Citation: Clin Transl Sci (2018) 11, 85–92; doi:10.1111/cts.12514

ARTICLE

Electronic Health Record Phenotypes for Precision Medicine: Perspectives and Caveats From Treatment of Breast Cancer at a Single Institution

Matthew K. Breitenstein[1,2,*], Hongfang Liu[3], Kara N. Maxwell[4], Jyotishman Pathak[5] and Rui Zhang[6,7]

Precision medicine is at the forefront of biomedical research. Cancer registries provide rich perspectives and electronic health records (EHRs) are commonly utilized to gather additional clinical data elements needed for translational research. However, manual annotation is resource-intense and not readily scalable. Informatics-based phenotyping presents an ideal solution, but perspectives obtained can be impacted by both data source and algorithm selection. We derived breast cancer (BC) receptor status phenotypes from structured and unstructured EHR data using rule-based algorithms, including natural language processing (NLP). Overall, the use of NLP increased BC receptor status coverage by 39.2% from 69.1% with structured medication information alone. Using all available EHR data, estrogen receptor-positive BC cases were ascertained with high precision ($P = 0.976$) and recall ($R = 0.987$) compared with gold standard chart-reviewed patients. However, status negation ($R = 0.591$) decreased 40.2% when relying on structured medications alone. Using multiple EHR data types (and thorough understanding of the perspectives offered) are necessary to derive robust EHR-based precision medicine phenotypes.
Clin Transl Sci (2018) **11**, 85–92; doi:10.1111/cts.12514.

Study Highlights

WHAT IS THE CURRENT KNOWLEDGE ON THE TOPIC?
✔ Targeted therapeutics are routinely prescribed in certain diseases, including breast cancer. However, knowledge regarding optimal ascertainment of precision medicine phenotypes using electronic health record (EHR) data is limited.

WHAT QUESTION DID THIS STUDY ADDRESS?
✔ As we pursue multi-institution biobank studies to advance translational science and develop precision medicine knowledge, we increasingly rely on EHRs to annotate necessary clinical phenotypes. This study addresses important considerations regarding development of robust EHR-based precision medicine phenotypes needed for translational science and precision medicine Research.

WHAT THIS STUDY ADDS TO OUR KNOWLEDGE
✔ In addition to the nuanced perspectives offered by structured and unstructured data sources, this study articulates necessary considerations for robust phenotyping in the precision medicine era.

HOW THIS MIGHT CHANGE CLINICAL PHARMACOLOGY OR TRANSLATIONAL SCIENCE
✔ This study demonstrates how EHR phenotyping might empower translational science. Potential research applications include: human genetics, epidemiology, heath outcomes and utilization, quality improvement, and pharmacoepidemiology, to name a few. In the future, potential for clinical pharmacology practice implementation exists. However, enhancements and thorough validation are required first.

Advancements in our understanding of breast cancer (BC) provide powerful personalized treatment opportunities not readily available in all cancer types. At present, we commonly define molecular subtypes (estrogen receptor (ER), progesterone receptor (PR), or human epidermal growth factor receptor2 (HER2) overexpression) as positive (+) or negative (–) receptor status in BC patients. Negative status of the molecular subtypes ER, PR, and HER2 are characterized as triple-negative breast cancer (TNBC). Standard treatment in the neoadjuvant, adjuvant, and metastatic settings is guided by expression of hormone receptors (typically ER) and HER2.[1] Generally, endocrine therapy

[1]Institute for Biomedical Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, USA; [2]Center for Pharmacoepidemiology Research and Training, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, USA; [3]Division of Biomedical Statistics and Informatics, Mayo Clinic, Rochester, Minnesota, USA; [4]Department of Medicine, Division of Hematology/Oncology, Perelman School of Medicine, University of Pennsylvania, Pennsylvania, USA; [5]Division of Health Informatics, Weill Cornell Medicine, Cornell University, New York, New York, USA; [6]Department of Pharmaceutical Care & Health Systems, College of Pharmacy, University of Minnesota, Minneapolis, Minnesota, USA; [7]Institute for Health Informatics, University of Minnesota, Minneapolis, Minnesota, USA. *Correspondence: Matthew K. Breitenstein (mkbreit@upenn.edu)
Received 2 May 2017; accepted 26 September 2017. doi:10.1111/cts.12514

is prescribed for ER+ (PR+ or –) patients, Her2 directed therapy for Her2+ patients, and chemotherapy alone for patients with TNBC.[2] Utilization of targeted therapies, in part, have achieved great reductions in BC mortality, with 5-year overall survival rates approaching 90% in all-comer BC patients and 99% in ER-positive localized, node-negative cases.[3]

Cancer registries provide a wealth of cancer-specific information (e.g., stage, lymph node involvement); augmenting cancer registries with electronic health record (EHR) data to gain additional perspectives is an exciting frontier in cancer control research.[4] While the structure, depth, and longitudinal nature of EHR data present considerable strengths and opportunities, EHR-based phenotypes have potential to provide an incomplete (or inaccurate) perspective. Historically, development of EHR-based phenotyping algorithms has focused on maximizing ascertainment of true case or control disease status. While this approach facilitated the development of reproducible and portable phenotypes,[5] a limited focus was placed on identifying granular phenotype perspectives beyond high-level true case or control status. Applications of rule-based natural language processing (NLP) have effectively annotated breast cancer clinical data elements from pathology reports.[6] Further, machine learning-enhanced NLP has shown that annotation, including BC receptor status, of pathology notes can be developed with limited supervision using only a few training notes.[7] However, critical phenotype considerations remain unclear. These include: treatment perspectives contained across multiple EHR data sources, ascertaining negation of receptor status by evaluating surrounding contextual features contained within the clinical dialogs, and the impact of alternative EHR data source perspectives. With the advancement of precision medicine, disease phenotypes[8] are becoming increasingly complex—ensuring robust EHR perspectives from multiple data sources develops increasing importance. A thorough understanding of the nuances needed for EHR phenotyping in the era of precision medicine is critical to advance translational science.

In our study, we sought to understand the caveats and data considerations for creating robust precision medicine phenotypes of BC subtypes (e.g., ER, PR, HER2, and TN status) from multiple EHR perspectives. We posit that a "precision medicine phenotype" (PMP) represents perspectives based on molecular subtypes or targeted therapeutic utilization for cancer patients. Further, data sources and extraction techniques utilized are likely to have a pronounced impact on the perspective of the PMP. We aim to understand the different combinations of clinical data sources and informatics approaches necessary, including developed NLP infrastructure, to establish a robust BC PMP. We evaluate PMP perspectives by i) clinical data source coverage, and ii) performance relative to a baseline of expert-ascertained receptor status. Finally, we highlight caveats and potential biases regarding cancer PMPs.

METHODS

Patient subgroups and data extraction

Female patients with newly diagnosed BC between 1998 and 2011 were selected for inclusion in this study under minimal risk IRB 15–003347. Cancer registry data were linked with local EHR data for these BC patients ($n = 13{,}162$) at the Mayo Clinic, Rochester, Minnesota. Patients with localized BC with more than one primary tumor (bilateral or ipsilateral disease) and ductal carcinoma *in situ* (DCIS) were excluded. Patients not treated at the Mayo Clinic and seen for second opinions only were excluded. A combined study design, high-level pseudo code, and study nomenclature is included for reference (**Figure 1**). A subset of patients, encompassing our "Gold Standard Cohort," were selected at random for manual chart review (in addition to EHR phenotyping) to ascertain receptor status. The Gold Standard Cohort was then randomly split into approximately equal-sized "training" and "testing" subgroups for use in algorithm development and evaluation of performance. Receptor status was initially identified by multiple reviewers and confirmed by a single expert. BC-specific information was ascertained from the Mayo Clinic Tumor Registry, a hospital-based cancer registry manually curated by the Mayo Clinic Cancer Center. Structured and unstructured (i.e., clinical notes) EHR data augmented the cancer registry to ascertain BC receptor status. Specifically, rule-based informatics phenotyping methodologies were utilized to extract structured clinical data from the enterprise data trust (EDT),[9] a local data warehouse, and unstructured clinical and pathology notes as free text from the local EHR. Phenotyping details are described in-depth below. PMPs of receptor status were compared by perspective of prescribed medications, clinical narratives, or both.

Prescribed medication perspective

The *prescribed medication perspective* focused on identifying receptor status based on targeted medication utilization. Our corpus of prescribed medication data was gathered from structured medication orders in EDT[9] and clinical notes. The medication extraction and normalization system MedXN[10] was linked to a copy of the EHR to extract medication prescription information contained within clinical notes. Receptor status was ascertained from prescribed medication by brand and generic drug names, hand-curated by experts, targeting that specific receptor. ER+ patients were identified by prescriptions for antiestrogen therapy, aromatase inhibitors (AIs), or selective estrogen response modulators (SERMs). AIs included exemestane (Aromasin), anastrozole (Arimidex), or letrozole (Femara) prescriptions. SERM prescriptions included: tamoxifen (Nolvadex, Soltamox) or raloxifene (Evista, Keoxifene). HER2-positive patients are commonly prescribed trastuzumab (Herceptin) and were identified accordingly. Using these perspectives, a patient was classified as positive for one or more receptor status if corresponding neoadjuvant or adjuvant therapies were prescribed. Similarly, patients were classified as negative for a particular receptor status if medication prescription information was available for a patient, but no prescriptions for a corresponding targeted therapy were identified.

Clinical narrative perspective

The *clinical narrative perspective* focused on elucidating BC receptor status based on the unstructured text dialog

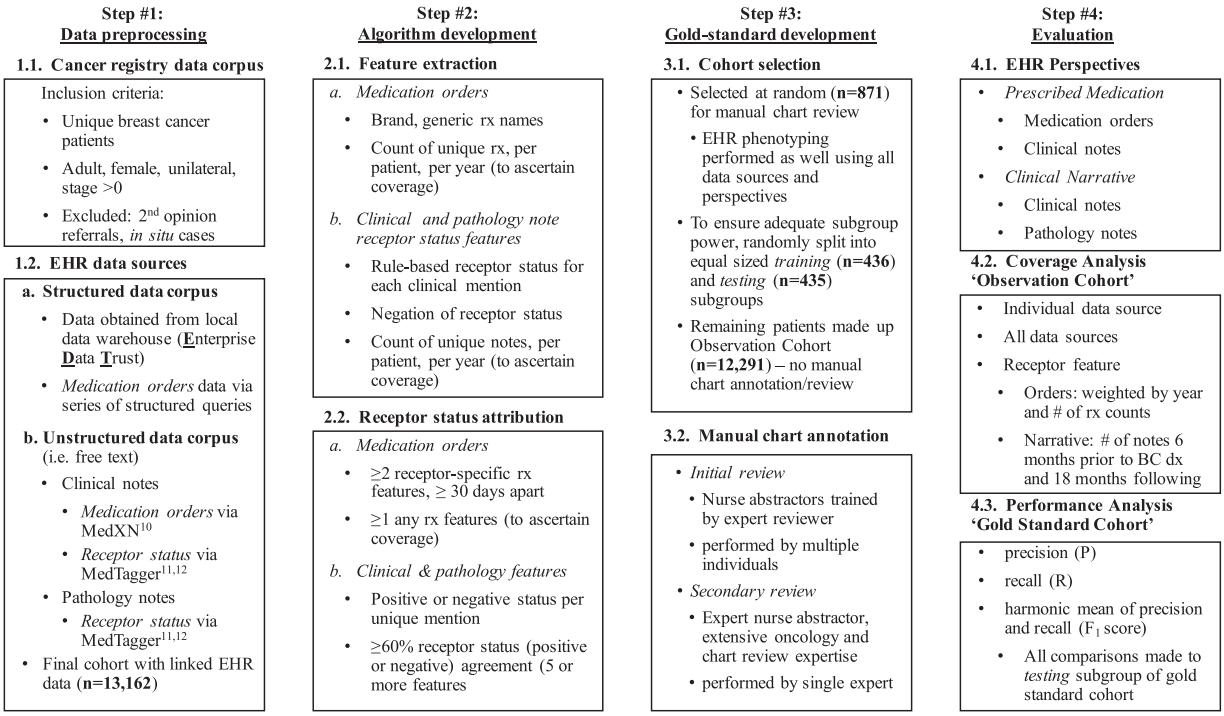

Figure 1 Overview of cohorts and pseudo code. A series of steps were taken to develop the breast cancer precision medicine phenotype; these included: 1) Data preprocessing, where data were extracted from both structured and unstructured (i.e., free text notes) electronic health record (EHR) data sources and linked with cancer registry data; 2) Ascertainment of receptor status from multiple EHR perspectives, initiated with extraction of necessary data features and subsequently attributed via a series of rules; 3) Development of a Gold Standard Cohort, consisting of patients manually chart reviewed and annotated for receptor status, to evaluate performance of the EHR rule-based algorithm; and 4) Perspectives and methodologies utilized to evaluate performance of the EHR rule-based algorithm.

contained within the patient's EHR. Specifically, clinical diagnosis of receptor status was ascertained from the clinical narrative, clinical notes, and pathology reports, via a series of rule-based NLP algorithms as follows: Pattern-based information extraction functionality was utilized to extract receptor information from the clinical narrative using the open-source pipeline MedTagger.[11,12] Development of the NLP algorithm included training and testing phases. Again, to ensure adequate power for each subtype, the "training" subgroup included approximately half of the total (n = 871) patients within the Gold Standard Cohort. The *training* subgroup included 10,182 (median: 18; range: 1–290) clinical notes and 9,077 (median: 20; range: 2–113) pathology reports. Training of the algorithm initiated using full receptor names (i.e., estrogen, progesterone, epidermal growth factor receptor 2) and standard abbreviations (e.g., ER, PR, HER2). Training results were recursively compared with the Gold Standard Cohort training subgroup to identify the most appropriate regular expression patterns. Algorithm training ceased when further manipulation of regular expression patterns did not improve the recall (i.e., sensitivity), or true negative rate (i.e., specificity). Following generation of the trained NLP algorithm, additions were made to accommodate negation (i.e., surrounding text that designates the opposite of the object) of the receptor status.

Many instances of receptor status-positive or -negative were identified per patient following deployment of the NLP BC receptor status algorithm. Commonly, a patient would have some level of "chatter" between individual receptor status-positive or -negative. Eventually, this chatter would resolve into a definitive clinical diagnosis. Anecdotally, chatter within clinical notes was an artifact of the diagnostic process and not necessarily reflective of clinical disagreement. For example, chatter may reflect discordance in receptor status between the diagnostic biopsy and the final surgical specimen. Following optimization, we applied a minimum threshold of 60% agreement of receptor status-positive or -negative to resolve individual receptor statuses and minimize false-negative or false-positive receptor status ascertainment. Patients below this threshold were designated as *nonresolvable* (or "nr"), requiring manual chart review to resolve and removed from our analysis.

Evaluation

The Gold Standard Cohort "testing" subgroup was utilized to elucidate algorithm performance by clinical data source. Comparisons of performance were made between the Observation Cohort and the Gold Standard Cohort testing subgroup. Performance was evaluated by phenotype perspectives of individual and combinations of EHR clinical data sources. A family of measures characterized performance, including: precision (P), recall (R), and harmonic mean of precision and recall (F_1 score).[13]

Table 1 Gold Standard Cohort receptor status

		ER			PR			HER2			TNBC		
Cohort	***N***	**(+)**	**(–)**	**na**	**(+)**	**(–)**	**na**	**(+)**	**(–)**	**na**	**yes**	**no**	**na**
Gold standard	871	751	113	7	678	187	6	116	453	302	44	682	145
Training	436	379	54	3	345	90	1	64	216	156	20	346	70
Validation	435	372	59	4	333	97	5	52	237	146	24	336	75

ER = estrogen receptor, PR = progesterone receptor, HER2 = human epidermal growth factor receptor, TN = triple negative, (+) = status positive, (–) = status negative, na = not available.

RESULTS

Gold Standard Cohort

Overall, within the Gold Standard Cohort receptor status was unavailable for a proportion of patients during manual chart review (**Table 1**). Likely a reflection of clinical practice standards at this facility, only a small minority of patients were missing data for ER and PR, with the missing percentage varying little over time. The majority of absent data was for HER2 status, consistent with the clinically heterogeneous adoption of HER2-directed therapy[14] over the course of the study period (1998–2011). As expected for this study period, HER2 status was unavailable for the majority of patients (75.1%). A detailed list of receptor status coverage ascertained for the Gold Standard Cohort can be found in **Table 1**.

Observation cohort

After removing patients contained within the Gold Standard Cohort, our "Observation Cohort" consisted of patients (n = 12,291) who were phenotyped using EHR data and did not undergo manual review. The median age at BC diagnosis was 57 years (range 18–98); 37.9% of the patients were Stage III–IV. Patient records were utilized to evaluate two fundamentals of EHR phenotyping: i) data reliability via clinical attribute coverage of a patient cohort by EHR perspectives, and ii) performance measurement (precision, recall, and F_1 score) of the EHR perspectives. Coverage and performance are important characteristics of robust precision phenotypes: development of approaches that maximize data ascertainment and return accurate, reliable predictions are imperative.

Coverage by clinical data source

In our application, cohort coverage refers to patients within the Observation Cohort having sufficient EHR data to ascertain a phenotype via the specified EHR perspective (i.e., some level of necessary data elements were present). Similarly, individual receptor status coverage (e.g., ER, PR, HER2, and TNBC) represents the ability to ascertain positive or negative status for that particular receptor status via an EHR perspective. Observation Cohort coverage (i.e., available EHR data) ranged from 63–92% (**Table 2**) by individual data sources. First, within the *prescribed medications perspective*, Observation Cohort coverage based on structured data alone was unexpectedly low, at 69.1%, potentially an artifact of legacy system integration. Inclusion of *clinical notes* to EDT, *prescribed medications perspectives* increased Observation Cohort coverage to 77.1% from 69.1%. Second, the *clinical narrative perspective,* ascertained via the rule-based NLP algorithm using both clinical and pathology notes, increased Observation Cohort coverage to 92.1% (~30% increase when compared with a baseline of structured medication information alone of 69.1%). When all data sources were utilized, representing the combined *prescribed medication* and *clinical narrative perspectives*, Observation Cohort coverage exceeded 96%.

The following observations and posited justifications were ascertained from the coverage analysis (**Table 2**): i) Overall, ER+ feature coverage was low (between 42.6% and 45.6%) for *prescribed medications perspectives*, the maximum ability to identify positive cases, while having a concerning 69.1–77.1% data source coverage. Approximately 70% of women diagnosed with BC are known to be ER+.[3] Alternatively, our developed NLP methodology identified between 68.8% and 74.5% women as being ER+. Further, utilizing all EHR data sources (including *prescribed medications* and *clinical narrative perspectives*) we identified 73.8% of our cohort as being ER+, with 96.2% EHR data source coverage. ii) PR was not resolvable via *prescribed medication perspective*: Treatment decisions are typically made by ER status alone. The developed NLP methodology was needed to identify PR+ cases from the *clinical narrative perspective*, where between 52% to 64% were identified as PR-positive. iii) The *clinical narrative perspective* was needed to resolve HER2 status: HER2 status had satisfactory coverage utilizing our developed NLP methodology within clinical or pathology note clinical narratives, implying the signal was contained within the EHR. However, medication order data sources were ineffective at identifying HER2 status, suggestive of potential nomenclature (e.g., combination therapy acronyms) or data source (e.g., intravenous vs. oral delivery route) issues. Further, iv) TNBC was unreliable due to poor coverage: Reliable TNBC status could only be ascertained from *clinical narrative perspectives,* potentially due to the limited reliability of status negation by the medication orders data sources. We identified between 5.9% and 8.4% of patients as TNBC utilizing NLP to ascertain receptor status directly from the clinical narrative, which is close to the anticipated TNBC prevalence. v) Overall, alternative perspectives helped resolve our PMPs. For individual receptor status coverage, alternative perspectives, or combinations of perspectives helped to resolve precision phenotypes. vi) The ability to disambiguate between sparse EHR data and lack of prescription is critical when assigning a targeted therapy status based on prescribed therapeutics. In future work, inverse weighting by measures of data sparseness are recommended.

Performance evaluation in Gold Standard Cohort

For algorithm performance, receptor status phenotype precision measures were higher in *clinical narrative perspective* and lower in the *prescribed medication perspective* (**Table 3**). We found ER status to be a reliable precision medicine

Table 2 Coverage of individual EHR data sources and phenotypes

			Prescribed medications perspective			Clinical narrative perspective			
Receptor status		**Data source**	**EDT**	**Clinical notes**	**EDT or clinical notes**	**Clinical notes**	**Pathology notes**	**Clinical or pathology**	**All EHR perspective**
Observation Cohort coverage		*n*	8,826	8,078	9,851	11,287	10,236	11,766	12,291
		%	*69.1%*	*63.3%*	*77.1%*	*88.4%*	*80.2%*	*92.1%*	*96.2%*
ER feature coverage	(+)	*n*	3,761	3,507	4,491	8,305	7,045	8,771	9,069
		%	*42.6%*	*43.4%*	*45.6%*	*73.6%*	*68.8%*	*74.5%*	*73.8%*
	(–)	*n*	5,065	4,571	5,360	2,214	1,921	2,374	2,241
		%	*57.4%*	*56.6%*	*54.4%*	*19.6%*	*18.8%*	*20.2%*	*18.2%*
	nr	*n*	0	0	0	768	1,270	621	981
		%	*0.0%*	*0.0%*	*0.0%*	*6.8%*	*12.4%*	*5.3%*	*8.0%*
PR feature coverage	(+)	*n*	—	—	—	7,205	5,341	7,531	7,531
		%	—	—	—	*63.8%*	*52.2%*	*64.0%*	*61.3%*
	(–)	*n*	—	—	—	3,030	2,728	3,255	3,255
		%	—	—	—	*26.8%*	*26.7%*	*27.7%*	*26.5%*
	nr	*n*	8,826	8,078	9,851	1,052	2,167	980	1,505
		%	*100.0%*	*100.0%*	*100.0%*	*9.3%*	*21.2%*	*8.3%*	*12.2%*
HER2 feature coverage	(+)	*n*	6	121	121	1,611	1,438	1,770	1,786
		%	*0.1%*	*1.5%*	*1.2%*	*14.3%*	*14.0%*	*15.0%*	*14.5%*
	(–)	*n*	8,820	7,957	9,730	5,398	4,589	5,903	5,897
		%	*99.9%*	*98.5%*	*98.8%*	*47.8%*	*44.8%*	*50.2%*	*48.0%*
	nr	*n*	0	0	0	4,278	4,209	4,093	4,608
		%	*0.0%*	*0.0%*	*0.0%*	*37.9%*	*41.1%*	*34.8%*	*37.5%*
TNBC feature coverage	yes	*n*	0	0	0	1,014	606	1,102	1,035
		%	*0.0%*	*0.0%*	*0.0%*	*9.0%*	*5.9%*	*9.4%*	*8.4%*
	no	*n*	3,763	3,556	4,538	7,162	5,415	7,582	9,876
		%	*42.6%*	*44.0%*	*46.1%*	*63.5%*	*52.9%*	*64.4%*	*80.4%*
	nr	*n*	5,063	4,522	5,313	3,111	4,215	3,082	1,380
		%	*57.4%*	*56.0%*	*53.9%*	*27.6%*	*41.2%*	*26.2%*	*11.2%*

Receptor status phenotype coverage by clinical data source Note: total cohort size $n = 12{,}770$; cohort coverage refers to coverage of that clinical data source out of the total cohort size. ER = estrogen receptor, PR = progesterone receptor, HER2 = human epidermal growth factor receptor 2, TNBC = triple negative, nr = true missing or unable to resolve status; % of patients with relevant EHR data source coverage for individual receptor status phenotypes.

phenotype when all data sources were utilized in our Gold Standard Cohort: Ascertainment of ER+ status using all available data sources was excellent ($P = 0.98$, $R = 0.99$, $F = 0.98$) and provided very high coverage (96.2%). However, negation performance was noticeably lower ($P = 0.97$, $R = 0.61$, $F = 0.75$) for *prescribed medications* as opposed *to clinical narrative* ($P = 0.99$, $R = 0.98$, $F = 0.99$) perspectives. Further, we found PR precision medicine phenotypes to be reliable when NLP of clinical narratives was utilized: Using the perspective of the *clinical narrative,* PR performance was very high ($P = 0.99$, $R = 0.94$, $F = 0.96$). An important note: While coverage may be conditionally independent of performance, specifically precision, increases in overall cohort coverage are linked to increased *R* and *F* scores (i.e., the ability to negate an individual receptor status).

While HER2 ($P = 0.70$, $R = 0.67$, $F = 0.68$) and TNBC ($P = 0.72$, $R = 0.68$, $F = 0.70$) status performances were acceptable, they remained noticeably poorer than ER or PR status. Related, we found the HER2 precision medicine phenotype to be unreliable and complicated by both defining and resolving applicable synonyms. Reducing the observation period, controlling for potential bias due to treatment advancements, to contain the years 2008 through 2011, only slightly enhanced algorithm performance ($P = 0.58$, $R = 0.78$, $F = 0.67$ and $P = 0.75$, $R = 0.86$, $F = 0.80$). The many combinations and synonyms for HER2 status (e.g., HER2/neu, HER2) complicated training. Complications and data quality issues arising from lack of standardization are likely to be encountered with other precision medicine therapeutic targets, data sources, and synonymous naming conventions.

PR and TNBC status could not be ascertained utilizing medication orders due to low coverage. However, NLP could identify most cases of either status from clinical notes. To control for potential confounding by indication (antiestrogen therapies target reactions relevant to both ER and PR status) during performance evaluation, we collapsed ER and PR status within the Gold Standard Cohort. While ER+ receptor status had a very high predictive power utilizing medication orders, the recall and the harmonic mean (F_1 score) remained extremely poor. Further complicating these efforts was the limited cohort coverage (77%) encountered via *medication order*s. Negation of any receptor status remained poor; negative receptor status was not reliably ascertained utilizing medication orders. Due to the likely nonrandom gaps in coverage, strong potential exists to introduce sampling bias.

Table 3 Receptor status phenotype performance compared within manual chart reviewed Gold Standard Cohort—testing subset

		Prescribed medications			Clinical narratives			
Receptor status		**EDT**	**Clinical notes**	**EDT or clinical notes**	**Clinical**	**Pathology**	**Clinical or pathology**	**All EHR sources**
n		374	335	374	377	360	377	435
Coverage		*86.0%*	*77.0%*	*86.0%*	*86.7%*	*82.8%*	*86.7%*	*100%*
ER	P	0.9849	0.9702	0.9710	0.9877	0.9861	0.9877	0.9758
	R	0.5909	0.5909	0.6091	0.9786	0.9100	0.9847	0.9877
	F	0.7386	0.7345	0.7486	0.9831	0.9465	0.9862	0.9818
PR	P	—	—	—	0.9784	0.9730	0.9857	0.9857
	R	—	—	—	0.9347	0.7780	0.9418	0.9418
	F	—	—	—	0.9561	0.8657	0.9632	0.9632
HER2	P	0.0000	1.0000	1.0000	0.7750	0.4583	0.6977	0.6977
	R	0.0000	0.0294	0.0222	0.6889	0.5116	0.6667	0.6667
	F	0.0000	0.0571	0.0435	0.7294	0.4835	0.6818	0.6818
TN	P	—	—	—	0.6522	0.8462	0.7000	0.7222
	R	—	—	—	0.7895	0.5790	0.7368	0.6842
	F	—	—	—	0.7143	0.6875	0.7180	0.7027

All comparisons made to "gold standard" validation cohort ($n = 435$); P = precision, R = recall, F = F1 score (harmonic mean of precision and recall; ER = estrogen receptor, PR = progesterone receptor, HER2 = human epidermal growth factor receptor 2, TN = triple negative. PR and TN are blank because they cannot be directly inferred from a prescribed medications perspective.

Future work should include evaluations of data source coverage and density to control for these potential biases.

DISCUSSION

As precision medicine matures, incorporating EHR-driven precision medicine phenotypes into clinical pharmacology research endeavors will become increasingly important. Further, the need for informatics phenotyping approaches to develop PMPs will continue to grow.

Precision medicine treatment opportunities for BC are relatively advanced compared with other cancer sites, with mature utilization of targeted therapies. However, other cancer types, including diffuse large B-cell lymphoma, non-small cell lung cancer, and multiple myeloma, are rapidly increasing targeted therapy offerings. Further, potential to target alterations in select cancer pathways with therapeutics, regardless of the cancer type, remain on the near horizon.[15] The clinical reality of precision medicine will become increasingly complex, requiring disciplined application of informatics approaches to ascertain robust precision phenotypes. While we demonstrated potential utility as a research application, rigorous evaluation and expert clinician input is needed prior to implementation considerations for clinical practice applications.

Caveats of precision medicine phenotypes

While related, the concepts molecular subtype, receptor status, and precision phenotype referenced throughout this article have nuanced differences. Molecular subtype refers to altered (i.e., increased) molecular expression of a variant or wildtype gene product of known etiological significance. In BC, immunohistochemistry staining of a biopsy is performed and a pathologist establishes receptor status in accordance with the established clinical guideline.[16] While receptor status and treatment with the corresponding targeted therapy are typically congruent, in some instances a patient may be positive (or borderline) for multiple receptors types. When ambiguity exists, a tumor board, consisting of relevant oncology expertise, will convene to develop a diagnosis and assign a corresponding treatment plan. However, these nuances are likely to be unclear, or lost as noise, from the perspective of naïve phenotyping algorithm. In these instances, the precision phenotype might represent either the most clinically relevant receptor status/molecular subtype or to which targeted therapy was first prescribed, or potentially something more nuanced in between. Anecdotally, we identified a certain level of "noise," occurring primarily during the diagnostic process, while training the NLP algorithm within the clinical narrative. Our rule-based approach was iteratively validated to resolve receptor status variation across the clinical dialog data sources at the specified threshold. While validated in our algorithm for appropriateness, inappropriate selection of this threshold could potentially impact the receptor status phenotype and introduce bias—validation is required prior to adjustments. Appropriate representation of the varying EHR perspectives (and corresponding source data integrity) is a critical design consideration. Related, more comprehensive incorporation of clinical acronyms for combination therapies also holds potential to enhance these perspectives.[17]

Insights from our BC precision medicine phenotype

In our study, clinical or pathology notes alone or together provided the broadest cohort coverage and clinical notes alone provided the most precise measure of receptor status. The pathology note data source outperformed the clinical note data source for TNBC status. Utilizing all EHR data sources provided the largest cohort coverage. High coverage was accompanied by some of the highest individual receptor status performance by every clinical data source. Augmenting structured EHR data with NLP increased coverage and performance for BC PMP for both case identification and negation at our institution. We increased data coverage

when NLP of EHR notes augmented the structured EHR data. Structured data alone might be insufficient to ascertain a robust precision medicine phenotype when complete patient data coverage is unavailable. Overall, our NLP approaches increased coverage, with comparable results for applications in chronic conditions.[18] As we demonstrated, balancing a combination of perspectives was necessary to ascertain a high-quality EHR-derived phenotype.

Overcoming potential gaps in EHR data coverage

Despite gains with NLP, negation remained problematic throughout our study, introducing potential for bias. Specifically, identifying true negation of a targeted therapy status was problematic when coverage for a patient's EHR data was poor—improvements are ongoing. When longitudinal drug exposures can be captured with a high level of coverage, we posit that negation will become more reliable. Minimizing potential bias begins by ensuring data integrity within local research data warehouses. In our study, the EDT, our local data warehouse, was documented to represent all clinical data from the local EHR.[9] However, documentation available to the researchers regarding handling of data from legacy EHR systems and maintenance/integrity was incomplete—a potential study limitation. This can be particularly problematic when certain therapeutics (e.g., trastuzumab treatment for HER2+ breast cancer vs. anastrozole for ER+ breast cancer) have varying administration routes (infusion vs. oral) and corresponding data will be captured via different clinical workflows. The integration of external data into the EHR perspective would allow for a precision phenotype to include the perspective of filled medication orders—integration of insurance claims perspectives hold great promise to overcome biases resulting from incomplete coverage. For example, in cases where longitudinal EHR coverage is initially poor, augmenting the EHR perspective with that of filled medication orders from insurance claims, to represent care received at external healthcare facilities, might increase PMP performance. Conversely, in instances where a patient is known to have near complete coverage in the EHR, performance of that phenotype will potentially be enhanced, with increased reliability. Finally, novel computational phenotyping approaches are likely needed to account for data coverage within a patient's EHR compared with patients with similar disease states and drug exposures.

Informatics opportunities

We posit informatics approaches will be critical to ascertaining PMPs in an evolving landscape of molecular targets and corresponding therapeutic agents: First, clinical utilization of evolving molecular subtypes and newly discovered targeted therapies will likely outpace mandates for annotation of applicable clinical data elements in population research registries. For example, national organizations such as the Surveillance, Epidemiology, and End Results Program (SEER) mandate collection of only specific clinical data elements. While curating these registries is of tremendous value, gaps in data (or clinical data elements of insufficient granularity) needed to define molecular subtypes and identify targeted therapy utilization may exist. Informatics approaches offer the opportunity to resolve granularity gaps with the wealth of clinical knowledge contained within the EHR. Second, curating these cancer registries commonly rely on resource-intensive manual abstraction. As informatics-based phenotyping advances, future opportunities to develop automated or semiautomated data annotation procedures might exist. Third, while rule-based algorithms might sufficiently phenotype most cases, in certain instances medical oncology and pathology diagnostic decisions might remain too "unclear," from an informatics perspective, for an EHR phenotyping algorithm to disambiguate case status. In retrospective research, extraction processes that refer select "unclear" cases for expert clinical review might provide an optimal means to prioritize the manual annotation efforts needed for registry inclusion. Rule-based approaches may not be readily scalable "out of the box"; advancements in deep-learning techniques are needed to ensure feature extraction adaptable to nuances contained within multiple care delivery perspectives. Fourth, prescribed medications, medical oncology, and pathology data sources provide unique perspectives that might reflect certain aspects of clinical reality. While different clinical perspectives are certainly valuable for research applications and conceptually scalable to clinical informatics applications, they should not inappropriately offer a prescriptive perspective for guiding clinical care decisions. Specifically, based on observed performance and limited ability (i.e., unreliability) to negate or identify a "negative" status, improvements are needed prior to implementation of a similar algorithm. Close clinical collaboration and validation is needed prior to consideration of phenotyping algorithms for clinical care applications.

In the future, the complexity of informatics approaches needed to ascertain precision phenotypes will likely vary between simple rule-based and novel computational approaches. Formally trained informaticians can help guide application and development of the methodologies needed to ascertain robust precision phenotypes. Cancer subtypes, such as TNBC and basal-like subtypes, are frequently heterogeneous,[19] and characterized using multiple types of data, complicating potential EHR phenotyping. In research endeavors, phenotyping approaches that span genetic, transcriptional, histological,[20] and clinical features will likely be necessary to resolve these heterogeneous cases,[21] which hold potential to uncover novel biology.[22]

Cancer registries are a particularly robust resource for personalized medicine discovery. Augmenting cancer registry data with clinical data elements from the local EHR offers a profound opportunity to gain knowledge that may have previously been hidden. For certain cancer registries, additional biological specimens are collected (e.g., tumor slides and blocks, germline DNA, somatic tumor DNA, plasma, serum) that correspond to the clinical intervention. This enables pursuit of translational bioinformatics research endeavors spanning the richness of information contained within both the local EHR and biological specimens.[23] Indeed, development of robust precision medicine phenotypes is critical for translational bioinformatics to empower pursuit of clinical pharmacology knowledge.

CONCLUSION

As precision medicine phenotypes grow increasingly complex in the era of precision medicine, nuanced informatics applications that account for multiple EHR perspectives are needed. A thorough understanding of EHR data source perspectives, data source coverage, and potential for bias are imperative to the development of robust precision medicine phenotypes.

Acknowledgments. This work was supported by the National Cancer Institute-sponsored Mayo Clinic Cancer Genetic Epidemiology Training Program (**R25 CA092049)**. The authors thank James R. Cerhan, MD, PhD, for the substantial editorial feedback provided in the development of this article. Further, the researchers thank the nurse abstraction group led by Wendy Gay for their contributions to chart review and cohort integrity assurance, and Xiaoyang Ruan, PhD, for assistance in deployment of natural language processing algorithms.

Author Contributions. M.K.B., H.L., K.N.M., J.P., and R.Z. wrote the article; M.K.B. and R.Z. designed the research; M.K.B. performed the research; M.K.B. analyzed the data; M.K.B., H.L., and J.P. contributed new reagents/analytical tools.

Conflict of Interest/Disclosure. The authors declared no conflicts of interest.

1. Harris, L.N. *et al.* Use of biomarkers to guide decisions on adjuvant systemic therapy for women with early-stage invasive breast cancer: American Society of Clinical Oncology clinical practice guideline. *J. Clin. Oncol.* **34**(10), 1134–1150 (2016).
2. Carels, N., Spinassé, L.B., Tilli, T.M. & Tuszynski, J.A. Toward precision medicine of breast cancer. *Theor. Biol. Med. Model.* **13**(1), 7 (2016).
3. Howlader, N. *et al.* US incidence of breast cancer subtypes defined by joint hormone receptor and HER2 status. *J. Natl. Cancer Inst.* **106**(5), dju055 (2014).
4. Abernethy, A.P. *et al.* Rapid-learning system for cancer care. *J. Clin. Oncol.* **28**(27), 4268–4274 (2010).
5. Gottesman, O. *et al.* The electronic medical records and genomics (eMERGE) network: past, present, and future. *Genet. Med.* **15**(10), 761.
6. Buckley, J.M. *et al.* The feasibility of using natural language processing to extract clinical information from breast pathology reports. *J. Pathol. Informat.* **3**, 23 (2012).
7. Yala, A. *et al.* Using machine learning to parse breast pathology reports. *Breast Cancer Res. Treat.* **161**(2), 203–211 (2017).
8. Wei, W.Q. & Denny, J.C. Extracting research-quality phenotypes from electronic health records to support precision medicine. *Genome Med.* **7**(1), 41 (2015).
9. Chute, C.G., Beck, S.A., Fisk, T.B. & Mohr, D.N. The enterprise data trust at Mayo Clinic: A semantically integrated warehouse of biomedical data. *J. Am. Med. Inform. Assoc.* **17**(2), 131–135 (2010).
10. Sohn, S., Clark, C., Halgrim, S.R., Murphy, S.P., Chute, C.G. & Liu, H. MedXN: An open source medication extraction and normalization tool for clinical text. *J. Am. Med. Inform. Assoc.* **21**(5), 858–865 (2014).
11. Torii, M., Wagholikar, K. & Liu, H. Using machine learning for concept extraction on clinical documents from multiple data sources. *J. Am. Med. Inform. Assoc.* **18**(5), 580–587 (2011).
12. Liu, H. *et al.* An information extraction framework for cohort identification using electronic health records. *AMIA Summits Transl. Sci. Proc.* **2013**, 149 (2013).
13. Goutte, C. & Gaussier, E. A probabilistic interpretation of precision, recall and F-score, with implication for evaluation. *Eur. Conf. Inform. Retriev. Res.* **2005**, 345–359 (2005).
14. Nahta, R.E. & Esteva, F.J. Trastuzumab: triumphs and tribulations. *Oncogene.* **26**(25), 3637–3643 (2007).
15. Renfro, L.A., Mallick, H., An, M.W., Sargent, D.J. & Mandrekar, S.J. Clinical trial designs incorporating predictive biomarkers. *Cancer Treat. Rev.* **43**, 74–82 (2016).
16. Hammond, M.E.H. *et al.* American Society of Clinical Oncology/College of American Pathologists guideline recommendations for immunohistochemical testing of estrogen and progesterone receptors in breast cancer. *J. Clin. Oncol.* **28**(16), 2784–2795 (2010).
17. Warner, J.L., Cowan, A.J., Hall, A.C. & Yang, P.C. HemOnc.org: A collaborative online knowledge platform for oncology professionals. *J. Oncol. Pract.* **11**(3), e336–e350 (2015).
18. Carroll, R.J. *et al.* Portability of an algorithm to identify rheumatoid arthritis in electronic health records. *J. Am. Med. Inform. Assoc.* **19**(e1), e162–e169 (2012).
19. Rakha, E.A., Reis-Filho, J.S. & Ellis, I.O. Basal-like breast cancer: A critical review. *J. Clin. Oncol.* **26**(15), 2568–2581 (2008).
20. Kothari, S., Phan, J.H., Stokes, T.H. & Wang, M.D. Pathology imaging informatics for quantitative analysis of whole-slide images. *J. Am. Med. Inform. Assoc.* **20**(6), 1099–1108 (2013).
21. Pareja, F., Geyer, F.C., Marchiò, C., Burke, K.A., Weigelt, B., Reis-& Filho, J.S. Triple-negative breast cancer: the importance of molecular and histologic subtyping, and recognition of low-grade variants. *Breast Cancer.* **2**, 16036 (2016).
22. Ritchie, M.D., Holzinger, E.R., Li, R., Pendergrass, S.A. & Kim, D. Methods of integrating data to uncover genotype-phenotype interactions. *Nat. Rev. Genet.* **16**(2), 85–97 (2015).
23. Tenenbaum, J.D. *et al.* An informatics research agenda to support precision medicine: seven key areas. *J. Am. Med. Inform. Assoc.* **23**(4), 791–795 (2016).