.

In Chapter 4, the first case study investigated an observational single-arm

trial that evaluated the effectiveness of erector spinae plane block in reducing post-operative pain in breast cancer surgery. We constructed a historical control arm using data from a previous clinical trial. We first evaluated the comparability of the new trial and historical data using Pocock criteria in terms of standard treatment, patient eligibility, treatment evaluation, distributions of patient characteristics and completion time. We identified between-trial unbalance in two of the three confounders selected based on medical knowledge. Then we successfully balanced these characteristics with well-implemented propensity score matching analysis: propensity score estimation using pre-selected confounders, balance diagnostics and sensitivity analysis for propensity score estimation. This case study illuminated how rigorous data screening and apt statistical methods can eliminate discrepancies from external data, establishing a robust historical control.

Additionally, the inherent strengths of this study were evident: consistent eligibility criteria between the new and historical trials, and superior quality and accessibility of historical control data from the prior trial. These benefits permitted us to center on the pivotal steps of using historical controls: appropriate data selection and propensity score analysis execution. However, an inherent limitation was that our analysis was retrospective, making it supplementary to the concluded clinical trial.

Based on the insights from the first case study, we aimed to advance our investigation into a more prospective setting and to examine the feasibility of historical controls under more intricate data sources. We believe that historical controls shouldn't merely supplement clinical trials but should be integrated into clinical trial design and methodologies. With the theoretical backing of the potential outcome framework and systematic guidance from the Estimand Framework (EF), recent studies have proposed the Target Trial Framework (TTF) for observational studies (Hernán et al., 2022). Therefore, we sought to incorporate the TTF into historical controls while broadening our data source from clinical trial data to RWD.

In Chapter 5, the case study focuses on a randomized controlled trial examining Olaparib plus Bevacizumab as first-line maintenance therapy in ovarian cancer. The objective is to emulate the control arm by leveraging observational data from the real-world database ESME within the target trial framework. Though the RWD database has the advantage of larger sample size, it presents challenges in data management and data preprocessing. Working with clinicians, we are currently engaged in the identification of debulking surgery data with the aim of refining our selection of patients who meet the PAOLA-1 eligibility criteria.

Meanwhile, it's pertinent to note that while the key issue with historical controls is attributed to bias from inconsistent internal and external patients, our journey through this thesis highlighted another "real-world" challenge – data access.

Our primary obstacle was accessing target trial data. Each case study required a target trial in need of historical control support, but soliciting such trials for feasibility evaluation was complex. Given the collaborative nature of this thesis between a corporate entity and a research institute, we went through a more intricate legal and administrative process within the regulatory framework surrounding data governance and patient confidentiality. Additionally, there's the elephant in the room – commercial confidentiality. Initially the thesis set on using a clinical trial associated with one of Sanofi's oncology products as the target trial. We identified several potential candidates from Sanofi's oncology pipeline: Clinical trial NCT04191382, which investigated the drug Amcenestrant for breast cancer; Clinical trial NCT02990338, which investigated Isatuximab for refractory or relapsed and refractory multiple myeloma; and Clinical trial NCT03367819 which explored the efficacy of Isatuximab in combination with REGN2810 for prostate cancer or non-small cell lung cancer. However, due to constraints related to data privacy and time limitations, none of these projects could be materialized. Consequently, we shifted to the PAOLA-1 study as our target trial.

The secondary obstacle was the acquisition of historical control data. In our first case study, we used data from a previous clinical trial as the historical control data and thus faced the same challenge mentioned above. For the second case study, we considered the real-world databases that Sanofi had licensed, such as Optum's Real-World Data (https://www.optum.com/business/life-sciences/real-world-data.html/) and Flatiron's Clinico-Genomic Database (https://flatiron.com/real-world-evidence/clinico-genomic-database-cgdb/). However, the publication of findings gained from these vendor databases required further deliberation. Therefore, we ultimately turned to the ESME database to conduct our case study.

In retrospect, these challenges highlight the intricate issues faced when academic institutions partner with corporations in healthcare research. On one hand, research aims for open access and transparency. On the other, businesses and regulators often require confidentiality and thorough checks. Finding the right balance between these differing needs is crucial. Our experience in this thesis serves as a small example of navigating these complexities. It emphasizes the importance of open communication,

adaptable regulations, and innovative approaches to overcome such hurdles in the future.

In conclusion, the exploration of historical control arms in clinical trials underscores the dynamic and evolving landscape of drug development. While they offer a promising opportunity to advance clinical research, the practicalities of their integration into clinical trials face challenges — from data access to statistical methodologies. This thesis, rooted in collaboration between academia and corporate spheres, contributes to advancing our understanding of the feasibility and applicability of historical control arms, shedding light on their potential benefits and appropriate applications in the field of oncology drug development.

6.2 Perspectives

In our studies, our primary tool was propensity score analysis tailored for individual data. Further statistical methods can be considered. Alternative methods to deduce efficacy involve comparing outcomes from single-arm patients with machine-learning projections for control patient outcomes, such as G-computation, doubly robust estimation and doubly debiased machine learning (Chernozhukov et al., 2018; Funk et al., 2011; Snowden et al., 2011).

G-computation operates on a counterfactual framework, suggesting that patient outcomes can be predicted if they had been part of the control rather than the experimental group, paving a way to infer causality (Snowden et al., 2011). Doubly robust estimation combines propensity score method and outcome regression method (Funk et al., 2011). Outcome regression method estimates the treatment effect by modeling the outcome as a function of the treatment and the confounders. Doubly robust estimator is consistent (meaning the estimate converges to the true effect) if either the propensity score model or the outcome regression model is correctly specified. The modern evolution of causal inference in machine learning has reinvigorated interest in the counterfactual framework, leading to the methods like doubly debiased machine learning, aiming to rectify the bias in machine learning estimates (Chernozhukov et al., 2018).

A comparative evaluation of various statistical techniques like propensity score matching, inverse probability of treatment weighting, G-computation, and doubly debiased machine learning was conducted (Loiseau et al., 2022). This evaluation illu-

minated that doubly debiased machine learning exhibited minimal bias, outperforming G-computation. Furthermore, techniques rooted in outcome prediction models surpassed propensity score approaches in reducing estimation errors and bolstering statistical power. Therefore, these methods can be applied in our future work.

Besides, notable advancements were recently observed in the regulatory landscape pertaining to using of historical control arms for drug authorization. In 2023, key regulatory bodies have issued publications highlighting the pivotal role of historical control arms in the evaluation of medicinal products. The EMA published a reflection paper on single-arm trials as pivotal evidence for the authorisation of medicines in the EU (EMA, 2023). In parallel, the FDA issued a guidance document on considertaions for design and conduct of externally controlled trials (FDA, 2023). Additionally, Haute Autorité de Santé (HAS) published a paper on rapid access to innovative medicinal products mentioning historical control arms (Vanier et al., 2023).

These publications collectively highlight a growing interest in integrating historical control data into the drug development process. They stress the importance of establishing a comprehensive framework that involves all stakeholders, including regulatory authorities, industry experts, researchers, and statisticians. These regulatory documents resonate closely with the core messages and objectives of this thesis. It is crucial to note that historical control arms are not intended to completely replace randomized control arms but rather to complement them in appropriate clinical settings. Therefore, we propose the development of a relevant framework to ensure the proper utilization and validation of historical control arms. This framework aims to enhance the efficiency and effectiveness of clinical trials and expedite the authorization of innovative medicines.

Further projects

Within the context of PhD CIFRE, in addition to my thesis projects, I worked at the Sanofi Development Real-World Evidence team, guided by the expertise of Dr. Ramon Hernandez. The team conducts RWE projects within four strategic pillars: indication identification, optimization of clinical development plan, drug combination and supporting regulatory submissions.

I contributed to the data analysis segment for diverse RWE projects spanning indications in oncology, neurology, and immunology. Data anal-

ysis included real-world cohort creation and comparisons, survival analysis (Kaplan-Meier plots, log-rank test, Cox proportional hazards regression), etc. The analyses were conducted using developed RWE analytics platforms, such as the Aetion Evidence Platform (https://aetion.com/) and TriNetX (https://trinetx.com/). The real-world databases I worked with include Flatiron's Clinico-Genomic Database (https://flatiron.com/real-world-evidence/clinico-genomic-database-cgdb/) and Optum's Real-World Data (https://www.optum.com/business/life-sciences/real-world-data.html/).

Appendix A

Published article 1

PLOS ONE

RESEARCH ARTICLE

Erector spinae plane block versus thoracic paravertebral block for the prevention of acute postsurgical pain in breast cancer surgery: A prospective observational study compared with a propensity score-matched historical cohort

Antoine Premachandra[1‡*], **Xiaomeng Wang**[2,3‡], **Mary Saad**[1,2,4], **Sahar Moussawy**[1], **Roman Rouzier**[2,5], **Aurélien Latouche**[2,4☯], **Aline Albi-Feldzer**[1☯]

1 Department of Anaesthesiology, Institut Curie, PSL Research University, Saint-Cloud, France, **2** INSERM, U900, Institut Curie, PSL Research University, Saint-Cloud, France, **3** Department of Research and Development, Sanofi, Chilly Mazarin, France, **4** Conservatoire National des Arts et Métiers, Paris, France, **5** Department of Surgical Oncology, Centre François Baclesse, Caen, France

☯ These authors contributed equally to this work.
‡ These authors are joint first authors on this work. XW and AP also contributed equally to this work.
* antoine.premachandra@gmail.com

Citation: Premachandra A, Wang X, Saad M, Moussawy S, Rouzier R, Latouche A, et al. (2022) Erector spinae plane block versus thoracic paravertebral block for the prevention of acute postsurgical pain in breast cancer surgery: A prospective observational study compared with a propensity score-matched historical cohort. PLoS ONE 17(12): e0279648. https://doi.org/10.1371/journal.pone.0279648

Editor: Silvia Fiorelli, Sapienza University of Rome: Universita degli Studi di Roma La Sapienza, ITALY

Received: August 27, 2022

Accepted: December 12, 2022

Published: December 30, 2022

Data Availability Statement: The data that support the findings of this study are owned by Institut Curie and are available on request. Data can be asked to the CRI of the Curie Institut in Saint Cloud, then the RIPH Group (recherche impliquant la personne humaine) will evaluate the request. Contact: alexia.savignoni@curie.fr.

Abstract

Background

Preventing acute postsurgical pain (PSP) following breast cancer surgery is a major issue. Thoracic paravertebral block (TPVB) has been widely studied for this indication. Erector spinae plane block (ESPB) has been assumed to be effective. We aimed to compare the efficacy and safety of ESPB over TPVB in preventing acute PSP.

Methods

In this prospective observational study, 120 patients admitted for unilateral major oncologic breast surgery received T2/T3 ESPB (ropivacaine 0.75%, 0.35 ml.kg^{-1}), and 102 were analysed. Then, the ESPB cohort was compared to a TPVB cohort from the experimental arm of a randomized controlled study with the same protocol (NCT02408393) using propensity score matching analysis. The primary outcome was the need for morphine consumption in the PACU. Secondary outcomes were the morphine total dose, the incidence of ESPB and TPVB complications, and discontinuous visual analogue scale measurement trends at rest and at mobilization in the 24 hours after surgery.

Results

A total of 102 patients completed the study between December 2018 and August 2019. Propensity score matching formed 94 matched pairs. The proportion of morphine titration in the PACU was higher in the ESPB group than in the TPVB group (74.5% vs. 41.5%, p<0.001),

Funding: The authors received no specific funding for this work.

Competing interests: The authors have declared that no competing interests exist.

with a between-group difference of 33.0% (95% CI [19.3%, 46.7%]). No ESPB-related complications were observed.

Conclusion

ESPB is less effective in preventing morphine consumption in the PACU than TPVB. Our findings do not support the use of ESPB as the first-line regional anaesthesia for major breast cancer surgery. Randomized trials comparing ESPB and TPVB are needed.

Introduction

The incidence of acute postsurgical pain (PSP) following breast cancer surgery is as high as 70% [1]. Acute PSP impacts quality of life and may increase the risk of chronic PSP [2,3]. Thus, any PSP reducing strategy is highly beneficial for patients.

Recent evidence suggests that regional anaesthesia (RA) in a multimodal analgesia program can efficiently minimize acute PSP and opioid consumption after breast cancer surgery [4–6]. Thoracic paravertebral block (TPVB) has been widely studied and considered the "gold standard" for major breast cancer surgery. In addition, this technique has recently been proven to be safe when performed under ultrasound guidance [1,7,8]. Erector spinae plane block (ESPB), described in 2016 as an alternative to TPVB, is an interfacial block performed in the plane of the spine erector muscles, which is more superficial than the paravertebral space [9]. Clinical and cadaveric studies suggest that ESPB may act on the ventral rami of the spinal nerves in the paravertebral space via a diffusion process through the costotransverse foramen and the costotransverse ligament [10,11]. ESPB is more advantageous than TPVB because it is a simpler, faster procedure [12], and because of easily identifiable landmarks, it may be safer because of more distant injection sites from the pleura and perimedullary space [13] (Fig 2).

Although ESPB has been shown to be efficient in preventing acute PSP after spine and thoracic surgery [14,15], the reliability of this diffusion process remains controversial, and ESPB has been proposed for breast surgery without evidence supporting its efficacy [16–18]. Moreover, the clinical relevance of its benefits has been questioned [19].

In breast surgery, ESPB has been shown to be superior to general anaesthesia alone and to placebo [16,20–23].

However, there is no large study comparing ESPB to TPVB for major breast surgery. The available studies comparing ESPB to TPVB have shown analyses of a relatively low number of patients, and their results are pooled in several meta-analyses that have shown conflicting results [24–27]. Thus, the role of ESPB in the strategy of analgesia for breast surgery is not clearly defined, and the comparison of ESPB performance with that of TPVB will provide some answers.

Hence, we conducted a prospective observational study to evaluate the safety and efficacy of ESPB with ropivacaine for minimizing morphine consumption in the postanesthesia care unit (PACU) after major breast cancer surgery. Then, we aimed to compare these results to an external control arm of TPVB by leveraging a historical cohort from a multicentric randomized trial, the MIRs03 study (NCT02408393) [1], which compared the efficacy of TPVB with ropivacaine to a thoracic paravertebral injection of saline in preventing acute and chronic PSP (Institut Curie, Saint-Cloud, France, Numéro EudraCT: 2014-002436-13). For our primary aim, we tested the hypothesis that ESPB increases the incidence of acute PSP when compared to TVPB. We performed a propensity score matching analysis to compare the endpoints of interest while accounting for between-study heterogeneity.

Methods

Study design

This prospective observational study was approved and registered by the Institutional Review Board (IRB) of the Curie Institute of Paris-Saint Cloud in December 2018. Information was provided orally, and oral consent was obtained. The IRB waived the need for written consent.

To allow the comparison to TPVB, we compared the outcomes of the ESPB cohort to those of patients in the experimental arm of the MIRs03 study [1].

Patient management

We used the same inclusion criteria as MIRs03: female patients aged 18–85 years with an American Society of Anaesthesiologists status of I to III who were admitted for mastectomy with or without axillary lymph node or sentinel lymph node dissection or partial mastectomy with axillary lymph node dissection.

The exclusion criteria were as follows: male sex; life expectancy less than 2 years; active malignant disease; pregnancy; breastfeeding; bilateral surgery; ipsilateral breast surgery in the past 3 years; chronic pain; allergy to local anaesthetics (LA), steroids and morphine; reported history of substance abuse; local skin inflammation at the puncture area; and inability to comply with the protocol for any reason.

Procedure

The medical protocol of the ESPB study was the same as that of MIRs03 and is summarized in Fig 1.

No premedication was given.

Upon arrival at the PACU, ECG, NIBP and SpO_2 were installed, and oxygen (2 $l.min^{-1}$) was delivered. Patients were placed in the lateral position and received target-controlled infusion of remifentanil with a targeted 2 $ng.ml^{-1}$ effect-site concentration.

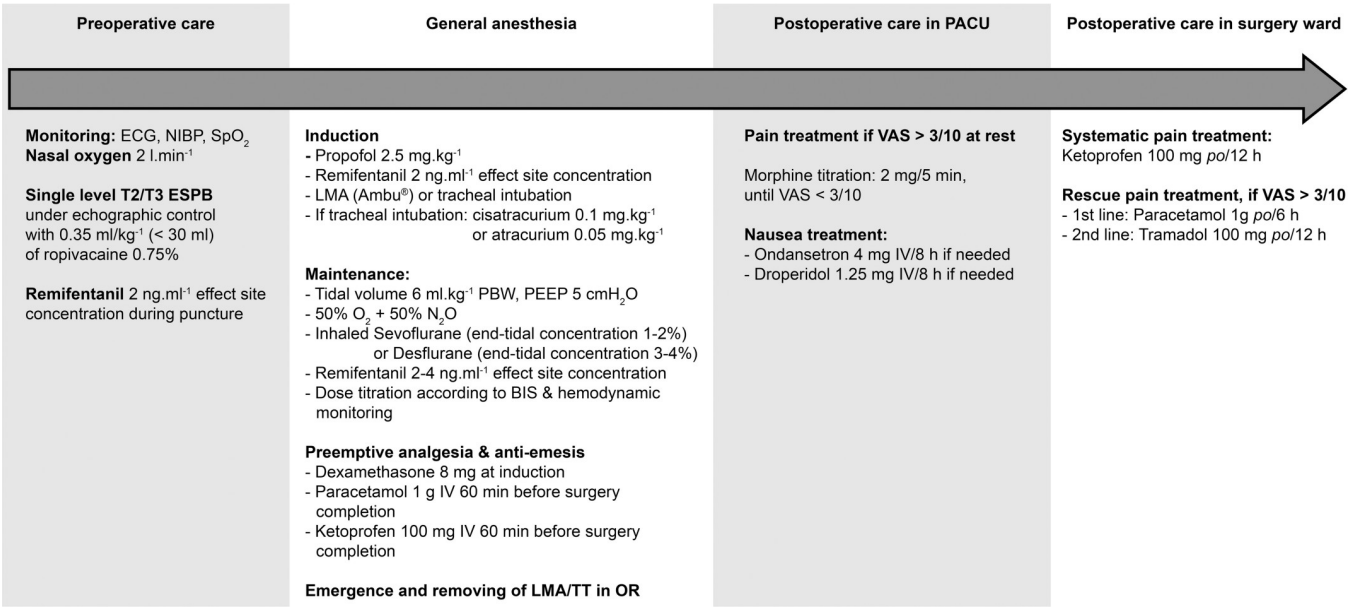


Fig 1. Study protocol.

https://doi.org/10.1371/journal.pone.0279648.g001

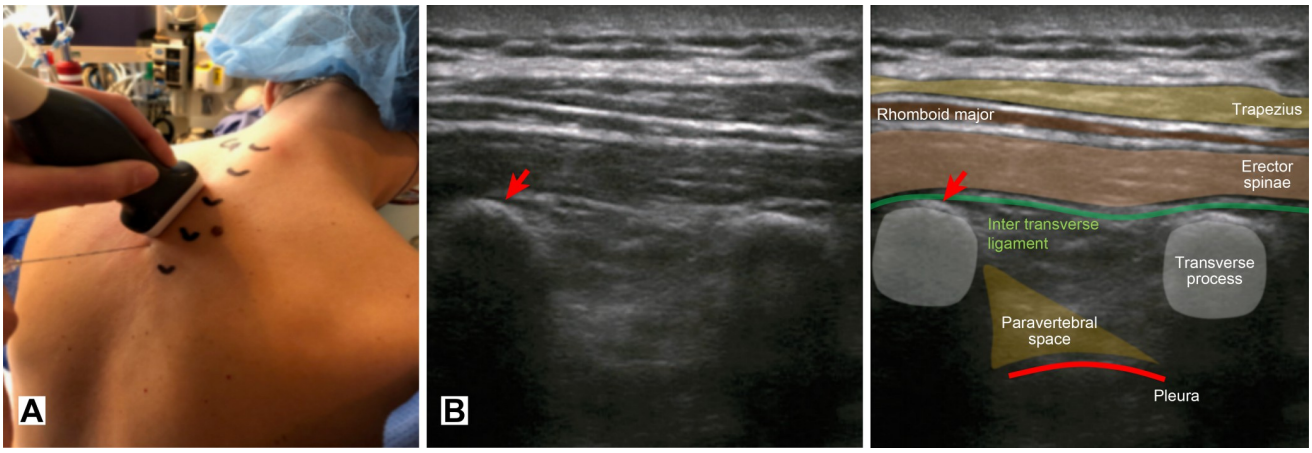


Fig 2. Technique description. *A*. Probe and needle placement. *B*. Echoanatomy and injection site (red arrow).

https://doi.org/10.1371/journal.pone.0279648.g002

Under aseptic conditions, single-level T2 or T3 (T2/T3) ESPB was placed by senior clinicians with significant experience in thoracic wall blocks. The probe (Model Alpinion E-cube i7 with a 2–5 MHz ultrasound probe linear array L3-8H) was placed in the parasagittal plane at a 90-degree angle to the transverse process after determining the T2 and T3 transverse processes by ultrasound. The needle (22-gauge 80-mm Pajunk, SonoTAP) was advanced in the plane of the ultrasound (US) beam with the bevel oriented in a cranial direction (Fig 2). When the needle tip was positioned between the erector spinae muscle and transverse process, a hydrodissection was carried out with 1–3 ml of saline solution to confirm erector spinae muscle fascia plane dissection. Then, 0.35 $ml.kg^{-1}$ ropivacaine 0.75% without exceeding 30 ml was injected (Fig 2).

During the MIRs03 trial, the TPVB was performed as follows: the patients were placed in the lateral position on the opposite side from surgery, and remifentanil administration was started with an IV targeted effect-site concentration objective to reach a concentration of 2 $ng.ml^{-1}$. The second thoracic paravertebral space (T2) was scanned by ultrasonography (Model Alpinion E-cube i7 [Alpinion Medical Systems, Korea]) with a 2- to 5-MHz ultrasound probe (linear array L3-8H). The probe was positioned on the transverse plane against the spinal process. Under aseptic conditions, a 22-gauge 80-mm needle (SonoTAP [Pajunk, Germany]) was advanced in an "in-plane" direction towards the paravertebral space, immediately above the pleura and below the costotransverse ligament. The position of the needle was confirmed by the descent of the pleura when injecting 2 to 3 ml of saline solution for hydrolocalization. Then, 0.35 $ml.kg^{-1}$ ropivacaine 0.75% was injected with intermittent negative aspiration tests every 5 ml, without exceeding a total of 30 ml or an equivalent volume of saline. Immediately after the paravertebral block injection procedure was completed in the preoperative holding area, remifentanil injection was discontinued, and the patients were transferred to the operating room 30 min later [1].

The anaesthesia management is detailed in Fig 1. After completion of the surgery, all patients were awake, breathed spontaneously and transferred to the PACU.

PSP intensity at rest and upon elevation of the arm ipsilateral to the surgery was measured upon arrival in the PACU and then every 30 minutes during the first 2 hours and every 6 hours for the first 24 hours using a VAS ranging from 0 (no pain at all) to 10 (worst imaginable pain). In the case of a resting VAS $>$ 3/10 in the PACU, intravenous morphine titration was administered using boluses of 2 mg every 5 minutes (no upper limit of dose) until the VAS

dropped $\leq$ 3/10. All patients stayed at least 2 h, and then they were allowed to leave the PACU if VAS was $\leq$ 3/10 for 30 min and the modified Aldrete Score reached at least 9. The PSP management is detailed in Fig 1.

The concentration of ropivacaine, the injected volume, the sedation and general anaesthesia protocol, the acute PSP management protocol and the postoperative nausea and vomiting (PONV) management protocol were the same as those performed with TPVB in the MIRs03 study.

Data collection

The recorded data included age, weight and height, the type of surgery, the injected volume and dose of ropivacaine, the occurrence of ESPB-related complications, the rest and mobilization VAS measured during the hospital stay and 24 hours after surgery, the need for rescue morphine, and the overall morphine dose (mg) administered in the PACU.

The same data on TPVB were collected during the MIRs03 study. The consent given for the MIRs03 study included the reuse of data.

Outcomes

The primary outcome was the percentage of patients who needed morphine titration in the PACU. The secondary outcomes were the total dose of morphine in the PACU, the incidence of RA complications, and discontinuous VAS measurement trends at rest and at mobilization during the first 24 hours after surgery.

Statistical analysis

Evidence shows that approximately 25% of patients required morphine titration after TPVB [7]. The sample size was calculated based on the accuracy of the estimate of the efficacy. For an expected rate of patients requiring morphine titration of 50%, the inclusion of 106 patients produces a two-sided 95% confidence interval with a width equal to 20%.

Baseline and outcome comparisons were performed by chi-square or Fisher's exact test for categorical variables and Student's t test for continuous variables. The ESPB and TPVB cohorts were compared for age, body mass index (BMI), and surgery type, which are potential prognostic factors of acute PSP [28].

To balance the patient characteristics between the ESPB and TPVB cohorts, we conducted a propensity score matching analysis. The propensity scores were estimated by a multivariable logistic regression model in which the probability of receiving the intervention (ESPB vs. TPVB) was regressed conditional on age, BMI, and surgery type.

Patients were matched on the logit of the propensity score using a calliper of width equal to 0.2 of the standard deviation of the logit of the estimated propensity scores [29]. Matching was performed without replacement (i.e., each subject was available for matching only once) in a greedy manner (i.e., at each step in the matching process, the nearest TPVB subject was selected for matching to the given ESPB subject). The balance of covariates between the two arms was checked using standardized mean differences (SMDs) before and after matching. A standardized mean difference of less than 0.1 is considered to indicate a negligible difference in the mean or prevalence of a covariate between groups [29].

The risk difference in acute PSP with a 95% confidence interval was estimated as the difference between the probability of receiving morphine titration of TPVB patients and that of ESPB patients in the matched sample. The standard errors were estimated using cluster-robust standard errors to account for pair membership [30].

In addition, we built a random forest model in addition to the logistic regression model to assess the sensitivity of the matching result to the propensity score estimation [31]. The random forest model was constructed using the intervention (ESPB vs. TPVB) as the output and the baseline characteristics as inputs. Matching was performed with the same parameters as described above.

Comparisons of patient characteristics of ESPB placement details according to the need for morphine titration were performed with the Mann–Whitney or Student's t test after testing for normality with the Shapiro–Wilk test.

The outcomes and baseline characteristics of patients were compared according to the injected volume ($<$25 ml vs. $\geq$25 ml).

All tests were two-sided. A P value of less than 0.05 was considered to indicate statistical significance. All analyses were conducted using R statistical software, version 4.0.2 (R Foundation for Statistical Computing, Vienna, Austria).

Results

In the ESPB cohort, 120 patients were enrolled between December 2018 and August 2019, and 102 patients were included in the analysis. The reasons for secondary exclusion of the 18 patients are detailed in Fig 3.

As there was no significant between-centre difference in the incidence of morphine consumption in the MIRs03 study (S4 Table), data from all 178 patients in the experimental arm were employed, of which 13 patients were excluded because of missing data on morphine consumption.

The study sample consisted of 102 ESPB patients and 165 TPVB patients. There were statistically significant differences in baseline characteristics regarding breast surgery type (ESPB patients underwent more total mastectomies, $p<0.001$ and SMD > 0.1) and BMI (ESPB patients had lower BMI, SMD > 0.1).

Propensity score matching formed 94 matched pairs, which means that 94 of 102 ESPB patients were matched with a TPVB patient. Eight ESPB patients who received axillary lymph node dissection were thus excluded from further analysis. The distributions of the estimated propensity scores by logistic regression are presented in the Supplementary Materials (S2 Fig).

The baseline characteristics of ESPB and TPVB patients in the propensity score-matched sample are described in Table 1. The mean and prevalence of continuous and categorical variables were very similar between the two groups (all SMDs < 0.1).

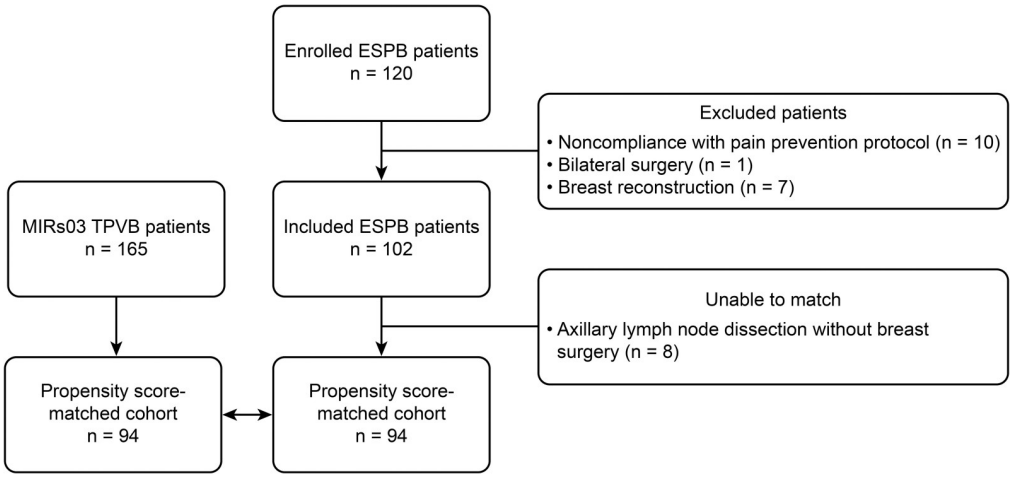


Fig 3. Flow chart.

https://doi.org/10.1371/journal.pone.0279648.g003

Table 1. Baseline characteristics of patients before and after propensity score matching.

	All patients				Propensity score-matched patients			
	ESPB (n = 102)	**TPVB (n = 165)**	***P***	**SMD**	**ESPB (n = 94)**	**TPVB (n = 94)**	***P***	**SMD**
Age (years), mean (SD)	56.5 (12.9)	57.3 (14.1)	0.65	0.057	56.0 (12.7)	55.2 (14.3)	0.72	0.053
Weight (kg), mean (SD)	67.20 (13.83)	68.85 (13.87)	0.35	0.119	67.05 (14.08)	67.65 (13.59)	0.77	0.043
BMI (kg/m^2), mean (SD)	25.2 (5.0)	25.8 (5.0)	0.34	0.122	25.1 (5.1)	24.9 (4.5)	0.78	0.041
Surgery, n (%)			< 0.001	0.466			0.71	0.080
Mastectomy	74 (72.5)	143 (86.7)			74 (78.7)	77 (81.9)		
Tumorectomy	20 (19.6)	22 (13.3)			20 (21.3)	17 (18.1)		
Axillary lymph node dissection	8 (7.8)	0 (0.0)			0 (0.0)	0 (0.0)		

https://doi.org/10.1371/journal.pone.0279648.t001

The primary endpoint of this study was the effect of ESPB on the need for morphine titration after breast surgery in the PACU (indicated in both cohorts when VAS > 3). In the propensity score-matched sample, the percentage of patients who required morphine titration was significantly higher in the ESPB group than in the TPVB group (74.5% vs. 41.5%, p<0.001). The observed difference between the two groups was 33.0% (95% confidence interval [CI] 19.3%, 46.7%).

Regarding secondary outcomes, as shown in Table 2, among all the propensity score-matched patients, the overall morphine dose was significantly higher in the ESPB group than in the TPVB group (3.7 mg vs. 2.2 mg, p = 0.02). Among those who received morphine titration, there was no difference in the morphine dose (5.1 mg vs. 6.1 mg, p = 0.07).

The mean injected volume and dose of ropivacaine in the ESPB and TVPB groups were 22.4 ml ± 4.2 vs. 23.3 ml ± 3.6 (p = 0.03) and 166 mg ± 35 vs. 174 mg ± (p = 0.07), respectively.

In the ESPB cohort, the VAS score was reported every 30 minutes in the PACU during the first 2 hours and then every 6 hours during 48 hours; the highest score was reported 30 minutes after surgery (mean VAS 3.9 ± 2.2) (S1 Fig).

No ESPB-related complications were observed. In the TPVB cohort, 5 cases of Claude Bernard Horner syndrome, 1 case of nausea and 1 case of refractory hypotension during surgery were observed (p = 0.001).

There was no significant difference in the incidence of the need for morphine between the LA volume <25 ml group and the LA volume ≥25 ml group (70.0% vs. 81.0%, p = 0.32; S2 Table).

In the sensitivity analysis, matching of propensity scores estimated by the random forest model obtained consistent results with those of the logistic regression model (S3 Fig).

Discussion

Our study is one of the largest clinical studies in which the efficacy and safety of ESPB in preventing acute PSP after major breast cancer surgery are evaluated [25,32].

Table 2. Primary and secondary outcomes in propensity score-matched patients.

	ESPB (n = 94)	**TPVB (n = 94)**	***P***
Need for morphine titration, n (%)	70 (74.5)	39 (41.5)	< 0.001
Morphine dose (mg), mean (SD)	3.7 (3.3)	2.2 (3.2)	0.02
RA placement complication incidence, n (%)	0 (0)	7 (7.4)	0.001

https://doi.org/10.1371/journal.pone.0279648.t002

Regarding efficacy, ESPB was less likely to prevent morphine consumption in the PACU than TPVB, with an observed difference of 33%. The incidence of morphine titration was as high as 74.5% after ESPB, and the overall morphine dose was significantly higher in the ESPB group than in the TPVB group (3.7 mg vs. 2.2 mg, p = 0.02).

Although there are no placebo-controlled studies evaluating the effect of ESPB on preventing acute PSP after breast surgery, many controlled studies comparing ESPB to standard care were conducted and showed that both TPVB and ESPB were superior to their control groups. Because ESPB seems to be safer than TPVB and takes less time for novice practitioners to learn [12], many authors have proposed ESPB as a standard-of-care RA for breast surgery.

Of interest, there are some reports suggesting the efficacy of ESPB in preventing acute PSP in this indication; some of them are randomized. In three meta-analyses, researchers compared the effect of ESPB to that of TPVB, but 2 also included thoracic surgery patients [25,27]. One meta-analysis included patients undergoing total mastectomies and found no statistically significant difference in morphine consumption at 24 hours [24]. In a more recently published randomized study, researchers were unable to demonstrate the noninferiority of ESPB to TPVB in minor breast surgery [33]. All these studies were conducted on small groups of patients.

The diffusion process of LAs to the paravertebral space has been shown to be impacted by the injected volume [10,11], but it seems to be inconsistent [34].

In the present T2/T3 ESPB evaluation, the median LA-injected volume was $>$ 20 ml, which is quite a large volume when compared to other studies. Regarding the impact of the injected volume, receiving more or less than 25 ml of LAs did not influence the morphine titration incidence (70% vs. 81%, p = 0.32). Additionally, the injected volume was similar between ESPB patients who received and those who did not receive morphine titration. Rather than testing fractionated volumes of LAs at multiple levels, we decided to use a single-site large LA volume injected at T2/T3 to reinforce a possible volume effect allowing sensory nerve roots arising from T2 to be blocked.

Moreover, the LA concentration seems to matter [17], suggesting that a large volume and high concentration should be used to increase the probability of efficacy. In this setting and considering its rapid and extensive rate of absorption [35], safety remains to be demonstrated.

With no complications reported, US-guided ESPB placement seems to be a safe technique. However, regarding the low rate of RA technique complications, including TPVB, our sample size may be insufficient to exclude the possibility of rare complications.

Overall, there are few studies comparing ESPB to TPVB, and the results are conflicting. These differences may be attributed to several factors. First, regarding the type of surgery, we only studied the analgesic effect following major breast surgery (e.g., mastectomy with or without axillary node dissection), rather than that following minor breast surgery (e.g., lumpectomy and partial mastectomy). Second, the pain treatment strategy in the PACU is as follows: the threshold to trigger morphine titration in our centre is a VAS score $\geq$ 3, while some authors used a VAS score $\geq$4 and others used patient-controlled analgesia with or without continuous infusion of opioids. Third, regarding the concentration and volume of the LAs used, we used ropivacaine 0.75%, and one may hypothesize that using a higher volume of a solution with a lower concentration may change the results.

Different limitations in this work should be noted. First, this is an observational cohort study compared with a historical group. When involving historical data, between-study differences can be a major concern [36]. In our study, the identical design (same eligibility criteria and protocol for perioperative management) of the ESPB and TPVB trials is the main advantage supporting comparable patient characteristics, intervention effects, and outcome measurements. In addition, we balanced three important prognostic factors of acute PSP (age,

BMI and breast surgery type) between the groups by conducting a propensity score matching analysis. To the best of our knowledge, there were no remaining systematic differences in the baseline covariates that could be prognostic of acute PSP in propensity score-matched subjects. The completion dates of the two studies were 2018 and 2019, so we assumed that there was no substantial evolution of clinical practice. In addition, we obtained consistent results when matching propensity scores estimated by the logistic regression model and random forest model, which showed the robustness of our conclusion. On the other hand, unmeasured confounders in observational studies may cause bias. For instance, although the care protocols used in the ESPB cohort were identical to those used in the MIRs03 study, the multicentre nature of the latter could theoretically generate heterogeneity of practice. However, in the MIRs03 study, the centre had no impact on morphine consumption (S4 Table). Of note, in the MIRs03 study, patients were recruited from March 27, 2015, to June 3, 2018, and recruitment for the ESPB cohort began in December 2018. Thus, from December 2018 to August 2019, all patients admitted to our centre for major breast surgery were treated with ESPB.

Second, our primary endpoint, the incidence of the need for morphine consumption in the PACU, can be discussed. Such a criterion allows us to evaluate the effectiveness of the block in preventing low early postoperative pain peaks. To completely understand the effects of this block, other parameters could be evaluated, such as the late consumption of analgesics or late mobilization. Indeed, some authors have hypothesized a delayed diffusion of the local anaesthetic from the injection zone to the paravertebral space with spontaneous respiratory movements, which could be responsible for delayed efficiency [37]. Next, in this study, we did not evaluate functional criteria or patient satisfaction. Finally, the other RA techniques used in breast surgery, such as PECs or serratus blocks, were not evaluated in this study.

These data suggest that ESPB should not be proposed as the first-line treatment over TPVB for the prevention of acute low-peak PSP after major breast cancer surgery. However, because of these limitations, a randomized trial is necessary to confirm these results. Hence, we are conducting a multicentric double-blind randomized trial to test the non-inferiority of ESPB compared to TPBV (ER-One, NCT04827030). The process of patient inclusion has already started, and the estimated completion date is August 2023.

Conclusions

In this comparative study using a propensity score matching analysis with a historical arm, US-guided ESPB at the T2/T3 level was not effective in preventing morphine consumption in the PACU after major breast surgery compared with TPVB.

Despite its easy implementation, the use of ESPB as the standard of care for radical breast cancer surgery is not justified over TPVB.

Supporting information

S1 Fig. VAS boxplot. Both rest and mobilization VAS peaks were encountered at 30 min of PACU stay.
(DOCX)

S2 Fig. Distribution of the estimated propensity scores using a logistic regression model. The overlapping area of propensity scores of the two groups implies that there are patients who share similar propensity scores and can thus be considered matched pairs.
(DOCX)

S3 Fig. Distribution of estimated propensity scores using the random forest model. S3 Fig presents the distributions of the estimated propensity scores using the random forest model,

which are similar to those estimated by the logistic regression model in S2 Fig.
(DOCX)

S1 Table. Baseline characteristics and injected volume according to morphine titration need status in the ESPB cohort. There was no statistically significant difference in patient characteristics or ESPB injected volume between patients who required morphine titration and those who did not. [a]t test when the Shapiro–Wilk test and q-q plots do not reject normality. [b]Mann–Whitney test when the Shapiro–Wilk test or q-q plots reject normality.
(DOCX)

S2 Table. Baseline characteristics and outcomes according to the injected volume in the ESPB cohort. There was no significant difference in the need for morphine titration, overall morphine dosage or VAS at rest or mobilization according to BMI. VAS, visual analog scale. [a]t test when the Shapiro–Wilk test and q-q plots do not reject normality. [b]Mann–Whitney test when the Shapiro–Wilk test or q-q plots reject normality.
(DOCX)

S3 Table. Baseline characteristics and outcomes of patients matched on propensity scores estimated by the random forest model. This table shows the comparisons of the baseline characteristics and outcomes of patients matched on propensity scores estimated by the random forest model. Ninety-five out of 102 ESPB patients were matched with a TPVB patient. Seven ESPB patients who received axillary lymph node dissection but no breast surgery were excluded from matching. Across the baseline covariates, the absolute SMDs of age and BMI were below 0.1, indicating a negligible difference. The SMD of the performed surgery type was 0.169, which slightly exceeded the preset threshold of 0.1 but was lower than the value of 0.466 before matching. The matching process created two groups of patients with more comparable covariates. The percentage of patients who required morphine titration was significantly higher in the ESPB group than in the TPVB group (74.7% vs. 38.9%, $p<0.001$). The observed difference between the two groups was 35.8% (95% CI [22.7%, 48.9%]). Among the patients who received morphine titration, the overall morphine doses were similar between the two groups (5.1 ml vs. 5.8 ml, $p = 0.14$). The results of propensity score matching analysis with the random forest model are consistent with those of the logistic regression model.
(DOCX)

S4 Table. Incidence of morphine titration among centers in the MIRs03 study. [a] *Kruskal-Wallis test*. S4 Table shows the number of patients on the experimental arm who received postoperative morphine titration at the five centers in the MIRs03 study. There was no significant difference in the incidences of morphine consumption between centers in the MIRs03 study.
(DOCX)

Acknowledgments

The authors would like to extend their sincere thanks to the entire anaesthesiology team, the PACU nurse team and the Department of Surgery of Institut Curie Saint Cloud.

Author Contributions

Conceptualization: Aline Albi-Feldzer.

Data curation: Antoine Premachandra, Aline Albi-Feldzer.

Formal analysis: Antoine Premachandra, Xiaomeng Wang, Aurélien Latouche, Aline Albi-Feldzer.

Methodology: Xiaomeng Wang, Aline Albi-Feldzer.

Project administration: Aline Albi-Feldzer.

Software: Xiaomeng Wang.

Supervision: Mary Saad, Aurélien Latouche, Aline Albi-Feldzer.

Validation: Mary Saad, Sahar Moussawy, Roman Rouzier, Aurélien Latouche, Aline Albi-Feldzer.

Writing – original draft: Antoine Premachandra, Xiaomeng Wang, Aline Albi-Feldzer.

Writing – review & editing: Antoine Premachandra, Xiaomeng Wang, Mary Saad, Sahar Moussawy, Roman Rouzier, Aurélien Latouche, Aline Albi-Feldzer.

References

1. Albi-Feldzer A, Dureau S, Ghimouz A, Raft J, Soubirou JL, Gayraud G, et al. Preoperative Paravertebral Block and Chronic Pain after Breast Cancer Surgery: A Double-blind Randomized Trial. Anesthesiology. 1 déc 2021; 135(6):1091-103. https://doi.org/10.1097/ALN.0000000000003989 PMID: 34618889
2. Gärtner R, Jensen MB, Nielsen J, Ewertz M, Kroman N, Kehlet H. Prevalence of and Factors Associated With Persistent Pain Following Breast Cancer Surgery. JAMA. 11 nov 2009; 302(18):1985. https://doi.org/10.1001/jama.2009.1568 PMID: 19903919
3. Perkins FM, Kehlet H. Chronic Pain as an Outcome of Surgery. Anesthesiology. 1 oct 2000; 93 (4):1123-33.
4. Jacobs A, Lemoine A, Joshi GP, Van de Velde M, Bonnet F, the PROSPECT Working Group collaborators, et al. PROSPECT guideline for oncological breast surgery: a systematic review and procedure-specific postoperative pain management recommendations. Anaesthesia. mai 2020; 75(5):664-73.
5. Zinboonyahgoon N, Vlassakov K, Lirk P, Spivey T, King T, Dominici L, et al. Benefit of regional anaesthesia on postoperative pain following mastectomy: the influence of catastrophising. Br J Anaesth. août 2019; 123(2):e293-302. https://doi.org/10.1016/j.bja.2019.01.041 PMID: 31331591
6. Santonastaso DP, de Chiara A, Russo E, Musetti G, Lucchi L, Sibilio A, et al. Single shot ultrasound-guided thoracic paravertebral block for opioid-free radical mastectomy: a prospective observational study. J Pain Res. 2019; 12:2701-8. https://doi.org/10.2147/JPR.S211944 PMID: 31571975
7. Qian B, Fu S, Yao Y, Lin D, Huang L. Preoperative ultrasound-guided multilevel paravertebral blocks reduce the incidence of postmastectomy chronic pain: a double-blind, placebo-controlled randomized trial. JPR. févr 2019;Volume 12:597-603. https://doi.org/10.2147/JPR.S190201 PMID: 30787636
8. Pace MM, Sharma B, Anderson-Dam J, Fleischmann K, Warren L, Stefanovich P. Ultrasound-Guided Thoracic Paravertebral Blockade: A Retrospective Study of the Incidence of Complications. Anesthesia & Analgesia. avr 2016; 122(4):1186-91. https://doi.org/10.1213/ANE.0000000000001117 PMID: 26756911
9. Forero M, Adhikary SD, Lopez H, Tsui C, Chin KJ. The Erector Spinae Plane Block: A Novel Analgesic Technique in Thoracic Neuropathic Pain. Regional Anesthesia and Pain Medicine. 2016; 41(5):621-7. https://doi.org/10.1097/AAP.0000000000000451 PMID: 27501016
10. Schwartzmann A, Peng P, Maciel MA, Forero M. Mechanism of the erector spinae plane block: insights from a magnetic resonance imaging study. Can J Anesth/J Can Anesth. oct 2018; 65(10):1165-6. https://doi.org/10.1007/s12630-018-1187-y PMID: 30076575
11. Damjanovska M, Stopar Pintaric T, Cvetko E, Vlassakov K. The ultrasound-guided retrolaminar block: volume-dependent injectate distribution. JPR. févr 2018;Volume 11:293-9. https://doi.org/10.2147/JPR.S153660 PMID: 29445296
12. Moustafa M, Alabd A, Ahmed AM, Deghidy E. Erector spinae versus paravertebral plane blocks in modified radical mastectomy: Randomised comparative study of the technique success rate among novice anaesthesiologists. Indian J Anaesth. 2020; 64(1):49. https://doi.org/10.4103/ija.IJA_536_19 PMID: 32001909
13. El Ghamry M, Amer A. Role of erector spinae plane block versus paravertebral block in pain control after modified radical mastectomy. A prospective randomised trial. Indian J Anaesth. 2019; 63 (12):1008. https://doi.org/10.4103/ija.IJA_310_19 PMID: 31879425

14. Tsui BCH, Fonseca A, Munshey F, McFadyen G, Caruso TJ. The erector spinae plane (ESP) block: A pooled review of 242 cases. Journal of Clinical Anesthesia. mars 2019; 53:29-34. https://doi.org/10.1016/j.jclinane.2018.09.036 PMID: 30292068

15. Nagaraja P, Ragavendran S, Singh N, Asai O, Bhavya G, Manjunath N, et al. Comparison of continuous thoracic epidural analgesia with bilateral erector spinae plane block for perioperative pain management in cardiac surgery. Ann Card Anaesth. 2018; 21(3):323. https://doi.org/10.4103/aca.ACA_16_18 PMID: 30052229

16. Gürkan Y, Aksu C, Kuş A, Yörükoğlu UH. Erector spinae plane block and thoracic paravertebral block for breast surgery compared to IV-morphine: A randomized controlled trial. Journal of Clinical Anesthesia. févr 2020; 59:84-8. https://doi.org/10.1016/j.jclinane.2019.06.036 PMID: 31280100

17. Altıparmak B, Korkmaz Toker M, Uysal Aİ, Gümüş Demirbilek S. Comparison of the efficacy of erector spinae plane block performed with different concentrations of bupivacaine on postoperative analgesia after mastectomy surgery: ramdomized, prospective, double blinded trial. BMC Anesthesiol. déc 2019; 19(1):31. https://doi.org/10.1186/s12871-019-0700-3 PMID: 30832580

18. Kim JA, Wahlster S, LaBuzetta JN, Nobleza COS, Johnson NJ, Rubinos C, et al. Focused Management of Patients With Severe Acute Brain Injury and ARDS. Chest. janv 2022; 161(1):140-51. https://doi.org/10.1016/j.chest.2021.08.066 PMID: 34506794

19. Hussain N, Brull R, Noble J, Weaver T, Essandoh M, McCartney CJ, et al. Statistically significant but clinically unimportant: a systematic review and meta-analysis of the analgesic benefits of erector spinae plane block following breast cancer surgery. Reg Anesth Pain Med. janv 2021; 46(1):3-12. https://doi.org/10.1136/rapm-2020-101917 PMID: 33168651

20. Elewa AM, Faisal M, Sjöberg F, Abuelnaga ME. Comparison between erector spinae plane block and paravertebral block regarding postoperative analgesic consumption following breast surgery: a randomized controlled study. BMC Anesthesiol. déc 2022; 22(1):189. https://doi.org/10.1186/s12871-022-01724-3 PMID: 35717148

21. Singh S, Kumar G, Akhileshwar. Ultrasound-guided erector spinae plane block for postoperative analgesia in modified radical mastectomy: A randomised control study. Indian J Anaesth. 2019; 63(3):200. https://doi.org/10.4103/ija.IJA_758_18 PMID: 30988534

22. Gürkan Y, Aksu C, Kuş A, Yörükoğlu UH, Kılıç CT. Ultrasound guided erector spinae plane block reduces postoperative opioid consumption following breast surgery: A randomized controlled study. Journal of Clinical Anesthesia. nov 2018; 50:65-8. https://doi.org/10.1016/j.jclinane.2018.06.033 PMID: 29980005

23. Yao Y, Li H, He Q, Chen T, Wang Y, Zheng X. Efficacy of ultrasound-guided erector spinae plane block on postoperative quality of recovery and analgesia after modified radical mastectomy: randomized controlled trial. Reg Anesth Pain Med. janv 2020; 45(1):5-9.

24. Leong RW, Tan ESJ, Wong SN, Tan KH, Liu CW. Efficacy of erector spinae plane block for analgesia in breast surgery: a systematic review and meta-analysis. Anaesthesia. mars 2021; 76(3):404-13. https://doi.org/10.1111/anae.15164 PMID: 32609389

25. Huang W, Wang W, Xie W, Chen Z, Liu Y. Erector spinae plane block for postoperative analgesia in breast and thoracic surgery: A systematic review and meta-analysis. Journal of Clinical Anesthesia. nov 2020; 66:109900. https://doi.org/10.1016/j.jclinane.2020.109900 PMID: 32502778

26. Hung KC, Liao SW, Sun CK. Comparable analgesic efficacy between erector spinae plane and thoracic paravertebral blocks for breast and thoracic surgeries? Journal of Clinical Anesthesia. août 2021; 71:110200. https://doi.org/10.1016/j.jclinane.2021.110200 PMID: 33609853

27. Xiong C, Han C, Zhao D, Peng W, Xu D, Lan Z. Postoperative analgesic effects of paravertebral block versus erector spinae plane block for thoracic and breast surgery: A meta-analysis. Farag E, éditeur. PLoS ONE. 25 août 2021; 16(8):e0256611. https://doi.org/10.1371/journal.pone.0256611 PMID: 34432822

28. Katz J, Poleshuck EL, Andrus CH, Hogan LA, Jung BF, Kulick DI, et al. Risk factors for acute pain and its persistence following breast cancer surgery. Pain. déc 2005; 119(1-3):16-25. https://doi.org/10.1016/j.pain.2005.09.008 PMID: 16298063

29. Austin PC. An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. Multivariate Behavioral Research. 31 mai 2011; 46(3):399-424. https://doi.org/10.1080/00273171.2011.568786 PMID: 21818162

30. Abadie A, Spiess J. Robust Post-Matching Inference. Journal of the American Statistical Association. 14 janv 2021;1-13.

31. Lee BK, Lessler J, Stuart EA. Improving propensity score weighting using machine learning. Statist Med. 10 févr 2010; 29(3):337-46. https://doi.org/10.1002/sim.3782 PMID: 19960510

32. Zhang Y, Liu T, Zhou Y, Yu Y, Chen G. Analgesic efficacy and safety of erector spinae plane block in breast cancer surgery: a systematic review and meta-analysis. BMC Anesthesiol. déc 2021; 21(1):59. https://doi.org/10.1186/s12871-021-01277-x PMID: 33610172

33. Swisher MW, Wallace AM, Sztain JF, Said ET, Khatibi B, Abanobi M, et al. Erector spinae plane versus paravertebral nerve blocks for postoperative analgesia after breast surgery: a randomized clinical trial. Reg Anesth Pain Med. avr 2020; 45(4):260-6. https://doi.org/10.1136/rapm-2019-101013 PMID: 31969443

34. Gadsden J, Gonzales J, Chen A. Relationship between injectate volume and disposition in lumbar erector spinae plane block: a cadaveric study. In: Presented as an Abstract in the 46th annual Regional Anesthesiology & Acute Pain Medicine Meeting. Lake Buena Vista, USA.

35. De Cassai A, Bonanno C, Padrini R, Geraldini F, Boscolo A, Navalesi P, et al. Pharmacokinetics of lidocaine after bilateral ESP block. Reg Anesth Pain Med. janv 2021; 46(1):86-9. https://doi.org/10.1136/rapm-2020-101718 PMID: 32868484

36. Hatswell A, Freemantle N, Baio G, Lesaffre E, van Rosmalen J. Summarising salient information on historical controls: A structured assessment of validity and comparability across studies. Clinical Trials. déc 2020; 17(6):607-16. https://doi.org/10.1177/1740774520944855 PMID: 32957804

37. Schwartzmann A, Peng P, Maciel MA, Alcarraz P, Gonzalez X, Forero M. A magnetic resonance imaging study of local anesthetic spread in patients receiving an erector spinae plane block. Can J Anaesth. août 2020; 67(8):942-8.

Appendix B

Published article 2

Current Perspectives for External Control Arms in Oncology Clinical Trials: Analysis of EMA approvals 2016-2021

Xiaomeng Wang[a,b], Flavio Dormont[b], Christelle Lorenzato[b], Aurélien Latouche[a,c], Ramon Hernandez[b], Roman Rouzier[d]

[a] *INSERM, U900, Institut Curie, PSL Research University, Saint-Cloud, France*
[b] *Department of Research and Development, Sanofi, Chilly-Mazarin, France*
[c] *Conservatoire National des Arts et Métiers, Paris, France*
[d] *Department of Surgical Oncology, Centre François Baclesse, Caen, France*

ABSTRACT

Leveraging external control data, especially real-world data (RWD), has drawn particular attention in recent years for facilitating oncology clinical development and regulatory decision-making. Medical regulators have published guidance on accelerating the use of RWD and external controls. However, few systematic discussions have been conducted on external controls in cancer drug submissions and regulatory feedback.

This study aimed to identify European oncology drug approvals using external control data to demonstrate clinical efficacy. We included 18 eligible submissions employing 24 external controls and then discussed the use of external control, data sources, analysis methods, and regulators' feedback.

The external controls have been actively submitted to the European Medical Agency (EMA) recently. We found that 17% of the EMA-approved cancer drugs in 2016-2021 used external controls, among which 37% of the cases leveraged RWD. However, nearly one-third of the external controls were not considered supportive evidence by EMA due to limitations regarding heterogeneous patient populations, missing outcome assessment in RWD, and inappropriate statistical analysis.

This study highlighted that proper use of external controls requires a careful assessment of clinical settings, data availability, and statistical methodology. For better use of external controls in oncology clinical trials, we recommend: prospective study designs to avoid selection bias, sufficient baseline data to ensure the comparability of study populations, consistent endpoint measurements to enable outcome comparison, robust statistical methodology for comparative analysis, and collaborative efforts of sponsors and regulators to establish regulatory frameworks.

KEYWORDS

external control; historical control; oncology; clinical trial; real-world data

1. Introduction

Randomized controlled trials (RCTs) are the gold standard for evaluating the efficacy of new treatments[1]. RCTs establish causal conclusions by randomly assigning patients to either an investigational or concurrent control treatment that usually consists of a placebo or standard of care. Nevertheless, ethical or feasibility issues can hinder the randomization process in oncology clinical trials. For example, when a promising new therapy is being studied in early-stage clinical trials for cancer with high unmet needs, enrolling patients into control arms may cause ethical issues and unwillingness to participate[2]. Besides, with the advances in molecular classification and precision medicine in oncology - further dividing patient populations into smaller groups - the number of patients available for a particular clinical trial may be insufficient to produce valid evidence [3,4]. Medical regulators acknowledge these challenges and have granted conditional approvals to submissions with single-arm trials[5,6].

External controls have been introduced as a pragmatic approach to knowledge production, which can improve the interpretation of single-arm trial results and reduce the bias due to the lack of randomization[6]. Data from historical clinical trials or real-world databases expand the use of external controls to minimize patient burden and facilitate drug development[7–9].

Recent technological advances and a dynamic policy landscape have created a fertile ground for using real-world data (RWD) to improve clinical evidence generation[10]. RWD are qualified as routinely collected data relating to patient health status and the delivery of health care other than traditional RCTs such as electronic health records (EHRs), claims, registries, or patient-generated data[11,12]. Clinical evidence regarding the usage and potential benefits or risks of a medical product derived from the analysis of RWD is then considered real-world evidence (RWE). The regulatory authorities have signaled their support for using RWD to generate clinical evidence. In 2018, the US Food and Drug Administration (FDA) published the framework for RWE underpinned by three pillars: whether RWD are fit for use, whether the study design can provide adequate evidence, and whether the study conduct meets regulatory requirements[11]. In 2019, the European Medicines Agency (EMA) published the OPTIMAL framework for RWE, consisting of operational, technical and methodological[12]. Besides, the EMA outlined the vision for establishing the value of RWE for various regulatory uses by 2025[13].

Recent evidence shows that RCT and RWE findings were not always matched despite attempts to emulate RCT design and confounder adjustment[14]. Thus, challenges remain before external controls can be an integrated part of decision-making[15]. On the other hand, researchers proposed the combination of RCTs and RWE data for clinical knowledge generation in the era of precision medicine[4]. External controls leveraging RWD are of particular interest in oncology drug development to foster patients' access to innovative therapy in the context of segmentation of tumor entities and targeted therapy [16,17]. Learning from previous external control applications can inform future studies and avoid common pitfalls. So far, however, few systematic discussions have been conducted about external controls in oncology drug development.

This study aims to systematically review the external controls in oncology clinical development and understand their impact on regulatory decision-making. Section 2 defines external control and reviews statistical considerations. Sections 3 and 4 describe the study methodology and outline the search findings. Section 5 discusses the statistical characteristics of external controls, their impact on regulatory decision-making, and recommendations for using external controls. Finally, Section 6 concludes the study.

2. External controls and Statistical considerations

In an externally controlled trial, patients receiving the investigational treatment are compared with patients external to the study. The external control can consist of patients treated earlier than the concurrent clinical trial (historical control) or during the same period on different clinical conditions[18]. External control is also known as synthetic control, implying the control data have been selected and processed before being incorporated into the analysis. "External control", "historical control", and "synthetic control" are used interchangeably in practice[7,19–22]. This study uses "external control" to refer to data derived outside the concurrent clinical trial.

In single-arm or parallel-group trials, external controls serve as the sole comparator to provide benchmarks and contextualize the new treatment[9,21]. In RCTs with unequal randomization, external controls can augment the concurrent control to form hybrid control arms composed of internal and external control data, which allows randomizing fewer patients to the control arm[3,23–25].

Clinical data for creating external controls fall into two major categories: clinical trial data and real-world data (RWD). The choice of data sources depends on the research questions. For example, for the endpoints of interest measured

in a specialized manner such as key biomarker testing, previous clinical trial data can be appropriate. On the other hand, RWD might be better suited for cancers with limited experience from clinical trials.

External controls are subject to systematic biases due to the absence of randomization. The validity of external control lies in its exchangeability with the internal patients, for which Pocock proposed six evaluation criteria in 1976: (1) the same defined standard treatment; (2) the same patient eligibility; (3) the same treatment evaluation; (4) comparable distributions of essential patient characteristics; (5) performed in the same organization; (6) no other observed confounders[22]. While these criteria are stringent in non-randomized clinical conditions, applying them to the most significant degree possible can minimize the risk of substantial biases. The FDA cited these criteria in the statistical review for drug approvals[26].

Statistical analysis methods have been developed to mitigate the biases of incorporating external controls. Frequentist approaches ensure between-group comparability of baseline characteristics by following two main steps: (1) estimate a balance score (e.g., propensity score, Mahalanobis distance) using a selection of covariates that can impact the treatment assignment and outcome, (2) employ the balance score to obtain comparable external and internal cohorts through analysis such as matching, inverse probability weighting, and covariate adjustment[27,28]. Bayesian approaches account for the outcome heterogeneity by incorporating and discounting the external control data into the new clinical trial, using power prior, commensurate prior, or meta-analytic prior[29,30].

3. Search methodology and data extraction

We searched European public assessment reports (EPARs) for human medicines (https://www.ema.europa.eu/en/medicines) to identify regulatory submissions referring to external control data in clinical efficacy. We included original marketing applications for cancer drugs in 2016-2021 and excluded drugs for diagnostic use only or withdrawn from the market. The search was conducted in 31 May 2022.

From January 1, 2016, to December 31, 2021, 113 cancer drugs were granted marketing authorization by EMA (Figure 1). Two drugs for diagnostic use and eight withdrawn from the market were excluded from the study. After screening the EPARs of the eligible 103 medicines, we identified 18 (17%) drug submissions using external controls. Figure 2 presents the number of screened cancer drugs and eligible external control cases.

We examined the characteristics of the drug approvals concerning: drug name, class of drug, approval year, marketing authorization holder, therapeutic area and indication. We scrutinized the clinical efficacy section in the EPAR of such approvals to extract information on pivotal studies and external controls: pivotal study design, use of external control, source of external control data, method of analysis, and EMA's decision on the external controls.

Figure 1 Selection of external control cases from cancer drug approvals

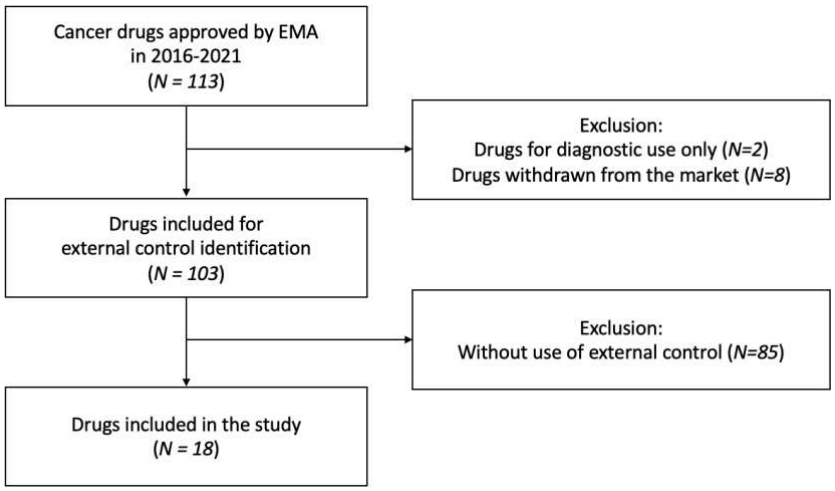
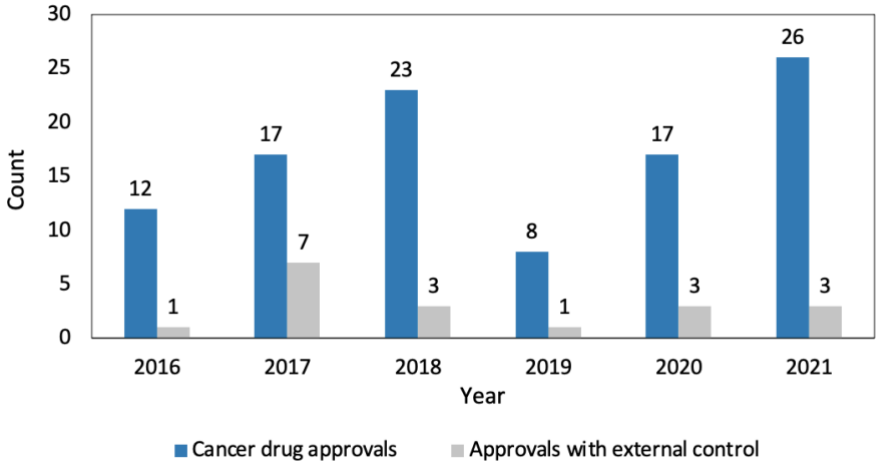


Figure 2 Summary of identification results in 2016-2021



4. Results

Table 1 lists the identified drugs and external controls. 7 (39%) of the 18 identified drugs were granted conditional approval, all of which were submitted after 2019. Note that six drug submissions have used multiple external controls for one pivotal study or have conducted pivotal studies for multiple indications; therefore, the 18 submissions leveraged 24 external controls. Though all drugs were approved by EMA, some external controls were not deemed supportive. We considered the EMA's decision "accepted" if the external controls were included as supportive evidence to demonstrate efficacy and "rejected" if the external controls were deemed inadequate for decision-making.

Table 1: Cancer drug approvals using external controls

Drug	Year	Class of drug	Therapeutic area	Conditional approval	Pivotal study design	Use of external control	Source of external control data	Method of analysis	EMA's decision
Minjuvi (tafasitamab)	2021	Monoclonal antibody	Lymphoma	Yes	Phase 2, single-arm, open-label, multicentre	Comparative efficacy analysis	RWD	Confounding adjustment	Rejected
Abecma (idecabtagene vicleucel)	2021	Autologous cellular immunotherapy	Multiple Myeloma	Yes	Phase 2, single-arm, open-label, multicentre	Comparative efficacy analysis	RWD	Confounding adjustment	Rejected
Enhertu (trastuzumab deruxtecan) - 1	2021	Antibody-drug conjugate	Breast Neoplasms	Yes	Phase 2, single-arm, open-label, multicenter, 2-part	Comparative efficacy analysis	RWD	Confounding adjustment	Rejected
Enhertu (trastuzumab deruxtecan) - 2	2021	Antibody-drug conjugate	Breast Neoplasms	Yes	Phase 2, single-arm, open-label, multicenter, 2-part	Understanding the natural history of disease	Published observational studies	Meta-analysis	Accepted
Blenrep (belantamab mafodotin)	2020	Antibody-drug conjugate	Multiple Myeloma	Yes	Phase 2, two-arm, randomized, open-label, multicentre	Historical benchmark	Published observational studies	Descriptive	Accepted
Rozlytrek (entrectinib)	2020	Kinase inhibitor	Non-Small-Cell Lung Cancer	Yes	Phase 2, single-arm, open-label, multicenter, basket study	Comparative efficacy analysis	RWD	Confounding adjustment	Rejected
Tecartus (brexucabtagene autoleucel) - 1	2020	Autologous cellular immunotherapy	Lymphoma	Yes	Phase 2, single-arm, open-label, multicentre	Historical benchmark	Published observational studies	Descriptive	Accepted
Tecartus (brexucabtagene autoleucel) - 2	2020	Autologous cellular	Lymphoma	Yes	Phase 2, single-arm, open-label, multicentre	Historical benchmark	Published observational studies	Meta-analysis	Rejected

		immunotherapy							
Libtayo (cemiplimab)	2019	Monoclonal antibody	Squamous Cell Carcinoma	Yes	Phase 2, single-arm, 3-group, multicenter	Comparative efficacy analysis	RWD	Descriptive	Accepted
Apealea (paclitaxel)	2018	Mitotic inhibitor	Ovarian Neoplasms	No	Phase 3, parallel, randomized, comparator-controlled, open-label, non-inferiority study	Defining margin of non-interiority	Historical clinical trials	Meta-analysis	Accepted
Verzenios (abemaciclib) - 1	2018	Kinase inhibitor	Breast Neoplasms	No	Phase 2, single-arm, open-label, multicentre	Historical benchmark	Unspecified	Descriptive	Accepted
Verzenios (abemaciclib) - 2	2018	Kinase inhibitor	Breast Neoplasms	No	Phase 2, single-arm, open-label, multicentre	Comparative efficacy analysis	RWD	Confounding adjustment	Rejected
Yescarta (axicabtagene ciloleucel)	2018	Autologous cellular immunotherapy	Lymphoma	No	Phase 2, single-arm, open-label, multicentre	Comparative efficacy analysis	RWD and historical clinical trials	Meta-analysis	Accepted
Bavencio (avelumabl)	2017	Monoclonal antibody	Neuroendocrine Tumors	No	Phase 2, single-arm, open-label, multicentre	Understanding the natural history of disease	RWD	Descriptive	Accepted
Qarziba (dinutuximab beta) - 1	2017	Monoclonal antibody	Neuroblastoma	No	retrospective data analysis under a compassionate use program	Comparative efficacy analysis	RWD	Descriptive	Accepted
Qarziba (dinutuximab beta) - 2	2017	Monoclonal antibody	Neuroblastoma	No	retrospective data analysis under a compassionate use program	Comparative efficacy analysis	Historical clinical trials	Descriptive	Accepted
Rydapt (midostaurin)	2017	Kinase inhibitor	Leukemia, Mastocytosis	No	Phase 2, single-arm, multicentre	Comparative efficacy analysis	Historical clinical trials	Confounding adjustment	Rejected
Tecentriq (atezolizumab) - 1	2017	Monoclonal antibody	Non-Small-Cell Lung Cancer	No	Phase 2, single-arm, multicentre	Historical benchmark	Historical clinical trials	Descriptive	Rejected
Tecentriq (atezolizumab) - 2	2017	Monoclonal antibody	Urologic Neoplasms	No	Phase 2, single-arm, multicentre, two-cohort	Historical benchmark	RWD	Confounding adjustment	Rejected
Ledaga (chlormethine)	2017	Alkylating agent	Mycosis Fungoides	No	phase 2, multicenter, randomized, comparator-controlled, third party (observer) blinded, non-inferiority study	Defining margin of non-interiority	Unspecified	Descriptive	Accepted
Blitzima (rituximab)	2017	Monoclonal antibody	Lymphoma, Leukemia	No	Phase 1, randomized, controlled, multicentre, 2-arm, parallel-group, double-blind	Comparative efficacy analysis	Historical clinical trials	Descriptive	Accepted
Truxima (rituximab) - 1	2017	Monoclonal antibody	Lymphoma, Leukemia	No	Phase 1, randomized, controlled, multicentre, 2-arm, parallel-group, double-	Comparative efficacy analysis	Historical clinical trials	Descriptive	Accepted
Truxima (rituximab) - 2	2017	Monoclonal antibody	Lymphoma, Leukemia	No	open-label, single-arm, maintenance study	Defining margin of non-interiority	Historical clinical trials	Descriptive	Accepted
Darzalex (daratumumab)	2016	Monoclonal antibody	Multiple Myeloma	No	Phase 2, open-label, multicentre, 2-arm	Historical benchmark	Published observational studies	Descriptive	Accepted

Table 2 summarizes the characteristics of external controls regarding use for clinical efficacy, data source, analysis method, and EMA's decision.

Table 2: Summary of external control characteristics

	n Total=24	%

Use of external control		
Comparative efficacy analysis	12	50%
Historical benchmark	7	29%
Defining margin of non-interiority	3	13%
Understanding nature history of disease	2	8%
Source of external control data		
Real-world data (RWD)	9	37%
Historical clinical trials	7	29%
RWD and historical clinical trials	1	4%
Published observational studies	5	21%
Unspecified	2	8%
Method of analysis		
Descriptive analysis	13	54%
Comparative analysis with confounding adjustment (matching, inverse probability weighting)	7	29%
Meta-analysis	4	17%
EMA's decision on external control		
Accepted	15	63%
Rejected	9	37%

Table 3 summarizes the EMA's decision on external controls by data sources and analysis methods.

Table 3: EMA's decision on external controls

	EMA's decision on external controls		
	Accepted	**Rejected**	**Total**
Source of external control data			
Real-world data (RWD)	3 (33%)	6 (67%)	9 (100%)
Historical clinical trials	5 (71%)	2 (29%)	7 (100%)
RWD and historical clinical trials	1 (100%)	0 (0%)	1 (100%)
Published observational studies	4 (80%)	1 (20%)	5 (100%)
Unspecified	2 (100%)	0 (0%)	2 (100%)
Method of analysis			
Descriptive	12 (92%)	1 (8%)	13 (100%)
Confounding adjustment	0 (0%)	7 (100%)	7 (100%)
Meta-analysis	3 (75%)	1 (25%)	4 (100%)
Total	15 (63%)	9 (37%)	24 (100%)

Table 4 presents the limitations cited by EMA for rejecting external controls to support clinical efficacy.

Table 4: Limitations identified by EMA on external controls

Drug	Year	Source of external control data	Method of analysis	Limitations identified by EMA
Minjuvi (tafasitamab)	2021	RWD	Confounding adjustment	Heterogeneous patient populations, differences in standard of care received during treatment, suboptimal statistical methodology
Abecma (idecabtagene vicleucel)	2021	RWD	Confounding adjustment	Selection bias of the study population, missing data of prognostic factors
Enhertu (trastuzumab deruxtecan) - 1	2021	RWD	Confounding adjustment	Selection bias of the study population, missing assessment of response, differences in the measurement of endpoint, not optimal statistical methodology
Rozlytrek (entrectinib)	2020	RWD	Confounding adjustment	Limitations of the study design, limited data
Tecartus (brexucabtagene autoleucel) - 2	2020	Published observational studies	Meta-analysis	Heterogeneous patient populations, limited information on the study design
Verzenios (abemaciclib) - 2	2018	RWD	Confounding adjustment	Heterogeneous patient populations
Rydapt (midostaurin)	2017	Historical clinical trials	Confounding adjustment	Limited information on the baseline characteristics, no correction for the time of initiation of treatment
Tecentriq (atezolizumab) - 1	2017	Historical clinical trials	Descriptive	Limited information on the determination of historical response rates
Tecentriq (atezolizumab) - 2	2017	RWD	Confounding adjustment	Heterogeneous patient populations

5. Discussion

We identified that 50% of external controls were used for comparative efficacy analysis, and 29% served as historical benchmarks in the superiority study design. Moreover, 8% of external controls provided contextual information on the natural history of disease in rare indications. Besides, 13% of external controls were included in defining the margin of non-inferiority for non-inferiority hypotheses. All external controls were employed in single-arm trials.

We did not find external controls supplementing the concurrent control arm in RCTs. Though there have been discussions on combining external with internal data in RCTs[21,31], the hybrid control arm has yet to be applied for EMA submissions. Nevertheless, the hybrid design allows one to evaluate population comparability and can be considered in future external control studies[25].

We observed that the external controls used most individual RWD (38%) and historical clinical trials (29%). 21% cited published studies, including prospective or retrospective observational studies. 8% did not specify the data source. One case (4%) employed individual RWD and historical clinical trial data. The results highlighted the broad usage of RWD in external controls in recent cancer drug approvals.

We found that 54% of the external controls used descriptive analysis and naive comparison, and 17% applied meta-analysis using aggregate or patient-level data. Besides, 29% performed comparative efficacy analysis with adjustment for confounding covariates, including matching and inverse probability weighting (based on propensity score or Mahalanobis distance) with individual patient data from RWD or historical clinical trials.

The EMA accepted 63% of the external controls for demonstrating clinical efficacy. However, Table 3 shows that only 33% of external controls using RWD were considered supportive evidence. By contrast, the majority of external controls using data from historical clinical trials (71%), both RWD and historical clinical trials (100%), published studies (80%), or unknown sources (100%) were accepted. Regarding the analysis method, the EMA accepted most external controls using descriptive analysis (92%) or meta-analysis (75%). In comparison, no case applying methods of confounding adjustment (0%) was considered supportive evidence. These results indicate that the EMA primarily considered the pivotal study results in decision-making, and various data sources or analysis methods did not mitigate the concerns about the systematic bias of external controls.

The limitations of external controls cited by EMA (Table 4) primarily concerned the Pocock criteria discussed in Section 2.2. Heterogeneity in the external and internal patients was a significant issue. It was caused by differences in the standard of care or limited information from the sponsors on baseline characteristics, especially on significant prognostic factors. The regulators also pointed out that missing or different assessment of endpoints was an essential downside of using RWD in external controls. Besides, unclear study design and selection bias hindered regulators' understanding and acceptance of external controls. Unluckily, these biases were hardly compensated through statistical methodology as the regulators deemed the sponsors applied suboptimal analysis.

To improve the use of external controls in regulatory decision-making, we first suggest a priori trial planning considering external control data. Sponsors should engage with regulators early to align the acceptability of external control for the target population, the study design, and the methodology to avoid selection bias. Secondly, we recommend evaluating external data using Pocock's criteria to confirm population comparability and outcome assessment. Thirdly, we advise robust statistical analysis with adjustment for confounders to reduce potential bias. For example, use doubly robust estimation for propensity score methods because propensity score estimation is subject to model misspecification[32,33]. Besides, using individual RWD to emulate target trials can help ensure fit-for-purpose data and confounding control[34]. Fourthly, we suggest performing sensitivity analysis to validate data quality and study results. Finally, we encourage sponsors to work with regulators to develop regulatory guidance on external controls for study design, data selection, and analysis plans.

Our study was conducted in a systematic manner with in-depth extractions of external control characteristics and objective discussions on the issues of external controls. Despite this, our study was subject to several limitations. First, we limited our scope with a prespecified inclusion period of 2016-2021 under the presumption that external control using RWD is a recently emerging research interest in oncology. Second, our data extraction was limited to the EMA's public assessment reports rather than the sponsors' original submission documents. As a result, some details on external controls may have been omitted. Finally, we did not scrutinize the impact of external control on regulatory pathways, such as conditional marketing authorization. Besides, the prespecified period of 2016-2021

may limit the assessment of post-authorization changes. It can be interesting for future work to study the use of external controls for diverse regulatory pathways.

6. Conclusion

There has been growing research interest in leveraging external controls to facilitate oncology clinical trials. Medical regulatory agencies have published frameworks on accelerating the use of RWD and external controls for regulatory decision-making. However, few systematic discussions were conducted about current applications of external controls in oncology drug development and regulatory feedback. This study identified 18 EMA-approved oncology drugs leveraging 24 external controls. We discussed the use of external control, data sources, analysis methods, and regulators' decision.

We found that external controls had been actively submitted to the EMA. 17% of the EMA-approved cancer drugs in 2016-2021 used external controls, among which 37% of the cases leveraged RWD. However, nearly one-third of the external controls were not considered supportive evidence by EMA due to limitations regarding heterogeneous patient populations, missing outcome assessment in RWD, and inappropriate statistical analysis. This study highlighted that the validity of external controls requires carefully assessing data availability and determining the optimal clinical and statistical methodology.

Our findings highlighted that proper use of external controls requires a careful assessment of clinical settings, data availability, and statistical methodology. For better use of external controls in oncology clinical development, we suggest: *a priori* study designs to avoid selection bias, sufficient baseline data to ensure the comparability of study populations, consistent endpoint measurements to enable outcome comparison, robust statistical methodology for comparative analysis, and collaborative efforts of sponsors and regulators to establish frameworks on external controls.

Declaration of Interest statement

XW receives Ph.D. funding from Sanofi and Association Nationale de la Recherche et de la Technologie (ANRT).

FD, RH, and CL are employees of Sanofi.

The other authors declare no conflict of interest.

Author Contributions

XW: Writing – original draft; Writing – review & editing.

FD: Writing – review & editing.

CL: Writing – review & editing.

RH: Writing – review; Supervision.

AL: Writing – review & editing; Supervision.

RR: Writing – review & editing; Supervision.

REFERENCES

[1] Schulz KF, Altman DG, Moher D. CONSORT 2010 Statement: updated guidelines for reporting parallel group randomised trials. BMJ 2010;340:c332. https://doi.org/10.1136/bmj.c332.

[2] Unger JM, Cook E, Tai E, Bleyer A. The Role of Clinical Trial Participation in Cancer Research: Barriers, Evidence, and Strategies. American Society of Clinical Oncology Educational Book 2016:185–98. https://doi.org/10.1200/EDBK_156686.

[3] Agarwala V, Khozin S, Singal G, O'Connell C, Kuk D, Li G, et al. Real-World Evidence In Support Of Precision Medicine: Clinico-Genomic Cancer Data As A Case Study. Health Affairs 2018;37:765–72. https://doi.org/10.1377/hlthaff.2017.1579.

[4] Eichler H-G, Pignatti F, Schwarzer-Daum B, Hidalgo-Simon A, Eichler I, Arlett P, et al. Randomized Controlled Trials Versus Real World Evidence: Neither Magic Nor Myth. Clinical Pharmacology & Therapeutics 2021;109:1212–8. https://doi.org/10.1002/cpt.2083.

[5] Hatswell AJ, Baio G, Berlin JA, Irs A, Freemantle N. Regulatory approval of pharmaceuticals without a randomised controlled study: analysis of EMA and FDA approvals 1999–2014. BMJ Open 2016;6:e011666. https://doi.org/10.1136/bmjopen-2016-011666.

[6] Goring S, Taylor A, Müller K, Li TJJ, Korol EE, Levy AR, et al. Characteristics of non-randomised studies using comparisons with external controls submitted for regulatory approval in the USA and Europe: a systematic review. BMJ Open 2019;9:e024895. https://doi.org/10.1136/bmjopen-2018-024895.

[7] Viele K, Berry S, Neuenschwander B, Amzal B, Chen F, Enas N, et al. Use of historical control data for assessing treatment effects in clinical trials. Pharmaceut Statist 2014;13:41–54. https://doi.org/10.1002/pst.1589.

[8] Lim J, Walley R, Yuan J, Liu J, Dabral A, Best N, et al. Minimizing Patient Burden Through the Use of Historical Subject-Level Data in Innovative Confirmatory Clinical Trials: Review of Methods and Opportunities. Drug Inf J 2018;52:546–59. https://doi.org/10.1177/2168479018778282.

[9] Davi R, Mahendraratnam N, Chatterjee A, Dawson CJ, Sherman R. Informing single-arm clinical trials with external controls. Nat Rev Drug Discov 2020;19:821–2. https://doi.org/10.1038/d41573-020-00146-5.

[10] Khozin S, Blumenthal GM, Pazdur R. Real-world Data for Clinical Evidence Generation in Oncology. JNCI: Journal of the National Cancer Institute 2017;109. https://doi.org/10.1093/jnci/djx187.

[11] US Food and Drug Adnimistration. Framework for FDA's Real-World Evidence Program 2018.

[12] Cave A, Kurz X, Arlett P. Real-World Data for Regulatory Decision Making: Challenges and Possible Solutions for Europe. Clinical Pharmacology & Therapeutics 2019;106:36–9. https://doi.org/10.1002/cpt.1426.

[13] Arlett P, Kjær J, Broich K, Cooke E. Real-World Evidence in EU Medicines Regulation: Enabling Use and Establishing Value. Clinical Pharmacology & Therapeutics 2022;111:21–3. https://doi.org/10.1002/cpt.2479.

[14] Franklin JM, Patorno E, Desai RJ, Glynn RJ, Martin D, Quinto K, et al. Emulating Randomized Clinical Trials With Nonrandomized Real-World Evidence Studies: First Results From the RCT DUPLICATE Initiative. Circulation 2021;143:1002–13. https://doi.org/10.1161/CIRCULATIONAHA.120.051718.

[15] Collins R, Bowman L, Landray M, Peto R. The Magic of Randomization versus the Myth of Real-World Evidence. N Engl J Med 2020;382:674–8. https://doi.org/10.1056/NEJMsb1901642.

[16] Rahman R, Ventz S, McDunn J, Louv B, Reyes-Rivera I, Polley M-YC, et al. Leveraging external data in the design and analysis of clinical trials in neuro-oncology. The Lancet Oncology 2021;22:e456–65. https://doi.org/10.1016/S1470-2045(21)00488-5.

[17] Skovlund E, Leufkens HGM, Smyth JF. The use of real-world data in cancer drug development. European Journal of Cancer 2018;101:69–76. https://doi.org/10.1016/j.ejca.2018.06.036.

[18] Group IEW. ICH Harmonised Tripartite Guideline: Choice of Control Group and Related Issues in Clinical Trials E10 2000.

[19] Burcu M, Dreyer NA, Franklin JM, Blum MD, Critchlow CW, Perfetto EM, et al. Real-world evidence to support regulatory decision-making for medicines: Considerations for external control arms. Pharmacoepidemiology and Drug Safety 2020;29:1228–35. https://doi.org/10.1002/pds.4975.

[20] Thorlund K, Dron L, Park JJ, Mills EJ. Synthetic and External Controls in Clinical Trials – A Primer for Researchers. CLEP 2020;Volume 12:457–67. https://doi.org/10.2147/CLEP.S242097.

[21] Schmidli H, Häring DA, Thomas M, Cassidy A, Weber S, Bretz F. Beyond Randomized Clinical Trials: Use of External Controls. Clin Pharmacol Ther 2020;107:806–16. https://doi.org/10.1002/cpt.1723.

[22] Pocock SJ. The combination of randomized and historical controls in clinical trials. Journal of Chronic Diseases 1976;29:175–88. https://doi.org/10.1016/0021-9681(76)90044-8.
[23] Chen J, Ho M, Lee K, Song Y, Fang Y, Goldstein BA, et al. The Current Landscape in Biostatistics of Real-World Data and Evidence: Clinical Study Design and Analysis. Statistics in Biopharmaceutical Research 2021;0:1–28. https://doi.org/10.1080/19466315.2021.1883474.
[24] Xu Y, Lu N, Yue L, Tiwari R. A Study Design for Augmenting the Control Group in a Randomized Controlled Trial: A Quality Process for Interaction Among Stakeholders. Ther Innov Regul Sci 2020;54:269–74. https://doi.org/10.1007/s43441-019-00053-x.
[25] Tan WK, Segal BD, Curtis MD, Baxi SS, Capra WB, Garrett-Mayer E, et al. Augmenting control arms with real-world data for cancer trials: Hybrid control arm methods and considerations. Contemporary Clinical Trials Communications 2022;30:101000. https://doi.org/10.1016/j.conctc.2022.101000.
[26] US Department of Health and Human Services Food and Drug Administration. Statistical review and evaluation of emtricitabine/ tenofovir alafenamide NDA 208215 2015. https://www.fda.gov/media/98523/download (accessed November 1, 2021).
[27] Austin PC. An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. Multivariate Behav Res 2011;46:399–424. https://doi.org/10.1080/00273171.2011.568786.
[28] De Maesschalck R, Jouan-Rimbaud D, Massart DL. The Mahalanobis distance. Chemometrics and Intelligent Laboratory Systems 2000;50:1–18. https://doi.org/10.1016/S0169-7439(99)00047-7.
[29] Hobbs BP, Carlin BP, Mandrekar SJ, Sargent DJ. Hierarchical Commensurate and Power Prior Models for Adaptive Incorporation of Historical Information in Clinical Trials. Biometrics 2011;67:1047–56. https://doi.org/10.1111/j.1541-0420.2011.01564.x.
[30] Schmidli H, Gsteiger S, Roychoudhury S, O'Hagan A, Spiegelhalter D, Neuenschwander B. Robust meta-analytic-predictive priors in clinical trials with historical control information: Robust Meta-Analytic-Predictive Priors. Biom 2014;70:1023–32. https://doi.org/10.1111/biom.12242.
[31] Gray CM, Grimson F, Layton D, Pocock S, Kim J. A Framework for Methodological Choice and Evidence Assessment for Studies Using External Comparators from Real-World Data. Drug Saf 2020;43:623–33. https://doi.org/10.1007/s40264-020-00944-1.
[32] King G, Nielsen R. Why Propensity Scores Should Not Be Used for Matching. Polit Anal 2019;27:435–54. https://doi.org/10.1017/pan.2019.11.
[33] Funk MJ, Westreich D, Wiesen C, Stürmer T, Brookhart MA, Davidian M. Doubly Robust Estimation of Causal Effects. American Journal of Epidemiology 2011;173:761–7. https://doi.org/10.1093/aje/kwq439.
[34] Hernán MA, Robins JM. Using Big Data to Emulate a Target Trial When a Randomized Trial Is Not Available. Am J Epidemiol 2016;183:758–64. https://doi.org/10.1093/aje/kwv254.

Résumé en français

L'essai contrôlé randomisé représente la méthode de référence pour établir l'effet causal des traitements expérimentaux par rapport aux traitements de contrôle ou aux placebos. Néanmoins, des problèmes éthiques ou de faisabilité peuvent entraver le processus de randomisation, en particulier dans le développement de médicaments en oncologie. Les autorités réglementaires reconnaissent ces défis et ont accordé des approbations conditionnelles pour des essais à bras unique, avec l'exigence de preuves confirmatoires ultérieures à partir d'études post-approbation. En outre, à l'ère du séquençage du génome et de la médecine de précision, les essais randomisés comparatifs peuvent devenir peu réalisables dans certaines situations, notamment lorsqu'il s'agit de rares anomalies génétiques.

Dans ce contexte, les bras de contrôle historique ont émergé comme une approche complémentaire aux essais cliniques. En fournissant des informations contextuelles et en améliorant l'interprétation des résultats des essais à bras unique, les bras de contrôle historique visent à réduire le biais dû au manque de randomisation. Bien que diverses méthodes statistiques dans le cadre de l'inférence causale aient été proposées, il existe actuellement un manque de directives régissant leur application dans le processus de développement de médicaments. L'objectif de cette thèse est d'évaluer la faisabilité des bras de contrôle historique, en mettant l'accent sur la disponibilité des données historiques et les méthodes d'analyse statistique appropriées.

Afin de réduire les biais lors de l'inclusion de contrôles externes dans les analyses statistiques, diverses méthodes ont été développées, comprenant à la fois des approches fréquentistes et bayésiennes. Les approches bayésiennes prennent en compte l'hétérogénéité des résultats et réduisent le poids des données de contrôle historique lors de leur incorporation dans le nouvel essai clinique, telles que la méthode "power prior", la méthode "commensurate prior", ou la méthode "meta-analytic prior". Étant donné que ces méthodes empruntent des informations aux données historiques, elles

sont souvent appelées “méthodes d’emprunt bayésiennes”. En revanche, les approches fréquentistes impliquent deux étapes principales. Tout d’abord, un score d’équilibre est estimé en utilisant des covariables sélectionnées qui peuvent affecter l’attribution du traitement et le résultat. Des exemples de scores d’équilibre incluent le score de propension et la distance de Mahalanobis. Ensuite, le score d’équilibre est utilisé pour créer des cohortes externes et internes comparables à l’aide de méthodes telles que l’appariement, la pondération par inverse de la probabilité et l’ajustement des covariables.

Dans un premier temps, nous avons réalisé une revue systématique de l’application des bras de contrôle historique dans le développement de médicaments en oncologie en Europe. Cette revue a été menée afin d’identifier les autorisations accordées par l’Agence européenne des médicaments (EMA) entre 2016 et 2021, basées sur l’utilisation de contrôles historiques pour démontrer l’efficacité en oncologie. Parmi les 113 autorisations délivrées par l’EMA au cours de cette période, nous avons identifié 18 soumissions utilisant 24 contrôles externes. Ces soumissions ont été analysées en examinant plusieurs aspects, tels que l’utilisation des contrôles externes, les sources de données, les méthodes d’analyse et les retours des régulateurs. Il est à noter que 37% des bras de contrôle historique ont utilisé des données de vie réelle. Cependant, près d’un tiers de ces contrôles externes n’ont pas été considérés par l’EMA comme apportant un niveau de preuve suffisant en raison de limitations liées à l’hétérogénéité des populations de patients, à l’absence d’évaluation des résultats dans les données de vie réelle et à une analyse statistique jugée inappropriée. Les limitations associées à cette revue ont également été identifiées. Nos résultats indiquent que, bien que les contrôles historiques aient été activement soumis aux autorités réglementaires, ils ne sont pas systématiquement considérés comme des preuves favorables. Nous avons identifié des limitations significatives et formulé des suggestions correspondantes concernant la conception de l’étude, la sélection des données et l’application des méthodes statistiques. Cette étude souligne que la validité des contrôles externes nécessite une évaluation minutieuse de la disponibilité des données et la détermination préalable d’une méthodologie et d’une stratégie d’analyse statistique optimale.

S’appuyant sur les enseignements tirés de la revue, nous avons ensuite réalisé deux études de cas pour approfondir notre compréhension de l’utilisation des bras de contrôle historique dans des contextes plus complexes. La première étude de cas examine un essai à bras unique observationnel qui a évalué l’efficacité du bloc du plan du muscle érecteur du rachis dans la réduction de la douleur post-opératoire

lors de la chirurgie du cancer du sein. Le bloc paravertébral thoracique (BPVT) est une technique d'anesthésie locorégionale qui a démontré son intérêt dans le cadre des chirurgies "extensives" pour le cancer du sein. Une technique alternative, dénommée bloc érecteur du rachis, a été développée au milieu des années 2010. Cette procédure est plus simple, plus rapide et moins risquée que le bloc paravertébral. Étant donné que le bloc érecteur du rachis était initialement controversé, une étude monobras prospective (ESPB) a été réalisée pour évaluer la sécurité et l'efficacité de cette approche, impliquant 120 patients. L'objectif de cette étude était de réaliser une comparaison historique à partir de l'étude ESPB et d'une étude randomisée antérieure évaluant deux modalités pour le bloc paravertébral (MIRO3). La première étape a consisté à appliquer les critères de Pocock pour évaluer la comparabilité des deux études. Les caractéristiques des patients ont été comparées entre les groupes à l'aide de la différence standardisée. Les facteurs de confusion associés au critère de jugement (le taux de patients recevant de la morphine en post-opératoire) ont été pris en compte. Un score de propension a été déterminé par régression logistique, suivi de l'appariement des patients à l'aide de la technique du caliper. Les différences standardisées moyennes (SMD) ont été présentées avant et après l'appariement. Les résultats ont montré que le taux de patients nécessitant de la morphine était plus élevé dans le groupe recevant le bloc érecteur du rachis. De plus, une analyse de sensibilité a été réalisée en utilisant une méthode de score de propension basée sur la forêt aléatoire. Cette étude a illustré comment un processus rigoureux de sélection des données et l'utilisation de méthodes statistiques appropriées peuvent éliminer les divergences des données externes, établissant ainsi un solide contrôle historique. Cependant, il convient de noter qu'une limitation inhérente était que notre analyse était rétrospective, ce qui la rend complémentaire à l'essai clinique conclu.

Basés sur les enseignements tirés de la première étude de cas, nous avons cherché à approfondir notre recherche dans un cadre plus prospectif et à examiner la faisabilité des contrôles historiques dans des sources de données plus complexes. Nous croyons que les contrôles historiques ne devraient pas simplement compléter les essais cliniques, mais devraient être intégrés dans la conception et les méthodologies des essais cliniques. Avec le soutien théorique du cadre des résultats potentiels et les directives systématiques du Cadre d'Estimand (EF), des études récentes ont proposé le Cadre d'Essai Cible (TTF) pour les études observationnelles. Par conséquent, nous avons cherché à incorporer le TTF dans les contrôles historiques tout en élargissant notre source de données des essais cliniques aux données de la vie réelle. La deuxième étude de cas se concentre sur un essai randomisé contrôlé examinant Olaparib en association

avec Bevacizumab en tant que traitement d'entretien de première ligne dans le cancer de l'ovaire. Ici, nous avons cherché à émuler le bras de contrôle en utilisant les données observationnelles de la base de données du monde réel ESME, en utilisant le cadre de l'essai cible. L'exécution de l'essai émulé est actuellement en cours. Bien que la base de données des données de la vie réelle présente l'avantage d'une plus grande taille d'échantillon, elle pose des défis en matière de gestion des données et de prétraitement des données. En collaboration avec les cliniciens, nous sommes actuellement engagés dans l'identification des données de chirurgie de débulking, dans le but de raffiner notre sélection de patients répondant aux critères d'éligibilité de l'étude PAOLA-1.

En conclusion, l'exploration des bras de contrôle historique dans les essais cliniques met en lumière le paysage dynamique et évolutif du développement de médicaments. Bien qu'ils offrent une opportunité prometteuse pour faire progresser la recherche clinique, les aspects pratiques de leur intégration dans les essais cliniques sont confrontés à des défis, allant de l'accès aux données aux méthodologies statistiques. Cette thèse, ancrée dans la collaboration entre le monde académique et l'industrie, contribue à faire avancer notre compréhension de la faisabilité et de l'applicabilité des bras de contrôle historique, éclairant ainsi leurs avantages potentiels et leurs applications appropriées dans le domaine du développement de médicaments en oncologie.

De plus, des avancées notables ont été récemment observées dans le paysage réglementaire concernant l'utilisation des bras de contrôle historique pour l'autorisation de médicaments. En 2023, les principales autorités réglementaires comme EMA, FDA et HAS, ont publié des documents mettant en évidence le rôle crucial des bras de contrôle historique dans l'évaluation des produits médicamenteux. Ces publications soulignent collectivement un intérêt croissant pour l'intégration des données de contrôle historique dans le processus de développement de médicaments. Elles insistent sur l'importance de mettre en place un cadre complet impliquant tous les acteurs, y compris les autorités réglementaires, les experts de l'industrie, les chercheurs et les statisticiens. Ces documents réglementaires font écho de près aux messages et objectifs fondamentaux de cette thèse.

Il est crucial de noter que les bras de contrôle historique ne sont pas destinés à remplacer complètement les bras de contrôle randomisés, mais plutôt à les compléter dans des contextes cliniques appropriés. Par conséquent, nous proposons le développement d'un cadre pertinent visant à garantir une utilisation adéquate et une validation des bras de contrôle historique. Ce cadre vise à améliorer l'efficacité et l'efficacité des essais cliniques et à accélérer l'autorisation de médicaments innovants.

En fin de compte, cette thèse illustre les progrès et les défis liés à l'utilisation des bras de contrôle historique dans le développement de médicaments en oncologie. Elle met en lumière l'importance de l'évaluation rigoureuse des données historiques, de la sélection appropriée des données et de l'application de méthodes statistiques adaptées. Grâce à une approche multidisciplinaire et à une collaboration entre les parties prenantes de l'industrie et de la recherche, cette thèse contribue à éclairer le potentiel des bras de contrôle historique pour améliorer la recherche clinique en oncologie et à ouvrir la voie à une utilisation plus répandue et efficace de ces approches novatrices dans le développement de médicaments.

En somme, les avancées dans le domaine des bras de contrôle historique témoignent de l'évolution constante du paysage de la recherche clinique et de la réglementation dans le domaine du développement de médicaments. Ces avancées promettent d'améliorer la validité et l'efficacité des essais cliniques en oncologie, tout en maintenant des normes de sécurité et d'efficacité rigoureuses.

References

Agarwala, V., Khozin, S., Singal, G., O'Connell, C., Kuk, D., Li, G., Gossai, A., Miller, V., & Abernethy, A. P. (2018). Real-World Evidence In Support Of Precision Medicine: Clinico-Genomic Cancer Data As A Case Study. *Health Affairs*, *37*(5), 765–772. https://doi.org/10.1377/hlthaff.2017.1579

Albi-Feldzer, A., Dureau, S., Ghimouz, A., Raft, J., Soubirou, J.-L., Gayraud, G., & Jayr, C. (2021). Preoperative Paravertebral Block and Chronic Pain after Breast Cancer Surgery: A Double-blind Randomized Trial. *Anesthesiology*, *135*(6), 1091–1103. https://doi.org/10.1097/ALN.0000000000003989

Albi-Feldzer, A., Mouret-Fourme E, E., Hamouda, S., Motamed, C., Dubois, P.-Y., Jouanneau, L., & Jayr, C. (2013). A Double-blind Randomized Trial of Wound and Intercostal Space Infiltration with Ropivacaine during Breast Cancer Surgery: Effects on Chronic Postoperative Pain. *Anesthesiology*, *118*(2), 318–326. https://doi.org/10.1097/ALN.0b013e31827d88d8

Andrillon, A., Pirracchio, R., & Chevret, S. (2020). Performance of propensity score matching to estimate causal effects in small samples. *Statistical Methods in Medical Research*, *29*(3), 644–658. https://doi.org/10.1177/0962280219887196

Arlett, P., Kjær, J., Broich, K., & Cooke, E. (2022). Real-World Evidence in EU Medicines Regulation: Enabling Use and Establishing Value. *Clinical Pharmacology & Therapeutics*, *111*(1), 21–23. https://doi.org/10.1002/cpt.2479

Austin, P. C. (2009a). Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Statistics in Medicine*, *28*(25), 3083–3107. https://doi.org/10.1002/sim.3697

Austin, P. C. (2009b). Type I Error Rates, Coverage of Confidence Intervals, and Variance Estimation in Propensity-Score Matched Analyses. *The International*

Journal of Biostatistics, *5*(1). https://doi.org/10.2202/1557-4679.1146

Austin, P. C. (2009c). The Relative Ability of Different Propensity Score Methods to Balance Measured Covariates Between Treated and Untreated Subjects in Observational Studies. *Medical Decision Making*, *29*(6), 661–677. https://doi.org/10.1177/0272989X09341755

Austin, P. C. (2011a). Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies. *Pharmaceutical Statistics*, *10*(2), 150–161. https://doi.org/10.1002/pst.433

Austin, P. C. (2011b). An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. *Multivariate Behavioral Research*, *46*(3), 399–424. https://doi.org/10.1080/00273171.2011.568786

Austin, P. C. (2014). The use of propensity score methods with survival or time-to-event outcomes: Reporting measures of effect similar to those used in randomized experiments. *Statistics in Medicine*, *33*(7), 1242–1258. https://doi.org/10.1002/sim.5984

Austin, P. C. (2017). Double propensity-score adjustment: A solution to design bias or bias due to incomplete matching. *Statistical Methods in Medical Research*, *26*(1), 201–222. https://doi.org/10.1177/0962280214543508

Austin, P. C., Grootendorst, P., & Anderson, G. M. (2007). A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: A Monte Carlo study. *Statistics in Medicine*, *26*(4), 734–753. https://doi.org/10.1002/sim.2580

Austin, P. C., & Mamdani, M. M. (2006). A comparison of propensity score methods: A case-study estimating the effectiveness of post-AMI statin use. *Statistics in Medicine*, *25*(12), 2084–2106. https://doi.org/10.1002/sim.2328

Austin, P. C., & Stuart, E. A. (2015). Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. *Statistics in Medicine*, *34*(28), 3661–3679. https://doi.org/10.1002/sim.6607

Bacic, J., Liu, T., Thompson, R. H., Boorjian, S. A., Leibovich, B. C., Golijanin, D., & Gershman, B. (2020). Emulating Target Clinical Trials of Radical Nephrectomy

With or Without Lymph Node Dissection for Renal Cell Carcinoma. *Urology*, *140*, 98–106. https://doi.org/10.1016/j.urology.2020.01.039

Banbeta, A., van Rosmalen, J., Dejardin, D., & Lesaffre, E. (2019). Modified power prior with multiple historical trials for binary endpoints. *Statistics in Medicine*, *38*(7), 1147–1169. https://doi.org/10.1002/sim.8019

Beaver, J. A., Howie, L. J., Pelosof, L., Kim, T., Liu, J., Goldberg, K. B., Sridhara, R., Blumenthal, G. M., Farrell, A. T., Keegan, P., Pazdur, R., & Kluetz, P. G. (2018). A 25-Year Experience of US Food and Drug Administration Accelerated Approval of Malignant Hematology and Oncology Drugs and Biologics: A Review. *JAMA Oncology*, *4*(6), 849–856. https://doi.org/10.1001/jamaoncol.2017.5618

Bigirumurame, T., Hiu, S. K. W., Teare, M. D., Wason, J. M. S., Bryant, A., & Breckons, M. (2023). Current practices in studies applying the target trial emulation framework: A protocol for a systematic review. *BMJ Open*, *13*(6), e070963. https://doi.org/10.1136/bmjopen-2022-070963

Bini, M., Quesada, S., Meeus, P., Rodrigues, M., Leblanc, E., Floquet, A., Pautier, P., Marchal, F., Provansal, M., Campion, L., Causeret, S., Gourgou, S., Ray-Coquard, I., Classe, J.-M., Pomel, C., De La Motte Rouge, T., Barranger, E., Savoye, A. M., Guillemet, C., … Joly, F. (2022). Real-World Data on Newly Diagnosed BRCA-Mutated High-Grade Epithelial Ovarian Cancers: The French National Multicenter ESME Database. *Cancers*, *14*(16, 16), 4040. https://doi.org/10.3390/cancers14164040

Burger, H. U., Gerlinger, C., Harbron, C., Koch, A., Posch, M., Rochon, J., & Schiel, A. (2021). The use of external controls: To what extent can it currently be recommended? *Pharmaceutical Statistics*, *20*(6), 1002–1016. https://doi.org/10.1002/pst.2120

Burns, P. B., Rohrich, R. J., & Chung, K. C. (2011). The Levels of Evidence and their role in Evidence-Based Medicine. *Plastic and Reconstructive Surgery*, *128*(1), 305–310. https://doi.org/10.1097/PRS.0b013e318219c171

Cave, A., Kurz, X., & Arlett, P. (2019). Real-World Data for Regulatory Decision Making: Challenges and Possible Solutions for Europe. *Clinical Pharmacology & Therapeutics*, *106*(1), 36–39. https://doi.org/10.1002/cpt.1426

Chan, P., Peskov, K., & Song, X. (2022). Applications of Model-Based Meta-Analysis in Drug Development. *Pharmaceutical Research*, *39*(8), 1761–1777. https://doi.org/10.1007/s11095-022-03201-5

Chen, M.-H., & Ibrahim, J. G. (2000). Power prior distributions for regression models. *Statistical Science*, *15*(1), 46–60. https://doi.org/10.1214/ss/1009212673

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, *21*(1), C1–C68. https://doi.org/10.1111/ectj.12097

Choi, J., Dekkers, O. M., & le Cessie, S. (2019). A comparison of different methods to handle missing data in the context of propensity score analysis. *European Journal of Epidemiology*, *34*(1), 23–36. https://doi.org/10.1007/s10654-018-0447-z

Collins, R., Bowman, L., Landray, M., & Peto, R. (2020). The Magic of Randomization versus the Myth of Real-World Evidence. *New England Journal of Medicine*, *382*(7), 674–678. https://doi.org/10.1056/NEJMsb1901642

Davi, R., Mahendraratnam, N., Chatterjee, A., Dawson, C. J., & Sherman, R. (2020). Informing single-arm clinical trials with external controls. *Nature Reviews Drug Discovery*, *19*(12), 821–822. https://doi.org/10.1038/d41573-020-00146-5

De Maesschalck, R., Jouan-Rimbaud, D., & Massart, D. L. (2000). The Mahalanobis distance. *Chemometrics and Intelligent Laboratory Systems*, *50*(1), 1–18. https://doi.org/10.1016/S0169-7439(99)00047-7

De Nonneville, A., Zemmour, C., Frank, S., Joly, F., Ray-Coquard, I., Costaz, H., Classe, J.-M., Floquet, A., De la Motte Rouge, T., Colombo, P.-E., Sauterey, B., Leblanc, E., Pomel, C., Marchal, F., Barranger, E., Savoye, A.-M., Guillemet, C., Petit, T., Pautier, P., … Sabatier, R. (2021). Clinicopathological characterization of a real-world multicenter cohort of endometrioid ovarian carcinoma: Analysis of the French national ESME-Unicancer database. *Gynecologic Oncology*, *163*(1), 64–71. https://doi.org/10.1016/j.ygyno.2021.07.019

Eichler, H.-G., Pignatti, F., Schwarzer-Daum, B., Hidalgo-Simon, A., Eichler, I., Arlett, P., Humphreys, A., Vamvakas, S., Brun, N., & Rasi, G. (2021). Randomized Controlled Trials Versus Real World Evidence: Neither Magic Nor Myth. *Clinical Pharmacology & Therapeutics*, *109*(5), 1212–1218. https://doi.org/10.1002/

cpt.2083

EMA. (2021). *Minjuvi : EPAR - Public assessment report*. https://www.ema.europa.eu/en/documents/assessment-report/minjuvi-epar-public-assessment-report_en.pdf

EMA. (2023, April 21). *Single-arm trials as pivotal evidence for the authorisation of medicines in the EU | European Medicines Agency*. https://www.ema.europa.eu/en/news/single-arm-trials-pivotal-evidence-authorisation-medicines-eu

European Medicines Agency. (2001). *ICH E10 Choice of control group in clinical trials*.

European Medicines Agency (EMA). (2020). *Enhertu : EPAR - Public assessment report*. https://www.ema.europa.eu/en/documents/assessment-report/enhertu-epar-public-assessment-report_en.pdf

FDA. (2015). *Statistical review and evaluation of emtricitabine/ tenofovir alafenamide NDA 208215*. https://www.fda.gov/media/98523/download

FDA. (2018). *Framework for FDA's Real-World Evidence Program*. https://www.fda.gov/media/120060/download

FDA. (2023, January 31). *Considerations for the Design and Conduct of Externally Controlled Trials for Drug and Biological Products*. FDA. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/considerations-design-and-conduct-externally-controlled-trials-drug-and-biological-products

Forero, M., Adhikary, S. D., Lopez, H., Tsui, C., & Chin, K. J. (2016). The Erector Spinae Plane Block: A Novel Analgesic Technique in Thoracic Neuropathic Pain. *Regional Anesthesia & Pain Medicine*, *41*(5), 621–627. https://doi.org/10.1097/AAP.0000000000000451

Franklin, J. M., Patorno, E., Desai, R. J., Glynn, R. J., Martin, D., Quinto, K., Pawar, A., Bessette, L. G., Lee, H., Garry, E. M., Gautam, N., & Schneeweiss, S. (2021). Emulating Randomized Clinical Trials With Nonrandomized Real-World Evidence Studies: First Results From the RCT DUPLICATE Initiative. *Circulation*, *143*(10), 1002–1013. https://doi.org/10.1161/CIRCULATIONAHA.120.051718

Franklin, J. M., & Schneeweiss, S. (2017). When and How Can Real World Data Analyses Substitute for Randomized Controlled Trials?: Real world evidence and RCTs. *Clinical Pharmacology & Therapeutics*, *102*(6), 924–933. https://doi.org/10.1002/cpt.857

Friends of Cancer Research. (2019). *Characterizing the use of external controls for augmenting randomized control arms and confirming benefit*. https://www.focr.org/sites/default/files/Panel-1_External_Control_Arms2019AM.pdf

Funk, M. J., Westreich, D., Wiesen, C., Stürmer, T., Brookhart, M. A., & Davidian, M. (2011). Doubly Robust Estimation of Causal Effects. *American Journal of Epidemiology*, *173*(7), 761–767. https://doi.org/10.1093/aje/kwq439

Gärtner, R., Jensen, M.-B., Nielsen, J., Ewertz, M., Kroman, N., & Kehlet, H. (2009). Prevalence of and Factors Associated With Persistent Pain Following Breast Cancer Surgery. *JAMA*, *302*(18), 1985. https://doi.org/10.1001/jama.2009.1568

Geoffrey, Marshall. (1948). Streptomycin Treatment of Pulmonary Tuberculosis. *British Medical Journal*, *2*(4582), 769–782. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2091872/

Ghadessi, M., Tang, R., Zhou, J., Liu, R., Wang, C., Toyoizumi, K., Mei, C., Zhang, L., Deng, C. Q., & Beckman, R. A. (2020). A roadmap to using historical controls in clinical trials – by Drug Information Association Adaptive Design Scientific Working Group. *Orphanet Journal of Rare Diseases*, *15*(1), 69. https://doi.org/10.1186/s13023-020-1332-x

Gökbuget, N., Kelsh, M., Chia, V., Advani, A., Bassan, R., Dombret, H., Doubek, M., Fielding, A. K., Giebel, S., Haddad, V., Hoelzer, D., Holland, C., Ifrah, N., Katz, A., Maniar, T., Martinelli, G., Morgades, M., O'Brien, S., Ribera, J.-M., … Kantarjian, H. (2016). Blinatumomab vs historical standard therapy of adult relapsed/refractory acute lymphoblastic leukemia. *Blood Cancer Journal*, *6*(9), e473. https://doi.org/10.1038/bcj.2016.84

Gokhale, M., Stürmer, T., & Buse, J. B. (2020). Real-world evidence: The devil is in the detail. *Diabetologia*, *63*(9), 1694–1705. https://doi.org/10.1007/s00125-020-05217-1

Goring, S., Taylor, A., Müller, K., Li, T. J. J., Korol, E. E., Levy, A. R., & Freemantle, N. (2019). Characteristics of non-randomised studies using comparisons with external controls submitted for regulatory approval in the USA and Europe: A systematic review. *BMJ Open*, *9*(2), e024895. https://doi.org/10.1136/bmjopen-2018-024895

Gray, C. M., Grimson, F., Layton, D., Pocock, S., & Kim, J. (2020). A Framework for Methodological Choice and Evidence Assessment for Studies Using External Comparators from Real-World Data. *Drug Safety*, *43*(7), 623–633. https://doi.org/10.1007/s40264-020-00944-1

Grignolo, A., & Pretorius, S. (2016). *Phase III Trial Failures: Costly, But Preventable*. *25*(8).

Gu, X. S., & Rosenbaum, P. R. (1993). Comparison of Multivariate Matching Methods: Structures, Distances, and Algorithms. *Journal of Computational and Graphical Statistics*, *2*(4), 405–420. https://doi.org/10.1080/10618600.1993.10474623

Hall, K. T., Vase, L., Tobias, D. K., Dashti, H. T., Vollert, J., Kaptchuk, T. J., & Cook, N. R. (2021). Historical Controls in Randomized Clinical Trials: Opportunities and Challenges. *Clinical Pharmacology & Therapeutics*, *109*(2), 343–351. https://doi.org/10.1002/cpt.1970

Hansen, B. B., & Klopfer, S. O. (2006). Optimal Full Matching and Related Designs via Network Flows. *Journal of Computational and Graphical Statistics*, *15*(3), 609–627. https://doi.org/10.1198/106186006X137047

Hariton, E., & Locascio, J. J. (2018). Randomised controlled trials—the gold standard for effectiveness research. *BJOG : An International Journal of Obstetrics and Gynaecology*, *125*(13), 1716. https://doi.org/10.1111/1471-0528.15199

Harter, P., Mouret-Reynier, M. A., Pignata, S., Cropet, C., González-Martín, A., Bogner, G., Fujiwara, K., Vergote, I., Colombo, N., Nøttrup, T. J., Floquet, A., El-Balat, A., Scambia, G., Guerra Alia, E. M., Fabbro, M., Schmalfeldt, B., Hardy-Bessard, A.-C., Runnebaum, I., Pujade-Lauraine, E., & Ray-Coquard, I. (2022). Efficacy of maintenance olaparib plus bevacizumab according to clinical risk in patients with newly diagnosed, advanced ovarian cancer in the phase III PAOLA-1/ENGOT-ov25 trial. *Gynecologic Oncology*, *164*(2), 254–264. https://doi.org/10.1016/j.ygyno.2021.12.016

Hatswell, A., Freemantle, N., Baio, G., Lesaffre, E., & van Rosmalen, J. (2020). Summarising salient information on historical controls: A structured assessment of validity and comparability across studies. *Clinical Trials*, 1740774520944855. https://doi.org/10.1177/1740774520944855

Hernan, M. A., & Robins, J. M. (2023). *Causal Inference: What If*.

Hernán, M. A., & Robins, J. M. (2016). Using Big Data to Emulate a Target Trial When a Randomized Trial Is Not Available. *American Journal of Epidemiology*, *183*(8), 758–764. https://doi.org/10.1093/aje/kwv254

Hernán, M. A., Wang, W., & Leaf, D. E. (2022). Target Trial Emulation: A Framework for Causal Inference From Observational Data. *JAMA*. https://doi.org/10.1001/jama.2022.21383

Hill, J., & Reiter, J. P. (2006). Interval estimation for treatment effects using propensity score matching. *Statistics in Medicine*, *25*(13), 2230–2256. https://doi.org/10.1002/sim.2277

Ho, D., Imai, K., King, G., & Stuart, E. A. (2011). MatchIt: Nonparametric Preprocessing for Parametric Causal Inference. *Journal of Statistical Software*, *42*, 1–28. https://doi.org/10.18637/jss.v042.i08

Hobbs, B. P., Carlin, B. P., Mandrekar, S. J., & Sargent, D. J. (2011). Hierarchical Commensurate and Power Prior Models for Adaptive Incorporation of Historical Information in Clinical Trials. *Biometrics*, *67*(3), 1047–1056. https://doi.org/10.1111/j.1541-0420.2011.01564.x

Hobbs, B. P., Sargent, D. J., & Carlin, B. P. (2012). Commensurate Priors for Incorporating Historical Information in Clinical Trials Using General and Generalized Linear Models. *Bayesian Analysis*, *7*(3), 639–674. https://doi.org/10.1214/12-BA722

Huang, W., Wang, W., Xie, W., Chen, Z., & Liu, Y. (2020). Erector spinae plane block for postoperative analgesia in breast and thoracic surgery: A systematic review and meta-analysis. *Journal of Clinical Anesthesia*, *66*, 109900. https://doi.org/10.1016/j.jclinane.2020.109900

Huitfeldt, A. (2015). *Emulation of Target Trials to Study the Effectiveness and Safety of Medical Interventions*. https://dash.harvard.edu/handle/1/23205172

ICH. (2021). *E9(R1) Statistical Principles for Clinical Trials: Addendum: Estimands and Sensitivity Analysis in Clinical Trials; International Council for Harmonisation; Guidance for Industry; Availability*. https://database.ich.org/sites/default/files/E9-R1_Step4_Guideline_2019_1203.pdf

Imbens, G. (2004). Nonparametric Estimation of Average Treatment Effects under Exogeneity: A Review. *Review of Economics and Statistics*.

Joffe, M. M., Have, T. R. T., Feldman, H. I., & Kimmel, S. E. (2004). Model Selection, Confounder Control, and Marginal Structural Models: Review and New Applications. *The American Statistician*, *58*(4), 272–279. https://www.jstor.org/stable/27643582

Kabisch, M., Ruckes, C., Seibert-Grafe, M., & Blettner, M. (2011). Randomized Controlled Trials. *Deutsches Ärzteblatt International*, *108*(39), 663–668. https://doi.org/10.3238/arztebl.2011.0663

Khozin, S., Blumenthal, G. M., & Pazdur, R. (2017). Real-world Data for Clinical Evidence Generation in Oncology. *JNCI: Journal of the National Cancer Institute*, *109*(11). https://doi.org/10.1093/jnci/djx187

King, G., & Nielsen, R. (2019). Why Propensity Scores Should Not Be Used for Matching. *Political Analysis*, *27*(4), 435–454. https://doi.org/10.1017/pan.2019.11

Lambert, J., Lengliné, E., Porcher, R., Thiébaut, R., Zohar, S., & Chevret, S. (2022). Enriching single-arm clinical trials with external controls: Possibilities and pitfalls. *Blood Advances*, bloodadvances.2022009167. https://doi.org/10.1182/bloodadvances.2022009167

Leong, R. W., Tan, E. S. J., Wong, S. N., Tan, K. H., & Liu, C. W. (2021). Efficacy of erector spinae plane block for analgesia in breast surgery: A systematic review and meta-analysis. *Anaesthesia*, *76*(3), 404–413. https://doi.org/10.1111/anae.15164

Lewis, C. J., Sarkar, S., Zhu, J., & Carlin, B. P. (2019). Borrowing From Historical Control Data in Cancer Drug Development: A Cautionary Tale and Practical Guidelines. *Statistics in Biopharmaceutical Research*, *11*(1), 67–78. https://doi.org/10.1080/19466315.2018.1497533

Lim, J., Walley, R., Yuan, J., Liu, J., Dabral, A., Best, N., Grieve, A., Hampson, L., Wolfram, J., Woodward, P., Yong, F., Zhang, X., & Bowen, E. (2018). Minimizing Patient Burden Through the Use of Historical Subject-Level Data in Innovative Confirmatory Clinical Trials: Review of Methods and Opportunities. *Therapeutic Innovation & Regulatory Science*, *52*(5), 546–559. https://doi.org/10.1177/2168479018778282

Loiseau, N., Trichelair, P., He, M., Andreux, M., Zaslavskiy, M., Wainrib, G., & Blum, M. G. B. (2022). External control arm analysis: An evaluation of propensity score approaches, G-computation, and doubly debiased machine learning. *BMC Medical Research Methodology*, *22*(1), 335. https://doi.org/10.1186/s12874-022-01799-z

Mack, C., Christian, J., Brinkley, E., Warren, E. J., Hall, M., & Dreyer, N. (2019). When Context Is Hard to Come By: External Comparators and How to Use Them. *Therapeutic Innovation & Regulatory Science*, 216847901987867. https://doi.org/10.1177/2168479019878672

McCaffrey, D. F., Griffin, B. A., Almirall, D., Slaughter, M. E., Ramchand, R., & Burgette, L. F. (2013). A Tutorial on Propensity Score Estimation for Multiple Treatments Using Generalized Boosted Models. *Statistics in Medicine*, *32*(19), 3388–3414. https://doi.org/10.1002/sim.5753

Meldrum, M. L. (2000). *A Brief History of the Randomized Controlled Trial*. https://doi.org/10.1016/s0889-8588(05)70309-9

Morgan, S. L., & Todd, J. J. (2008). A Diagnostic Routine for the Detection of Consequential Heterogeneity of Causal Effects. *Sociological Methodology*, *38*(1), 231–282. https://doi.org/10.1111/j.1467-9531.2008.00204.x

Murad, M. H., Asi, N., Alsawas, M., & Alahdab, F. (2016). New evidence pyramid. *Evidence Based Medicine*, *21*(4), 125–127. https://doi.org/10.1136/ebmed-2016-110401

Neuenschwander, B., Branson, M., & Spiegelhalter, D. J. (2009). A note on the power prior. *Statistics in Medicine*, *28*(28), 3562–3566. https://doi.org/10.1002/sim.3722

Normand, S.-L. T., Landrum, M. B., Guadagnoli, E., Ayanian, J. Z., Ryan, T. J., Cleary, P. D., & McNeil, B. J. (2001). Validating recommendations for coro-

nary angiography following acute myocardial infarction in the elderly. *Journal of Clinical Epidemiology*, *54*(4), 387–398. https://doi.org/10.1016/S0895-4356(00)00321-8

Pocock, S. J. (1976). The combination of randomized and historical controls in clinical trials. *Journal of Chronic Diseases*, *29*(3), 175–188. https://doi.org/10.1016/0021-9681(76)90044-8

Pullenayegum, E. M. (2011). An informed reference prior for between-study heterogeneity in meta-analyses of binary outcomes. *Statistics in Medicine*, *30*(26), 3082–3094. https://doi.org/10.1002/sim.4326

Rahman, R., Ventz, S., McDunn, J., Louv, B., Reyes-Rivera, I., Polley, M.-Y. C., Merchant, F., Abrey, L. E., Allen, J. E., Aguilar, L. K., Aguilar-Cordova, E., Arons, D., Tanner, K., Bagley, S., Khasraw, M., Cloughesy, T., Wen, P. Y., Alexander, B. M., & Trippa, L. (2021). Leveraging external data in the design and analysis of clinical trials in neuro-oncology. *The Lancet Oncology*, *22*(10), e456–e465. https://doi.org/10.1016/S1470-2045(21)00488-5

Ray-Coquard, I., Pautier, P., Pignata, S., Pérol, D., González-Martín, A., Berger, R., Fujiwara, K., Vergote, I., Colombo, N., Mäenpää, J., Selle, F., Sehouli, J., Lorusso, D., Guerra Alía, E. M., Reinthaller, A., Nagao, S., Lefeuvre-Plesse, C., Canzler, U., Scambia, G., … Harter, P. (2019). Olaparib plus Bevacizumab as First-Line Maintenance in Ovarian Cancer. *New England Journal of Medicine*, *381*(25), 2416–2428. https://doi.org/10.1056/NEJMoa1911361

Robins, J. M., Hernán, M. Á., & Brumback, B. (2000). Marginal Structural Models and Causal Inference in Epidemiology. *Epidemiology*, *11*(5), 550. https://journals.lww.com/epidem/fulltext/2000/09000/marginal_structural_models_and_causal_inference_in.11.aspx

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41–55. https://doi.org/10.1093/biomet/70.1.41

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, *66*(5), 688–701. https://doi.org/10.1037/h0037350

Rubin, D. B. (2004). On principles for modeling propensity scores in medical research. *Pharmacoepidemiology and Drug Safety*, *13*(12), 855–857. https://doi.org/10.1002/pds.968

Rubin, D. B. (2005). Causal Inference Using Potential Outcomes. *Journal of the American Statistical Association*, *100*(469), 322–331. https://doi.org/10.1198/016214504000001880

Rudolph, K. E., & Stuart, E. A. (2018). Using Sensitivity Analyses for Unobserved Confounding to Address Covariate Measurement Error in Propensity Score Methods. *American Journal of Epidemiology*, *187*(3), 604–613. https://doi.org/10.1093/aje/kwx248

Saad, E. D., Paoletti, X., Burzykowski, T., & Buyse, M. (2017). Precision medicine needs randomized clinical trials. *Nature Reviews Clinical Oncology*, *14*(5), 317–323. https://doi.org/10.1038/nrclinonc.2017.8

Schaubel, D. E., & Wei, G. (2011). Double inverse-weighted estimation of cumulative treatment effects under nonproportional hazards and dependent censoring. *Biometrics*, *67*(1), 29–38. https://doi.org/10.1111/j.1541-0420.2010.01449.x

Schmidli, H., Gsteiger, S., Roychoudhury, S., O'Hagan, A., Spiegelhalter, D., & Neuenschwander, B. (2014). Robust meta-analytic-predictive priors in clinical trials with historical control information: Robust Meta-Analytic-Predictive Priors. *Biometrics*, *70*(4), 1023–1032. https://doi.org/10.1111/biom.12242

Schmidli, H., Häring, D. A., Thomas, M., Cassidy, A., Weber, S., & Bretz, F. (2020). Beyond Randomized Clinical Trials: Use of External Controls. *Clinical Pharmacology & Therapeutics*, *107*(4), 806–816. https://doi.org/10.1002/cpt.1723

Schulz, K. F., Altman, D. G., & Moher, D. (2010). CONSORT 2010 Statement: Updated guidelines for reporting parallel group randomised trials. *BMJ*, *340*, c332. https://doi.org/10.1136/bmj.c332

Sekhon, J. S. (2011). Multivariate and Propensity Score Matching Software with Automated Balance Optimization: The **Matching** Package for *R*. *Journal of Statistical Software*, *42*(7). https://doi.org/10.18637/jss.v042.i07

Setoguchi, S., Schneeweiss, S., Brookhart, M. A., Glynn, R. J., & Cook, E. F. (2008). Evaluating uses of data mining techniques in propensity score estimation: A simulation study. *Pharmacoepidemiology and Drug Safety*, *17*(6), 546–555.

https://doi.org/10.1002/pds.1555

Snowden, J. M., Rose, S., & Mortimer, K. M. (2011). Implementation of G-Computation on a Simulated Data Set: Demonstration of a Causal Inference Technique. *American Journal of Epidemiology*, *173*(7), 731–738. https://doi.org/10.1093/aje/kwq472

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science : A Review Journal of the Institute of Mathematical Statistics*, *25*(1), 1–21. https://doi.org/10.1214/09-STS313

Thomas, Q. D., Boussere, A., Classe, J.-M., Pomel, C., Costaz, H., Rodrigues, M., Ray-Coquard, I., Gladieff, L., Rouzier, R., Rouge, T. D. L. M., Gouy, S., Barranger, E., Sabatier, R., Floquet, A., Marchal, F., Guillemet, C., Polivka, V., Martin, A.-L., Colombo, P.-E., & Fiteni, F. (2022). Optimal timing of interval debulking surgery for advanced epithelial ovarian cancer: A retrospective study from the ESME national cohort. *Gynecologic Oncology*, *167*(1), 11–21. https://doi.org/10.1016/j.ygyno.2022.08.005

Thorlund, K., Dron, L., Park, J. J., & Mills, E. J. (2020). Synthetic and External Controls in Clinical Trials – A Primer for Researchers. *Clinical Epidemiology*, *Volume 12*, 457–467. https://doi.org/10.2147/CLEP.S242097

Vanderbeek, A. M., Ventz, S., Rahman, R., Fell, G., Cloughesy, T. F., Wen, P. Y., Trippa, L., & Alexander, B. M. (2019). To randomize, or not to randomize, that is the question: Using data from prior clinical trials to guide future designs. *Neuro-Oncology*, *21*(10), 1239–1249. https://doi.org/10.1093/neuonc/noz097

VanderWeele, T. J. (2019). Principles of confounder selection. *European Journal of Epidemiology*, *34*(3), 211–219. https://doi.org/10.1007/s10654-019-00494-6

Vanier, A., Fernandez, J., Kelley, S., Alter, L., Semenzato, P., Alberti, C., Chevret, S., Costagliola, D., Cucherat, M., Falissard, B., Gueyffier, F., Lambert, J., Lengliné, E., Locher, C., Naudet, F., Porcher, R., Thiébaut, R., Vray, M., Zohar, S., … Guludec, D. L. (2023). Rapid access to innovative medicinal products while ensuring relevant health technology assessment. Position of the French National Authority for Health. *BMJ Evidence-Based Medicine*. https://doi.org/10.1136/bmjebm-2022-112091

Viele, K., Berry, S., Neuenschwander, B., Amzal, B., Chen, F., Enas, N., Hobbs, B., Ibrahim, J. G., Kinnersley, N., Lindborg, S., Micallef, S., Roychoudhury, S., & Thompson, L. (2014). Use of historical control data for assessing treatment effects in clinical trials. *Pharmaceutical Statistics*, *13*(1), 41–54. https://doi.org/10.1002/pst.1589

Wang, L., Guyatt, G. H., Kennedy, S. A., Romerosa, B., Kwon, H. Y., Kaushal, A., Chang, Y., Craigie, S., de Almeida, C. P. B., Couban, R. J., Parascandalo, S. R., Izhar, Z., Reid, S., Khan, J. S., McGillion, M., & Busse, J. W. (2016). Predictors of persistent pain after breast cancer surgery: A systematic review and meta-analysis of observational studies. *Canadian Medical Association Journal*, *188*(14), E352–E361. https://doi.org/10.1503/cmaj.151276

Westreich, D., Lessler, J., & Funk, M. J. (2010). Propensity score estimation: Neural networks, support vector machines, decision trees (CART), and meta-classifiers as alternatives to logistic regression. *Journal of Clinical Epidemiology*, *63*(8), 826–833. https://doi.org/10.1016/j.jclinepi.2009.11.020

Xiong, C., Han, C., Zhao, D., Peng, W., Xu, D., & Lan, Z. (2021). Postoperative analgesic effects of paravertebral block versus erector spinae plane block for thoracic and breast surgery: A meta-analysis. *PLOS ONE*, *16*(8), e0256611. https://doi.org/10.1371/journal.pone.0256611

Xu, Y., Lu, N., Yue, L., & Tiwari, R. (2020). A Study Design for Augmenting the Control Group in a Randomized Controlled Trial: A Quality Process for Interaction Among Stakeholders. *Therapeutic Innovation & Regulatory Science*, *54*(2), 269–274. https://doi.org/10.1007/s43441-019-00053-x

Yuan, J., Liu, J., Zhu, R., Lu, Y., & Palm, U. (2019). Design of randomized controlled confirmatory trials using historical control data to augment sample size for concurrent controls. *Journal of Biopharmaceutical Statistics*, *29*(3), 558–573. https://doi.org/10.1080/10543406.2018.1559853

Zhao, S., Van Dyk, D. A., & Imai, K. (2020). Propensity score-based methods for causal inference in observational studies with non-binary treatments. *Statistical Methods in Medical Research*, *29*(3), 709–727. https://doi.org/10.1177/0962280219888745