applied sciences

MDPI

Article

A Causal Inference Methodology to Support Research on Osteopenia for Breast Cancer Patients

Niki Kiriakidou [1,*], Aristotelis Ballas [1], Cristina Meliá Hernando [2], Anna Miralles [2], Teta Stamati [1], Dimosthenis Anagnostopoulos [1] and Christos Diou [1]

[1] Department of Informatics and Telematics, Harokopio University, 177-78 Athens, Greece; aballas@hua.gr (A.B.); teta@hua.gr (T.S.); dimosthe@hua.gr (D.A.); cdiou@hua.gr (C.D.)
[2] Instituto de Investigación Sanitaria (INCLIVA), 46010 Valencia, Spain; c.hernandomelia@gmail.com (C.M.H.); amiralles@incliva.es (A.M.)
* Correspondence: kiriakidou@hua.gr

Abstract: Breast cancer is the most common cancer in the world. With a 5-year survival rate of over 90% for patients at the early disease stages, the management of side-effects of breast cancer treatment has become a pressing issue. Observational, real-world data such as electronic health records, insurance claims, or data from wearable devices have the potential to support research on the quality of life (QoL) of breast cancer patients (BCPs), but care must be taken to avoid errors introduced due to data quality and bias. This paper proposes a causal inference methodology for using observational data to support research on the QoL of BCPs, focusing on the osteopenia of patients undergoing treatment with aromatase inhibitors (AIs). We propose a machine learning-based pipeline to estimate the average and conditional average treatment effects (ATE and CATE). For evaluation, we develop a Structural Causal Model for the osteopenia of BCPs and rely on synthetically generated data to study the effectiveness of the proposed methodology under various data challenges. A set of studies were designed to estimate the effect of high-intensity exercise on bone mineral density loss using synthetic datasets of BCPs under AI treatment. Four observational study scenarios were evaluated, corresponding to synthetically generated data of 1000 BCPs with (a) no bias, (b) sampling bias, (c) hidden confounder bias, and (d) bias due to unobserved mediator. In all cases, evaluations were performed under both complete and missing data scenarios. In particular, machine learning-based models based on tree ensembles and neural networks achieved a lower estimation error by 23.8–51.3% and 32.4–89.3% for ATE and CATE, respectively, compared to direct estimation using sample averages. The proposed approach shows improved effectiveness in treatment effect estimation in the presence of missing values and sampling bias, compared to a "traditional" statistical analysis workflow. This suggests that the application of causal effect estimation methods for the study of BCPs' quality of life using real-world data is promising and worth pursuing further.

Keywords: causal inference; treatment effect estimation; osteopenia; breast cancer

Citation: Kiriakidou, N.; Ballas, A.; Hernando, C.M.; Miralles, A.; Stamati, T.; Anagnostopoulos, D.; Diou, C. A Causal Inference Methodology to Support Research on Osteopenia for Breast Cancer Patients. *Appl. Sci.* **2024**, *14*, 9700. https://doi.org/10.3390/app14219700

Academic Editor: Luigi Portinale

Received: 4 September 2024
Revised: 15 October 2024
Accepted: 20 October 2024
Published: 24 October 2024

1. Introduction

Breast cancer is one of the most common malignancies affecting women worldwide. Several effective treatment options have become available during the past years, which significantly improve the disease outlook, depending on tumor subtype and disease stage [1]. Despite their proven effectiveness, cancer treatments and especially long-term adjuvant therapies can negatively impact the quality of life of breast cancer patients (BCPs) [2], including physical and emotional side-effects. One particular example is hormone receptor-positive BCPs undergoing adjuvant endocrine therapy, such as aromatase inhibitor (AI) treatment [3,4]. It has been shown that these treatments lead to bone mineral density (BMD) loss [5], significantly increasing the incidence of fractures in this patient group. Non-pharmacological exercise- and lifestyle-based preventive strategies are preferred, since

these do not affect the patients' outcomes; however, their effectiveness is not clear and further study has been suggested [6]. Real-world data (RWD) offer a rich and valuable information source that can help support clinical research in this direction, especially when considering the effect of non-pharmacological and lifestyle factors on the quality of life of patients.

RWD refers to data produced and collected for non-research purposes, which can also be exploited for providing valuable insights into patients' characteristics, possible treatment patterns, progression of diseases as well as treatment outcomes. Additionally, RWD provide researchers, clinicians and policymakers with the opportunity to identify patterns, trends and associations, which may not be evident in controlled settings, enabling the generation of real-world evidence [7]. This is especially relevant when it comes to research questions related to the quality of life of patients suffering from chronic diseases, as the outcome is determined by multiple interacting variables that are difficult to control.

The most common sources of RWD include information from electronic health records (EHRs) of patients' interaction with the healthcare system, insurance claims, population health surveys, registry and biobank data as well as data obtained via wearable and Internet-of-Things devices, which are used for lifestyle or telehealth purposes [8–10]. It is worth mentioning that such RWD are observational by nature, since they are acquired from existing records, with researchers having no control over the interventions performed on the subjects (if any), in contrast to data generated in controlled clinical trials [11]. On the other hand, RWD present the opportunity to study larger populations, including diverse patient groups.

Over the last years, the potential of RWD has been increasingly explored for inferring causal relationships and measuring treatment effects between variables of interest (see [12] as well as the overviews [13,14]). Despite the wealth of information offered by RWD, they are also subject to several types of bias and quality issues that present challenges in drawing safe and trustworthy inferences. Examples include representation bias (i.e., underrepresentation of certain population subgroups in the available data), measurement bias, missing data, variations in data collection practices, coding errors and inconsistencies across different sources, which are some of the challenges that may affect the validity of the analysis results. Another challenge when using RWD for causal inference is the presence of confounding factors (observed or unobserved), which affect both the outcome and the treatment variables and can lead to incorrect treatment effect estimation if they are not properly handled. Thus, ensuring data quality through data curation processes [15] is a necessary step for reliable analysis and causal inference drawn from RWD.

The REBECCA project (https://rebeccaproject.eu/, accessed on 1 April 2021) aims at leveraging different types of RWD to facilitate research on the quality of life of breast cancer patients (BCPs). This also includes the study of osteopenia and osteoporosis from AI treatment in BCPs, as previously mentioned. In this work, we present the REBECCA data analysis workflow, a general methodology for handling real-world and observational data for research on breast cancer-induced chronic conditions as well the quality of life of BCPs and use the study of osteopenia in BCPs undergoing AI treatment as a use case. The goal is to develop a process for estimating causal effects from observational data suffering from data quality issues, as is the case with RWD.

To facilitate the evaluation of the proposed methodology, it is important to know the data generating process and have access to 'ground truth' data. For this reason, the methodology presented in this paper is evaluated on synthetically generated data. In detail, we generated a synthetic dataset of BCPs under AI treatment using a model developed based on the bibliography, in collaboration with experts from the INCLIVA Health Research Institute in Valencia, Spain.

The main contributions of this work include the following:

- A proposal and detailed presentation of a workflow for handling noisy observational data (such as RWD) to facilitate research on BCPs' quality of life, including machine learning-based (ML) methods for causal effect estimation.

- A causal graph associated with determinants of osteopenia and osteoporosis in patients undergoing AI treatment, which is used for synthetic data generation in the experiments.
- A set of experiments simulating different data analysis challenges, to demonstrate the application and effectiveness of the proposed workflow.

The remainder of this paper is organized as follows: Section 2 introduces the necessary background and a summary of relevant research, while Section 3 provides a description of all stages of the proposed REBECCA data analysis workflow. Section 4 outlines the causal directed acyclic graph on osteopenia/osteoporosis, including the development process. Section 5 provides a thorough presentation of numerical experiments based on a set of use case scenarios for the demonstration of the proposed data analysis workflow. Finally, Section 6 summarizes and discusses the main contributions and limitations of the present paper, along with several interesting directions for future work.

2. Background and Relevant Research

2.1. Causal Effects

The main objective of the proposed methodology is to provide accurate estimations of causal effects from RWD. In detail, we are interested in calculating the effect of an intervention or a treatment on the outcome of interest. For the purpose of our study, we consider the case of a binary treatment variable.

Formally, let $\mathbf{x}_i \in \mathcal{X}$ be the covariates in the covariate space, $t_i \in \{0, 1\}$ the (binary) treatment assignment and $y(\mathbf{x}_i, t_i)$ the outcome of interest for subject or unit i, with $i = 1, 2, \ldots, n$. Each unit can be assigned into a control or treatment group, denoted as $t_i = 0$ and $t_i = 1$, respectively. Under the Neyman–Rubin potential outcome framework [16] and depending on the unit's assigned group, we can either measure the outcome $y(\mathbf{x}_i, 0)$ or the outcome $y(\mathbf{x}_i, 1)$, which stands for the factual outcome for i. The fundamental problem of causal inference is that we cannot observe and measure what the outcome would have been if unit i had been assigned to the other group. This outcome is known as the counterfactual outcome, or simply counterfactual, and needs to be estimated to calculate the effect of the treatment on an individual level. Specifically, the individual treatment effect (ITE) for unit i is defined as the difference of the outcomes in the treatment and control groups,

$$\text{ITE} = y(\mathbf{x}_i, 1) - y(\mathbf{x}_i, 0) \tag{1}$$

The challenge in this case lies in providing an accurate estimate, $\hat{y}(\mathbf{x}_i, t_i)$, of the unobserved counterfactual outcome. For measuring the accuracy of the estimated ITE, we use the Precision in Estimation of Heterogeneous Effect (PEHE), which is defined as follows:

$$\text{PEHE} = \frac{1}{n} \sum_{i=1}^{n} [(y(\mathbf{x}_i, 1) - y(\mathbf{x}_i, 0)) - (\hat{y}(\mathbf{x}_i, 1) - \hat{y}(\mathbf{x}_i, 0))]^2 \tag{2}$$

Given that each sample in the dataset is represented by its covariates, this problem can be approximated via the conditional average treatment effect (CATE), that is

$$\text{CATE}(\mathbf{x}) = \mathbb{E}[y(X, 1) - y(X, 0) \mid X = \mathbf{x}] \tag{3}$$

where X is a random variable corresponding to the vector of covariates for each sample. If estimators $\hat{y}(\mathbf{x}, 0)$ and $\hat{y}(\mathbf{x}, 1)$ are available, then the CATE can be estimated as

$$\hat{\text{CATE}}(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} (\hat{y}(\mathbf{x}_i, 1) - \hat{y}(\mathbf{x}_i, 0)) \tag{4}$$

Additionally, when interested in measuring the causal effects on a population level, the average treatment effect (ATE) is defined by

$$\text{ATE} = \mathbb{E}[y(X, 1) - y(X, 0)] \tag{5}$$

which can be estimated as

$$\tilde{\text{ATE}} = \frac{1}{n_1}\sum_{i=1}^{n_1} y(\mathbf{x}_i, 1) - \frac{1}{n_0}\sum_{i=1}^{n_0} y(\mathbf{x}_i, 0) \tag{6}$$

where n_1 and n_0 are the number of samples in the treatment and the control group, respectively. After the estimation of the missing counterfactual outcomes, the average treatment effect is estimated as

$$\hat{\text{ATE}} = \frac{1}{n}\sum_{i=1}^{n}[\hat{y}(\mathbf{x}_i, 1) - \hat{y}(\mathbf{x}_i, 0)] \tag{7}$$

To quantify the effectiveness of causal effect estimation at the population level (assuming both outcomes are available in the evaluation data), we use the absolute error on the ATE, defined by

$$|\epsilon_{\text{ATE}}| = \left|\frac{1}{n}\sum_{i=1}^{n}[y(\mathbf{x}_i, 1) - y(\mathbf{x}_i, 0)] - \frac{1}{n}\sum_{i=1}^{n}[\hat{y}(\mathbf{x}_i, 1) - \hat{y}(\mathbf{x}_i, 0)]\right| \tag{8}$$

2.2. Randomized Controlled Trials for Causal Inference

Randomized Controlled Trials (RCTs) are considered to be the gold standard for inferring cause-and-effect relationships between a treatment or an intervention and an outcome of interest [17], since the randomization of participants in the control and treatment group cancels out the effects of confounding variables. This ensures that any observed differences between the two groups are due to the applied intervention.

There are several limitations and disadvantages to conducting RCTs. They require significant cost in terms of resources and time [18], introducing limitations on the number of enrolled participants as well as on the total duration of the studies. Furthermore, RCTs are often conducted in controlled settings, which may not reflect real-world conditions, leading to unrepresentative results for the whole population [19]. Finally, ethical restrictions are also a significant barrier for conducting RCTs, especially if the assignment of participants in the control group could result in mistreatment [17] (e.g., if the intervention already has a known benefit for participants).

Such limitations are particularly pronounced in studies related to the quality of life of BCPs, since they largely depend on the behavior of patients in their daily lives, e.g., their physical activity habits and the available psychosocial support [2], which cannot be easily controlled. Therefore, researchers seek ways to leverage readily available observational RWD to study factors affecting the quality of life of BCPs and to estimate the effects of interventions.

2.3. Real-World Data

RWD are often accompanied by various challenges and complications, spanning from data quality management to data analysis methods. A characteristic of RWD is that they are observational, in contrast to data gathered from RCTs, and we are able to induce various forms of bias in the analyses.

2.3.1. RWD and Bias

RWD can suffer from several types of biases including (but not limited to) systematic, sampling/representation and confounding bias.

A representative example of systematic bias is the underestimation of physical activity if monitoring devices (such as mobile phones and smartwatches) are not used continuously. Sampling or representation bias may also be included in the data, since subgroups of patients in the underlying population are often under-represented or over-represented.

An additional type of bias commonly encountered in observational RWD is confounding bias [20]. A confounder is a variable which affects both the intervention or treatment variable as well as the study variable and, hence, there is a distortion in the measure of

association between the variables of interest due to the additional effect of the confounder. It is also important to be aware of possible unobserved confounders, which might affect the causal relationship of the intervention and study variables.

To address these issues, several methods for bias estimation have been proposed in the literature [21]. Through bias estimation, the sources of bias can be identified and appropriate mitigation measures can be considered. These methods enable us to quantify bias in an inference model and assess the reliability of the model.

2.3.2. RWD and Data Quality

Another big challenge in working with RWD is missing data, which can occur due to both human (e.g., omission, neglect, and lack of time) and technical limitations or deficiencies (e.g., device communication and stopped recordings). A common example in the case of research on BCP quality of life includes patients who choose to not answer quality of life questionnaires, or answer only a subset of the questions. These types of errors are often handled using statistical and machine learning imputation strategies.

Another common issue is data heterogeneity, where data originating from different structured or unstructured sources do not use the same encoding, even when referring to the same variables. Such data need to be mapped to a common representation format for analysis. Data heterogeneity may also originate from the use of different measurement methods and/or devices, creating the need for robust and reliable frameworks [22]. For example, IMU sensors of commercial smartwatches have different sampling rates, which may lead to differences in physical activity estimates.

2.4. Structural Causal Models

One approach to the analysis of observational data, including RWD, is the use of Structural Causal Models (SCMs) [23]. SCMs are mathematical models that answer questions of causality and conceptualize the hidden underlying "causal story" of a dataset. SCMs consist of two sets of variables, the exogenous variables U, whose values are determined by external factors, and the endogenous variables V, whose values are determined by the model. Moreover, SCMs include a set of functions f that assigns each variable in V a value, based on the values of the other variables in the model [24]. SCMs correspond to directed acyclic graphs (DAGs), where each node of the SCM represents a variable and each edge connecting two nodes represents a causal relationship between the variables. Using a SCM, a practitioner or a researcher can translate how and which features for a specific case interact with each other.

SCMs have been previously applied for the analysis of observational data in various disciplines. For example, Arif and MacNeil [25] used SCMs as a framework to overcome the limitations of inferring causality from observational data in the field of ecology. They highlighted that the use of statistical analysis of observational data leads to biased estimation of causal effects and instead proposed the use of DAGs to represent the causal structure of each problem. By applying graphical rules, such as the backdoor and frontdoor criteria [23], researchers can determine the necessary statistical adjustments for establishing causal relationships from observational data. Finally, they present simulated examples in which they demonstrate how the use of these criteria can provide accurate causal estimates in the field of ecology, without relying on randomized experiments.

Reinhold et al. [26] presented the application of SCMs in precision medicine. In particular, they highlighted the necessity of answering causal questions (e.g., what is the response of a patient when different treatments are issued) and mentioned that SCMs can be leveraged for causal reasoning. The authors focused on the development of a SCM that captures the interactions between demographic information, disease covariates and magnetic resonance (MR) images of the brain in patients with multiple sclerosis. Using the SCM, the authors generated counterfactual images depicting how an MR image of the brain would appear if certain demographic or disease factors were altered. These counterfactual

images have applications in modeling disease progression and can be utilized in downstream image processing tasks where it is necessary to control for confounding variables.

Petersen and van der Laan [27] highlighted the importance of asking causal questions in the field of epidemiology and underlined the potential benefits of using formal frameworks for causal inference. In particular, the significance of causality lies in the fact that we can identify and intervene with the best way in the root cause of something, rather than simply spotting a pattern or a behavior. Following this reasoning, the authors focused on using SCMs, since their flexibility enables them to state valid assumptions and, thus, lead to the development of a causal model which can describe the true data-generating process. Overall, they aimed to emphasize the utility of causal thinking when conducting a statistical analysis, to clarify the capabilities and limitations of SCMs as well as to provide an overview of SCMs as a powerful tool offered to be used in the field of epidemiology. Finally, it is clearly stated that the judicious use of a causal framework can significantly enhance the quality of epidemiological research and improve research in various disciplines relying on statistics to understand how the world functions.

2.5. Machine Learning Models for Treatment Effect Estimation

Alongside SCMs, purely data-driven machine learning models have been proposed in the bibliography to directly estimate the various types of treatment effects, introduced in Section 2.1, without the use of DAGs.

Following a Bayesian non-parametric approach, Chipman et al. [28] developed a Bayesian "sum-of-trees" model, named Bayesian Additive Regression Trees (BART). The BART approach fits a parameter-rich model by using a strongly influential prior distribution. Specifically, the model uses the sum of trees to approximate the average value of the outcomes given a set of covariates, $\mathbb{E}[Y \mid \mathbf{x}]$. The main idea behind BART is to impose a prior, which regularizes the fit by keeping the individual tree effects small in order to elaborate the sum-of-trees model. Additionally, BART uses a tailored version of Bayesian back-fitting Markov Chain Monte Carlo [29] for fitting the sum-of-trees model.

Künzel et al. [30] proposed two methodologies, named S-learner and T-learner, for the accurate estimation of treatment effects. S-learner estimates the outcomes by using all the features of the dataset, along with the treatment as an additional feature. Notice that the treatment indicator is handled by the base learner like any other feature; hence, it does not play any special role in the estimation of the effects. In contrast, T-learner utilizes two independent models: one trained on the treated data to predict outcomes under treatment and the other one on the control data to predict outcomes without treatment. Next, the treatment effect is then estimated by subtracting the predicted outcome for the control model from the predicted outcome for the treatment model. In the literature, a variety of causal inference models were proposed based on T-learner and S-learner methodologies using a variety of ML models as base learners, providing some interesting results [30,31].

Shalit at al. [32] proposed a new framework for estimating individual treatment effects, named Counterfactual Regression (CFR). This framework uses a prediction model, focusing on learning a balanced representation of the control and treatment groups. Under this balanced representation, the distributions of the two groups are considered to be similar. Specifically, the authors used two different integral probability metrics, Wasserstein (Wass) distance [33] and Maximum Mean Discrepancy (MMD) [34], for calculating the distances between the treatment and control distributions, as well as proposed a generalization bound for estimating the individual treatment effect. In addition, the authors demonstrated the performance of the proposed models, CFR (MMD) and CFR (Wass), which stand for MMD and Wass distances, respectively, and compared their performance with state-of-the-art models. Finally, they proposed a variant without balance regularization, the neural network-based model TARNet.

Shi et al. [35] proposed a neural network-based model for the estimation of treatment effects, named Dragonnet. The authors mainly focused on the use of observational data

for treatment effect estimation, giving special attention to the average and individual level. The proposed model aimed to improve these estimations by exploiting the propensity score's sufficiency. Additionally, the authors proposed targeted regularization, which is based on non-parametric estimation theory and focuses on reducing the bias of the estimator, therefore further improving the estimation of treatment effects. The presented experiments provided evidence about the superiority of Dragonnet over TARNet, CFR (MMD) and CFR (Wass) on a variety of challenging causal inference benchmark datasets.

In more recent works, Kiriakidou and Diou [36,37] proposed NN-Dragonnet for estimating treatment effects, which consists of a modification of the neural network-based model Dragonnet. The rationale behind the development of this model is to capture information not only from the covariates of the samples, but also from the average outcomes of neighboring instances from both treatment and control groups. In simple words, the proposed model utilizes the average of the nearest outcomes of each instance from both control and treatment groups along with the covariates as inputs. The authors evaluated the performance of NN-Dragonnet on three collections of datasets using three different Minkowski distance metrics (i.e., Euclidean, Manhattan and Chebyshev) for the calculation of nearest neighboring instances. Their numerical experiments demonstrated that NN-Dragonnet achieves lower PEHE values than Dragonnet and TARNet models, which implies that it can be used for improved estimation of the CATE.

3. Proposed Methodology

To address the limitations and challenges underlined in Section 2, we propose a methodology for using data collected in uncontrolled settings, as is the case with RWD, for clinical research purposes. This methodology has been developed in the context of the REBECCA project and is summarized in Figure 1.

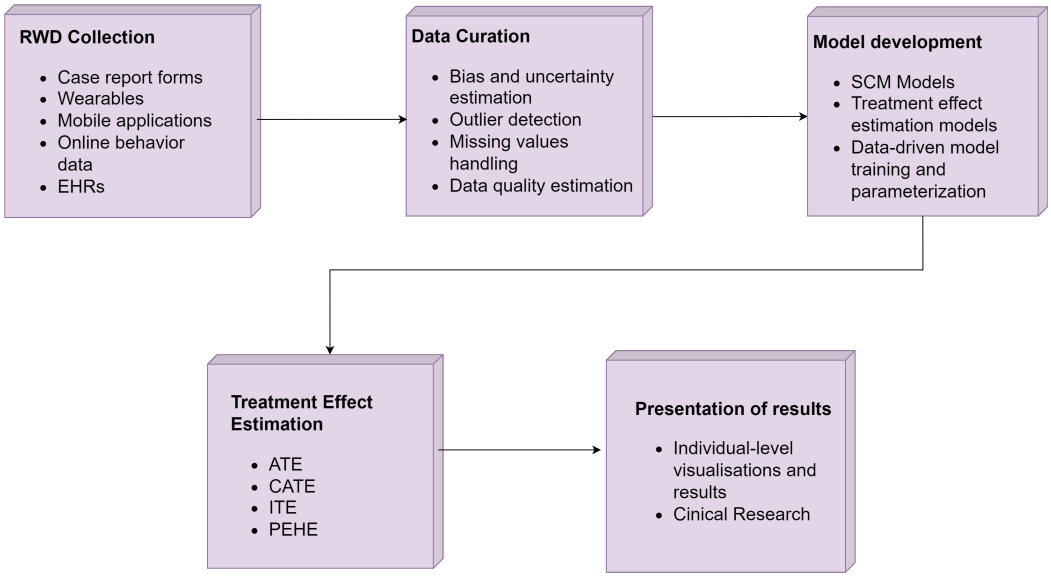

Figure 1. REBECCA data analysis workflow.

3.1. Stage 1: Collection of Real-World Data

The first stage of the proposed data workflow is *"RWD collection"*. Data sources include case report forms, smartwatch and smartphone measurements as well as elements of online activity data collected from BCPs.

Specifically, the case report forms consist of clinical variables providing background information for each patient, including demographic data, tumor characteristics and the associated treatment, medication information, additional medical and immunization history and finally results of clinical and medical examinations. In addition, data can include

questionnaires, which are used to measure multidimensional aspects of BCPs' quality of life, including their physical, social, emotional as well as functional well-being.

Regarding the smartwatch, smartphone and online activity monitoring data, these are used along with indicator extraction algorithms [38–40] to provide objective measurements of the individual physical and online behavior of the patients.

Data collection may include different sources with different data representations. All data are mapped to a common data model and representation format, which includes all variables of interest.

3.2. Stage 2: Data Curation

The second stage of the proposed data workflow is the "*Data curation*" stage. This is an essential step in the process of data handling, since it involves removing errors and inconsistencies in the data to ensure that they are accurate and reliable so the analysis can be continued. These errors may include typos, missing values and outliers, while inconsistencies can include duplicate data or an inconsistent data format. Many of these issues can be addressed in an automated way. The following functionalities can facilitate this stage:

1. Missing data imputation: Several machine learning-based imputation strategies exist that attempt to estimate the missing values, such as Bayesian Ridge Regression [41], *k*-Nearest Neighbors [42], Random Forests [43] and Extra-Trees [44].
2. Outlier detection: These methods identify data anomalies, which significantly differ from other observations or data trends and can be attributed to measurement errors or data entry errors. In the literature, the Isolation Forest [45], Local Outlier Factor [46] and One-Class Support Vector Machine [47] algorithms constitute some of the most popular outlier detection methods.
3. Partial measurement estimation: In several cases, the overall behavior of an individual needs to be inferred from partial measurements. For instance, users may provide measurements through smartwatches for only part of the day, or may not use their smartwatch for some activities, thus requiring estimation of overall physical activity.
4. Measurement error and bias quantification: This component augments measurements with metadata related to the possible error and bias of the measurements. This information will assist in subsequent data analysis tasks, e.g., by providing confidence intervals through sensitivity analysis.

The outcome of the "Data curation" stage is curated datasets, which can be used for data analysis and statistical and causal inference.

3.3. Stage 3: Model Development

The third stage of the REBECCA workflow is "*Model Development*", which is focused on encoding existing domain knowledge into a directed acyclic graph (DAG), which in turn represents the causal associations between the variables under consideration in the study. The developed DAG is then used in combination with the data available from the previous stage to develop a Structural Causal Model (SCM). The resulting SCM quantifies, in a functional way, the causal variable relations.

This stage may also incorporate purely data-driven ML models, which are trained for estimating the ATE and CATE, using the available collected data.

3.3.1. SCM for Osteopenia and Osteoporosis

Figure 2 presents the developed DAG for the "Osteopenia and Osteoporosis" use case, which aims to study various factors contributing to bone mineral density (BMD) loss (and increased risk of fractures) in BCPs, especially for those who undergo AI treatment. For measuring the variable BMD loss, we consider the *T*-score. The value of the *T*-score [48] on a patient's bone density report suggests the difference, in standard deviations, from the bone mineral density of the average of healthy 30-year-old women. According to the World Health Organization (WHO), normal ranges are between −1.0 and +4, while a *T*-score

between −2.5 and −1.0 is considered to indicate low bone density, or else osteopenia. Finally, a *T*-score below −2.5 indicates that the patient has osteoporosis. The functional relationships between the linked variables are learned from the available data, resulting in a model that can be used to assess the effect of various QoL-related parameters on BMD loss. Details regarding the DAG are provided in Section 4.1 and in the Appendix A. In this paper, the DAG is used for data generation purposes, to synthesize the dataset used in the experiments.

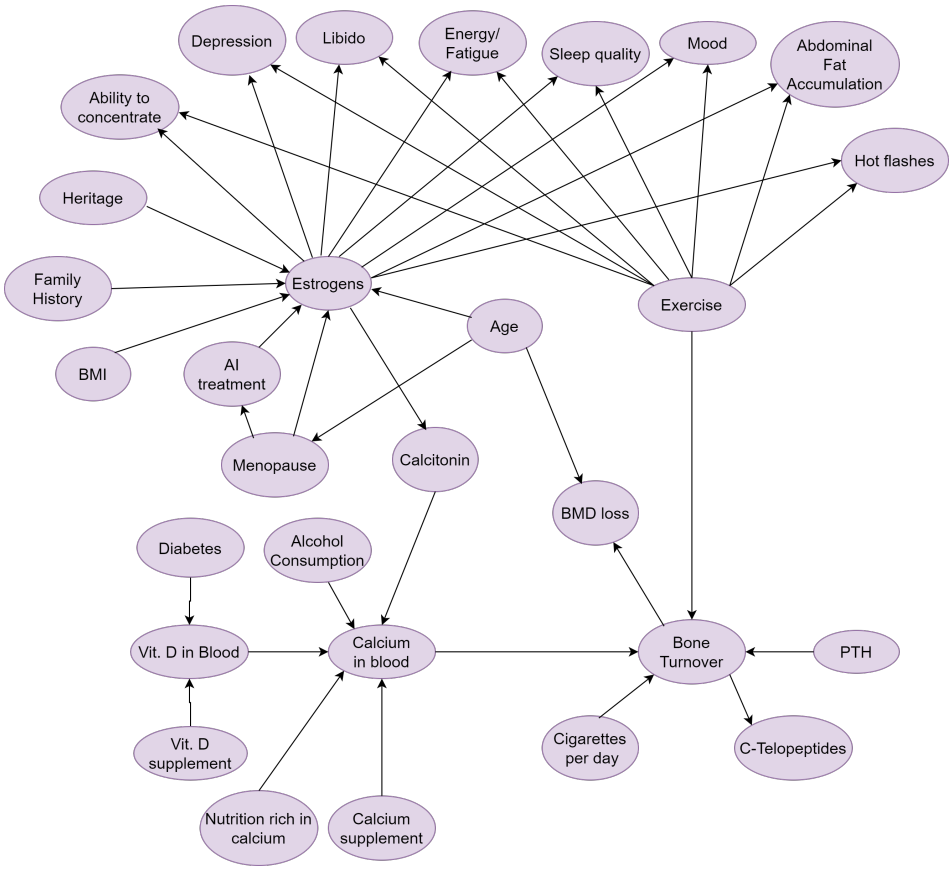

Figure 2. Causal DAG for osteopenia/osteoporosis. A causal directed acyclic graph (DAG) is a graphical tool used for visually representing the causal connections between a set of variables. In our case, the set of variables are relevant to the clinical study for osteopenia/osteoporosis, as a result of treating breast cancer with aromatase inhibitors, in the context of the REBECCA project.

3.3.2. Data-Driven Causal Inference Models

Besides the SCMs, the proposed workflow also includes ML-based approaches for treatment effect estimation, ranging from traditional linear models to tree-based (R-Forest and BART) and neural network-based models (TARNet, Dragonnet and NN-Dragonnet). In contrast to SCMs, ML models do not directly encode any knowledge about the causal relationships between variables and aim at directly estimating the values of the outcome with and without treatment, given a set of known covariates. Depending on the level of existing knowledge about the underlying causal mechanisms, these values may prove to be more accurate than SCMs for the estimation of causal effects, at the cost of lower interpretability. The decision on the approach to use (SCM, possibly with model training to learn the function at each node vs purely data-driven models) depends on the specific problem and on data availability.

3.4. Stage 4: Treatment Effect Estimation

The fourth stage of the REBECCA workflow is the "*Treatment Effect Estimation*" stage. In this stage, the trained models developed in Stage 3 are used to estimate the ATE (at population level) or CATE (for specific patients or patient subgroups).

3.5. Stage 5: Presentation of Results

The fifth and final stage of the REBECCA workflow is the "*Presentation of results*". We distinguish between the following cases:

- **Clinical research:** In this case, the models are used directly to provide estimates of the ATE/CATE. Additional information may also be presented to the researcher, such as descriptive and inferential statistics for different population subgroups based on the available data.
- **Patient management:** In this case, the models' estimations can support clinicians in identifying patients who are in need of provider contact (e.g., have a high estimated BMD loss), as well as to identify optimal treatments for those patients (e.g., lifestyle changes or pharmacological interventions). Offering such resources regarding patients for review to clinicians was proposed by several researchers [49–51].

4. Use Case: Breast Cancer-Related Osteopenia and Osteoporosis

As a representative use case of the proposed RWD analysis workflow, we focus on the issues of osteopenia and osteoporosis as a result of adjuvant breast cancer treatment with AIs [52]. This treatment, commonly provided to postmenopausal women with breast cancer, lowers their estrogen levels and leads to decreased bone density [52,53]. Consequentially, such patients have more fragile bones compared to women who do not receive AI treatment and are therefore susceptible to bone fractures. The use case of osteopenia and osteoporosis in breast cancer has been discussed in several research works as in [54–57].

4.1. Directed Acyclic Graph

Figure 2 presents the developed DAG for the use case of osteopenia/osteoporosis. For the development of the DAG, both researchers and specialized clinicians collaborated to model existing knowledge regarding the interactions between the studied variables.

Firstly, a literature review was performed regarding the comorbidity of osteopenia and osteoporosis in BCPs treated with AIs. Next, an interdisciplinary team of technical and clinical research scientists participating in the REBECCA project discussed the main findings of this review and defined the variables that should be considered for studying BMD loss.

All the information and knowledge discussed between the technical and clinical partners of REBECCA were transformed into the DAG presented in Figure 2. Each node in the DAG represents a variable and each edge indicates the direction of its relationship. The reader can refer to Appendix A for a detailed description of the links between the connected variables.

It is worth emphasizing that the underlying mechanism of bone turnover, which is the ratio of bone formation to bone resorption, is not fully understood. Therefore, the proposed model is a simplified representation aimed to act as a decision support and analysis tool and not as a model offering a detailed depiction of reality. Finally, the developed DAG is used only for the generation of the datasets used in Section 5 and not as a tool for the analysis in the presented experiments.

4.2. Synthetic Datasets

In the context of the REBECCA project, we developed a synthetic data generator (SDG) for the generation of synthetic but realistic real-world datasets.

To generate data, the first step is to define the functional relationships between the variables of the DAG of Figure 2. By taking into account the causal connections between

the variables of the DAG and by defining their functional relationship, the generator yields datasets, which can simulate RWD distributions and observational data.

One of the basic functionalities of the SDG is that each dataset entry, which corresponds to a simulated individual, is associated with the patient's "personal graph". By storing the values of each variable in a personal graph, we are able to simulate and generate counterfactual data. This functionality also enables us to intervene on an arbitrary number of variables, which in turn triggers the recalculation of each value connected to and affected by the intervention.

5. Experimental Analysis

In this section, we present a series of experiments for evaluating the ability of the proposed approach for supporting the QoL of patients with osteopenia undergoing treatment with AIs. Our scope is to compare the performance of state-of-the-art ML models, ranging from traditional linear models to tree-based and neural network-based models, in estimating treatment effects.

For conducting the experiments, we use synthetic datasets, which were produced in a completely controlled setting by the SDG, based on the causal DAG developed in the "Model Development" stage (see Section 4.2), for the case of osteopenia and osteoporosis studied as a comorbidity of breast cancer treatment with aromatase inhibitors.

At this point, it is worth recalling that the primary reason for utilizing synthetic datasets in our research is that it is impossible to observe the causal effect of a single unit, since only the factual outcome can be measured, which constitutes the fundamental problem of causal inference [16]. In detail, a BCP can either receive or not receive a specific treatment or an intervention and, thus, we cannot calculate the real causal effect of the particular treatment on the patient. Therefore, when using RWD for causal inference, we cannot observe but only estimate the missing counterfactual outcomes and then estimate the causal effects [58,59].

Next, we present the models, which are included and evaluated in the REBECCA workflow and are characterized as some of the most effective and widely used models for the estimation of causal effects:

- "LR1", which stands for the "S-learner" methodology proposed by Künzel et al. [30], using Linear Regression as a base learner [60].
- "LR2", which stands for the "T-learner" methodology proposed by Künzel et al. [30], using Linear Regression as a base learner [60].
- "R-Forest", which stands for the "S-learner" methodology proposed by Künzel et al. [30], using Random Forest as a base learner [43].
- "BART", which stands for the model proposed by Chipman et al. [28].
- "TARNet", which stands for the model proposed by Shalit et al. [32].
- "Dragonnet", which stands for the Dragonnet model proposed by Shi et al. [35].
- "NN-Dragonnet (C)", which stands for the model proposed by Kiriakidou et al. [36,37], using Chebyshev distance for calculating the average of the nearest instances' outcomes.
- "NN-Dragonnet (E)", which stands for the model proposed by Kiriakidou et al. [36,37], using Euclidean distance for calculating the average of the nearest instances' outcomes.
- "NN-Dragonnet (M)", which stands for the model proposed by Kiriakidou et al. [36,37], using Manhattan distance for calculating the average of the nearest instances' outcomes.

The implementation code was written in Python 3.10 and executed on a PC (3.2 GHz Quad-Core processor, 32 GB RAM) using the Windows operating system. We experimented with several configurations of each causal inference model through a grid hyperparameter search. For example, all neural network-based models (i.e., TARNet, Dragonnet and NN-Dragonnet) were evaluated using a varying number of neurons in the hidden layers as well as different learning rates. In addition, all versions of NN-Dragonnet were evaluated with several values of parameter k (number of neighboring instances) ranging from 5 to 15, while R-Forest was evaluated with different values of the "number of trees" and "max depth" hyperparameters, which are reported in Appendix C.

The performance of all causal inference models was measured using the absolute error of the ATE $|\epsilon_{\text{ATE}}|$ and the Precision in Estimation of Heterogeneous Effect (PEHE) metrics, which are, respectively, defined by (2) and (8). It is worth noticing that the PEHE assesses the precision of the causal inference models for estimating the ITE by measuring the accuracy of individualized predictions. A lower value indicates that the model has better precision in predicting the benefits of a treatment on an individual; hence, clinicians are able to provide more informed decisions for targeted personalized care.

The primary aim is to provide an answer to the following research question: "*What is the effect of exercising with high intensity on the bone mineral density loss of BCPs?*". By taking into consideration the concept of treatment effect estimation for the above statement, we consider the variable "Exercise" as the treatment variable T and the variable "BMD loss" as the outcome variable Y.

Notice that since we generated synthetic datasets for studying the case of osteopenia and osteoporosis, it is possible to calculate the *real* causal effect, as the counterfactual outcome of each individual is calculated using the SDG (see Section 4.2).

This is referenced as *ground truth* throughout the section. Following the ground truth calculation, we conduct several experiments simulating different use case scenarios that include ideal datasets, with either biases in the data, as well as consider hidden confounders or a hidden mediator.

5.1. Experimental Setup

Using the synthetic data generator described in Section 4.2, we generated synthetic datasets consisting of 500 patients and also calculated the corresponding counterfactual outcomes of those BCPs. Notice that for calculating the counterfactual outcome of each sample, we re-calculated all of its variable values by assigning a change in the intensity of their physical activity, which is the treatment variable, while keeping the same value of the noise variables. In detail, we changed from high-intensity exercise to low-intensity exercise and vice versa, depending on the initial assignment of each individual. The ground truth effect of high-intensity exercise on the BMD loss of BCPs is calculated as 0.58, i.e., intense physical activity inhibits the loss of BMD, as has been underlined in the literature [61–63].

We direct the reader to Appendix B for a detailed description of the implemented functional relationships of the variables as well as to https://github.com/kiriakidou/A-Causal-Inference-Methodology-to-Support-Research-on-Osteopenia-for-Breast-Cancer-Patients (accessed on 1 April 2021), for the generated datasets. At this point, it is worth mentioning that since no RWD are available, the utilized noises in our research were defined using input from domain experts in the REBECCA project.

For each generated dataset, we consider two cases: (a) a non-imputed dataset in which no missing values exists and (b) an imputed dataset in which 30% of entries from the "estrogens" and "calcitonin" variables were randomly removed for simulating the case of the missing data problem, commonly encountered in practice when working with RWD. Lacking any additional evidence, we did not make any assumptions about the mechanisms underlying missingness in the generated data, so the data are Missing Completely at Random (MCAR) [64], i.e., the likelihood of a data point being missing is independent of both observed and unobserved data. This approach is commonly used when evaluating ML models with missing data [65].

As regards data imputation, we experimented with several imputation techniques such as utilization of the mean value for each feature, k-NN for various values of parameter k and Random Forest; nevertheless, Extra-Trees was selected, since it provided the best overall performance for all causal inference models. In addition, the employment of any imputation technique does not affect the obtained findings in our experiments since similar conclusions could be made from the employment of any of the utilized imputation techniques ("Data Curation" stage of the proposed workflow). Finally, we highlight that 80% of the data were used for training the models (in-sample), while the rest 20% was used as a hold-out for evaluation (out-of-sample) [66].

In the following text, we design and perform experiments considering four different case studies for highlighting the complications that occur when using RWD for causal inference.

5.1.1. Case 1: Unbiased Sampling

For the design of this experiment, we generated two similar groups of BCPs and randomly assigned them into control and treatment groups. This assignment resembles an RCT study, where the first group consists of 500 individuals with no exercise or low-intensity exercise habits, while the second group consists of 500 individuals who are performing high-intensity exercise sessions.

Apart from the physical activity level, all variables follow the same distribution across the two groups. Additionally, for this unbiased case the value of the ATE is calculated to be 0.58, which is identical to the ground truth effect and, hence, $|\epsilon_{\text{ATE}}| = 0$.

Finally, the PEHE of the average is 3.42 and 4.25 for the train and the test set of the non-imputed dataset and 2.67 and 3.56 for the train and the test set of the imputed dataset, respectively.

5.1.2. Case 2: Observational Study with Bias

This case simulates a common scenario, in which data suffer from selection bias. For the generation of the dataset, we created two groups of patients, each one comprising 500 individuals. The difference compared to the unbiased scenario is that 80% of the patients that belong to the high-intensity exercise group (i.e., treatment group) have been prescribed calcium supplementation. In contrast, only 20% of the patients who belong to the low-intensity exercise group (i.e., control group) receive calcium supplementation. In general, clinicians and medical researchers suggest that a reasonable proportion of patients who should receive calcium supplementation is typically between 50% and 80% [67], especially in populations at higher risk of bone density loss, such as older adults or postmenopausal women. The clinical experts from the REBECCA project suggested that in the generated synthetic data, 80% of the patients that belong to the high-intensity exercise group (i.e., treatment group) should be receiving calcium supplementation, as this aligns with current guidelines and practices aimed at maximizing bone health outcomes in these vulnerable populations [68].

In this case, the direct calculation of the ATE is 0.87, which is an overestimation of the ATE compared to the real 0.58 value; hence, $|\epsilon_{\text{ATE}}| = 0.29$. This result is reasonable, since in this case the ATE measures the combined effect of a higher concentration of calcium in the blood along with high-intensity physical activity, leading to a higher reduction in BMD loss.

Finally, the PEHE is 2.74 and 2.72 for the train and the test set for both non-imputed and imputed datasets, respectively.

For the selection of the variables used to conduct the experiments based on Cases 1 and 2, we use the DAG in Figure 2, developed for the use case of osteopenia and osteoporosis. In detail, our decision rests on the condition that the selected variables should satisfy the backdoor criterion [23].

The subset of the selected variables are as follows: “Exercise”, “PTH” , “Age”, “Calcium in blood”, “C-Telopeptides”, “Alcohol consumption”, “Calcium supplement”, “Calcitonin”, “Nutrition rich in calcium”, “Vit.D in Blood” as well as “Cigarettes per day”.

Finally, the target variable is the Bone Turnover, which directly leads to BMD loss.

5.1.3. Case 3: Observational Study with Hidden Confounders

With regard to the third experiment, we use the biased dataset produced for the previous case (Case 2), but with the main difference that we do not measure the variables “Calcium supplementation”, “Nutrition rich in calcium”, “Alcohol consumption” and “Vit. D in Blood”.

As shown in Figure 2, all these variables directly affect the variable "Calcium in blood", but since they are not measured they are considered to be hidden confounders. The result of direct computation of the ATE equals 0.87 and therefore, $|\epsilon_{\text{ATE}}| = 0.29$.

5.1.4. Case 4: Observational Study with Unobserved Mediator

Regarding this experiment, we are not measuring the variable "Calcium in Blood", which is a mediator between the variable "Bone Turnover" and "Calcium supplement", "Alcohol consumption", "Calcitonin", "Vit. D in Blood" as well as "Nutrition rich in calcium". These variables can be proxies for "Calcium in blood" and, therefore, can be used to capture the bias present in the dataset. The direct estimation of the ATE equals 0.87, which implies that $|\epsilon_{\text{ATE}}| = 0.29$.

5.2. Numerical Experiments

Tables 1 and 2 report the performance of the ML models in terms of $|\epsilon_{\text{ATE}}|$ and the PEHE in both in-sample (i.e., counterfactual estimation of the training samples) and out-of-sample (i.e., factual and counterfactual estimation on the held-out test data), respectively, relative to the four different cases presented in the experimental setup.

Table 1. Experimental results on train set. Top results are specified in bold, while second best are underlined. The calculated ground truth ATE of "intense exercise" on "BMD loss" is 0.58.

Model	**Case 1** $\|\epsilon_{\text{ATE}}\| = 0$		**Case 2** $\|\epsilon_{\text{ATE}}\| = 0.29$		**Case** 3 $\|\epsilon_{\text{ATE}}\| = 0.29$		**Case 4** $\|\epsilon_{\text{ATE}}\| = 0.29$	
	$\|\epsilon_{\text{ATE}}\|$	**PEHE**	$\|\epsilon_{\text{ATE}}\|$	**PEHE**	$\|\epsilon_{\text{ATE}}\|$	**PEHE**	$\|\epsilon_{\text{ATE}}\|$	**PEHE**
LR1	0.154	3.506	0.141	2.727	0.155	2.727	0.189	2.727
LR2	**0.007**	3.305	0.150	2.342	0.158	2.386	0.194	2.653
R-Forest	0.066	3.062	0.175	**2.186**	0.177	**2.183**	0.118	**2.265**
BART	0.427	3.221	0.160	2.490	0.165	2.435	0.169	2.727
TARnet	0.055	3.295	0.166	2.262	0.150	2.281	0.095	2.641
Dragonnet	0.050	**3.041**	0.221	2.334	0.196	2.249	0.079	2.645
NN-Dragonnet (C)	0.159	3.263	0.106	2.260	0.140	2.240	0.093	2.734
NN-Dragonnet (E)	0.246	3.389	0.104	2.261	0.208	2.216	0.110	2.740
NN-Dragonnet (M)	0.257	3.390	**0.094**	2.261	**0.126**	2.242	**0.031**	2.656

(a) Non-Imputed Dataset

Model	**Case 1** $\|\epsilon_{\text{ATE}}\| = 0$		**Case 2** $\|\epsilon_{\text{ATE}}\| = 0.29$		**Case** 3 $\|\epsilon_{\text{ATE}}\| = 0.29$		**Case 4** $\|\epsilon_{\text{ATE}}\| = 0.29$	
	$\|\epsilon_{\text{ATE}}\|$	**PEHE**	$\|\epsilon_{\text{ATE}}\|$	**PEHE**	$\|\epsilon_{\text{ATE}}\|$	**PEHE**	$\|\epsilon_{\text{ATE}}\|$	**PEHE**
LR1	0.154	2.750	0.143	2.727	0.157	2.727	0.191	2.727
LR2	**0.019**	2.328	0.152	2.341	0.160	2.386	0.196	2.648
R-Forest	0.064	2.103	0.175	**2.183**	0.174	**2.190**	0.141	**2.292**
BART	0.462	2.508	0.147	2.378	0.169	2.477	0.150	2.727
TARnet	0.059	2.338	0.164	2.252	0.148	2.271	0.087	2.624
Dragonnet	0.048	2.091	0.223	2.320	0.233	2.289	**0.077**	2.63
NN-Dragonnet (C)	0.106	2.238	0.105	2.248	0.140	2.250	0.101	2.726
NN-Dragonnet (E)	0.027	**2.076**	0.104	2.254	0.134	2.248	0.111	2.733
NN-Dragonnet (M)	0.053	2.096	**0.097**	2.250	**0.127**	2.210	0.102	2.725

(b) Imputed Dataset

Regarding Case 1, the direct estimation of the ATE using Equation (7) leads to an estimated ATE of 0.58. In detail, in Case 1 there are no biases in the dataset and thus the estimated effect coincides with the ground truth ATE, as it would have been calculated through an RCT. On the other hand, the direct calculation of the effect of high-intensity exercise on BMD loss for the biased datasets is 0.87, which is an overestimation of the ATE compared to the real 0.58 value. The deviation in results compared to the ground truth is reasonable, as in this case the ATE measures the combined effect of a higher concentration

of calcium in the blood, along with high-intensity physical activity, leading to a higher inhibition of BMD loss.

The interpretation of Tables 1 and 2 reveals that ML models achieve better estimation of the ATE (and CATE) in observational studies, as opposed to the direct estimation of the causal effect from the dataset. Therefore, the benefit of using these algorithms for causal effect estimation is apparent, as the estimation errors are considerably lower. Regarding to the estimation of the PEHE of each model, they are all close to the calculation of the PEHE for the average of the population, as it is calculated on the unbiased and biased dataset. Furthermore, it is worth noticing that in both non-imputed and imputed datasets, there is no significant difference in the estimation of the ATE in Cases 2 and 3, where the "Calcium supplementation", "Nutrition rich in calcium", "Alcohol consumption" and "Vit. D in Blood" are confounders, as there is still in the dataset the variable "Calcium in blood" acting as a proxy variable. Finally, the estimations of the ATE in Cases 2 and 4 are also very close, as the hidden variable "Calcium in blood" is represented by the variables "Calcium supplement", "Alcohol consumption", "Calcitonin", "Vit. D in Blood" and "Nutrition rich in calcium" in the dataset.

In practice, these results indicate that direct calculation of the ATE using the datasets of Cases 2–4 would lead to a significant overestimation of the effect of high-intensity physical activity on BMD loss, due to its confounding with calcium supplementation, while methods such as NN-Dragonnet would lead to much lower estimation error (e.g., 0.094 instead of 0.29 for Case 2). Such differences are important, especially given that BC patients often suffer from additional conditions, such as cardiovascular disease or chronic fatigue. Having accurate information about the risks and benefits of each treatment on each of the multiple existing conditions enables clinicians to reach optimal treatment decisions.

Table 2. Experimental results on test set. Top results are specified in bold, while second best are underlined. The calculated ground truth ATE of "intense exercise" on "BMD loss" is 0.58.

Model	**Case 1** $\lvert\epsilon_{\text{ATE}}\rvert = 0$		**Case 2** $\lvert\epsilon_{\text{ATE}}\rvert = 0.29$		**Case** 3 $\lvert\epsilon_{\text{ATE}}\rvert = 0.29$		**Case 4** $\lvert\epsilon_{\text{ATE}}\rvert = 0.29$	
	$\lvert\epsilon_{\text{ATE}}\rvert$	**PEHE**	$\lvert\epsilon_{\text{ATE}}\rvert$	**PEHE**	$\lvert\epsilon_{\text{ATE}}\rvert$	**PEHE**	$\lvert\epsilon_{\text{ATE}}\rvert$	**PEHE**
LR1	0.154	4.378	**0.141**	2.762	**0.155**	2.756	0.189	**2.743**
LR2	0.062	3.896	0.201	**2.336**	0.220	2.330	0.157	2.790
R-Forest	0.128	3.912	0.281	2.453	0.285	2.463	0.088	3.179
BART	0.427	3.966	0.175	2.565	0.182	2.496	0.169	2.751
TARnet	0.034	4.020	0.227	2.373	0.224	2.406	0.056	2.833
Dragonnet	**0.011**	**3.752**	0.265	2.440	0.237	2.365	0.045	2.800
NN-Dragonnet (C)	0.112	4.096	0.182	2.404	0.229	2.396	0.141	3.050
NN-Dragonnet (E)	0.197	4.220	0.176	2.386	0.275	2.342	0.151	3.085
NN-Dragonnet (M)	0.216	4.306	0.154	2.409	0.221	**2.324**	**0.001**	2.903

(a) Non-Imputed Dataset

Model	**Case 1** $\lvert\epsilon_{\text{ATE}}\rvert = 0$		**Case 2** $\lvert\epsilon_{\text{ATE}}\rvert = 0.29$		**Case** 3 $\lvert\epsilon_{\text{ATE}}\rvert = 0.29$		**Case 4** $\lvert\epsilon_{\text{ATE}}\rvert = 0.29$	
	$\lvert\epsilon_{\text{ATE}}\rvert$	**PEHE**	$\lvert\epsilon_{\text{ATE}}\rvert$	**PEHE**	$\lvert\epsilon_{\text{ATE}}\rvert$	**PEHE**	$\lvert\epsilon_{\text{ATE}}\rvert$	**PEHE**
LR1	0.154	3.668	**0.143**	2.762	**0.157**	2.755	0.191	**2.742**
LR2	0.053	2.928	0.205	**2.334**	0.223	**2.326**	0.168	2.770
R-Forest	0.120	2.970	0.277	2.462	0.277	2.460	0.114	3.147
BART	0.462	3.345	0.164	2.451	0.186	2.601	0.150	2.756
TARnet	0.034	3.124	0.237	2.358	0.232	2.393	0.061	2.772
Dragonnet	0.009	2.807	0.271	2.429	0.285	2.434	**0.051**	2.764
NN-Dragonnet (C)	0.047	2.961	0.184	2.387	0.239	2.402	0.141	3.023
NN-Dragonnet (E)	0.031	**2.707**	0.181	2.374	0.233	2.357	0.141	3.049
NN-Dragonnet (M)	**0.004**	2.762	0.170	2.390	0.229	2.366	0.128	3.025

(b) Imputed Dataset

6. Discussion and Conclusions

In this work, we propose a data analysis workflow, which is developed in the context of the REBECCA project. The main objective is to present a methodology on how to handle RWD and provide accurate and reliable estimations of causal effects for research on the quality of life of BCPs. The proposed methodology consists of the stages of data collection, data curation (including data imputation and outlier detection), model development, treatment effect estimation and finally, the presentation of results. Issues related to data quality as well as sampling, confounding and other biases have an impact on the accuracy of treatment effect estimation.

To demonstrate and evaluate the proposed approach, we developed a causal DAG for the case of osteopenia and osteoporosis, which is a comorbidity of AI treatment in postmenopausal BCPs. Specifically, we used the DAG and an associated SCM to generate synthetic datasets and then we applied the proposed workflow to several use case scenarios, including cases with missing values as well as sampling and confounding biases. Based on the results of the experimental analysis, the proposed approach achieves significantly higher estimation accuracy, compared to direct computation of the ATE, in cases where bias is present. This implies that the use of ML models for causal inference is beneficial, since they provide lower estimation errors. The experimental analysis results suggest that there is promise in using automated methods for data quality control and causal effect estimation for studying factors affecting the quality of life of BCPs.

The synthetic data used in this study were carefully designed based on established relationships from the literature on clinical trials, as well as input from domain experts (see Appendixes A and B). This approach helped in ensuring that the generated data faithfully represent realistic clinical scenarios and the underlying mechanisms involved. Nevertheless, even though the generated data may not capture some of the complexity and the nuances of real patient data, they highlight the challenges that arise when attempting to estimate treatment effects on the quality of life of BCPs.

Generally, synthetic data offer substantial benefits in addressing the lack of real-world datasets, especially in cases such as the inherent absence of the counterfactual outcomes [69]. However, the synthetic data generation process carries the risk of introducing biases, either from the assumptions made in defining relationships or from the algorithm used to generate the data. To the best of our knowledge, there are no available causal inference datasets for studying the comorbidity of osteopenia and osteoporosis; hence, an evaluation for ensuring that key statistical properties of the RWD are preserved by the generated synthetic is impossible. This can be considered as a limitation of this approach, since the generated synthetic data may not fully capture the complexity of RWD. An elegant approach for addressing this difficulty will certainly be included in our future research.

In this research, we focused our attention on the estimation of the ATE and CATE, which constitute quantitative measures of how treatments perform on average and for specific subgroups. It is worth highlighting the importance of translating the findings made from the estimation of these metrics into hypotheses for future research as well as clinical practice for supporting clinicians in making more informed, patient-specific treatment decisions. However, this is not possible due to the lack of available real-world data.

In our future work, we intend to seek and obtain real BC patient data from clinical trials for assessing the quality of the generated synthetic data. In addition, we will focus our attention on developing methods that combine the data-driven causal effect estimation models with Structural Causal Models, as well as methods for model validation and testing. Finally, another interesting direction is the enhancement of SCMs for multiple comorbidities of breast cancer, which need to be developed to support holistic, multi-level interventions to improve the quality of life of patients.

Author Contributions: Conceptualization, N.K., A.B. and C.D.; methodology, N.K., A.B. and C.D.; software, N.K. and A.B.; validation, N.K., A.B. and C.D.; formal analysis, N.K., A.B., C.M.H., A.M., T.S., D.A. and C.D.; investigation, N.K., A.B. and C.D.; resources, N.K., A.B. and C.D.; data curation,

N.K. and A.B.; writing—original draft, N.K., A.B., C.M.H., A.M., T.S., D.A. and C.D.; writing—review and editing, N.K., A.B., C.M.H., A.M., T.S., D.A. and C.D.; visualization, N.K.; supervision N.K. and C.D.; project administration, C.D.; funding acquisition, C.D. All authors have read and agreed to the published version of the manuscript.

Funding: The work leading to these results received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement No. 965231, project REBECCA (REsearch on BrEast Cancer induced chronic conditions supported by Causal Analysis of multi-source data).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original data presented in the study are openly available in GitHub at https://github.com/kiriakidou/A-Causal-Inference-Methodology-to-Support-Research-on-Osteopenia-for-Breast-Cancer-Patients (accessed on 19 October 2024).

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. DAG Variables

Next, we present all the links between the variables of the DAG along with an explanation for the connection of the corresponding nodes.

- Age → Menopause: This is an obvious relationship, since age in the majority of the cases indicates whether the woman is pre-menopausal or post-menopausal.
- Age → Estrogens: According to Lephart [70] estrogen levels peak in the mid-to-late 20s in women and then decline by 50% by 50 years of age and dramatically decrease further after menopause.
- Age → Bone Mineral Density (BMD) loss: BMD loss increases as a person gets older [71].
- Menopause → Aromatase Inhibitor (AI) treatment: Only post-menopausal women undergo AI treatment [72].
- Menopause → Estrogens: Estrogen levels in post-menopausal women drop and they no longer ovulate [70].
- AI treatment → Estrogens: AI treatment specifically targets estrogens and, thus, accelerates their deprivation [70].
- Family History (patient's ethnicity) → Estrogens: According to Visvanathan and Yager [73], there are variations in the estrogen levels of breast cancer patients among different ethnicities.
- Body Mass Index (BMI) → Estrogens: Clinicians advise was that BMI is a good indication of whether the patients have better eating and exercising behavior.
- Heritage → Estrogens: Clinicians informed us that genetics, mainly if the mother of the patient had osteoporosis or had a limb fracture, has an effect on estrogen levels, which afterwards lead to a reduction in BMD.
- Estrogens → Calcitonin: Calcitonin is proposed as mediator of estrogen action, as is mentioned in [74].
- Estrogens → Mood: Based on Thompson and Reilly [75], a lack of estrogens worsens the patient's mood.
- Estrogens → Abdominal Fat Accumulation: Low levels of estrogens can contribute to women gaining fat in the belly area [76].
- Estrogens → Energy/Fatigue: Clinicians informed us that a lack of estrogens lowers the patient's levels of energy and, thus, causes fatigue.
- Estrogens → Sleep quality: Clinicians highlighted that information about sleep is crucial for studying the case of osteopenia and osteoporosis. As is indicated in Gava et al. [77], low estrogen levels cause sleep disturbances.

- Estrogens → Ability to concentrate: The research of Hara et al. [78] suggests that estrogen levels have an impact on memory and cognition. Hence, low estrogen levels lead to difficulty in concentrating.
- Estrogens → Hot flashes: Clinicians underlined that a lack of estrogens have a negative impact on patients' hot flashes .
- Estrogens → Libido: Clinicians highlighted that a lack of estrogens decreases patients' libido.
- Estrogens → Depression: The research of Studd [79] indicates that low estrogen levels are associated with depression.
- Exercise → Mood: Research of Hoffman and Hoffman [80] reveals that exercise leads to an improvement in mood.
- Exercise → Abdominal Fat Accumulation: Clinicians discussed that if a person does more intense exercise, then there is probability for less abdominal fat accumulation.
- Exercise → Energy/Fatigue: Intense exercise lead to fatigue, as is stated in [81].
- Exercise → Sleep quality: The research of Hargens et al. [82] indicates that exercise decreases patients' sleep complaints as well as insomnia.
- Exercise → Ability to concentrate: The research of Falkai et al. [83] suggests that mild exercise can improve the ability to concentrate.
- Exercise → Hot flashes: The research of Romani et al. [84] indicates that higher levels of physical activity are significantly associated with increasing odds of moderate or severe hot flashes.
- Exercise → Libido: Moderate exercise is linked to increases in libido, while excessive exercise is linked to lower libido, as is supported in [85].
- Exercise → Depression: According to Mead et al. [86], exercise seems to improve the symptoms of depression.
- Exercise → Bone Turnover: Results of the study of Gombos et al. [87] are consistent with previous reports in the literature indicating that the force generated by muscle contraction contribute to stimulating bone resorption.
- Calcitonin → Calcium in blood: The main function of calcitonin is the decreasing of calcium levels in the blood [88].
- Calcium supplement → Calcium in blood: When the patient receives calcium supplementation, it leads to higher calcium levels in the blood.
- Vit. D in Blood → Calcium in blood: Low vitamin D levels inhibit the absorption of calcium in the blood and, hence, lead to a low level of calcium in the blood [89].
- Nutrition rich in calcium → Calcium in blood: Receiving more calcium through nutrition leads to higher calcium levels in the blood.
- Alcohol consumption → Calcium in blood: One of the clinical symptoms of chronic alcohol consumption the decrease in calcium in the blood, as is supported in the research in [90].
- Diabetes → Vit. D in Blood: Vitamin D deficiency is associated with a decreased insulin release, based on Mitri et al. [91]'s research.
- Vit. D supplement → Vit. D in Blood: When the patient receives vitamin D supplementation, this leads to higher vitamin D levels in the blood.
- Calcium in blood → Bone Turnover: Calcium is essential for bone formation, as is reported in [92].
- Cigarettes per day → Bone Turnover: According to Trevisan et al. [93], smoking negatively affects bone health and, hence, reduces the formation of bones.
- Parathormone(PTH) → Bone Turnover: PTH stimulates the release of calcium in an indirect process through osteoclasts, which ultimately leads to the resorption of the bones, as is supported in [94].
- Bone Turnover → C-Telopeptides: Increased levels of C-Telopeptides indicate increased bone resorption, based on Ju et al. [95].

Appendix B. Functional Relationships of Synthetic Data Generator

Each variable in the DAG is associated with one of the following four functions:

- *identity* → The value of this variable is set beforehand;
- *randint* → A random integer is selected from a range of predefined values;
- *normal* → A random value drawn from a normal gaussian distribution;
- *variable name* → Custom functions defined specifically for the generation of these variables;
- *parametric* → The value of the variable depends on the generated values of all incoming nodes. Specifically, given that some variables $(x_1, x_2, \ldots, x_n)$ directly affect a variable y in the causal DAG, the output value of y is of course dependent on the outputs of its parents, along with some inherent to the parent variables and exogenous noise factors u. Therefore, we generate the output of y as $y = \sum_i^n a_i x_i u_i + U$.

The parameters of each function were based on values found in the literature, taking into consideration the context of the use cases described in the manuscript (i.e., menopausal women under aromatase inhibitor treatment). All relationships are described in the Table A1, where $N(\mu, \sigma^2)$ refers to noise drawn from a normal Gaussian distribution. In addition, the custom functions for each variable mentioned in Table A1 are defined below.

For the *BMD Loss* function, we differentiate between the following cases:

Case 1: Low Exercise → the function is as follows:

$$\begin{aligned}\text{bmd loss} \quad = \quad & -0.1 \cdot \frac{\text{calcium}}{10} - \frac{\text{formation}}{3} \cdot \mathcal{N}(0.5, 1) + \frac{\text{resorption}}{3} + \frac{\text{age} - 50}{50} \cdot \mathcal{N}(0.5, 1) \\ & + \mathcal{N}(0.5, 0.5) + \frac{\max(\text{pa}, 1)}{4} \cdot \mathcal{N}(1, 0.5) + 0.5\end{aligned}$$

Case 2: Intense Exercise → the function is as follows:

$$\begin{aligned}\text{bmd loss} \quad = \quad & -0.1 \cdot \frac{\text{calcium}}{10} - \frac{\text{formation}}{3} \cdot \mathcal{N}(0.5, 1) + \frac{\text{resorption}}{3} + \frac{\text{age} - 50}{50} \cdot \mathcal{N}(0.5, 1) \\ & + \mathcal{N}(0.5, 0.5) - \frac{\text{pa}}{4} \cdot \mathcal{N}(0.5, 0.5)\end{aligned}$$

For *Back Pain*:

$$\text{back pain} = \begin{cases} \text{randint}(0, 1), & \text{if bmd loss} = 0 \\ \text{randint}(2, 3), & \text{if bmd loss} = 1 \\ \text{randint}(3, 4), & \text{if bmd loss} = 2 \end{cases}$$

For the *Calcium in Blood* function, we differentiate between the following cases:

Case 1: No Calcium Supplementation

$$\text{calc in blood} = \max\left((-0.9)^{\text{alcohol}} \cdot (1.1)^{\text{nutri}} \cdot \frac{\text{vit\_d}}{20} + \mathcal{N}(1, 1) \cdot \frac{\text{bone\_res}}{2} + \mathcal{N}(1, 1), \mathcal{N}(0.5, 1)\right)$$

Case 2: Calcium Supplementation

$$\begin{aligned}\text{calc in blood} \quad = \quad & \max\left(\left(6.5 + \left(\frac{10.2 - 8.5}{50 - 20}\right) \cdot (-0.9)^{\text{alcohol}} \cdot (1.1)^{\text{nutri}} \cdot \frac{\text{vit\_d}}{20}\right.\right. \\ & \left.\left. + \mathcal{N}(1, 1) \cdot \frac{\text{bone\_res}}{2} + \mathcal{N}(1, 1)\right), 7.5 + \mathcal{N}(1, 1)\right)\end{aligned}$$

For the generation of the *Estrogens* level, we calculate the following. Let ($g(f, b, r,$ aroma, factor) be defined as follows:

$$g(f, b, r, \text{aroma}, \text{factor}) = -\text{factor} \cdot (1.1)^f \cdot (1.1)^r \cdot (0.9)^b \cdot (1.9)^{\text{aroma}}$$

Case 1: Pre-menopausal patient

$$\begin{aligned} u &= \max(\text{abs}(\mathcal{N}(50,5) - \text{age}), 1) \\ f_1 &= 400 + g(\text{family}, \text{bmi}, \text{heritage}, \text{ai}, 9.3) \cdot u \end{aligned}$$

Case 2: Menopausal patient

$$\begin{aligned} u &= \max(\text{abs}(\text{age} - \mathcal{N}(60,10)), 1) \\ f_1 &= 30 + g(\text{family}, \text{bmi}, \text{heritage}, \text{ai}, 2.31) \cdot u \end{aligned}$$

Table A1. Functional relationships in synthetic data generator.

Node	Parent Node x_i	Function	Parameters a_i, u_i and U or Value
Abdominal Fat Accumulation	Estrogens	parametric	$a_1 = -1/370$, $u_1 = 1$, $U \sim N(1,2)$
	Exercise	parametric	$a_2 = -1$, $u_2 = 1$, $U \sim N(1,2)$
Ability to concentrate	Estrogens	parametric	$a_1 = 1/370$, $u_1 = 1$, $U \sim N(0,2)$
	Exercise	ordinal	$a_2 = 1$, $u_2 = 1$, $U \sim N(0,2)$
Age		normal	$N(60,10)$
AI treatment	Menopause	identity	1
Alcohol Consumption		randint	$(0,4)$
Back pain	BMD loss	Back pain	Function described below
BMD Loss	Calcium in blood, Bone formation, Bone resorption, Age, Exercise, Calcium supplement	BMD loss	Function described below
BMI		randint	$(0,2)$
Bone Formation	Calcium in blood	parametric	$a_1 = 2$, $u_1 = 2$, $U \sim N(1,1)$
	Calcitonin	parametric	$a_2 = -1/4$, $u_2 = 1$, $U \sim N(1,1)$
	Cigarettes per day	parametric	$a_3 = -1$, $u_3 = U \sim N(0.5,1)$, $U \sim N(1,1)$
Bone resorption	Calcitonin	parametric	$a_1 = 1/2$, $u_1 = 1$, $U \sim N(0,1)$
	Exercise	parametric	$a_2 = 2$, $u_2 = 1$, $U \sim N(0,1)$
	PTH	parametric	$a_3 = 1$, $u_3 = U \sim N(0.5,1)$, $U \sim N(0,1)$
Calcitonin	Estrogens	parametric	$a_1 = 10/370$, $u_1 = 1$, $U \sim N(2,1)$
Calcium in Blood	Alcohol consumption, Calcium Supplement, Calcitonin, Nutrition rich in calcium, Vit D in Blood	calcium in blood	Function described below
Calcium Supplement		identity	0
Cigarettes per day		randint	$(0,3)$
Clinical Symptoms	Estrogens, Exercise	randint	$(0,1)$
C-Telopeptides	Bone resorption	normal	$N(500,100)$

Table A1. *Cont.*

Node	Parent Node x_i	Function	Parameters a_i, u_i and U or Value
Depression	Estrogens Exercise	parametric parametric	$a_1 = 1/370$, $u_1 = 1$, $U \sim N(0,2)$ $a_2 = -1$, $u_2 = 1$, $U \sim N(0,2)$
Diabetes		randint	$(0, 1)$
Energy Loss/Fatigue	Estrogens, Exercise	randint	$(0, 4)$
Estrogens	Family history, BMI, Heritage, AI treatment, Menopause, Age	estrogens	Function described below
Family history		randint	$(0, 1)$
Heritage		randint	$(0, 1)$
Libido	Estrogens Exercise	parametric parametric	$a_1 = 1/370$, $u_1 = 1$, $U \sim N(0,2)$ $a_2 = 1/370$, $u_2 = 1$, $U \sim N(0,2)$
Menopause	Age	identity	1
Mood	Estrogens, Exercise	normal	$N(0,1)$
Number of fractures	BMD Loss	fractures	Function described below
Nutrition rich in calcium		randint	$(0, 4)$
Exercise		randint	$(0, 3)$
PTH		normal	$N(35,10)$
Sleep quality	Estrogens Exercise	parametric ordinal parametric ordinal	$a_1 = 1$, $u_1 = 1$, $U \sim N(-1,1)$ $a_2 = 1/370$, $u_2 = 1$, $U \sim N(-1,1)$
Vit D in Blood	Vit D Supplement, Diabetes	Vit D in blood	Function described below
Vit D Supplement		randint	$(0, 1)$

The final estrogen level is given by

$$\text{estrogens} = \max((f_1), 0 + \text{abs}(\mathcal{N}(0.5, 1)))$$

For *Number of fractures*:

$$\text{number of fractures} = \begin{cases} \text{randint}(0,1), & \text{if bmd loss} = 0 \\ \text{randint}(2,3), & \text{if bmd loss} = 1 \\ \text{randint}(3,4), & \text{if bmd loss} = 2 \end{cases}$$

For the generation of the *Vit. D in Blood* values, we calculate the following:
Case 1: No Vit. D supplementation

$$\begin{aligned} a &= \frac{200 - 140}{15 - 5} \\ f &= \max((20 - a \cdot \text{diabetes}) - \mathcal{N}(2,1), 0) \end{aligned}$$

Case 2: Vit. D Supplementation

$$\begin{aligned} a &= \frac{140 - 10}{50 - 20} \\ f &= \max(30 - a \cdot \text{diabetes} + \mathcal{N}(0.5,1), 0) \end{aligned}$$

The final levels of vitamin D are given by

$$\text{vit d in blood} = \max(f, \text{abs}(\mathcal{N}(0.5, 1)))$$

Appendix C. Hyperparameter Values Used in the Experiments

Table A2. Hyperparameter values of each method in the experiments.

Model	Hyperparameters
R-Forest	n_estimators = 100, criterion = 'gini', max_depth = None, min_samples_split = 2 and min_samples_leaf = 1.
BART	n_trees = 100, n_chains = 4, $\sigma_a = 10^{-3}$, $\sigma_b = 10^{-3}$, $\alpha = 0.95$ and $\beta = 2$.
TARnet	Topology: Three dense layers of 200 neurons with ELU activation function for producing the representation layer $Z(X)$. The output of $Z(X)$ is further processed by two dense layers of 100 neurons each with ELU activation function and kernel regularizer of 10^{-2} for the prediction of the outcome of the control group and a similar branch for the treatment group.
Dragonnet	Topology: Three dense layers of 200 neurons with ELU activation function for producing the representation layer $Z(X)$. The output of $Z(X)$ is further processed by two dense layers of 100 neurons each with ELU activation function and kernel regularizer of 10^{-2} for the prediction of the outcome of the control group and a similar branch for the treatment group. In addition, the shared representation $Z(X)$ is used for predicting the propensity score, through the use of a simple linear map followed by a sigmoid activation function.
NN-Dragonnet	Topology: Three dense layers of 200 neurons with ELU activation function for producing the representation layer $Z(X)$. Next, the output of $Z(X)$ is concatenated with the average of the neighboring instances of control(treatment) group and the combined information is further processed by two dense layers of 100 neurons each with ELU activation function and kernel regularizer of 10^{-2} for the prediction of the outcome of the control(treatment) group. In addition, the shared representation $Z(X)$ is used for predicting the propensity score, through the use of a simple linear map followed by a sigmoid activation function.

Notice that all neural network-based models were trained for 100 epochs using ADAM (Adaptive Moment Estimation) as the optimizer with a learning rate of 10^{-3} and then for another 300 epochs using SGD (Stochastic Gradient Descent) as the optimizer with a learning rate of 10^{-5} and momentum $m = 0.9$. In addition, the utilized loss function was t-reg with $\alpha = 1$ and $\beta = 1$, while 20% of the training data were used as the validation set.

References

1. Waks, A.G.; Winer, E.P. Breast Cancer Treatment: A Review. *JAMA* **2019**, *321*, 288–300. [CrossRef] [PubMed]
2. Mokhatri-Hesari, P.; Montazeri, A. Health-related quality of life in breast cancer patients: Review of reviews from 2008 to 2018. *Health Qual. Life Outcomes* **2020**, *18*, 338. [CrossRef] [PubMed]
3. Akram, M.; Iqbal, M.; Daniyal, M.; Khan, A.U. Awareness and current knowledge of breast cancer. *Biol. Res.* **2017**, *50*, 33. [CrossRef] [PubMed]
4. Macedo, L.F.; Sabnis, G.; Brodie, A. Aromatase inhibitors and breast cancer. *Ann. N. Y. Acad. Sci.* **2009**, *1155*, 162–173. [CrossRef] [PubMed]
5. Diana, A.; Carlino, F.; Giunta, E.F.; Franzese, E.; Guerrera, L.P.; Di Lauro, V.; Ciardiello, F.; Daniele, B.; Orditura, M. Cancer treatment–induced bone loss (CTIBL): State of the art and proper Management in Breast Cancer Patients on endocrine therapy. *Curr. Treat. Options Oncol.* **2021**, *22*, 45. [CrossRef]
6. Nikander, R.; Sievänen, H.; Heinonen, A.; Daly, R.M.; Uusi-Rasi, K.; Kannus, P. Targeted exercise against osteoporosis: A systematic review and meta-analysis for optimising bone strength throughout life. *BMC Med.* **2010**, *8*, 47. [CrossRef]
7. Chen, J.; Ho, M.; Lee, K.; Song, Y.; Fang, Y.; Goldstein, B.A.; He, W.; Irony, T.; Jiang, Q.; van der Laan, M.; et al. The current landscape in biostatistics of real-world data and evidence: Clinical study design and analysis. *Stat. Biopharm. Res.* **2021**, *15*, 29–42. [CrossRef]
8. Bellows, B.K.; Kuo, K.L.; Biltaji, E.; Singhal, M.; Jiao, T.; Cheng, Y.; McAdam-Marx, C. Real-world evidence in pain research: A review of data sources. *J. Pain Palliat. Care Pharmacother.* **2014**, *28*, 294–304. [CrossRef]
9. Dhruva, S.S.; Ross, J.S.; Akar, J.G.; Caldwell, B.; Childers, K.; Chow, W.; Ciaccio, L.; Coplan, P.; Dong, J.; Dykhoff, H.J.; et al. Aggregating multiple real-world data sources using a patient-centered health-data-sharing platform. *NPJ Digit. Med.* **2020**, *3*, 60. [CrossRef]
10. Pintelas, E.; Livieris, I.E.; Barotsis, N.; Panayiotakis, G.; Pintelas, P. An autoencoder convolutional neural network framework for sarcopenia detection based on multi-frame ultrasound image slices. In *Proceedings 17, Proceedings of the Artificial Intelligence Applications and Innovations: 17th IFIP WG 12.5 International Conference, AIAI 2021, Hersonissos, Crete, Greece, 25–27 June 2021*; Springer: Berlin/Heidelberg, Germany, 2021; pp. 209–219.

11. Concato, J. Observational versus experimental studies: What's the evidence for a hierarchy? *NeuroRx* **2004**, *1*, 341–347. [CrossRef]
12. Schneeweiss, S.; Rassen, J.A.; Glynn, R.J.; Avorn, J.; Mogun, H.; Brookhart, M.A. High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. *Epidemiology* **2009**, *20*, 512. [CrossRef] [PubMed]
13. Bica, I.; Alaa, A.M.; Lambert, C.; Van Der Schaar, M. From real-world patient data to individualized treatment effects using machine learning: Current and future methods to address underlying challenges. *Clin. Pharmacol. Ther.* **2021**, *109*, 87–100. [CrossRef] [PubMed]
14. Liu, F.; Panagiotakos, D. Real-world data: A brief review of the methods, applications, challenges and opportunities. *BMC Med. Res. Methodol.* **2022**, *22*, 287. [CrossRef] [PubMed]
15. Liu, B.; Zhou, X.; Wang, Y.; Hu, J.; He, L.; Zhang, R.; Chen, S.; Guo, Y. Data processing and analysis in real-world traditional Chinese medicine clinical data: Challenges and approaches. *Stat. Med.* **2012**, *31*, 653–660. [CrossRef] [PubMed]
16. Rubin, D.B. Causal inference using potential outcomes: Design, modeling, decisions. *J. Am. Stat. Assoc.* **2005**, *100*, 322–331. [CrossRef]
17. Listl, S.; Jürges, H.; Watt, R.G. Causal inference from observational data. *Community Dent. Oral Epidemiol.* **2016**, *44*, 409–415. [CrossRef]
18. Hariton, E.; Locascio, J.J. Randomised controlled trials—The gold standard for effectiveness research. *BJOG Int. J. Obstet. Gynaecol.* **2018**, *125*, 1716. [CrossRef]
19. Rothwell, P.M. Factors that can affect the external validity of randomised controlled trials. *PLoS Clin. Trials* **2006**, *1*, e9. [CrossRef]
20. Groenwold, R.H.; Van Deursen, A.M.; Hoes, A.W.; Hak, E. Poor quality of reporting confounding bias in observational intervention studies: A systematic review. *Ann. Epidemiol.* **2008**, *18*, 746–751. [CrossRef]
21. Mehrabi, N.; Morstatter, F.; Saxena, N.; Lerman, K.; Galstyan, A. A survey on bias and fairness in machine learning. *ACM Comput. Surv. (CSUR)* **2021**, *54*, 1–35. [CrossRef]
22. Ballas, A.; Diou, C. Towards Domain Generalization for ECG and EEG Classification: Algorithms and Benchmarks. *IEEE Trans. Emerg. Top. Comput. Intell.* **2023**, *8*, 44–54. [CrossRef]
23. Pearl, J. *Causality*; Cambridge University Press: Cambridge, UK, 2009.
24. Glymour, M.; Pearl, J.; Jewell, N.P. *Causal Inference in Statistics: A Primer*; John Wiley & Sons: Hoboken, NJ, USA, 2016.
25. Arif, S.; MacNeil, M.A. Applying the structural causal model framework for observational causal inference in ecology. *Ecol. Monogr.* **2023**, *93*, e1554. [CrossRef]
26. Reinhold, J.C.; Carass, A.; Prince, J.L. A structural causal model for MR images of multiple sclerosis. In *Proceedings, Part V 24, Proceedings of the Medical Image Computing and Computer Assisted Intervention—MICCAI 2021: 24th International Conference, Strasbourg, France, 27 September–1 October 2021*; Springer: Berlin/Heidelberg, Germany, 2021; pp. 782–792.
27. Petersen, M.L.; van der Laan, M.J. Causal models and learning from data: Integrating causal modeling and statistical estimation. *Epidemiology* **2014**, *25*, 418. [CrossRef] [PubMed]
28. Chipman, H.A.; George, E.I.; McCulloch, R.E. BART: Bayesian additive regression trees. *Ann. Appl. Stat.* **2010**, *4*, 266–298. [CrossRef]
29. Hastie, T.; Tibshirani, R. Bayesian backfitting (with comments and a rejoinder by the authors. *Stat. Sci.* **2000**, *15*, 196–223. [CrossRef]
30. Künzel, S.R.; Sekhon, J.S.; Bickel, P.J.; Yu, B. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proc. Natl. Acad. Sci. USA* **2019**, *116*, 4156–4165. [CrossRef]
31. Alaa, A.; Schaar, M. Limits of estimating heterogeneous treatment effects: Guidelines for practical algorithm design. In Proceedings of the International Conference on Machine Learning, PMLR, Stockholm, Sweden, 10–15 July 2018; pp. 129–138.
32. Shalit, U.; Johansson, F.D.; Sontag, D. Estimating individual treatment effect: Generalization bounds and algorithms. In Proceedings of the 34th International Conference on Machine Learning, Sydney, Australia, 6–11 August 2017; Volume 70, pp. 3076–3085.
33. Villani, C. *Optimal Transport: Old and New*; Springer: Berlin/Heidelberg, Germany, 2009; Volume 338.
34. Gretton, A.; Borgwardt, K.M.; Rasch, M.J.; Schölkopf, B.; Smola, A. A kernel two-sample test. *J. Mach. Learn. Res.* **2012**, *13*, 723–773.
35. Shi, C.; Blei, D.; Veitch, V. Adapting neural networks for the estimation of treatment effects. In Proceedings of the Advances in Neural Information Processing Systems 32, Vancouver, BC, Canada, 8–14 December 2019; Volume 32.
36. Kiriakidou, N.; Diou, C. An improved neural network model for treatment effect estimation. In *Proceedings, Part I, Proceedings of the Artificial Intelligence Applications and Innovations: 18th IFIP WG 12.5 International Conference, AIAI 2022, Hersonissos, Crete, Greece, 17–20 June 2022*; Springer: Berlin/Heidelberg, Germany, 2022; pp. 147–158.
37. Kiriakidou, N.; Diou, C. Integrating nearest neighbors on neural network models for treatment effect estimation. *Int. J. Neural Syst.* **2023**, *33*. [CrossRef]
38. Diou, C.; Kyritsis, K.; Papapanagiotou, V.; Sarafis, I. Intake monitoring in free-living conditions: Overview and lessons we have learned. *Appetite* **2022**, *176*, 106096. [CrossRef]

39. Kyritsis, K.; Diou, C.; Delopoulos, A. A data driven end-to-end approach for in-the-wild monitoring of eating behavior using smartwatches. *IEEE J. Biomed. Health Inform.* **2020**, *25*, 22–34. [CrossRef]
40. Sarafis, I.; Diou, C.; Delopoulos, A. Behaviour profiles for evidence-based policies against obesity. In Proceedings of the 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Berlin, Germany, 23–27 July 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 3596–3599.
41. Loesgen, K. A generalization and Bayesian interpretation of ridge-type estimators with good prior means. *Stat. Pap.* **1990**, *31*, 147–154. [CrossRef]
42. Fix, E.; Hodges, J.L. Discriminatory analysis. Nonparametric discrimination: Consistency properties. *Int. Stat. Rev. Int. Stat.* **1989**, *57*, 238–247. [CrossRef]
43. Breiman, L.; Cutler, A. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
44. Geurts, P.; Ernst, D.; Wehenkel, L. Extremely randomized trees. *Mach. Learn.* **2006**, *63*, 3–42. [CrossRef]
45. Liu, F.T.; Ting, K.M.; Zhou, Z.H. Isolation forest. In Proceedings of the 2008 Eighth IEEE International Conference on Data Mining, Pisa, Italy, 15–19 December 2008; IEEE: Piscataway, NJ, USA, 2008; pp. 413–422.
46. Breunig, M.M.; Kriegel, H.P.; Ng, R.T.; Sander, J. LOF: Identifying density-based local outliers. In Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, Dallas, TX, USA, 15–18 May 2000; pp. 93–104.
47. Schölkopf, B.; Platt, J.C.; Shawe-Taylor, J.; Smola, A.J.; Williamson, R.C. Estimating the support of a high-dimensional distribution. *Neural Comput.* **2001**, *13*, 1443–1471. [CrossRef]
48. Carey, J.J.; Delaney, M.F. T-scores and Z-scores. *Clin. Rev. Bone Miner. Metab.* **2010**, *8*, 113–121. [CrossRef]
49. Ferstad, J.O.; Vallon, J.J.; Jun, D.; Gu, A.; Vitko, A.; Morales, D.P.; Leverenz, J.; Lee, M.Y.; Leverenz, B.; Vasilakis, C.; et al. Population-level management of type 1 diabetes via continuous glucose monitoring and algorithm-enabled patient prioritization: Precision health meets population health. *Pediatr. Diabetes* **2021**, *22*, 982–991. [CrossRef]
50. Prahalad, P.; Ding, V.Y.; Zaharieva, D.P.; Addala, A.; Johari, R.; Scheinker, D.; Desai, M.; Hood, K.; Maahs, D.M. Teamwork, targets, technology, and tight control in newly diagnosed type 1 diabetes: The pilot 4T study. *J. Clin. Endocrinol. Metab.* **2022**, *107*, 998–1008. [CrossRef]
51. Zaharieva, D.P.; Bishop, F.K.; Maahs, D.M. Advancements and future directions in the teamwork, targets, technology, and tight control—The 4T study: Improving clinical outcomes in newly diagnosed pediatric type 1 diabetes. *Curr. Opin. Pediatr.* **2022**, *34*, 423–429. [CrossRef]
52. Smith, I.E.; Dowsett, M. Aromatase inhibitors in breast cancer. *N. Engl. J. Med.* **2003**, *348*, 2431–2442. [CrossRef]
53. Brueggemeier, R.W.; Hackett, J.C.; Diaz-Cruz, E.S. Aromatase inhibitors in the treatment of breast cancer. *Endocr. Rev.* **2005**, *26*, 331–345. [CrossRef] [PubMed]
54. Martinez, V.; Navarro, C.; Cano, C.; Fajardo, W.; Blanco, A. DrugNet: Network-based drug–disease prioritization by integrating heterogeneous data. *Artif. Intell. Med.* **2015**, *63*, 41–49. [CrossRef] [PubMed]
55. Zheng, K.; Harris, C.E.; Jennane, R.; Makrogiannis, S. Integrative blockwise sparse analysis for tissue characterization and classification. *Artif. Intell. Med.* **2020**, *107*, 101885. [CrossRef]
56. Amkrane, Y.; El Adoui, M.; Benjelloun, M. Towards breast cancer response prediction using artificial intelligence and radiomics. In Proceedings of the 2020 5th International Conference on Cloud Computing and Artificial Intelligence: Technologies and Applications (CloudTech), Marrakesh, Morocco, 24–26 November 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 1–5.
57. Shaikh, K.; Krishnan, S.; Thanki, R.M. *Artificial Intelligence in Breast Cancer Early Detection and Diagnosis*; Springer: Berlin/Heidelberg, Germany, 2021.
58. Louizos, C.; Shalit, U.; Mooij, J.M.; Sontag, D.; Zemel, R.; Welling, M. Causal effect inference with deep latent-variable models. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; Volume 30.
59. Kiriakidou, N.; Livieris, I.E.; Pintelas, P. Mutual information-based neighbor selection method for causal effect estimation. *Neural Comput. Appl.* **2024**, *36*, 9141–9155. [CrossRef]
60. Montgomery, D.C.; Peck, E.A.; Vining, G.G. *Introduction to Linear Regression Analysis*; John Wiley & Sons: Hoboken, NJ, USA, 2021.
61. Manaye, S.; Cheran, K.; Murthy, C.; Bornemann, E.A.; Kamma, H.K.; Alabbas, M.; Elashahab, M.; Abid, N.; Franchini, A.P.A.; Elashahab, M. The Role of High-intensity and High-impact Exercises in Improving Bone Health in Postmenopausal Women: A Systematic Review. *Cureus* **2023**, *15*, e34644. [CrossRef]
62. Zehnacker, C.H.; Bemis-Dougherty, A. Effect of weighted exercises on bone mineral density in post menopausal women a systematic review. *J. Geriatr. Phys. Ther.* **2007**, *30*, 79–88. [CrossRef]
63. Cadore, E.L.; Brentano, M.A.; Kruel, L.F.M. Effects of the physical activity on the bone mineral density and bone remodelation. *Rev. Bras. Med. Esporte* **2005**, *11*, 373–379. [CrossRef]
64. Aleksić, D. A novel test of missing completely at random: U-statistics-based approach. *Statistics* **2024**, *58*, 1004–1023. [CrossRef]
65. Barata, A.P.; Takes, F.W.; van den Herik, H.J.; Veenman, C.J. Imputation methods outperform missing-indicator for data missing completely at random. In Proceedings of the 2019 International Conference on Data Mining Workshops (ICDMW), Beijing, China, 8–11 November 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 407–414.
66. Livieris, I.E. A novel forecasting strategy for improving the performance of deep learning models. *Expert Syst. Appl.* **2023**, *230*, 120632. [CrossRef]

67. Kistler-Fischbacher, M.; Yong, J.S.; Weeks, B.K.; Beck, B.R. High-intensity exercise and geometric indices of hip bone strength in postmenopausal women on or off bone medication: The MEDEX-OP randomised controlled trial. *Calcif. Tissue Int.* **2022**, *111*, 256–266. [CrossRef]
68. Cosman, F.; de Beur, S.J.; LeBoff, M.; Lewiecki, E.; Tanner, B.; Randall, S.; Lindsay, R. Clinician's guide to prevention and treatment of osteoporosis. *Osteoporos. Int.* **2014**, *25*, 2359–2381. [CrossRef] [PubMed]
69. Kiriakidou, N.; Livieris, I.E.; Diou, C. C-XGBoost: A tree boosting model for causal effect estimation. In Proceedings of the IFIP International Conference on Artificial Intelligence Applications and Innovations, Corfu, Greece, 27–30 June 2024; Springer: Cham, Switzerland, 2024; pp. 58–70.
70. Lephart, E.D. A review of the role of estrogen in dermal aging and facial attractiveness in women. *J. Cosmet. Dermatol.* **2018**, *17*, 282–288. [CrossRef]
71. Khosla, S. Pathogenesis of age-related bone loss in humans. *J. Gerontol. Ser. Biomed. Sci. Med. Sci.* **2013**, *68*, 1226–1235. [CrossRef] [PubMed]
72. Gibson, L.; Lawrence, D.; Dawson, C.; Bliss, J. Aromatase inhibitors for treatment of advanced breast cancer in postmenopausal women. *Cochrane Database Syst. Rev.* **2009**, *2009*, CD003370.
73. Visvanathan, K.; Yager, J.D. Ethnic Variations in Estrogen and Its Metabolites: Sufficient to Explain Differences in Breast Cancer Incidence Rates? *J. Natl. Cancer Inst.* **2016**, *108*, djw147. [CrossRef] [PubMed]
74. Agnusdei, D.; Civitelli, R.; Camporeale, A.; Gennari, C. Calcitonin and estrogens. *J. Endocrinol. Investig.* **1990**, *13*, 625–630. [CrossRef] [PubMed]
75. Thompson, E.A.; Reilly, D. The homeopathic approach to the treatment of symptoms of oestrogen withdrawal in breast cancer patients. A prospective observational study. *Homeopathy* **2003**, *92*, 131–134. [CrossRef] [PubMed]
76. Babaei, P.; Dastras, A.; Tehrani, B.S.; Roudbaneh, S.P. The effect of estrogen replacement therapy on visceral fat, serum glucose, lipid profiles and apelin level in ovariectomized rats. *J. Menopausal Med.* **2017**, *23*, 182–189. [CrossRef]
77. Gava, G.; Orsili, I.; Alvisi, S.; Mancini, I.; Seracchioli, R.; Meriggiola, M.C. Cognition, mood and sleep in menopausal transition: The role of menopause hormone therapy. *Medicina* **2019**, *55*, 668. [CrossRef]
78. Hara, Y.; Waters, E.M.; McEwen, B.S.; Morrison, J.H. Estrogen effects on cognitive and synaptic health over the lifecourse. *Physiol. Rev.* **2015**, *95*, 785–807. [CrossRef]
79. Studd, J. Estrogens and depression in women. In *Women's Health in Menopause: Behaviour, Cancer, Cardiovascular Disease, Hormone Replacement Therapy*; Springer Science & Business Media: Dordrecht, The Netherlands, 1994; pp. 229–231.
80. Hoffman, M.D.; Hoffman, D.R. Exercisers achieve greater acute exercise-induced mood enhancement than nonexercisers. *Arch. Phys. Med. Rehabil.* **2008**, *89*, 358–363. [CrossRef] [PubMed]
81. Ament, W.; Verkerke, G.J. Exercise and fatigue. *Sport. Med.* **2009**, *39*, 389–422. [CrossRef] [PubMed]
82. Hargens, T.A.; Kaleth, A.S.; Edwards, E.S.; Butner, K.L. Association between sleep disorders, obesity, and exercise: A review. *Nat. Sci. Sleep* **2013**, *5*, 27–35. [CrossRef] [PubMed]
83. Falkai, P.; Schmitt, A.; Rosenbeiger, C.P.; Maurus, I.; Hattenkofer, L.; Hasan, A.; Malchow, B.; Heim-Ohmayer, P.; Halle, M.; Heitkamp, M. Aerobic exercise in severe mental illness: Requirements from the perspective of sports medicine. *Eur. Arch. Psychiatry Clin. Neurosci.* **2022**, *272*, 643–677. [CrossRef] [PubMed]
84. Romani, W.A.; Gallicchio, L.; Flaws, J.A. The association between physical activity and hot flash severity, frequency, and duration in mid-life women. *Am. J. Hum. Biol. Off. J. Hum. Biol. Assoc.* **2009**, *21*, 127–129. [CrossRef]
85. Guay, A.T.; Spark, R.F.; Bansal, S.; Cunningham, G.R.; Goodman, N.F.; Nankin, H.R.; Petak, S.M.; Perez, J.B.; Law, B.; Garber, J.R. American Association of Clinical Endocrinologists medical guidelines for clinical practice for the evaluation and treatment of male sexual dysfunction: A couple's problem–2003 update. *Endocr. Pract.* **2003**, *9*, 77–95. [CrossRef]
86. Mead, G.E.; Morley, W.; Campbell, P.; Greig, C.A.; McMurdo, M.; Lawlor, D.A. Exercise for depression. *Cochrane Database Syst. Rev.* **2008**, *8*, *CD004366*.
87. Gombos, G.C.; Bajsz, V.; Pék, E.; Schmidt, B.; Sió, E.; Molics, B.; Betlehem, J. Direct effects of physical training on markers of bone metabolism and serum sclerostin concentrations in older adults with low bone mass. *BMC Musculoskelet. Disord.* **2016**, *17*, 254. [CrossRef]
88. Copp, D.H.; Cameron, E.; Cheney, B.A.; Davidson, A.G.F.; Henze, K. Evidence for calcitonin—a new hormone from the parathyroid that lowers blood calcium. *Endocrinology* **1962**, *70*, 638–649. [CrossRef]
89. Staud, R. Vitamin D: More than just affecting calcium and bone. *Curr. Rheumatol. Rep.* **2005**, *7*, 356–364. [CrossRef]
90. Rylander, R.; Megevand, Y.; Lasserre, B.; Amstutz, W.; Granbom, S. Moderate alcohol consumption and urinary excretion of magnesium and calcium. *Scand. J. Clin. Lab. Investig.* **2001**, *61*, 401–405. [CrossRef] [PubMed]
91. Mitri, J.; Muraru, M.; Pittas, A. Vitamin D and type 2 diabetes: A systematic review. *Eur. J. Clin. Nutr.* **2011**, *65*, 1005–1015. [CrossRef] [PubMed]
92. Dorozhkin, S.V. Calcium orthophosphates. *J. Mater. Sci.* **2007**, *42*, 1061–1095. [CrossRef]
93. Trevisan, C.; Alessi, A.; Girotti, G.; Zanforlini, B.M.; Bertocco, A.; Mazzochin, M.; Zoccarato, F.; Piovesan, F.; Dianin, M.; Giannini, S. The impact of smoking on bone metabolism, bone mineral density and vertebral fractures in postmenopausal women. *J. Clin. Densitom.* **2020**, *23*, 381–389. [CrossRef]

94. Raisz, L.G. Bone resorption in tissue culture. Factors influencing the response to parathyroid hormone. *J. Clin. Investig.* **1965**, *44*, 103–116. [CrossRef]
95. Ju, H.S.J.; Leung, S.; Brown, B.; Stringer, M.A.; Leigh, S.; Scherrer, C.; Shepard, K.; Jenkins, D.; Knudsen, J.; Cannon, R. Comparison of analytical performance and biological variability of three bone resorption assays. *Clin. Chem.* **1997**, *43*, 1570–1576. [CrossRef]