Automated discovery of drug treatment patterns for endocrine therapy of breast cancer within an electronic medical record

Guergana K Savova,1,2 Janet E Olson,1 Sean P Murphy,1 Victoria L Cafourek,1 Fergus J Couch,1 Matthew P Goetz,1 James N Ingle,1 Vera J Suman,1 Christopher G Chute,1 Richard M Weinshilboum1

ABSTRACT
Objective To develop an algorithm for the discovery of drug treatment patterns for endocrine breast cancer therapy within an electronic medical record and to test the hypothesis that information extracted using it is comparable to the information found by traditional methods.

Materials The electronic medical charts of 1507 patients diagnosed with histologically confirmed primary invasive breast cancer.

Methods The automatic drug treatment classification tool consisted of components for: (1) extraction of drug treatment-relevant information from clinical narratives using natural language processing (clinical Text Analysis and Knowledge Extraction System); (2) extraction of drug treatment data from an electronic prescribing system; (3) merging information to create a patient treatment timeline; and (4) final classification logic.

Results Agreement between results from the algorithm and from a nurse abstractor is measured for categories: (0) no tamoxifen or aromatase inhibitor (AI) treatment; (1) tamoxifen only; (2) AI only; (3) tamoxifen before AI; (4) AI before tamoxifen; (5) multiple AIs and tamoxifen cycles in no specific order; and (6) no specific treatment dates. Specificity (all categories): 96.14%–100%; sensitivity (categories (0)–(4)): 90.27%–99.83%; sensitivity (categories (5)–(6)): 99.93%; positive predictive values: 80%–97.38%; negative predictive values: 96.91%–99.93%.

Discussion Our approach illustrates a secondary use of the electronic medical record. The main challenge is event temporality.

Conclusion We present an algorithm for automated treatment classification within an electronic medical record to combine information extracted through natural language processing with that extracted from structured databases. The algorithm has high specificity for all categories, high sensitivity for five categories, and low sensitivity for two categories.

BACKGROUND AND SIGNIFICANCE
The electronic medical record (EMR) has traditionally been viewed by medical practitioners as a place to document the medical care of their patients, provide a record of billable services, and protect the legal interests of both patients and healthcare providers. A less appreciated purpose of the EMR is its potential for medical research. A broad EMR implementation will result in the accumulation of vast quantities of patient data that can be mined to improve the quality, safety, efficiency, and efficacy of healthcare, and also advance research and public health. These developments will necessitate exposing enormous computational and analytical tools directly to practitioners and investigators. At the national level, the US Office of the National Coordinator of Health Information Technology (ONC) has emphasized the importance of the use of information technology in healthcare by leading the Strategic Health IT Advanced Research Projects (SHARP)7 initiative. One of the four SHARP projects focuses exclusively on secondary data use of information arising from EMR (SHARPn).2

Several national efforts, supported by the US National Institutes of Health, demonstrate the secondary use of the EMR for research purposes by focusing on ‘developing, disseminating and applying approaches to research that combine DNA biorepositories with the EMR systems for large-scale, high-throughput biomedical research’3 such as the electronic Medical Record and Genomics consortium4 and the Pharmacogenomics Research Network (PGRN).5 Wilke and colleagues6 overview the convergence of healthcare and genotyping technologies. In a 2010 JAMA manuscript on the association between CYP2D6 polymorphisms and outcomes among women with early stage breast cancer treated with tamoxifen, Schroth and colleagues7 point out that the results of their study ‘once more underscore the need for high-powered data sets.’ A major bottleneck is patient data abstraction, traditionally a time-consuming, effort intense review of patients’ medical records by trained abstractors to find relevant phenotype nuggets.

The EMR offers the computational environment for automated or semi-automated information extraction across data residing in structured databases and in the free-text clinical narratives generated by the practitioners at the point of care. Information extraction from structured databases requires standard querying techniques, while information extraction from the clinical narrative calls for Natural Language Processing (NLP) methods. EMR algorithms for high-throughput phenotyping require merged information extracted from multiple sources, for example, prescribing systems, laboratory results, and clinical notes. Many academic healthcare centers now recognize...
Research and applications

the secondary uses of the EMR by creating operational EMR mirrors for research purposes as exemplified by the Mayo Clinic Enterprise Data Trust (EDT)\(^8\) and Informatics for Integrating Biology and the Bedside (i2b2) datamarts.\(^9\) In an August, 2011 JAMA editorial, Jha\(^10\) discusses the promises of the EMR, emphasizing the importance of NLP as an enabling tool for accessing the vast information residing in the EMR, and stated that ‘federal government can play a helpful role by funding the basic research needed to launch this field forward.’

Our study contributes to the growing body of research demonstrating the secondary use of the EMR data enabled by NLP methods for mining the clinical narrative and its integration with other types of EMR information. In this study, we used the manually abstracted treatment regimens from a prior breast cancer pharmacogenomic study to test the hypothesis that automatically extracted information from the EMR originally created by the practitioner at the point of care, coupled with computational techniques, is comparable to the information found by traditional manual abstraction methods. Our model study is part of a larger PGRN-supported breast cancer pharmacogenomic study designed to evaluate the role of genetic variation related to endocrine therapy, clinical outcomes, and drug-related side effects. We chose this study as a ‘model system’ because: (1) it is an example of medical research combining large-scale genomic and phenotype data in an attempt to advance breast cancer drug therapy; and (2) it presents technical challenges typical for information extraction from the EMR, specifically mining treatment events and relating them temporally. The study objectives were: (1) to develop an automated EMR endocrine breast cancer treatment classification algorithm; (2) to measure the agreement between the data abstracted by the algorithm and the traditional manual abstraction method; and (3) to perform a detailed error analysis to outline the next steps. The purpose of this study was not to reach conclusions about the ‘pharmacogenetics’ of the endocrine therapy of breast cancer, but rather to test an enabling high-throughput EMR automated tool to obtain endocrine treatment patterns.

MATERIALS AND METHODS

**Patient cohort**

The Mayo Clinic Breast Cancer study is an on-going cohort study initiated in February 2001 at Mayo Clinic, Rochester, Minnesota. Patients (women aged 18 years or over from Minnesota, Iowa, Wisconsin, Illinois, North Dakota, and South Dakota diagnosed with histologically confirmed primary invasive breast cancer and no prior history of another cancer (excluding non-melanoma skin cancer)) were enrolled within 6 months of diagnosis. Patients’ written informed consent was obtained. All aspects of these studies were approved by the Mayo IRB. The cohort used in this analysis consisted of 1507 women with stage I, II, or III breast cancer.

**Manual data abstraction**

A trained registered nurse abstracted the treatment information for tamoxifen and/or aromatase inhibitor (AI) cycles from the patient’s chart. Table 1 lists the relevant treatments. Information for 14 cases was abstracted from paper histories, typically notes from the local physicians who treated patients after the patients returned to their homes. Based on the information in these sources, each patient was assigned to one treatment category:

- **Category 0:** no tamoxifen or AI
- **Category 1:** Tamoxifen only
- **Category 2:** AI only
- **Category 3:** Tamoxifen then AI
- **Category 4:** AI then tamoxifen
- **Category 5:** multiple AIs and tamoxifen cycles in no specific order
- **Category 6:** no specific dates could be attributed to the AI and tamoxifen treatment/s
- **Category 7:** no confirmation of whether the patient has received either tamoxifen or AI treatment.

### Table 1 List of tamoxifen and aromatase inhibitor medications

<table>
<thead>
<tr>
<th>Tamoxifen drugs and synonyms (yes/no inclusion in RxNorm)</th>
<th>Aromatase inhibitors (yes/no inclusion in RxNorm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tamoxifen (yes)</td>
<td>Anastrozole (yes)</td>
</tr>
<tr>
<td>Fentamox (yes)</td>
<td>Tamifen (yes)</td>
</tr>
<tr>
<td>Nolvadex (yes)</td>
<td>Femara (yes)</td>
</tr>
<tr>
<td>Emblox (yes)</td>
<td>Letrozole (yes)</td>
</tr>
<tr>
<td>Tamofen (yes)</td>
<td>Nolvadex (yes)</td>
</tr>
<tr>
<td>Oestrifen (yes)</td>
<td>Anastrozole (yes)</td>
</tr>
<tr>
<td>Noctam (yes)</td>
<td>Anastrozole (yes)</td>
</tr>
<tr>
<td>Fentamox (yes)</td>
<td>Letrozole (yes)</td>
</tr>
<tr>
<td>Tamox (no)</td>
<td>Femara (yes)</td>
</tr>
<tr>
<td>Tam (no)</td>
<td>Oestrifen (yes)</td>
</tr>
</tbody>
</table>

**Electronic medical record**

The Mayo Clinic has utilized a comprehensive EMR since 2002. Some EMR data types are the free-text clinical narrative generated by the practitioners at the point of care, laboratory results, and the electronic prescribing database (Orders97) with fields for drug name, generic drug name, prescription dates, dose, refills, quantity, frequency, route, and duration. We used the research mirror of Mayo’s EMR, the EDT.\(^8\) Orders97 provides a medication history as part of outpatient clinical notes and is a means to enter prescriptions related to all types of services including ophthalmology (eye glasses) and other general applications (wheel chairs), in addition to medications.

**Clinical Text Analysis and Knowledge Extraction System (cTAKES)**

cTAKES (http://ohnlp.sourceforge.net/viewvc/ohnlp/trunk/) is an open-source, general purpose, modular, extensible NLP software for processing the clinical narrative.\(^11\) cTAKES applications to mining the clinical narrative are presented elsewhere.\(^12\–15\) cTAKES processes clinical notes and identifies the following clinical named entity mentions (NEs): Drugs, Diseases/Disorders, Signs/Symptoms, Anatomical sites, and Procedures. Each discovered NE is assigned attributes such as text span, ontology mapping code (UMLS,\(^16\) SNOMED CT,\(^17\) RxNorm\(^18\)), context (family history of, current, unrelated to patient), and a negation indicator. Version 1.1.0 of the cTAKES platform consists of the following annotators in this specific order:

- Sentence boundary detector, a wrapper around OpenNLP’s sentence boundary detector,\(^19\) but trained on clinical data
- Rule-based tokenizer to separate punctuations from words
- Normalizer, a wrapper around LVG\(^20\) to standardize, for example, morphologically different phrases with the same meaning, for example, ‘infection,’ ‘infesting,’ ‘infects’ normalize to the same form of ‘infect’

---

\(^{1}\) JAMA editorial, Jha (2011) discusses the promises of the EMR, emphasizing the importance of NLP as an enabling tool for accessing the vast information residing in the EMR, and stated that ‘federal government can play a helpful role by funding the basic research needed to launch this field forward.’

\(^{2}\) Patients (women aged 18 years or over from Minnesota, Iowa, Wisconsin, Illinois, North Dakota, and South Dakota diagnosed with histologically confirmed primary invasive breast cancer and no prior history of another cancer (excluding non-melanoma skin cancer)) were enrolled within 6 months of diagnosis. Patients’ written informed consent was obtained. All aspects of these studies were approved by the Mayo IRB. The cohort used in this analysis consisted of 1507 women with stage I, II, or III breast cancer.

\(^{3}\) A trained registered nurse abstracted the treatment information for tamoxifen and/or aromatase inhibitor (AI) cycles from the patient’s chart. Table 1 lists the relevant treatments. Information for 14 cases was abstracted from paper histories, typically notes from the local physicians who treated patients after the patients returned to their homes.

\(^{4}\) Each discovered NE is assigned attributes such as text span, ontology mapping code (UMLS, SNOMED CT, RxNorm), context (family history of, current, unrelated to patient), and a negation indicator. Version 1.1.0 of the cTAKES platform consists of the following annotators in this specific order.

---

**References**

1. **Category 0:** no tamoxifen or AI
2. **Category 1:** Tamoxifen only
3. **Category 2:** AI only
4. **Category 3:** Tamoxifen then AI
5. **Category 4:** AI then tamoxifen
6. **Category 5:** multiple AIs and tamoxifen cycles in no specific order
7. **Category 6:** no specific dates could be attributed to the AI and tamoxifen treatment/s
8. **Category 7:** no confirmation of whether the patient has received either tamoxifen or AI treatment.
Context dependent tokenizer grouping tokens to create Date, Time, Fraction, Range, Measurement, and PersonTitle groupings

Part-of-speech tagger, a wrapper around OpenNLP’s but trained on clinical data. The module assigns a part-of-speech label to each word

Phrasal chunker, a wrapper around OpenNLP’s but trained on clinical data to detect phrases such as noun phrases, verb phrases, and prepositional phrases

Dictionary lookup annotator which performs a permutational lookup against a dictionary database. The permutations are computed off the components within the user-specified lookup window (usually the noun phrase)

Context annotator based on NegEx21 to link information with the patient or family members

Negation detector based on NegEx to discover whether a concept is negated or not

Dependency parser to detect dependency relationships

Drug mention annotator populating a drug template

Figure 1 cTAKES expansion for drug-specific attributes at the document level.

The associated code primary attribute is reserved for the RxNorm18 code while the associated code secondary attribute represents an additional mapping to medication classes, for example, the National Drug File — Reference Terminology25 class. RxNorm terminology provides normalized names for clinical drugs used in pharmacy management. Each drug links to a unique code, for example, tamoxifen is assigned an RxNorm code of 10324. RxNorm, however, does not provide a list of unofficial abbreviations that practitioners might use in their daily practice and care management documentation, for example, the commonly used in clinical notes abbreviation ‘tam.’ The list of tamoxifen and AI medications consisted of 11 and seven items, respectively (table 1). Two terms (‘tam’ and ‘tamox’) not found in the RxNorm (UMLS26 version 2009AB) were added to the dictionary.

Figure 1 cTAKES expansion for drug-specific attributes at the document level.

Drug mention class
- drug mention text: Drug Mention Element
- associated code primary: Associated Code Element
- associated code secondary: Associated Code Element
- context: Context Element
- negation: boolean
- start date: Start Date Element
- end date: End Date Element
- dosage: Dosage Element
- frequency: Frequency Element
- frequency unit: Frequency Unit Element
- duration: Duration Element
- route: Route Element
- form: Form Element
- change status: Drug Change Status Element
- strength: Strength Element

Values extracted from text:
Tamoxifen 20 mg po daily started on March 1, 2005
- drug mention text: Tamoxifen
- associated code primary: C0351245
- associated code secondary: null
- context: current
- negation: false
- start date: March 1, 2005
- end date: null
- dosage: 1.0
- frequency: 1.0
- frequency unit: daily
- duration: null
- route: Enteral Orally
- form: null
- change status: noChange
- strength: 20 mg
Research and applications

Extraction of dosage attribute
Dosage refers to how many of each drug the patient is taking. Any numeric text description or a number value adjacent to a Strength or Frequency mention will populate Dosage with its value. For instance, in the phrase ‘Take two 325-mg aspirin tablets’ the ‘two’ would indicate the Dosage attribute and ‘325 mg’ would indicate the Strength.

Extraction of frequency attribute
Frequency describes how often the patient needs to take the drug. It consists of a frequency number and frequency unit. Periodic phrases coupled with numeric text values or number values adjacent to Drug named entity, Strength, and Frequency terms are discovered as Frequency. This also includes common Latin terms (eg, ‘b.i.d.,’ ‘p.r.n.,’ and ‘qhs’) and hyphenated phrases (eg, ‘every-other-evening,’ ‘as-needed,’ and ‘once-a-day’).

Extraction of route attribute
Route refers to the way that the drug is introduced into the body. Phrases that are equivalent or synonymous of the following are included: topical, oral, gastric, rectal, intravenous, intra-arterial, intramuscular, intracardiac, subcutaneous, intrathecal, intraperitoneal, transdermal, and transmucosal.

Extraction of form attribute
Form describes the physical appearance of a drug. Terms include aerosol, capsule, cream, elixir, emulsion, enema, gel, implant, inhalant, injection, liquid, lotion, lozenge, ointment, patch, pill, powder, shampoo, soap, solution, spray, suppository, syrup, and tablet.

Extraction of change status attribute
Change status refers to whether the medication is currently being taken or not and the change associated with this. Its values are started, stopped, increased, decreased, and no change. For example, the extraction from ‘increased Zoloft from 5 mg to 10 mg’ results in two drug annotations: (1) Drug mention=Zoloft,’ context=current, change status=increased, strength=10 mg; and (2) Drug mention=Zoloft,’ context=history of, change status=null, strength=5 mg.

Extraction of strength attribute
Strength is a two word text span of, typically, number and unit. It is subdivided into strength number and strength unit, for example in the span ‘325 mg,’ 325 is the number and mg is the unit.

Estimation of confidence in data extracted
A confidence metric that the medication has been prescribed is used for the Impression/Report/Plan (I/R/P) and History of Present Illness (HPI) sections based on the presence of drug-related attributes near the drug mention. For instance, if the I/R/P section contains ‘tamoxifen 20 mg daily,’ the presence of both ‘20 mg’ and ‘daily’ boosts the confidence. Any drug-related named entity will be assigned a base value of 0.05 for the Confidence initially; however, this value will be set to 0.15 if there is an adjacent Strength discovered. Additionally, 0.05 will be added if the presence of the Dosage attribute is determined. This value will increase with the addition of other drug elements if discovered. ‘Confidence’ will increase by a factor of 1.5 if Form, Route, or Duration attributes are found, and it will increase by a factor of 1.5 if Frequency and Frequency Unit are discovered. The maximum confidence would be just under 1.0 (0.989=(0.05 +0.15)×1.3×1.3×1.3×1.5×1.5) in the event that all elements of the drug named entity were discovered (see figure 2 for examples). The weights of these attributes are used to reflect the scale of the likelihood that this attribute relates to a drug term, so the presence of Form/Route/Duration is slightly less weighted than Frequency/Frequency Unit. The ‘Confidence’ value threshold that will determine if a mention will be extracted is 0.15 and above. The confidence scaling was determined empirically based on observations of the likelihood that a drug mention was related to the patient and not a hypothetical general discussion. The more detail about the drug mention, the higher the confidence that the mention is related to a present or recent prescription. In the absence of a medication dosage and strength information, at least four other signature elements are required to achieve the desired threshold.

Figure 2 for the resulting scores when different signature elements are discovered and the effect of each element on the final score (1.5 or 1.5 factor). The only combination not represented in the figure would be a drug mention with dosage and no strength, which is atypical.

Consideration of negation terms
The list of discovered medication mentions was additionally pruned to filter out negated AI and tamoxifen mentions. We extended the list of the negation words incorporated within the generic cTAKES negation detection algorithm.21 Terms signaling non-confirmed treatments included ‘suggest,’ ‘doubt,’ ‘discuss,’ ‘decide,’ ‘recommend,’ ‘talk,’ ‘plan,’ ‘think,’ ‘consider,’ and ‘considers’ along with their variants. For example, in ‘I discussed
with the patient tamoxifen treatment,’ ‘tamoxifen’ is flagged as a medication, but filtered out as it was in the context of a discussion only and not a confirmed treatment.

Component two: information extraction from the medication prescribing system

This component extracted prescription information from the electronic prescribing database (Orders97) through standard database queries. The same window used for the clinical notes was applied to obtain all prescription entries starting 2 weeks prior to the diagnosis of breast cancer to allow for slight deviations in the date of diagnosis. The start date for an Orders97 medication was the time the medication was prescribed. As typical for electronic prescription interfaces, Orders97 specifies no end date.

Component three: information merging

This component unified the information from the first two components to create the patient-level medication history. The final patient-level start date for a medication was determined by its first Orders97 entry if there was one, otherwise by the first start date of that medication as discovered from the clinical notes. If a start date existed for both the clinical document and the Orders97 entry, then the Orders97 date was used, unless the Orders97 date was more than 2 weeks earlier than the diagnosis date for this particular patient. In the event that the Orders97 information was earlier than 2 weeks, the clinical document date was used. The final end date for a medication was determined by the last stop date attribute if available, otherwise by the date of the last mention within the clinical notes if there was no 7+ month gap of inactivity. If there was a 7+ month gap, then a new start/end cycle was begun. The 7-month timeframe was considered the most reasonable from a clinical practice standpoint in that 6-month intervals are most common.

Component four: automatic treatment classification

This component applied the final logic based on the evidence from the third component. The algorithm collapsed into one category the data points for category 0 and 7 from the manually abstracted data because the distinctions were not entirely clear. Patients with no medication data are classified into category 0.

Evaluation metrics

We report results in terms of sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV). Accuracy is the ratio between the total number of system assignments that agree with the abstracter assignments and the total number of cases.

RESULTS

A total of 96835 clinical notes were identified for the 1507 patients, and 1384 (91.84%) patients had notes with at least one medication mention. The clinical notes of 619 (41.07%) patients contained at least one tamoxifen mention (6358 tamoxifen instances; mean 10.24, median 7). The clinical notes of 559 (35.77%) patients contained at least one aromatase inhibitor (AI) mention (6739 AI instances; mean 12.54, median 8). For patients with any Orders97 orders (1578 patients (91.48%)), there were a total of 20328 medication orders. Overall, 510 (35.86%) patients had at least one tamoxifen order (1557 tamoxifen orders, mean 2.62, median 2) and 446 (29.54%) patients had at least one AI order (1253 AI orders; mean 2.76, median 2).

Each clinical note document was processed through the enhanced cTAKES to discover medications from the following five sections: Impression/Report/Plan, History of Present Illness, Current medications, Admission medications, and Discharge medications. Processing took approximately 3 h and 15 min (Dual AMD Opteron 248 processors (64 bit) with 16 GB system RAM running the Linux-based Fedora Core 6 (64 bit) OS). Manual data abstraction is estimated to proceed at a rate of 25 charts per week by an experienced registered nurse abstracter, which for this study totals 15 weeks.

The accuracy of overall agreement for the results for treatment classification was 0.925. For categories 0, 1, 2, 3, and 4, the sensitivity and specificity ranged from 90.27% to 99.83%, while the predictive values ranged from 80.00% to 99.93% with lower results for categories 0 and 0 where complex treatment timelines presented a challenge (table 2).

Due to time limitations, we randomly selected 131 (63.6%) from the total of 206 disagreements between the automated system and the traditional manual abstraction method for review. Table 3 summarizes the disagreements. Some 75% (96 instances) were true algorithmic errors, and 27% (35 instances) were not algorithmic errors, of which 11% (14 instances) were due to the lack of data source access (eg, scanned documents or paper histories), 8% (10 instances) were misses by the abstracter, and 8% (11 instances) were disagreements due to the last manual abstraction date. The algorithmic errors were related to the discovery of complex temporality, medication attributes outside of the window, and confidence metric assignment. The cut-off date disagreements are such that the system abstracted information from documents after the last manual abstraction date.

DISCUSSION

Within an EMR environment, we built an algorithm for mining endocrine breast cancer treatment patterns that combines NLP and standard database queries. The algorithm has high
specificity for all categories, high sensitivity for five of the categories, and low sensitivity for two of the categories. Approaches like ours could assist accrual of large EMR cohorts for research as they are scalable and can be used both retrospectively and prospectively.

The EMR environment we used, EDT,\(^8\) is grounded in interoperability principles such as Health Level 7 Standards to ‘improve care delivery, optimize workflow, reduce ambiguity and enhance knowledge transfer among all stakeholders’ (http://www.hl7.org/). A direct implication of that

<table>
<thead>
<tr>
<th>Table 2</th>
<th>Agreement results between the two methods: (1) traditional manual abstraction (manual abstraction columns) and (2) automated tool combining information extracted from clinical free-text through natural language processing with information extracted from an electronic prescribing system (automated tool, all components rows)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Manual abstraction</td>
<td>True</td>
</tr>
<tr>
<td>Category 0 and 7: no tamoxifen or AI treatment, or no confirmation for tamoxifen or an AI</td>
<td>Automated tool (all components)</td>
</tr>
<tr>
<td>True</td>
<td>573</td>
</tr>
<tr>
<td>False</td>
<td>1</td>
</tr>
<tr>
<td>Total</td>
<td>574</td>
</tr>
<tr>
<td>Category 1: tamoxifen only</td>
<td>Automated tool (all components)</td>
</tr>
<tr>
<td>True</td>
<td>334</td>
</tr>
<tr>
<td>False</td>
<td>36</td>
</tr>
<tr>
<td>Total</td>
<td>370</td>
</tr>
<tr>
<td>Category 2: aromatase inhibitors only</td>
<td>Automated tool (all components)</td>
</tr>
<tr>
<td>True</td>
<td>305</td>
</tr>
<tr>
<td>False</td>
<td>28</td>
</tr>
<tr>
<td>Total</td>
<td>333</td>
</tr>
<tr>
<td>Category 3: tamoxifen followed by aromatase inhibitors</td>
<td>Automated tool (all components)</td>
</tr>
<tr>
<td>True</td>
<td>160</td>
</tr>
<tr>
<td>False</td>
<td>13</td>
</tr>
<tr>
<td>Total</td>
<td>173</td>
</tr>
<tr>
<td>Category 4: aromatase inhibitors followed by tamoxifen</td>
<td>Automated tool (all components)</td>
</tr>
<tr>
<td>True</td>
<td>4</td>
</tr>
<tr>
<td>False</td>
<td>13</td>
</tr>
<tr>
<td>Total</td>
<td>17</td>
</tr>
<tr>
<td>Category 5: multiple aromatase inhibitors and tamoxifen cycles in no specific order</td>
<td>Automated tool (all components)</td>
</tr>
<tr>
<td>True</td>
<td>0</td>
</tr>
<tr>
<td>False</td>
<td>4</td>
</tr>
<tr>
<td>Total</td>
<td>4</td>
</tr>
<tr>
<td>Category 6: no specific dates can be attributed to treatment, thus no sequencing timelines</td>
<td>Automated tool (all components)</td>
</tr>
<tr>
<td>True</td>
<td>0</td>
</tr>
<tr>
<td>False</td>
<td>4</td>
</tr>
<tr>
<td>Total</td>
<td>4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 3</th>
<th>Distribution of manually reviewed disagreements (a sample of a randomly selected 131 disagreements from a total of 206 disagreements)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatment category</td>
<td>Manually reviewed: algorithm errors</td>
</tr>
<tr>
<td>Category 0 and 7</td>
<td>25</td>
</tr>
<tr>
<td>Category 1</td>
<td>35</td>
</tr>
<tr>
<td>Category 2</td>
<td>18</td>
</tr>
<tr>
<td>Category 3</td>
<td>5</td>
</tr>
<tr>
<td>Category 4</td>
<td>0</td>
</tr>
<tr>
<td>Category 5</td>
<td>9</td>
</tr>
<tr>
<td>Category 6</td>
<td>4</td>
</tr>
<tr>
<td>Total</td>
<td>96</td>
</tr>
</tbody>
</table>
implementation is the clear boundaries of the clinical note sections enabling unambiguous phenotype extraction, for example, limiting medication extraction to Impression/Report/Plan, History of Present Illness, Current Medications, Admission Medications, and Discharge Medications sections. Such an environment allows data sharing and phenotype extraction not only across patients, but across institutions and healthcare providers. Our approach is disease-agnostic and could be applied to other diseases after a modification of the medication dictionary. We also hope to contribute to a shift where each practitioner starts to view the EMR as a valuable data resource with a life that extends beyond purely patient care and administrative tasks.

Our study emphasizes important technical challenges for information extraction from the clinical narrative such as clinical events temporality. Another challenge includes incomplete event documentation. If there is insufficient evidence and/or heavy reliance on background knowledge and inference, the current NLP-based system does not have any recovery functionalities.

CONCLUSION

Our goal in this study was to demonstrate a secondary and meaningful use of the EMR data generated at the point of care, coupled with information technologies for high-throughput phenotyping for clinical translational research. We chose a breast cancer pharmacogenomic study for our model system. We showed that data harvested from the EMR in an automated fashion are comparable to those manually abstracted by a trained registered nurse abstractor. Next steps will include the incorporation of additional modules for chemotherapy, radiation therapy, breast cancer recurrence, and other types of cancer diagnosis extraction as well as drugs that are CYP2D6 inhibitors, which are known to affect tamoxifen metabolism—all these extracted from the EMR using NLP and standard database queries. Our hope is that studies like ours will enable the EMR to achieve its full potential as an indispensable analytical aid for both practitioners and biomedical investigators to improve patient care and accelerate translational research.

Funding

This work was funded by an NCI DCCPS supplement grant (P Olson), grants U01 GM61388 (P Weihsliboum), U01 HG 04599 (PI Chute), RO1 CA122340 (PI Couch), Breast SPORE CA 116201 (PI Igle), and a SHARP award (90TR0002) (PI Chute).

Competing interests

None.

Provenance and peer review

Not commissioned; externally peer reviewed.

Data sharing statement

The data used in the study has been shared within the team. The software code is released open source at the cTAKES website: http://chelp.svn.sourceforge.net/viewvc/ohnlp/trunk/cTAKES/

REFERENCES