Discovery and Clinical Decision Support for Personalized Healthcare

Jinsung Yoon, Camelia Davtyan, *MD, FACP*, and Mihaela van der Schaar, *Fellow, IEEE*

Abstract— **With the advent of electronic health records, more data is continuously collected for individual patients and more data is available for review from past patients. Despite this, it has not yet been possible to successfully use this data to systematically build clinical decision support systems that can produce personalized clinical recommendations to assist clinicians in providing individualized healthcare. In this paper, we present a novel approach, Discovery Engine (DE), that discovers which patient characteristics are most relevant for predicting the correct diagnosis and/or recommending the best treatment regimen for each patient. We demonstrate the performance of DE in two clinical settings: diagnosis of breast cancer and personalized recommendation for a specific chemotherapy regimen for breast cancer patients. For each distinct clinical recommendation, different patient features are relevant; DE can discover these different relevant features and use them to recommend personalized clinical decisions. For treatment (chemotherapy regimens) recommendations, we determine *transfer rewards* based on the outcomes of external knowledge (published literature and clinical practice guidelines) describing the effect of the same treatments on "similar" patients. DE approach achieves a 16.6% improvement over existing state-of-the-art recommendation algorithms in terms of kappa coefficients for recommending the personalized chemotherapy regimens. For diagnostic predictions, DE approach achieves a 2.18% and 4.24% improvement over existing state-of-the-art prediction algorithms in terms of prediction error rate and false positive rate, respectively. We also demonstrate that the performance of our approach is robust against missing information, and that the relevant features discovered by DE are confirmed by clinical references.**

Index Terms— **Clinical decision support systems, diagnosis decision support systems, relevant feature selection, healthcare informatics, personalized treatment.**

I. Introduction

CLINICIANS are routinely faced with the practical challenge of integrating high-dimensional clinical data in order to recommend the most appropriate clinical decision for a given patient [1]. As the understanding of complex diseases progresses, the types of available tests and treatments diversify and, as a result, the difficulty of recommending the optimal clinical decision for a particular patient increases as well. Current clinical decisions continue to rely on clinical practice guidelines which, in cases when scientific analysis and evidence is scarce, are largely based on clinical experience and opinion. Also, current clinical practice guidelines are aimed at a "representative" patient rather than an individual patient who may display other relevant characteristics. Such "representative" guidelines may thus miss the opportunity to consider personal traits when recommending clinical decisions. For example, the American Cancer Society (ACS) recently issued new guidelines which suggested that women with an average risk of breast cancer should start having mammograms at an age of 45 (five years later than ACS had previously recommended) [2]. However, women who have certain risk factors (family history of breast cancer, no children, etc.) have a higher risk of developing breast cancer, and they would benefit from having mammograms at an earlier age. In cases such as these, the ACS guidelines recommend that a high risk patient consults with her physician to determine an appropriate screening age and interval, which is based on that particular physician's experience and opinion. Moreover, statistics show that diagnostic errors result in 10% of patient deaths and represent the most frequent type of medical malpractice claims in the United States [3]. This reality highly underscores the urgent need for building smart clinical decision support systems (CDSS) and diagnosis decision support systems (DDSS) that can assist clinicians in making accurate, personalized clinical recommendations [4]. It has been recently recognized [5] that medical informatics tools and machine learning techniques can be successfully used to provide recommendations for personalized diagnosis and treatment.

The goal of this paper is to develop methods that will enable CDSS and DDSS to personalize their recommendations based on individual patient characteristics. The wealth of information being routinely collected as part of the electronic health record (EHR) provides an unprecedented opportunity to discover appropriate clinical recommendations for patients given historical information about the clinical decisions administered to similar patients and their actual outcomes [6]. However, using this information is difficult precisely because there is so much of it; the solution is to extract only the relevant information for the particular patient and the relevant clinical decisions previously used for similar patients among the wealth of available information. Extracting only the relevant information is important because using irrelevant features can significantly hurt the performance of the system, unnecessarily increase its complexity, and decrease its learning/adaptation speed [7]. Furthermore, efficient discovery of relevant patient

J. Yoon, and M. van der Schaar are with the Department of Electrical Engineering, University of California Los Angeles, Los Angeles, CA 90095, USA (email: jsyoon0823@ucla.edu, mihaela@ee.ucla.edu).

C. Davtyan is with the Department of Medicine, Geffen School of Medicine at UCLA Los Angeles, CA 90095, USA (email: CDavtyan@mednet.ucla.edu).

features can help clinicians focus on the relevant information available about the patient without having to sift through a large patient record.

In this paper, we present a novel approach called Discovery Engine which optimizes clinical recommendations by identifying the features in the patient record that differentiate the individuals who receive a certain clinical decision and respond positively from those who do not. Our approach utilizes the available contextual information about patients and learns from the large quantities of observational clinical data to inform clinical recommendations and makes better decisions by learning from similar patients. We show that our DE approach consistently outperforms existing state-of-the-art machine learning algorithms in terms of matching individual patients to the optimal treatment regimens, and diagnosis accuracy.

One of the biggest challenges faced by this class of recommendation systems is that the rewards/actual outcomes of clinical decisions (e.g. five-year recurrence free survival) are usually not available [8]. Moreover, even if rewards/actual outcomes of the clinical decisions were available, the counterfactuals – rewards/actual outcomes of alternative clinical decisions that were not used – are never available. What is available, however, is a large medical literature that reports the results of a wide range of clinical studies, including different types of patients, different patient characteristics, different types of clinical decisions, and the actual outcomes of these decisions. We use the results of these studies to construct transfer rewards, which we use as proxies for rewards. This allows us to train the DE algorithms as well as to evaluate their performance in comparison to existing methods when the actual outcomes cannot be achieved. The four primary contributions of this paper are as follows:

- We describe a novel approach for discovering the most relevant information from the EHR that distinguishes between patients that should receive one particular clinical decision and the patients who should receive another. For instance, premenopausal breast cancer patients are more likely to respond to a specific type of chemotherapy such as CEF [9].
- Using the past records in the EHR and external knowledge from the medical literature, our approach discovers the optimal personalized clinical decision based on the discovered relevant information (e.g. their clinical test result, treatment history, and outcomes).
- In lieu of having actual reward values associated with clinical decisions, we define the transfer rewards, a method for estimating actual outcomes described in external knowledge based on their similarities to individual patients given reported characteristics.
- We apply DE to two medical applications: 1. Personalized treatment recommendations (chemotherapy regimens) for breast cancer patients and 2. Diagnosis of breast cancer. DE is used to discover which features are relevant to make a distinct clinical decision and then uses this knowledge to build a clinical decision recommendation system. We evaluate the performance of DE in the context of breast-cancer diagnosis and treatment, and show that it consistently and significantly outperforms state-of-the-art machine learning algorithms.

II. Related Works

A. *Personalized Clinical Decision Support System*

Current medical practice relies on manually curated systematic reviews of the available scientific evidence and clinical guidelines that provide recommendations for large groups of patients rather than personalized recommendations that are tailored to individual patients. Clinical decision support systems have been proposed before, but many of them do not consider the specific characteristics of patients and do not provide personalized clinical recommendations; hence, they are not very accurate and have only limited applicability in practice [10-11]. Moreover, clinicians often refer to the medical literature available through Medline/PubMed, VisualDX, and UpToDate to help them associate observed finding with possible conditions and recommended decisions. However, these resources are also not customized to a specific patient's case.

The advent of Big Data has been identified as both an opportunity and a challenge for the design of CDSS [12]. A large literature has used data-driven approaches to develop *representative* rather than *personalized* healthcare systems [13]. A smaller literature is dedicated to developing personalized CDSS. However, most existing papers in this strand of literature [14]-[20] either just discuss opportunities rather than propose a concrete algorithm or apply off-the-shelf machine learning techniques to the considered medical problem while disregarding the unique characteristics and challenges of developing personalized CDSS. Diagnosis decision support systems have been developed for cardiovascular diseases and diabetes using ensemble learning [21], SVMs [22], artificial neural networks [23] or rule-based algorithms [24]. Although some diagnosis decision support systems issue accurate diagnostic recommendations for certain specialized diseases, most of them are based on a small number of *manually* selected features [25-26]. Whenever the number of features (contexts) is large, these methods fail to perform well [27].

Most importantly, most of the proposed CDSS solely focus on diagnosis recommendations and do not provide solutions for the equally important problem of treatment recommendation. A small number of studies does attempt to propose CDSS for treatment recommendation [28]. However, these CDSS differ significantly from DE. For instance, [29] proposes antibiotics recommendation systems for representative patients and it does not use a data-driven approach. [30] proposes a CDSS for personalized medicine recommendation, but which uses similarity information among drugs and patients that is specific to the study at hand and cannot be easily applied to other diseases - the similarity between patients is solely based on the ICD 9 feature and the similarity between drugs is based on their chemical structure and protein structure. Hence, the methods in [31] are not widely applicable to a diverse set of patients and treatments.

In contrast, DE can robustly issue an accurate clinical decision (e.g. treatment or diagnosis) for patients for a variety of complex diseases. As we will show in our experimental section, it can perform well even when the number of features based on which it needs to make a decision is large because it adopts a novel method for discovering the different features that are most relevant to consider when deciding on different diagnosis or treatment options for a specific patient. For this, we develop a customized feature selection and decision making system for which we show that it significantly outperforms existing off-the-shelf techniques. Importantly, based on the authors' knowledge, DE is the first personalized CDSS which is able to discover which different features of a patient are indicative of the success of a treatment in a complex disease such as breast cancer.

Our work is also related to the field of medical informatics dedicated to improving breast cancer diagnosis and treatment. However, our work is distinct from these works. For instance, some of the works analyze the impact of specific patient features such as genetic information or imaging information extracted from mammograms or other imaging to improve the performance of CDSS [31-33]. However, these CDSS rely on conventional machine learning methods such as decision trees, Naïve Bayes, SVMs, multi-armed bandits, which are shown to work poorly in our setting. Moreover, we would like to highlight that the presented DE system is based on a novel set of machine learning methods developed especially for personalized diagnosis and treatment discovery which are shown in later sections to significantly outperform existing methods that are generic. Moreover, DE is applicable not only to breast cancer, but it is general and can be applied to discover personalized treatment for other complex diseases.

B. Relevant Feature Selection

Another strand of literature related to this work is relevant feature selection algorithms including Correlation Feature Selection (CFS) and Mutual Information Feature Selection (MIFS) [34-35]. These are related to our feature selection approach. However, our clinical decision dependent feature selection algorithm (CDFS) is very different from existing feature selection algorithms which focus on the patients' characteristics and not on how these characteristics distinctively impact different clinical decisions: our approach is capable of discovering *different* features that are relevant to distinct *different* clinical decisions. This makes CDFS similar to our prior work [36-38] – the RELEAF algorithm. However, unlike RELEAF, which is very slow because it must compare all possible combinations of features, CDFS is able to discover the relevant features in a very fast and efficient manner because it adopts a sequential feature selection approach. This sequential approach significantly reduces the complexity of the feature selection algorithm with pairwise dependent features.

C. Machine Learning Techniques

The other strand of literature related to this work is that on machine learning techniques, including Support Vector Machines (SVMs), AdaBoost, logistic regression, decision trees etc. which are not able to accurately capture the nuanced relationships between specific patient characteristics and specific clinical decisions, and are thereby not able to issue accurate personalized decision recommendations. (Comparisons against these methods are performed in the experiments sections [38].) On the other hand, our DE approach can learn the relevant features of patients that make them respond to specific clinical decisions, achieving better performance as a result.

Our method also exhibits similarities to the contextual multi-armed bandit problem (MAB). However, contextual MABs are very inefficient when the number of contexts (in our case patient features) is large (see the experimental section of [37]). DE is able to successfully deal with the curse of dimensionality by discovering what information is relevant and making clinical decisions based only on relevant contexts rather than the entire set of contexts that can be extracted from the EHR, much of which is irrelevant to the decisions of whether to administer a specific treatment or not.

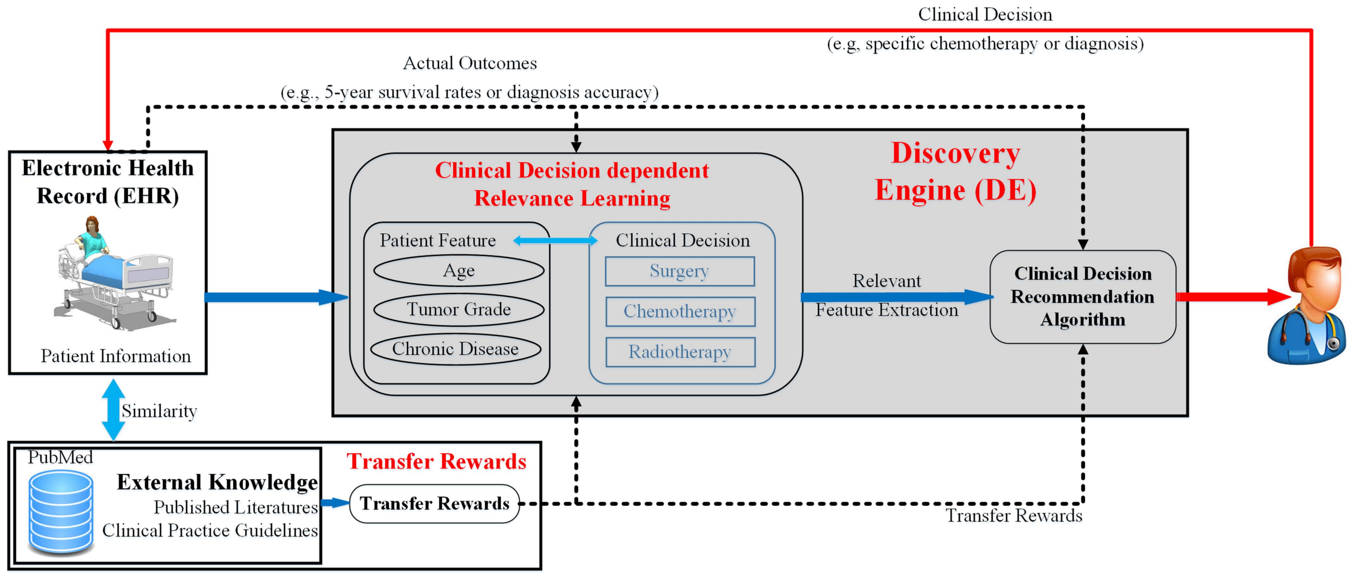


Fig. 1. Personalized clinical decision support system using discovery engine (DE)

III. System Model

In this section, we introduce our method - Discovery Engine. Fig. 1. depicts the proposed system, which issues a personalized clinical recommendation to the physician about certain patients. If the actual outcomes of certain clinical decisions are available, DE uses these outcomes as a reward to train itself. Otherwise, DE relies on a transfer reward (which is discussed in Section V) to measure the estimated reward of certain clinical decisions. While the proposed system is applicable in general, we illustrate its use in the context of breast cancer. Let $\boldsymbol{x} = (x_1, x_2, \dots\dots, x_D)$ denote the patient information where D is the total number of patient features such as age, tumor size, estrogen receptor information etc.; $a \in A \triangleq \{a_1, a_2, \dots\dots, a_K\}$ denotes the clinical decision (e.g. chemotherapy regimes or breast cancer diagnosis) that is recommended to the patient. Each patient feature is denoted as$f \in F \triangleq \{f_1, f_2, \dots\dots, f_D\}$. The reward y would ideally be derived as the five-year survival rate of a patient given the treatment regimens if it is available. However, obtaining these rewards is usually difficult in practice; instead, we use a transfer reward measure which is discussed in a subsequent section. Let $\boldsymbol{x}(n), a(n), y(n)$ be the patient information, clinical decision and reward of *n-th* patient and $\boldsymbol{\mathcal{H}}_N = (\boldsymbol{x}(n), a(n), y(n))_{n=1}^{N}$ be the information available for the N previously seen patients. (This represents the training set.)

The outcomes of a clinical decision a do not depend on all the features [39]: we assume that the outcomes of a clinical decision a depend only on a subset of features $\mathcal{R}(a) \subseteq F$ which we call the relevant features. Let $\mathcal{R} = \cup_{a \epsilon A} \mathcal{R}(a)$ be the set of all relevant features. A key challenge is that the relevant features for predicting the outcomes of a clinical decision are not known a priori; they need to be discovered. Note that such a discovery method is very different from existing feature selection algorithms which focus only on the patients' characteristics and not on how these characteristics differently impact/inform the outcomes of different clinical decisions: our approach is capable of discovering different features that are relevant to different clinical decisions. We say that $\mathcal{R}(a)$ is relevant/informative for clinical decision a if the expected reward only depends on the information contained in $\mathcal{R}(a)$.

Our goal is to discover the relevant features of each clinical decision a (this may be different for each decision) and recommend the optimal clinical decision that corresponds to the discovered relevant patient information. The optimal recommended clinical decision is given by

$$a^*(\boldsymbol{x_R}) \triangleq arg \max_a \mathbb{E}_{y|a,\boldsymbol{x}_{\mathcal{R}(a)}}(y|a, \boldsymbol{x}_{\mathcal{R}(a)})$$

where $a^*(\boldsymbol{x_R})$ is the clinical decision that yields the maximum expected reward (e.g., best expected patient outcome) for a patient characterized by the relevant features $\boldsymbol{x_R}$. Nomenclature table in Appendix is the summary of variables used in the system model and the algorithm sections.

IV. Algorithms

Discovery engine (DE) consists of two algorithms: a clinical decision dependent feature selection algorithm (CDFS) and a clinical decision recommendation algorithm. First, DE discovers different relevant features for different clinical decisions using CDFS. Based on the discovered features by CDFS, DE recommends the optimal clinical decision for each patient. Detailed steps of each algorithm are described in the following subsections.

A. Clinical Decision Dependent Feature Selection (CDFS)

The proposed CDFS algorithm sequentially discovers the relevant features which yield maximum relevance to the specific clinical decision with minimum redundancy (compared with the previously discovered relevant features).

To describe CDFS, we start by introducing a few notations. Let $\hat{y}_a^S(\boldsymbol{x_S})$ and $N_a^S(\boldsymbol{x_S})$ be the sample mean rewards estimator and the number of patients (whose feature information contains $\boldsymbol{x_S}$ and was provided clinical decision a). Let $\hat{y}_a$ and N_a be the sample mean rewards estimator and the number of patients who received the clinical decision a.

First, we define a relevance metric $h_f^r(a)$ which measures how the expected reward for patients having the feature x_f differs from that obtained for the entire set of patients in $\boldsymbol{\mathcal{H}}_N$(previously defined in system model section) when clinical decision a is assigned. We formalize this as:

$$h_f^r(a) \triangleq \sum_{x_f} \frac{N_a^f(x_f)}{N_a} | \hat{y}_a^f(x_f) - \hat{y}_a| \tag{1}$$

Second, we define a redundancy metric $h_{f,s}^d(a)$ which measures how the expected reward made for a given patient is affected by considering an additional feature x_f when clinical decision a is recommended. We formalize this as:

$$h_{f,s}^d(a) = -\sum_{x_f, x_s} \frac{N_a^{f,s}(x_f, x_s)}{N_a} [\hat{y}_a^{f,s}(x_f, x_s) - \hat{y}_a^s(x_s)] \tag{2}$$

Third, we use the minimum redundancy maximum relevance (mRMR) criterion [34] to combine the above metrics to sequentially discover relevant feature sets for each clinical decision. Let us define $\mathcal{U}_f(a)$ as the utility obtained if feature x_f is additionally selected as a relevant feature for clinical decision a. If $\hat{\mathcal{R}}(a)$ is the relevant features set previously discovered by CDFS for clinical decision a, the utility function $\mathcal{U}_f(a)$ is defined as:

$$\mathcal{U}_f(a) = h_f^r(a) - \frac{1}{|\hat{\mathcal{R}}(a)|} \sum_{s \in \hat{\mathcal{R}}(a)} h_{f,s}^d(a) \tag{3}$$

where $1/|\hat{\mathcal{R}}(a)|$ is used as a normalization factor. The main steps of the CDFS are outlined below:

Step 1: Define $\hat{\mathcal{R}}(a)$ as the relevant feature set discovered by CDFS for clinical decision a and $\hat{\mathcal{R}}^c(a)$ as the complementary set of $\hat{\mathcal{R}}(a)$. G and H are defined as the selected relevant features in step 2 and step 3, respectively. For each clinical decision a, initialize $\hat{\mathcal{R}}(a)$ as the empty set (i.e. $\emptyset$) and $\hat{\mathcal{R}}^c(a)$ as the set of all features (i.e. F).

Step 2: The algorithm selects the first relevant feature which maximizes the relevance metric ($h_f^r(a)$), i.e.,

$$\mathrm{G} = arg \max_{f \in \hat{\mathcal{R}}^c(a)} h_f^r(a)$$

$$\hat{\mathcal{R}}(a) \leftarrow \hat{\mathcal{R}}(a) \cup \mathrm{G}$$

Step 3: The algorithm finds the subsequent relevant feature that maximizes utility function ($\mathcal{U}_f(a)$), i.e.,

$$\mathrm{H} = arg \max_{f \in \hat{\mathcal{R}}^c(a)} \mathcal{U}_f(a)$$

$$\hat{\mathcal{R}}(a) \leftarrow \hat{\mathcal{R}}(a) \cup \mathrm{H}$$

Step 4: The algorithm iteratively runs Step 3 until the utility function $\mathcal{U}_f(a)$ is less than threshold cost C, where C is an input parameter for the algorithm, i.e.,

$$\text{If} \max_{f \in \hat{\mathcal{R}}^c(a)} \mathcal{U}_f(a) < C,$$
$$\text{then,} \quad \hat{\mathcal{R}}(a) = \hat{\mathcal{R}}(a)$$

Therefore, the number of relevant features for each clinical decision is depend on the threshold cost C. The pseudo-code of CDFS is given in the upper part of the Algorithm 1.

Algorithm 1 Discovery Engine (DE)

```
Input: $C_{TH}, C$
Initialize: $\hat{\mathcal{R}}(a) = \emptyset, \hat{\mathcal{R}}^c(a) = \{f_1, f_2, \dots \dots, f_D\}$ for each $a$
CDFS:
for each clinical decision $a$
    $\mathrm{G} = arg \max_{f \in \hat{\mathcal{R}}^c(a)} h_f^r(a)$
    $\hat{\mathcal{R}}(a) \leftarrow \hat{\mathcal{R}}(a) \cup \mathrm{G}$
    while ($\max_{f \in \hat{\mathcal{R}}^c(a)} \mathcal{U}_f(a) > C$)
        $\mathrm{H} = arg \max_{f \in \hat{\mathcal{R}}^c(a)} \mathcal{U}_f(a)$
        $\hat{\mathcal{R}}(a) \leftarrow \hat{\mathcal{R}}(a) \cup \mathrm{H}$
    end while
end for
Optimal clinical decision recommendation:
$U = \{a \in A | N_a^{\hat{\mathcal{R}}(a)}(\boldsymbol{x}_{\hat{\mathcal{R}}(a)}) < C_{TH} \cdot \log(n)\}$
if ($U = \emptyset$)
    $\hat{a}(\boldsymbol{x}) = arg \max_a \hat{y}_a^{\hat{\mathcal{R}}(a)}(\boldsymbol{x}_{\hat{\mathcal{R}}(a)})$
end if
Update: $N_a^{\hat{\mathcal{R}}(a)}(\boldsymbol{x}_{\hat{\mathcal{R}}(a)}), \hat{y}_a^{\hat{\mathcal{R}}(a)}(\boldsymbol{x}_{\hat{\mathcal{R}}(a)})$ based on the rewards
```

B. Clinical Decision Recommendation Algorithm

The proposed clinical decision recommendation algorithm is a modified contextual multi-armed bandit algorithm [40-41] which only uses the relevant contexts (features) discovered by CDFS to recommend the optimal clinical decision which maximizes the estimated patient outcomes. The main steps of the recommendation algorithm are outlined below:

Step 1: Find the set of unresolved clinical decisions (*U*) for the patient with information vector $\boldsymbol{x}_{\hat{\mathcal{R}}(a)}$:

$$U = \left\{a \in A \middle| N_a^{\hat{\mathcal{R}}(a)}(\boldsymbol{x}_{\hat{\mathcal{R}}(a)}) < C_{TH} \cdot \log(n)\right\} \tag{4}$$

where $C_{TH} \cdot \log(n)$ is a control function. C_{TH} is an input parameter and *n* is the total number of previous patients. If there exist a set of unresolved clinical decisions, DE abstains from making clinical decision recommendation and only updates $N_a^{\hat{\mathcal{R}}(a)}(\boldsymbol{x}_{\hat{\mathcal{R}}(a)})$ and $\hat{y}_a^{\hat{\mathcal{R}}(a)}(\boldsymbol{x}_{\hat{\mathcal{R}}(a)})$ based on the rewards obtained from post-examination or transfer rewards estimation. In other words, DE only issues recommendations when it is sufficiently confident about its clinical recommendations and it abstains otherwise.

Step 2: If there is no unresolved clinical decisions for the patient with information vector $\boldsymbol{x}_{\hat{\mathcal{R}}(a)}$, the optimal clinical decision with respect to the relevant feature set $\hat{\mathcal{R}}(a)$ is determined as

$$\hat{a}(\boldsymbol{x}) = arg \max_a \hat{y}_a^{\hat{\mathcal{R}}(a)}(\boldsymbol{x}_{\hat{\mathcal{R}}(a)}) \tag{5}$$

This optimization selects the clinical decision with the maximum estimated reward for the patient with relevant information vector $\boldsymbol{x}_{\hat{\mathcal{R}}(a)}$. The computational complexity of DE is $O(ND^2)$; hence, DE has a relatively low run-time complexity with high dimensional datasets. The pseudo-code of recommendation algorithm is given in the lower part of Algorithm 1.

C. DE with Missing Information

Electronic health records, more often than not, lack some information; hence, DE must be able to operate properly even with missing information.

Suppose that the dataset contains missing information. We can divide the feature information vector $\boldsymbol{x}$ into two components: the available features ($\boldsymbol{x}^{av}$) and the missing features ($\boldsymbol{x}^m$). Thus, $\boldsymbol{x} = \{\boldsymbol{x}^{av}, \boldsymbol{x}^m\}$. First, the relevance metric of CDFS is solely computed based on the available information:

$$h_f(a) \triangleq \sum_{x_f^{av}} \frac{N_a^f(x_f^{av})}{N_a} |\hat{y}_a^f(x_f^{av}) - \hat{y}_a|$$

Therefore, if the feature f is frequently missing, $h_f(a)$ decreases, and as a result the feature f is rarely selected as a relevant feature.

Second, it should be also noted that we can estimate the reward ($\hat{y}_a^{\hat{\mathcal{R}}(a)}(\boldsymbol{x}_{\hat{\mathcal{R}}(a)}^{av})$)with missing information based on a given patient's available relevant information, $\boldsymbol{x}_{\hat{\mathcal{R}}(a)}^{av}$, for each clinical decision a. More specifically, it can be estimated as:

$$\hat{y}_a^{\hat{\mathcal{R}}(a)}(\boldsymbol{x}_{\hat{\mathcal{R}}(a)}^{av}) = \mathrm{E}(\hat{y}_a^{\hat{\mathcal{R}}(a)}(\boldsymbol{x}_{\hat{\mathcal{R}}(a)}^{av}, \boldsymbol{x}_{\hat{\mathcal{R}}(a)}^{m}) | \boldsymbol{x}_{\hat{\mathcal{R}}(a)}^{av})$$
$$= \sum_{\boldsymbol{x}_{\hat{\mathcal{R}}(a)}^{m}} \hat{y}_a^{\hat{\mathcal{R}}(a)}(\boldsymbol{x}_{\hat{\mathcal{R}}(a)}^{av}, \boldsymbol{x}_{\hat{\mathcal{R}}(a)}^{m}) \cdot P(\boldsymbol{x}_{\hat{\mathcal{R}}(a)}^{m} | \boldsymbol{x}_{\hat{\mathcal{R}}(a)}^{av})$$
$$= \sum_{\boldsymbol{x}_{\hat{\mathcal{R}}(a)}^{m}} \hat{y}_a^{\hat{\mathcal{R}}(a)}(\boldsymbol{x}_{\hat{\mathcal{R}}(a)}) \cdot P(\boldsymbol{x}_{\hat{\mathcal{R}}(a)}^{m} | \boldsymbol{x}_{\hat{\mathcal{R}}(a)}^{av})$$

We can estimate the conditional probability, $P(\boldsymbol{x}_{\hat{\mathcal{R}}(a)}^{m} | \boldsymbol{x}_{\hat{\mathcal{R}}(a)}^{av})$, based on the probability distribution of the features in training set. Based on this estimation rule, we can robustly identify the optimal clinical decision even if there is missing information.

V. Transfer Rewards

As discussed in the introduction, the most valuable rewards for most clinical decision support systems are, in theory, the actual patient outcomes (e.g. 5 year survival rates or recurrence rates). However, these outcomes are very difficult to obtain in practice. Instead, we use a proxy for outcomes based on external knowledge which consists of published literature and clinical practice guidelines. We refer to all external knowledge simply as *references* in the remainder of the paper.

The idea is to match patients to appropriate relevant *references*. For each patient and each *reference*, we define the term *similarity* as the amount of information that *reference* provides about that patient. *Similarity* is computed by calculating the posterior probability of that patient feature belonging to the population demography from the *reference*. Then we aggregate the actual outcomes of certain clinical decision for each *reference* according to the *similarity* (posterior probability) and use this as a *transfer reward* for that clinical decision when applied to that patient. The system model for transfer reward estimation is illustrated in Fig. 2.

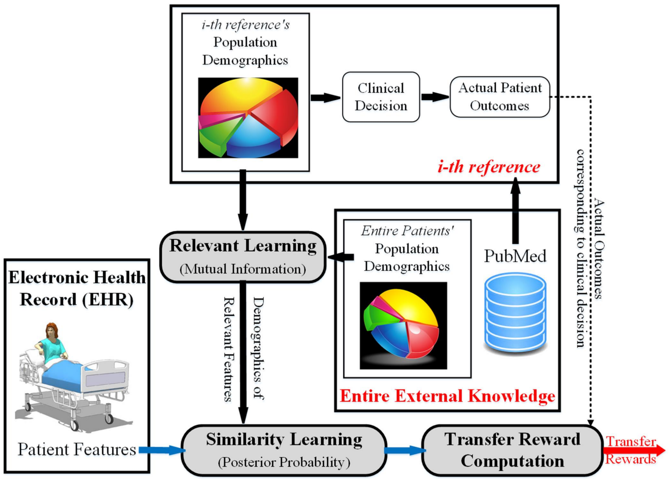


Fig. 2. System model of transfer reward estimation

To compute the transfer rewards, we first estimate the *similarity* between a patient and a *reference*. The first step of estimating this *similarity* is to find the relevant patient features for each *reference*; we do this using a sequential feature selection algorithm based on the mutual information in order to deal with population demography [35]. The mutual information between the *i-th reference* (E_i) and the *k-th* feature (f_k) is defined as:

$$\mathrm{I}(f_k; E_i) = \sum\nolimits_{x\in\chi_k} P(x|E_i)\log\frac{P(x|E_i)}{P(x)} \qquad (6)$$

where $P(x|E_i)$ is the probability of feature x in *reference i*, $P(x)$ is the probability of feature x across the entire set of *references*, and χ_k is context space of f_k. Let $\hat{\mathcal{R}}(E_i)$ define as the discovered relevant features set for *reference* E_i and the utility function $\mathcal{U}_f(E_i)$ is determined as:

$$\mathcal{U}_f(E_i) = \mathrm{I}(x_f; E_i) - \frac{1}{|\hat{\mathcal{R}}(E_i)|}\sum\nolimits_{s\in\hat{\mathcal{R}}(a)} I(x_f|E_i; x_s|E_i) \qquad (7)$$

where $1/|\hat{\mathcal{R}}(E_i)|$ is used as a normalization factor. This utility function measures an increment of mutual information between relevant feature set and the *reference* when feature f is additionally selected as relevant feature.

The algorithm selects the first relevant feature which maximizes the mutual information with E_i, i.e.,

$$\mathrm{G} = arg\max\nolimits_{f\in\hat{\mathcal{R}}^c(E_i)} I(x_f; E_i)$$
$$\hat{\mathcal{R}}(E_i) = \hat{\mathcal{R}}(E_i) \cup \mathrm{G}$$

Then, the algorithm finds the subsequent relevant feature that maximizes utility function $\mathcal{U}_f(E_i)$, i.e.,

$$\mathrm{H} = arg\max_{f\in\hat{\mathcal{R}}^c(E_i)} \mathcal{U}_f(E_i)$$
$$\hat{\mathcal{R}}(E_i) = \hat{\mathcal{R}}(E_i) \cup \mathrm{H}$$

The algorithm iteratively adds new relevant features in $\hat{\mathcal{R}}(E_i)$ until the maximum utility function $\mathcal{U}_f(a)$ becomes less than zero.

The second step for estimating the *similarity* is to compute a posterior probability that a patient feature set belonging to the population demography from the *reference*. Given the *n-th* patient, characterized by the feature vector $\boldsymbol{x_n} = \{x_1(n), \dots\dots x_D(n)\}$, we compute the posterior probability that a patient with these features belonging to the population demography from the *reference*; we express this value as $P\big(E_i\big|X_1 = x_1(n), \dots, X_D = x_D(n)\big)$. We compute this via Bayes rule; it is computationally convenient to take logarithms:

$$\log(P\big(E_i\big|X_1 = x_1(n), \dots, X_D = x_D(n)\big)$$
$$= \log\frac{(P(X_1 = x_1(n), \dots, X_D = x_D(n)|E_i)) \cdot P(E_i))}{P(X_1 = x_1(n), \dots, X_D = x_D(n))}$$
$$\cong \log P(E_i) + \sum\nolimits_{l\in\hat{\mathcal{R}}(E_i)} \log\frac{P(X_l = x_l(n)|E_i)}{P(X_l = x_l(n))}$$

where we write $P(E_i)$ as the probability of selecting the *i-th reference* as the best clinical decision for the entire population. We define this approximation of posterior probability as a *similarity* between *n-th* patient and *i-th reference* ($Sim_{E_i}(\boldsymbol{x_n})$).

Second, we compute the estimated transfer reward of each clinical decision a for *n-th* patient as a weighted sum of actual outcomes of clinical decisions in each *reference* according to the *similarity* ($Sim_{E_i}(\boldsymbol{x})$). i.e.,

$$\widehat{Sim}_{a|E_i}(\boldsymbol{x}) = \frac{Sim_{E_i}(\boldsymbol{x})}{\sum_{i:a\in A(E_i)} Sim_{E_i}(\boldsymbol{x})}$$
$$tr_a(\boldsymbol{x}) = \sum\nolimits_{i:a\in A(E_i)} \widehat{S}\ \ {}_{a|E_i}(\boldsymbol{x}) \cdot r(\,a|E_i)$$

where $r(a|E_i)$ is an actual patient outcome for clinical decision a in *i-th reference*, ($\widehat{Sim}_{a|E_i}(\boldsymbol{x})$) is a normalized *similarity* for clinical decision *a*, and $A(E_i)$ is the set of clinical decisions considered in *i-th reference.* We define this estimated reward as transfer reward ($tr_a(\boldsymbol{x})$) and these provide a complete ranking of each clinical decision for each patient. The clinical decision with the highest transfer reward is the optimal clinical decision for the given patient. The pseudo-code for estimating transfer reward is given in Algorithm 2.

Algorithm 2 Transfer Rewards Estimation

```
Initialize: $\hat{\mathcal{R}}(E_i) = \emptyset$, $\hat{\mathcal{R}}^c(E_i) = \{f_1, f_2, \dots\dots, f_D\}$ for each reference $E_i$
for each reference $E_i$
    $\mathrm{G} = arg\max_{f\in\hat{\mathcal{R}}^c(E_i)} I(x_f; E_i)$
    $\hat{\mathcal{R}}(E_i) = \hat{\mathcal{R}}(E_i) \cup \mathrm{G}$
    while ($\max_{f\in\hat{\mathcal{R}}^c(E_i)} \mathcal{U}_f(E_i) > 0$)
        $\mathrm{H} = arg\max_{f\in\hat{\mathcal{R}}^c(E_i)} \mathcal{U}_f(E_i)$
        $\hat{\mathcal{R}}(E_i) = \hat{\mathcal{R}}(E_i) \cup \mathrm{H}$
    end while
end for
for each reference $E_i$
    $Sim_{E_i}(\boldsymbol{x}) = \log P(E_i) + \sum_{l\in\hat{\mathcal{R}}(E_i)} \log\frac{P(X_l=x_l(n)|E_i)}{P(X_l=x_l(n))}$
end for
for each clinical decision $a$
    $\widehat{Sim}_{a|E_i}(\boldsymbol{x}) = \frac{Sim_{E_i}(x)}{\sum_{i:a\in A(E_i)} Sim_{E_i}(x)}$
    $tr_a(\boldsymbol{x}) = \sum_{i:a\in A(E_i)} \widehat{Sim}_{a|E_i}(\boldsymbol{x}) \cdot r(\,a|E_i)$
end for
```

VI. Experiment I: Chemotherapy Recommendation for Breast Cancer Patients

While our DE algorithm can be applicable in general clinical decision support systems, we illustrate its use in the context of a personalized recommendation system of chemotherapy regimens for breast cancer patients in this section. Fig. 3. illustrates the system model of this application.

A. Data Description

From an initial set of 2,353 *references* (performing a narrow search of chemotherapies using PubMed Clinical Queries), six chemotherapies listed in Table I, the standard chemotherapy regimens for breast cancer [7, 42-43], described in 32 references were selected for further analysis. The sample size of reported references ranged from 50 to 3,934 individuals. The individuals in each reference are mutually exclusive. A summary of the population demographics, chemotherapies and actual outcomes in *references* is provided in Appendix.

TABLE I
Code for Each Chemotherapy Regimen

Code	Specific Chemotherapy Regimen
AC	Doxorubicin+ Cyclophosphamide
ACT	Doxorubicin+ Cyclophosphamide+ Taxanes
AT	Doxorubicin+ Taxanes
CAF	Cyclophosphamide+ Doxorubicin+ 5-Fluorouracil
CEF	Cyclophosphamide+ Epirubicin+ 5-Fluorouracil
CMF	Cyclophosphamide+ Methotrexate+ 5-Fluorouracil

We evaluate our DE algorithm and benchmarks on a de-identified data set of 10,000 patient cases which underwent screening and testing at large academic medical center. The patient data is characterized by 15 features summarized in Table II and those are also corresponding with the patient features in 32 *references*.

TABLE II
Summary of Patient Information Features

Feature	Range	Features	Range
Age	30s ~ 60+	PLNC	0 ~ 10+
Menopausal	Pre/Post	Lymph Node Status	Pos/Neg
Race	White/Black/ Other	WHO Score	0 ~ 5
Estrogen Receptor	Pos/Neg	Surgery Type	BCT/MRM/ No
Progesterone Receptor	Pos/Neg	Prior Radiotherapy	Exp / No
HER2NEU	Pos/Neg/Neu	Prior Chemotherapy	Exp / No
Tumor Stage	T1 ~ T4	Histology	Ductal / Mix / Lobular
Tumor Grade	G1 ~ G3		

*PLNC: positive axillary lymph node count
*BCT: breast conservative therapy. MRM: modified radical mastectomy
*Features with categorical values are changed mutually exclusive binary indicator for the evaluation.

We iteratively evaluated the performance of the algorithms based on 10 rounds with 10 different training sets and report the average performance as a final performance of each algorithm. In each round, we used a randomly selected training set of 4,000 patients among 10,000 entire patients and a disjoint testing set of 6,000 patients. In other words, no training data were used during testing of the model, but 10 different models were used to derive the average performance. We select 4,000 patients to be in the training set, since the performance of all algorithms (besides ACL) saturated beyond this number of patients.

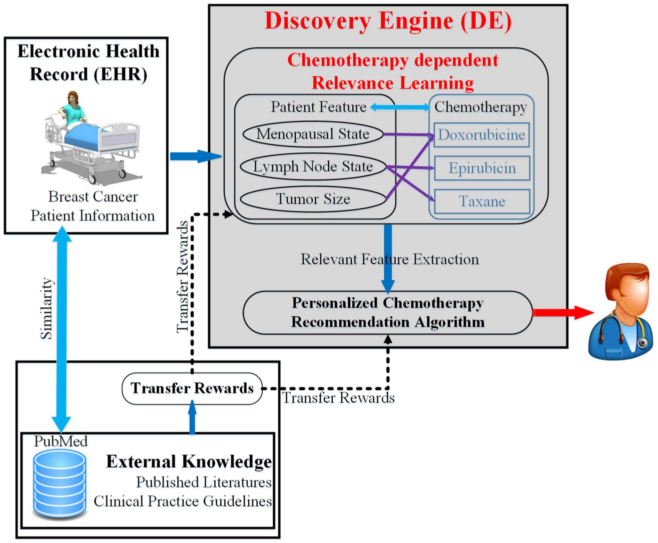


Fig. 3. Personalized chemotherapy recommendation system for breast cancer patients using DE algorithm

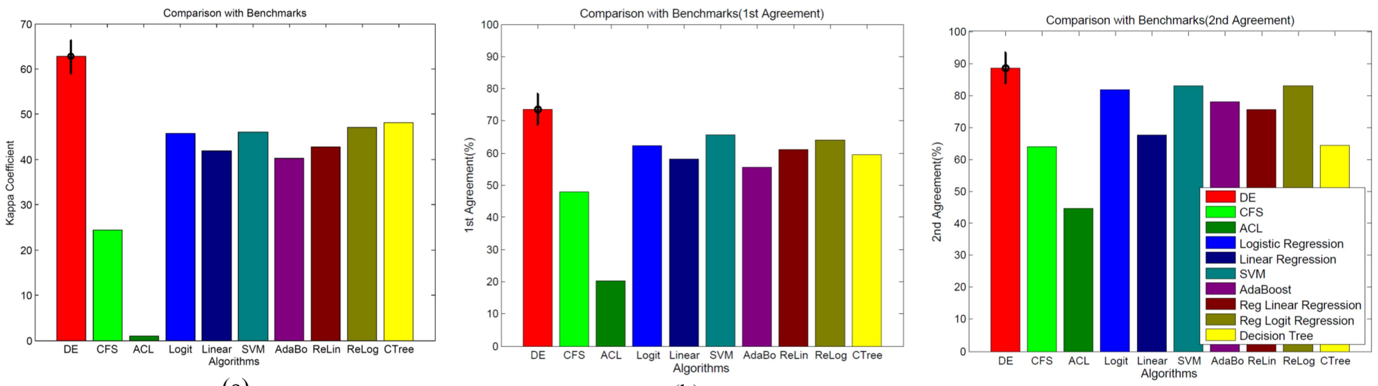
B. Benchmarks

We compare the performance of DE with nine existing state-of-the-art machine learning techniques and feature selection algorithms described below:

- Correlation Feature Selection (CFS): a well-known feature selection algorithm [34];
- All Contextual Learning (ACL): a well-known contextual learning algorithm which uses all features. This is a modified offline version of the contextual bandit algorithm of Slivkins [40];
- Multivariate Logistic Regression (Logit);
- Linear Regression (Linear);
- Multivariate Support Vector Machines (SVM); we use a radial basis function (RBF) kernel SVM;
- Adaptive Boosting (AdaBo);
- Classification Tree (CTree);
- Regularized Multivariate Logistic Regression using Lasso (ReLog);
- Regularized Linear Regression using Lasso (ReLin);

C. Success of the optimal chemotherapy recommendation for breast cancer patients

Given a patient, both our algorithm and the benchmark algorithm recommend a chemotherapy regimen corresponding to particular *references*. If the recommended chemotherapy has the highest estimated transfer reward for the patient among entire six chemotherapy regime, we regard the algorithm in question as making the correct recommendation for that patient; i.e. it has recommended the best course of treatment. (Notice that the best course of treatment may not promise a good outcome: some cancers are not treatable.) We take the



(a) (b) (c)

Fig. 4. Performance analysis with benchmark algorithms (a) Kappa coefficient, (b) 1st simple percent agreement, (c) 2nd simple percent agreement

fraction/percentage of correct recommendations to be the success rate for the algorithm in question.

Given the success rate for the algorithm, we apply two performance metrics: the simple percent agreement and the Cohen's kappa coefficient [44]. Simple percent agreement (p_0) is the success rate (the fraction of times the personalized treatment prediction coincides with the recommendation provided in the medical literature for the patient). Cohen's kappa coefficient (κ) is a metric which measures inter-rater agreement. It is usually considered a more robust measure than a simple percent agreement (p_0), because κ measures the improvement over chance agreements. If p_e is the probability of agreement by chance, then, kappa coefficient is defined as $\kappa = \frac{p_0 - p_e}{1 - p_e}$.

The bar graphs in Fig. 4. show that the first chemotherapy recommendation of DE is successful (as defined above) 73.4% of the time and one of the first two recommendations is successful 88.4% of the time. This is 7.7% better than the second best approach (SVM) in terms of selecting the optimal chemotherapy on its first choice, and 5.6% better in terms of matching the optimal chemotherapy within the first 2 choices. This is already a significant improvement, however, in terms of kappa coefficients, the improvement is even greater: DE works 16.6% better than SVM. This is because SVM indiscriminately recommends the popular chemotherapies and is not robust when classifying the less popular chemotherapies. Given robustness considerations, which are essential in medical treatment recommendations, kappa coefficients are more often used as a performance metric in medical informatics.

When comparing our algorithm with other algorithms that rely on feature selection, we again see a significant improvement. Again, note that while other algorithms use feature selection, they do not select relevant features for specific chemotherapies, and it is through this selection that our algorithm achieves improvement. CFS achieves only a 48% of simple percent agreement because it cannot use the efficacy of the chemotherapy to discriminate the relevant features and hence the technique is entirely unsupervised. ACL succumbs to the "curse of dimensionality" because there are 15 features with different ranges, resulting in over 7 million combinations to explore. Logistic regression, linear regression, and SVM perform worse than DE because they do not consider the relevant features for selecting chemotherapies at all.

D. Relevant Features for Each Chemotherapy

TABLE III
DISCOVERED RELEVANT FEATURE FOR EACH CHEMOTHERAPY

Chemotherapy Code	1st Relevant Feature	2nd Relevant Feature	3rd Relevant Feature	4th Relevant Feature
AC	PLNC	Tumor Stage	Estrogen Receptor	Age
ACT	Tumor Stage	Prior Chemotherapy	PLNC	Estrogen Receptor
AT	Prior Chemotherapy	PLNC	Surgery Type	Age
CAF	Surgery Type	Tumor Stage	Age	Tumor Grade
CEF	PLNC	Estrogen Receptor	Tumor Stage	Age
CMF	Estrogen Receptor	PLNC	Radio therapy	Tumor Stage

Table III shows the top 4 ranked relevant features discovered by CDFS - tumor stage, positive axillary lymph node number (PLNC), estrogen receptor etc.- for recommending AC, ACT, AT, CAF, CEF and CMF chemotherapy. As it can be seen from Table III, CDFS is able to discover the different relevant features that are relevant for different chemotherapy.

It is important to note the features discovered by DE are indeed confirmed to be relevant by clinical studies. Firstly, note that the six considered chemotherapies are commonly recommended to node positive breast cancer patients, i.e. patients where cancer has been found in the lymph nodes [43]. It is extremely important to know whether lymph nodes are positive or negative. PLNC tells us both the number of nodes and whether lymph nodes are positive or negative. For instance, zero PLNC implies node negative breast cancer, while otherwise indicates node positive breast cancer. Hence, PLNC is selected as a relevant feature by CDFS. Secondly, the menopausal status is considered important because medications affect cancer differently in premenopausal and postmenopausal women [9]. More specifically, the CEF chemotherapy is only recommended to premenopausal women. Although the menopausal status is not included in this relevant feature set, women over the age of 50 are usually considered postmenopausal [9]. Therefore, age was correctly identified by DE to be a discriminative feature for selecting among chemotherapy regimens. Thirdly, tumor stage is another important feature to consider when deciding among chemotherapy regimens as described in reference [42]. Medications A(Doxorubicin), T(Taxotere), E(Epirubicin) are recommended for advanced breast cancer and our top six chemotherapy regimens include more than one of these medications. Therefore, DE has correctly discovered that the features that are relevant for these chemotherapy regimens contain tumor stage information. Finally, the medication T(Taxotere) is usually recommended to breast cancer patients who do not respond to their current chemotherapy. Thus, the prior chemotherapy information is correctly discovered by DE to be relevant for AT and ACT therapies.

E. *Performance when Patient Information is Missing*

As explained before, patient information is often missing from the EHR. Moreover, studies have shown that the missing information is often not random [45]. For example, the age of the patient is easy to record and blood pressure is often verified several times by a nurse when a patient is seen by a clinician, so it is typically neither missing nor incorrect. However, HER2NEU may not be recorded depending on the diagnostic tests ordered and capabilities of a medical center. Fig. 5. is achieved by evaluating the performance of DE when features are missing with various rates. It shows the performance degradation of DE and of the benchmark algorithms as a function of the average *degree of incompleteness* [45]. (We did not use the percentage of missing features as a metric since the features are not randomly missing. The percentages of missing features corresponding to each *degree of incompleteness* are described in Appendix; these percentages were computed based on statistics extracted from medical records of patients.) Fig. 5. shows that the performance of DE degrades from 73.4% to 63.0% (when the average *degree of incompleteness* is 50%). However, even with missing information, DE continues to outperform the other methods. DE discovers relevant features with low missing probability, and is able to estimate the missing feature information based on the available feature information. As a result, the impact of missing information is minimized. In fact, DE performs better than most other algorithms even when DE misses significant amounts of information from the EHR while the other algorithms make their decision with full information. Hence we can indeed see that the performance of DE is robust even when information is missing.

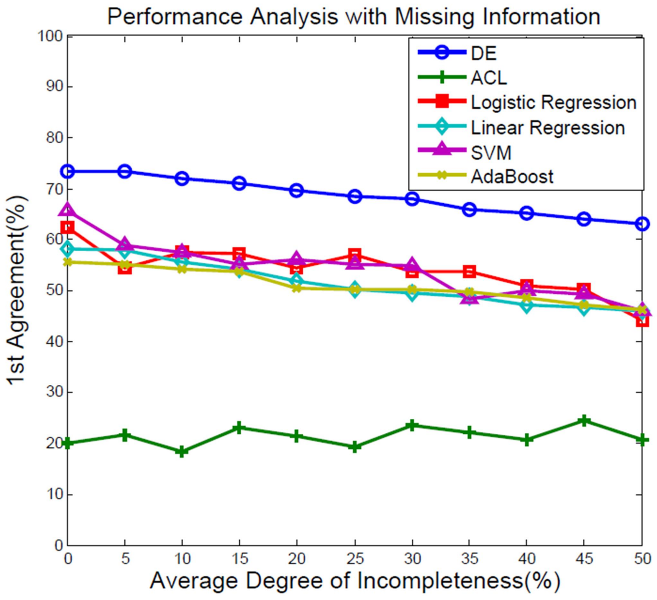


Fig. 5. Performance analysis with missing information

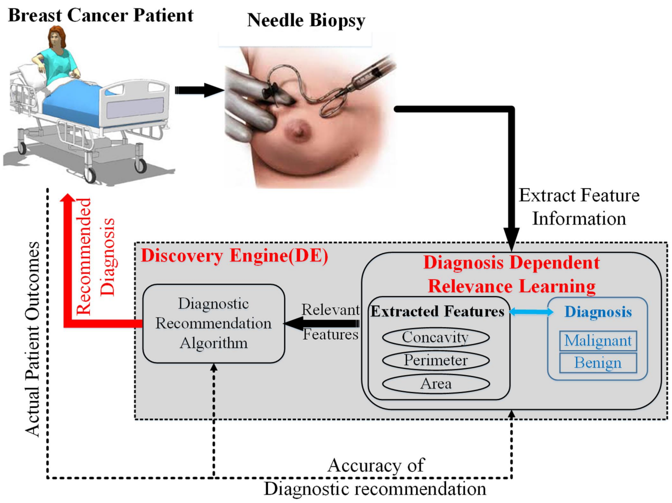


Fig. 6. Personalized diagnostic recommendation system for breast cancer patients using DE algorithm

VII. Experiment II: Diagnosis Decision Support System for Breast Cancer Patients

In this section, we illustrate how the DE algorithm can be used for breast cancer diagnosis. In this case, we can directly use patients' actual outcomes as the rewards. Fig. 6. describes the system model of this application.

A. *Data Description*

TABLE IV
Summary of The Feature Information

	Information Type	Explanation
1	Radius	Mean of distance from center to points on the perimeter
2	Texture	Standard deviation of gray-scale values
3	Perimeter	The perimeter of tumor cell nucleus
4	Area	The area of tumor cell nucleus
5	Smoothness	Local variation in radius lengths
6	Compactness	Perimeter^2 / area -1
7	Concavity	Severity of concave portions of the contour
8	Concave Points	Number of concave portions of the contour
9	Symmetry	Symmetricity of tumor cell nucleus
10	Fractal Dimension	Coastline approximation – 1

*Features consist of mean, standard errors and worst of above 10 info types.
*Each info type is computed real value feature for each tumor cell nucleus.

In this section we evaluate the performance of DE for breast cancer diagnosis using the well-known UCI dataset [46]. The dataset contains 30 patient features extracted from needle biopsy features such as radius, compactness, or smoothness of tumor cell nucleus. Table IV summarizes the details of 30 patient features. The number of instances in this dataset is 569 and the diagnosis (label) for each instance is either malignant or benign.

B. Benchmarks

We compare the performance of DE algorithms with seven existing machine learning algorithms and four existing feature selection algorithms:

- Logistic Regression (Logit);
- Linear Regression (Linear);
- Support Vector Machines (SVM); we use a radial basis function (RBF) kernel SVM;
- Adaptive Boosting (AdaBo);
- Classification Tree (CTree);
- Regularized Logistic Regression using Lasso (ReLog);
- Regularized Linear Regression using Lasso (ReLin);
- Correlation Feature Selection (CFS): a well-known feature selection algorithm based on correlation [34];
- Mutual Information Feature Selection (MIFS): a well-known feature selection algorithm based on mutual information [35];
- Relevance Learning with Feedback (RELEAF): an action dependent relevance learning algorithm based on the expected rewards [36-37];
- Principal Component Analysis (PCA): a statistical procedure to discover linearly uncorrelated variables based on orthogonal transformation [47]

C. Experiments Setup

First, we compare our DE algorithm against state-of-the-art machine learning algorithms: logistic/linear regressions, rbf kernel SVM, adaptive boosting algorithms, classification tree and regularized logistic/linear regressions. We use 10-fold cross-validation in order to evaluate the performance of algorithms. We performed 10 independent cross validation runs and report the average performance of 10 runs.

To highlight the impacts of CDFS, we performed two additional sets of experiments. In the first set, we compared the performance of our DE system using CDFS with the performance of the DE system where CDFS was replaced with one of the four different feature selection algorithms: CFS, MIFS, RELEAF, and PCA. This comparison shows the impact of CDFS on the overall performance of the DE. Other experiment settings are exactly the same as the first experiment.

In the second set of additional experiments, we use our feature selection algorithm CDFS in conjunction with the diagnostic recommendation made by the benchmark algorithms - linear regression, logistic regression, SVM - to highlight the specific impact of our feature selection algorithm. Other experiment settings are exactly the same as the first experiment.

D. Performance of diagnostic recommendation for breast cancer patients

Given a patient, DE and the other benchmark algorithms classify the tumors as malignant or benign. To quantify their performance, we apply three different performance metrics: the prediction error rate (*PER*), the false positive rate (*FPR*), and the false negative rate (*FNR*). *PER* is defined as the fraction of times the recommended diagnosis of our algorithm is different from the actual label. *FPR* and *FNR* are defined as the diagnosis error rate for benign instances and for malignant instances, respectively. The main goal of DDSS is to minimize the FPR given an allowable threshold for FNR as selected by the clinicians. In practice, this is often set to be below 2% [25]. Therefore, in this experiment, the FNR threshold is also set to be 2%. Using this threshold, we can characterize our performance metrics as follows.

$$\begin{aligned} \text{minimize} \quad & FPR \\ \text{subject to} \quad & FNR \leq 2\% \end{aligned}$$

Comparison with machine learning algorithms:

TABLE V
COMPARISON WITH TYPICAL MACHINE LEARNING ALGORITHMS

%	PER	FPR	FNR
DE	2.23	2.62	1.92
Logit	11.77	18.3	1.96
Linear	8.47	13.55	1.98
SVMs	4.41	6.82	1.99
AdaBo	9.12	14.86	1.91
CTree	11.45	18.64	1.95
ReLog	6.71	10.11	1.92
ReLin	5.51	9.15	1.93

As the Table V shows, our DE algorithm has 2.23% prediction error rates and 2.62% false positive rates which is 2.18% and 4.24% better than the second best algorithm (SVM) when the tolerable threshold of FNR is set to below 2%. There are two reasons for the outstanding performance of the DE approach. First, our diagnostic recommendation algorithm yields high accuracy for classification, because it is able to provide personalized diagnosis, while other comparable algorithms apply the same model for all patients. Second, DE can discover different relevant features for different diagnosis based on CDFS, while the other algorithms (Logistic/Linear Regression, SVMs, AdaBoost and Classification Tree) base their decisions on all the features without relevant feature discovery.

Comparison with feature selection algorithms:

In this subsection, we demonstrate the impact of the CDFS algorithm on the DE system. We compare the performance of the DE using CDFS with the performance of DE using different feature selection algorithms.

TABLE VI
PERFORMANCE OF DE WITH OTHER FEATURE SELECTION METHODS

%	CDFS	RELEAF	CFS	MIFS	PCA
PER	2.23	18.37	2.76	3.19	3.94
FPR	2.62	24.11	3.90	3.99	6.44
FNR	1.92	1.96	1.98	1.90	1.94

As seen in Table VI, CDFS outperforms all other feature selection algorithms when the tolerable threshold of FNR is set to below 2%. This is because CDFS is capable of discovering diagnosis relevant features based on their impact on the expected diagnosis accuracies. Although RELEAF also considers the dependence between diagnosis accuracy and feature selection, it is extremely slow when the number of features is large, as is the case for this and many other medical datasets. Furthermore, combinatorial approach (RELEAF) requires a relatively large amount of training sets to accurately discover the relevant feature, which is not the case for the medical dataset available to us.

TABLE VII
IMPACT OF THE CDFS IN CONJUNCTION WITH ALTERNATIVE MACHINE LEARNING ALGORITHMS

	PER (%)		**FPR (%)**	
	CDFS	**w/o CDFS**	**CDFS**	**w/o CDFS**
Linear	5.49	8.47	8.88	13.55
Logit	7.84	11.77	12.7	18.3
SVMs	4.01	4.41	5.51	6.82

Next, we replace DE's recommendation algorithms with various existing machine learning algorithms in order to demonstrate the impact of the CDFS component of DE on the diagnostic decisions. As seen in Table VII, CDFS improves the performance of all benchmark algorithms because it is able to accurately discover and select the (different) features that are relevant for different diagnosis.

E. Relevant Features for Diagnostic Decision

TABLE VIII
DISCOVERED RELEVANT FEATURES FOR EACH DIAGNOSIS

	Malignant	**Benign**
1st Relevant Feature	Worst perimeter	Worst concave points
2nd Relevant Feature	Worst concave points	Mean concave points
3rd Relevant Feature	Worst radius	Worst perimeter
4th Relevant Feature	Mean concavity	Worst radius
5th Relevant Feature	Worst area	Mean concavity

Table VIII shows the top 5 ranked relevant features discovered by CDFS – worst perimeter, concave points concavity, radius and area etc.- for diagnosing malignant or benign cancer among all of the features summarized in Table IV. As seen in Table V, CDFS is able to discover the different features that are relevant for different diagnosis.

It should be noted that the relevance of the features discovered by DE is confirmed by clinical studies. For instance, studies of breast biopsies [48] state that the 3 most important factors to diagnose tumor cell nuclei as malignant or benign are the relative size ratio between nucleus and cytoplasm, irregular shape, and irregular chromatin. However, because chromatin feature information is not available in our dataset, only the relative size and irregular shape can be potential candidates as the relevant features. The size related features are radius, perimeter, and area, and the shape related features are the concavity and concave points [48]. The top 5 features found to be relevant by DE to classify malignant and benign tumor cell nuclei are all related to the tumor shape and relative size, which is in accordance to reference [48]. Features such as texture, smoothness, compactness, symmetry, and fractal dimension are not found to be relevant by DE and are not mentioned as important features in reference [48]. Hence, we can conclude that DE can discover the relevant features for making a correct diagnosis without prior medical knowledge.

VIII. DISCUSSION AND FUTURE WORKS

We proposed a novel approach that discovers the relevant information from the EHR that is important in determining which clinical decision to recommend for a patient and used this information to provide personalized recommendation to assist physicians in their decision making process. Our results demonstrate that DE outperforms existing machine learning, prediction, and feature selection methods in both CDSS and DDSS applications. *This superior recommendations are extremely important because they have the potential to prevent medical errors and thus improve the quality of medical care.* We also showed that our method is robust against missing information, which is important in numerous clinical settings.

Future work will consider that feature information such as the tumor size, PLNC number and tumor radius may change over time. The time dependence may influence the duration of a therapy and the selection of future therapies. Therefore, DE approach with time analysis can also be an important extension. Currently, we only consider a single decision, identifying what information is important in deciding on that single decision. Another DE extension will consider the global sequence of treatment decisions that optimize long-term outcomes (e.g. overall survival rate or 5-year recurrence rate).

In conclusion, we believe that our proposed contextual learning approach demonstrates promise towards providing useful personalized clinical recommendations. As new types of treatment are evaluated and approved for use, clinicians will have an increasingly difficult time determining which clinical decisions are most effective for individual patients. DE provides a pathway towards providing computational methods for personalized clinical decision recommendations.

APPENDIX

APPENDIX TABLE I
NOMENCLATURE TABLE

Notation	Interpretation	Notation	Interpretation
n	Current patient number	N	Total patient number
$\boldsymbol{x}$	Patient information vector	D	Total number of features
a	Clinical decision	f	Patient feature
y	Rewards	$\boldsymbol{\mathcal{H}}_N$	Available information for *N* previous patients
$\mathcal{R}(a)$	Relevant feature set for clinical decision *a*	$\mathcal{R}$	Relevant feature set for all *a*
$a^*(\boldsymbol{x})$	Optimal recommended clinical decision	$\hat{y}_a^S(\boldsymbol{x}_s)$	Sample mean reward estimator for patients contains $\boldsymbol{x}_s$
$N_a^S(\boldsymbol{x}_s)$	Number of patients contains $\boldsymbol{x}_s$	$\hat{y}_a$	Sample mean reward estimator of clinical decision *a*
N_a	Number of patients with clinical decision *a*	h_f^r	Relevance metric
$h_{f,s}^d$	Redundancy metric	U_f	Utility function
C	Threshold cost	C_{TH}	Coefficient of control function
x_f^{av}	Available features	x_f^m	Missing features
E_i	*i-th reference*	$P(x\|E_i)$	Probability of feature *x* in *i-th reference*
$Sim_{E_i}(\boldsymbol{x})$	*Similarity* between E_i and $\boldsymbol{x}$	$\widehat{Sim}_{a\|E_i}(\boldsymbol{x})$	Normalized *similarity* for clinical decision *a*
$tr_a(\boldsymbol{x})$	Transfer reward of *a* given $\boldsymbol{x}$		

APPENDIX TABLE II
SUMMARY OF THE INFORMATION IN REFERENCES

Medical Paper No		Paper 1	Paper 2	…	Paper 32	Entire Papers
Population Demography	Age	**30s:** 21% **40s:** 32% **50s:** 27% **60s:** 20%	**30s:** 15% **40s:** 35% **50s:** 30% **60s:** 20%	…	**30s:** 16% **40s:** 47% **50s:** 22% **60s:** 15%	**30s:** 16% **40s:** 40% **50s:** 27% **60s:** 16%
	Estrogen Receptor	**Pos:** 61% **Neg:** 22% **Miss:**17%	**Missing**	…	**Pos:** 26% **Neg:** 74%	**Pos:** 63% **Neg:** 33% **Miss:** 4%
	Tumor Grade	**Missing**	**Missing**	…	**G1:** 7% **G2:** 47% **G3:** 40% **Miss:** 6%	**G1:** 15% **G2:** 41% **G3:** 38% **Miss:** 6%
	…					
	Tumor Size	**Missing**	**T1:** 19% **T2:** 58% **T3:** 11% **T4:** 11% **Miss:** 1%	…	**T1:** 41% **T2:** 51% **T3:** 4% **T4:** 4%	**T1:** 37% **T2:** 40% **T3:** 12% **T4:** 10% **Miss:** 1%
	Lymph Node Status	**Pos:** 99% **Neg:** 0% **Miss:** 1%	**Missing**	…	**Pos:** 66% **Neg:** 36%	**Pos:** 34% **Neg:** 66%
Chemotherapy		AC CMF	AT CAF	…	CEF CMF	
Actual Outcomes (5 year survival rates)		AC:95% CMF:92%	AT: 72% CAF:78%	…	CEF:82% CMF:81%	

*We can use various actual outcomes from references. (e.g. 5 year recurrence free survival rates)

APPENDIX TABLE III
DEGREE OF INCOMPLETENESS IN SECTION VI

		Corresponding *degree of incompleteness* per feature (%)				
Average *Degree of Incompleteness*		**10%**	**20%**	**30%**	**40%**	**50%**
Feature	**Age**	0.43	0.87	1.3	1.73	2.16
	Menopausal	10.53	21.06	31.59	42.12	52.65
	Race	9.12	18.23	27.35	36.47	45.58
	Estrogen Receptor	5.9	11.81	17.71	23.61	29.51
	Progesterone Receptor	8.63	17.26	25.89	34.52	43.15
	HER2NUE	13.72	27.44	41.16	54.87	68.59
	Tumor Stage	5.25	10.5	15.75	20.99	26.24
	Tumor Grade	12.5	25.01	37.51	50.02	62.52
	PLNC	8.29	16.57	24.86	33.14	41.43
	Lymph Node Status	10.17	20.35	30.52	40.7	50.87
	WHO Score	15.38	30.76	46.14	61.52	76.9
	Surgery Type	6.62	13.24	19.86	26.48	33.1
	Prior Radiotherapy	13.57	27.14	40.71	54.28	67.84
	Prior Chemotherapy	14.21	28.43	42.64	56.86	71.07
	Histology	15.67	31.35	47.02	62.7	78.37

REFERENCES

[1] T. D. Richmond, “The current status and future potential of personalized diagnostics: streamlining a customized process,” *Bio-technology annual review*, vol. 14, pp. 411-422, 2008.

[2] American Cancer Society Guidelines for the Early Detection of Cancer, http://www.cancer.org/healthy/findcancerearly/cancerscreeningguidelines/american-cancer-society-guidelines-for-the-early-detection-of-cancer.

[3] M. Frellick, “Landmark Report Urges Reform to Avert Diagnostic Errors,” *Medscape*, 2015.

[4] E.S. Berner, “Clinical Decision Support Systems,” *Springer* , New York, 2007.

[5] M. Akay, D. Fotiadis, K. S. Nikita and R. W. Williams, “Guest Editorial: Biomedical Informatics in Clinical Environments,” *Biomedical and Health Informatics, IEEE Journal of*, vol. 19, no. 1, pp. 149-150, 2015.

[6] E. H. Shortliffe, and J. J Cimino, “Biomedical informatics: computer applications in health care and biomedicine,” *Springer Science and Business Media*, 2013.

[7] A. L. Blum and P. Langley, “Selection of relevant features and examples in machine learning,” *Artificial intelligence*, vol. 97, no. 1, pp. 245-271, 1997.

[8] C. C. Bennett, and K. Hauser, “Artificial intelligence framework for simulating clinical decision-making: A Markov decision process approach,” *Artificial intelligence in medicine*, vol. 57, no. 1, pp. 9-19, 2013.

[9] L, Mark, “Clinical practice guidelines for the care and treatment of breast cancer: adjuvant systemic therapy for node-positive breast cancer (summary of the 2001 update),” *Canadian Medical Association Journal*, pp. 644-646, 2001.

[10] S. Tsumoto, “Automated extraction of medical expert system rules from clinical databases based on rough set theory,” *Information sciences*, vol. 112, no. 1, pp. 67-84, 1998.

[11] J. Xu, D. Sow, D. Turaga and M. van der Schaar, “Online Transfer Learning for Differential Diagnosis Determination,” *AAAI Workshop on the World Wide Web and Public Health Intelligence*, 2014.

[12] M. Viceconti, P. Hunter, and D. Hose, "Big data, big knowledge: big data for personalised healthcare," *IEEE J. Biomedical and Health Informatics,* 2015.

[13] S. Ram, W. Zhang, M. Williams and Y. Pengetnze, “Predicting Asthma-Related Emergency Department Visits Using Big Data,” *IEEE J. Biomedical and Health Informatics,* 2015.

[14] Douali, Nassim, and M. Jaulent. "Genomic and personalized medicine decision support system." *Complex Systems (ICCS), 2012 International Conference on. IEEE*, 2012.

[15] I. Kouris, C. Tsirmpas, S. G. Mougiakakou, D. Iliopoulou and D. Koutsouris, “E-Health towards ecumenical framework for personalized medicine via Decision Support System,” *In Engineering in Medicine and Biology Society (EMBC), 2010 Annual International Conference of the IEEE,* pp. 2881-2885, 2010.

[16] K. Donsa, S. Spat, P. Beck, T. R. Pieber and A. Holzinger, “Towards personalization of diabetes therapy using computerized decision support and machine learning: some open problems and challenges,” *In Smart Health Springer International Publishing*, pp. 237-260, 2015.

[17] C. Bennett, T. Doub, A. Bragg, J. Luellen, C. Van Regenmorter, J. Lockman and R. Reiserer, “Data mining session-based patient reported outcomes (PROs) in a mental health setting: toward data-driven clinical decision support and personalized treatment,” *In Healthcare Informatics, Imaging and Systems Biology (HISB), 2011 First IEEE International Conference on*, pp. 229-236, 2011.

[18] H. Shin and M. K. Markey, “A machine learning perspective on the development of clinical decision support systems utilizing mass spectra of blood samples,” *Journal of Biomedical Informatics*, vol. 39, no. 2, pp. 227-248, 2006.

[19] M. McNamara, K. Sarma, D. R. Aberle, A. A. Bui and C. Arnold, “Data Model for Personalized Patient Health Guidelines: An Exploratory Study,” *American Medical Informatics Association*, vol. 2014, pp. 1835, 2014.

[20] A. B. McCoy, L. R. Waitman, J. B. Lewis, J. A. Wright, D. P. Choma, R. A. Miller and J. F. Peterson, “A framework for evaluating the appropriateness of clinical decision support alerts and responses,” *Journal of the American Medical Informatics Association*, vol. 19, no. 3, pp. 346-352, 2012.

[21] J. H. Eom, S. C. Kim and B. T. Zhang, “AptaCDSS-E: A classifier ensemble-based clinical decision support system for cardiovascular disease level prediction,” *Expert Systems with Applications*, vol. 34, no. 4, pp. 2465-2479, 2008.

[22] E. Çomak, A. Arslan and İ. Türkoğlu, “A decision support system based on support vector machines for diagnosis of the heart valve diseases,” *Computers in Biology and Medicine*, vol. 37, no. 1, pp. 21-27, 2007.

[23] K. Zarkogianni and K. S. Nikita, “Personal health systems for diabetes management, early diagnosis and prevention,” *Handbook of Research on Trends in the Diagnosis and Treatment of Chronic Conditions*, 2015.

[24] P. K. Anooj, “Clinical decision support system: Risk level prediction of heart disease using weighted fuzzy rules,” *Journal of King Saud University-Computer and Information Sciences,* vol. 24, no. 1, pp. 27-40, 2012.

[25] K. Polat, and S. Güneş, “Breast cancer diagnosis using least square support vector machine,” *Digital Signal Processing*, vol. 17, no. 4, pp. 694-701, 2007

[26] L. Song, W. Hsu, J. Xu and M. van der Schaar, “Using Contextual Learning to Improve Diagnostic Accuracy: Application in Breast Cancer Screening*,” IEEE J. Biomedical and Health Informatics*, 2015

[27] P. Hall, J. S. Marron and A. Neeman, “Geometric representation of high dimension, low sample size data*,” Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 67, no. 3, pp. 427-444, 2005.

[28] E. C. Karvounis, K. Stefanou, T. P. Exarchos, A. T. Tzallas, M. Tsipouras and D. Fotiadis, “A treatment decision support system for patients receiving Ventricular Assist Device (VAD) therapy,” *In Biomedical and Health Informatics (BHI), 2012 IEEE-EMBS International Conference on,* pp. 695-698, 2012.

[29] R. Tsopra, A. Venot and C. Duclos, “An Algorithm Using Twelve Properties of Antibiotics to Find the Recommended Antibiotics, as in CPGs,” *American Medical Informatics Association*, pp. 1115, 2014.\

[30] P. Zhang, F. Wang, J. Hu and R. Sorrentino, “Towards personalized medicine: Leveraging patient similarity and drug similarity analytics,” *AMIA Summits on Translational Science Proceedings*, 2014.

[31] Y. Wu, J. Liu, D. Page, P. Peissig, C. McCarty, A. Onitilo and E. S. Burnside, “Comparing the value of mammographic features and genetic variants in breast cancer risk prediction,”. *American Medical Informatics Association*, vol. 2014, pp. 1228, 2014.

[32] F. Kuusisto, I. Dutra, M. Elezaby, E. A. Mendonça, J. Shavlik, and E. S. Burnside, “Leveraging Expert Knowledge to Improve Machine-Learned Decision Support Systems,” *AMIA Summits on Translational Science Proceedings*, vol. 87, 2015.

[33] J. Liu, D. Page, P. Peissig, C. McCarty, A. A. Onitilo, A. Trentham-Dietz and E. S. Burnside, “New genetic variants improve personalized breast cancer diagnosis,” *AMIA Summits on Translational Science Proceedings*, vol. 83, 2014.

[34] M. A. Hall, “Correlation-based feature selection for machine learning,” *Doctoral dissertation*, The University of Waikato, 1999

[35] H. Peng, F. Long and C. Ding, “Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy,” *Pattern Analysis and Machine Intelligence, IEEE Transactions,* vol. 27, no. 8, pp. 1226-1238, 2005.

[36] C. Tekin, and M. van der Schaar, “Discovering, learning and exploiting relevance,” *In Advances in Neural Information Processing Systems*, pp. 1233-1241, 2014.

[37] C. Tekin, and M. van der Schaar, “RELEAF: An algorithm for learning and exploiting relevance,” *IEEE Journal of Selected Topics in Signal Processing, Special Issue on Signal Processing for Big Data,* vol. 9, no. 4, pp. 716-727, 2015.

[38] T. A. Peterson, E. Doughty and M. G. Kann, “Towards precision medicine: advances in computational approaches for the analysis of human variants,” *Journal of molecular biology*, vol. 425, no. 11, pp. 4047-4063, 2013.

[39] L. J. Goldstein, A. O'Neill, J. A. Sparano, E. A. Perez, L. N. Shulman, S. Martino and N. E. Davidson, “Concurrent doxorubicin plus docetaxel is not more effective than concurrent doxorubicin plus cyclophosphamide in operable breast cancer with 0 to 3 positive axillary nodes: North American Breast Cancer Intergroup Trial E 2197,” *Journal of Clinical Oncology,* vol. 26, no. 25, pp. 4092-4099, 2008.

[40] A. Slivkins, “Contextual bandits with similarity information,” *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 2533-2568, 2014.

[41] A. Badanidiyuru, J. Langford, and A. Slivkins, “Resourceful contextual bandits,” *The 27th Conference on Learning Theory*, pp. 1109-1134, 2014.

[42] F. M. Fleegler, J. Griggs, B. Reiner, B. Reville, S. F. Schnall, M. C. Weiss and L. Weissmann, “Chemotherapy Medicine,” Retrieved from http://www.breastcancer.org/treatment/chemotherapy/medicines, 2015.

[43] P. J. Hamel and P. Johnson, “Chemotherapy Regimen for Breast Cancer,” Retrieved from http://www.healthcentral.com/breast-cancer/chemo-regimen.html, 2015.

[44] Cohen, J. A, “Coefficient of agreement for nominal scales,” *Educational and psychological measurement*, vol. 20, no. 1, pp. 37-46, 1960.

[45] T. Botsis, G. Hartvigsen, F. Chen and C. Wen, “Secondary use of EHR: data quality issues and informatics opportunities,” *AMIA summits on translational science proceedings*, 2010.

[46] K. Bache and M. Lichman, “UCI machine learning repository,” 2013. [Online]. Available: http://archive.ics.uci.edu/ml

[47] Pearson, K, "On Lines and Planes of Closest Fit to Systems of Points in Space," *Philosophical Magazine* vol. 2 no. 11, pp. 559–572, 1901.

[48] S. R. Orell, G. F. Sterrett, and D. Whitaker, “Fine needle aspiration cytology,” *Elsevier Churchill Livingstone*, pp. 49-55, 2005.

[49] C. Zhou, Y. L. Wu, G. Chen, J. Feng, X. Q. Liu, C. Wang and C. You, “Erlotinib versus chemotherapy as first-line treatment for patients with advanced EGFR mutation-positive non-small-cell lung cancer (OPTIMAL, CTONG-0802): a multicentre, open-label, randomised, phase 3 study*,” The lancet oncology*, vol. 12, no. 8, pp. 735-742, 2011.

[50] C. I. Flowers, C. O’Donoghue, D. Moore, A. Goss, D. Kim, and J.H. Kim, “Reducing false-positive biopsies: a pilot study to reduce benign biopsy rates for BI-RADS 4A/B assessments through testing risk stratification and new thresholds for intervention,” *Breast Cancer Research and Treatment*, vol. 139, no. 3, pp. 769-777, 2013.

Jinsung Yoon received the B.S. degree in electrical engineering from Seoul National University, Seoul, Korea. He is currently pursuing his PhD degree in electrical engineering at University of California, Los Angeles. His research interests include machine learning algorithms, medical informatics, data mining, recommendation systems, and prediction systems. He was a recipient of the Graduate Study Abroad Scholarship from the National Research Foundation of Korea in 2014.

Mihaela van der Schaar is Chancellor Professor of Electrical Engineering at University of California, Los Angeles. Her research interests include network economics and game theory, online learning, dynamic multi-user networking and communication, multimedia processing and systems, real-time stream mining. She is an IEEE Fellow, a Distinguished Lecturer of the Communications Society for 2011-2012, the Editor in Chief of IEEE Transactions on Multimedia and a member of the Editorial Board of the IEEE Journal on Selected Topics in Signal Processing. She received an NSF CAREER Award (2004), the Best Paper Award from IEEE Transactions on Circuits and Systems for Video Technology (2005), the Okawa Foundation Award (2006), the IBM Faculty Award (2005, 2007, 2008), the Most Cited Paper Award from EURASIP: Image Communications Journal (2006), the Gamenets Conference Best Paper Award (2011) and the 2011 IEEE Circuits and Systems Society Darlington Award best Paper Award. She received three ISO awards for her contributions to the MPEG video compression and streaming international standardization activities and holds 33 granted US patents.

Camelia Davtyan is a Clinical Professor of Medicine at the Geffen School of Medicine at UCLA and Director of Women’s Health at the UCLA Comprehensive Health Program.