The Use of Oncology Electronic Health Record Databases to Assess the Effectiveness of Breast Cancer Treatment

HARVARD UNIVERSITY

Graduate School of Arts and Sciences

DISSERTATION ACCEPTANCE CERTIFICATE

The undersigned, appointed by the
Committee on Higher Degrees in Population Health Sciences,
have examined a dissertation entitled

"The Use of Oncology Electronic Health Record Databases to Assess the Effectiveness of Breast Cancer Treatment"

presented by

DAVID MEROLA

candidate for the degree of Doctor of Philosophy
and hereby certify that it is worthy of acceptance.

Dr. Sebastian Schneeweiss, M.D., Sc.D., Committee Chair,
Harvard T.H. Chan School of Public Health, Harvard Medical School

J. KLIN (Jan 11, 2022 16:27 EST)

Dr. Joshua Kueiyu Lin, M.D., Sc.D., Harvard Medical School

Jessica Young
Jessica Young (Jan 11, 2022 16:31 EST)

Dr. Jessica Gerald Young, Ph.D., Harvard T.H. Chan School of Public Health, Harvard Medical School

Dr. Deborah Schrag, M.D., Harvard Medical School

Date: 11 January 2022

The Use of Oncology Electronic Health Record Databases to Assess the Effectiveness of Breast Cancer Treatment

A dissertation presented

by

David Merola

to

The Department of Population Health Sciences

Graduate School of Arts and Sciences

and

The Department of Epidemiology

Harvard T.H. Chan School of Public Health

In partial fulfillment of the requirements

For the degree of

Doctor of Philosophy

In the subject of

Population Health Sciences (Field of Study: Epidemiology)

Harvard University

Cambridge, MA

January 2022

Dissertation Advisor: Sebastian Schneeweiss David Merola

The use of oncology electronic health record databases to assess the effectiveness of breast cancer treatments

Abstract

Background: Non-experimental studies using large healthcare databases may be well-suited for addressing relevant questions in clinical oncology that pertain to the safety and effectiveness of medications. They complement randomized trials by including frail and complex patients seen in routine care that reflect real-world practice patterns and treatment adherence. Historically, pharmacoepidemiology research in the oncology setting has been limited, mainly due to poor capture of important confounding factors in real-world data sources (e.g., tumor grade, histology, and location, laboratory values, biomarkers, and performance status). However, more recently, quality and availability of secondary data in oncology have been emerging in specialized electronic health record (EHR) systems. These longitudinal databases are derived from several major sources of clinical information: 1) Physician medication ordering systems, 2) Physician notes from outpatient oncology encounters, 3) Molecular diagnostics, 4) Structured fields within the health record. Collectively, such data sources permit ascertainment of patients' demographics, cancer types, treatment history, and an array of confounders and health outcomes necessary for comparative effectiveness studies of oncology drugs. Despite these advancements, the use of oncology EHR databases still poses many challenges that stem from a lack of linkage to alternative data sources, such as claims or high-quality tumor registries. This results in poor capture of out-of-network encounters, medical procedures, or inpatient encounters, as well as missing data. Consequently, it is unknown whether these challenges can be overcome with currently available epidemiological and statistical methods, and ultimately if these data are suitable for clinical investigations.

The objectives of this body of work are to: 1) explore the utility of specialty oncology EHR databases in comparative effectiveness research; 2) build a framework that will support drawing causal conclusions from EHR-based studies in the oncology setting in light of the limitations of EHRs; and 3) identify and implement markers for data quality and study validity that can be used to assess confidence in findings. To achieve these objectives, two comparative effectiveness studies of first-line treatments for advanced breast cancer were conducted and calibrated against randomized clinical trials—the PALOMA-2 trial and the PARSIFAL trial. Additionally, an algorithm was constructed to predict completeness in an EHR-based oncology cohort, which was subsequently implemented in the two comparative effectiveness studies as a sensitivity analysis. In particular, effect estimates in the non-randomized studies were calculated among subjects with increasingly higher levels of predicted data completeness to see if the estimates converged to the randomized trial estimates. In this way, predicted completeness was assessed as a potential tool to improve study validity.

Methods: To construct the prediction algorithm for data completeness, a Medicare-linked EHR database derived from two academic medical centers in Massachusetts was used. This linked database was constructed from many sources of clinical information; namely, healthcare claims (inpatient, outpatient, and pharmacy), physician drug orders, unstructured notes, and billing codes from medical procedures and inpatient or outpatient provider encounters. This permitted ascertainment of patient demographics, vitals, height and weight, medical procedures, medications, timing of provider encounters, and diagnoses, which were used to create candidate predictors of data completeness. The study population consisted of subjects that had a year of continuous enrollment in Medicare, were at least 65 years old, and had one or more outpatient oncology encounter in the EHR system. Data completeness was quantified by the "continuity ratio," defined

as the yearly proportion of outpatient encounters reported to Medicare that were captured by EHR data. Least absolute shrinkage and selection operator (LASSO) regression was used to select candidate predictors, which were regressed on continuity ratio. The performance of the final model was assessed using the coefficient of determination and Spearman's correlation of predicted vs. observed EHR-continuity. We quantified misclassification of several comorbidities and medications within deciles of continuity ratio by calculating the ratio and standardized difference of the proportion of subjects classified as having each covariate when using outpatient EHR data alone vs. outpatient EHR data and claims.

For the first comparative effectiveness study, an oncology EHR database derived from outpatient oncology practices within the US Oncology Network was used to estimate the rate of time-to-next-treatment (TTNT) in palbociclib-letrozole users versus letrozole-only users. TTNT was chosen as an endpoint because it was well-observed in the EHR database and appeared to serve as a meaningful surrogate for treatment effectiveness in the PALOMA-2 trial. All eligibility criteria, treatments, and outcome variables were defined to mimic the trial as closely as possible. Patients with evidence of a breast cancer subtype inconsistent with the PALOMA-2 study population (i.e., hormone-negative, HER-2 positive) were excluded. To address missing data, 50 complete datasets were constructed using multiple imputation by chained equations. In each of the imputed datasets, a Cox proportional hazards model was fit to estimate the hazard ratio of TTNT in an intention-to-treat analysis analogous to the trial. All 50 estimates were subsequently pooled.

In the second comparative effectiveness study, a similar approach was undertaken. We used the same longitudinal EHR data from outpatient oncology practices across the US to emulate the PARSIFAL trial in its treatments and selection criteria as closely as possible. Multiple imputation was employed to account for missing data in patient characteristics. Baseline characteristics were compared and hazard ratios with 95% confidence intervals for overall survival were estimated fitting

a multivariable proportional hazards model. Findings in both comparative effectiveness studies were compared to their respective RCTs result with qualitative assessment and standardized difference estimates.

Results: In the PALOMA-2 emulation study, there were 3,836 study-eligible advanced breast cancer subjects. The hazard ratio for TTNT in the observational study (HR: 0.62; 95% CI: 0.56-0.68) was closely aligned with that of the randomized trial (HR: 0.64; 95% CI: 0.52-0.78) (*Standardized Difference* = -0.05). In the PARSIFAL trial emulation, 1,886 subjects were selected into the study cohort following application of all eligibility criteria. Although the 3-year survival was meaningfully lower in clinical practice (59%) compared to the RCT (78%), the relative effect size was HR=1.07 (95% CI: 0.86 – 1.35), similar to the RCT (HR=1.00; 0.68 – 1.48, *Standardized Difference* = 0.04). Restriction of the study cohort by increasing levels of continuity ratio did not appreciably influence effect estimates in the PALOMA-2 trial emulation, but shifted the effect estimate of the PARSIFAL trial emulation away from the RCT estimate with wider confidence intervals.

Conclusion: This body of work calls for more emulations using a principled approach and methods for addressing the various threats to validity that can arise from the use of oncology EHR databases. Likewise, agreed-upon reporting standards can facilitate summarization of global efforts in advancing the use of RWD in clinical oncology. In the context of comparative effectiveness studies of oncology drugs, confounding may not be the most critical issue given the current data density in oncology EHR systems. Rather, it may be that more complete data will be needed for specific outcomes and possibly biomarkers. Overall, the field of real-world evidence in oncology is developing in a very positive direction as we are applying causal inference methods and as data sources continue to evolve and become richer in data granularity and continuity.

Table of Contents

List of Tables

List of Figures

Acknowledgements

As I approach the end of this journey that is my dissertation, I'm filled with the deepest sense of gratitude for all those that made this experience possible and enabled my growth as a person and professional.

To Sebastian: Your guidance, vision, and shared enthusiasm has not only inspired me to take a leap of faith in times of doubt, it has also made our work incredibly rewarding. Thank you for teaching me how to think "big-picture" and for your encouragement in our meetings. I could not have asked for a better mentor and advisor.

To Jessica, Deb, and Josh: Hearing each of your patterns of thinking and approach to conducting research has helped me adopt my own. Thank you for all your keen and incisive feedback throughout this process, and for the time you took to educate me. I feel incredibly fortunate to have worked alongside each of you—it has been an honor and privilege.

To Ken Shermock and Robbie Kidd: You have had a tremendous impact on my life by introducing me to research and the scientific method early on during my pharmacy training. I would not have found my passion and reached this point had it not been for your willingness to teach and support me with applied experiences beyond the classroom. Thank you for extending yourselves to help me grow.

Dedication

This dissertation is dedicated to Imma, Abba, Jonathan, and Bernice. Your love and guidance are the most meaningful forces in my life and none of this would be possible without you. Thank you for your unwavering support and presence during my successes and failures, the good times and the trying times.

Chapter I. An algorithm to predict data completeness in oncology electronic medical records for comparative effectiveness research

David Merola, PharmD, MS[1,2], Sebastian Schneeweiss, MD, ScD[1,2], Deborah Schrag, MD, MPH[3], Joyce Lii, MS, MA[1], Kueiyu Joshua Lin, MD, ScD[1]

[1] Division of Pharmacoepidemiology and Pharmacoeconomics, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA

[2] Department of Epidemiology, Harvard TH Chan School of Public Health, Boston, MA, USA

[3] Department of Medicine, Memorial Sloan Kettering Cancer Center, Weill Cornell Medical School New York, NY, USA

ABSTRACT

Introduction: Electronic health record (EHR) discontinuity (missing out-of-network encounters) can lead to information bias. We sought to construct an algorithm that identifies high EHR-continuity among oncology patients.

Methods: The study population consisted of subjects that had a year of continuous enrollment in Medicare, were at least 65 years old, and had one or more outpatient oncology encounters in one EHR system of a large tertiary care academic medical center. Using a linked Medicare-EHR database and regression, we sought to 1) measure how often Medicare claims for outpatient encounters were substantiated by visits recorded in the EHR, and 2) predict continuity ratio, defined as the yearly proportion of outpatient encounters reported to Medicare that were captured by EHR data. Factors typically available in EHR databases that are derived from outpatient oncology practices were selected as predictors. The prediction model's performance was evaluated with the coefficient of determination and Spearman's correlation of predicted vs. observed continuity ratio. We quantified misclassification of several comorbidities and medications within deciles of continuity ratio by calculating the ratio and standardized difference of the proportion of subjects classified as having each covariate when using outpatient EHR data alone vs. outpatient EHR data and claims.

Results: A total of 79,678 subjects met all eligibility criteria, of which half (n = 39,839) were used for model training and the remaining half for validation. Predicted and observed continuity ratio were highly correlated ($\sigma_{Spearman} = 0.86$). Patients with top 50% of predicted EHR-continuity had 2.19-fold (95% CI: 2.15 – 2.24) greater sensitivity than the remaining population for cancer diagnosis. The corresponding estimate was 5.88 (95% CI: 5.57 – 6.21) for non-cancer co-morbidities, 5.59 (95% CI: 5.37 – 5.81) for chemotherapy, and 3.32 (95% CI: 3.24 – 3.41) for non-

chemotherapy drug variables. Patients across levels of EHR-continuity have a similar comorbidity profile.

Conclusion: In the oncology setting, restricting EHR-based study cohorts to subjects with high continuity may reduce misclassification without greatly impacting representativeness. Further work is needed to elucidate the best manner of implementing continuity prediction rules in cohort studies.

BACKGROUND & INTRODUCTION

Non-experimental studies using large healthcare databases may be well-suited for addressing relevant clinical questions that pertain to the safety and effectiveness of medications. They complement randomized trials by including frail and complex patients seen in routine care that reflect real-world practice patterns and treatment adherence.

In oncology, the quality and availability of secondary data have been improving with the use of specialized electronic health record (EHR) systems. Oncology EHR databases[1, 2] draw upon several sources of clinical information, including chemotherapy and other medication orders, outpatient notes from oncology-related encounters, and biomarkers/molecular diagnostics. Collectively, these data sources allow investigators to define populations of interest, as well as ascertain key variables needed for studies of cancer treatment effects. However, the validity of investigations that employ these databases may be moderated by a lack of linkage to additional data sources (e.g., healthcare claims) that capture out-of-network care.

Out-of-network encounters, in which patients receive new diagnoses, medicines, or procedures, may not be recorded in a specialized oncology EHR system, particularly if EHR systems are not linked across a health information exchange or if EHR system linkages are uncoupled from the EHR database.[3] Such data leakage has been shown to cause a substantial amount of information bias (i.e., misclassification of study variables).[4] We have previously demonstrated that information bias due to EHR discontinuity, defined as "receiving care outside of reach of the study EHR," can be substantially reduced by applying a prediction rule to identify patients who receive a high degree of within-network care.[5] However, it is unclear whether this prediction model is generalizable to the oncology population since oncology EHR databases are typically comprised of only records from

outpatient community practices. Also, patients with active malignancy may have unique health-seeking behavior patterns, confounders, and effect modifiers relevant to CER in oncology, such as tumor stage or other cancer-specific factors.

In this study, we sought to extend prior work into the oncology setting by developing and validating an oncology-specific algorithm to identify patients with high EHR-continuity among patients with malignancy. We built this prediction model using outpatient-only EHRs to mimic oncology EHR databases that are derived exclusively from outpatient oncology practices. We evaluated the impact of EHR-continuity on misclassification of study variables relevant for a typical oncology CER and the representativeness of the patient co-morbidity profiles of those deemed to have high versus low EHR-continuity.

METHODS

Data Resources

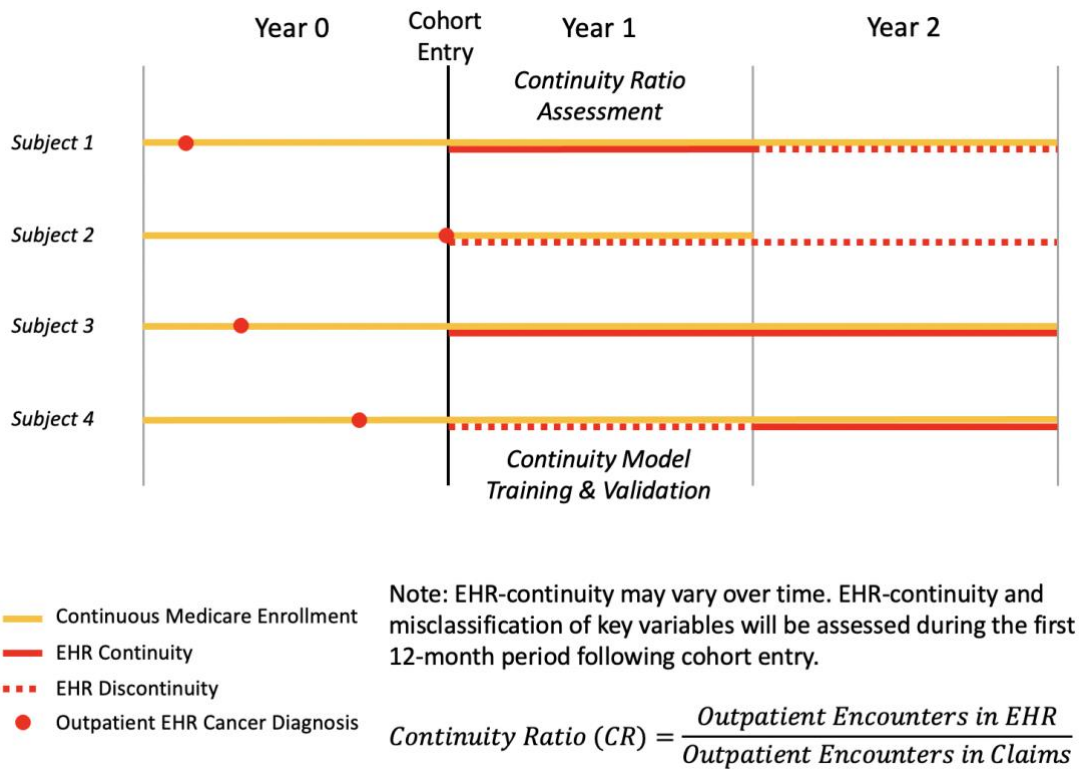
This study utilized an EHR database (2007-2017) comprised of records from two academic medical centers in Massachusetts. One of the centers included records from 1 tertiary hospital, 2 community hospitals, and 19 primary care centers, while the other network included records from 1 tertiary hospital, 1 community hospital, and 18 primary care centers. Medicare claims data were linked to the EHR database to facilitate measurement of data continuity, defined below. Collectively, the linked database was comprised of information from multiple sources, including healthcare claims (inpatient, outpatient, and pharmacy), physician drug orders, unstructured notes, and billing codes from medical procedures and inpatient or outpatient provider encounters. Through these data sources, patient demographics, vitals, height and weight, medical procedures, medications, timing of

provider encounters, and diagnoses were available, which were used to create candidate predictors of EHR-continuity.

Study Design and Cohort Selection

The study cohort consisted of those aged 65 years or older with continuous enrollment in Medicare fee-for-service Parts A and B (i.e., medical coverage), and D (i.e., prescription coverage) for 365 days, and had at least one outpatient encounter during this enrollment period. Furthermore, subjects that had a different sex reported in each database were excluded to improve the accuracy of the linkage between the two databases. In an attempt to enhance the generalizability of our final prediction model to specialty oncology EHR databases that exclusively contain outpatient records, only outpatient EHR records were used and subjects were required to have at least 1 cancer diagnosis, defined by International Classification of Diseases codes in the EHR, during the continuous Medicare enrollment period described above. The cohort entry date (henceforth, "index date") was defined as the first day at which all eligibility criteria were met. Follow-up proceeded for 365 days after the index date and patients that experienced loss to follow-up due to death, disenrollment in the claims database, or end of the study period (12/31/2017) were censored from our analysis to ensure all subjects analyzed had an equivalent assessment of candidate predictors with respect to time. A diagram of cohort selection and study design is shown in Figure 1.1. We allowed EHR-continuity to vary for each 12-month period and developed the prediction model using data from the first year following the index date.

Figure 1.1. A schematic of the study design and cohort selection



EHR-Continuity and Candidate Predictors

Continuity in the EHR database was measured in the one-year follow-up period after cohort entry and defined as the proportion of outpatient health record encounters identified in the Medicare database (i.e., "continuity ratio"). One year was chosen for continuity assessment because it is a common time frame used to measure baseline characteristics in comparative effectiveness studies, which is one setting that the final model could be applied. Furthermore, longer time periods than one year may result in continuity measurements that are crude, while shorter time periods may result in unstable estimates by not allowing sufficient time for candidate predictor to occur in the data.

$$\text{Continuity Ratio } (CR) = \frac{Outpatient\ Encounters\ in\ EHR}{Outpatient\ Encounters\ in\ Claims}$$

Within the EHR database, encounter types were differentiated based on clinic names, presence of certain billing codes, and dates of admission/discharge to ensure exclusion of laboratory-only, inpatient, non-acute institutional stays (i.e., rehabilitation center), radiology-only, and pharmacy-only visits from the CR calculation. Encounters identified as outpatient in the EHR database were matched to outpatient Medicare encounters by admission and discharge date. Outpatient EHR encounters that couldn't be matched to outpatient Medicare encounters by date were excluded from the CR calculation; however, Medicare encounters that could not be matched to EHR encounters were retained for the purposes of CR calculation.

Like the CR, candidate predictors of the CR were also measured during the first year following cohort entry (Table 1.1). These variables were chosen as they were thought to be predictive of engagement with the health system and EHR-continuity. Generally, they could be classified into the following categories: 1) Medication and diagnosis recordings in the EHR; 2) Preventive interventions; and 3) Markers of an encounter. These predictors were considered to be available in most oncology EHR databases, which often do not contain information on medical procedures or inpatient care.

Table 1.1. Patient Characteristics and Candidate Predictors of EHR Continuity

Patient Characteristics	Training Dataset (n = 39,839)	Validation Dataset (n = 39,839)
Demographics		
Age; mean (sd)	72.9 (6.7)	73.0 (6.7)
Female Sex; n (%)	22267 (55.9)	22,359 (56.1)
Cancer Type[a]		
Lung; n (%)	1467 (3.7)	1435 (3.6)
Breast; n (%)	3918 (9.8)	3895 (9.8)
Prostate; n (%)	3297 (8.3)	3256 (8.2)
Melanoma; n (%)	1088 (2.7)	1084 (2.7)
Stomach; n (%)	149 (0.4)	115 (0.3)
Pancreatic; n (%)	221 (0.6)	215 (0.5)
Colorectal; n (%)	1135 (2.8)	1136 (2.9)
Uterine; n (%)	572 (1.4)	590 (1.5)
Leukemia; n (%)	576 (1.4)	598 (1.5)
Non-Hodgkin Lymphoma; n (%)	1282 (3.2)	1401 (3.5)
Multiple Myeloma; n (%)	450 (1.1)	528 (1.3)
EHR Continuity Candidate Predictors[a]		
BMI Recorded; n (%)	18432 (46.3)	18538 (46.5)
At least 1 outpatient visit; n (%)	33540 (84.2)	33456 (84.0)
At least 2 outpatient visits; n (%)	28927 (72.6)	28764 (72.2)
At least 1 basic fact[b]; n (%)	20661 (51.9)	20706 (52.0)
At least 2 basic facts[b]; n (%)	14689 (36.9)	14869 (37.3)
At least 1 diagnosis recorded; n (%)	30759 (77.2)	30738 (77.2)
At least 2 diagnoses recorded; n (%)	25216 (63.3)	25205 (63.3)
At least 1 medication recorded; n (%)	27203 (68.3)	27234 (68.4)
At least 2 medications recorded; n (%)	22341 (56.1)	22264 (55.9)
At least 1 vaccination record; n (%)	4900 (12.3)	4992 (12.5)
Markers of Severe Disease[a]		
Chemotherapy Recorded; n (%)	3785 (9.5)	3674 (9.2)
MRI Recorded; n (%)	1076 (2.7)	1026 (2.6)
Metastatic Cancer; n (%)	2642 (6.6)	2522 (6.3)

[a] Factors considered for building the prediction model for high EHR continuity in patients with malignancy

[b] Basic routine care facts include: height, weight, diastolic or systolic blood pressure, body mass index (BMI), and smoking status

Prediction Model Development

The study cohort was first split at random into two equal parts that were used for training and validation of our prediction model, respectively. The data were split evenly because there was a large initial sample size and the variances of the model's parameters and performance statistics were of equal priority. A random sample was chosen to maximize the heterogeneity of the training data and resultant generalizability of the prediction model to different databases. In our prior work using EHR data from a general population, the data were not randomly divided and allocated by institution for training and validation purposes. However, allocating the data in a similar manner for this study would result in an under-represented training dataset. In particular, this would result in an oncology department of a tertiary cancer center being placed in a separate dataset from the cancer center it is affiliated with.

Using the training data, CR was regressed on 26 candidate predictors (Table 1.1) using a least absolute shrinkage and selection operator (LASSO) linear regression model. Five-fold cross-validation was used to choose the optimal tuning parameter. LASSO regression was chosen for two reasons. First, the model would yield a set of coefficients that could be applied by investigators to EHR-based oncology study cohorts. Second, the model permits shrinkage of coefficients to zero, allowing selection of a subset of candidate variables to optimize prediction accuracy and reduce the chance of over-fitting.

Two statistics were used to assess model fit in the validation dataset: (1) Spearman's Correlation Coefficient ($\sigma_{Spearman}$) for observed vs. predicted CR, and (2) the coefficient of determination (R^2). Collectively, these metrics were chosen because they illuminate how well the model ranks patients

based on their observed CR (i.e., monotonicity), while also providing a measure of how much variability in CR is explained by the model's covariates.

Misclassification Assessment

We assessed variable misclassification by decile of predicted CR. Since claims data capture information across the healthcare continuum, the classification based on EHR plus claims data was used as the benchmarking proportion against which the misclassification of the prevalence based on EHR data only was quantified:

1) Sensitivity:

$$\text{Sensitivity} = \frac{\Pr(X|based\ on\ outpatient\ EHR)}{\Pr\left(X \,|\, based\ on\ outpatient\ EHR\ \&\ claims\ Data\right)}$$

where "X" is a binary marker of the presence of a comorbidity or medication. These variables were all measured and evaluated in the same one-year period as CR and candidate predictors.

2) Standardized differences (SDiff):

$$\text{Standardized Difference} = \frac{Pr(X|Based\ on\ outpatient\ EHR) - Pr\left(X \,|\, based\ on\ outpatient\ EHR\ \&\ claims\ data\right)}{\sqrt{\left[\sigma^2_{Outpatient\ EHR} + \sigma^2_{Outpatient\ EHR\ \&\ Claims\ Data}\right]/2}}$$

where σ^2 denotes the variance of each proportion specified.

Sensitivity and SDiff were chosen because they highlight different dimensions of misclassification, which may be more apparent on the multiplicative vs. additive scale or vice versa. In keeping with convention, an absolute value of SDiff >0.1 was deemed as a relevant difference between two groups.[6] A total of 11 cancers (i.e., lung, breast, prostate, melanoma, stomach, pancreatic, colorectal, uterine, multiple myeloma, leukemia, and non-Hodgkin lymphoma), 19 non-cancer comorbidities, 1 chemotherapy, and 39 non-chemotherapy medication-related variables were assessed, which were

chosen because they are prognosticators of mortality, proxies for overall health, or potential confounders in comparative effectiveness studies of oncology therapies.[7-9]

Representativeness Assessment

In addition to misclassification, it is possible that restriction to subjects with high CR may affect representativeness. To evaluate this, SDiff was calculated for demographics as well as the comorbidities specified above based on both EHR and claims data between subjects in the top 50th percentile of CR versus the remaining population. As a summary measure, an unweighted mean of all absolute standardized differences (MSD) was calculated among distinct demographics, comorbidities, and medications.

RESULTS

Study Population

Among 348,199 subjects with at least 365 days of enrollment in the Medicare dataset and one or more outpatient oncology encounters in the EHR database, 79,678 met all study eligibility criteria (Supplementary Table 1.1). Of these, 39,839 were allocated to training and validation of the prediction model, respectively. In the overall study cohort, the mean age was 73 years (standard deviation: 6.7) and 56% of subjects were female (n = 44,626) (Table 1.1). There were no notable differences in measured characteristics between the training and validation datasets (Table 1.1).

Model Development and Performance

Among the 26 candidate predictors, 10 variables were selected into the final LASSO regression model, with the presence of a medication record and routine care fact(s) being the greatest predictors of continuity (**Table 1.2**). In the validation dataset, a strong correlation between observed and predicted continuity ratio was observed ($\sigma_{Spearman} = 0.86$) and the model explained 69% of the variability in CR ($R^2 = 0.69$). Similar performance was observed in the training dataset ($\sigma_{Spearman} = 0.86$ and $R^2 = 0.68$).

Table 1.2. Parameter Estimates from Continuity Ratio LASSO Regression Model (Training Data)

Variable	Coefficient
(Intercept)	0.016
BMI recorded	0.073
At least 1 diagnosis	0.013
At least 2 diagnoses	0.054
At least 1 OP visit	0.050
At least 2 OP visits	0.051
At least 1 med	0.048
At least 2 meds	0.132
At least 1 basic fact	0.139
At least 2 basic facts	0.208
Influenza or Pneumococcal vaccine	0.157

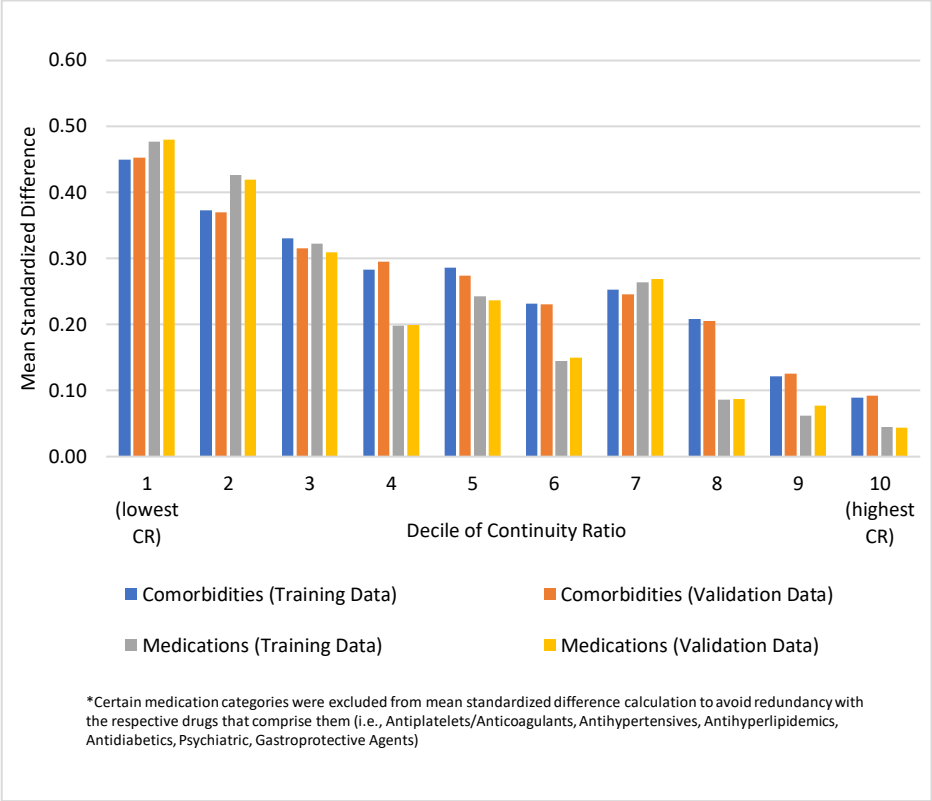
The dependent and all independent variables were ascertained in the 1-year time interval following cohort entry.
The model was estimated using half (n=39,839) of the study population, allocated for training.
Adjusted R-squared of the final model in the validation dataset was 0.69
Spearman's Correlation of observed vs. predicted continuity in the validation dataset was 0.86

Caption: Coefficients from final Least Absolute Shrinkage and Selection Operator (LASSO) regression model, predicting continuity ratio (CR).

Variable Misclassification Assessment

The assessment of MSD by decile of CR revealed a clear trend of improvement in classification for medication-related variables and comorbidities measured in both the training and validation datasets (Figure 1.2). Sensitivity also improved incrementally with higher levels of predicted continuity (Figure 1.3). These trends in MSD and sensitivity were not only apparent overall, but for nearly each individual variable as well (Supplementary Table 1.2 and Supplementary Table 1.3). On average across all variables measured, MSD was reduced by a factor of $1/7^{th}$ and sensitivity was improved 35-fold comparing subjects in the highest vs. lowest decile of CR.

Figure 1.2. Mean Standardized Difference of Select Comorbidities and Medications by Predicted Continuity Ratio

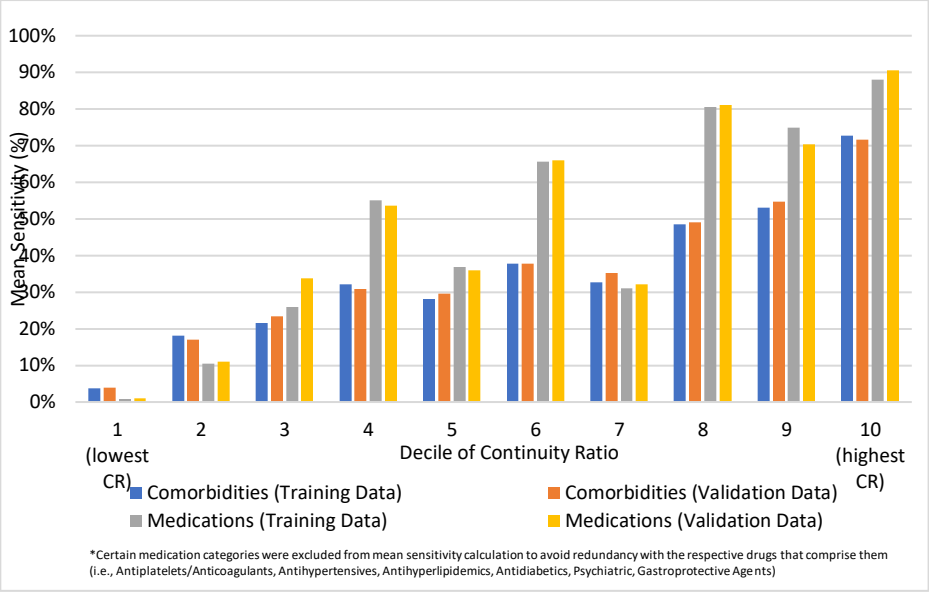


Caption: Mean standardized difference of 30 comorbidities and 40 medication-related variables within levels of predicted continuity ratio (CR) in the training and validation datasets. Lower numbers indicate less misclassification when comparing variables classified using EHR data only vs. EHR and claims data together.

Patients with top 50% of predicted EHR-continuity had 2.19-fold (95% CI: 2.15 – 2.24) greater sensitivity than the remaining population for cancer diagnosis (Table 1.3). The corresponding

estimate was 5.88 (95% CI: 5.57 – 6.21) for non-cancer co-morbidities, 5.59 (95% CI: 5.37 – 5.81) for chemotherapy, and 3.32 (95% CI: 3.24 – 3.41) for non-chemotherapy drug variables (Table 1.3). When quantified by MSD, misclassification also improved for these groups of variables when selected for high EHR-continuity (Table 1.3).

Figure 1.3. Mean Sensitivity of Select Comorbidities and Medications by Predicted Continuity Ratio



Caption: Mean sensitivity of 30 comorbidities and 40 medication-related variables within levels of predicted continuity ratio (CR) in the training and validation datasets. Higher numbers indicate less misclassification when comparing variables classified using EHR data only vs. EHR and claims data together.

Representativeness of Subjects with High EHR-continuity

Overall, the prevalence of measured patient characteristics did not appear to be appreciably different between subjects in the top 50th percentile of CR vs. the remaining population (Supplementary Table 1.4). The MSD for comorbidities (MSD = 0.03) and demographics (MSD = 0.02) were all below the pre-specified cutoff of 0.1 (Supplementary Table 1.4) in the validation set. We found a similar pattern in the training set (data not shown).

Table 1.3. Misclassification of Oncology-Related vs. Non-Oncology Variables (Validation Dataset)

Misclassification metric	Predicted EHR-continuity	Cancer diagnoses	Non-cancer comorbidities	Chemotherapies	Non-chemotherapies
Sensitivity*	Top 50%	0.73	0.38	0.65	0.80
	Lower 50%	0.33	0.07	0.12	0.24
	Relative ratio comparing top vs. lower 50% (95% CI)	2.19 (2.15 - 2.24)	5.88 (5.57 - 6.21)	5.59 (5.37 - 5.81)	3.32 (3.24 - 3.41)
Mean Standardized Difference**	Top 50%	0.07	0.26	0.22	0.09
	Lower 50%	0.19	0.47	0.63	0.34
	Relative ratio comparing top vs. lower 50% (95% CI)	0.36 (0.34 – 0.37)	0.55 (0.54 – 0.56)	0.36 (0.34 – 0.37)	0.27 (0.26 – 0.28)

* Excludes 'Metastatic Cancer' and 'General Cancer' variables from calculation to avoid redundancy
**Certain medication categories were excluded from calculation to avoid redundancy with the respective drugs that comprise them (i.e., Antiplatelets/Anticoagulants, Antihypertensives, Antihyperlipidemics, Antidiabetics, Psychiatric, Gastroprotective Agents). Mean of absolute values of standardized differences were taken.

DISCUSSION

Information bias due to EHR-discontinuity presents a major challenge in comparative effectiveness studies of drugs that are based on EHR data in oncology. This may be particularly problematic when administrative claims that record healthcare encounters across networks are not available for linkage, as is often the case due to patient privacy concerns. In this study, we identified predictors for EHR-continuity that can be used to mitigate variable misclassification in clinical investigations relying solely on outpatient oncology EHR databases. We found a consistent trend in which higher predicted EHR-continuity corresponded to less variable misclassification. These findings suggest that restricting analyses to those with high EHR-continuity in EHR-based oncology CER may improve the validity of effect estimates by reducing information bias (i.e., misclassification of the

study variables) in pre-exposure confounder measurement, as well as outcome surveillance. Furthermore, key characteristics of subjects with high versus low CR were similar, indicating that generalizability would not be greatly impacted with respect to those measured comorbidities and demographics.

The performance of our CR prediction model in oncology patients was similar to a previously developed model in a general population, with some notable differences. The oncology-based model had 4 fewer variables (16 vs. 20), explained more variability in CR (69% in the oncology specific population compared to 48% in the general population), and exhibited a similar correlation between observed and predicted CR (oncology model: $\sigma_{Spearman} = 0.86$; general model: $\sigma_{Spearman} = 0.82$), relative to the general model. The fact that we can use fewer variables to explain a larger proportion of the variability in EHR-continuity could be due to the particular care-seeking behavior among patients with malignancy. Since the presence of cancer and its treatment often affect multiple organ systems, it is possible that cancer patients tend to seek care in the same network for other medical needs, so that their cancer related information can be seen by other providers. Indeed, we found the CR has a mean of 0.45 (standard deviation: 0.40) whereas the corresponding mean capture proportion in the general population was 0.18 (standard deviation: 0.19). The difference in health-seeking behavior evident in the distribution of CR also supports our approach to develop an oncology-specific prediction model for EHR-continuity.

In comparing oncology-related variables to other variables, cancer diagnoses seemed to have less misclassification for the same level of continuity. This could be explained, in part, by the possibility of oncology diagnoses occurring predominantly in an outpatient setting, while other comorbidities might require more urgent, emergency department visit, and inpatient care. For example, myocardial

infarction, heart failure, or stroke were among non-cancer comorbidities measured that are all associated with emergent care. These variables tended to have greater misclassification, possibly due to worse capture in oncology EHR data that consists of outpatient encounters only. Despite this, the same general relationship between CR and misclassification was maintained for oncology-related variables.

In a cohort study design, one way of applying our algorithm is to use the predictors assessed in the baseline period prior to the cohort entry to identify the high EHR-continuity subjects to reduce misclassification of the covariates measured at baseline. Alternatively, our algorithm might be used as an artificial censoring rule during the follow-up period of a comparative effectiveness study to avoid including the person-time during follow-up when we have insufficient data to ascertain the outcome. However, if such censoring criteria is related to the treatment choice and outcome development, selection bias can occur and researchers should apply appropriate analytical strategies, such as marginal structural models, to mitigate the potential biases.[10, 11]

Our study has several limitations. First, the data resources and eligibility criteria we used may limit the generalizability of our findings to subjects who are Medicare beneficiaries aged at least 65 years. Although we attempted to mimic specialty oncology EHR databases by restricting our general EHR database to outpatient records and subjects with oncology diagnoses, our database was not exactly an oncology specialty EHR database and therefore may contain more diverse patient populations than does an oncology-focused group practice setting, such as that found in the *iKnowMed* EHR or Flatiron databases. Furthermore, our results may be highly sensitive to the data source used and, particularly, type of medical center(s) that our sample was drawn from. Continuity may be very

different among patients in smaller medical centers that are not the predominant regional referral center, such as that used in our study.

Second, our model was developed and tested using data from an academic healthcare network in a large metropolitan area. It is possible that the model may perform differently in other networks, particularly if reporting standards in alternate EHRs or health systems differ. Therefore, additional work is needed to validate the model's performance in EHR databases derived from other health systems that contain patient populations with different characteristics.

Third, this analysis only looked at a one-year time frame, while discontinuity may vary over many years. Despite this, our model's performance over a longer study period is not expected to differ since our prior work in a general population has shown consistency in performance for up to 7 separate years in the same cohort.

Fourth, as the tolerance of information bias may vary by research context, it is challenging to establish clear guidance on a cutoff of CR to achieve a unanimously acceptable degree of misclassification in EHR-based studies. Despite this, our model seems to rank the extent of misclassification between patients very well, making it a reasonable means of ranking subjects by their likelihood of having measurement bias relative to one another. For the purpose of reducing information bias in CER, based on existing literature,[12] using the patients with top 50% predicted EHR-continuity may be reasonable starting point but sensitivity analyses that vary such a cut-off are recommended.

CONCLUSION

Using outpatient EHR data derived from oncology patients, we created a prediction rule that can identify subjects receiving a high degree of within-network care in which information bias is much reduced. Oncology patients with high EHR-continuity have a comparable co-morbidity profile compared to other cancer patients. Future work is needed to validate our findings in other sources of oncology-based EHR data and to determine the actual impact of the algorithm-identified high-continuity cohort on the treatment effect estimation in real-world oncology comparative effectiveness research.

Author Contributions

Conception or design of the work: David Merola, Kueiyu Joshua Lin

Collection and assembly of data: Joyce Lii

Data analysis and interpretation: David Merola

Drafting the article: David Merola

Critical revision of the article: Sebastian Schneeweiss, Jessica Young, Deborah Schrag, Kueiyu Joshua Lin

Final approval of the version to be published: All authors

Chapter II: Exploring the application of a specialized electronic health record database for comparative effectiveness research in oncology: A clinical trial emulation

David Merola, PharmD, MS[1,2], Jessica Young, PhD, MPH[2,3,4], Deborah Schrag, MD, MPH[5], Kueiyu Joshua Lin, MD, ScD[1,6], Sarah Alwardt, PhD[7], Sebastian Schneeweiss, MD, ScD[1,2]

[1] Division of Pharmacoepidemiology and Pharmacoeconomics, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA

[2] Department of Epidemiology, Harvard TH Chan School of Public Health, Boston, MA, USA

[3] Department of Population Medicine, Harvard Medical School & Harvard Pilgrim Health Care Institute, Boston, MA, USA

[4] CAUSALab, Harvard TH Chan School of Public Health, Boston, MA, USA

[5] Department of Medicine, Memorial Sloan Kettering Cancer Center, Weill Cornell Medical School New York, NY, USA

[6] Department of Medicine, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA

[7] Ontada LLC, The Woodlands, TX, USA

ABSTRACT

Introduction: Oncology electronic health record (EHR) databases have increased in quality and availability over the past decade. Yet it remains unclear whether these secondary data resources can be used to conduct reliable comparative effectiveness studies. We sought to emulate a clinical trial with EHR data in the advanced breast cancer population and calibrate our results against the trial.

Methods: An EHR Database derived from outpatient oncology practices within the US Oncology Network (2005-2021) was used to emulate time-to-next treatment (TTNT), an exploratory endpoint reported in a follow-up study of participants in the PALOMA-2 trial. TTNT is well-defined in our data source and, therefore, more amenable for calibration against the randomized study results relative the PALOMA-2 trial's primary and secondary endpoints. In the non-randomized study, all eligibility criteria, treatments, and outcome variables were defined to mimic the trial as closely as possible. Patients with evidence of a breast cancer subtype inconsistent with the PALOMA-2 study population (i.e., hormone-negative, HER-2 positive) were excluded. To address missing data, 50 complete datasets were constructed using multiple imputation by chained equations. In each of the imputed datasets, a Cox proportional hazards model was fit to estimate the hazard ratio of TTNT in an intention-to-treat analysis analogous to the trial. All 50 estimates were subsequently pooled.

Results: There were 3,836 study-eligible advanced breast cancer patients. The hazard ratio for TTNT in the observational study (HR: 0.62; 95% CI: 0.56-0.68) was closely aligned with that of the randomized trial (HR: 0.64; 95% CI: 0.52-0.78).

Conclusion: Under our assumptions regarding missing data and comparability of the two study populations, our non-experimental study coincided with that of the randomized trial, lending support to the use of observational databases for causal inference when carefully analyzed.

BACKGROUND & INTRODUCTION

Legislative and technological changes over the past decade have given rise to the use of healthcare databases (e.g., administrative claims, electronic health records) in clinical research.[13, 14] Traditionally, the utility of these databases in the context of oncology has been limited due to their poor capture of key clinical characteristics (e.g., tumor stage, histology, performance status, etc.). In order for a data source to have utility in comparative effectiveness research of drug therapies, it must adequately capture patients' treatment history, outcome(s) of interest, and key prognosticators of the outcome(s).

New specialized electronic health record (EHR) databases[1, 2] that meet these standards are rapidly emerging. These data draw upon information from structured fields within the health record in outpatient notes of community oncology encounters, oncologist orders for chemotherapy and other medications, and biomarkers/molecular diagnostics, permitting researchers to ascertain key confounders, sequential treatment history, clinical outcomes, and other important longitudinal clinical information. Despite these improvements, specialized oncology EHR databases have limitations that stem from their lack of linkage to other data sources. For example, missing values and incomplete capture of encounters across the healthcare continuum are characteristic of these data because they exclusively rely on outpatient encounters from an EHR system with no input from high-quality tumor registries, other health care networks, or inpatient records. Consequently, the utility of these data in conducting comparative effectiveness research of oncology drug regimens has yet to be elucidated.

One approach to establishing whether specialized oncology EHR databases can be used for drug effectiveness research is to calibrate database studies against randomized clinical trials.[15, 16] If a

thoughtfully analyzed observational study's result is congruent with that of the clinical trial, assuming closely emulated treatment, outcome, and eligibility criteria, then such a finding would support the validity of using the database to carry out effectiveness studies in that particular setting.

We aimed to estimate treatment effects on time-to-next treatment (TTNT), an exploratory endpoint reported in a follow-up study of the PALOMA-2 trial (NCT01740427) participants, using an EHR database.[17] In particular, we sought to estimate the conditional relative hazard as well as cumulative risk of adding or switching to a second line therapy among subjects initiating palbociclib and letrozole versus letrozole only at baseline. The PALOMA-2 trial was a landmark Phase III study that examined the efficacy of palbociclib in combination with letrozole versus letrozole and placebo for the first-line treatment of estrogen receptor positive (ER+), human epidermal growth factor receptor type 2 negative (HER2-) advanced breast cancer.[17, 18] The primary efficacy endpoint of the trial, investigator-assessed progression-free survival, was not measurable in our data source due to a lack of imaging data. Consequently, we chose to estimate treatment effects on TTNT, which is well-captured in our EHR data source and therefore a more amenable marker for evaluating agreement between the non-randomized and randomized studies.

METHODS

Data sources

This study utilized data from the US Oncology Network McKesson iKnowMed EHR database (iKM), which is derived from outpatient medical records of over a hundred community oncology practices. The iKM EHR system is not linked to a high-quality tumor registry and is the sole source of data for the iKM EHR database. The data were drawn from various fields in the health record system, which were compiled into 11 structured tables for analysis. Detailed patient-level

information on demographics, as well as time-varying information on biomarkers, diagnoses, treatments, vitals, metastasis, laboratory results, and other key confounders are included in the database. More information on the properties of the data resource used may be found in the Chapter S2 and Supplementary Table 2.1.

Study population and follow-up

Patients were selected on the basis of eligibility criteria adapted from the PALOMA-2 trial. Within the iKM database, women at least 18 years of age with metastatic breast cancer and no evidence of prior treatment for metastatic disease were included. To evaluate the first-line advanced disease setting, cohort entry was defined by the first date in which palbociclib or letrozole were ordered following an initial record of metastatic disease. Patients were excluded if they had a record of any systemic breast cancer treatment between the date of first metastasis and start of palbociclib or letrozole use. Patients with evidence of hormone receptor (HR)-negative or human epidermal growth factor receptor-II (HER-2)-positive subtypes of breast cancer were excluded, while patients with confirmed HR+, HER-2-negative disease or missing biomarker data were included. Other eligibility criteria are listed alongside the PALOMA-2 trial criteria in Supplementary Table 2.2.[18] Follow-up began on the day of cohort entry and continued until the first of the following events: (1) outcome occurrence (i.e., addition of a second line therapy or death due to any cause), (2) loss to follow-up, defined by a 90-day period following the last treatment with no evidence of treatment, a laboratory test result, or vitals recording, or (3) administrative end of data (March 28, 2021).

Estimands

Treatment ascertainment

Treatment exposures were ascertained by identifying generic names of prescription drug orders by within-network providers, which were fully captured in the iKM database. Other medications, particularly those not prescribed by within-network providers, were not captured in the database. When patients were prescribed dual therapy with palbociclib and letrozole, the orders were recorded on the same day. Therefore, patients with incident orders for palbociclib and letrozole on the same day were compared to those with incident order(s) of letrozole only following the first record of metastasis.

All-cause mortality

The date of mortality was ascertained by provider recording of patients' vital status as 'deceased' in a structured field in the health record system. The completeness of mortality data in the database has not been formally assessed, but has been estimated to capture approximately 70% of deaths that occur.

Subsequent treatment measurement

The date of the first systemic anti-cancer therapy that was not the primary treatment regimen (i.e., letrozole alone or palbociclib and letrozole) following the index date was termed the "subsequent treatment," and used to define the TTNT outcome described below.

Time-to-next treatment (outcome) measurement

TTNT was defined as a composite outcome of all-cause mortality or initiation of a subsequent systemic anti-cancer therapy. TTNT was chosen as the outcome because it is well-observed in our EHR database and appeared to serve as a meaningful surrogate for treatment effectiveness in the PALOMA-2 trial population. In particular, the hazard ratio for progression-free survival (HR: 0.56;

95% CI: 0.46 – 0.69) was proximal to the TTNT estimate (HR: 0.64; 95% CI: 0.52 – 0.78) reported in an analysis of the trial data with extended follow-up.[17] In the metastatic breast cancer setting, there are several efficacious treatment choices available following failure of a first-line therapy, which further supports TTNT as a reasonable proxy for disease progression and treatment efficacy.[19, 20] As with all clinical endpoints, TTNT has limitations. For instance, extreme cases of treatment success and treatment failure may both contribute to long periods prior to initiation of subsequent lines of therapy. In choosing TTNT as our outcome, we assume that the reasons for initiating subsequent lines of therapy in our study match those observed in the randomized trial.

Baseline patient characteristics

Patient demographics (age, geographic region), clinical characteristics (smoking status, BMI, tumor stage, diagnosis date, family history of cancer, Karnofsky/ECOG performance status, site(s) of metastasis, disease-free interval, number of metastatic sites), medication use (anticoagulant use, bone remineralization therapies, antihypertensives, antidepressants, anxiolytics, anti-hyperlipidemics, immunizations, anti-diabetics), and comorbidities (anemia, renal disease, anxiety, arthritis, cardiovascular disease, COPD, diabetes, neutropenia, osteoporosis) were collected to characterize the study cohort, adjust for confounding, and/or facilitate comparison with the PALOMA-2 trial study population. These variables were all ascertained on or prior to the date of treatment start.

Missing data

Five key confounding variables had missing values, including body mass index (BMI) (2%), tumor stage (13%), smoking status (17%), performance status (27%), and number of metastatic sites (63%). The missing values were believed to be due to changes in EHR reporting standards that occurred among practices participating in the Oncology Care Model, which could be indirectly observed in

the data through a practice identifier variable.[21, 22] Therefore, we assumed that these variables followed a *missing at random* (MAR) mechanism and, in particular, that missingness was a function of practice ID, the exposure, outcome (i.e., "next treatment"), and all confounders adjusted for in the analysis.

Statistical analysis

Multiple imputation with chained equations (MICE) was used to impute missing values since this method is suitable to address data that are MAR.[23] Furthermore, MICE is flexible in its use of conditional models, which permit imputation of ordinal, nominal, and continuous variables.[23] The functional forms of the models specified for the imputations are shown in Supplementary Table 2.3. All variables included in the outcome regression model were also included in the imputation models, in addition to predictors of missingness to reduce bias.[24, 25] Predictive mean matching was used to estimate values of body mass index (BMI), while ordered logistic and multinomial logistic models were used to estimate missing values of ordinal (i.e., stage, performance status, and number of metastatic sites) and nominal (i.e., smoking status) variables, respectively. These models were used to generate 50 imputed datasets, which were analyzed individually using the methods illustrated below. Variables were imputed in the order of their degree of missingness (from least to most). To account for the uncertainty in estimates due to missingness, all 50 point and interval estimates were pooled using Rubin's Rules.[26, 27]

For the primary analysis, a multivariable Cox proportional hazards model was used to calculate the relative hazard of initiating a second line treatment or death among patients initiating palbociclib and letrozole vs. letrozole alone conditional on measured baseline confounders. The model was adjusted for 18 confounding variables believed to be prognosticators of the outcome (Supplementary Table

2.4). All of these variables were measured on or before the date of treatment initiation. The proportional hazards assumption was checked graphically with Schoenfeld residual plots. Lastly, using the first imputed dataset, a Kaplan-Meier plot was created in the inverse probability (IP) of treatment weighted study population for qualitative comparison to the event-free survival curve produced in the PALOMA-2 trial. The distribution of IP weights in the study population was examined by treatment group to identify extreme weights which suggest positivity violations.[28]

Non-randomized study vs. randomized trial agreement

We planned to qualitatively assess the magnitude and direction of any difference between the two studies' point and interval estimates of TTNT in the context of any potential sources of bias. The standardized difference between the log hazard ratio of TTNT from our emulation study with that reported in the PALOMA-2 follow-up study was used because it provides a measure of magnitude and direction of any deviation between the two studies, facilitating interpretation of the results.[6, 16]

Sensitivity analysis I: Approach to missing data and conditional versus marginal hazard ratios

To assess the robustness of our outcome model assumptions in the primary analysis, we conducted sensitivity analyses. First, only complete cases were analyzed in the same manner as the primary analysis. Then, a Cox proportional hazards model weighted by IP weights was used to estimate the marginal hazard ratio of TTNT in the complete cases only and "imputed" study populations. Analysis of complete cases only offers a way of gaining insight regarding our assumption of the missing data mechanism. In particular, the complete case analysis is expected to differ from the imputation-based analysis under the MAR assumption but may be similar if the data follow a *missing completely at random* (MCAR) mechanism. Additionally, the marginal hazard ratio, calculated with IP weights, was hypothesized to align more with the randomized trial result since the estimate produced

by the trial investigators was not conditional on the confounders in this study and non-collapsibility of the hazard ratio.[29] Despite this, the marginal effect estimate was conducted as a sensitivity analysis because analytically deriving confidence intervals for this estimator in the context of multiple imputation was challenging.

Sensitivity analysis II: Data discontinuity

Since EHR databases typically only contain information from a particular healthcare network, patients seeking out-of-network care may have diagnoses, treatments, and outcomes not recorded in the EHR system from which the data are derived These out-of-network encounters can lead to misclassification bias.[4] One way of handling this is by employing a published prediction rule to identify patients with high data-continuity in health records and restrict the study population to these patients with higher data completeness.[5] Therefore, in an exploratory analysis, we repeated our primary analysis among patients within the 25th, 50th, and 75th percentile of predicted EHR-continuity calculated during the 365 days prior to cohort entry (Supplementary Table 2.5). The continuity calculation used in this study was developed previously using an oncology cohort derived from a linked claims-EHR database.

Sensitivity analysis III: Surveillance bias

Outcome assessment among patients in the PALOMA-2 trial occurred every 3 months after randomization. However, in the emulation study it is possible that patients were surveilled at different rates among the treatment arms. This may lead to bias by allowing more opportunity for patients in one treatment arm to experience the outcome relative to the other. We assessed the potential for surveillance bias by estimating the mean rate of imaging procedures and office visits

(proxied by vitals measurements) per patient-day during the follow-up period for each treatment group.

Sensitivity analysis IV: Misclassification bias due to missing/incomplete biomarker data

Approximately 29% of patients receiving letrozole alone and 10% of patients receiving palbociclib-letrozole in the final study cohort had missing or incomplete biomarker (i.e., HR and/or HER-2 status) data. These patients were included in the primary analysis to conserve sample size under the implicit assumption that they had HR+/HER-2- disease. However, it is possible that some or all of these patients that received letrozole alone were in fact HER-2+ since letrozole may be used among patients with this subtype, while palbociclib is typically not. Given that HER-2+ disease is associated with a poorer prognosis[30] than HER-2–, our implicit assumption may have resulted in a bias away from the null in the primary analysis, favoring the palbociclib-letrozole regimen. In light of this, a sensitivity analysis was carried out by repeating our analyses among patients with confirmed HR+/HER-2– disease (i.e., complete biomarker data).

RESULTS

Study cohort selection

Among 246,752 women 18 years or older with a breast cancer diagnosis, 1,299 palbociclib-letrozole users and 2,537 letrozole only users met all study eligibility criteria (Supplementary Table 2.2). Baseline demographic and clinical characteristics of the study cohort are shown alongside those of PALOMA-2 trial participants in Table 2.1. The trial population differed substantially from the imputed study population (i.e., taking average over 50 imputed datasets), with emulation study participants tending to be classified as having newly metastatic disease, a shorter disease-free interval, Stage IV disease at initial diagnosis, and only one site of metastasis to a much greater extent

than trial participants. Upon cohort entry, patients in the letrozole only group had a median time since initial diagnosis with breast cancer of 1.5 years (IQR: 0.15 years – 7.5 years), while palbociclib and letrozole initiators had a median of 0.8 years (IQR: 0.1 years – 8.1 years) since initial diagnosis.

Table 2.1. Patient Demographic and Clinical Characteristics of Study Cohort

	Emulation Study[a] (Complete Cases)		Emulation Study[a] (Imputed Data)		PALOMA-2 Trial	
Characteristic	Palbociclib-Letrozole (n = 1,299)	Letrozole Only (n = 2,537)	Palbociclib-Letrozole (n = 1,299)	Letrozole Only (n = 2,537)	Palbociclib-Letrozole (n = 444)	Letrozole Only (n = 222)
Age[c]						
Median (range) - yr	66 (25-85)	68 (26-85)	66 (25-85)	68 (26-85)	62 (30-89)	61 (28-88)
<65 yr - no. (%)	584 (45.0)	970 (38.2)	584 (45.0)	970 (38.2)	263 (59.2)	141 (63.5)
≥65 yr - no. (%)	715 (55.0)	1567 (61.8)	715 (55.0)	1567 (61.8)	181 (40.8)	81 (36.5)
ECOG performance status or Karnofsky equivalent - no. (%)						
0	521 (56.6)	1219 (64.4)	729 (56.1)	1612 (63.5)	257 (57.9)	102 (45.9)
1	346 (37.6)	577 (30.5)	487 (37.5)	781 (30.8)	178 (40.1)	117 (52.7)
2	54 (5.9)	98 (5.2)	83 (6.4)	144 (5.7)	9 (2.0)	3 (1.4)
Data missing	378 (29.1)[d]	643 (25.3)[d]	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)
Disease stage at initial diagnosis - no. (%)						
I	169 (13.8)	551 (26.2)	181 (13.9)	658 (25.9)	51 (11.5)	30 (13.5)
II	236 (19.3)	469 (22.3)	251 (19.3)	567 (22.3)	137 (30.9)	68 (30.6)
III	121 (9.9)	226 (10.8)	130 (10.0)	276 (10.9)	72 (16.2)	39 (17.6)
IV	695 (56.9)	855 (40.7)	738 (56.8)	1035 (40.8)	138 (31.1)	72 (32.4)
Data missing	78 (6.0)[d]	436 (17.2)[d]	0 (0.0)	0 (0.0)	46 (10.4)	13 (5.9)
Recurrence type - no. (%)						
Distant or other	843 (64.9)	1616 (63.7)	843 (64.9)	1616 (63.7)	305 (68.7)	151 (68.0)
Newly diagnosed	456 (35.1)	921 (36.3)	456 (35.1)	921 (36.3)	139 (31.3)	71 (32.0)
Disease-free interval - no. (%)[e]						
Newly-metastatic disease	1002 (77.1)	1972 (77.7)	1002 (77.1)	1972 (77.7)	167 (37.6)	81 (36.5)
≤12 mo	105 (8.1)	349 (13.8)	105 (8.1)	349 (13.8)	99 (22.3)	48 (21.6)
>12 mo	192 (14.8)	216 (8.5)	192 (14.8)	216 (8.5)	178 (40.1)	93 (41.9)
Disease site - no. (%)						
Visceral	251 (19.3)	376 (14.8)	251 (19.3)	376 (14.8)	214 (48.2)	110 (49.5)
Nonvisceral	432 (33.3)	660 (26.0)	432 (33.3)	660 (26.0)	230 (51.8)	112 (50.5)
Unknown	616 (47.4)	1501 (59.2)	616 (47.4)	1501 (59.2)	103 (23.2)	48 (21.6)

Table 2.1 (continued)

No. of disease sites - no. (%)						
1	332 (61.8)	583 (67.4)	1062 (81.8)	2192 (86.4)	138 (31.1)	66 (29.7)
2	120 (22.3)	186 (21.5)	139 (10.7)	231 (9.1)	117 (26.4)	52 (23.4)
3	42 (7.8)	54 (6.2)	55 (4.2)	71 (2.8)	112 (25.2)	61 (27.5)
≥4	43 (8.0)	42 (4.9)	43 (3.3)	43 (1.7)	77 (17.3)	43 (19.4)
Data missing	762 (58.7)[d]	1672 (65.9)[d]	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)

[a] 735 (29%) of letrozole only initiators and 129 (10%) of palbociclib-letrozole initiators in emulation study had missing or incomplete biomarker data
[b] Absolute difference in percent between PALOMA-2 Trial and average of 50 imputed datasets
[c] Ages were not available for subjects ≥85 years to preserve privacy. Calculations assume these subjects are 85 years old
[d] For missing data categories, percentages are based on total subjects in the treatment arm
[e] Defined as the time interval between last cancer treatment received prior to initial metastasis and initial metastasis in the emulation study

Primary analysis

Parameter estimates from the primary and sensitivity analyses are shown in Table 2.2.

Table 2.2. Parameter Estimates of Cox Proportional Hazards Model by Method of Data Analysis

	Hazard Ratio	95% Confidence Interval	Standardized Difference [a]
PALOMA-2 Trial Result	0.64	(0.52, 0.78)	
Following Multiple Imputation (Adjusted by Stratification)	0.62	(0.56, 0.68)	-0.05
Following Multiple Imputation (Adjusted by IP Weighting)	0.66	(0.59, 0.73)	0.05
Complete Cases Only (Adjusted by Stratification)	0.48	(0.40, 0.58)	-0.40
Complete Cases Only (Adjusted by IP Weighting)	0.51	(0.43, 0.62)	-0.31

[a] Comparing PALOMA-2 Trial Result (top row) to real-world evidence analyses (remaining rows)

The hazard ratio for TTNT estimated in our primary analysis was 0.62 (95% CI: 0.56 – 0.68), which was in agreement with the PALOMA-2 trial result of 0.64 (95% CI: 0.52 – 0.78). The crude (unadjusted) hazard ratio was closer to the null (HR: 0.71; 95% CI: 0.64 – 0.78), which was consistent with the greater presence of negative prognostic factors observed in the palbociclib-

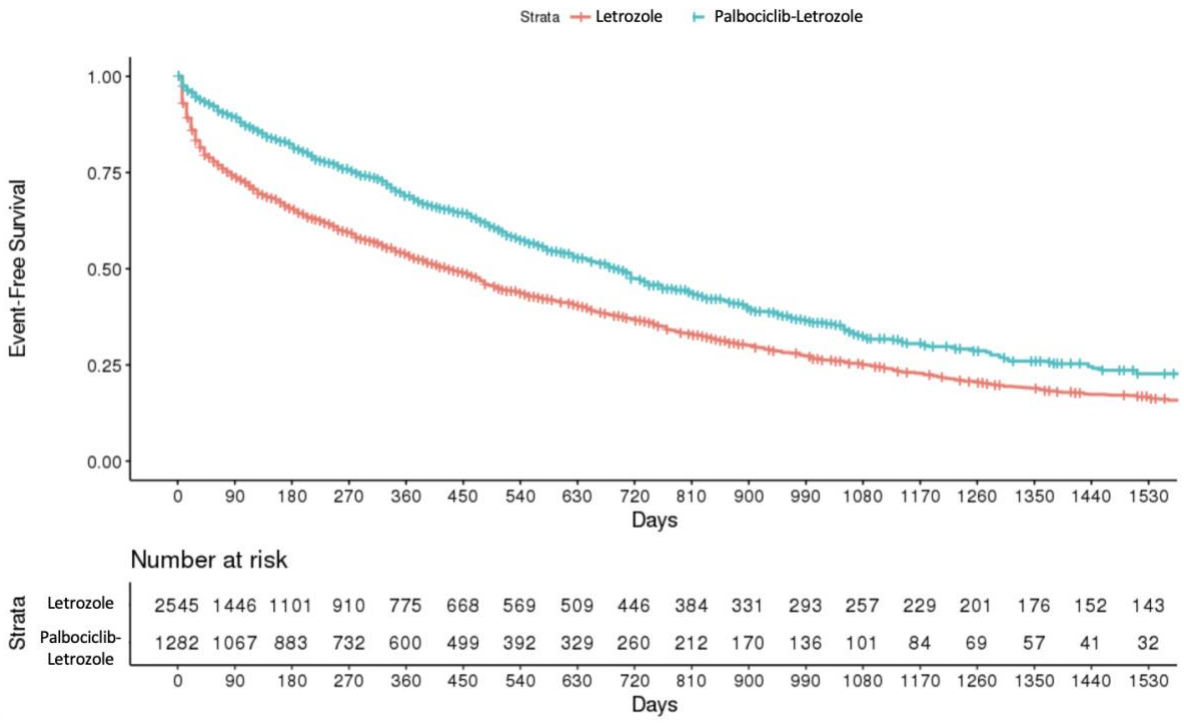
letrozole arm prior to adjustment (e.g., performance status, Stage IV diagnoses, and number of metastatic sites). Median event-free survival for TTNT in the emulation study was shorter than in the trial at 23.1 months (95% CI: 20.8 – 24.7) in the palbociclib-letrozole arm versus 14.2 months (95% CI: 12.8 – 15.9) in the letrozole only arm after adjustment using IP weights (Table 2.3**,** Figure 2.1).

Table 2.3. Median Time-To-Next Treatment in First Imputed Dataset Adjusted by IP Weights

	Palbociclib + Letrozole	Letrozole + Placebo
PALOMA-2 Trial, months	28.0 (95% CI: 23.6 - 29.6)	17.7 (95% CI: 14.3 - 21.5)
Real-World Evidence Study, months	23.1 (95% CI: 20.8 - 24.7)	14.2 (95% CI: 12.8 - 15.9)

Median times were calculated using the Kaplan-Meier estimator for event-free survival

Figure 2.1. Kaplan Meier Estimate of Event-Free Survival (IP-weighted in First Imputed Dataset)



The relationship between Kaplan-Meier event-free survival estimates in each treatment arm were similar between the non-randomized and randomized trial (Figure 2.1**,** Supplementary Figure 2.1).

Sensitivity analyses

The IP weight-based analysis in the imputed data was also similar to the primary analysis (Table 2.2). However, both, the IP weight-based and stratification-based complete case analyses were not in agreement and further from the null than the clinical trial result. When conducting the primary analysis among patients within the top 75th, 50th, and 25th percentiles of CR, effect estimates were not appreciably altered (Table 2.4).

Table 2.4. Parameter Estimates of Cox Proportional Hazards Model by Varying Levels of Restriction by Continuity Ratio

	Hazard Ratio	95% Confidence Interval	Standardized Difference[a]
PALOMA-2 Trial Result	0.64	(0.52, 0.78)	-
Following Multiple Imputation and Restriction to Top 25th Percentile CR	0.63	(0.52, 0.77)	-0.01
Following Multiple Imputation and Restriction to Top 50th Percentile CR	0.61	(0.54, 0.69)	-0.05
Following Multiple Imputation and Restriction to Top 75th Percentile CR	0.60	(0.53, 0.68)	-0.07
Following Multiple Imputation (Not Restricted by CR)	0.62	(0.56, 0.68)	-0.05

[a] Comparing PALOMA-2 Trial Result (top row) to real-world evidence analyses (remaining rows)

As displayed in Supplementary Table 2.6, there was some evidence of differential surveillance, with the mean rate of imaging procedures and office visits much greater in the letrozole only arm (0.027 imaging procedures/patient-day; 0.287 office visits/patient-day) vs. the palbociclib-letrozole arm (0.012 procedures/patient-day; 0.122 office visits/patient-day). Notably, imaging data were missing for the vast majority of patients. Lastly, in our sensitivity analysis restricting to patients with complete HR+/HER-2– biomarker data (n = 2,972), all of our effect estimates shifted slightly

further from the null relative to our main analyses and, overall, appeared to be robust to our implicit assumption regarding missing biomarker data (Supplementary Table 2.7).

DISCUSSION

In this clinical trial emulation study using EHR data from US oncology practices, we were able to successfully emulate the TTNT endpoint reported in the PALOMA-2 trial. Our results were robust to changes in analytic methods, supporting the soundness of our modelling assumptions. Our study size was over 5 times larger than the clinical trial, supporting adequate statistical power to emulate the treatment effect observed in the clinical trial and may allow for the analysis of more subgroups. This study addressed data discontinuity in an EHR-based cohort study and our results suggest that data discontinuity may be less prevalent in patients with advanced malignancy receiving active treatments. This is not surprising, as oncology care is typically integrated within a single network (e.g., US Oncology Network) and patients are less likely to seek cancer treatment across different health systems at the same time. In the analysis of complete cases only, a different result was observed compared to the primary analysis following multiple imputation. This is consistent with data that are MAR, where patients with complete data are systematically different than those with missing values.

Despite the advantages of our study, our confidence in the results of our emulation is tempered by the potential presence of differential surveillance and several assumptions that were made to account for missing values. Our analysis of imaging procedures and office visits revealed a greater than 2-fold higher rate of surveillance among letrozole only patients. Based on this, we would expect a much greater rate and frequency of outcome events in the letrozole only arm, resulting in a bias away from the null. However, approximately 7% of study patients had imaging data available and

office visits do not directly indicate surveillance for disease progression. Therefore, it is difficult to say whether surveillance bias could explain our results and more reliable markers of surveillance are needed.

In addition to assumptions concerning missing data and differential surveillance, it is a possibility that cancellation of biases, random chance, and emulation failures in eligibility criteria could explain our successful trial emulation. Patients in the clinical trial tended to be younger, have fewer patients with Stage IV diagnoses, have a less favorable performance status, a shorter disease-free interval, and more metastatic sites. Many of these differences are conflicting with respect to prognosis, and the extent that each difference may ultimately have on the outcome is unknown. Furthermore, the reasons for deciding to change a patients' treatment may be significantly different among treating physicians in the PALOMA-2 trial compared to routine practice. For instance, affordability of treatment and insurance coverage may not influence therapeutic decisions in the clinical trial, as treatments are typically provided by study sponsors. Lastly, since treatment indication is not directly observed in EHR data, it is possible that patients selected into our study were not consistent with our target population (e.g., receiving second or later lines of therapy, have HER-2+ disease, etc.). Our concerns here, however, are at least partially alleviated due to the robustness of our results to the sensitivity analysis of patients with confirmed HR+/HER-2- disease, as well as the relatively low percentage (~7%) of patients excluded for having HER-2+ disease in the original analysis (Supplementary Table 2.2).

CONCLUSIONS

Although the results of our non-randomized study coincided with that of the randomized trial, assumptions had to be made regarding comparability of the two study populations and the

mechanisms of missing data. The limitations set forth in this study illustrate the current challenges facing investigators using specialized EHR databases for comparative effectiveness research. Large-scale emulations of multiple randomized trials are needed in oncology similar to those in other fields[3,4,16] to gain predictable confidence in when and how treatment effects of oncology products can be studied with EHR databases.[31]

Author Contributions

Conception or design of the work: David Merola

Collection and assembly of data: David Merola

Data analysis and interpretation: David Merola

Drafting the article: David Merola

Critical revision of the article: Sebastian Schneeweiss, Jessica Young, Deborah Schrag, Kueiyu

Joshua Lin, Sarah Alwardt

Final approval of the version to be published: All authors

A special thank you to our collaborators at Ontada, Nicholas Robert and Sarah Alwardt, for providing the data resource used in this study and helpful feedback on our analytic approach.

Chapter III: Calibrating oncology drug effectiveness in clinical practice based on electronic health records compared against RCT evidence: The PARSIFAL trial emulation

David Merola, PharmD, MS[1,2], Jessica Young, PhD, MPH[1,3,4], Deborah Schrag, MD, MPH[5], Kueiyu Joshua Lin, MD, ScD[1,2,6], Nicholas Robert, MD,[7] Sebastian Schneeweiss, MD, ScD[1,2]

[1] Department of Epidemiology, Harvard TH Chan School of Public Health, Boston, MA, USA

[2] Division of Pharmacoepidemiology and Pharmacoeconomics, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA

[3] Department of Population Medicine, Harvard Medical School, Boston, MA, USA

[4] CAUSALab, Harvard TH Chan School of Public Health, Boston, MA, USA

[5] Department of Medicine, Memorial Sloan Kettering Cancer Center, Weill Cornell Medical School New York, NY, USA

[6] Department of Medicine, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA

[7] Ontada LLC, The Woodlands, TX, USA

ABSTRACT

Introduction: The use of electronic health records (EHR) data from clinical oncology to establish drug effectiveness in clinical practice has become of great interest to regulators, clinicians, and other healthcare stakeholders. However, the utility of EHR data in clinical effectiveness studies may be limited by missing data, unmeasured confounding, and imperfect outcome surveillance. We emulated the PARSIFAL trial using non-randomized specialty oncology EHR data to determine whether conclusions would be similar between the two studies.

Methods: We used longitudinal EHR data from outpatient oncology practices across the US to emulate the PARSIFAL trial in its treatments and selection criteria as closely as possible. Multiple imputation was employed to account for missing data in patient characteristics. Baseline characteristics were compared and hazard ratios with 95% confidence intervals for overall survival were estimated fitting a multivariable proportional hazards model. Findings were compared to the RCT result using a Wald test.

Results: Following application of all study eligibility criteria, 1,886 subjects were selected into the study cohort. Although the 3-year survival was meaningfully lower in clinical practice (59%) compared to the RCT (78%), the relative effect size was HR=1.07 (95% CI: 0.86 – 1.35), similar to the RCT (HR=1.00; 0.68 – 1.48, p-value for agreement = 0.613).

Conclusions: Despite common challenges encountered in EHR-based studies, it may be possible to achieve similar conclusions to randomized trials with the application of analytic tools and study design choices that address missing data, confounding, and selection bias. This is a promising finding in light of other emulations and ongoing efforts to improve causal inferences from existing data resources.

BACKGROUND AND INTRODUCTION

The use of real-world data (RWD) from clinical oncology[1, 2] to establish drug effectiveness has become of great interest to regulators, clinicians, and other healthcare stakeholders.[32-34] RWD have been defined by the Food and Drug Administration as "data relating to patient health status and/or the delivery of health care routinely collected from a variety of sources," including electronic health record (EHR) data.[35] As the quality and availability of RWD have increased over the past decade, the utility of these data sources in generating actionable clinical evidence on the effectiveness of medical products, i.e., "real-world evidence" or RWE, has become ever more promising.[36-39]

Databases derived from specialized oncology EHR systems[1, 2] contain rich information on patients' treatments and health outcomes, including performance status, tumor stage, and histology, that are critical for successful comparative effectiveness research of oncology medicines.[40] They draw upon several sources of clinical information, including medication and chemotherapy physician ordering systems, physician notes from outpatient oncology encounters, and molecular diagnostics/biomarkers found in EHRs. Despite advancements in EHR data quality, there remain limitations to these data that may hamper their utility in generating RWE. Among these, lack of physical treatment randomization, missing data, an inability to link records to inpatient records, high-quality tumor registries, or healthcare claims, and a lack of information captured across the health care continuum (i.e., "out-of-network" encounters), present major challenges for investigators interested in drawing causal conclusions on treatment effects.[4, 40] The extent of bias that may result in the context of these limitations and the utility of oncology EHR databases in clinical investigations is unknown.

To clarify whether RWD can be used for effectiveness research, some investigators have taken the approach of calibrating RWE against randomized clinical trials (RCT).[15, 16, 31, 41-45] In our study, we extend this framework to the oncology setting with the use of a specialized oncology EHR database. In a previous study, we successfully emulated the reduction in time-to-next treatment (TTNT) estimate (HR: 0.64; 95% CI: 0.52 – 0.78) reported in follow-up analysis of the PALOMA-2 trial (NCT01740427) study cohort using oncology EHR data.[17] The generalizable conclusions of this single trial emulation must be limited due to assumptions made regarding missing data, uncertainty in accurately identifying the first-line advanced breast cancer population, and the potential for unmeasured confounding and differential surveillance due to comparison of a common monotherapy to a dual therapy regimen.

Here, we emulated an alternative randomized clinical trial, the PARSIFAL trial (NCT02491983), which examined the efficacy of letrozole vs. fulvestrant in combination with palbociclib for the first-line treatment of advanced breast cancer and found no difference in overall survival over the 3-year study period (HR: 1.00; 95% CI: 0.68 – 1.48).[46] The PARSIFAL trial emulation complements our previous emulation of the PALOMA-2 trial, as it compares two dual therapies with similar indications and efficacy that began being used in practice around the same calendar time. This mitigates concerns regarding unmeasured confounding, surveillance bias, and differential reasons for missing data that may have been more likely in our PALOMA-2 emulation. Using EHR data form the US Oncology Network, we sought to estimate treatment effects on the relative hazard of overall survival, conditional on confounders, among fulvestrant and palbociclib vs. letrozole and palbociclib initiators, and to calibrate our results against the estimate reported in the PARSIFAL trial.

METHODS

Data source

The McKesson *iKnowMed* (iKM) EHR database is a large research database derived from outpatient oncology practices within the US Oncology Network (USON). The USON is comprised of over 400 practice sites and treats over 1,000,000 patients annually. Data in the iKM were drawn from structured fields in electronic health records, including key confounders such as performance status and tumor histology, as well as detailed treatment information. The data source only contains records from outpatient care; inpatient or medical procedure information is not available (see Chapter S2 and Supplementary Table 2.1).

Study population and follow-up

Cohort selection criteria were adapted from the PARSIFAL trial (Supplementary Table 3.1). Women at least 18 years old with a diagnosis of metastatic breast cancer and no evidence of prior treatment for metastatic disease were included. Patients with evidence of hormone receptor (HR) negative or human epidermal growth factor receptor-2 (HER-2) positive subtypes of breast cancer were excluded. In contrast to the trial, several eligibility criteria were not emulated due to incomplete capture in the database or limited relevance in RWE studies (e.g., safety criteria such as hypersensitivity to study drug or inability to swallow tablets; see details in Supplementary Table 3.1). Follow-up time was initiated on the cohort entry (index) date, which was the day all eligibility criteria were fulfilled and the treatment(s) of interest were initiated. The follow-up period proceeded until the earliest of the following events: (1) outcome occurrence (all-cause mortality); (2) loss to follow-up, defined by a >90-day period with no treatment, laboratory test result, or vitals recording after last evidence of treatment, or (3) administrative end of data (March 28, 2021).

Estimands

Treatment ascertainment

Prescription drug orders by within-network oncologists and associated prescribing dates were fully captured in the data resource and drawn from structured fields in the health record system. The primary exposure of interest was treatment initiation with palbociclib in combination with fulvestrant following initial metastasis, which was compared to initiation of palbociclib in combination with letrozole following initial metastasis. Treatment groups were operationalized in the database by identifying the occurrence of physician orders with the generic names of interest on the same day.

Outcome measurement

The primary outcome was overall survival, defined as the time from cohort entry to all-cause mortality. Mortality date was ascertained by provider recording of patients' vital status as 'deceased' in a structured field in the health record system. The completeness of mortality data in the database has not been formally assessed but has been estimated to be missing in 30% of patients.

Baseline patient characteristics

Patient demographics (age, geographic region), clinical characteristics (smoking status, BMI, tumor stage, diagnosis date, family history of cancer, Karnofsky/ECOG performance status, site(s) of metastasis, disease-free interval, number of metastatic sites), medication use (anticoagulant use, bone remineralization therapies, antihypertensives, antidepressants, anxiolytics, anti-hyperlipidemics, immunizations, anti-diabetics), and comorbidities (anemia, renal disease, anxiety, arthritis, cardiovascular disease, COPD, diabetes, neutropenia, osteoporosis) were collected to characterize

the study cohort and facilitate comparison with the PARSIFAL trial study population. These variables were all ascertained on or before the start date of the treatments of interest.

Missing data

Missing values were present in five confounding variables: body mass index (BMI) (2%), tumor stage (6%), smoking status (11%), performance status (28%), and number of metastatic sites (63%). Missingness in the data occurred, at least in part, due to changes in reporting standards among oncology practices over time. Based on this information, we assumed that missing data followed a *missing at random* (MAR) mechanism, which permits valid estimation through imputation-based procedures.[21] More specifically, we assumed that the missingness in each variable occurred as a function of a practice identifier, the outcome, treatment, and all confounding variables modeled in our primary analysis, described below.

Statistical analysis

Given our assumption of MAR and the variety of variable types (e.g., ordinal, continuous, etc.) with missingness, multiple imputation with chained equations (MICE) was used to create 50 imputed datasets.[23] Predictive mean matching, ordered logistic regression, and multinomial logistic regression were used to impute continuous, ordinal, and unordered categorical variables with missing data, respectively. All variables modeled in our outcome regression model were also placed in our imputation models, including the outcome and exposure, as well as indicators for practices associated with varying degrees of missingness. Variables were imputed in the order of their relative missingness—from least to most. The functional forms of the models specified for the imputations are shown in Supplementary Table 2.3. Point and interval estimates estimated within each of the 50 imputed datasets were pooled together using Rubin's Rules.[26, 27]

In the primary analysis, a Cox proportional hazards model[47] was used to calculate the relative hazard of all-cause mortality among patients treated with palbociclib and fulvestrant vs. palbociclib and letrozole. The model was adjusted for 18 pre-exposure risk factors for death that could be confounding variables. These variables were chosen for inclusion in the model because they were available in our data source and deemed to be prognosticators of survival (Supplementary Table 2.4).[48] Schoenfeld residual plots were used to assess the proportional hazards assumption. Lastly, using the first imputed dataset, a Kaplan-Meier plot was created in the inverse probability of treatment (IP) weighted study population for qualitative comparison to the overall survival curve reported in the PARSIFAL trial. The distribution of IP weights was evaluated by treatment group to check for the presence of any practical positivity violations, which can result in extreme weights.[28]

Non-randomized study vs. randomized trial agreement

To assess the compatibility of our study result with the PARSIFAL trial, standardized difference was calculated to compare the log hazard ratios of overall survival from both studies.[6, 16] This measure was chosen because it permits assessment of the magnitude and direction of any difference between the two studies' estimates to facilitate interpretation in the presence of any potential sources of bias.

Sensitivity analyses

We assessed the sensitivity of our results to modeling assumptions made in the primary analysis in several ways. First, the primary analysis was repeated among patients that had no missing data called *complete case analysis*.[49] This was done to indirectly evaluate our MAR assumption, since analyses of complete cases would likely differ from the multiple imputation-based analysis under the MAR assumption but may be similar if the data follow a *missing completely at random* (MCAR) mechanism.

Next, an IP-weighted Cox proportional hazards model was fit to the complete cases. IP-based estimates (marginal treatment effects) make different modeling assumptions than multivariable-adjusted models (conditional treatment effects) with respect to the relationships between the exposure, outcome, and confounders.[50] Furthermore, the marginal treatment effect may resemble the randomized trial result, which was not conditional on all of our measured confounders. Therefore, deviations in IP-based estimates and our primary analysis might indicate a sensitivity of our results to these different modeling approaches or may reflect the non-collapsibility of the hazard ratio.[29]

In addition to testing our assumptions on modeling and missingness, we conducted additional analyses to adjust for potential data discontinuity. Data discontinuity occurs when patients seek out-of-network care that may not be recorded in our data source, which exclusively contains outpatient records from the US Oncology Network. Encounters occurring outside of the iKM system may result in misclassification bias if they entail new diagnoses, treatments, or procedures.[4] To account for this, a previously validated prediction rule for discontinuity (Supplementary Table 2.5) was applied in the one-year period before cohort entry to characterize study patients in terms of their predicted EHR-continuity.[4] Then, the primary analysis will be repeated among patients in the 25th, 50th, and 75th percentile of predicted EHR-continuity.

Lastly, differential surveillance may indicate the presence of differences in the clinical care of patients between treatment groups and the presence of confounding, which may not be directly observed in the data. To explore the possibility of differential surveillance, the mean rate of imaging procedures and office visits per patient-day were calculated by treatment group as a proxy for unmeasured confounding.

RESULTS

Table 3.1. Patient and Demographic and Clinical Characteristics at Cohort Entry

	Emulation Study[a]		Imputed Data[a]		PARSIFAL Trial	
Characteristic	Palbociclib-Fulvestrant (n = 462)	Palbociclib-Letrozole (n = 1,424)	Palbociclib-Fulvestrant (n = 462)	Palbociclib-Letrozole (n = 1,424)	Palbociclib-Fulvestrant (n = 243)	Palbociclib-Letrozole (n = 243)
Age[c]						
Median (range) - yr	69 (32 - 85)	66 (25 - 85)	69 (32 - 85)	66 (25 - 85)	64 (25 - 88)	62 (35 - 90)
ECOG performance status or Karnofsky equivalent - no. (%)						
0	178 (52.8)	587 (57.2)	241 (52.2)[d]	789 (55.4)[d]	151 (62.1)	124 (51.0)
1	134 (39.8)	381 (37.1)	187 (40.5)[d]	545 (38.3)[d]	80 (32.9)	107 (44.0)
2	25 (7.4)	59 (5.7)	34 (7.4)[d]	90 (6.3)[d]	12 (4.9)	12 (4.9)
Data missing[e]	125 (27.1)	397 (27.9)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)
Recurrence type - no. (%)						
Recurrent	384 (83.1)	929 (65.2)	384 (83.1)	929 (65.2)	141 (58.0)	147 (60.5)
De Novo	78 (16.9)	495 (34.8)	78 (16.9)	495 (34.8)	102 (42.0)	96 (39.5)
Disease site - no. (%)						
Visceral	43 (9.3)	275 (19.3)	43 (9.3)	275 (19.3)	115 (47.3)	118 (48.6)
Nonvisceral	87 (18.8)	470 (33.0)	87 (18.8)	470 (33.0)	128 (52.7)	125 (51.4)
Unknown	332 (71.9)	679 (47.7)	332 (71.9)	679 (47.7)	0 (0.0)	0 (0.0)
No. of disease sites - no. (%)						
<3	91 (85.0)	500 (84.6)	394 (85.3)[d]	1230 (86.4)[d]	141 (58.0)	133 (51.4)
≥3	16 (15.0)	91 (15.4)	68 (14.7)[d]	194 (13.6)[d]	102 (42.0)	110 (48.6)
Data missing[e]	355 (76.8)	833 (58.5)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)
Previous treatment in early setting	355 (76.8)					
Chemotherapy	65 (14.1)	142 (10.0)	65 (14.1)	142 (10.0)	98 (40.3)	92 (37.9)
Tamoxifen only	30 (6.5)	125 (8.8)	30 (6.5)	125 (8.8)	48 (19.8)	59 (24.3)
Aromatase inhibitors only	129 (27.9)	167 (11.7)	129 (27.9)	167 (11.7)	26 (10.7)	21 (8.6)
Tamoxifen and aromatase inhibitors	54 (11.7)	47 (3.3)	54 (11.7)	47 (3.3)	39 (16.0)	31 (12.8)

[a] 135 (9%) of palbociclib-letrozole initiators and 57 (12%) of palbociclib-letrozole initiators in emulation study had missing or incomplete biomarker data

[b] Absolute difference in percent between PARSIFAL Trial and average of 50 imputed datasets

[c] Ages were not available for subjects ≥85 years to preserve privacy. Calculations assume these subjects are 85 years old

[d] Provided as average values over 50 imputed datasets. Totals may not add to 100% due to rounding

[e] For missing data, percentages are based on total subjects in the treatment arm

Study population

Following application of all study eligibility criteria, 1,886 patients were selected into the study cohort—462 initiators of palbociclib and fulvestrant, and 1,424 initiators of palbociclib and letrozole (Supplementary Table 3.1; Table 3.1). Relative to the PARSIFAL trial, patients in our study tended to be older, have fewer metastatic sites, a less favorable performance status among palbociclib-fulvestrant initiators, and a more favorable performance status among palbociclib-letrozole users. All these patterns were maintained following missing data imputation. Upon cohort entry, patients in the palbociclib-letrozole group had a median time since initial diagnosis with breast cancer of 1.1 years (IQR: 0.1 years – 8.1 years), while palbociclib-fulvestrant initiators had a median of 4.8 years (IQR: 1.4 years – 9.5 years) since initial diagnosis.

Comparison of overall survival in non-randomized study vs. PARSIFAL trial

In the primary analysis, the hazard ratio for overall survival was 1.07 (95% CI: 0.86 – 1.35), which was congruent with the clinical trial result (HR: 1.00; 95% CI: 0.68 – 1.48) (Table 3.2).

Table 3.2. Parameter Estimates of Cox Proportional Hazards Model by Method of Data Analysis

	Parameter Estimate	95% Confidence Interval	Standardized Difference[a]
PARSIFAL Trial Result	1.00	(0.68, 1.48)	-
Following Multiple Imputation (Adjusted by Stratification)	1.07	(0.86, 1.35)	0.04
Following Multiple Imputation (Adjusted by IP-Weighting)	1.13	(0.87, 1.48)	0.07
Complete Cases Only (Adjusted by Stratification)	1.56	(0.98, 2.47)	0.20
Complete Cases Only (Adjusted by IP-Weighting)	1.23	(0.73, 2.09)	0.09

[a] Comparing PARSIFAL Trial Result (top row) to real-world evidence analyses (remaining rows)

The crude (unadjusted) hazard ratio was 1.24 (95% CI: 1.02 – 1.51), which was aligned with our observation of more negative prognosticators in the palbociclib-fulvestrant arm with respect to performance status, disease recurrence, number of metastatic sites, and age. We observed a

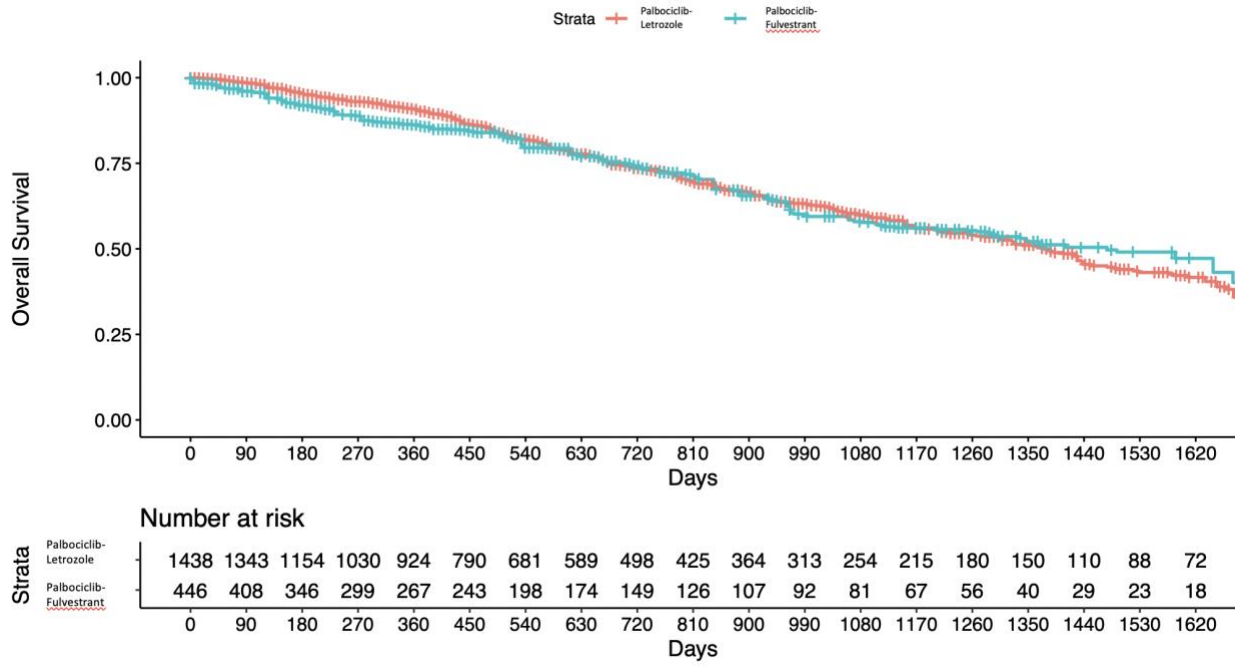
substantially higher mortality rate in the RWD study than in the randomized trial. In our study, the 3-year overall survival was 59.5% (95% CI: 55.5 – 63.7) vs. 57.7 (95% CI: 49.6 – 67.1) in the palbociclib/letrozole and palbociclib/fulvestrant groups, respectively (Table 3.3). This is compared to a 3-year overall survival of 77.1% (95% CI: 70.2 – 82.5) and 79.4 (95% CI: 73.1 – 84.4) in the palbociclib/letrozole and palbociclib/fulvestrant arms of the PARSIFAL trial, respectively (Table 3.3). Kaplan-Meier survival estimates over the study period were aligned between the randomized and non-randomized studies (Figure 3.1; Supplementary Figure 3.1).

Table 3.3. Estimates of 3-Year Overall Survival

	Palbociclib + Letrozole	Palbociclib + Fulvestrant
PARSIFAL Trial		
Number of Subjects	243	243
Number of Deaths, n (%)	51 (21.0)	51 (21.0)
Follow-up Time, median days (IQR)	960 (726 - 1,191)	
3-Year Survival Probability (95% CI)[a]	79.4 (73.1 - 84.4)	77.1 (70.2 - 82.5)
Real-World Evidence Study		
Number of Subjects	1,424	462
Number of Deaths, n (%)	372 (26.1)	136 (29.4)
Follow-up Time, median days (IQR)	511 (231-909)	507 (213-880)
3-Year Survival Probability (95% CI)[a]	59.5 (55.5 - 63.7)	57.7 (49.6 - 67.1)

[a] Kaplan-Meier estimate of survival probability at 3 years in the first imputed dataset, adjusted by inverse probability of treatment weights. Estimates from the remaining imputed datasets were very similar.

Figure 3.1. Kaplan-Meier Estimates of Overall Survival (IP-Weighted, Using First Imputed Dataset)



Sensitivity analyses

Relative to our primary analysis, our complete case analyses had point and interval estimates further from the randomized trial result (Table 3.2) and did not appear to be aligned regardless of adjustment method (i.e., IP-weighting or stratification). However, our imputation-based analysis using IP-weighting for confounding adjustment did agree with the trial result (Table 3.2). In our analysis adjusting for data continuity, point estimates grew further from the null with wider confidence intervals as higher levels of restriction by continuity ratio were imposed (Table 3.4). There was no evidence of surveillance bias in our estimates based on the observed rates of imaging procedures over the study period or office visits in each treatment arm (Supplementary Table 3.2). Notably, imaging data were missing in approximately 83% of patients in either study arm.

Table 3.4. Parameter Estimates of Cox Proportional Hazards Model by Varying Levels of Restriction by Continuity Ratio

	Hazard Ratio	95% Confidence Interval	Standardized Difference[a]
PARSIFAL Trial Result	1.00	(0.68, 1.48)	-
Following Multiple Imputation and Restriction to Top 25th Percentile CR	1.35	(0.87, 2.11)	0.14
Following Multiple Imputation and Restriction to Top 50th Percentile CR	1.16	(0.82, 1.65)	0.08
Following Multiple Imputation and Restriction to Top 75th Percentile CR	1.08	(0.85, 1.37)	0.05
Following Multiple Imputation (Not Restricted by CR)	1.07	(0.86, 1.35)	0.04

All estimates were adjusted by stratification (i.e., multivariable adjustment only)
[a] Comparing PARSIFAL Trial Result (top row) to real-world evidence analyses (remaining rows)

DISCUSSION

In this non-randomized comparative effectiveness study comparing the effect of palbociclib and fulvestrant on palbociclib and letrozole, we found very similar effect estimates and reached the same clinical conclusion as a recently completed randomized Phase 2 trial. Our study had 3.8 times more patients than the trial leading to more precise effect estimates and was representative of clinical practice in the US. An array of sensitivity analyses lend support to the validity of our statistical modeling assumptions and the mechanisms of missing data. A major strength of this study was that it examined the effectiveness of two regimens that had an equivalent evidence base for their use in the first-line advanced breast cancer setting. Consequently, provider prescribing preference is thought to be a stronger deciding force of treatment choice than patient characteristics and disease severity, resembling random treatment allocation of randomized trials. This can reduce the potential for unmeasured confounding and differential treatment of patients by design, which is supported by the non-differential rate of imaging procedures recorded in the EHR system between both treatment groups.

This study had a similar result to our previous emulation of the PALOMA-2 trial with fewer indicators of potential bias. Both trial emulations exhibited some differences between the randomized trial and EHR study populations. These differences, however, did not appear to influence the observed effect estimates, assuming the trial estimate was unbiased. One reason for this could be that baseline characteristics in the RWD are poorly captured, leading to apparent discrepancies in the randomized non-randomized study populations when in fact they were more similar. For example, previous treatments of newer patients may not be completely recorded in the US Oncology Network's health record system, resulting in a greater resemblance of each study populations than observed. Alternatively, it is possible that the patient characteristics that did differ were not strong effect measure modifiers. In fact, there was no evidence of strong effect measure modification in the PARSIFAL trial in any pre-specified subgroups.[46] This context could explain why our results were so similar to the PARSIFAL trial's results.

Our study has several limitations. First, due to the non-randomized nature of our study, unmeasured or residual confounding always remains a possibility, particularly for non-oncology related prognosticators of survival that were poorly captured in our database. We estimated that an unmeasured confounder would have to have an independent association of 1.64 or 0.61 with both the exposure and outcome to explain our null finding if the true hazard ratio for overall survival was 1.36 or 0.85, respectively; beyond the bounds of the 95% confidence interval of the primary analysis results.[51, 52] Despite this possibility, we believe that an unmeasured confounder of this magnitude is unlikely, particularly because its association with the outcome and exposure would have to be this strong independent of all other measured confounders. Second, some types of measurement error, such as non-differential exposure misclassification, can result in a bias towards the null. Although this could also explain our observed result, our prior emulation of the PALOMA-2 trial

demonstrated a consistent conclusion for a non-null effect estimate, which strengthens our confidence in this study's findings. Third, our study only investigated effectiveness measures in the first-line advanced breast cancer population, limiting generalizability of our study's results to this setting. It is possible that studies investigating different treatment settings and/or outcomes may have a greater sensitivity to confounders that were not measured in this study, for example cardiovascular morbidity, rendering our conclusions less relevant to those contexts. In particular, this may be more important in studies of earlier-stage disease, where cancer may not be the most probable cause of death for patients. Furthermore, our analytic strategy may also not be generalizable to studies that employ alternative data sources, as the quantity and mechanism of missing information may vary.

CONCLUSIONS

Common challenges of using EHR databases for comparative effectiveness research are highlighted by our study. Despite these challenges, we demonstrated that it may be possible to achieve similar conclusions to randomized trials in line with several other emulation projects when we apply analytic tools and rigorous study designs that address missing data, confounding, and selection bias. As more RCT emulation studies with oncology EHR data are becoming available we will gain confidence in RWE studies in oncology and will be able to differentiate how to conduct such studies and when they will likely be leading to valid findings.

Author Contributions

Conception or design of the work: David Merola

Collection and assembly of data: David Merola

Data analysis and interpretation: David Merola

Drafting the article: David Merola

Critical revision of the article: Sebastian Schneeweiss, Jessica Young, Deborah Schrag, Kueiyu

Joshua Lin, Nicholas Robert

Final approval of the version to be published: All authors

A special thank you to our collaborators at Ontada, Nicholas Robert and Sarah Alwardt, for providing the data resource used in this study and helpful feedback on our analytic approach.

Supplemental Material

Chapter S1

Supplementary Table 1.1. Attrition Table

Selection Criterion	N (excluded)	N (remaining)
Having at least one outpatient encounter in MGB and 365 days of Medicare Part A/B/D enrollment		348199
With at least one cancer diagnosis during CMS enrollment period	-242359	105840
Exclude subjects with different sex in EHR vs. CMS	-65	105775
Exclude subjects with different discrepant DOB in EHR vs. CMS	-51	105724
Age at least 65 years	-10075	95649
Exclude subjects with missing/incomplete data in first year	-15971	79678
Final Cohort		**79678**

Caption: Cohort attrition by application of each study eligibility criterion

Supplementary Table 1.2. Mean Standardized Difference of Select Comorbidities and Medication-Related Variables by Decile of Predicted Continuity

Decile of Predicted CR	Validation Dataset Flag (0 = Training; 1 = Validation)	N	Myocardial Infarction	Heart Failure	Peripheral Vascular Disease	Stroke	Dementia	Chronic Obstr. Pulm. Disorder	Rheumatoid Arthritis	Peptic Ulcer Disease	Liver Disease	Diabetes
1	0	8726	0.332	0.358	0.374	0.321	0.352	0.587	0.225	0.830	0.447	0.692
2	0	4362	0.328	0.348	0.387	0.323	0.304	0.547	0.201	0.789	0.413	0.653
3	0	2266	0.287	0.360	0.348	0.297	0.327	0.519	0.180	0.749	0.393	0.577
4	0	5059	0.303	0.337	0.333	0.296	0.261	0.448	0.156	0.660	0.351	0.475
5	0	733	0.259	0.267	0.322	0.286	0.314	0.412	0.158	0.751	0.372	0.500
6	0	3642	0.253	0.283	0.309	0.282	0.235	0.348	0.140	0.536	0.301	0.346
7	0	984	0.244	0.280	0.282	0.210	0.258	0.406	0.136	0.582	0.294	0.432
8	0	10339	0.257	0.276	0.293	0.229	0.213	0.337	0.113	0.512	0.259	0.296
9	0	155	0.094	0.212	0.253	0.202	0.211	0.140	0.035	0.171	0.111	0.071
10	0	3573	0.158	0.117	0.183	0.097	0.093	0.106	0.035	0.213	0.125	0.073
1	1	8931	0.335	0.374	0.377	0.318	0.360	0.570	0.220	0.846	0.444	0.702
2	1	4214	0.317	0.343	0.356	0.301	0.287	0.528	0.204	0.781	0.416	0.633
3	1	2253	0.309	0.356	0.365	0.292	0.277	0.505	0.166	0.737	0.362	0.537
4	1	4883	0.312	0.388	0.339	0.308	0.290	0.461	0.145	0.683	0.356	0.502
5	1	744	0.298	0.265	0.331	0.265	0.284	0.449	0.157	0.656	0.384	0.516
6	1	3607	0.278	0.296	0.297	0.244	0.218	0.353	0.138	0.535	0.319	0.345
7	1	968	0.266	0.249	0.293	0.257	0.193	0.355	0.123	0.561	0.295	0.507
8	1	10400	0.262	0.276	0.309	0.213	0.217	0.337	0.100	0.512	0.270	0.297
9	1	170	0.259	0.144	0.221	0.111	0.112	0.127	0.156	0.179	0.216	0.063
10	1	3669	0.160	0.114	0.195	0.123	0.089	0.126	0.033	0.213	0.139	0.066

Supplementary Table 1.2 (continued)

Decile of Predicted CR	Validation Dataset Flag	Renal Dysfunction	Cancer	Lung Cancer	Breast Cancer	Prostate Cancer	Melanoma	Stomach Cancer	Pancreatic Cancer	Colorectal Cancer	Uterine Cancer	Multiple Myeloma
1	0	0.727	1.558	0.310	0.500	0.414	0.331	0.139	0.126	0.298	0.205	0.184
2	0	0.670	0.751	0.190	0.289	0.219	0.232	0.060	0.089	0.186	0.114	0.121
3	0	0.635	0.594	0.120	0.202	0.240	0.208	0.085	0.085	0.127	0.106	0.064
4	0	0.597	0.425	0.088	0.158	0.169	0.176	0.063	0.059	0.128	0.074	0.067
5	0	0.573	0.436	0.144	0.182	0.183	0.189	0.100	0.068	0.132	0.065	0.034
6	0	0.476	0.416	0.106	0.147	0.129	0.166	0.061	0.057	0.118	0.054	0.041
7	0	0.469	0.312	0.095	0.093	0.119	0.186	0.077	0.052	0.109	0.081	0.066
8	0	0.439	0.293	0.069	0.104	0.113	0.129	0.042	0.029	0.084	0.057	0.041
9	0	0.283	0.294	0.038	0.116	0.020	0.161	0.114	0.000	0.105	0.000	0.000
10	0	0.185	0.185	0.031	0.060	0.059	0.064	0.009	0.018	0.051	0.032	0.015
1	1	0.724	1.574	0.308	0.480	0.416	0.341	0.122	0.129	0.306	0.213	0.181
2	1	0.656	0.715	0.194	0.255	0.214	0.251	0.097	0.114	0.177	0.110	0.119
3	1	0.621	0.558	0.122	0.181	0.191	0.177	0.069	0.064	0.137	0.098	0.053
4	1	0.610	0.442	0.119	0.186	0.191	0.186	0.087	0.052	0.126	0.096	0.041
5	1	0.593	0.433	0.121	0.147	0.168	0.146	0.030	0.072	0.099	0.082	0.010
6	1	0.463	0.392	0.074	0.127	0.135	0.178	0.052	0.040	0.096	0.054	0.061
7	1	0.528	0.342	0.044	0.117	0.116	0.165	0.046	0.087	0.087	0.033	0.044
8	1	0.422	0.290	0.068	0.103	0.115	0.120	0.046	0.027	0.078	0.048	0.029
9	1	0.260	0.250	0.031	0.050	0.038	0.128	0.000	0.000	0.086	0.078	0.154
10	1	0.178	0.197	0.027	0.063	0.060	0.069	0.014	0.022	0.038	0.040	0.007

Supplementary Table 1.2 (continued)

Decile of Predicted CR	Validation Dataset Flag (0 = Training; 1 = Validation)	Leukemia	Non-Hodgkin Lymphoma	Metastatic Cancer	Hypertension	Coagulation Disorder	Morbid Obesity	Anemia	Alcohol Abuse	Drug Abuse	Psychosis	Depression
1	0	0.236	0.329	0.469	2.551	0.415	0.185	0.880	0.173	0.123	0.232	0.624
2	0	0.145	0.175	0.418	1.826	0.396	0.195	0.857	0.177	0.131	0.221	0.613
3	0	0.142	0.157	0.260	1.199	0.359	0.190	0.828	0.182	0.130	0.207	0.606
4	0	0.092	0.142	0.309	0.655	0.378	0.169	0.781	0.166	0.131	0.237	0.546
5	0	0.104	0.094	0.212	0.693	0.343	0.216	0.718	0.141	0.158	0.238	0.579
6	0	0.070	0.089	0.214	0.404	0.336	0.145	0.632	0.147	0.129	0.174	0.445
7	0	0.076	0.054	0.165	0.980	0.281	0.142	0.610	0.157	0.120	0.209	0.476
8	0	0.056	0.068	0.174	0.286	0.323	0.139	0.587	0.160	0.125	0.184	0.435
9	0	0.000	0.051	0.101	0.095	0.218	0.051	0.212	0.154	0.114	0.246	0.179
10	0	0.029	0.045	0.066	0.115	0.135	0.060	0.191	0.071	0.040	0.091	0.183
1	1	0.225	0.345	0.477	2.581	0.417	0.191	0.892	0.179	0.144	0.240	0.607
2	1	0.178	0.205	0.395	1.785	0.415	0.179	0.873	0.188	0.128	0.226	0.582
3	1	0.147	0.139	0.296	1.164	0.379	0.193	0.785	0.143	0.130	0.190	0.572
4	1	0.099	0.140	0.286	0.662	0.395	0.164	0.806	0.155	0.156	0.238	0.567
5	1	0.096	0.074	0.272	0.748	0.334	0.165	0.719	0.176	0.138	0.203	0.511
6	1	0.098	0.101	0.196	0.434	0.320	0.144	0.622	0.164	0.129	0.179	0.421
7	1	0.084	0.049	0.195	0.950	0.294	0.121	0.550	0.186	0.129	0.195	0.454
8	1	0.057	0.070	0.165	0.284	0.309	0.134	0.582	0.164	0.115	0.178	0.421
9	1	0.033	0.070	0.092	0.146	0.260	0.000	0.314	0.089	0.089	0.078	0.270
10	1	0.035	0.047	0.058	0.092	0.169	0.053	0.200	0.060	0.060	0.092	0.188

Supplementary Table 1.2 (continued)

Decile of Predicted CR	Validation Dataset Flag (0 = Training; 1 = Validation)	Antiplat/ Anticoag	Anti-hypertens.	Anti-hyperlipid	Anti-arrhythmic	Insulin	Non-insulin	Benzo-diazepines	Anti-psychotics	Dementia	Anti-depressants
1	0	0.331	2.101	1.488	0.315	0.308	0.535	0.536	0.232	0.226	0.815
2	0	0.202	1.664	1.253	0.255	0.275	0.474	0.435	0.199	0.193	0.704
3	0	0.110	1.110	0.949	0.218	0.172	0.367	0.292	0.135	0.144	0.541
4	0	0.052	0.629	0.561	0.113	0.082	0.202	0.153	0.060	0.089	0.368
5	0	0.066	0.580	0.571	0.151	0.135	0.235	0.320	0.177	0.129	0.389
6	0	0.033	0.429	0.387	0.064	0.055	0.144	0.135	0.038	0.061	0.239
7	0	0.072	0.955	0.751	0.127	0.201	0.291	0.289	0.112	0.133	0.441
8	0	0.019	0.222	0.204	0.041	0.031	0.067	0.076	0.030	0.037	0.150
9	0	0.029	0.152	0.078	0.025	0.025	0.065	0.043	0.027	0.043	0.064
10	0	0.007	0.098	0.105	0.017	0.018	0.027	0.037	0.024	0.011	0.047
1	1	0.337	2.135	1.488	0.303	0.314	0.563	0.533	0.222	0.229	0.799
2	1	0.197	1.576	1.209	0.259	0.227	0.443	0.457	0.212	0.169	0.702
3	1	0.097	1.123	0.925	0.165	0.162	0.299	0.292	0.125	0.117	0.579
4	1	0.056	0.612	0.542	0.110	0.093	0.216	0.176	0.063	0.089	0.354
5	1	0.055	0.698	0.550	0.113	0.085	0.187	0.294	0.104	0.089	0.328
6	1	0.040	0.413	0.417	0.068	0.061	0.155	0.136	0.052	0.054	0.258
7	1	0.109	0.941	0.827	0.162	0.178	0.325	0.271	0.108	0.124	0.394
8	1	0.019	0.219	0.212	0.037	0.031	0.076	0.075	0.030	0.032	0.153
9	1	0.013	0.233	0.237	0.029	0.039	0.054	0.067	0.000	0.027	0.143
10	1	0.007	0.092	0.093	0.020	0.014	0.025	0.044	0.021	0.014	0.066

Supplementary Table 1.2 (continued)

Decile of Predicted CR	Validation Dataset Flag (0 = Training; 1 = Validation)	Anti-convulsants	Gastoprotective Agents	Hormone Agents	Osteoporosis Agents	Anti-biotics	Cortico-steroids	NSAIDs	COX2 Inhibitors	Opioid	Nitrates	Chemo-therapy
1	0	0.486	0.926	0.286	0.440	1.796	0.909	0.611	0.225	0.986	0.351	0.737
2	0	0.449	0.791	0.257	0.414	1.779	0.858	0.556	0.171	0.921	0.313	0.736
3	0	0.354	0.627	0.186	0.332	1.341	0.691	0.435	0.195	0.580	0.219	0.557
4	0	0.202	0.348	0.180	0.195	0.832	0.476	0.304	0.106	0.333	0.164	0.448
5	0	0.251	0.450	0.160	0.202	1.236	0.721	0.407	0.129	0.623	0.175	0.459
6	0	0.150	0.239	0.145	0.140	0.689	0.404	0.204	0.063	0.242	0.101	0.299
7	0	0.241	0.481	0.217	0.214	1.048	0.693	0.347	0.136	0.463	0.197	0.412
8	0	0.085	0.127	0.075	0.067	0.467	0.249	0.160	0.040	0.167	0.058	0.239
9	0	0.077	0.055	0.090	0.083	0.341	0.134	0.118	0.066	0.135	0.026	0.115
10	0	0.034	0.067	0.057	0.037	0.313	0.146	0.106	0.028	0.118	0.032	0.096
1	1	0.491	0.943	0.281	0.456	1.832	0.903	0.605	0.221	1.015	0.341	0.736
2	1	0.441	0.817	0.238	0.392	1.718	0.887	0.543	0.193	0.879	0.308	0.754
3	1	0.305	0.596	0.222	0.339	1.279	0.685	0.438	0.169	0.536	0.227	0.557
4	1	0.205	0.338	0.160	0.195	0.844	0.455	0.277	0.100	0.309	0.161	0.437
5	1	0.249	0.380	0.176	0.204	1.238	0.676	0.393	0.084	0.579	0.185	0.398
6	1	0.161	0.263	0.128	0.151	0.673	0.380	0.252	0.069	0.250	0.117	0.318
7	1	0.278	0.520	0.113	0.223	1.166	0.567	0.419	0.115	0.443	0.205	0.366
8	1	0.083	0.122	0.072	0.060	0.475	0.248	0.165	0.048	0.171	0.065	0.233
9	1	0.022	0.128	0.111	0.020	0.581	0.235	0.144	0.089	0.129	0.043	0.133
10	1	0.035	0.059	0.057	0.040	0.299	0.148	0.107	0.024	0.108	0.030	0.096

Supplementary Table 1.3. Sensitivity of Select Comorbidities and Medication-Related Variables by Decile of Predicted Continuity Ratio

Decile of Predicted CR	Validation Dataset Flag (0 = Training; 1 = Validation)	N	Myocardial Infarction	Heart Failure	Peripheral Vascular Disease	Stroke	Dementia	Chronic Obstr. Pulm. Disorder	Rheumatoid Arthritis	Peptic Ulcer Disease	Liver Disease	Diabetes
1	0	8726	0%	0%	0%	0%	1%	0%	0%	1%	1%	0%
2	0	4362	3%	2%	2%	6%	9%	5%	6%	6%	10%	2%
3	0	2266	9%	1%	1%	5%	8%	6%	14%	9%	8%	5%
4	0	5059	17%	10%	9%	12%	23%	15%	22%	21%	22%	20%
5	0	733	14%	4%	7%	10%	11%	15%	14%	11%	12%	9%
6	0	3642	22%	17%	10%	14%	31%	26%	27%	31%	28%	32%
7	0	984	12%	10%	13%	10%	10%	18%	25%	19%	24%	13%
8	0	10339	29%	30%	22%	29%	38%	33%	47%	36%	42%	46%
9	0	155	50%	36%	37%	33%	43%	68%	83%	72%	64%	85%
10	0	3573	49%	64%	43%	61%	66%	73%	83%	69%	65%	86%
1	1	8931	0%	0%	0%	0%	1%	0%	0%	1%	1%	0%
2	1	4214	4%	1%	2%	5%	13%	6%	6%	5%	10%	3%
3	1	2253	5%	0%	1%	5%	8%	7%	16%	11%	11%	7%
4	1	4883	14%	8%	8%	12%	19%	16%	24%	20%	20%	17%
5	1	744	5%	4%	5%	4%	10%	14%	14%	14%	12%	4%
6	1	3607	19%	17%	10%	19%	30%	24%	29%	30%	26%	32%
7	1	968	7%	8%	11%	8%	31%	21%	25%	20%	23%	13%
8	1	10400	26%	29%	19%	33%	37%	34%	52%	36%	40%	45%
9	1	170	13%	57%	40%	60%	73%	71%	33%	62%	30%	87%
10	1	3669	48%	61%	41%	52%	69%	68%	84%	69%	60%	87%

Supplementary Table 1.3 (continued)

Decile of Predicted CR	Validation Dataset Flag (0 = Training; 1 = Validation)	Renal Dysfunction	Cancer	Lung Cancer	Breast Cancer	Prostate Cancer	Melanoma	Stomach Cancer	Pancreatic Cancer	Colorectal Cancer	Uterine Cancer
1	0	1%	18%	9%	15%	19%	8%	2%	6%	9%	9%
2	0	9%	59%	45%	48%	53%	33%	44%	27%	39%	40%
3	0	10%	67%	61%	61%	50%	36%	26%	38%	52%	41%
4	0	18%	78%	71%	67%	62%	47%	49%	56%	56%	61%
5	0	14%	78%	61%	68%	62%	46%	25%	50%	57%	60%
6	0	27%	77%	64%	69%	70%	51%	38%	61%	57%	69%
7	0	24%	85%	74%	81%	76%	42%	42%	50%	64%	59%
8	0	39%	86%	81%	79%	74%	57%	63%	81%	71%	69%
9	0	48%	79%	80%	74%	95%	44%	0%	0%	57%	100%
10	0	69%	89%	87%	87%	84%	77%	88%	84%	79%	80%
1	1	1%	19%	9%	17%	18%	8%	4%	6%	9%	8%
2	1	8%	61%	41%	52%	55%	30%	21%	24%	36%	42%
3	1	9%	69%	61%	65%	56%	41%	20%	48%	50%	50%
4	1	17%	77%	62%	61%	58%	47%	31%	56%	58%	57%
5	1	12%	80%	64%	75%	68%	49%	50%	44%	64%	57%
6	1	28%	78%	73%	72%	69%	47%	48%	67%	62%	67%
7	1	18%	83%	87%	78%	75%	45%	33%	25%	69%	76%
8	1	39%	86%	81%	79%	74%	60%	60%	82%	74%	73%
9	1	54%	84%	86%	88%	89%	50%	0%	0%	67%	60%
10	1	70%	88%	89%	86%	84%	74%	85%	82%	84%	74%

Supplementary Table 1.3 (continued)

Decile of Predicted CR	Validation Dataset Flag (0 = Training; 1 = Validation)	Multiple Myeloma	Leukemia	Non-Hodgkin Lymphoma	Metastatic Cancer	Hypertension	Coagulation Disorder	Morbid Obesity	Anemia	Alcohol Abuse	Drug Abuse
1	0	6%	7%	13%	3%	1%	0%	1%	0%	0%	0%
2	0	41%	33%	43%	19%	15%	3%	2%	3%	4%	0%
3	0	66%	37%	47%	41%	34%	8%	2%	3%	0%	0%
4	0	60%	55%	50%	38%	62%	10%	13%	10%	15%	9%
5	0	75%	48%	64%	52%	55%	12%	9%	8%	17%	0%
6	0	69%	60%	62%	49%	77%	15%	17%	18%	25%	0%
7	0	63%	67%	81%	60%	42%	18%	18%	16%	0%	0%
8	0	77%	73%	75%	68%	86%	25%	27%	29%	25%	20%
9	0	0%	100%	67%	69%	95%	17%	67%	61%	25%	0%
10	0	89%	85%	82%	81%	95%	54%	66%	69%	62%	67%
1	1	8%	10%	13%	4%	1%	0%	0%	0%	1%	1%
2	1	48%	28%	40%	20%	14%	3%	1%	2%	5%	0%
3	1	72%	35%	51%	32%	36%	5%	2%	5%	12%	0%
4	1	75%	56%	52%	40%	62%	10%	17%	9%	14%	4%
5	1	92%	59%	75%	42%	52%	10%	13%	6%	7%	0%
6	1	61%	50%	59%	52%	75%	18%	24%	18%	15%	8%
7	1	77%	66%	82%	50%	42%	20%	33%	20%	13%	0%
8	1	83%	72%	74%	69%	85%	27%	29%	28%	22%	23%
9	1	0%	83%	67%	63%	91%	25%	100%	52%	50%	50%
10	1	95%	83%	81%	82%	96%	48%	67%	67%	66%	56%

Supplementary Table 1.3 (continued)

Decile of Predicted CR	Validation Dataset Flag (0 = Training; 1 = Validation)	Psychosis	Depression	Antiplatelets/Anticoagulants	Anti-hypertensives	Anti-hyperlipidemics	Anti-arrhythmics	Insulin	Non-insulin	Benzo-diazepines
1	0	0%	0%	3%	1%	0%	0%	1%	0%	1%
2	0	8%	3%	47%	12%	11%	11%	9%	9%	17%
3	0	5%	3%	74%	31%	26%	25%	33%	23%	49%
4	0	10%	13%	91%	60%	53%	61%	74%	52%	79%
5	0	4%	4%	88%	57%	50%	38%	45%	47%	51%
6	0	23%	24%	95%	72%	66%	75%	81%	62%	82%
7	0	0%	14%	83%	37%	33%	40%	27%	27%	47%
8	0	24%	31%	98%	87%	83%	86%	91%	83%	92%
9	0	22%	66%	95%	92%	93%	92%	92%	82%	93%
10	0	54%	66%	99%	95%	92%	93%	95%	94%	95%
1	1	0%	0%	3%	1%	0%	1%	0%	0%	0%
2	1	2%	3%	47%	13%	10%	12%	18%	12%	18%
3	1	6%	6%	79%	32%	28%	38%	38%	30%	48%
4	1	8%	12%	91%	61%	53%	62%	72%	51%	76%
5	1	0%	5%	89%	50%	49%	53%	58%	51%	48%
6	1	12%	26%	94%	72%	64%	73%	79%	60%	82%
7	1	17%	14%	76%	36%	32%	34%	29%	29%	49%
8	1	24%	31%	98%	87%	82%	87%	91%	81%	91%
9	1	60%	44%	98%	84%	79%	88%	89%	86%	89%
10	1	54%	67%	99%	95%	93%	92%	96%	94%	95%

Supplementary Table 1.3 (continued)

Decile of Predicted CR	Validation Dataset Flag (0 = Training; 1 = Validation)	Antipsychotics	Dementia	Anti-depressants	Anti-convulsants	Gastoprotective Agents	Hormonal Agents	Osteoporosis Agents	Anti-biotics	Cortico-steroids
1	0	0%	0%	1%	0%	1%	0%	0%	1%	0%
2	0	13%	6%	9%	10%	11%	9%	7%	5%	5%
3	0	34%	27%	25%	24%	28%	31%	20%	18%	20%
4	0	77%	49%	47%	55%	60%	36%	48%	48%	45%
5	0	29%	33%	41%	40%	44%	40%	45%	22%	19%
6	0	86%	64%	63%	65%	72%	52%	63%	54%	52%
7	0	42%	36%	34%	37%	35%	22%	46%	23%	15%
8	0	90%	78%	78%	82%	87%	73%	83%	72%	74%
9	0	90%	75%	88%	82%	92%	67%	79%	73%	80%
10	0	91%	93%	93%	92%	93%	82%	91%	78%	83%
1	1	0%	0%	0%	1%	1%	1%	0%	1%	0%
2	1	7%	9%	9%	8%	10%	10%	8%	6%	7%
3	1	36%	34%	21%	31%	30%	21%	21%	20%	20%
4	1	76%	51%	49%	55%	61%	41%	50%	48%	49%
5	1	33%	42%	45%	38%	49%	44%	42%	19%	18%
6	1	79%	63%	60%	63%	69%	56%	61%	55%	54%
7	1	37%	28%	38%	31%	33%	50%	37%	20%	23%
8	1	90%	81%	78%	83%	87%	75%	85%	71%	74%
9	1	100%	89%	76%	93%	82%	60%	94%	61%	69%
10	1	92%	92%	91%	92%	93%	80%	91%	79%	83%

Supplementary Table 1.3 (continued)

Decile of Predicted CR	Validation Dataset Flag (0 = Training; 1 = Validation)	Prescription NSAIDs	COX2 Inhibitors	Opioid	Nitrates	Chemotherapy
1	0	0%	0%	0%	0%	0%
2	0	7%	13%	9%	10%	4%
3	0	26%	10%	37%	30%	19%
4	0	52%	52%	68%	51%	32%
5	0	27%	33%	30%	36%	32%
6	0	67%	66%	75%	65%	47%
7	0	32%	22%	39%	34%	27%
8	0	76%	80%	86%	82%	67%
9	0	77%	50%	83%	91%	69%
10	0	83%	83%	89%	90%	82%
1	1	0%	0%	0%	0%	0%
2	1	8%	10%	10%	10%	6%
3	1	26%	14%	40%	26%	17%
4	1	56%	52%	71%	52%	31%
5	1	26%	40%	33%	34%	39%
6	1	58%	65%	75%	61%	45%
7	1	25%	35%	39%	29%	35%
8	1	75%	78%	85%	80%	67%
9	1	72%	50%	82%	87%	68%
10	1	83%	86%	89%	91%	81%

Supplementary Table 1.4. Medication Use between Subjects with High vs Low Predicted CR in the Validation Dataset

Variable	Top 50th Percentile Predicted CR (n = 18,814)	Bottom 50th Percentile Predicted CR (n = 21,025)	Standardized Difference
Demographics			
Age; mean (sd)	72.8 (6.8)	73.2 (6.7)	-0.06
Female Sex; n (%)	10,565 (56.2)	11,794 (56.1)	0.00
Mean Standardized Difference[a]			**0.03**
Comorbidities			
Myocardial Infarction	0.069	0.057	0.05
Heart Failure	0.079	0.068	0.04
Peripheral Vascular Disease	0.073	0.066	0.03
Stroke	0.058	0.052	0.03
Dementia	0.070	0.062	0.03
Chronic Obstr. Pulm. Disorder	0.138	0.143	-0.02
Rheumatoid Arthritis	0.031	0.023	0.05
Peptic Ulcer Disease	0.302	0.280	0.05
Liver Disease	0.113	0.098	0.05
Diabetes	0.178	0.183	-0.01
Renal Dysfunction	0.243	0.221	0.05
Cancer	0.812	0.796	0.04
Lung Cancer	0.085	0.068	0.06
Breast Cancer	0.169	0.172	-0.01
Prostate Cancer	0.142	0.142	0.00
Melanoma	0.068	0.073	-0.02
Stomach Cancer	0.009	0.008	0.00
Pancreatic Cancer	0.016	0.011	0.04
Colorectal Cancer	0.063	0.058	0.02
Uterine Cancer	0.025	0.030	-0.03
Multiple Myeloma	0.022	0.025	-0.02
Leukemia	0.035	0.036	-0.01
Non-Hodgkin Lymphoma	0.057	0.073	-0.07
Metastatic Cancer	0.155	0.124	0.09
Hypertension	0.812	0.782	0.07
Coagulation Disorder	0.091	0.085	0.02

Supplementary Table 1.4 (continued)

Variable	Top 50th Percentile Predicted CR (n = 18,814)	Bottom 50th Percentile Predicted CR (n = 21,025)	Standardized Difference
Morbid Obesity	0.022	0.019	0.02
Anemia	0.286	0.288	-0.01
Alcohol Abuse	0.025	0.017	0.06
Drug Abuse	0.013	0.011	0.02
Psychosis	0.030	0.028	0.01
Depression	0.196	0.166	0.08
Mean Standardized Difference[a,b]			**0.02**

All calculations were made using the validation dataset (n = 39,839). Variables and continuity ratio were ascertained using complete data in the first year following cohort entry.

[a] Calculated as the simple average of the absolute value of all standardized differences in each group within table

[b] Excludes "Cancer" sub-category variable and "Metastatic Cancer" variable from mean standardized difference calculation to avoid redundancy

Chapter S2

The validity of real-world evidence studies is inextricably linked to the reliability and relevance of the data resource used including its provenence.[53] Real-world data may be derived from a number of different sources with varying degrees of completeness (see **Supplementary Table 2.1**). To have utility in comparative effectiveness research (CER) of medical products, data resources must, at a minimum, contain sufficient information to ascertain treatments, outcomes, prognostic factors, and eligibility criteria, as well as be capable of establishing temporality between the treatment and outcome.[54, 55] The greater the number and completeness of the data sources used to construct a healthcare database, the more likely the database is to give rise to valid inferences on treatment effectiveness.[56]

The data resource used in this study, the iKnowMed (iKM) EHR database from the US Oncology Network, draws information from structured fields in an outpatient health record system used exclusively by oncology practices (**Supplementary Table 2.1**).[57] The database includes patient-level information on all chemotherapy and other medication orders prescribed by in-network oncologists, physician notes from encounters in the oncology practices, and molecular diagnostics. All disease state information, including tumor stage and histology, is drawn from structured fields in the health record, with a varying degree of completeness between diseases. The iKM database is limited by a lack of linkage with out-of-network records/claims and no capture of inpatient care or medical procedures (e.g., diagnostic imaging, surgery). These properties permit ascertainment of treatments and certain outcomes (e.g. time-to-next treatment) that patients are initiated on within-network (e.g., chemotherapy) with a high degree of validity. Furthermore, temporality between treatment initiation and certain outcomes can be established due to the longitudinal nature of the health records. As with any clinical database, challenges may arise in the application of certain study eligibility criteria that are based on biomarker data that might have missingness or treatment history/line-of-therapy due to the inability to observe out-of-network encounters that may have been captured before the patient began seeking care in the database network. Non-cancer related comorbidities and other confounding variables may also be challenging to capture for these reasons as well. Such comorbidities, however, can be ascertained indirectly by non-oncology medications used by the patients and recorded during medication reconciliations in unstructured outpatient notes.

Supplementary Table 2.1. Database origins and completeness

Data Source	Availability	Completeness
Tumor registry data with NAACR standards	No	N/A
Structured fields in EHR indicating cancer type	Yes	High; diagnostic billing codes indicating cancer type are available and complete for all patients
Structured fields in EHR indicating disease state information	Yes	Moderate; stage, histology, and metastasis information are reported with a moderate degree of missingness (approx. 10%-30% missingness)
Outpatient claims	No	N/A
Inpatient claims	No	N/A
Pharmacy claims	No	N/A
Chemotherapy orders	Yes	High; chemotherapy orders from within-network physicians are complete
Outpatient notes	Yes	Moderate; outpatient oncology notes are completely captured from within-network encounters only
Inpatient notes	No	N/A
Procedure notes	No	N/A
Radiology notes	No	N/A
Medication orders	Yes	Moderate; medication ordering records are available and complete from within-network physicians only
Molecular diagnostics and laboratory data	Yes	Moderate; molecular diagnostics have a moderate degree of missingness (approx. 10%-30% missingness)
Capture of out-of-network encounters	No	N/A

Abbreviations: NAACR = North American Association of Central Cancer Registries; EHR = Electronic Health Record

Supplementary Table 2.2. Eligibility Criteria of PALOMA-2 Trial and Real-World Evidence Study

PALOMA-2 Trial Eligibility Criteria	RWE Eligibility Criteria			Notes
Inclusion Criteria	Inclusion Criteria	n	Excluded	
Age at least 18 years with Diagnosis of Breast Cancer	Age at least 18 years with Diagnosis of Breast Cancer	246,752		
Incident User of Palbociclib or Letrozole	Incident Use of Palbociclib or Letrozole following Date of Initial Metastasis	15,638	231,114	
Locoregionally Recurrent or Metastatic Disease	History of Metastatic Disease (Prior to Index Date)	N/A		Subjects without metastatic disease have already been excluded at this point
No Prior Systemic Anti-Cancer Therapy for Advanced ER+ Disease	No Systemic Anti-Cancer Therapy* Following Initial Record Indicating Metastasic Disease and Prior to Index Date	11,238	4,400	Assuming that any therapies for advanced disease occurred following metastasis was recorded in the data
Confirmed ER+ Histology	No Records Indicating ER- Histology Prior to or on Index Date	11,022	216	Subjects only excluded if evidence for ER- (subjects were included if missing or if ER+)
Adequate Organ and Marrow Function		N/A		This is a safety criterion for the trial and less relevant to real-world treatment setting
Measurable Disease as per RECIST v1.1 or Bone-Only Disease		N/A		We are assuming that all subjects in RWD have measurable disease if they are receiving treatment
Postmenopausal Status		N/A		Not captured in RWD, but likely to be satisfied in RWD since this is consistent with the indication for the exposures of interest
Exclusion Criteria	Exclusion Criteria			
Confirmed Diagnosis with HER-2 Positive Disease	Confirmed Diagnosis with HER-2 Positive Disease	10,285	737	
Known Uncontrolled or Symptomatic CNS Metastases	Record of CNS Metastases Prior to Index Date	10,208	77	
Prior (neo)adjuvant treatment with letrozole or anastrozole with DFI <=12 Months from Completion of Treatment	Treatment with letrozole or anastrozole <=12 Months from index date	6,613	3,595	
Prior Treatment with any CDK 4/6 Inhibitor	Prior Treatment with ribociclib, palbociclib, abemaciclib	6,238	375	

Supplementary Table 2.2 (continued)

PALOMA-2 Trial Eligibility Criteria	RWE Eligibility Criteria			Notes
ECOG >2	ECOG >2 or Karnofsky <50	6,193	45	Subjects only excluded if evidence of ECOG > 2 (or equivalent Karnofsky <50, if ECOG unavailable). Subjects with missing performance status were included.
	Received Palbociclib without Letrozole	4,389	1,804	
	Male Gender	4,360	29	
	Record of Systemic Anti-Cancer Therapy* Other Than Exposure of Interest on Index Date	3,836	524	

*Includes the following: tamoxifen, exemestane, goserelin, toremifene, doxorubicin, paclitaxel, capecitabine, gemcitabine, vinorelbine, eribulin, carboplatin, cisplatin, cyclophosphamide, docetaxel, epirubicin, ixabepilone, sacituzumab govitecan-hziy, cyclophosphamide, methotrexate, fluorouracil, olaparabib talazoparib, alpelisib, atezolizumab, larotrectinib, entrectinib, pembrolizumab (for incident letrozole users only, also includes use of abemaciclib and ribociclib on the index date)

Note: To facilitate analysis of RWD, ECOG Criteria was operationalized as an exclusion criteria of subjects with ECOG >2

Supplementary Table 2.3. Regression Models Specified to Estimate Imputed Values

Variable with Missingness	Imputation Method Used	Model	Additional Information
Body Mass Index	Predictive Mean Matching	$Body\ mass\ index = \beta_0 + \beta_1(event\ time) + \beta_2 I(event = 1) + \beta_3 I(treatment = 1) + \beta_4 I(age = 45-54) + \beta_5 I(age = 55-64) + \beta_6 I(age = 65-74) + \beta_7 I(age = 75-84) + \beta_8 I(age = 85+) + \beta_9 I(family\ history\ cancer = 1) + \beta_{10}(diagnosis\ year) + \beta_{11} I(visceral\ disease = 1) + \beta_{12} I(nonviscral\ disease = 1) + \beta_{13}(disease\ free\ interval) + \beta_{14} I(osteoporosis\ medications = 1) + \beta_{15} I(antihypertensive\ medications = 1) + \beta_{16} I(anemia\ hx = 1) + \beta_{17} I(cardiovascular\ disease\ hx = 1) + \beta_{18} I(osteporosis\ hx = 1) + \beta_{19} I(region = northeast) + \beta_{20} I(region = south) + \beta_{21} I(region = west) + \beta_{22} I(newly\ diagnosed = 1) + \beta_{23} I(stage = 2) + \beta_{24} I(stage = 3) + \beta_{25} I(stage = 4) + \beta_{26} I(smoking = former) + \beta_{27} I(smoking = never) + \beta_{28} I(smoking\ other = 1) + \beta_{29} I(performance\ status = 1) + \beta_{30} I(performance\ status = 2) + \beta_{31}(number\ metastatic\ sites) + \beta_{32} I(practice\ id\ bmi = 1) + \beta_{33} I(practice\ id\ stage = 1) + \beta_{34} I(practice\ id\ stage = 2) + \beta_{35} I(practice\ id\ stage = 3) + \beta_{36} I(practice\ id\ stage = 4) + \beta_{37} I(practice\ id\ smoking = 1) + \beta_{38} I(practice\ id\ smoking = 2) + \beta_{39} I(practice\ id\ smoking = 3) + \beta_{40} I(practice\ id\ performance\ status = 1) + \beta_{41} I(practice\ id\ performance\ status = 2) + \beta_{42} I(practice\ id\ performance\ status = 3) + \beta_{43} I(practice\ id\ performance\ status = 4) + \beta_{44} I(practice\ id\ metastatic\ sites = 1) + \beta_{45} I(practice\ id\ metastatic\ sites = 2)$	Each case with a missing BMI value was matched with 5 non-missing values with the closest predicted BMIs.
Tumor Stage	Ordered Logistic Regression	$logit[Pr(stage \le j)] = \beta_0 \pm (\beta_1(event\ time) + \beta_2 I(event = 1) + \beta_3 I(treatment = 1) + \beta_4 I(age = 45-54) + \beta_5 I(age = 55-64) + \beta_6 I(age = 65-74) + \beta_7 I(age = 75-84) + \beta_8 I(age = 85+) + \beta_9 I(family\ history\ cancer = 1) + \beta_{10}(diagnosis\ year) + \beta_{11} I(visceral\ disease = 1) + \beta_{12} I(nonviscral\ disease = 1) + \beta_{13}(disease\ free\ interval) + \beta_{14} I(osteoporosis\ medications = 1) + \beta_{15} I(antihypertensive\ medications = 1) + \beta_{16} I(anemia\ hx = 1) + \beta_{17} I(cardiovascular\ disease\ hx = 1) + \beta_{18} I(osteporosis\ hx = 1) + \beta_{19} I(region = northeast) + \beta_{20} I(region = south) + \beta_{21} I(region = west) + \beta_{22} I(newly\ diagnosed = 1) + \beta_{23}(BMI) + \beta_{24} I(smoking = former) + \beta_{25} I(smoking = never) + \beta_{26} I(smoking\ other = 1) + \beta_{27} I(performance\ status = 1) + \beta_{28} I(performance\ status = 2) + \beta_{29}(number\ metastatic\ sites) + \beta_{30} I(practice\ id\ bmi = 1) + \beta_{31} I(practice\ id\ stage = 1) + \beta_{32} I(practice\ id\ stage = 2) + \beta_{33} I(practice\ id\ stage = 3) + \beta_{34} I(practice\ id\ stage = 4) + \beta_{35} I(practice\ id\ smoking = 1) + \beta_{36} I(practice\ id\ smoking = 2) + \beta_{37} I(practice\ id\ smoking = 3) + \beta_{38} I(practice\ id\ performance\ status = 1) + \beta_{39} I(practice\ id\ performance\ status = 2) + \beta_{40} I(practice\ id\ performance\ status = 3) + \beta_{41} I(practice\ id\ performance\ status = 4) + \beta_{42} I(practice\ id\ metastatic\ sites = 1) + \beta_{43} I(practice\ id\ metastatic\ sites = 2))$	Where j=1 when stage 1, j=2 when stage 2, j=3 when stage 3, j=4 when stage 4

Supplementary Table 2.3 (continued)

Variable with Missingness	Imputation Method Used	Model	Additional Information
Performance Status	Ordered Logistic Regression	$logit[Pr(performance\ status \leq j)] = \beta_0 \pm (\beta_1(event\ time) + \beta_2 I(event = 1) + \beta_3 I(treatment = 1) + \beta_4 I(age = 45-54) + \beta_5 I(age = 55-64) + \beta_6 I(age = 65-74) + \beta_7 I(age = 75-84) + \beta_8 I(age = 85+) + \beta_9 I(family\ history\ cancer = 1) + \beta_{10}(diagnosis\ year) + \beta_{11} I(visceral\ disease = 1) + \beta_{12} I(nonviscral\ disease = 1) + \beta_{13}(disease\ free\ interval) + \beta_{14} I(osteoporosis\ medications = 1) + \beta_{15} I(antihypertensive\ medications = 1) + \beta_{16} I(anemia\ hx = 1) + \beta_{17} I(cardiovascular\ disease\ hx = 1) + \beta_{18} I(osteporosis\ hx = 1) + \beta_{19} I(region = northeast) + \beta_{20} I(region = south) + \beta_{21} I(region = west) + \beta_{22} I(newly\ diagnosed = 1) + \beta_{23}(BMI) + \beta_{24} I(smoking = former) + \beta_{25} I(smoking = never) + \beta_{26} I(smoking\ other = 1) + \beta_{27} I(stage = 2) + \beta_{28} I(stage = 3) + \beta_{29} I(stage = 4) + \beta_{30}(number\ metastatic\ sites) + \beta_{31} I(practice\ id\ bmi = 1) + \beta_{32} I(practice\ id\ stage = 1) + \beta_{33} I(practice\ id\ stage = 2) + \beta_{34} I(practice\ id\ stage = 3) + \beta_{35} I(practice\ id\ stage = 4) + \beta_{36} I(practice\ id\ smoking = 1) + \beta_{37} I(practice\ id\ smoking = 2) + \beta_{38} I(practice\ id\ smoking = 3) + \beta_{39} I(practice\ id\ performance\ status = 1) + \beta_{40} I(practice\ id\ performance\ status = 2) + \beta_{41} I(practice\ id\ performance\ status = 3) + \beta_{42} I(practice\ id\ performance\ status = 4) + \beta_{43} I(practice\ id\ metastatic\ sites = 1) + \beta_{44} I(practice\ id\ metastatic\ sites = 2))$	Where j=0 when performance status is 0, j=1 when performance status is 1 j=2 when performance status is 2
Number of Metastatic Sites	Ordered Logistic Regression	$logit[Pr(number\ metastatic\ sites \leq j)] = \beta_0 \pm (\beta_1(event\ time) + \beta_2 I(event = 1) + \beta_3 I(treatment = 1) + \beta_4 I(age = 45-54) + \beta_5 I(age = 55-64) + \beta_6 I(age = 65-74) + \beta_7 I(age = 75-84) + \beta_8 I(age = 85+) + \beta_9 I(family\ history\ cancer = 1) + \beta_{10}(diagnosis\ year) + \beta_{11} I(visceral\ disease = 1) + \beta_{12} I(nonviscral\ disease = 1) + \beta_{13}(disease\ free\ interval) + \beta_{14} I(osteoporosis\ medications = 1) + \beta_{15} I(antihypertensive\ medications = 1) + \beta_{16} I(anemia\ hx = 1) + \beta_{17} I(cardiovascular\ disease\ hx = 1) + \beta_{18} I(osteporosis\ hx = 1) + \beta_{19} I(region = northeast) + \beta_{20} I(region = south) + \beta_{21} I(region = west) + \beta_{22} I(newly\ diagnosed = 1) + \beta_{23}(BMI) + \beta_{24} I(smoking = former) + \beta_{25} I(smoking = never) + \beta_{26} I(smoking\ other = 1) + \beta_{27} I(stage = 2) + \beta_{28} I(stage = 3) + \beta_{29} I(stage = 4) + \beta_{30} I(performance\ status = 1) + \beta_{31} I(performance\ status = 2) + \beta_{32} I(practice\ id\ bmi = 1) + \beta_{33} I(practice\ id\ stage = 1) + \beta_{34} I(practice\ id\ stage = 2) + \beta_{35} I(practice\ id\ stage = 3) + \beta_{36} I(practice\ id\ stage = 4) + \beta_{37} I(practice\ id\ smoking = 1) + \beta_{38} I(practice\ id\ smoking = 2) + \beta_{39} I(practice\ id\ smoking = 3) + \beta_{40} I(practice\ id\ performance\ status = 1) + \beta_{41} I(practice\ id\ performance\ status = 2) + \beta_{42} I(practice\ id\ performance\ status = 3) + \beta_{43} I(practice\ id\ performance\ status = 4) + \beta_{44} I(practice\ id\ metastatic\ sites = 1) + \beta_{45} I(practice\ id\ metastatic\ sites = 2))$	Where j=1 when number of metastatic sites is 1, j=2 when number of metastatic sites is 2, j=3 when number of metastatic sites is 3, j=4 when number of metastatic sites is 4 or greater

Supplementary Table 2.3 (continued)

Variable with Missingness	Imputation Method Used	Model	Additional Information
Smoking Status	Multinomial Logistic Regression	$log\left[\frac{Pr(smoking=j)}{Pr(smoking=current)}\right] = \beta_0 \pm (\beta_1(event\ time) + \beta_2 I(event = 1) + \beta_3 I(treatment = 1) + \beta_4 I(age = 45-54) + \beta_5 I(age = 55-64) + \beta_6 I(age = 65-74) + \beta_7 I(age = 75-84) + \beta_8 I(age = 85+) + \beta_9 I(family\ history\ cancer = 1) + \beta_{10}(diagnosis\ year) + \beta_{11} I(visceral\ disease = 1) + \beta_{12} I(nonviscral\ disease = 1) + \beta_{13}(disease\ free\ interval) + \beta_{14} I(osteoporosis\ medications = 1) + \beta_{15} I(antihypertensive\ medications = 1) + \beta_{16} I(anemia\ hx = 1) + \beta_{17} I(cardiovascular\ disease\ hx = 1) + \beta_{18} I(osteporosis\ hx = 1) + \beta_{19} I(region = northeast) + \beta_{20} I(region = south) + \beta_{21} I(region = west) + \beta_{22} I(newly\ diagnosed = 1) + \beta_{23}(BMI) + \beta_{24}(number\ metastatic\ sites) + \beta_{25} I(stage = 2) + \beta_{26} I(stage = 3) + \beta_{27} I(stage = 4) + \beta_{28} I(performance\ status = 1) + \beta_{29} I(performance\ status = 2) + \beta_{30} I(practice\ id\ bmi = 1) + \beta_{31} I(practice\ id\ stage = 1) + \beta_{32} I(practice\ id\ stage = 2) + \beta_{33} I(practice\ id\ stage = 3) + \beta_{34} I(practice\ id\ stage = 4) + \beta_{35} I(practice\ id\ smoking = 1) + \beta_{36} I(practice\ id\ smoking = 2) + \beta_{37} I(practice\ id\ smoking = 3) + \beta_{38} I(practice\ id\ performance\ status = 1) + \beta_{39} I(practice\ id\ performance\ status = 2) + \beta_{40} I(practice\ id\ performance\ status = 3) + \beta_{41} I(practice\ id\ performance\ status = 4) + \beta_{42} I(practice\ id\ metastatic\ sites = 1) + \beta_{43} I(practice\ id\ metastatic\ sites = 2))$	Where j=1 when smoking status is former, j=2 when smoking status is never, j=3 when smoking status is other

Supplementary Table 2.4. Key Confounders Adjusted for in the Clinical Trial Emulation Study

Variable Name	Definition	Operationalization
Age Category	Age at index date, grouped into 6 categories	Categorical: 25-44, 45-54, 55-64, 65-74, 75-84, and 85+
Family History	Family history of cancer	Binary: Yes/No
Performance Status	ECOG performance status or equivalent Karnofsky Score at index date	Categorical: 0 (ECOG 0 or Karnofsky 100), 1 (ECOG 1 or Karnofsky 80-90)
Tumor Stage	Tumor Stage	Categorical: I, II, III, IV
Year of Diagnosis	Number of years between cohort entry and initial breast cancer diagnosis	Ordinal: 0-35 years (1-year increments)
Visceral Disease	Cancer spread to visceral tissue (i.e., adrenal gland, bronchus, cervical, cervix, digestive, fallopian, gastrointestinal, genital, intestine, leptomeninges, liver, lung, mediastinum, omentum, orbit, respiratory, urinary, ovary, pancreas, peritoneum, pleura, rectum, retroperitoneum, uterus	Binary: Yes/No
Non-Visceral Disease	Cancer spread to non-visceral tissue (i.e., bone, chest, lymph nodes, spleen, skin)	Binary: Yes/No
Newly Diagnosed	If initial diagnosis was within 7 days of first metastasis record	Binary: Yes/No
Disease-Free Interval	Time between last treatment prior to first record of metastasis and date of initial metastasis; if no treatment present prior to metastasis then classified as "newly metastatic"	Categorical: ≤12 months, >12 months, newly metastatic
Number of Metastatic Sites	Number of distinct tissues in which disease has spread	Categorical: 1, 2, 3, 4+
Osteoporosis Medications	Physician ordering of osteoporosis medicines in year prior to index date	Binary: Yes/No
Antihypertensive Medications	Physician ordering of antihypertensive medicines in year prior to index date	Binary: Yes/No
Anemia History	History of anemia in year prior to index date	Binary: Yes/No
Cardiovascular Disease History	History of cardiovascular disease in year prior to index date	Binary: Yes/No
Osteoporosis History	History of osteoporosis in year prior to index date	Binary: Yes/No
Smoking History	Most recent smoking history on or before index date	Categorical: Current, Former, Never, Other
Body Mass Index	Most recent body mass index recorded on or before index date	Continuous
Region	Geographic region in the United States	Categorical: Midwest, South, Northeast, West

Supplementary Table 2.5. Continuity Ratio (CR) Calculation

Factor	Contribution to CR if Factor Present
(Intercept)	0.02
BMI recorded	0.07
At least 1 diagnosis	0.01
At least 2 diagnoses	0.05
At least 1 OP visit	0.05
At least 2 OP visits	0.05
At least 1 med	0.05
At least 2 meds	0.13
At least 1 basic fact	0.14
At least 2 basic facts	0.21
Influenza or Pneumococcal vaccine	0.16

CR = Sum of All Factors Present During the 1-Year Baseline Period

Supplementary Table 2.6. Mean Rate of Imaging Procedures and Office Visits During Follow-Up by Treatment Arm (Surveillance Bias Assessment)

Radiologic Imaging Procedures			
Treatment Group	N	%N Missing[a]	Mean Number of Imaging Procedures per Patient-Day
Letrozole Only	120	95%	0.027
Palbociclib and Letrozole	157	88%	0.012
Office Visits			
Treatment Group	N	%N Missing	Mean Encounters[b] per Patient-Day
Letrozole Only	2411	5%	0.287
Palbociclib and Letrozole	1292	1%	0.122

[a] Imaging data was missing and/or incomplete for the vast majority of patients

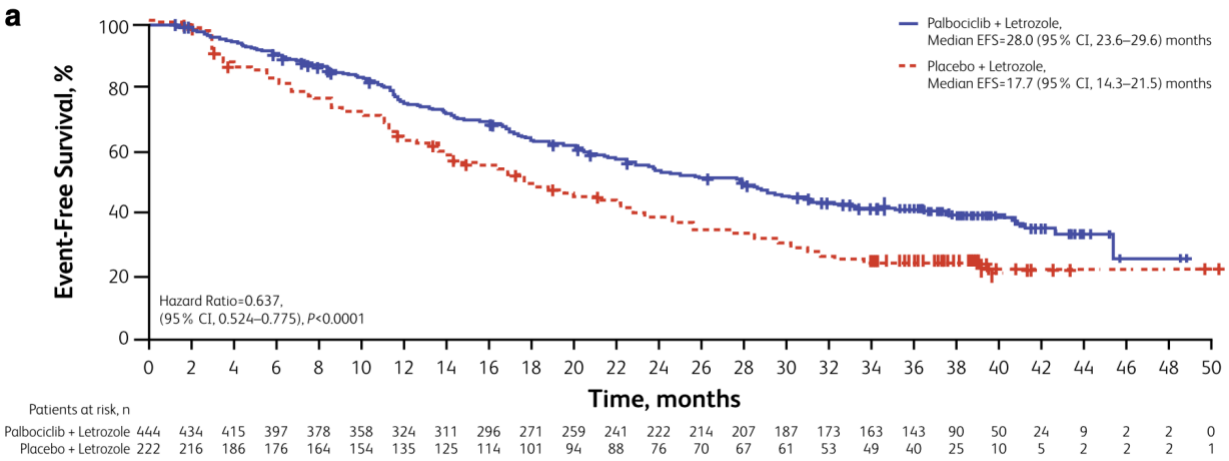
[b] Proxied by vitals (blood pressure/pulse) recordings

Supplementary Table 2.7. Parameter Estimates of Cox Proportional Hazards Model by Method of Data Analysis Among Subjects with Complete Biomarker Data (Sensitivity Analysis)

	Hazard Ratio	95% Confidence Interval	Standardized Difference[a]
PALOMA-2 Trial Result	0.64	(0.52, 0.78)	-
Following Multiple Imputation (Adjusted by Stratification)	0.55	(0.50, 0.62)	-0.18
Following Multiple Imputation (Adjusted by IP-Weighting)	0.59	(0.53, 0.67)	-0.09
Complete Cases Only (Adjusted by Stratification)	0.46	(0.38, 0.56)	-0.33
Complete Cases Only (Adjusted by IP-Weighting)	0.48	(0.40, 0.59)	-0.28

[a] Comparing PALOMA-2 Trial Result (top row) to real-world evidence analyses (remaining rows)

Supplementary Figure 2.1. Kaplan Meier Estimates of Event-Free Survival (PALOMA-2)



Supplementary Table 3.1. Eligibility Criteria of PARSIFAL Trial and Real-World Evidence Study

PARSIFAL Trial Eligibility Criteria	RWE Eligibility Criteria	n	Excluded	Notes
Inclusion Criteria	Inclusion Criteria			
Age at least 18 years with Locally Advanced Breast Cancer Not Amenable to Curative Therapy	Age at least 18 years with Diagnosis of Breast Cancer	246,752		
	Incident Use of Palbociclib Following Date of Initial Metastasis	9,182	237,570	
No Prior Chemotherapy in the Metastatic Setting	No Systemic Anti-Cancer Therapy* Following Initial Record Indicating Metastasic Disease and Prior to Index Date	6,383	2,799	Assuming that any therapies for advanced disease occurred following metastasis was recorded in the data
Confirmed ER+ Histology	No Records Indicating ER- Histology Prior to or on Index Date	6,290	93	Subjects only excluded if evidence for ER- (subjects were included if missing or if ER+)
Adequate Organ and Marrow Function, Resolution of All Toxic Effects of Prior Therapy or Surgical Procedures		N/A		This is a safety criterion for the trial and less relevant to real-world treatment setting
Measurable Disease as per RECIST v1.1 or Non-Measurable Disease		N/A		We are assuming that all subjects in RWD have measurable disease if they are receiving treatment
Postmenopausal Status		N/A		Not captured in RWD, but likely to be satisfied in RWD since this is consistent with the indication for the exposures of interest
Resolution of all acute toxic effects of prior anti-cancer therapy or surgical procedures to NCI-CTCAE version 4.0 Grade equal or minor than 1		N/A		This is a safety criterion for the trial and less relevant to real-world treatment setting
Exclusion Criteria	Exclusion Criteria			
Confirmed HER-2 Positive Disease	Confirmed Diagnosis with HER-2 Positive Disease	6,086	204	
Known Uncontrolled or Symptomatic CNS Metastases	Record of CNS Metastases Prior to Index Date	6,046	40	
Prior (neo)adjuvant endocrine therapy with DFI <=12 Months from Completion of Treatment	Treatment with letrozole or anastrozole <=12 Months from index date	3,647	2,399	
	Prior Treatment with any CDK 4/6 inhibitor (ribociclib, palbociclib, abemaciclib)	3,325	322	

Supplementary Table 3.1 (continued)

PARSIFAL Trial Eligibility Criteria	RWE Eligibility Criteria			Notes
ECOG >2	ECOG >2 or Karnofsky <50	3,295	30	Subjects only excluded if evidence of ECOG > 2 (or equivalent Karnofsky <50, if ECOG unavailable). Subjects with missing performance status were included. Eligibility criteria in trial excluded subjects with ECOG>1, however, there were subjects included with an ECOG>1 per PARSIFAL manuscript
	Received Palbociclib without Letrozole or Fulvestrant	1,908	1,387	
	Male Gender	1,886	22	
Major surgery within 4 weeks of start of study drug		N/A		This is a safety criterion for the trial and less relevant to real-world treatment setting
Patients with an active, bleeding diathesis		N/A		This is a safety criterion for the trial and less relevant to real-world treatment setting
Serious concomitant systemic disorder incompatible with the study		N/A		This is a safety criterion for the trial and less relevant to real-world treatment setting
Known hypersensitivity to letrozole, fulvestrant or any of their excipients, or to any palbociclib excipients		N/A		This is a safety criterion for the trial and less relevant to real-world treatment setting
Are unable to swallow tablets		N/A		This is a safety criterion for the trial and less relevant to real-world treatment setting
QTc > 480 msec on basal assessments, personal history of long or short QT syndrome, Brugada syndrome or known history of QTc prolongation, or Torsade de Pointes		N/A		This is a safety criterion for the trial and less relevant to real-world treatment setting
Uncontrolled electrolyte disorders that can compound the effects of a QTc-prolonging drug		N/A		This is a safety criterion for the trial and less relevant to real-world treatment setting
Patients with rapidly progressive visceral disease or visceral crisis		N/A		This is a safety criterion for the trial and less relevant to real-world treatment setting
Locally advanced breast cancer candidate for a radical treatment		N/A		This is a safety criterion for the trial and less relevant to real-world treatment setting
Chronic daily treatment with corticosteroids with a dose of at least 10 mg/day methylprednisolone equivalent		N/A		This is a safety criterion for the trial and less relevant to real-world treatment setting; Dosing is also not available for non-onco treatments

*Includes the following: tamoxifen, exemestane, goserelin, toremifene, doxorubicin, paclitaxel, capecitabine, gemcitabine, vinorelbine, eribulin, carboplatin, cisplatin, cyclophosphamide, docetaxel, epirubicin, ixabepilone, sacituzumab govitecan-hziy, cyclophosphamide, methotrexate, fluorouracil, olaparabib talazoparib, alpelisib, atezolizumab, larotrectinib, entrectinib, pembrolizumab (for incident letrozole users only, also includes use of abemaciclib and ribociclib on the index date)
Note: To facilitate analysis of RWD, ECOG Criteria was operationalized as an exclusion criteria of subjects with ECOG >2 or Karnofsky <50

Supplementary Table 3.2. Differential Surveillance Assessment During Follow-Up

Imaging Procedures

Treatment Group	N	%N Missing[a]	Mean Radiologic Imaging Procedures per Patient-Day
Palbociclib and Fulvestrant	88	81%	0.01
Palbociclib and Letrozole	231	84%	0.01

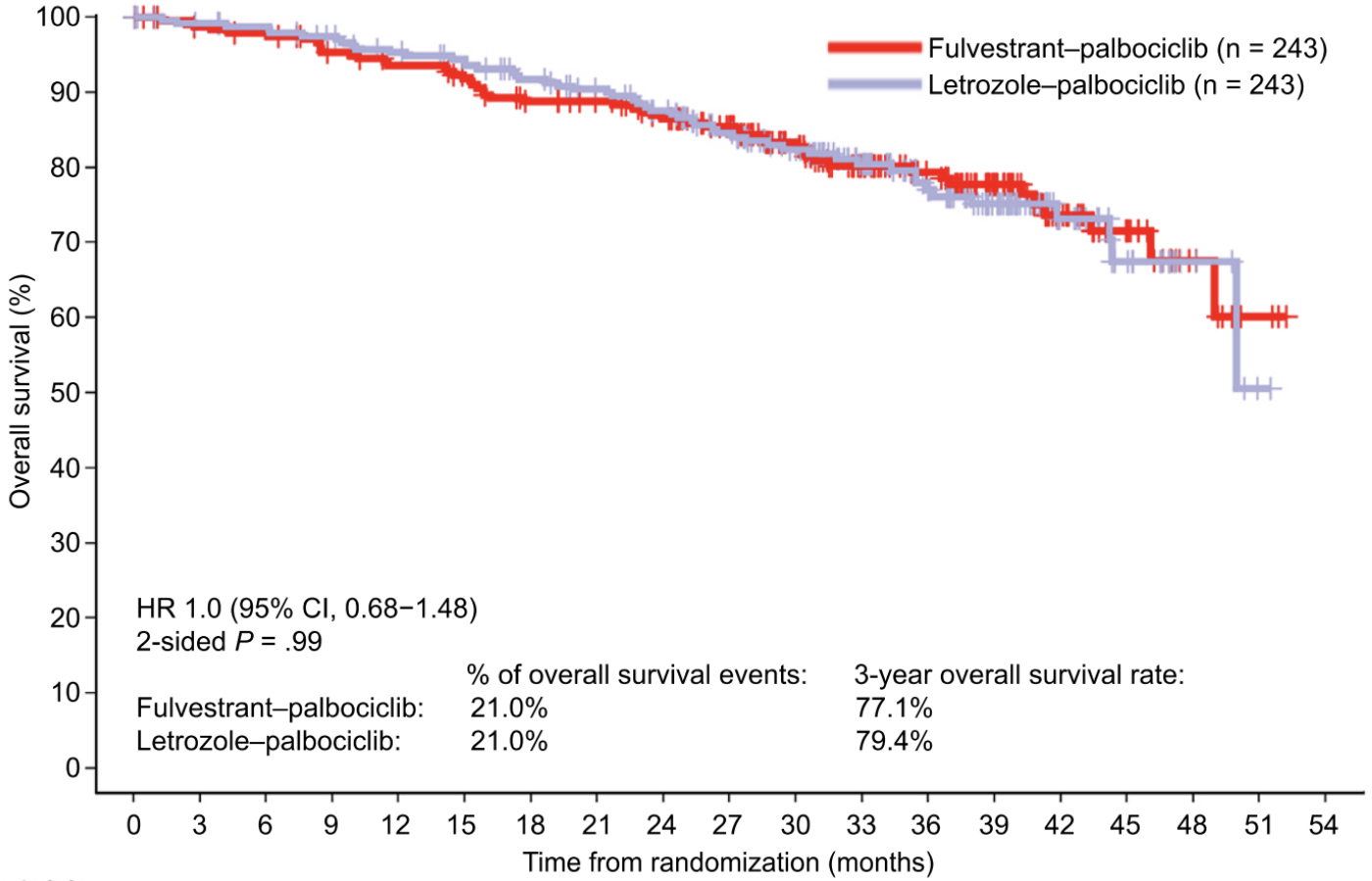
Office Visits

	N	%N Missing	Mean Encounters[b] per Patient-Day
Palbociclib and Fulvestrant	458	0.9%	0.12
Palbociclib and Letrozole	1417	0.5%	0.11

[a] Imaging data was missing and/or incomplete for the vast majority of patients

[b] Office visit encounters proxied by blood pressure/pulse readings

Supplementary Figure 3.1. Kaplan-Meier Estimates of Overall Survival (PARSIFAL)



Number at risk

Fulvestrant-Palbociclib	243	235	232	222	216	209	198	193	182	167	141	114	96	71	44	24	10	3	0
Letrozole-Palbociclib	243	236	233	227	219	214	204	194	184	165	141	116	87	64	35	20	6	1	0

Abbreviations: 95% CI, 95% of confidence interval; HR, hazard ratio; OS, overall survival.

Bibliography

1. Cartwright TH, Clayton M, Garey JS, Boehm KA. Use of an electronic health record (iKnowMed), in conjunction with evidence-based pathways, for data capture and outcome measurement in colorectal cancer patients treated with first-line therapy in the US Oncology Network. Journal of Clinical Oncology. 2010;28(15_suppl):3626-. doi: 10.1200/jco.2010.28.15_suppl.3626.

2. Parikh RB, Galsky MD, Gyawali B, Riaz F, Kaufmann TL, Cohen AB, et al. Trends in Checkpoint Inhibitor Therapy for Advanced Urothelial Cell Carcinoma at the End of Life: Insights from Real-World Practice. Oncologist. 2019;24(6):e397-e9. Epub 2019/04/05. doi: 10.1634/theoncologist.2019-0039. PubMed PMID: 30944183; PubMed Central PMCID: PMCPMC6656487 article.

3. Ruley M, Walker V, Studeny J, Coustasse A. The Nationwide Health Information Network: The Case of the Expansion of Health Information Exchanges in the United States. Health Care Manag (Frederick). 2018;37(4):333-8. Epub 2018/09/21. doi: 10.1097/HCM.0000000000000231. PubMed PMID: 30234636.

4. Lin KJ, Glynn RJ, Singer DE, Murphy SN, Lii J, Schneeweiss S. Out-of-system Care and Recording of Patient Characteristics Critical for Comparative Effectiveness Research. Epidemiology. 2018;29(3):356-63. Epub 2017/12/29. doi: 10.1097/EDE.0000000000000794. PubMed PMID: 29283893; PubMed Central PMCID: PMCPMC5882528.

5. Lin KJ, Singer DE, Glynn RJ, Murphy SN, Lii J, Schneeweiss S. Identifying Patients With High Data Completeness to Improve Validity of Comparative Effectiveness Research in Electronic Health Records Data. Clin Pharmacol Ther. 2018;103(5):899-905. Epub 2017/09/03. doi: 10.1002/cpt.861. PubMed PMID: 28865143; PubMed Central PMCID: PMCPMC6026022.

6. Austin PC. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. Stat Med. 2009;28(25):3083-107. Epub 2009/09/17. doi: 10.1002/sim.3697. PubMed PMID: 19757444; PubMed Central PMCID: PMCPMC3472075.

7. Li N, Hao Y, Xie J, Lin PL, Koo V, Ohashi E, et al. Everolimus-Based Therapy versus Chemotherapy among Patients with HR+/HER2- Metastatic Breast Cancer: Comparative Effectiveness from a Chart Review Study. Int J Breast Cancer. 2015;2015:240750. Epub 2015/06/17. doi: 10.1155/2015/240750. PubMed PMID: 26078883; PubMed Central PMCID: PMCPMC4452841.

8. Paleri V, Wight RG, Silver CE, Haigentz M, Jr., Takes RP, Bradley PJ, et al. Comorbidity in head and neck cancer: a critical appraisal and recommendations for practice. Oral Oncol. 2010;46(10):712-9. Epub 2010/09/21. doi: 10.1016/j.oraloncology.2010.07.008. PubMed PMID: 20850371.

9. Peyton CC, Tang D, Reich RR, Azizi M, Chipollini J, Pow-Sang JM, et al. Downstaging and Survival Outcomes Associated With Neoadjuvant Chemotherapy Regimens Among Patients Treated

With Cystectomy for Muscle-Invasive Bladder Cancer. JAMA Oncol. 2018;4(11):1535-42. Epub 2018/09/05. doi: 10.1001/jamaoncol.2018.3542. PubMed PMID: 30178038; PubMed Central PMCID: PMCPMC6248089.

10. Greenland S. Quantifying biases in causal models: classical confounding vs collider-stratification bias. Epidemiology. 2003;14(3):300-6. Epub 2003/07/16. PubMed PMID: 12859030.

11. Robins JM, Hernan MA, Brumback B. Marginal structural models and causal inference in epidemiology. Epidemiology. 2000;11(5):550-60. Epub 2000/08/24. doi: 10.1097/00001648-200009000-00011. PubMed PMID: 10955408.

12. Joshua Lin K, Jin Y, Gagne J, Glynn RJ, Murphy SN, Tong A, et al. Longitudinal Data Discontinuity in Electronic Health Records and Consequences for Medication Effectiveness Studies. Clin Pharmacol Ther. 2021. Epub 2021/08/24. doi: 10.1002/cpt.2400. PubMed PMID: 34424534.

13. Bonamici RS. H.R. 34 - 114th Congress (2015-2016): 21st Century Cures Act 2016. Available from: **https://www.congress.gov/bill/114th-congress/house-bill/34**.

14. Platt R, Brown JS, Robb M, McClellan M, Ball R, Nguyen MD, et al. The FDA Sentinel Initiative - An Evolving National Resource. N Engl J Med. 2018;379(22):2091-3. Epub 2018/11/30. doi: 10.1056/NEJMp1809643. PubMed PMID: 30485777.

15. Franklin JM, Patorno E, Desai RJ, Glynn RJ, Martin D, Quinto K, et al. Emulating Randomized Clinical Trials With Nonrandomized Real-World Evidence Studies: First Results From the RCT DUPLICATE Initiative. Circulation. 2021;143(10):1002-13. Epub 2020/12/18. doi: 10.1161/CIRCULATIONAHA.120.051718. PubMed PMID: 33327727; PubMed Central PMCID: PMCPMC7940583.

16. Franklin JM, Pawar A, Martin D, Glynn RJ, Levenson M, Temple R, et al. Nonrandomized Real-World Evidence to Support Regulatory Decision Making: Process for a Randomized Trial Replication Project. Clin Pharmacol Ther. 2020;107(4):817-26. Epub 2019/09/22. doi: 10.1002/cpt.1633. PubMed PMID: 31541454.

17. Rugo HS, Finn RS, Dieras V, Ettl J, Lipatov O, Joy AA, et al. Palbociclib plus letrozole as first-line therapy in estrogen receptor-positive/human epidermal growth factor receptor 2-negative advanced breast cancer with extended follow-up. Breast Cancer Res Treat. 2019;174(3):719-29. Epub 2019/01/12. doi: 10.1007/s10549-018-05125-4. PubMed PMID: 30632023; PubMed Central PMCID: PMCPMC6438948.

18. Finn RS, Martin M, Rugo HS, Jones S, Im SA, Gelmon K, et al. Palbociclib and Letrozole in Advanced Breast Cancer. N Engl J Med. 2016;375(20):1925-36. Epub 2016/12/14. doi: 10.1056/NEJMoa1607303. PubMed PMID: 27959613.

19. Walker B, Boyd M, Aguilar K, Davies K, Espirito J, Frytak J, et al. Comparisons of Real-World Time-to-Event End Points in Oncology Research. JCO Clin Cancer Inform. 2021;5:45-6. Epub 2021/01/08. doi: 10.1200/CCI.20.00125. PubMed PMID: 33411622.

20. National Comprehensive Cancer Network. Breast Cancer (Version 1.2022). **https://www.nccn.org/professionals/physician_gls/pdf/breast.pdf**. Accessed December 2, 2021.

21. Rubin DB. Inference and Missing Data. Biometrika. 1976;63(3):581-92. doi: 10.2307/2335739.

22. Services USCfMM. Oncology Care Model | CMS Innovation Center [December 9, 2021]. Available from: **https://innovation.cms.gov/innovation-models/oncology-care**.

23. Azur MJ, Stuart EA, Frangakis C, Leaf PJ. Multiple imputation by chained equations: what is it and how does it work? Int J Methods Psychiatr Res. 2011;20(1):40-9. Epub 2011/04/19. doi: 10.1002/mpr.329. PubMed PMID: 21499542; PubMed Central PMCID: PMCPMC3074241.

24. Schafer JL, Olsen MK. Multiple Imputation for Multivariate Missing-Data Problems: A Data Analyst's Perspective. Multivariate Behav Res. 1998;33(4):545-71. Epub 1998/10/01. doi: 10.1207/s15327906mbr3304_5. PubMed PMID: 26753828.

25. Sterne JA, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ. 2009;338:b2393. Epub 2009/07/01. doi: 10.1136/bmj.b2393. PubMed PMID: 19564179; PubMed Central PMCID: PMCPMC2714692.

26. Rubin DB. Multiple Imputation after 18+ Years. Journal of the American Statistical Association. 1996;91(434):473-89. doi: 10.1080/01621459.1996.10476908.

27. Rubin DB. Multiple imputation for nonresponse in surveys. Hoboken, N.J. ;: Wiley-Interscience; 2004. xxix, 287 p. p.

28. Zhu Y, Hubbard RA, Chubak J, Roy J, Mitra N. Core concepts in pharmacoepidemiology: Violations of the positivity assumption in the causal analysis of observational data: Consequences and statistical approaches. Pharmacoepidemiol Drug Saf. 2021;30(11):1471-85. Epub 2021/08/11. doi: 10.1002/pds.5338. PubMed PMID: 34375473; PubMed Central PMCID: PMCPMC8492528.

29. Sjolander A, Dahlqwist E, Zetterqvist J. A Note on the Noncollapsibility of Rate Differences and Rate Ratios. Epidemiology. 2016;27(3):356-9. Epub 2015/12/19. doi: 10.1097/EDE.0000000000000433. PubMed PMID: 26680297.

30. Toikkanen S, Helin H, Isola J, Joensuu H. Prognostic significance of HER-2 oncoprotein expression in breast cancer: a 30-year follow-up. J Clin Oncol. 1992;10(7):1044-8. Epub 1992/07/01. doi: 10.1200/JCO.1992.10.7.1044. PubMed PMID: 1351537.

31. Franklin JM, Glynn RJ, Martin D, Schneeweiss S. Evaluating the Use of Nonrandomized Real-World Data Analyses for Regulatory Decision Making. Clin Pharmacol Ther. 2019;105(4):867-77. Epub 2019/01/14. doi: 10.1002/cpt.1351. PubMed PMID: 30636285.

32. Administration FaD. Real-World Evidence 2021. Available from: **https://www.fda.gov/science-research/science-and-research-special-topics/real-world-evidence**.

33. Califf RM, Robb MA, Bindman AB, Briggs JP, Collins FS, Conway PH, et al. Transforming Evidence Generation to Support Health and Health Care Decisions. N Engl J Med. 2016;375(24):2395-400. Epub 2016/12/16. doi: 10.1056/NEJMsb1610128. PubMed PMID: 27974039.

34. Eichler HG, Koenig F, Arlett P, Enzmann H, Humphreys A, Petavy F, et al. Are Novel, Nonrandomized Analytic Methods Fit for Decision Making? The Need for Prospective, Controlled, and Transparent Validation. Clin Pharmacol Ther. 2020;107(4):773-9. Epub 2019/10/02. doi: 10.1002/cpt.1638. PubMed PMID: 31574163; PubMed Central PMCID: PMCPMC7158212.

35. Administration FaD. Oncology Real World Evidence Program 2021. Available from: **https://www.fda.gov/about-fda/oncology-center-excellence/oncology-real-world-evidence-program**.

36. Brufsky A, Liu X, Li B, McRoy L, Layman RM. Real-World Tumor Response of Palbociclib Plus Letrozole Versus Letrozole for Metastatic Breast Cancer in US Clinical Practice. Target Oncol. 2021;16(5):601-11. Epub 2021/08/03. doi: 10.1007/s11523-021-00826-1. PubMed PMID: 34338965; PubMed Central PMCID: PMCPMC8484164.

37. Carrigan G, Whipple S, Capra WB, Taylor MD, Brown JS, Lu M, et al. Using Electronic Health Records to Derive Control Arms for Early Phase Single-Arm Lung Cancer Trials: Proof-of-Concept in Randomized Controlled Trials. Clin Pharmacol Ther. 2020;107(2):369-77. Epub 2019/07/28. doi: 10.1002/cpt.1586. PubMed PMID: 31350853; PubMed Central PMCID: PMCPMC7006884.

38. Paul SK, Montvida O, Best JH, Gale S, Pethoe-Schramm A, Sarsour K. Effectiveness of biologic and non-biologic antirheumatic drugs on anaemia markers in 153,788 patients with rheumatoid arthritis: New evidence from real-world data. Semin Arthritis Rheum. 2018;47(4):478-84. Epub 2017/09/28. doi: 10.1016/j.semarthrit.2017.08.001. PubMed PMID: 28947313.

39. Tan K, Bryan J, Segal B, Bellomo L, Nussbaum N, Tucker M, et al. Emulating Control Arms for Cancer Clinical Trials Using External Cohorts Created From Electronic Health Record-Derived Real-World Data. Clin Pharmacol Ther. 2021. Epub 2021/07/02. doi: 10.1002/cpt.2351. PubMed PMID: 34197637.

40. Booth CM, Karim S, Mackillop WJ. Real-world data: towards achieving the achievable in cancer care. Nat Rev Clin Oncol. 2019;16(5):312-25. Epub 2019/02/01. doi: 10.1038/s41571-019-0167-7. PubMed PMID: 30700859.

41. Jemielita T, Li XN, Piperdi B, Zhou W, Burke T, Chen C. Overall Survival With Second-Line Pembrolizumab in Patients With Non-Small-Cell Lung Cancer: Randomized Phase III Clinical Trial Versus Propensity-Adjusted Real-World Data. JCO Clin Cancer Inform. 2021;5:56-65. Epub 2021/01/14. doi: 10.1200/CCI.20.00099. PubMed PMID: 33439727.

42. Jemielita T, Widman L, Fox C, Salomonsson S, Liaw KL, Pettersson A. Replication of Oncology Randomized Trial Results using Swedish Registry Real World-Data: A Feasibility Study. Clin Pharmacol Ther. 2021;110(6):1613-21. Epub 2021/09/23. doi: 10.1002/cpt.2424. PubMed PMID: 34549809.

43. Soni PD, Hartman HE, Dess RT, Abugharib A, Allen SG, Feng FY, et al. Comparison of Population-Based Observational Studies With Randomized Trials in Oncology. J Clin Oncol. 2019;37(14):1209-16. Epub 2019/03/22. doi: 10.1200/JCO.18.01074. PubMed PMID: 30897037; PubMed Central PMCID: PMCPMC7186578.

44. Hernan MA, Sauer BC, Hernandez-Diaz S, Platt R, Shrier I. Specifying a target trial prevents immortal time bias and other self-inflicted injuries in observational analyses. J Clin Epidemiol. 2016;79:70-5. Epub 2016/05/31. doi: 10.1016/j.jclinepi.2016.04.014. PubMed PMID: 27237061; PubMed Central PMCID: PMCPMC5124536.

45. Caniglia EC, Zash R, Jacobson DL, Diseko M, Mayondi G, Lockman S, et al. Emulating a target trial of antiretroviral therapy regimens started before conception and risk of adverse birth outcomes. AIDS. 2018;32(1):113-20. Epub 2017/11/08. doi: 10.1097/QAD.0000000000001673. PubMed PMID: 29112066; PubMed Central PMCID: PMCPMC5718935.

46. Llombart-Cussac A, Perez-Garcia JM, Bellet M, Dalenc F, Gil-Gil M, Ruiz-Borrego M, et al. Fulvestrant-Palbociclib vs Letrozole-Palbociclib as Initial Therapy for Endocrine-Sensitive, Hormone Receptor-Positive, ERBB2-Negative Advanced Breast Cancer: A Randomized Clinical Trial. JAMA Oncol. 2021. Epub 2021/10/08. doi: 10.1001/jamaoncol.2021.4301. PubMed PMID: 34617955; PubMed Central PMCID: PMCPMC8498933.

47. Cox DR. Regression models and life-tables. J R Stat Soc Ser B Methodol. 1972;34:187–220.

48. Brookhart MA, Schneeweiss S, Rothman KJ, Glynn RJ, Avorn J, Sturmer T. Variable selection for propensity score models. Am J Epidemiol. 2006;163(12):1149-56. Epub 2006/04/21. doi: 10.1093/aje/kwj149. PubMed PMID: 16624967; PubMed Central PMCID: PMCPMC1513192.

49. White IR, Carlin JB. Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values. Stat Med. 2010;29(28):2920-31. Epub 2010/09/16. doi: 10.1002/sim.3944. PubMed PMID: 20842622.

50. Schulte PJ, Mascha EJ. Propensity Score Methods: Theory and Practice for Anesthesia Research. Anesth Analg. 2018;127(4):1074-84. Epub 2018/05/12. doi: 10.1213/ANE.0000000000002920. PubMed PMID: 29750691.

51. Mathur MB, Ding P, Riddell CA, VanderWeele TJ. Web Site and R Package for Computing E-values. Epidemiology. 2018;29(5):e45-e7. Epub 2018/06/19. doi: 10.1097/EDE.0000000000000864. PubMed PMID: 29912013; PubMed Central PMCID: PMCPMC6066405.

52. VanderWeele TJ, Ding P. Sensitivity Analysis in Observational Research: Introducing the E-Value. Ann Intern Med. 2017;167(4):268-74. Epub 2017/07/12. doi: 10.7326/M16-2607. PubMed PMID: 28693043.

53. Hershman DL, Wright JD. Comparative effectiveness research in oncology methodology: observational data. J Clin Oncol. 2012;30(34):4215-22. Epub 2012/10/17. doi: 10.1200/JCO.2012.41.6701. PubMed PMID: 23071228.

54. Schneeweiss S, Patorno E. Conducting Real-world Evidence Studies on the Clinical Outcomes of Diabetes Treatments. Endocr Rev. 2021;42(5):658-90. Epub 2021/03/13. doi: 10.1210/endrev/bnab007. PubMed PMID: 33710268; PubMed Central PMCID: PMCPMC8476933.

55. Huybrechts K, Schneeweiss S. Secondary Data. In: Modern Epidemiology, 4th edition, pp. 247-262, by Rothman K, Lash T, Greenland S, VanderWeele T (Eds.), Wolters Kluwer, Philadelphia, PA, 2021.

56. Schaumberg DA, McDonald L, Shah S, Stokes M, Nordstrom BL, Ramagopalan SV. Evaluation of comparative effectiveness research: a practical tool. J Comp Eff Res. 2018;7(5):503-15. Epub 2018/02/22. doi: 10.2217/cer-2018-0007. PubMed PMID: 29463115.

57. EHR for Oncology Practices | Ontada 2021 [December 23, 2021]. Available from: **https://www.ontada.com/Providers-Solutions/iKnowMed/**.