

A Deep-Learning-Enabled Workflow to Estimate Real-World Progression-Free Survival in Patients with Metastatic Breast Cancer: Study Using De-identified Electronic Health Records

Gowtham Varma, Rohit Kumar Yenukoti, Praveen Kumar M, Bandlamudi Sai Ashrit, K Purushotham, Subash C, Sunil Kumar Ravi, Verghese Kurien, Avinash Aman, Mithun Manoharan, Shashank Jaiswal, Akash Anand, Rakesh Barve, Viswanathan Thiagarajan, Patrick Lenehan, Scott A. Soefje, Venky Soundararajan

Submitted to: JMIR Cancer
on: July 24, 2024

Disclaimer: © The authors. All rights reserved. This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on its website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressly prohibit redistribution of this draft paper other than for review purposes.

Table of Contents

Original Manuscript..... 5

Supplementary Files..... 25

Figures 26

Figure 0..... 27

Figure 0..... 28

Figure 1..... 29

Figure 2..... 30

Figure 3..... 31

Figure 4..... 32

Figure 5..... 33

Multimedia Appendixes 34

Multimedia Appendix 1..... 35

CONSORT (or other) checklists..... 36

CONSORT (or other) checklist 0..... 36

A Deep-Learning-Enabled Workflow to Estimate Real-World Progression-Free Survival in Patients with Metastatic Breast Cancer: Study Using De-identified Electronic Health Records

Gowtham Varma¹ MBBS; Rohit Kumar Yenukoti¹ MBBS; Praveen Kumar M¹ MD, DM; Bandlamudi Sai Ashrit¹ MBBS; K Purushotham¹ MBBS; Subash C¹ MBBS; Sunil Kumar Ravi¹ MBBS; Verghese Kurien¹ MBBS; Avinash Aman¹ BTech; Mithun Manoharan¹ MBBS, MMST; Shashank Jaiswal¹ B.Tech; Akash Anand¹ BE; Rakesh Barve¹ BTech, PhD; Viswanathan Thiagarajan¹ B.Tech; Patrick Lenehan² MD, PhD; Scott A. Soefje³ BCOP, MBA, FCCP, PharmD; Venky Soundararajan² PhD

¹ Nference Bangalore IN

² Nference Cambridge US

³ Mayo Clinic Rochester US

Corresponding Author:

Praveen Kumar M MD, DM

Nference

4th Floor, IndiQube

Bangalore

IN

Abstract

Background: Progression-free survival (PFS) is a crucial endpoint in cancer drug research. Real-world evidence (RWE) is increasingly accepted to augment traditional clinical trial findings to better understand the effectiveness of oncological interventions. RWE can be leveraged to improve novel therapy development programs and provide better post-market surveillance of approved therapies.

Objective: To develop and validate a novel semi-automated workflow that estimates real-world progression-free survival (rwPFS) in patients with metastatic breast cancer (mBC).

Methods: This study analyzes de-identified EHR data using the nference® nSights platform. The overall cohort included 316 patients with HR-positive, and HER2-negative mBC, who were started on Palbociclib and Letrozole combination therapy between January 1, 2015, and December 31, 2021. We curated ground-truth datasets to evaluate the workflow's performance at sentence and patient levels. NLP-captured-progression or a change in the line of therapy were considered outcome events. Death, loss to follow-up, and end of study period were considered censoring events for computing rwPFS.

Results: At the sentence-level, progression events were captured from clinical notes and radiology reports with a sensitivity of 99.8%, specificity of 96.7%, and accuracy of 98.2%. At the patient-level, initial progression was correctly captured within a window of ± 30 days with a sensitivity of 92.5%, specificity of 83.0%, and accuracy of 88.0%. In a sample of 100 patients, the median rwPFS was determined to be 22 months (95% CI; 15-35 months) by the computational workflow and 25 months (95% CI; 15-35 months) by manual curation.

Conclusions: A semi-automated workflow enabled rapid and reliable determination of rwPFS in mBC patients receiving a combination therapy.

(JMIR Preprints 24/07/2024:64697)

DOI: <https://doi.org/10.2196/preprints.64697>

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

Please make my preprint PDF available to anyone at any time (recommended).

Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.

✓ **No, I do not wish to publish my submitted manuscript as a preprint.**

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).

Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain visible to all users.

✓ **Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in a JMIR journal, my full manuscript will be made available to all users.**

No. Please do not make my accepted manuscript PDF available to anyone.



Original Manuscript

A Deep-Learning-Enabled Workflow to Estimate Real-World Progression-Free Survival in Patients with Metastatic Breast Cancer: Study Using De-identified Electronic Health Records

Gowtham Varma¹, Rohit Kumar Yenukoti¹, Praveen Kumar-M^{*1}, Bandlamudi Sai Ashrit¹, Purushotham K¹, Subash C¹, Sunil Kumar Ravi¹, Verghese Kurien¹, Avinash Aman¹, Mithun Manoharan¹, Shashank Jaiswal¹, Akash Anand¹, Rakesh Barve¹, Viswanathan Thiagarajan¹, Patrick J. Lenehan², Scott A. Soefje³, Venky Soundararajan²

*- Corresponding author

Praveen Kumar-M,
Nference Labs Pvt Ltd, Bengaluru, India.
Email - pkumar@nference.net.

1- Nference Labs Pvt Ltd, Bengaluru, India;

2-Nference Inc. Cambridge, MA;

3-Mayo Clinic, Rochester, MN;

Keywords: real-world evidence, data-driven oncology, real-world progression-free survival, metastatic breast cancer, deep-learning, natural language processing

Abstract

Background: Progression-free survival (PFS) is a crucial endpoint in cancer drug research. The clinician-confirmed cancer progression, namely real-world PFS (rwPFS) in unstructured text (i.e. clinical notes) has been shown to serve as a reasonable surrogate for real-world indicators in ascertaining progression endpoints. Response Evaluation Criteria in Solid Tumors (RECIST) is traditionally used in clinical trials using serial imaging evaluations, which is not practical when working with real-world data. Manual abstraction of clinical progression from unstructured notes continues to be the gold standard. However, this process is a resource-intensive and time-consuming process. Natural Language processing (NLP), a subdomain of machine learning, has shown promise in accelerating the extraction of tumor progression from real world data in recent years.

Objectives: We aim to configure a pre-trained, general-purpose healthcare NLP framework to transform free-text clinical notes and radiology reports into structured progression events for studying rwPFS on metastatic breast cancer (mBC) cohorts.

Methods: This study developed and validated a novel semi-automated workflow to estimate rwPFS in patients with mBC using de-identified EHR data from the nference nSights platform. The developed workflow was validated in a cohort of 316 patients with hormone receptor-positive, human epidermal growth factor receptor 2 (HER2)-negative mBC, who were started on Palbociclib and Letrozole combination therapy between January 2015 and December 2021. Ground-truth datasets were curated to evaluate the workflow's performance at both the sentence and patient levels. NLP-captured progression or a change in therapy line were considered outcome events, while death, loss to follow-up, and end of study period were considered censoring events for rwPFS computation. Peak reduction and cumulative decline in Patient-Health-Questionnaire-8 (PHQ-8) scores were analyzed in the progressed and non-progressed patient subgroups.

Results: The configured clinical NLP engine achieved a sentence-level progression capture accuracy of 98.2%. At the patient level, initial progression was captured within ± 30 days with 88% accuracy. The median real-world progression-free survival (rwPFS) for the study cohort (N=316) was 20 months (95% CI: 18.0–25.0). In a validation subset (N=100), rwPFS determined by manual curation was 25 months (95% CI: 15–35 months), closely aligning with the computational workflow's 22 months (95% CI: 15–35 months). A sub-analysis revealed rwPFS estimates of 30 months (95% CI: 24.0–39.0) from radiology reports and 23 months (95% CI: 19.0–28.0) from clinical notes, highlighting the importance of integrating multiple note sources. External validation also demonstrated high accuracy (92.5%-sentence-level; 90.2%-patient-level). Sensitivity analysis revealed stable rwPFS estimates across varying levels of missing source data and event definitions. Peak reduction and cumulative decline in PHQ-8 scores during the study period highlighted significant associations between patient-reported outcomes and disease progression.

Conclusions: This workflow enables rapid and reliable determination of rwPFS in mBC patients receiving combination therapy. Further validation across more diverse external datasets and other cancer types is needed to ensure broader applicability and generalizability.

Introduction

Background and Significance

Real-world evidence (RWE) is increasingly accepted to augment traditional clinical trial findings to better understand the effectiveness of oncological interventions. RWE can be leveraged to improve novel therapy development programs and provide better post-market surveillance of approved therapies.[1–3]

Response Evaluation Criteria in Solid Tumors (RECIST) is broadly utilized to ascertain disease progression in clinical trials. However, assessing RECIST in retrospective Electronic Health Record (EHR) data is challenging due to its strict assessment indicators.[4] RECIST considers changes in the size of individual target lesions over time and the presence or absence of new lesions to categorize disease status into complete/partial response, stable disease, or progression.[5] A similar assessment paradigm is adopted in routine clinical practice, where clinicians document the occurrence of progression through serial clinical and radiological examinations. This clinician-confirmed cancer progression in unstructured text (i.e. clinical notes) has been shown to serve as a reasonable surrogate for real-world indicators in ascertaining progression endpoints. This is also more practical for real-world studies than purely RECIST-based approaches.[6]

In earlier studies across different types of solid tumors, real-world progression was captured either through manual abstraction from unstructured data or proxy measures were evaluated based solely on structured drug data. [7,8]. Recent studies have also utilized machine learning models specifically trained to automate the capture and characterization of clinician documentation of tumor response. These specialized models have shown varying accuracies. [9,10]. In the past decade, healthcare NLP frameworks like Google's Healthcare Natural Language API, Amazon

Comprehend Medical, IBM Watson Health, and Microsoft Text Analytics for Health have emerged and shown promise in clinical concept recognition, entity linking, and sentiment analysis. However, these general-purpose NLP frameworks have shown varying degrees of performance on different data sources.[11–13] While large language models (LLMs) are rapidly advancing, they currently have limitations in clinical concept identification and medical relation extraction as structured outputs for direct application. Even specialized clinical LLMs require further fine-tuning for such use cases. [14]

We aim to configure a pre-trained, general-purpose healthcare NLP framework to transform free-text clinical notes and radiology reports into structured progression events. By combining these with structured drug records and encounter data, we will compute real-world progression-free survival (rwPFS) for metastatic breast cancer. This work can also guide other researchers in configuring a general-purpose healthcare NLP model to capture rwPFS. Developing a standardized and automated path for ascertaining real-world progression (rwP) could help scale rwPFS computation across diverse subsets of solid tumors and maintain a better agreement across real-world studies.

Objectives

We aim to: (1) develop and validate a novel semi-automated workflow that estimates real-world progression-free survival (rwPFS) in patients with metastatic breast cancer (mBC), (2) explore the essentiality of each model in the general purpose NLP framework through ablation analysis, and (3) investigate additional factors influencing rwPFS, such as the source of clinician-documented progressions (radiology reports versus routine clinical notes) and the presence of prior or concurrent medications during the observation period.

Methodology

Data source

This study analyzed de-identified EHR data from a network of tertiary clinical centers tied to an academic medical center in the United States through the nference nSights Analytics Platform. In-house tools were used for the De-identification (De-ID) of electronic health records (EHR). The tool performs with a recall of 0.992 and a precision of 0.979 on the i2b2 2014 dataset at replacing protected health information (PHI) with plausible but fictional surrogates.[15] Overall, the platform hosts data from approximately 7 million patients from across the United States of America, with about 1.8 million patients having a mention of cancer across the structured data. The platform hosts an array of multimodal data, both structured and unstructured, such as clinical notes, radiology reports, DICOM headers, and pathology reports.

Figure 1. Methodology flow diagram illustrating the workflow

Ensemble of deep-learning-based multi-BERT framework

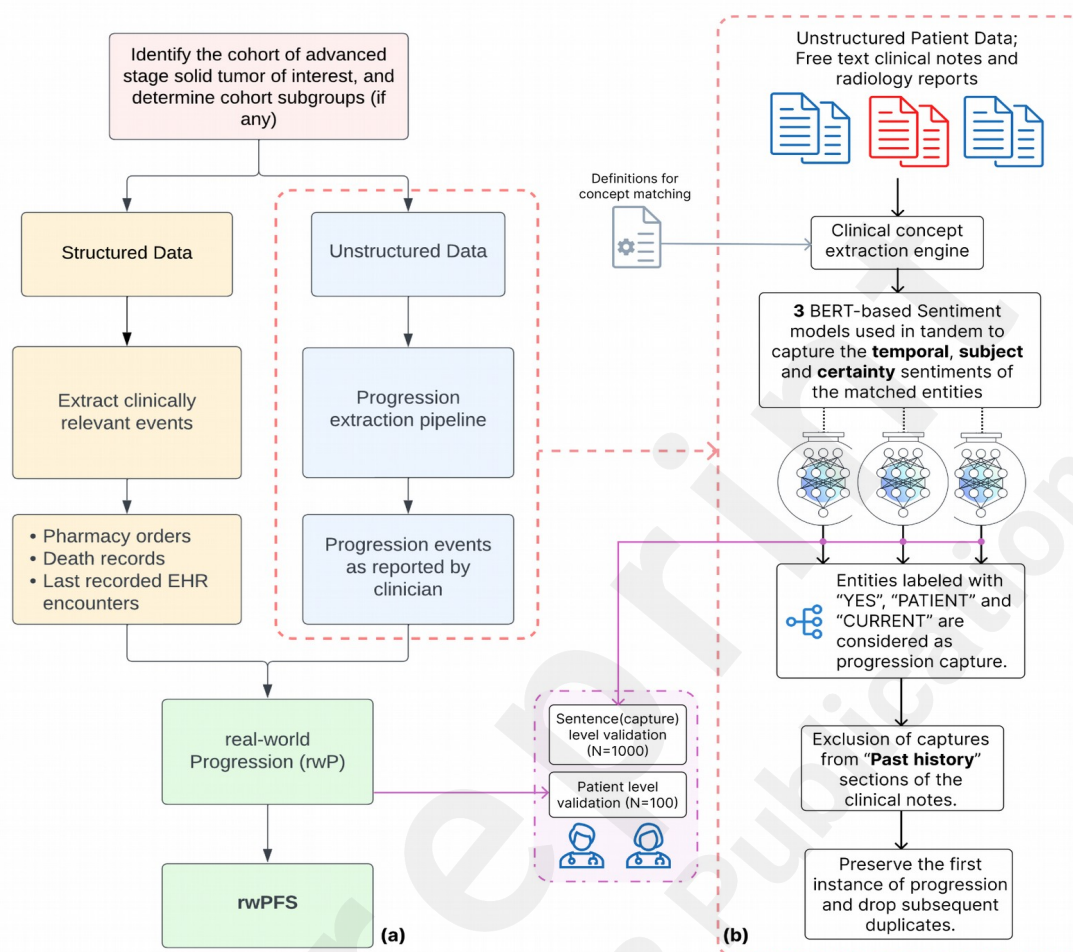


Figure 1. Methodology flow diagram illustrating the workflow: (a) Illustrates the workflow for real-world progression (rwP) extraction and determining the rwPFS. (b) Illustrates the methodology for capturing progression from unstructured texts in routine clinical documents and radiology reports using Nference's clinical NLP engine that performs clinical concept recognition, association, and sentiment analysis.

Study Design and Definitions

This retrospective observational study demonstrates the estimation of real-world progression-free survival with a workflow that integrates clinician-reported progression events from free-text (unstructured data) in clinical and radiology documents with structured patient events like drug orders, clinic/hospital encounters, and mortality records. The workflow was developed to leverage the pre-trained, general-purpose, deep-learning-based healthcare NLP framework developed at Nference called the Clinical NLP engine, which enables clinical concept recognition, sentiment analysis, and linking associated concepts. The models that are a part of the clinical NLP engine were initially trained on human-annotated datasets, and later further augmented by additional ground truth datasets generated by large language model (LLM) agents from the same parent EHR data source. A high-level overview of the workflow is illustrated in Figure 1.

Data Extraction and Augmentation

Breast cancer disease was identified using structured diagnosis codes 174 (ICD-9), C50 (ICD-10), and NLP-based positive model confirmations (augmented curation) of the disease-related terms in clinical notes.[16] A similar approach was undertaken for identifying metastatic disease using structured codes 197, 198 (ICD-9), C78, and C79 (ICD-10). Further cohort attrition for the study population is outlined in Figure 2. Hormone receptor (HR) status, human epidermal growth factor receptor-2 (HER-2) status, and ECOG (Eastern Cooperative Oncology Group) scores were

captured from clinical notes using the clinical NLP engine. A rule-based approach was utilized to identify the initiation date of first-line therapy in metastatic breast cancer by analyzing drug orders and administration records. To ensure reliability, only orders appearing for the first time after metastasis diagnosis were included. The same methodology was extended to identify second-line therapies.

Patient Population

The study cohort (N=316) consisted of female patients aged ≥ 18 and diagnosed with metastatic breast cancer with HR-positive and HER2-negative status with confirmed concurrent exposure to palbociclib and letrozole from Jan 1, 2015, to Dec 31, 2021 (study period), and with ECOG performance scores of less than three around the start of the therapy (± 60 days).

Figure 2. Cohort attrition diagram

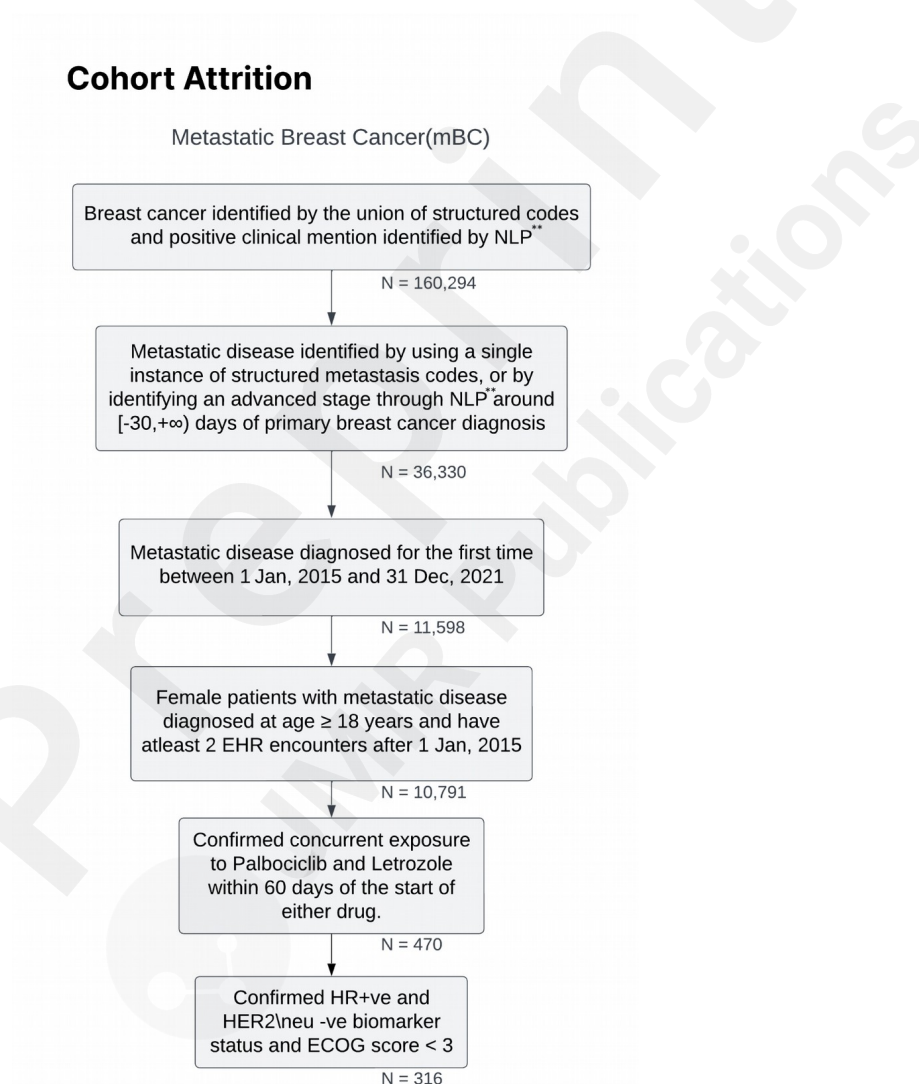


Figure 2. Cohort attrition diagram: Structured codes 174* (ICD-9), and C50* (ICD-10) or >4 positive disease sentiments from the augmented curation disease diagnosis model were used for breast cancer. For evidence of metastasis 197*, 198* (ICD-9), C78*, and C79* (ICD-10) in conjunction with augmented curation was used; *(single asterisk) represents all the children codes within the parent code. **(double asterisk) NLP - Natural Language Processing.

Demographics and baseline characterization of the study cohort, such as prior exposure to therapies, disease stage at the start of the study, stage at first cancer diagnosis, and other relevant metrics for the solid tumor of interest, were also documented.

Extracting Progression Events from Unstructured Text

To develop and evaluate our workflow, an initial rule-definition set of 200 cases from the overall mBC cohort (N=10,791) cohort was sampled, and a preliminary manual abstraction was performed. This evaluation aimed to systematically identify a set of rules for configuring the baseline clinical concept extraction model to capture progression. Authors PKM, SKR, VK and MM utilized internal clinical document exploration tools to cluster sentences based on initial pattern matches and iteratively refined these clusters to identify progression-indicative phrases. Independent reviews of oncology and radiology notes were conducted to extract commonly occurring phrases indicative of clinical progression. Authors GV and RHY subsequently collated these identified patterns into a set of regex search patterns. These patterns were tested on clinical notes to ensure they captured the appropriate progression-related contexts while removing duplicate or irrelevant verbatim such as general report headers, unrelated phrases (e.g., "CR" for complete response or "PD" for progressive disease without patient-specific context), and noise. Downstream NLP models were applied to validate the extracted patterns by mapping the right set of label combinations that accurately reflected the progression status. This process was repeated iteratively until a consensus was reached among the authors, ensuring a robust set of rules for capturing progression events. The final progression capture configuration is detailed in Supplementary Table 1.

The real-world progression (rwP) events were captured by configuring the Clinical NLP Engine. This baseline workflow is an ensemble of deep-learning-based multi-BERT framework trained on unstructured patient data like clinical documents and radiology reports to perform named entity recognition (NER) and predict the sentiment labels for *subject*, *temporality*, and *certainty* of the captured named entities. [17] The ensemble also infers associations between related entities like disease-severity, drug-disease, disease-anatomical_structure among others. These proprietary models are fine-tuned versions of SciBERT-cased[23] a domain-specific transformer model pretrained on scientific text. The base models underwent further supervised fine-tuning for classification tasks on annotated sentences from clinical document texts of the overall nference nSights database, but not specifically on mBC patient note database. The details regarding their architecture, training and performance of the individual models of the clinical NLP engine are detailed in Supplementary Note 1 and Supplementary Table 2. The Clinical NLP Engine also utilizes a section header model to identify the clinical note sections from which the named entities were captured. The rules determined during the initial abstraction were employed to capture cancer progression.

Real-world Progression (rwP) definitions

Real-world Progression (rwP) events were identified by the earliest documentation of disease progression in a clinical or radiology note or by advancement to a new line of therapy. The addition of a new chemotherapy, endocrine therapy, or targeted therapy drug after 60 days of exposure to the initiating therapy of interest is considered line advancement. The following list of drugs were considered potential second-line drug candidates in the study period: Tamoxifen, Fulvestrant, Elacestrant, Paclitaxel, Carboplatin, Abemaciclib, Docetaxel, Cyclophosphamide, Capecitabine, Ribociclib, Alpelisib, Everolimus, Doxorubicin, Epirubicin, 5-Fluorouracil, Olaparib, Talazoparib, Ixabepilone, Raloxifene, Toremifene. Censoring events included death, loss of follow-up, and the end of the observation period. Progression events captured within the first 30 days of therapy initiation were excluded as they occurred too early to reflect treatment effectiveness. The real-world progression-free survival (rwPFS) was also assessed with variations in the origin of unstructured data, comparing radiology reports and clinical documents as data sources. The key contributing survival variables used for rwPFS were also stratified to understand the source of events.

Progression Capture Validation

For validation of the clinical progression captures, the manual review and abstraction were performed at two levels:

1. The raw progression captures were evaluated for accuracy independent of their temporality to the observation period. 1000 captures were sampled from the overall pool of progression captures for the sentence-level progression capture analysis.
2. A stratified sample (mBC validation set; N=100) was selected from the overall study cohort (N=316), to match the progression event occurrence observed in the overall set. This approach ensures that the sample mirrors the broader cohort's characteristics for a valid comparison in patient-level evaluation for progression capture. These patients were not part of the initial rule-definition set and were evaluated for their first progression events. For the patient-level analysis, the elements of the confusion matrix were defined to account for temporality. We classified the captures into four categories:
 - a. True Positives: Automated Progression captured is within ± 30 days of Manual Capture
 - b. False Positives: Progression was not found through manual capture, but automated progression was captured at any time (or) Automated progression was captured >30 days before manual capture.
 - c. True Negatives: Progression was not found through manual review, nor was picked up by automated

- capture.
- d. False Negatives: Progression was identified through manual capture, but the automated method has not identified any progression (or) automated method captured progression >30 days after the date captured by manual review.

Ablation Analysis of the Progression Capture Pipeline

To evaluate the contribution of each workflow component to the overall performance, an ablation study was performed at five strategic points: 1) Temporal Model Ablation i.e., the removal of the temporality assessment model. 2) Subject Model Ablation i.e., the removal of the subject assessment model. 3) Certainty Model Ablation i.e., the certainty assessment model was removed and 4) All three assessment models were removed 5) Post-processing Ablation where the post-processing steps, specifically the exclusion of specific note sections and dropping subsequent duplicate mentions of the same note-contexts were removed. Each ablation was analyzed in isolation to quantify its respective contribution to the final output's accuracy, aiding in identifying critical components of the pipeline and potential areas for optimization. This step is further illustrated in Supplementary-Figure 2.

Validation on External Dataset

To further assess the generalizability and robustness of the progression capture pipeline, external validation was performed using data from a different partner academic medical center (AMC). In the external dataset, 63 mBC patients on first line therapy of the metastatic disease with palbociclib with or without aromatase inhibitors were identified for this analysis (See Supplementary-Figure 4 for cohort attrition). Similar NLP based data augmentation techniques were applied on the external dataset for cohort identification. The proposed progression extraction workflow was applied on the external dataset for extraction of progression events during the defined study period(60 months) of the patients. For validation, similar to the primary cohort, a two-tier approach was applied, including sentence-level and patient-level validation. For patient-level validation, the time of the first progression event patient cohort was manually abstracted and validated. Performance metrics such as precision, recall, accuracy, and F1 score were calculated to evaluate the alignment of automated progression captures with manual annotations.

Sensitivity Analysis on rwPFS estimates

Two sensitivity analyses were conducted to assess the robustness of real-world progression-free survival (rwPFS) estimates: (1) To evaluate the effect of data incompleteness, 10%, 20%, and 30% of rows capturing progression events(missingness at random) were systematically removed from the Kaplan-Meier source data. Median rwPFS and survival probabilities were descriptively analyzed to quantify variations introduced by missing records (2) The impact of treating death as a progression event versus censoring was assessed by generating Kaplan-Meier curves under both scenarios. Differences were evaluated using the log-rank test, with comparisons of median rwPFS and survival probabilities at predefined time points. These analyses ensured the robustness of rwPFS estimates by addressing potential biases from data structure and event definitions.

Patient-Reported Outcomes and Clinical Progression: A PHQ-8 Analysis

Patient-reported outcomes were integrated by analyzing PHQ-8 scores to complement clinician-documented progression events. Cumulative declines and peak reductions in PHQ-8 scores were compared between progressed and non-progressed patients using a t-test. Peak reduction was determined as the largest decrease observed between any two recorded scores during the study period. Cumulative decline, representing the total improvement over time, was calculated as the sum of all positive reductions(decline in PHQ8 value) in scores across all pairwise comparisons during the study period. This approach aimed to provide a holistic perspective by linking patient-reported mental health outcomes to clinical progression.

Outcomes Assessment

The primary outcome was rwPFS, calculated as the time between the start of the intervention of interest and the first rwP event captured or a change in the line of therapy for the patient. The secondary outcome was real-world overall survival (rwOS), calculated as the time between the start of the intervention of interest and the date of death. A sub-group analysis was performed to assess the impact of prior and other concomitant medications on rwPFS and rwOS in the metastatic setting. The validation metrics of the primary outcome were reported as sensitivity, specificity, accuracy, precision, and F1 scores.

Median follow-up time was computed from the date of start of therapy till their last encounter in the EHR system. Time to treatment after the diagnosis of advanced disease (first evidence of metastasis), and follow-up after the start of therapy were imputed using *the date of the first evidence of metastasis* (Identified by structured disease code or

NLP positive confirmation), and *date of first structured drug order for the combination therapy* (palbociclib and letrozole) as the anchor dates respectively.

Ethical Considerations

nference, in collaboration with the academic medical center data partner (AMC) that provided the de-identified data for this study, has established a secure data environment, hosted by and within the AMC, that houses the AMC's de-identified patient data. The provisioning of and access to this data are governed by an expert determination that satisfies the HIPAA Privacy Rule requirements for the de-identification of protected health information. Each AMC's de-identified data environment is specifically designed and operated to enable access to and analysis of de-identified data without the need for Institutional Review Board (IRB) oversight, approval, or an exemption confirmation. The data is accessible only to authorized users subject to a robust credentialing and authentication process. Data shown and reported in the manuscript has been extracted from this environment using an established protocol for data extraction, aimed at preserving patient privacy. The data has been de-identified pursuant to an expert determination in accordance with the HIPAA Privacy Rule.

Statistical Analysis

Data hosted on nference's nSights environment was imported on demand into the secure code workspaces deployed with Python version 3.10.6. Missing data imputation was not undertaken. The analysis workflow utilizes proprietary Python packages with APIs for database querying and data standardization. The descriptive statistics were reported as n (%) and median, interquartile ranges (IQR). Loss to followup was considered as a censoring event for survival estimates. The Kaplan-Meier estimator from the lifelines package 0.27.7 was utilized in this analysis. The median rwPFS and rwOS were reported, with a 95% confidence interval (CI).

Results

Workflow Configuration

We applied three selection conditions to ensure that the progression captures from the clinical NLP engine are relevant and up-to-date with the patient's current status with respect to the clinical note/report. First, entities labeled with 'YES', 'PATIENT', and 'CURRENT', each having a sentiment prediction confidence of ≥ 0.9 , were deemed relevant. Second, to address the issue of "copy-forwarding" in clinical notes, only the first chronological instance of each extracted sentence was retained. Finally, sentences from "Past History" sections were excluded, as they are unlikely to reflect events occurring at the time of documentation. The workflow configuration is illustrated in Figure 1.

Performance Evaluation

The accuracy of the progression capture was evaluated at two levels (Table 1). Level-1; Sentence-level progression capture validation yielded an accuracy of 98.2% for the relevant progression captures, and level-2; patient-level validation yielded an accuracy of 88.0%. Ablation analysis revealed the essentiality of the individual components of the clinical NLP engine for progression capture and selection conditions. All steps except the subject sentiment model labels substantially contribute to the overall model performance. This model can be disabled and the workflow performance remains the same. The patient-level workflow performance at each ablation step is outlined in Table 2.

Table 1: Manual validation was performed for progression at two levels		
Level 1: Manual Validation of Sampled Raw Progression Captures (N = 1000)		
Progression analysis	capture	Sensitivity
		99.8%
		Specificity
		96.7%
		Precision
		96.6%
		Accuracy
		98.2%
		F1 Score
		98.2%
Level 2: Manual Validation of First Progression (N = 100)		
First progression capture	Sensitivity	92.5%

analysis	Specificity	83.0%
	Precision	86.0%
	Accuracy	88.0%
	F1 Score	89.1%

Level 1: Sentence level review to validate the capture of progression sentiments at sentence level. At this level, we reviewed the extracted sentences to ascertain the validity of the progression capture at a sentence level. **Level 2:** Patient level review to identify first progression date. Here, we undertook a full review of patient records to ascertain the first progression capture of the metastatic disease.

Table 2. Output of the ablation analysis showcasing the performance metrics at each step

	Validation against Manually-Abstracted Patient-Level Dataset (N=100)			
	Accuracy (%)	Sensitivity (%)	Specificity (%)	Median PFS (months; 95% CI)
Overall workflow	88.0	92.5	83.0	20 (18-26)
Temporal model ablation	79.0	91.5	67.9	19 (15-23)
Subject model ablation	88.0	92.5	83.0	20 (18-26)
Certainty model ablation	42.0	100	20.5	7 (6-8)
All 3 sentiment models ablated	35	100	15.6	6 (5-7)
Post-processing ablation	87.0	96.2	76.6	19 (16-23)
PFS - progression-free survival.				

Cohort Description

The baseline characteristics are detailed in Table 3. The median age at metastasis was 59 (IQR; 50.5 - 69) years. The median follow-up time after metastasis diagnosis was 43.3 months (IQR; 28.1 - 61.2) and the median follow-up time after the start of the therapy was 39.8 months (IQR; 25.5 - 57.9). The starting dose of palbociclib and letrozole was available for 53.2% and 61%, respectively. The median number of drug orders for palbociclib and letrozole was 6 (IQR; 3-12) and 4 (IQR; 2-7), respectively. The treatment characteristics, including prior and concomitant exposure to other chemotherapy agents and a history of prior radiotherapy and breast surgery, are also detailed in Table 3. The breakdown of outcomes and censoring events that contributed to the Kaplan-Meier survival estimates are further detailed in Table 4.

Table 3. Study cohort characteristics

Category	Variable	Overall mBC Cohort (N = 316)	mBC validation set (N = 100)
Demographics	Age at metastasis (years), median (IQR)	59 (50.5-68.2)	59.1 (47.8-69.2)
	Female gender, n (%)	316 (100%)	100 (100%)
Ethnicity	Not Hispanic or Latino	298 (94.3%)	94 (94%)
	Hispanic or Latino	11 (3.5%)	3 (3%)
	Unknown or Choose Not to Disclose	7 (2.2%)	3 (3%)
Race	Caucasian	293 (92.8%)	95 (95%)
	Asian	6 (1.8%)	1 (1%)
	African American or Other	17 (5.4%)	4 (4%)
Tumor markers	HR-positive, n (%)	316 (100%)	100 (100%)
	ER positive and PR positive	223 (70.6%)	68 (68%)

Table 3. Study cohort characteristics

Category	Variable	Overall mBC Cohort (N = 316)	mBC validation set (N = 100)
	ER positive and PR negative	27 (8.5%)	11 (11%)
	PR positive and ER negative	50 (15.8%)	15 (15%)
	ER, PR status unknown	16 (5.1%)	6 (6%)
	HER2/neu-negative, n (%)	316 (100%)	100 (100%)
Disease severity	Patients with confirmed Stage IV [-30,+30] within 30 days of primary diagnosis, n (%)**	196 (61.1%)	62 (62%)
	ECOG performance score < 3, n (%)	316 (100%)	100 (100%)
Disease-related follow-up	Follow-up after metastasis in months, median (IQR)	43.3 (28.1 - 61.2)	50.8 (38.6 - 64.2)
Treatment	Prior systemic therapy, n (%)	22 (7.5%)	6 (6%)
	Prior radiotherapy, n (%)	125 (39.5%)	38 (38%)
	Prior surgical resection, n (%)	128 (40.5%)	37 (37%)
	Other concomitant systemic therapy, n (%)	36 (11%)	6 (6%)
Treatment Follow-up	Follow-up after start of treatment in months, median (IQR)	39.8 (25.5 - 57.9)	48.6 (37.8 - 61.4)
	Time to treatment after advanced disease diagnosis in months, median (IQR)	0.5 (0.2 - 1.7)	0.6 (0.2 - 1.7)

** (double asterisk) - All patients included in the study are stage 4 cancer. The provided numbers represent those diagnosed within the stated period; IQR: Interquartile Range, HR: Hormone Receptor, HER2: Human epidermal growth factor receptor 2, ECOG: Eastern Cooperative Oncology Group.

Table 4. Breakdown analysis of outcomes and censoring events

mBC Cohort (N = 316)			
	Source of Capture		n (%)
Breakdown of progression events (N = 199)	The first event is a progression capture from the pooled sources (n = 152)	Radiology Reports	78 (51.3%)
		Clinical Documents	74 (48.7%)
	The first event is the start of a second-line drug (n = 47)	Capecitabine	13 (27.7%)
		Everolimus	11 (23.4%)
		Abemaciclib	9 (19.1%)
		Ribociclib	9 (19.1%)
		5-Fluorouracil	3 (6.4%)
		Olaparib	1 (2.1%)
Breakdown of censoring events (N = 117)	Cyclophosphamide		1 (2.1%)
	Last encounter date		76 (65%)
	Patient death date		22 (18.8%)
	End of study period		19 (16.2%)

mBC - metastatic breast cancer.

Outcomes

In the study cohort (N=316), 199 (62.9%) patients progressed during the observation period (60 months starting from Jan 1, 2015). Out of the progressed patients, 152 (48.1%) were based on progression captures from unstructured data, and 47 (14.8%) were based on changes in the line of therapy from structured data. The median rwPFS for the

Figure 3. Kaplan-Meier Survival Plots for the overall and validation cohorts



Figure 4. Kaplan-Meier Survival plots for rwPFS based on the patient note source.

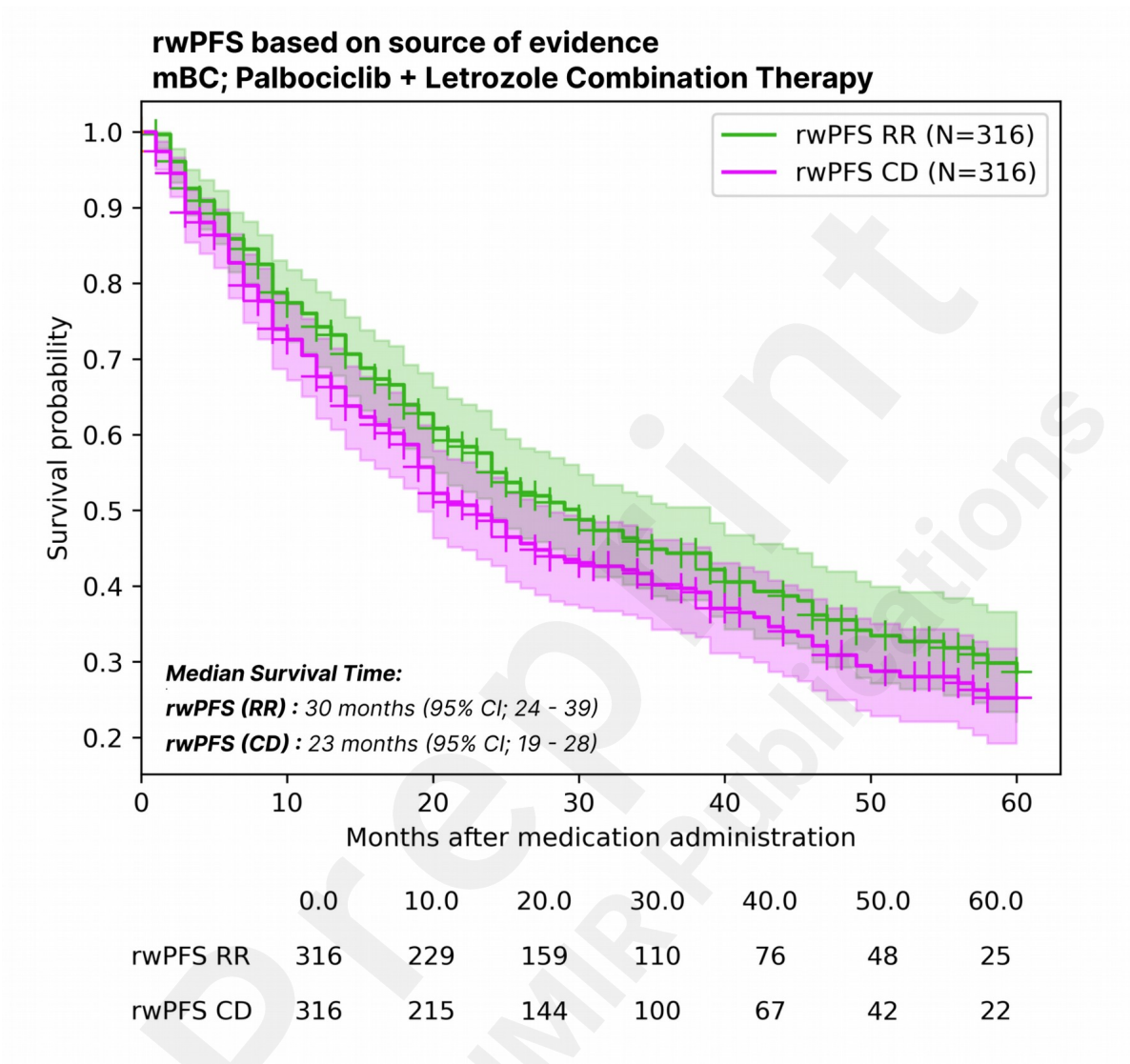


Figure 4. Kaplan-Meier Survival plots for rwPFS based on the patient note source. Survival plots indicating the real-world Progression Free Survival (rwPFS) with progressions captured from solitary sources of Radiology reports (RR) and Routine clinical documents (CD)

Figure 5. Kaplan-Meier Survival Curves for Sub-group Analysis.

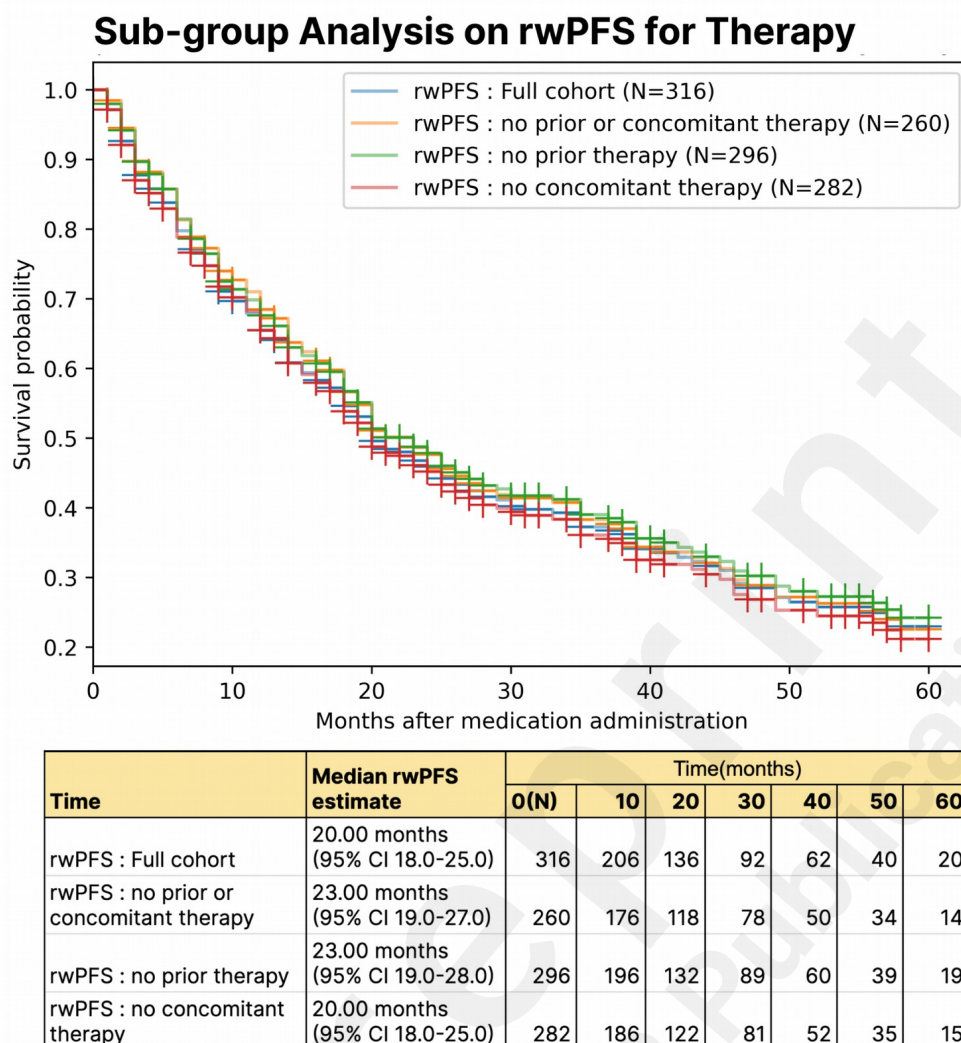


Figure 5. Kaplan-Meier Survival Curves for Sub-group Analysis. Each of the sub-groups account for different variations in treatment patterns. The survival curves and risks table showcase the effect of other prior or concomitant systemic therapies on the median rwPFS.

In the mBC validation set of 100 patients, the median rwPFS was determined to be 25 months (95% CI; 15-35 months) by manual curation and 22 months (95% CI; 15-35 months) by the computational workflow outlined in Figure 3(b). Based on the data source for progression capture, the rwPFS estimated exclusively from radiology reports was 30 months (95% CI; 24.0 - 39.0), compared to 23 months (95% CI; 19.0 - 28.0) when estimated exclusively from clinical documents, as represented in Figure 4. Sub-group analysis on the rwPFS based on prior or concomitant therapies is detailed in Figure 5.

External Dataset Validation

External validation was conducted using data from a partner academic medical center (AMC), representing a health system distinct from the main study, to assess the generalizability and robustness of the progression capture pipeline. At Level-1, manual validation of 200 sampled raw progression captures achieved an accuracy of 92.5% and a F1-score of 92.8%, while Level-2 validation of the first progression in 61 patients reported an accuracy of 90.2% and a F1-score of 92.5%. Two patients were excluded from performance metrics because the textual evidence identified during manual abstraction of the first progression event was unavailable to the automated extraction pipeline. Comprehensive performance metrics, including sensitivity, specificity, and F1 scores, are detailed in Table 5.

Validation Step	Sensitivity	Specificity	Accuracy	Precision	F1 score
-----------------	-------------	-------------	----------	-----------	----------

Level-1; Manual Validation of Sampled Raw Progression Captures (N = 200)	89.7%	95.7%	92.5%	96%	92.8%
Level-2; Manual Validation of First Progression (N = 61)	97.4%	78.3%	90.2%	88.1%	92.5%
Table 5: Manual validation of the progression capture workflow on external dataset.					

Sensitivity Analysis on rwPFS estimates

The sensitivity analysis demonstrated that real-world progression-free survival (rwPFS) estimates were robust under varying conditions. Systematic removal of 10%, 20%, and 30% of progression events resulted in median rwPFS values of 20 months (95% CI: 18–26), 20 months (95% CI: 18–27), 22 months (95% CI: 18–28), and 23 months (95% CI: 19–29) for the complete, 10%, 20%, and 30% datasets, respectively, with widening confidence intervals indicating increased uncertainty. Similarly, median rwPFS estimates were comparable when death was treated as censorship versus a progression event, with values of 20 months (95% CI: 18–25) and 18 months (95% CI: 15–21) in the main dataset. These findings, along with overlapping confidence intervals, indicate that the rwPFS estimates were not meaningfully affected by missing data or event definitions. See Supplementary Figure 5 for Kaplan-Meier curves and event tables.

PHQ-8 Outcomes and Disease Progression

Of the 316 patients, 94 had at least two PHQ-8 scores recorded during the study period, including 30 non-progressed and 64 progressed patients. Non-progressed patients showed greater mean peak reduction (5.57 ± 5.90 vs. 2.95 ± 4.30 , $t = -2.397$, $p = 0.018$) and cumulative decline (8.00 ± 12.68 vs. 3.66 ± 6.40 , $t = -2.201$, $p = 0.03$) in PHQ-8 scores compared to progressed patients. See Supplementary Table 3 for details.

Discussion

The study showcases the development and validation of a novel semi-automated workflow for estimating real-world progression-free survival (rwPFS) in patients with metastatic breast cancer (mBC) using de-identified electronic health records (EHR). One of its key strengths lies in the integration of natural language processing (NLP) techniques to extract clinician-reported progression events from unstructured data sources such as clinical notes and radiology reports, combined with structured patient data like drug orders and clinical encounters. This approach enhances the accuracy and comprehensiveness of capturing progression events, as evidenced by the high sensitivity (99.8%) and specificity (96.7%) at the sentence level with good patient-level accuracy (88%).

While our initial goal was to develop a fully automated workflow for capturing disease progression, we have successfully implemented a semi-automated approach. This is advantageous because the semi-automated method allows for disease-specific adjustments to clinical concept recognition configurations, ensuring relevance across various cancer types. This flexibility underscores the potential for broader applicability beyond mBC, making it a valuable tool for oncological research and real-world evidence generation.

Principal Findings

The median rwPFS of 20 months (95% CI: 18–25) reported in this study is comparable to those reported in previous real-world studies and clinical trial results, validating the workflow's reliability and accuracy.[18,19] Sub-group analysis revealed the impact of prior and other concomitant medications on median rwPFS. For instance, the PALOMA-2 trial reported a median progression-free survival of 24.8 months (95% CI: 22.1 - inf), which is comparable to the real-world observation in the study sub-cohort (N=260) of patients who received palbociclib and letrozole in the first-line metastatic setting (patients with no other prior or concomitant drugs) which was 23.00 months.[20] A matched comparison between the study cohort and other real-world cohorts or clinical trials could further establish the concordance between the survival estimates. Integrating progression events from both radiology reports and routine clinical/oncology notes standardizes the identification of disease progression, mitigating biases and overestimation that can arise from relying exclusively on a single data source. Ablation analysis also revealed the futility of using the subject

sentiment analysis model in the workflow as physicians are unlikely to describe the progression status of a family member or a blood relative in the patient's notes. While this model is useful for extracting other concepts using the clinical NLP engine, it has shown no benefit in its usage for progression capture.

The external validation further demonstrated the robustness and generalizability of the progression capture workflow across health systems. Manual validation on this dataset also achieves high accuracy at sentence level (92.5%) and at capturing the first progression event (90.2%). Sensitivity analysis confirmed that rwPFS estimates were stable, regardless of whether death was treated as censorship or an event, with overlapping confidence intervals observed across both scenarios. Sensitivity analysis confirmed that rwPFS estimates were stable across varying levels of missing data and event definitions, with slight increases in median rwPFS and wider confidence intervals under data incompleteness, and overlapping intervals when treating death as censorship or an event. Furthermore, integration of PHQ-8 outcomes revealed significant associations between patient-reported mental health and progression status, highlighting the potential of patient-reported outcomes to provide complementary insights.

Comparison to Prior Work

The study also highlights the importance of source data used for determining rwPFS. Relying solely on radiology reports overestimated the median rwPFS, compared to estimates derived from both clinical notes and radiology reports combined. The median rwPFS from the pool of free-text data excluding radiology reports (23 months) was closer to the median survival of the overall study cohort with all available pooled free-text data (20 months), when compared to the median rwPFS computed from free-text data exclusively from radiology reports (30 months). This discrepancy can be explained by the observation that patients can undergo radiological evaluations outside the EHR data network, with their findings being documented by treating physicians within the EHR network in patients' routine clinical notes. Similar findings were observed in a previous study that analyzed the impact of source data on real-world survival estimates.[6] Additionally, relying solely on structured data like drug records (time to discontinuation or time to next treatment) as a surrogate for rwPFS has been shown to underestimate the median rwPFS substantially in a prior study.[21]. Patient-reported outcomes (PROs) provide direct insight into a patient's symptoms and quality of life and have been linked to progression-free and overall survival in prior studies. Although direct comparisons with alternative workflows were not performed, our method demonstrates performance metrics that are in line with those reported in previous studies, warranting further comparative analyses in future work.

Among other computational techniques for characterizing cancer response in real-world data(RWD), the use of large language models (LLMs) has also shown promise. A prior study evaluating this has shown GatorTron to be the best-performing model, achieving an accuracy of 89% at the radiology report level upon fine-tuning.[22] However, applying LLMs across a broader patient corpus needs further investigation to fully ascertain their validity and generalizability. Patient-reported outcomes (PROs) provide direct insight into a patient's symptoms and quality of life and have been linked to progression-free and overall survival in prior studies.[24-26] We have observed significant associations between the decline in PHQ-8 scores and patient's progression status.

Limitations

There are however limitations to this study. First, the reliance on clinician-reported events means that the accuracy of the workflow is reliant on the quality and completeness of clinical documentation. Incomplete or inconsistent documentation could lead to underestimation or overestimation of progression events. To mitigate this, careful validation of extraction patterns and data completeness checks were implemented. Second, although the semi-automated workflow reduces the resource-intensive nature of manual abstraction, it requires initial manual rule-definition and configuration, which could introduce biases based on the selected rules and criteria. Representative evaluation samples were curated across the breast cancer cohorts to reduce the biases. Third, sensitivity analyses were limited primarily to variations in clinical text sources and censoring definitions. Expanding sensitivity analyses to include demographic factors, alternative definitions of progression, and data-source reliability could further strengthen the robustness of the findings. Fourth, augmenting the mortality data with commercial and federal death registries could enhance the accuracy of survival estimates. This was not feasible in the current analysis but represents an important area for future improvement. Fifth, integration of patient-reported outcomes could provide a more comprehensive understanding of patient well-being in relation to progression events. However, demonstrating this in real-world data was challenging due to limited availability of patient-reported records. Finally, the ensemble deep-learning engine's performance was evaluated within a specific cohort of mBC patients; thus, further validation across more diverse external datasets and different cancer types is necessary to truly establish the generalizability of the workflow.

Future Direction

These findings align with the growing body of research advocating for integrating AI and machine learning in healthcare data analysis, as these technologies can substantially enhance the speed, accuracy, and breadth of data processing capabilities. Future work will explore more advanced text data analysis and extraction methods such as adaptive machine learning techniques and large language models to minimize manual import and enhance scalability. Furthermore, by employing federated learning, insights and patterns from diverse populations across various institutions can be pooled securely, enriching the model's generalizability and performance across different healthcare settings. The successful implementation of this automated workflow demonstrates its potential to streamline the data extraction process from EHRs from various health systems. It also paves the way for its application in other oncological studies, where similar challenges in data abstraction exist.

Conclusion

Developing a practical and scalable method for capturing real-world progression from EHR data is crucial to improving oncological research and patient care. Overall, this technology represents a step forward in realizing the full potential of EHR data in oncology. Our findings establish a workflow for automated data capture to provide a more efficient, and scalable method than traditional manual processes, particularly in handling complex, unstructured EHR data. Although the principles of progression capture remain the same across other cancer types, further research across other types of solid tumors is needed to ascertain the generalizability of the workflow.

Competing Interest

No competing interest in the findings of the study.

Funding

No external funding was obtained for conduct of the study.

References

1. Klonoff DC. The New FDA Real-World Evidence Program to Support Development of Drugs and Biologics. *J Diabetes Sci Technol* 2020 Mar;14(2):345–349. PMID:30862182
2. Feinberg BA, Gajra A, Zettler ME, Phillips TD, Phillips EG Jr, Kish JK. Use of Real-World Evidence to Support FDA Approval of Oncology Drugs. *Value Health* 2020 Oct;23(10):1358–1365. PMID:33032780
3. Khozin S, Blumenthal GM, Pazdur R. Real-world Data for Clinical Evidence Generation in Oncology. *J Natl Cancer Inst* 2017 Nov 1;109(11). PMID:29059439
4. Grimaldi S, Terroir M, Caramella C. Advances in oncological treatment: limitations of RECIST 1.1 criteria. *Q J Nucl Med Mol Imaging* 2018 Jun;62(2):129–139. PMID:29166754
5. Eisenhauer EA, Therasse P, Bogaerts J, Schwartz LH, Sargent D, Ford R, Dancey J, Arbuck S, Gwyther S, Mooney M, Rubinstein L, Shankar L, Dodd L, Kaplan R, Lacombe D, Verweij J. New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1). *Eur J Cancer* 2009 Jan;45(2):228–247. PMID:19097774
6. Griffith SD, Tucker M, Bowser B, Calkins G, Chang C-HJ, Guardino E, Khozin S, Kraut J, You P, Schrag D, Miksad RA. Generating Real-World Tumor Burden Endpoints from Electronic Health Record Data: Comparison of RECIST, Radiology-Anchored, and Clinician-Anchored Approaches for Abstracting Real-World Progression in Non-Small Cell Lung Cancer. *Adv Ther* 2019 Aug;36(8):2122–2136. PMID:31140124
7. Meyers DE, Stukalin I, Vallerand IA, Lewinson RT, Suo A, Dean M, North S, Pabani A, Cheng T, Heng DYC, Bebb DG, Morris DG. The Lung Immune Prognostic Index Discriminates Survival Outcomes in Patients with Solid Tumors Treated with Immune Checkpoint Inhibitors. *Cancers* 2019 Nov 2;11(11). PMID:31684111
8. Griffith SD, Miksad RA, Calkins G, You P, Lipitz NG, Bourla AB, Williams E, George DJ, Schrag D, Khozin S, Capra WB, Taylor MD, Abernethy AP. Characterizing the Feasibility and Performance of Real-World Tumor Progression End Points and

- Their Association With Overall Survival in a Large Advanced Non–Small-Cell Lung Cancer Data Set. *JCO Clin Cancer Inform* Wolters Kluwer; 2019 Dec 1;(3):1–13.
9. Amorrtu R, Garcia M, Zhao Y, El Naqa I, Balagurunathan Y, Chen D-T, Thieu T, Schabath MB, Rollison DE. Overview of approaches to estimate real-world disease progression in lung cancer. *JNCI Cancer Spectr* 2023 Oct 31;7(6). PMID:37738580
 10. Torres AZ, Nussbaum NC, Parrinello CM, Bourla AB, Bowser BE, Wagner S, Tabano DC, George D, Miksad RA. Analysis of a Real-World Progression Variable and Related Endpoints for Patients with Five Different Cancer Types. *Adv Ther* 2022 Jun;39(6):2831–2849. PMID:35430670
 11. Suárez-Paniagua V, Dong H, Casey A. A multi-BERT hybrid system for named entity recognition in Spanish radiology reports. *CLEF 2021 – Conference and Labs of the Evaluation Forum 2021*. p. 846–856.
 12. Bai L, Mulvenna MD, Wang Z, Bond R. Clinical Entity Extraction: Comparison between MetaMap, cTAKES, CLAMP and Amazon Comprehend Medical. 2021 32nd Irish Signals and Systems Conference (ISSC) IEEE; 2021. p. 1–6.
 13. Shah-Mohammadi F, Finkelstein J. A Comparison of Automated Named Entity Recognition Tools Applied to Clinical Text. 2022 7th International Conference on Intelligent Informatics and Biomedical Science (ICIIBMS) IEEE; 2022. p. 227–231.
 14. Yang X, Chen A, PourNejatian N, Shin HC, Smith KE, Parisien C, Compas C, Martin C, Flores MG, Zhang Y, Magoc T, Harle CA, Lipori G, Mitchell DA, Hogan WR, Shenkman EA, Bian J, Wu Y. GatorTron: A Large Clinical Language Model to Unlock Patient Information from Unstructured Electronic Health Records. *arXiv [csCL]*. 2022. Available from: <http://arxiv.org/abs/2203.03540>
 15. Murugadoss K, Rajasekharan A, Malin B, Agarwal V, Bade S, Anderson JR, Ross JL, Faubion WA Jr, Halamka JD, Soundararajan V, Ardhanari S. Building a best-in-class automated de-identification tool for electronic health records through ensemble learning. *Patterns (N Y)* 2021 Jun 11;2(6):100255. PMID:34179842
 16. Wagner T, Shweta F, Murugadoss K, Awasthi S, Venkatakrishnan AJ, Bade S, Puranik A, Kang M, Pickering BW, O'Horo JC, Bauer PR, Razonable RR, Vergidis P, Temesgen Z, Rizza S, Mahmood M, Wilson WR, Challener D, Anand P, Liebers M, Doctor Z, Silvert E, Solomon H, Anand A, Barve R, Gores G, Williams AW, Morice WG 2nd, Halamka J, Badley A, Soundararajan V. Augmented curation of clinical notes from a massive EHR system reveals symptoms of impending COVID-19 diagnosis. *Elife* 2020 Jul 7;9. PMID:32633720
 17. Devlin J, Chang M-W, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* Minneapolis, Minnesota: Association for Computational Linguistics; 2019. p. 4171–4186.
 18. DeMichele A, Cristofanilli M, Brufsky A, Liu X, Mardekian J, McRoy L, Layman RM, Emir B, Torres MA, Rugo HS, Finn RS. Comparative effectiveness of first-line palbociclib plus letrozole versus letrozole alone for HR+/HER2- metastatic breast cancer in US real-world clinical practice. *Breast Cancer Res* 2021 Mar 24;23(1):37. PMID:33761995
 19. Wilkie J, Schickli MA, Berger MJ, Lustberg M, Reinbolt R, Noonan A, Ramaswamy B, Sardesai S, VanDeusen J, Wesolowski R, Williams N, Stover DG, Li J, Vargo CA. Progression-Free Survival for Real-World Use of Palbociclib in Hormone Receptor-Positive Metastatic Breast Cancer. *Clin Breast Cancer* 2020 Feb;20(1):33–40. PMID:31451366
 20. Finn RS, Martin M, Rugo HS, Jones S, Im S-A, Gelmon K, Harbeck N, Lipatov ON, Walshe JM, Moulder S, Gauthier E, Lu DR, Randolph S, Diéras V, Slamon DJ. Palbociclib and Letrozole in Advanced Breast Cancer. *N Engl J Med* 2016 Nov 17;375(20):1925–1936. PMID:27959613
 21. Walker MS, Herms L, Miller PJE. Performance of time to discontinuation and time to next treatment as proxy measures of progression-free survival, overall and by treatment group. *J Clin Orthod* Wolters Kluwer; 2020 May 20;38(15_suppl):e19135–e19135.
 22. Tan RSYC, Lin Q, Low GH, Lin R, Goh TC, Chang CCE, Lee FF, Chan WY, Tan WC, Tey HJ, Leong FL, Tan HQ, Nei WL, Chay WY, Tai DWM, Lai GGY, Cheng LT-E, Wong FY, Chua MCH, Chua MLK, Tan DSW, Thng CH, Tan IBH, Ng HT. Inferring cancer disease response from radiology reports using large language models with data augmentation and prompting. *J Am Med Inform Assoc* 2023 Sep 25;30(10):1657–1664. PMID:37451682
 23. Beltagy I, Lo K, Cohan A. SciBERT: A pretrained language model for scientific text. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* Stroudsburg, PA, USA: Association for Computational Linguistics; 2019. doi: 10.18653/v1/d19-1371
 24. Bhatt AS, Schabath MB, Hoogland AI, Jim HSL, Brady-Nicholls R. Patient-reported outcomes as interradiographic predictors of response in non-small cell lung cancer. *Clin Cancer Res* 2023 Aug 15;29(16):3142–3150. PMID:37233986

25. Balitsky AK, Rayner D, Britto J, Lionel AC, Ginsberg L, Cho W, Wilfred AM, Sardar H, Cantor N, Mian H, Levine MN, Guyatt GH. Patient-reported outcome measures in cancer care: An updated systematic review and Meta-analysis. JAMA Netw Open American Medical Association (AMA); 2024 Aug 1;7(8):e2424793. PMID:39136947
26. Badaoui S, Kichenadasse G, Rowland A, Sorich MJ, Hopkins AM. Patient-reported outcomes predict progression-free survival of patients with advanced breast cancer treated with abemaciclib. Oncologist Oxford University Press (OUP); 2021 Jul;26(7):562–568. PMID:33914991

Figure Legends

Figure 1. Methodology flow diagram illustrating the workflow: (a) Illustrates the workflow for real-world progression(rwP) extraction and determining the rwPFS. (b) Illustrates the methodology for capturing progression from unstructured texts in routine clinical documents and radiology reports using Nference's clinical NLP engine that performs clinical concept recognition, association, and sentiment analysis.

Figure 2. Cohort attrition diagram: Structured codes 174*(ICD-9), and C50* (ICD-10) or >4 positive disease sentiments from the augmented curation disease diagnosis model were used for breast cancer. For evidence of metastasis 197*, 198* (ICD-9), C78*, and C79* (ICD-10) in conjunction with augmented curation was used; *(single asterisk) represents all the children codes within the parent code. **(double asterisk) NLP - Natural Language Processing.

Figure 3. Kaplan-Meier Survival Plots for the overall study cohort and validation sets: (a) Kaplan-Meier Survival plots indicating the real-world Progression Free Survival (rwPFS) and real-world Overall Survival (rwOS) in the study cohort of patients with metastatic breast cancer using pooled note sources. (b) Patient-level validation of first progression capture and comparing outcomes estimated by computational workflow with manual curation.

Figure 4. Kaplan-Meier Survival plots for rwPFS based on the patient note source. Survival plots indicating the real-world Progression Free Survival (rwPFS) with progressions captured from solitary sources of Radiology reports (RR) and Routine clinical documents (CD)

Figure 5. Kaplan-Meier Survival Curves for Sensitivity Analysis. Each of the sub-groups account for different variations in treatment patterns. The survival curves and risks table showcase the effect of other prior or concomitant systemic therapies on the median rwPFS.

Author Contributions

GV led the data curation, formal analysis, visualization, and writing of the original draft, writing review, and editing, and contributed to the methodology. RKY contributed to the methodology, data curation, formal analysis, investigation, and visualization. PK was responsible for the conceptualization, methodology, lead investigation, funding acquisition, lead supervision, and writing review and editing. BSA contributed to data curation and validation. KP contributed to data curation, methodology, and investigation. SC contributed to data curation and validation. SKR contributed to data curation and validation. VK contributed to data curation and validation. AvA developed the clinical NLP engine software for NLP models inference. MM contributed to data curation, investigation, and validation. SJ developed the software. Aka provided software development and feedback. RB provided feedback and project administration. VT provided resources, feedback, and project administration. PL contributed to validation, feedback, and writing review and editing. SAS provided resources, feedback, and writing review and editing. VS provided supervision. All authors read and approved the final manuscript.

Competing Interests

The authors declare no competing interests.

Data Availability

Any data beyond what is reported in the manuscript, including but not limited to the raw EHR data, cannot be shared or released due to the parameters of the expert determination to maintain the data de-identification. Contact corresponding authors for additional details regarding the nference nSights Analytics Platform.

Code Availability

The underlying code for this study [and training/validation datasets] is not publicly available for proprietary reasons.

Appendix-1

Supplementary File contains a description about the clinical NLP engine, its developmental process, and validation metrics. Additionally, five supplementary figures are enclosed. Supplementary-Figure 1 describes the entity identification and labelling by the computational workflow. Supplementary-Figure 2 describes the framework of ablation analysis, and Supplementary-Figure 3 provides a high-level overview of the entire workflow. Supplementary Figure 4 describes the cohort attrition for the external validation dataset and Supplementary Figure 5 describes the sensitivity analysis on rwPFS estimates.

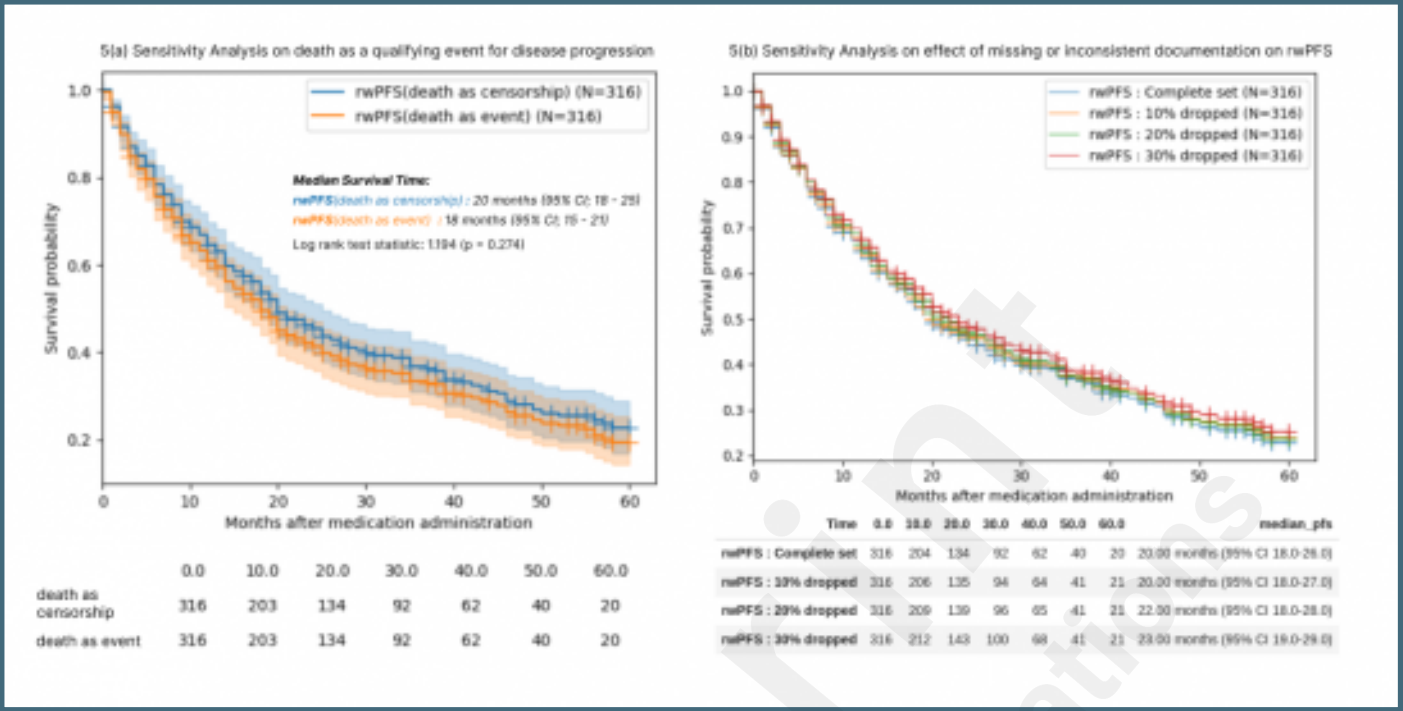
Abbreviations

- **AI:** Artificial Intelligence
- **BERT:** Bidirectional Encoder Representations from Transformers
- **CD:** Clinical Documents
- **CI:** Confidence Interval
- **DICOM:** Digital Imaging and Communications in Medicine
- **ECOG:** Eastern Cooperative Oncology Group
- **EHR:** Electronic Health Record
- **HER2:** Human Epidermal Growth Factor Receptor 2
- **HIPAA:** Health Insurance Portability and Accountability Act
- **HR:** Hormone Receptor
- **IQR:** Interquartile Range
- **LLM:** Large Language Model
- **mBC:** Metastatic Breast Cancer
- **NER:** Named Entity Recognition
- **NLP:** Natural Language Processing
- **PHI:** Protected Health Information
- **PHQ-8:** Patient Health Questionnaire -8
- **PFS:** Progression-Free Survival
- **RECIST:** Response Evaluation Criteria in Solid Tumors
- **RWD:** Real-World Data
- **RWE:** Real-World Evidence
- **rwP:** Real-World Progression
- **rwPFS:** Real-World Progression-Free Survival
- **rwOS:** Real-World Overall Survival

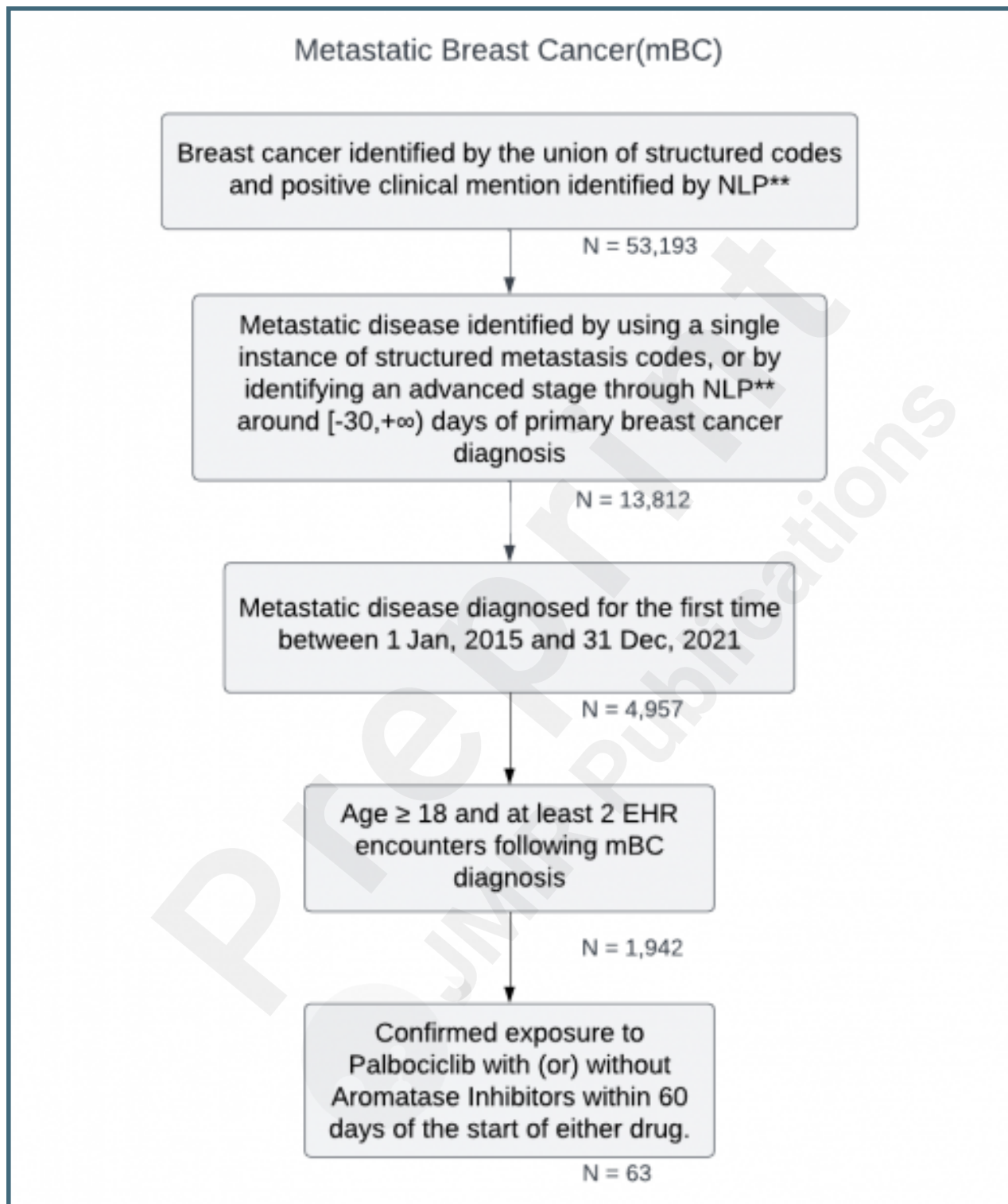
Supplementary Files

Figures

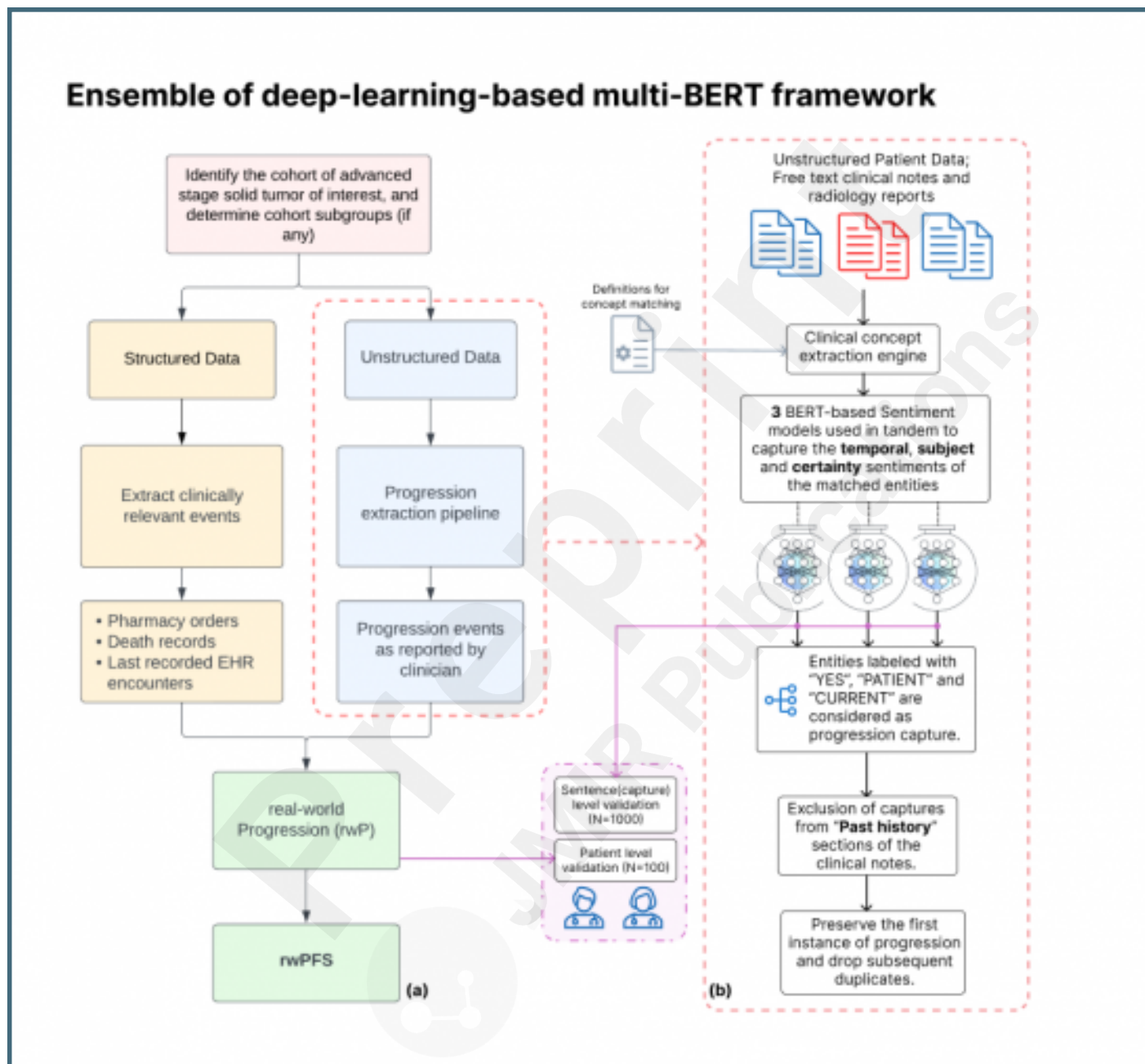
Supplementary Figure 5, Sensitivity Analysis.



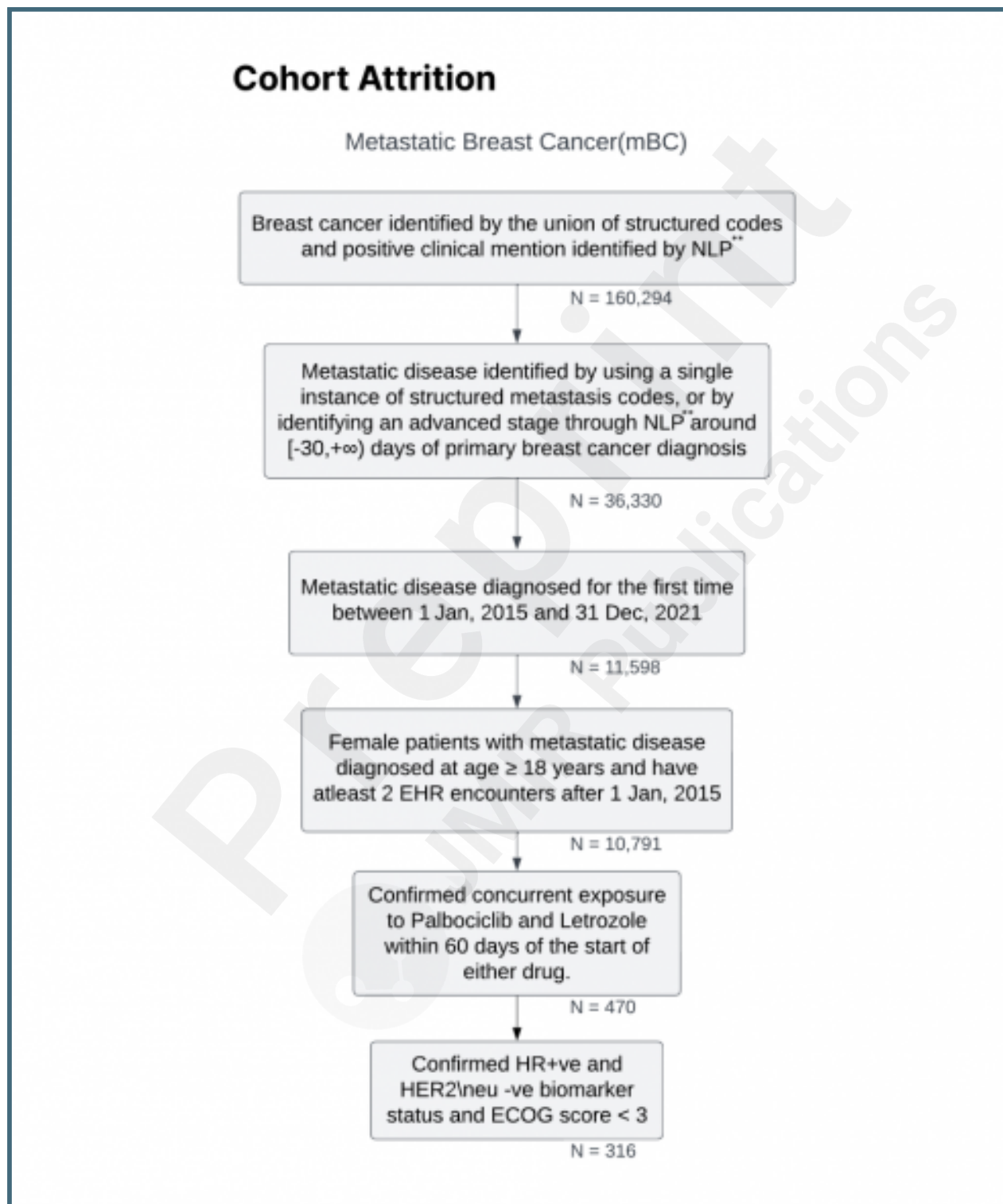
Cohort Attrition of patients in the External validation Dataset.



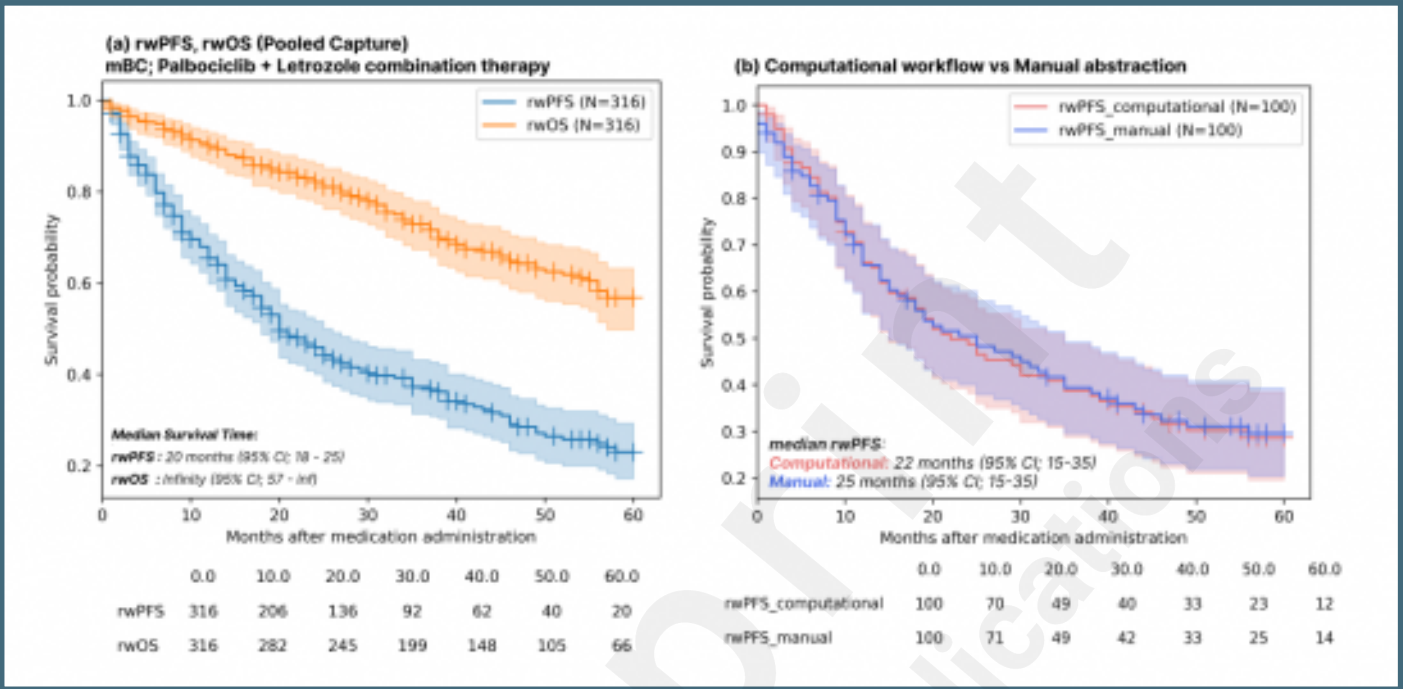
Methodology flow diagram illustrating the workflow: (a) Illustrates the workflow for real-world progression (rwP) extraction and determining the rwPFS. (b) Illustrates the methodology for capturing progression from unstructured texts in routine clinical documents and radiology reports using Nference's clinical NLP engine that performs clinical concept recognition, association, and sentiment analysis.



Cohort attrition diagram: Structured codes 174*(ICD-9), and C50* (ICD-10) or >4 positive disease sentiments from the augmented curation disease diagnosis model were used for breast cancer. For evidence of metastasis 197*, 198* (ICD-9), C78*, and C79* (ICD-10) in conjunction with augmented curation was used; *(single asterisk) represents all the children codes within the parent code. **(double asterisk) NLP - Natural Language Processing.



Kaplan-Meier Survival Plots for the overall study cohort and validation sets: (a) Kaplan-Meier Survival plots indicating the real-world Progression Free Survival (rwPFS) and real-world Overall Survival (rwOS) in the study cohort of patients with metastatic breast cancer using pooled note sources. (b) Patient-level validation of first progression capture and comparing outcomes estimated by computational workflow with manual curation.

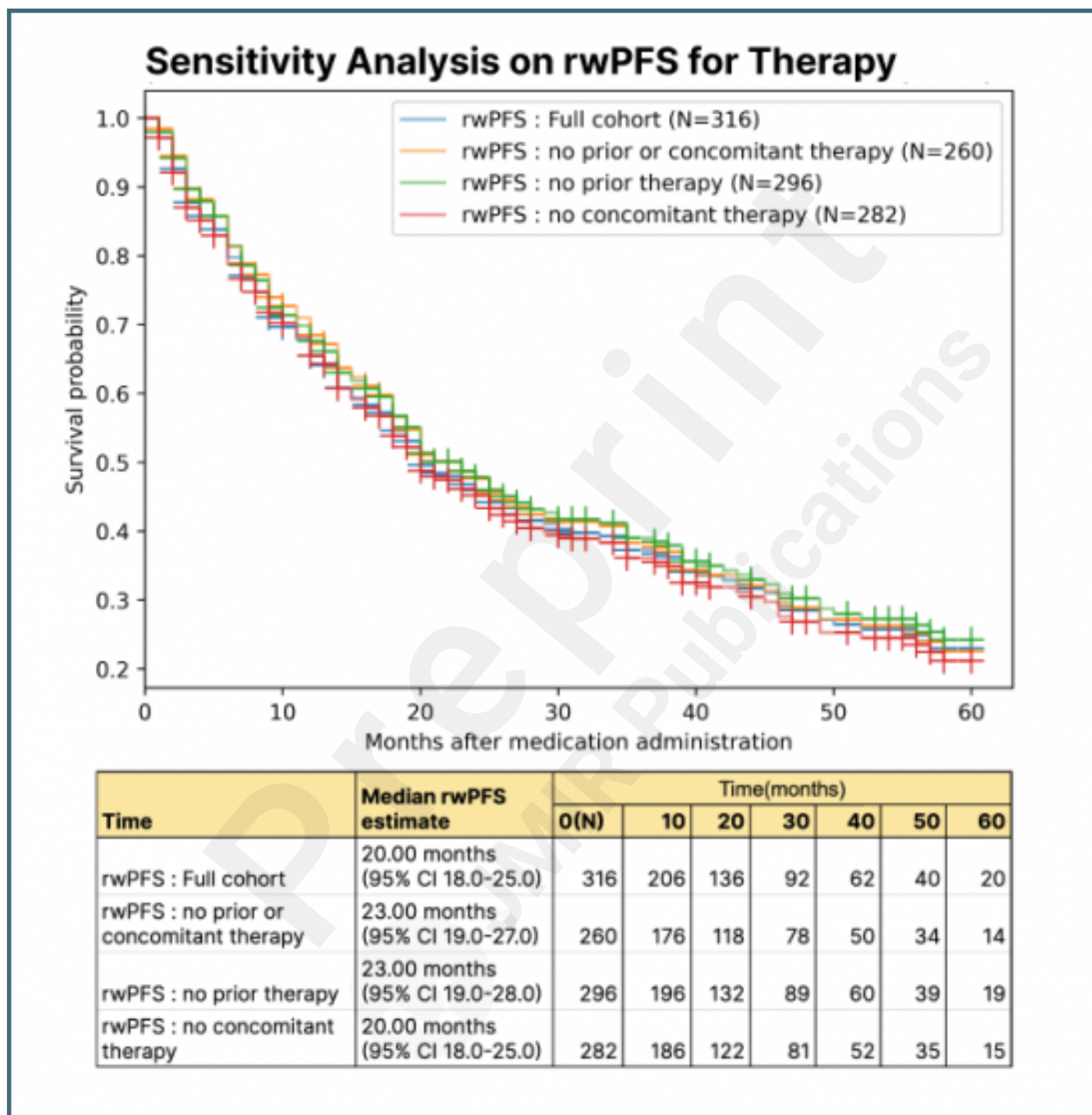


rwPFS based on source of evidence
mBC; Palbociclib + Letrozole Combination Therapy

Median Survival Time:
rwPFS (RR) : 30 months (95% CI; 24 - 39)
rwPFS (CD) : 23 months (95% CI; 19 - 28)

	0.0	10.0	20.0	30.0	40.0	50.0	60.0
rwPFS RR	316	229	159	110	76	48	25
rwPFS CD	316	215	144	100	67	42	22

Kaplan-Meier Survival Curves for Sensitivity Analysis. Each of the sub-groups account for different variations in treatment patterns. The survival curves and risks table showcase the effect of other prior or concomitant systemic therapies on the median rwPFS.



Multimedia Appendixes

Supplementary Note with Tables and Figures.

URL: <http://asset.jmir.pub/assets/bd0473c2f8cf4e8bb7287f78fcacb861.docx>



CONSORT (or other) checklists

RECORD checklist.

URL: <http://asset.jmir.pub/assets/a99ecf9a92c9bb1e642c0c3b04bb980e.pdf>