ARTICLE TYPE: Original Article

TITLE: Machine learning and natural language processing (NLP) approach to predict early progression to first-line treatment in real-world hormone receptor-positive (HR+)/HER2-negative advanced breast cancer patients

AUTHORS:

N. Ribelles[1#], J.M. Jerez[2#], P. Rodriguez-Brazzarola[2], B. Jimenez[1], T. Diaz-Redondo[1], H. Mesa[2], A. Marquez [1], A. Sanchez-Muñoz[1], B. Pajares[1], F. Carabantes[1], M.J. Bermejo[1], E. Villar[1], M.E. Dominguez-Recio[1], E. Saez[1], L. Galvez [1], A. Godoy[1], L. Franco [2], S. Ruiz-Medina[1], I. Lopez [1], E. Alba [1]

[1]Medical Oncology Intercenter Unit. Regional and Virgen de la Victoria University Hospitals. IBIMA. Málaga, Spain

[2]University of Málaga, Department of Languages and Computer Science, E.T.S.I. Computing. Málaga. Spain.

[#]Contributed equally to this article.

ABSTRACT:

Background: CDK4/6 inhibitors plus endocrine therapies are the current standard of care in the first-line treatment of HR+/HER2-negative metastatic breast cancer, but there are no well-established clinical or molecular predictive factors for patient response. In the era of personalized oncology, new approaches for developing predictive models of response are needed.

Materials and Methods: Data derived from the electronic health records (EHRs) of real-world patients with HR+/HER2-negative advanced breast cancer were used to develop predictive models for early and late progression to first-line treatment. Two machine learning approaches were used: a classic approach using a data set of manually extracted features from reviewed (EHR) patients, and a second approach using natural language processing (NLP) of free-text clinical notes recorded during medical visits.

Results: Of the 610 patients included, there were 473 (77.5%) progressions to first-line treatment, of which 126 (20.6%) occurred within the first 6 months. There were 152 patients (24.9%) who showed no disease progression before 28 months from the onset of first-line treatment. The best predictive model for early progression using the manually extracted data set achieved an area under the curve (AUC) of 0.734 (95% CI 0.687-0.782). Using the NLP free-text processing approach, the best model obtained an AUC of 0.758 (95% CI 0.714-0.800). The best model to predict long responders using manually extracted data obtained an AUC of 0.669 (95% CI 0.608-0.730). With NLP free-text processing, the best model attained an AUC of 0.752 (95%CI 0.705-0.799).

Conclusions: Using machine learning methods, we developed predictive models for early and late progression to first-line treatment of HR+/HER2-negative metastatic breast cancer, also finding that NLP-based machine learning models are slightly better than predictive models based on manually obtained data.

INTRODUCTION.

The treatment of HR+/HER2-negative metastatic breast cancer has drastically improved with the approval of cyclin-dependent kinase (CDK) 4 and 6 inhibitors. Data from pivotal trials assessing combinations of different CDK4/6 inhibitors plus endocrine therapies (ET) have shown concordant results with regard to progression-free survival (PFS) in the first-line treatment of both postmenopausal [1-6] and premenopausal [7-9] patients. The addition of CDK4/6 inhibitors increases the median PFS in the first-line setting by 60–90% compared to that obtained with ET alone. This benefit was found to be maintained in all subgroups analyzed according to the location of the disease, previous neoadjuvant and adjuvant treatments, number of metastatic sites, age, ECOG or time elapsed from the end of hormone adjuvant treatment until the diagnosis of distant recurrence.

All of the variables mentioned above are considered prognostic factors for overall survival in patients with metastatic breast cancer, as verified in different series of patients [10-15] Although it could be assumed that these prognostic factors are also prognostic for PFS in first-line treatment for metastatic disease, there are no specific data to confirm this point. In fact, available data indicates that PFS and overall survival are only moderately correlated in metastatic breast cancer [16]. Furthermore, it is possible that other variables collected in medical records could be useful for the definition of PFS prognostic patient subgroups.

In the current context of advanced breast cancer management, treatments should be personalized and indicate a certain treatment in those patients with the greatest expected benefit. Therefore, it is of considerable interest to define different subgroups of patients with different risks of progression to first-line treatment for metastatic disease in order to be able to select the best therapeutic option with the greatest accuracy and efficiency.

In recent years, artificial intelligence (AI) has begun revolutionizing several industries, including healthcare. Healthcare delivery organizations have invested a considerable amount of time and effort in the development of AI driven medical tools and research. The goal of applying machine learning, a branch of AI, is to identify patterns in data in order to find a model that best generalizes beyond the data seen. Although it is intimately connected to traditional statistical methods, machine learning often seeks nonlinear relationships among the independent variables. In the traditional approach of data analysis one begins with a statistical model and the data as input to the computation, whereas machine learning differs since it is a data driven approach that generalizes a model from the data in order to obtain a model that can be applied to new data [17].

Electronic health records (EHRs) include large amounts of data from real-world patients collected during regular clinical practice. Although most of these data are recorded in an unstructured text form, the application of machine learning techniques enables the implementation of algorithms to classify features or predict events based on these clinical text notes. Moreover, it is possible to obtain information to identify patients with higher sensitivity and specificity, even more so than that obtained from structured data [18, 19]. Furthermore, these techniques have been applied to diagnostic specialties [20, 21] and to predict events in cancer patients [22-24].

Our aim was to develop a machine learning model to predict early progression to first-line treatment in metastatic HR+/HER2-negative breast cancer through natural language processing (NLP)-based analysis of free-text clinical notes from EHRs and predict the risk according to the different treatments used in this setting. In addition, we evaluate if the performance of such model is at least the same as that obtained through the traditional approach using information structured within a database. To further verify the validity of this approach, we also set out to develop a predictive model for

long-term responders to first-line treatment, as this is also an important issue to consider when deciding initial therapy.

MATERIALS AND METHODS

Data source and patient selection

This is an observational and longitudinal study in which the data were derived from the EHRs of patients with metastatic HR+/HER2-negative breast cancer treated at Hospital Regional Universitario and Hospital Universitario Virgen de la Victoria in Malaga (Spain). Both hospitals use the same information system called Galen, comprising a database of patients and their EHRs, among other utilities [25]. The EHRs from patients who had received at least one line of treatment between 1991 and 2019 were identified from the database, and those patients with clinical notes in their EHR were included in the study. The EHR contains both structured and unstructured data. The structured fields comprise demographic data, first symptom date, first diagnosis date, tumor characteristics at initial breast cancer diagnosis (histology, tumor size, nodal status, stage), first treatment date, type and intention of first treatment, last control date and last control status. The unstructured data consist of free-text clinical notes recorded by the oncologist at each medical visit. In the development of the machine learning model, all the EHR clinical notes collected up to the start date of the first-line treatment were used, because this information is available in clinical practice when deciding the best therapeutic option for a patient.

Patient data were deidentified. This study was approved by the local research ethics committee.

Outcomes of interest

Two outcomes of interest were considered. The first one was early progression within 6 months after the first-line treatment. The second outcome was late progression after 28

months of starting the first-line treatment. This cut-off was chosen because it was the median PFS reported in a pooled analysis of patients treated with CDK4/6 inhibitors plus aromatase [26].

Expert-reviewed patient data set

Two certified medical oncologists (NR, BJ) manually reviewed deidentified EHR from selected patients. This data set includes 43 variables covering demographic data, first symptom date, first diagnosis date, tumor characteristics at initial breast cancer diagnosis, first treatment date, details of first treatment, recurrence date, recurrence disease characteristics, first-line treatment, second disease progression date, last control date and last control status (Supplementary Table S1). First-line treatments were categorized into four groups: chemotherapy (CT) alone, CT plus maintenance ET, ET alone, and ET plus CDK4/6 inhibitors. This data set was used to develop the predictive model through the classic approach. It was also used in a supervised manner to verify the accuracy of the predictive machine learning model developed from the unstructured free text.

Model building and validation

The first step was to collect the Spanish free-text clinical notes to generate a text corpus (i.e., a large structured set of texts) from the information stored within the Galen system.

We proceeded to the text processing step after having gathered a corpus of Spanish documents. This step comprises the following tasks: (1) remove irrelevant words and characters, (2) convert all characters to lowercase, (3) combine some misspelled or alternately spelled words into a single representation, and (4) stemming. This last step reduces inflectional and sometimes derivationally related forms of a word into a common base form by removing suffixes or prefixes used with a word.

Subsequently, we needed to represent these preprocessed text documents with a numeric representation as machine learning models take numerical values as input. To achieve this, we proceeded to build a vocabulary of all the unique words in our data set and associate a unique index to each term. Thus, each text document was represented as a list of indexes as long as the number of distinct words in the text. These lists were used to generate a Document-Term Matrix (DTM), used to store a statistical measure that represents the relevance of a word to a document. This numerical representation of this text ignores the order of words in the documents and is known as a BoW model [27].

Having populated the BoW model with one of the previous statistical measures, we arrived at one of the challenges of this modelling approach: the vast number of features. Accordingly, we applied different feature selection methods to reduce the dimensionality of our dataset. The selected methods applied were (1) Analysis of Variance (ANOVA), (2) Levene's Test, (3) Correlation-based Feature Selection (CFS), and (4) Principal Component Analysis (PCA).

Afterwards, we proceeded to feed the filtered dataset into fold cross validation of 10 folds in order to perform a robust estimation of the prediction error because in real-world problems, they cannot be exactly calculated. This technique divides the dataset into k folds, creates a classifier using k-1 fold for training, and an error value is calculated by testing the classifier in the remaining fold. Then the error is estimated by taking the average value of the error for each fold [28]. Thus, enabling the performance assessment of the following machine learning algorithms: Naive Bayes (NB), Linear Discriminant Analysis (LDA), Decision Trees (DT), Support Vector Machines (SVM), Lasso Regression, Ridge Regression, Elastic Net, Generalized Linear Boosting (GLMBoost), Adaptive Boosting (ADA), Gradient Boosting Machine (GBM), Bayesian Additive Regression Trees (BART) and Random Forests (RF) [29].

To evaluate the quality of these estimates, we applied a robust inference based on resampling methods [30] in order to obtain confidence intervals for the cross-validation AUC results. To establish statistically significant differences between the models, the non-parametric Wilcoxon signed-rank test was employed to compare the distribution of pairwise cross-validation AUC values [31].

We performed all statistical analysis using R version 3.6.1 and Python version 3.7.6.

RESULTS

Cohort characteristics

Of the 665 patients diagnosed with HR+/HER2-negative advanced breast cancer during the study period, 55 were excluded from the study because the follow-up period after the onset of first-line treatment was shorter than 6 months. Thus, 610 patients were included in the final analysis, with a total of 17,426 clinical visits from which free-text notes were collected. The median follow-up for metastatic disease of the whole cohort was 32.2 months. When considering each group, the mean follow-up period was 39.8 months for ET treatment, 36.3 months for CT plus maintenance ET, 19.7 months for CT alone and 18.7 months for ET plus CDK4/6 inhibitors.

The mean age of patients was 52 years (range 22–89 years) and 23.4% were classified as stage IV at diagnosis (Table 1). Regarding the classification of tumors, 19.5% were luminal A due to a Ki67 value <14%, 37.2% were luminal B, and the Ki67 value was unknown in 43.3% of cases. Approximately half (51%) of the patients were treated with ET, 19.5% received CT alone, 19.2% were treated with CT plus maintenance ET and 10.3% with ET plus CDK4/6 inhibitors.

There were 473 (77.5%) progressions to first-line treatment, with 126 (20.6%) occurring within the first 6 months. Of these early progressions, 57 patients had received ET (9.3%), 54 CT (8.9%), 4 CT plus ET maintenance (0.7%) and 11 ET plus CDK4/6 inhibitors (1.8%).

There were 152 long-responder patients (24.9%) in our cohort, who did not show disease progression earlier than 28 months from the onset of first-line treatment. Of these patients, 103 had been treated with ET (16.9%), 9 with CT (1.5%), 37 with CT plus maintenance ET (6.1%) and 3 with ET plus CDK4/6 inhibitors (0.5%).

Model performance for early progressions

Using the data set of manually extracted features of the reviewed patients, the model that yielded the best result was Elastic Net, which showed an AUC of 0.734 (95%CI 0.687-0.782). With the NLP free-text processing approach, the best model was GLMBoost, achieving an AUC of 0.758 (95%CI 0.714-0.800) (Table 2). Figure 1 shows an example of the report obtained when predicting the early progressions for each of the proposed treatments.

Model performance for long responders

The algorithm that showed best performance for predicting long responses with the data set of manually extracted patient features was Elastic Net, which obtained an AUC of 0.669 (95% CI 0.608-0.730). With the NLP free-text processing approach, GBM was the most relevant algorithm, attaining an AUC of 0.752 (95%CI 0.705-0.799) (Table 2). Figure 2 shows an example of the report obtained when predicting long-duration responses.

DISCUSSION

The treatment of HR+/HER2-negative metastatic breast cancer has evolved substantially in recent years, and will certainly continue to do so in the future [32, 33]. However, we do not currently have robust predictive factors to help choose the best treatment in a specific patient in daily practice, and attempts to identify predictive molecular factors have not yielded the desired results [34, 35]. In this context, new approaches for developing predictive models of patients' response to different available therapeutic options are needed.

The innovation and development of artificial intelligence medical tools and research are able to provide effective and efficient predictive models that improve the generalization capacity of traditional statistical models since they are capable of capturing complex relationships between the prognostic factors and the study variables. However, although these tools are very valuable, they require time and computational resources to develop, in addition to the interpretation difficulties as the models are more complex. Thus, additional techniques are required to extract valuable information and insight from the models, also known as explainable machine learning.

We developed a machine learning predictive model for early progression to first-line treatment of patients with HR+/HER2-negative advanced breast cancer based on analysis of the unstructured information contained in the free-text notes of EHRs using NLP techniques. Using the same methodology, we were also able to develop a predictive model to identify long-responder patients. Our approach used the information collected during medical visits as a whole, without the need to transform it into structured data for analysis. To the best of our knowledge, our study is the first application of NLP techniques in the development of a predictive model for breast cancer in this setting. Other authors have used machine learning approaches to predict the efficacy of CDK4/6 inhibitors in HR+/HER2-negative metastatic breast cancer [36].

Patient data pooled from eight clinical trials were included in the study, but only structured characteristics of the disease and baseline patient status were analyzed. The CDK4/6 inhibitor model produced a prediction accuracy of 69.2%, and the ET alone model had an accuracy of 70.6%. Machine learning methods have also been used to develop predictive models to address other issues but transforming EHR information into structured variables [23, 24, 37]. Moreover, our results demonstrate that NLP-based machine learning models are slightly better at predicting early and late events than manually curated data-based predictive models. The relevance of this type of model is noteworthy as they improve the predictive capacity by highlighting details not revealed by classic manual extraction [22]. Furthermore, the efficiency is increased by reducing the time and expense required to review medical records [38]. Likewise, another strength of our model is that it is developed from real-world patients. Interest in the use of data obtained from real-world patients is growing as it provides information from a much broader population than that included in clinical trials, and is therefore more relevant from a healthcare point of view [39].

Our study has several limitations. As it is a retrospective study of two unique institutions, it is possible that there were data selection, measurement biases and missing data. Regarding the laboratory and pathology reports, only the information entered into the EHR by the oncologist was used. It is possible that the use of original reports may have influenced the results of our predictive model. Although our model was subjected to an accurate internal validation process, it is necessary to verify that it can be applied to other patients with other EHRs through appropriate external validation. In addition, we were unable to adjust the model to predict the risk of early progression for each of the four categories of treatment with enough accuracy to be clinically relevant. The low number of patients and the somewhat short length of follow-up in each of the categories contributed to this issue. For example, the number of

patients treated with ET plus CDKI was 63 and the median follow-up for metastatic disease of these patients was 18.7 months.

In conclusion, we successfully developed a NLP-based machine learning model to predict two very different types of events (i.e., early progression and long responders) that are of great importance when deciding the best therapeutic approach for patients with metastatic HR+/HER2-negative breast cancer.

REFERENCES

[1] Finn RS, Martin M, Rugo HS, Jones S, Im SA, Gelmon K, et al. Palbociclib and Letrozole in Advanced Breast Cancer. The New England journal of medicine. 2016;375:1925-36.

[2] Cristofanilli M, Turner NC, Bondarenko I, Ro J, Im SA, Masuda N, et al. Fulvestrant plus palbociclib versus fulvestrant plus placebo for treatment of hormone-receptor-positive, HER2-negative metastatic breast cancer that progressed on previous endocrine therapy (PALOMA-3): final analysis of the multicentre, double-blind, phase 3 randomised controlled trial. The Lancet Oncology. 2016;17:425-39.

[3] Hortobagyi GN, Stemmer SM, Burris HA, Yap YS, Sonke GS, Paluch-Shimon S, et al. Updated results from MONALEESA-2, a phase III trial of first-line ribociclib plus letrozole versus placebo plus letrozole in hormone receptor-positive, HER2-negative advanced breast cancer. Annals of oncology : official journal of the European Society for Medical Oncology. 2019;30:1842.

[4] Slamon DJ, Neven P, Chia S, Fasching PA, De Laurentiis M, Im SA, et al. Overall Survival with Ribociclib plus Fulvestrant in Advanced Breast Cancer. The New England journal of medicine. 2020;382:514-24.

[5] Sledge GW, Jr., Frenzel M. Analysis of Overall Survival Benefit of Abemaciclib Plus Fulvestrant in Hormone Receptor-Positive, ERBB2-Negative Breast Cancer-Reply. JAMA oncology. 2020.

[6] Johnston S, Martin M, Di Leo A, Im SA, Awada A, Forrester T, et al. MONARCH 3 final PFS: a randomized study of abemaciclib as initial therapy for advanced breast cancer. NPJ breast cancer. 2019;5:5.

[7] Loibl S, Turner NC, Ro J, Cristofanilli M, Iwata H, Im SA, et al. Palbociclib Combined with Fulvestrant in Premenopausal Women with Advanced Breast Cancer and Prior Progression on Endocrine Therapy: PALOMA-3 Results. The oncologist. 2017;22:1028-38.

[8] Tripathy D, Im SA, Colleoni M, Franke F, Bardia A, Harbeck N, et al. Ribociclib plus endocrine therapy for premenopausal women with hormone-receptor-positive, advanced breast cancer (MONALEESA-7): a randomised phase 3 trial. The Lancet Oncology. 2018;19:904-15.

[9] Sledge GW, Jr., Toi M, Neven P, Sohn J, Inoue K, Pivot X, et al. MONARCH 2: Abemaciclib in Combination With Fulvestrant in Women With HR+/HER2- Advanced Breast Cancer Who Had Progressed While Receiving Endocrine Therapy. Journal of clinical oncology : official journal of the American Society of Clinical Oncology. 2017;35:2875-84.

[10] Chang J, Clark GM, Allred DC, Mohsin S, Chamness G, Elledge RM. Survival of patients with metastatic breast carcinoma: importance of prognostic markers of the primary tumor. Cancer. 2003;97:545-53.

[11] Clark GM, Sledge GW, Jr., Osborne CK, McGuire WL. Survival from first recurrence: relative importance of prognostic factors in 1,015 breast cancer patients. Journal of clinical oncology : official journal of the American Society of Clinical Oncology. 1987;5:55-61.

[12] Hortobagyi GN, Smith TL, Legha SS, Swenerton KD, Gehan EA, Yap HY, et al. Multivariate analysis of prognostic factors in metastatic breast cancer. Journal of clinical oncology : official journal of the American Society of Clinical Oncology. 1983;1:776-86.

[13] Largillier R, Ferrero JM, Doyen J, Barriere J, Namer M, Mari V, et al. Prognostic factors in 1,038 women with metastatic breast cancer. Annals of oncology : official journal of the European Society for Medical Oncology. 2008;19:2012-9.

[14] Puente J, Lopez-Tarruella S, Ruiz A, Lluch A, Pastor M, Alba E, et al. Practical prognostic index for patients with metastatic recurrent breast cancer: retrospective analysis of 2,322 patients from the GEICAM Spanish El Alamo Register. Breast cancer research and treatment. 2010;122:591-600.

[15] Regierer AC, Wolters R, Ufen MP, Weigel A, Novopashenny I, Kohne CH, et al. An internally and externally validated prognostic score for metastatic breast cancer: analysis of 2269 patients. Annals of oncology : official journal of the European Society for Medical Oncology. 2014;25:633-8.

[16] Michiels S, Saad ED, Buyse M. Progression-Free Survival as a Surrogate for Overall Survival in Clinical Trials of Targeted Therapy in Advanced Solid Tumors. Drugs. 2017;77:713-9.

[17] Wiens J, Shenoy ES. Machine Learning for Healthcare: On the Verge of a Major Shift in Healthcare Epidemiology. Clinical infectious diseases : an official publication of the Infectious Diseases Society of America. 2018;66:149-53.

[18] Carrell DS, Halgrim S, Tran DT, Buist DS, Chubak J, Chapman WW, et al. Using natural language processing to improve efficiency of manual chart abstraction in research: the case of breast cancer recurrence. American journal of epidemiology. 2014;179:749-58.

[19] Ling AY, Kurian AW, Caswell-Jin JL, Sledge GW, Jr., Shah NH, Tamang SR. Using natural language processing to construct a metastatic breast cancer cohort from linked cancer registry and electronic medical records data. JAMIA open. 2019;2:528-37.

[20] Ehteshami Bejnordi B, Veta M, Johannes van Diest P, van Ginneken B, Karssemeijer N, Litjens G, et al. Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer. Jama. 2017;318:2199-210.

[21] Yala A, Lehman C, Schuster T, Portnoi T, Barzilay R. A Deep Learning Mammography-based Model for Improved Breast Cancer Risk Prediction. Radiology. 2019;292:60-6.

[22] Banerjee I, Bozkurt S, Caswell-Jin JL, Kurian AW, Rubin DL. Natural Language Processing Approaches to Detect the Timeline of Metastatic Recurrence of Breast Cancer. JCO clinical cancer informatics. 2019;3:1-12.

[23] Elfiky AA, Pany MJ, Parikh RB, Obermeyer Z. Development and Application of a Machine Learning Approach to Assess Short-term Mortality Risk Among Patients With Cancer Starting Chemotherapy. JAMA network open. 2018;1:e180926.

[24] Parikh RB, Manz C, Chivers C, Regli SH, Braun J, Draugelis ME, et al. Machine Learning Approaches to Predict 6-Month Mortality Among Patients With Cancer. JAMA network open. 2019;2:e1915997.

[25] Ribelles N, Jerez JM, Urda D, Subirats JL, Marquez A, Quero C, et al. Galén: Sistema de Información para la gestión y coordinación de procesos en un servicio de Oncología. Revista eSalud. 2010;6:1-12.

[26] Gao JJ, Cheng J, Bloomquist E, Sanchez J, Wedam SB, Singh H, et al. CDK4/6 inhibitor treatment for patients with hormone receptor-positive, HER2-negative, advanced or metastatic breast cancer: a US Food and Drug Administration pooled analysis. The Lancet Oncology. 2020;21:250-60.

[27] Collobert R, Weston J, Bottou L, Karlen M, Kavukcuoglu K, Kuksa P. Natural language processing (almost) from scratch. . Journal of Machine Learning Research. 2011;12:2493-537.

[28] Rodriguez JD, Perez A, Lozano JA. Sensitivity analysis of k-fold cross validation in prediction error estimation. IEEE transactions on pattern analysis and machine intelligence. 2009;32:569-75.

[29] Bischl B, Lang M, Kotthoff L, Schiffner J, Richter J, Studerus E, et al. mlr: Machine Learning in R. Journal of Machine Learning Research. 2016;17:5938-42.

[30] LeDell E, Petersen M, van der Laan M. Computationally efficient confidence intervals for cross-validated area under the ROC curve estimates. Electronic journal of statistics. 2015;9:1583-607.

[31] Demšar J. Statistical comparisons of classifiers over multiple data sets. Journal of Machine Learning Research. 2006;7:1-30.

[32] André F, Ciruelos E, Rubovszky G, Campone M, Loibl S, Rugo HS, et al. Alpelisib for PIK3CA-Mutated, Hormone Receptor-Positive Advanced Breast Cancer. The New England journal of medicine. 2019;380:1929-40.
[33] Jones RH, Casbard A, Carucci M, Cox C, Butler R, Alchami F, et al. Fulvestrant plus capivasertib versus placebo after relapse or progression on an aromatase inhibitor in metastatic, oestrogen receptor-positive breast cancer (FAKTION): a multicentre, randomised, controlled, phase 2 trial. The Lancet Oncology. 2020;21:345-57.
[34] Turner NC, Liu Y, Zhu Z, Loi S, Colleoni M, Loibl S, et al. Cyclin E1 Expression and Palbociclib Efficacy in Previously Treated Hormone Receptor-Positive Metastatic Breast Cancer. Journal of clinical oncology : official journal of the American Society of Clinical Oncology. 2019;37:1169-78.
[35] Tolaney SM, Toi M, Neven P, Sohn J, Grischke E, Llombart-Cussac A, et al. Clinical significance of PIK3CA and ESR1 mutations in ctDNA and FFPE samples from the MONARCH 2 study of abemaciclib plus fulvestrant. American Association for Cancer Research; Atlanta, GA, USA; March 29–April 3, 2019 (abstr 4458).
[36] Mason J, Gong Y, Amiri-Kordestani L, Wedam S, Gao JJ, Singh H, et al. Prediction of CDK inhibitor efficacy in ER+/HER2- breast cancer using machine learning algorithms [abstract]. In: Proceedings of the 2019 San Antonio Breast Cancer Symposium; 2019 Dec 10-14; San Antonio, TX. Philadelphia (PA): AACR. Cancer Res 2020;80(4 Suppl):Abstract nr PD2-07. 2019.
[37] Bertsimas D, Dunn J, Pawlowski C, Silberholz J, Weinstein A, Zhuo YD, et al. Applied Informatics Decision Support Tool for Mortality Predictions in Patients With Cancer. JCO clinical cancer informatics. 2018;2:1-11.
[38] Rajkomar A, Oren E, Chen K, Dai AM, Hajaj N, Hardt M, et al. Scalable and accurate deep learning with electronic health records. NPJ digital medicine. 2018;1:18.
[39] Booth CM, Karim S, Mackillop WJ. Real-world data: towards achieving the achievable in cancer care. Nature reviews Clinical oncology. 2019;16:312-25.

[40] Rodriguez-Brazzarola P, Ribelles N, Jerez JJ, Trigo J, Cobo M, Ramos-Garcia I, et al. Predicting the risk of visit emergency department (ED) in lung cancer patients using machine learning. J Clin Oncol 38: 2020 (suppl; abstr 2042).

TABLES

Table 1. Patient characteristics.

Characteristic	n
n	610
Medical encounters analyzed	17,426
Age at diagnosis (years; median, range)	52 (22–89)
Menopausal status at diagnosis	
Premenopausal	297 (48.7%)
Postmenopausal	272 (44.6%)
Unknown	41 (6.7%)
Stage at diagnosis	
I	60 (9.8%)
II	198 (32.5%)
III	186 (30.5%)
IV	143 (23.4%)
Unknown	23 (3.8%)
Grade at diagnosis	
1	63 (10.3%)
2	244 (40.0%)
3	122 (20.0%)
Unknown	181 (29.7%)
Neo/adjuvant chemotherapy	
Anthracyclines	119 (19.5%)
Anthracyclines-taxanes	168 (27.5%)
Taxanes	5 (0.8%)
CMF	57 (9.3%)
Unknown	16 (2.6%)
No chemotherapy	245 (40.2%)
Adjuvant hormone therapy	
Tamoxifen	279 (45.7%)
Aromatase inhibitors	59 (9.7%)
Tamoxifen-aromatase inhibitors	61 (10.0%)
Unknown	11 (1.8%)
IHQ phenotype	
Luminal A	119 (19.5%)
Luminal B	227 (37.2%)
Luminal Unknown	264 (43.3)

IHQ, immunohistochemical.

Table 1. Patient characteristics (continuation).

Characteristic	n
First-line treatment	
Hormone therapy	311 (51.0%)
Hormone therapy plus CDKI4/6 inhibitors	63 (10.3%)
Chemotherapy	119 (19.5%)
Chemotherapy plus ET maintenance	117 (19.2%)
First-line treatment total progressions	
Hormone therapy	246 (40.3%)
Hormone therapy plus CDKI4/6 inhibitors	34 (5.6%)
Chemotherapy	102 (16.7%)
Chemotherapy plus ET maintenance	91 (14.9%)
First-line treatment early progressions	
Hormone therapy	57 (9.3%)
Hormone therapy plus CDKI4/6 inhibitors	11 (1.8%)
Chemotherapy	54 (8.9%)
Chemotherapy plus ET maintenance	4 (0.7%)
First-line treatment long-responders	
Hormone therapy	103 (16.9%)
Hormone therapy plus CDKI4/6 inhibitors	3 (0.5%)
Chemotherapy	9 (1.5%)
Chemotherapy plus ET maintenance	37 (6.1%)

ET, Endocrine therapy

Table 2. Performance metrics of machine learning models.

	Dataset	Best model	AUC (95% CI)	TPR (95% CI)	TNR (95% CI)
Early progressions	Manually extracted	Elastic Net	0.734 (0.687-0.782)	0.736 (0.729-0.743)	0.713 (0.707-0.719)
	NLP	GLMBoost	0.758 (0.714-0.800)	0.758 (0.751-0.764)	0.707 (0.701-0.714)
Long responders	Manually extracted	Elastic Net	0.669 (0.608-0.730)	0.664 (0.654-0.674)	0.658 (0.649-0.667)
	NLP	GBM	0.752 (0.705-0.799)	0.717 (0.710-0.723)	0.766 (0.760-0.773)

AUC, area under the curve; NLP, natural language processing; TPR, true positive rate; TNR, true negative rate.