ConfidentCare: A Clinical Decision Support System for Personalized Breast Cancer Screening

Ahmed M. Alaa, *Member, IEEE*, Kyeong H. Moon, William Hsu, *Member, IEEE*, and Mihaela van der Schaar, *Fellow, IEEE*

***Abstract*—Breast cancer screening policies attempt to achieve timely diagnosis by regularly screening healthy women. Various clinical decisions are needed to manage the screening process: selecting initial screening tests, interpreting test results, and deciding if further diagnostic tests are required. Such decisions are currently guided by clinical practice guidelines (CPGs), which represent a "one-size-fits-all" approach, designed to work well (on average) for a population. Since the risks and benefits of screening tests are functions of each patient's features, *personalized screening policies* tailored to the features of individuals are desirable. To address this issue, we developed *ConfidentCare*: a computer-aided clinical decision support system that learns a personalized screening policy from electronic health record (EHR) data. By "personalized screening policy", we mean a clustering of women's features, and a set of customized screening guidelines for each cluster. ConfidentCare operates by computing clusters of patients with similar features, then learning the "best" screening procedure for each cluster using a supervised learning algorithm. ConfidentCare utilizes an iterative algorithm that applies risk-based clustering of the women's feature space, followed by learning an active classifier for every cluster. The algorithm ensures that the learned screening policy satisfies a predefined accuracy requirement with a high level of confidence for every cluster. By applying ConfidentCare to real-world data, we show that it outperforms the current CPGs in terms of cost-efficiency and false positive rates.**

***Index Terms*—Breast cancer, Confidence measures, Clinical decision support, Personalized medicine, Supervised learning.**

I. Introduction

PERSONALIZED medicine is a healthcare paradigm that aims to move beyond the current "one-size-fits-all" approach to medicine that does not take into account the features and traits of individual patients (e.g. their micro-biomes, environments, and lifestyles) [1]-[3]. Vast attention has been dedicated to research in personalized medicine that builds on data science and machine learning techniques to customize healthcare policies. For instance, the White House has led the "precision medicine initiative" [4], which is scheduled for discussion in the American Association for the Advancement of Science annual meeting for the year 2016 [5]. Breast cancer screening is an important healthcare process that can benefit from personalization. Screening is carried out in order to diagnose a woman with no apparent symptoms in a timely manner [6]-[10]. However, the screening process entails both benefits and costs that can differ from one patient to another [11]. This signals the need for personalized screening policies that balance such benefits and costs in a customized manner.

In this paper we present ConfidentCare: a clinical decision support system (CDSS) that is capable of learning and implementing a personalized screening policy for breast cancer. The personalized screening policy is learned from data in the electronic health record (EHR), and is aimed to issue recommendations for different women with different features on which sequence of screening tests they should take. ConfidentCare discovers subgroups of "similar" patients from the EHR data, and learns how to construct a screening policy that will work well for each subgroup with a high level of confidence. Our approach can provide significant gains in terms of both the cost-efficiency, and the accuracy of the screening process as compared to other "one-size-fits-all" approaches suggested by current clinical practice guidelines (CPGs) that apply the same policy on all patients.

A. Breast cancer screening and the need for personalization

While breast cancer screening is believed to reduce mortality rates [10], it is associated with the risks of "overscreening", which leads to unnecessary costs, and "overdiagnosis", which corresponds to false positive diagnoses that lead the patients to receive unnecessary treatments [11]. While different patients have different levels of risks for developing breast cancer [12]-[16], different tests have different monetary costs, and different levels of accuracy that depend on the features of the patient [17], common CPGs are aimed at populations, and are not typically tailored to specific individuals or significant subgroups [18]-[21].

Being designed to work well on "average" for a population of patients, following CPGs may lead to overscreening or overdiagnosis for specific subgroups of patients, such as young women at a high risk of developing breast cancer, or healthy older women who may have a relatively longer expected lifespan [22]. Moreover, some screening tests may work well for some patients, but not for others (e.g. a mammogram test will exhibit low accuracy for patients with high breast density [17]), which can either lead to "underdiagnosis" or poor tumor detection performance. Migrating from the "one-size-fits-all" screening and diagnosis policies adopted by CPGs to more individualized policies that recognize and approach subgroups of patients with different features is the essence of applying the personalized medicine paradigm to the breast cancer screening process [17], [22]-[25].

A. M. Alaa, K. H. Moon, and M. van der Schaar are with the Department of Electrical Engineering, University of California Los Angeles, UCLA, Los Angeles, CA, 90024, USA (e-mail: ahmedmalaa@ucla.edu, mihaela@ee.ucla.edu). This work was supported by the NSF.

W. Hsu is with the Department of Radiological Sciences, UCLA, Los Angeles, CA 90024, USA (email: willhsu@mii.ucla.edu).

B. Contributions

ConfidentCare is a computer-aided clinical decision support system that assists clinicians in making decisions on which (sequence of) screening tests a woman should take given her features. ConfidentCare resorts to the realm of supervised learning in order to learn a personalized screening policy that is tailored to granular subgroups of patients. In particular, the system recognizes different subgroups of patients, learns the policy that fits each subgroup, and prompts recommendations for screening tests and clinical decisions that if followed, will lead to a desired accuracy requirement with a desired level of confidence. Fig. 1 offers a system-level illustration for ConfidentCare[1]. The system operates in two stages: an offline stage in which it learns from the EHR data how to cluster patients, and what policy to follow for every cluster, and an execution stage in which it applies the learned policy to every woman by first matching her with the closest cluster of patients in the EHR, and then approach her with the policy associated with that cluster. The main features of ConfidentCare are:

- ConfidentCare discovers a set of patients' subgroups. Given accuracy requirements and confidence levels that are set by the clinicians, ConfidentCare ensures that every subgroup of patients experiences a diagnostic accuracy, and a confidence level on that accuracy, that meets these requirements. Thus, unlike CPGs that perform well only on average, ConfidentCare ensures that the accuracy is high for every subgroup of patients.
- ConfidentCare ensures cost-efficiency, i.e. patients are not overscreened, and the sequence of recommended screening tests minimizes the screening costs.

The design of ConfidentCare is grounded to a new theoretical framework for supervised learning which entails the following technical contributions:

- We develop a new formulation for supervised learning problems where the learning task entails ensuring a high confidence level on the performance of the learner for different, disjoint partitions of the feature space, rather than the conventional formulation of supervised learning which focuses only on the average performance.
- We introduce a new notion of learnability that suits the scenarios where the goal is to carry out a constrained minimization of a cost function.
- We develop an iterative algorithm that uses breast cancer risk assessment to partition the feature space and learns a cost-sensitive, high-confidence screening policy for every partition.

We show that ConfidentCare can improve the screening cost-efficiency when compared with CPGs, and can offer performance guarantees for individual subgroups of patients with a desired level of confidence. Moreover, we show that ConfidentCare can achieve a finer granularity in its learned policy with respect to the patients feature space when it is provided with more training data. Our results emphasize the value of personalization in breast cancer screening process, and represent a first step towards individualizing breast cancer screening, diagnosis and treatment.

[1]We will revisit this figure and give a more detailed explanation for the system components in the next Section

C. Related works

1) Personalized (precision) medicine: While medical studies investigate the feasibility, potential and impact of applying the concepts of personalized medicine in the breast cancer screening process [1]-[3], [17]-[27], [30][31], none of these works provided specific tools or methods for building a personalized healthcare environment. For instance, in [17], it was shown that CPGs, which recommend screening tests only based on the age ranges, such as the European Society for Medical Oncology (ESMO) CPG and the American Cancer Society (ACS) CPG, are not cost-efficient for many subgroups of patients, where cost-efficiency was measured in terms of "costs per quality-adjusted life-year", and the authors recommended that screening should be personalized on the basis of a patient's age, breast density, history of breast biopsy, and the family history of breast cancer [28][29]. Similar results were reported in other medical studies [25]-[27], all suggesting that personalization using dimensions other than the age can yield more cost efficiency.

2) Dynamic treatment regimes: The work that relates most to ours is that on Dynamic treatment regimes (DTRs) [33]-[37]. A DTR is typically a sequence of decision rules, with one rule per stage of clinical intervention, where each rule maps up-to-date patient information to a recommended treatment [33]. DTRs aim to find an "optimal treatment policy": a sequential mapping of the patient's information to recommended treatments that would maximize the patient's long term reward. Such policies are constructed via reinforcement learning techniques, such as Q-learning. However, these works profoundly differ from the setting we consider in the following aspects: 1) DTRs are only focused on recommending treatments and do not consider screening and diagnoses; 2) DTRs does not consider cost-efficiency in the design of policies since they only consider the "value of information" in recommending treatments; 3) DTRs' complexity becomes huge when the number of patient "states" increases; 4) while confidence measures can be computed for policies in DTRs [35], the policies themselves are not designed in a way that guarantees to the clinician a certain level of reliability for every subgroup of patients.

3) Active classification for medical diagnosis: Screening and diagnostic clinical decisions typically involve "purchasing costly information" for the patients, which relates to the paradigm of active learning [38]-[45]. We note that in our setting, clinicians "purchase" costly features of the patients rather than purchasing unobserved labels, which makes our setting different from the conventional active learning framework [38]-[40]. Classification problems in which some features are costly are referred to as "active classification" [41], or "active sensing" [44]. Such problems have been addressed in the context of medical diagnosis in [41]-[45], but all these works correspond to solving an unconstrained optimization problem

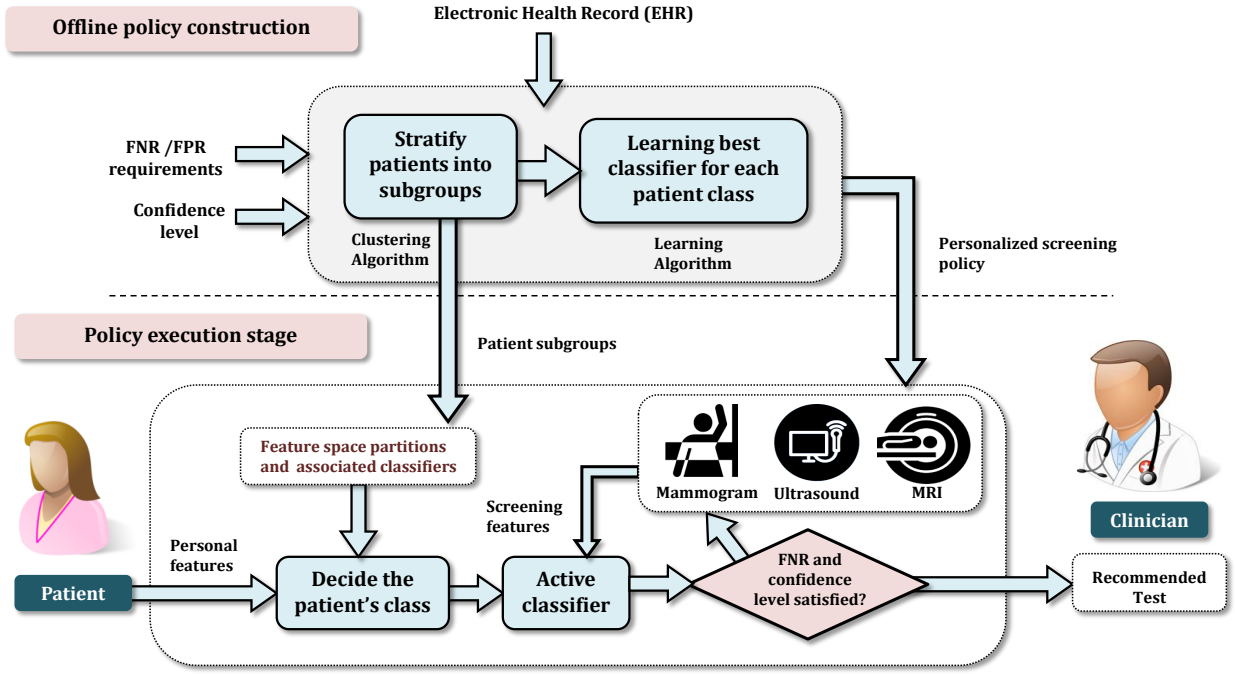

Fig. 1: Schematic of ConfidentCare described in Section II illustrating the offline policy construction and policy execution stage for new patients.

TABLE I: Comparison against existing literature

Method	Personalization	Accuracy and confidence guarantees	Cost-efficiency
DTRs	Yes	No	No
Active classification	No	No	Yes
ConfidentCare	Yes	Yes	Yes

that targets the whole population, for which no personalized accuracy or confidence guarantees can be provided. Table II positions our paper with respect to the existing literature by considering various aspects.

The rest of the paper is organized as follows. In Section II, we present a high-level view for the system components and operation of ConfidentCare. Next, in Section III, we present the technical problem formulation. In Section IV, we present the ConfidentCare algorithm, and we carry out various experiments using a dataset collected at the UCLA medical center in Section V. Finally, in Section VI, we draw our conclusions.

II. ConfidentCare: system components and operation

A. System operation

ConfidentCare is a computer-aided clinical decision support system that learns a personalized screening policy from the EHR data. By a "personalized screening policy" we mean: a procedure for recommending an action for the clinician to take based on the individual features of the patient, and the outcomes of the screening tests taken by that patient. An action can be: letting the patient take an additional screening test, proceed to a diagnostic test (e.g. biopsy), or just recommend a regular follow-up.

The tasks that ConfidentCare carries out can be summarized as follows:

- **Discover the granularity of the patient's population**: The system is provided with training data from the EHR that summarizes previous experiences of patients in terms of the screening tests they took, their test results, and their diagnoses. From such data, ConfidentCare recognizes different *subgroups* or *clusters* of patients who are similar in their features and can be approached using the same screening policy.
- **Learn the best policy for each subgroup of patients**: Having discovered the distinct subgroups of patients from the training data, ConfidentCare finds the best screening policy for each of these subgroups; by a "best" policy we mean: a policy that minimizes the screening costs while maintaining a desired level of diagnostic accuracy, with a high level of confidence that is set by the clinicians. The more training data provided to ConfidentCare, the more "granular" the learned policy leading to increased personalized recommendations for patients.
- **Identify the incoming patients' subgroups and execute their personalized policies**: After being trained, ConfidentCare handles an incoming patient by observing her

features, identifying the subgroup to which she belongs, and suggests the appropriate screening policy.

ConfidentCare can be thought of as an algorithm that stratifies patients into clusters, and automatically generates multiple CPGs, one for each cluster, in order to issue the best customized guidelines to follow for each cluster. The algorithm ensures that the accuracy of clinical decisions for each cluster satisfy a certain requirement with a certain confidence level.

B. Idiosyncrasies of breast cancer screening

Patients' features fall into two categories: *personal features*, and *screening features*. Personal features are observable at no cost, and are accessible without the need for taking any screening tests, for that they are provided by the patient herself via a questionnaire, etc. The personal features include numerical and categorical features such as: age, age at menarche, number of previous biopsies, breast density, age at first child birth, and the family history [17].

Screening tests reveal a set of costly features for the patient, which we call: the screening features. The screening features comprise the radiological assessment of breast images, usually encoded in the form of BI-RADS (Breast Imaging Report and Data System) scores [28]. The BI-RADS scores take values from the set $\{1, 2, 3, 4A, 4B, 4C, 5, 6\}$, the interpretation of which is given in Table II. BI-RADS scores of 3 or above are usually associated with followup tests or biopsy. The descriptions of all the personal and screening features are shown in Table III.

ConfidentCare considers three possible multimedia-based screening tests in the screening stage, which represent three different imaging modalities: mammogram (MG), ultrasound (US), and magnetic resonance imaging (MRI). Every screening test is associated with different costs and risks, which are functions of the patients' personal features. We consider a general cost function that incorporates both the misclassification costs in addition to the monetary costs (the detailed cost model is provided in the next subsection) [27]. ConfidentCare together with the theoretical framework developed in this section can operate upon a general class of features and tests, including genetic tests.

ConfidentCare recommends an action upon observing the outcome of a specific screening test. The actions can include: recommend a regular (1 year) followup, recommend a diagnostic test (biopsy), or an intermediate recommendation for an additional (costly) screening test (short-term followup). The final action recommended by the screening policy is either to proceed to a diagnostic test, or to take a regular followup (screening) test after 1 or 2 years. The accuracy measures that we adopt in this paper are: the false positive rate (FPR) and the false negative rate (FNR), which are defined as follows: the FPR is the probability that a patient with a negative true diagnosis (benign or no tumor) is recommended to proceed to a diagnostic test, whereas the FNR is the probability that a patient with a positive true diagnosis (malignant tumor) is recommended to take a regular followup screening test [32].

TABLE II: BI-RADS scores interpretation

Score	Interpretation
0	Incomplete.
1	Negative.
2	Benign.
3	Probably benign.
4A	Low suspicion for malignancy.
4B	Intermediate suspicion of malignancy.
4C	Moderate concern.
5	Highly suggestive of malignancy.
6	Known biopsy – proven malignancy.

TABLE III: Personal and screening features

Personal feature	Description and range of values
Age information	Age at screening test time-age at menarche-age at first child birth.
Family history	Number of first degree relatives who developed breast cancer (First degree relatives are: mother, sister, and daughter).
Number of previous biopsies	An integer number of biopsies.
Screening features	**Description**
Breast density	Described by four categories: • **Category 1**: The breast is almost entirely fat (fibrous and glandular tissue $< 25\%$). • **Category 2**: There are scattered fibro-glandular densities (fibrous and glandular tissue 25% to 50%). • **Category 3**: The breast tissue is heterogeneously dense (fibrous and glandular tissue 50% to 75%). • **Category 4**: The breast tissue is extremely dense (fibrous and glandular tissue $> 75\%$).
MG BI-RADS	Radiological assessment of the mammogram imaging.
US BI-RADS	Radiological assessment of the ultrasound test.
MRI BI-RADS	Radiological assessment of the MRI test.

C. System components

ConfidentCare is required to deal with the environment specified above and carry out the three tasks mentioned earlier, which are: discovering the granularity of the patients' population, learning the appropriate policies for each subgroup of patients, and handling incoming patients by executing the learned, personalized policy that best matches their observed features and traits. In the following, we describe the ConfidentCare algorithm, which implements those tasks using supervised learning.

The algorithm requires the following inputs from the clinician:

- A training set comprising a set of patients with their associated features, screening tests taken, and their true diagnoses.
- A restrictions on the maximum tolerable FNR.

- A desired confidence level on the FNR in the diagnoses issued by the system.

Provided by the inputs above, ConfidentCare operates through two basic stages:

- **Offline policy construction stage:** Given the training data and all the system inputs, ConfidentCare implements an iterative algorithm to cluster the patients' personal feature space, and then learns a separate *active classifier* for each cluster of patients. Each active classifier associated with a cluster of patients is designed such that it minimizes the overall screening costs, and meets the FNR and confidence requirements. The algorithm runs iteratively until it maximizes the number of patient clusters for which there exist active classifiers that can guarantee the performance and confidence requirements set by the clinician. This ensures the maximum level of personalization, i.e. ensure that the space of all patients' personal features is segmented into the finer possible set of partitions, where the performance requirements hold for each of such partitions.

- **Policy execution stage:** Having learned a policy based on the training data, ConfidentCare executes the policy by observing the personal features of an incoming patient, associates her with a cluster (and consequently, an already learned active classifier), and then the classifier associated to that cluster handles the patient by recommending screening tests and observing the test outcomes, until a final action is recommended.

Fig. 1 illustrates the components and operation of ConfidentCare. In the *offline policy construction stage*, ConfidentCare is provided with training data from the EHR, the maximum tolerable FNR, and the desired level of confidence. ConfidentCare runs an iterative algorithm that clusters the patients' personal feature space, and learns the best active classifier (the most cost-efficient classifier that meets the FNR accuracy and confidence requirements) for each cluster. In the *policy execution stage*, ConfidentCare observes the personal features of the incoming patient, associates her with a patients cluster, and then recommends a sequence of screening tests to that patient until it issues a final recommendation.

To clarify the operation of ConfidentCare, consider the following illustrative example. Assume that the set of personal features are given by a tuple *(Age, breast density, number of first degree relatives with breast cancer)*. A patient with a personal features vector (55, 40%,0) is approached by ConfidentCare. The system associates the patient with a certain cluster of patients that it has learned from the EHR data. Let the best policy for screening patients in that cluster, as computed by ConfidentCare, is to start with mammogram. If the clinician followed such a recommendation, ConfidentCare observed the mammogram BI-RADS score, say a score of 1, and then it decides to issue a final recommendation for a regular followup. If the BI-RADS score was higher, say a score of 4A, then the system recommends an additional imaging test, e.g. an MRI, and then observes the BI-RADS score of the MRI before issuing further recommendations. The process proceeds until a final recommendation is issued.

III. The personalized screening policy design

ConfidentCare uses supervised learning to learn a personalized screening policy from the EHR. In this subsection, we formally present the learning model under consideration.

1) Patients' features: Let $\mathcal{X}_d$, $\mathcal{X}_s$, and $\mathcal{Y}$ be three spaces, where $\mathcal{X}_d$ is the patients' d-dimensional personal feature space, $\mathcal{X}_s = \mathcal{B}^s$ is the s-dimensional space of all screening features, where $\mathcal{B} = \{1, 2, 3, 4A, 4B, 4C, 5, 6\}$, and $\mathcal{Y}$ is the space of all possible diagnoses, i.e. $\mathcal{Y} = \{0, 1\}$, where 0 corresponds to a *negative* diagnosis, and 1 corresponds to a *positive* diagnosis. The patients' feature space is $(d+s)$-dimensional and is given by $\mathcal{X} = \mathcal{X}_d \times \mathcal{X}_s$. Each instance in the feature space is a $(d+s)$-dimensional vector $\mathbf{x} = (\mathbf{x}_d, \mathbf{x}_s) \in \mathcal{X}, \mathbf{x}_d \in \mathcal{X}_d, \mathbf{x}_s \in \mathcal{X}_s$, the entries of which correspond to the personal and screening features listed in Table III, and are drawn from an unknown stationary distribution $\mathcal{D}$ on $\mathcal{X} \times \mathcal{Y}$, i.e. $(\mathbf{x}, y) \sim \mathcal{D}$, where $y \in \mathcal{Y}$, and $\mathcal{D}_x$ is the marginal distribution of the patients' features, i.e. $\mathbf{x} \sim \mathcal{D}_x$. The set of s available tests is denoted by $\mathcal{T}$, where $|\mathcal{T}| = s$.

The personal features are accessible by ConfidentCare with no cost, whereas the screening features are costly, for that the patient needs to take screening tests to reveal their values. Initially, the entries of $\mathbf{x}_s$ are blocked, i.e. they are all set to an unspecified value $\langle * \rangle$, and they are observable only when the corresponding screening tests are taken, and their costs are paid. We denote the space of all possible screening test observations as $\mathcal{X}_s^* = \{\mathcal{B}, \langle * \rangle\}^s$. ConfidentCare issues recommendations and decisions based on both the fully observed personal features $\mathbf{x}_d$, and a partially observed version of $\mathbf{x}_s$, which we denote as $\mathbf{x}_s^* \in \mathcal{X}_s^*$. The screening feature vector $\mathbf{x}_s$ can indeed be fully observed, but this would be the case only if all the screening tests were carried out for a specific patient.

In order to clarify the different types of features and their observability, consider the following illustrative example. Assume that we only have two personal features: the age and the number of first degree relatives who developed breast cancer, whereas we have three screening tests $\mathcal{T} = \{\text{MG}, \text{MRI}, \text{US}\}$. That is, we have that $d = 2$ and $s = 3$. Initially, ConfidentCare only observes the personal features, e.g. observing a feature vector $(42, 1, \langle * \rangle, \langle * \rangle, \langle * \rangle)$ means that the patient's age is 42 years, she has one first degree relative with breast cancer, and she took no screening tests. Based on the learned policy, ConfidentCare then decides which test should the patient take. For instance, if the policy decides that the patient should take a mammogram test, then the feature vector can then be updated to be $(42, 1, 2, \langle * \rangle, \langle * \rangle)$, which means that the BI-RADS score of the mammogram is 2. ConfidentCare can then decide what action should be recommended given that the BI-RADS score of the mammogram is 2: classify the patient as one who needs to proceed to a diagnostic test, or classify the patient as one who just needs to take a regular followup test in a 1 year period, or request an additional screening test result in order to be able to issue a confident classification for the patient.

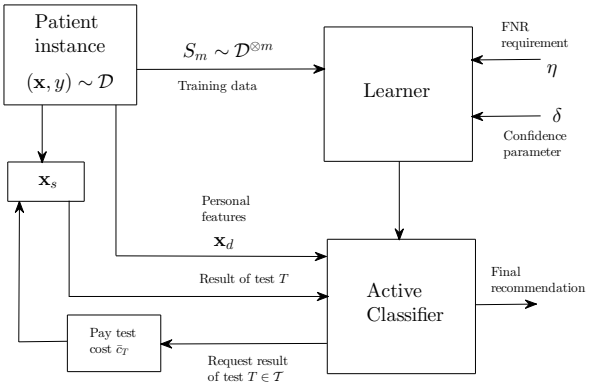

Fig. 2: Framework for the active classifier construction and operation.

2) Active classification: The process described in the previous subsection is a typical active classification process: a classifier aims to issue either a positive or a negative diagnosis (biopsy or regular followup) for patients based on their costly features (test outcomes). Such a classifier is active in the sense that it can query the clinician for costly feature information rather than passively dealing with a given chunk of data [41]. This setting should not be confused with conventional *active learning*, where labels (and not features) are the costly piece of information which the classifier may need to purchase [38][39]. In the following, we formally define an *active classifier*.

Definition 1: **(Active classifier)** An active classifier is a hypothesis (function)

$$h : \mathcal{X}_s^* \to \mathcal{Y} \cup \mathcal{T}. \quad \blacksquare$$

Thus, the active classifier either recommends a test in $\mathcal{T}$, or issues a final recommendation $y \in \mathcal{Y}$, where $y = 1$ corresponds to recommending a biopsy (positive screening test result) and $y = 0$ is recommending a regular followup (negative screening test result), given the current, partially observed screening feature vector $\mathbf{x}_s^* \in \mathcal{X}_s^*$. Whenever a test is taken, the screening feature vector is updated, based upon which the classifier either issues a new recommendation.

For instance, the range of the function h in our setting can be $\{0, 1, \text{MG}, \text{MRI}, \text{US}\}$, i.e. $\mathcal{Y} = \{0, 1\}$ and $\mathcal{T} = \{\text{MG}, \text{MRI}, \text{US}\}$. If $h(\mathbf{x}_s^*) = 0$ (or 1), then the classifier issues -with high confidence on the accuracy- a final recommendation for a biopsy or a regular followup for the patient with a screening feature vector $\mathbf{x}_s^* \in \mathcal{X}_s^*$, whereas if $h(\mathbf{x}_s^*) = \text{MG}$, then the classifier recommends the patient with a screening feature vector $\mathbf{x}_s^*$ to take a mammogram test. Note that if $h((\langle * \rangle, \langle * \rangle, \langle * \rangle)) = 0$, then the classifier recommends no tests for any patient.

3) Designing active classifiers: Designing an active classifier for the breast cancer screening and diagnosis problem under consideration cannot rely on conventional loss functions. This is because the classification problem involves costly decision making under uncertainty, and different types of diagnostic errors (false negatives and false positives) have very different consequences. Hence, our notion of learning needs to be *decision-theoretic*, and new objective functions and learning algorithms need to be defined and formulated.

We use an *inductive bias* approach for designing the active classifier; we restrict our learning algorithm to pick one hypothesis h from a specific hypothesis class $\mathcal{H}$. That is, we compensate our lack of knowledge of the stationary distribution $\mathcal{D}$ by inducing a prior knowledge on the set of possible hypothesis that the learning algorithm can output: a common approach for designing *agnostic* learners [47]. Unlike the conventional supervised learning paradigm which picks a hypothesis that minimizes a loss function, we will design a learning algorithm that picks a hypothesis from $\mathcal{H}$, such that the overall cost of screening is minimized, while maintaining the FNR to be below a predefined threshold, with a desired level of confidence; a common design objective for breast cancer clinical systems [29]. The screening cost involves both the monetary costs of the screening tests, as well as the *misclassification cost* reflected by the FPR. The FNR experienced by the patients when using an active classifier h is given by

$$\text{FNR}(h) = \mathbb{P}\left(h(\mathbf{x}_s^*) = 0 \,|h(\mathbf{x}_s^*) \in \mathcal{Y}, y = 1\right), \qquad (1)$$

whereas the FPR is given by

$$\text{FPR}(h) = \mathbb{P}\left(h(\mathbf{x}_s^*) = 1 \,|h(\mathbf{x}_s^*) \in \mathcal{Y}, y = 0\right). \qquad (2)$$

That is, the FNR is the probability that classifier h recommends a regular followup (outputs a 0) for a screening feature vector $\mathbf{x}_s$, when the patient takes all the recommended tests, given that the true diagnosis was 1, whereas the FPR is the probability that the classifier recommends a biopsy (outputs a 1) when the true diagnosis is 0. Both types of error are very different in terms of their implications, and one can easily see that the FNR is more crucial, since it corresponds to misdiagnosing a patient with breast cancer as being healthy [30]. Thus, the system must impose restrictions on the maximum tolerable FNR. On the other hand, the FPR is considered as a misclassification cost that we aim at minimizing given a constraint on the FNR [27].

Now we define the screening cost function. Let c_T be the monetary cost of test $T \in \mathcal{T}$, which is the same for all patients, and let $\bar{c}_T$ be the normalized monetary cost of test T, given by $\bar{c}_T = \frac{c_T}{\sum_{T' \in \mathcal{T}} c_{T'}}$. Let $\bar{c}(h(\mathbf{x}_s))$ be the total (normalized) monetary test costs that classifier h will pay in order to reach a final recommendation for a patient with screening feature vector $\mathbf{x}_s$. The average monetary cost of a hypothesis h is denoted as $\bar{c}(h)$, and is given by $\bar{c}(h) = \mathbb{E}\left[\bar{c}(h(\mathbf{x}_s))\right]$, where the expectation is taken over the randomness of the screening test results. To illustrate how the cost of a hypothesis is computed, consider the following example. Let the normalized costs of MG, US, and MRI be 0.1, 0.2 and 0.7 respectively. Initially, the classifier observes $\mathbf{x}_s^* = (\langle * \rangle, \langle * \rangle, \langle * \rangle)$. Assume a hypothesis h_1 and a patient with a screening features vector $\mathbf{x}_s = (3, 1, 1)$. The hypothesis h_1 has the following functional form: $h_1((\langle * \rangle, \langle * \rangle, \langle * \rangle)) = \text{MG}$, i.e. it initially recommends a

mammogram for every patient, $h_1((3, \langle * \rangle, \langle * \rangle)) = \text{MRI}$, and $h_1((3, 1, \langle * \rangle)) = 0$. Hence, using h_1, the screening cost is 0.8. Let h_2 be another hypothesis with $h_2((\langle * \rangle, \langle * \rangle, \langle * \rangle)) = \text{MG}$, $h_2((3, \langle * \rangle, \langle * \rangle)) = 0$. In this case, we have that $\bar{c}(h_2) = 0.1$, which is less than $\bar{c}(h_1) = 0.8$, yet it is clear that h_2 has a higher risk for a false negative diagnosis.

Let $C(h)$ be the *cost function* for hypothesis h, which incorporates both the average monetary costs and the average misclassification costs incurred by h. Formally, the cost function is defined as

$$C(h) = \gamma\, \text{FPR}(h) + (1 - \gamma)\, \bar{c}(h), \tag{3}$$

where $\gamma \in [0, 1]$ is a parameter that balances the importance of the misclassification costs compared to the monetary cost. $\gamma = 0$ means that ConfidentCare builds the classifiers by solely minimizing monetary costs, whereas $\gamma = 1$ means that ConfidentCare cares only about the misclassification costs. An optimal active classifier is denoted by h^*, and is the one that solves the following optimization problem

$$\begin{aligned} \min_{h \in \mathcal{H}} \quad & C(h) \\ \textit{s.t.} \quad & \text{FNR}(h) \leq \eta. \end{aligned} \tag{4}$$

Obtaining the optimal solution for (4) requires knowledge of the distribution $\mathcal{D}$, in order to compute the average FNR and cost in (4). However, $\mathcal{D}$ is not available for the (agnostic) learner. Instead, the learner relies on a size-m training sample $S_m = (\mathbf{x}_i, y_i)_{i \in [m]}$, with $S_m \overset{\text{i.i.d}}{\sim} \mathcal{D}^{\otimes m}$, where $\mathcal{D}^{\otimes m}$ is the product distribution of the m patient-diagnosis instances $(\mathbf{x}_i, y_i)_{i \in [m]}$. The training sample S_m feeds a learning algorithm $\mathcal{A} : \mathcal{S}_m \to \mathcal{H}$, where $\mathcal{S}_m$ is the space of all possible size-m training samples. The learning algorithm $\mathcal{A}$ simply tries to solve (4) by picking a hypothesis in $\mathcal{H}$ based only on the observed training sample S_m, and without knowing the underlying distribution $\mathcal{D}$. Fig. 2 depicts the framework for learning and implementing an active classifier.

4) Learnability of active classifiers: In order to evaluate the learner, and its ability to construct a reasonable solution for (4), we define a variant of the *probably approximately correct* (PAC) criterion for learning active classifiers that minimize the classification costs with a constraint on the FNR (conventional definitions for PAC-learnability can be found in [41] and [47]). Our problem setting, and our notion of learning depart from conventional supervised learning in that the learner is concerned with finding a feasible, and (almost) optimal solution for a constrained optimization problem, rather than being concerned with minimizing an unconstrained loss function.

In the following, we define a variant for the notion of PAC-learnability, the *probably approximately optimal* (PAO) learnability, of a hypothesis set $\mathcal{H}$ that fits our problem setting.

Definition 2: **(PAO-learning of active classifiers)** We say that active classifiers drawn from the hypothesis set $\mathcal{H}$ are *PAO-learnable* using an algorithm $\mathcal{A}$ if:

- $\mathcal{H}^* = \{h : \forall h \in \mathcal{H}, \text{FNR}(h) \leq \eta\} \neq \emptyset$, with $h^* = \arg\inf_{h \in \mathcal{H}^*} C(h)$, and $h^* \in \mathcal{H}^*$.
- For every $(\epsilon_c, \epsilon, \delta) \in [0, 1]^3$, there exists a polynomial function $N^*_{\mathcal{H}}(\epsilon, \epsilon_c, \delta) = poly(\frac{1}{\epsilon_c}, \frac{1}{\epsilon}, \frac{1}{\delta})$, such that for every $m \geq N^*_{\mathcal{H}}(\epsilon, \epsilon_c, \delta)$, we have that

$$\mathbb{P}_{S_m \sim \mathcal{D}^{\otimes m}} \left(C\left(\mathcal{A}\left(S_m\right)\right) \geq C(h^*) + \epsilon_c \right) \leq \delta,$$

$$\mathbb{P}_{S_m \sim \mathcal{D}^{\otimes m}} \left(\text{FNR}(\mathcal{A}\left(S_m\right)) \geq \text{FNR}(h^*) + \epsilon \right) \leq \delta,$$

  where $N^*_{\mathcal{H}}(\epsilon, \epsilon_c, \delta)$ is the *sample complexity* of the classification problem. ■

PAO-learnability reflects the nature of the learning task of the active classifier; a learning algorithm is "good" if it picks the hypothesis that, with a probability $1-\delta$, is within an ϵ from the region of feasible region, and within an ϵ_c from the optimal solution. In that sense, a hypothesis set is PAO-learnable if there exists a learning algorithm that can find, with a certain level of confidence, a probably approximately feasible and optimal solution to (4).

The sample complexity $N^*_{\mathcal{H}}(\epsilon, \epsilon_c, \delta)$ does not depend on η, yet the feasibility of the optimization problem in (4), and hence the learnability of the hypothesis class, depends on both the value of η and the hypotheses in $\mathcal{H}$. From a *bias-variance decomposition* point of view, one can view η as a restriction on the amount of inductive bias a hypothesis set can have with respect to the FNR, whereas ϵ, ϵ_c and δ are restrictions on the true cost and accuracy estimation errors that the agnostic learner would encounter. The threshold η qualifies or disqualifies the whole hypothesis set $\mathcal{H}$ from being a feasible set for learning the active classifier, whereas the tuple $(\epsilon, \epsilon_c, \delta)$ decides how many training samples do we need in order to learn a qualified hypothesis set $\mathcal{H}$. The notion of PAO-learnability can be thought of as a decision-theoretic variant of the conventional PAC-learnability, since the learner is effectively solving a constrained cost-minimization problem.

5) Patients feature space partitioning: ConfidentCare learns a different classifier separately for every subgroup of "similar" patients, which is the essence of personalization. However, the clustering of patients into subgroups is not an input to the system, but rather a task that it has to carry out; ConfidentCare has to bundle patients into M subgroups, and to each subgroup a different active classifier that is tailored to the features of the patients in that subgroup. The value of M reflects the level of personalization, i.e. the larger M is, the larger is the number of possible classifiers that are customized for every subgroup. Partitioning the patient's population into subgroups is carried out on the basis of the personal features of the patients; patients are categorized based on their personal, fully observable features.

Let $(\mathcal{X}_d, d_x)$ be a *metric space* associated with the personal feature space $\mathcal{X}_d$, where d_x is a *distance metric*, i.e. $d_x : \mathcal{X}_d \times \mathcal{X}_d \to \mathbb{R}_+$. We define an M-partitioning $\pi_M(\mathcal{X}_d, d_x)$ over the metric space $(\mathcal{X}_d, d_x)$ as a set of disjoint subsets of $\mathcal{X}_d$, i.e. $\pi_M(\mathcal{X}_d, d_x) = \{\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_M\}$, where $\mathcal{C}_i \subseteq \mathcal{X}_d$, $\bigcup_{i=1}^{M} \mathcal{C}_i = \mathcal{X}_d$, and $\mathcal{C}_j \bigcap \mathcal{C}_i = \emptyset, \forall i \neq j$. We define a function $\pi_M(\mathcal{X}_d, d_x; \mathbf{x}_d)$ as a map from the patient's personal feature vector $\mathbf{x}_d$ to the index of the partition to which she belongs, i.e. $\pi_M(\mathcal{X}_d, d_x; \mathbf{x}_d) = j$ if $\mathbf{x}_d \in \mathcal{C}_j$.

Each partition is simply a subgroup of patients who are believed to be "similar", where similarity is quantified by a

distance metric. By "similar" patients, we mean patients who have similar risks of developing breast cancer, and experience similar levels of accuracy for the different screening tests.

6) Personalization and ConfidentCare's optimization problem: A personalized screening policy is a tuple $(\pi_M(\mathcal{X}_d, d_x), [h_j]_{j=1}^M)$, i.e. a set of partitions over the personal feature space and the screening guidelines associated with each partition. Given a certain partitioning $\pi_M(\mathcal{X}_d, d_x)$ of the personal feature space, the task of the learner is to learn an active classifier $h_j \in \mathcal{H}$ for each partition $\mathcal{C}_j$, that provides (average) performance guarantees for the patients in that partition if the size of the training set is large enough, i.e. larger than the sample complexity[2]. This may not be feasible if the size of the training sample is not large enough in every partition, or if the hypothesis set has no feasible hypothesis that have a true FNR less than η for the patients in that partition. The following definition captures the extent of granularity with which a screening policy can handle the patient's population.

Definition 3: **(M-personalizable problems)** We say that the problem $(\mathcal{H}, S_m, \delta, \epsilon, \epsilon_c, \mathcal{D})$ is *M-personalizable* if there exists an M-partitioning $\pi_M(\mathcal{X}_d, d_x)$, such that for every partition $\mathcal{C}_j \in \pi_M(\mathcal{X}_d, d_x)$, $\mathcal{H}$ is PAO-learnable, and we have that $m_j \geq N^*_{\mathcal{H}}(\epsilon, \epsilon_c, \delta)$, where $m_j = \left|\mathcal{S}_m^j\right|$, and $\mathcal{S}_m^j = |\{(\mathbf{x}_i, y_i) : i \in [m], \mathbf{x}_{i,d} \in \mathcal{C}_j\}|$. ■

That is, a problem is M-personalizable if $\mathcal{H}$ has a non-empty set of feasible hypotheses for every partition, and the number of training samples in every partition is greater than the sample complexity for learning $\mathcal{H}$.

ConfidentCare constructs a feature space partitioning, i.e. the system recognizes the maximum number of patient subgroups for which it can construct separate active classifiers that meet the accuracy requirements. Designing a personalized screening policy involves partitioning $\mathcal{X}_d$ and designing an active classifier for every partition is equivalent to . Fig. 3 depicts the envisioned output of ConfidentCare for a 2D personal feature space: the feature space is partitioned into 4 partitions, and with each partition, an active classifier (a decision tree) is associated.

Let Π be the set of all possible partitioning maps for the feature space as defined in (5). ConfidentCare aims at maximizing the granularity of its screening policy by partitioning the feature space into the maximum possible number of patient subgroups, such that the active classifier associated with each subgroup of patients ensures that the FNR of this subgroup does not exceed η, with a confidence level of $1-\delta$. Thus, ConfidentCare is required to solve the optimization problem in (6). Once the optimal partitioning $\pi^*_M(\mathcal{X}_d, d_x)$ is found by solving (6), the associated cost-optimal classifiers are constructed by solving (4).

Designing a screening policy computation algorithm is equivalent to designing a partitioning algorithm $\mathcal{A}^{part} : \mathcal{S}_m \rightarrow \Pi$, and a learning algorithm $\mathcal{A} : \mathcal{S}_m^j \rightarrow \mathcal{H}$. ConfidentCare would operate by running the partitioning algorithm $\mathcal{A}^{part}$ to create a set of partitions of the personal feature space, and then running the learning algorithm $\mathcal{A}$ once for each partition in order to find the appropriate hypothesis for that partition. ConfidentCare computes an optimal screening policy if the partitioning found by $\mathcal{A}^{part}$ is a solution to (6).

[2]Note that the training set S_m is drawn from the total population of patients, but each active classifier associated with a certain partition is trained using training instances that belong to that partition only.

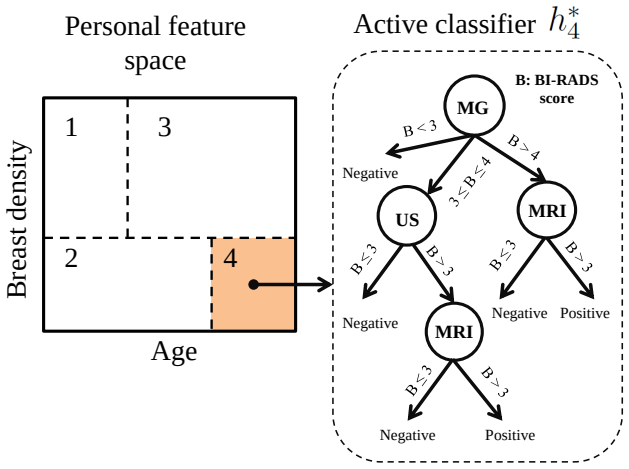

Fig. 3: An exemplary decision tree designed for a specific patient subgroup.

IV. ConfidentCare Algorithm: Analysis and Design

In this section we introduce the optimal screening policy and the ConfidentCare Algorithm.

A. Optimal screening policies: analysis and technical challenges

Theorem 1 provides an upper bound on the maximum number of clusters that can be constructed for a given dataset

Theorem 1: The maximum level of personalization that can be achieved for the problem $(\mathcal{H}, S_m, \epsilon, \epsilon_c, \delta, \mathcal{D})$ is upper-bounded by

$$M^* \leq \left\lfloor \frac{m}{N^*_{\mathcal{H}}(\delta, \epsilon, \epsilon_c)} \right\rfloor,$$

where M^* is the solution for (6).

Proof See Appendix A. ■

Theorem 1 captures the intuitive dependencies of the level of personalization on m and $(\epsilon, \epsilon_c, \delta)$. As the training sample size increases, a finer granularity of the screening policy can be achieved, whereas decreasing any of $(\epsilon, \epsilon_c, \delta)$ will lead to a coarser policy that has less level of personalization.

While Theorem 1 gives an upper-bound on the possible level of personalization, it does not tell whether such a bound is indeed achievable, i.e. is there a computationally-efficient partitioning algorithm $\mathcal{A}^{part}$, and a learning algorithm $\mathcal{A}$, through which we can we construct an optimal personalized screening policy given a hypothesis set $\mathcal{H}$ and a training sample S_m? In fact, it can be shown that for any hypothesis class $\mathcal{H}$, the problem of finding the maximum achievable level of personalization in (6) is NP-hard. Thus, there is no

$$\Pi = \left\{ \pi_M(\mathcal{X}_d, d_x) = \{\mathcal{C}_1, \ldots, \mathcal{C}_M\} \middle| \forall \mathcal{C}_i \cap \mathcal{C}_j = \emptyset, \bigcup_{i=1}^{M} \mathcal{C}_i = \mathcal{X}_d, \mathcal{C}_i \, \forall M \in \{1, 2, \ldots, |\mathcal{X}_d|\} \right\}. \tag{5}$$

$$\begin{array}{ll} \max\limits_{\pi_M(\mathcal{X}_d, d_x) \in \Pi} & M \\ \textit{s.t.} & (\mathcal{H}, S_m, \epsilon, \delta, \epsilon_c, \mathcal{D}) \text{ is } M\text{-personalizable over } \pi_M(\mathcal{X}_d, d_x). \end{array} \tag{6}$$

efficient polynomial-time algorithm $\mathcal{A}^{part}$ that can find the optimal partitioning of the personal feature space, and hence ConfidentCare has to discover the granularity of the personal feature space via a heuristic algorithm as we will show in the next subsection.

Given that we have applied a heuristic partitioning algorithm $\mathcal{A}^{part}$ to the training data, and obtained a (suboptimal) partitioning $\pi_M(\mathcal{X}_d, d_x)$, what hypothesis set $\mathcal{H}$ should we use, and what learning algorithm $\mathcal{A}$ should we chose in order to learn the best active classifier for every partition? In order to answer such a question, we need to select both an appropriate hypothesis set and a corresponding learning algorithm. We start by studying the learnability of a specific class of hypothesis sets.

Theorem 2: A finite hypothesis set $\mathcal{H}$, with $|\mathcal{H}| < \infty$, is PAO-learnable over a partition $\mathcal{C}_j \in \pi_M(\mathcal{X}_d, d_x)$ if and only if $\inf_{h \in \mathcal{H}} \text{FNR}_j(h) \leq \eta$, where FNR_j is the FNR of patients in partition $\mathcal{C}_j$.

Proof See Appendix B.

While the finiteness of the hypothesis set $\mathcal{H}$ is known to the designer, one cannot determine whether such a hypothesis set can support an FNR that is less than η since the distribution $\mathcal{D}$ is unknown to the learner. Thus, the learnability of a hypothesis set can only be determined in the learner's training phase, where the learner can infer from the training FNR estimate whether or not $\inf_{h \in \mathcal{H}} \text{FNR}(h) \leq \eta$. Theorem 2 also implies that solving the FNR-constrained cost minimization problem using the empirical estimates of both the cost and the FNR will lead to a solution that with probability $1 - \delta$ will be within ϵ_c from the optimal value, and within ϵ from the FNR constraint. Thus, an algorithm $\mathcal{A}$ that solves the constrained optimization problem in (4) "empirically" is a "good" learner for the hypothesis set $\mathcal{H}$. The key for the result of Theorem 2 is that if $|\mathcal{H}| < \infty$, then the FNR and cost functions are *Glivenko-Cantelli* classes [47], for which the *uniform convergence* property is satisfied, i.e. every large enough training sample can be used to obtain a "faithful" estimate of the costs and the accuracies of all the hypotheses in the set $\mathcal{H}$. We call the class of algorithms that solve optimization problem in (4) using the empirical cost and FNR measures as *empirical constrained cost-minimizers* (ECCM).

B. ConfidentCare design rationale

Based on Theorem 2 and the fact that (6) is NP-hard, we know that ConfidentCare will comprise a heuristic partitioning algorithm $\mathcal{A}^{part}$ that obtains an approximate solution for (6), and an ECCM learning algorithm $\mathcal{A}$ that picks a hypothesis in $\mathcal{H}$ for every partition. Since problem (6) is NP-hard, we use a *Divide-and-Conquer* approach to partition the feature space: we use a simple risk assessment-based 2-mean clustering algorithm $\mathcal{A}^{part}$ to split the a given partition in the personal feature space, and we iteratively construct a decision tree using $\mathcal{A}$ for each partition of the feature space, and then split all partitions using $\mathcal{A}^{part}$, until the algorithm $\mathcal{A}$ finds no feasible solution for (7) for any of the existing partitions if they are to be split further.

The algorithm $\mathcal{A}$ can be any ECCM algorithm, i.e. $\mathcal{A}$ solves the following optimization problem

$$\begin{aligned} &\mathcal{A}(S_m^j) = \arg\min_{h \in \mathcal{H}} \frac{1}{m_j} \sum_{(\mathbf{x}, y) \in S_m^j} \bar{c}\left(h(\mathbf{x}_s)\right) \\ &\textit{s.t.} \; \frac{\sum_{(\mathbf{x}, y) \in S_m^j} \mathbb{I}_{\{h(\mathbf{x}_s) \neq y, y=1\}}}{\sum_{(\mathbf{x}, y) \in S_m^j} \mathbb{I}_{\{y=1\}}} \leq \eta - \sqrt{\frac{\log\left(|\mathcal{H}|\right) + \log\left(\frac{4}{\delta}\right)}{2m_j}}, \end{aligned} \tag{7}$$

where the constraint in (7) follows from the sample complexity of $\mathcal{H}$, which is $N_{\mathcal{H}}^*(\epsilon, \epsilon_c, \delta) = \frac{\log(4|\mathcal{H}|/\delta)}{2\min\{\epsilon^2, \epsilon_c^2\}}$.

C. ConfidentCare algorithm

The inputs to ConfidentCare algorithm can be formally given by

- the size-m training data set $S_m = (\mathbf{x}_i, y_i)_{i \in [m]}$.
- the FNR restriction η.
- the confidence level $1 - \delta$.

The operation of ConfidentCare relies on a clustering algorithm that is a variant of Lloyd's K-means clustering algorithm [48]. However, our clustering algorithm will be restricted to splitting an input space into two clusters, thus we implement a risk assessment-based 2-means clustering algorithm, for which we also exploit some prior information on the input space. That is, we exploit the risk assessments computed via the *Gail model* in order to initialize the clusters centroids [13]-[16], thereby ensuring fast convergence. Let $\mathbf{G} : \mathcal{X}_d \to [0, 1]$ be Gail's risk assessment function, i.e. a mapping from a patient's personal feature to a risk of developing breast cancer. Moreover, we use a distance metric that incorporates the risk assessment as computed by the Gail model in order to measure the distance between patients. The distance metric used by our algorithm is

$$d(x, x') = \sum_{i=1}^{d} \beta_i |\mathbf{x}_{i,d} - \mathbf{x}'_{i,d}| + \beta_{d+1} |\mathbf{G}(\mathbf{x}_d, \tau) - \mathbf{G}(\mathbf{x}'_d, \tau)|,$$

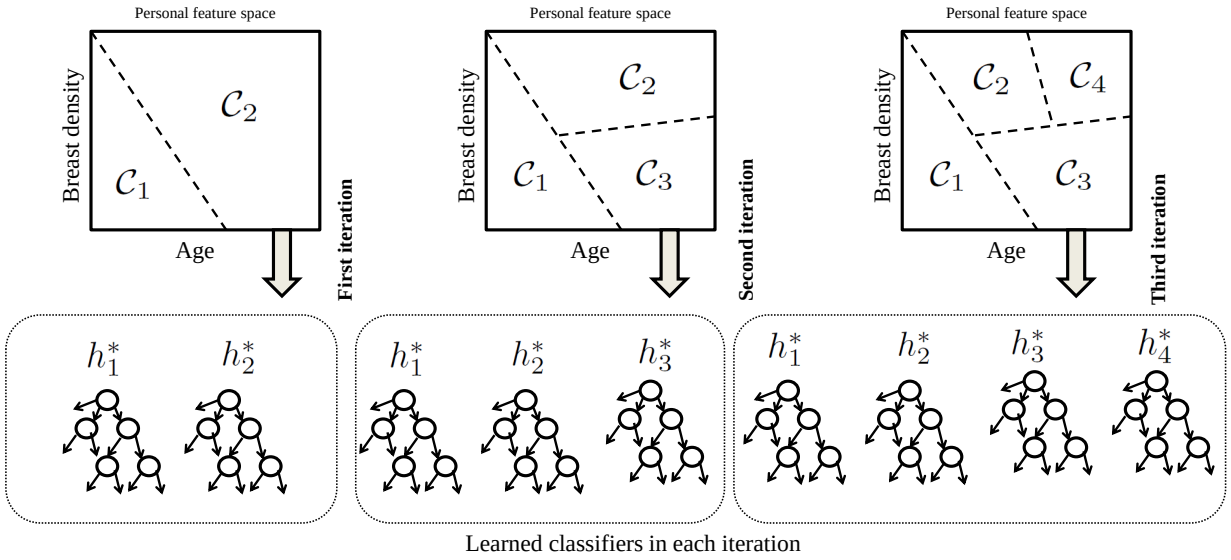

Fig. 4: Demonstration for the operation of ConfidentCare iterative algorithm. In each iteration, the personal feature space is split and a decision tree is learned for the newly emerging partition of the space.

where $\mathbf{G}(\mathbf{x}_d^i, \tau)$ is the probability that a patient with a feature vector $\mathbf{x}_d$ would develop a breast cancer in the next τ years. The parameter β quantifies how much information from the Gail model is utilized to measure the similarity between patients. In our algorithm, we adopt a risk-based clustering approach, which assigns explicitly a weight of β to the ℓ_1-norm of the difference between feature values, and a weight of $1-\beta$ to the difference in their risk assessments. Thus, the distance metric can be written as follows

$$d(x, x^{'}) = \beta||x - x^{'}|| + (1-\beta)|\mathbf{G}(x,\tau) - \mathbf{G}(x^{'},\tau)|.$$

Such a formulation explicitly merges the information extracted from the data (feature values), and the information extracted from medical domain-knowledge (risk assessment models), using a single parameter β. The value of the parameter β indicates to what extent we rely on prior (domain-knowledge) information in clustering the patients. Setting $\beta = 0$ is equivalent to stratifying the risk space, whereas $\beta = 1$ is equivalent to stratifying the feature space. The value of β needs to be learned as we show later in Section V-B.

Our clustering function, which we call $Split(\bar{\mathcal{X}}_d, d_x, \tau, \Delta)$ takes as inputs: a size-N subset of the personal feature space (training set) $\bar{\mathcal{X}}_d = \{\mathbf{x}_d^1, \mathbf{x}_d^2, \ldots, \mathbf{x}_d^N\} \subset \mathcal{X}_d$, a distance metric d_x, a Gail model parameter τ, and a precision level Δ. The function carries out the following steps:

- Compute the risk assessments $\left\{\mathbf{G}(\mathbf{x}_d^i, \tau)\right\}_{i=1}^N$ for all vectors in the (finite) input space using the Gail model. The parameter τ corresponds to the time interval over which the risk is assessed.
- Set the initial centroids to be $\mu_1 = \mathbf{x}_d^{i_*}$, where $i_* = \arg\min_i \mathbf{G}(\mathbf{x}_d^i, \tau)$, and $\mu_2 = \mathbf{x}_d^{i^*}$, where $i^* = \arg\max_i \mathbf{G}(\mathbf{x}_d^i, \tau)$.
- Create two empty sets $\mathcal{C}_1$ and $\mathcal{C}_2$, which represent the members of each cluster.
- Until convergence (where the stopping criterion is determined by Δ), repeat the following: assign every vector $\mathbf{x}_d^i$ to $\mathcal{C}_1$ if $d_x(\mathbf{x}_d^i, \mu_1) < d_x(\mathbf{x}_d^i, \mu_2)$, and assign it to $\mathcal{C}_2$ otherwise. Update the clusters' centroids as follows

$$\mu_j = \frac{1}{|\mathcal{C}_j|} \sum_{i=1}^{N} \mathbb{I}_{\mathbf{x}_d^i \in \mathcal{C}_j} \mathbf{x}_d^i, j \in \{1, 2\}.$$

- Return the clusters' centroids μ_1 and μ_2.

The rationale behind selecting the initial centroids as being the feature vectors with maximum and minimum risk assessments is that those two patients' features are more likely to end up residing in different clusters. A detailed pseudocode for the clustering function is given in Algorithm 1. As we will show later, ConfidentCare will utilize this function to iteratively partition the personal feature space.

For a given feature space partitioning, ConfidentCare builds an active classifier that emulates a "virtual CPG" for the set of patients within the partition. Designing the active classifier is equivalent to: following an inductive bias approach in which a specific hypothesis class $\mathcal{H}$ is picked, and designing an algorithm $\mathcal{A}$ that takes the training set S_m as an input and picks the "best" hypothesis in $\mathcal{H}$, i.e. $\mathcal{A}(S_m) \in \mathcal{H}$.

Adopting decision trees as a hypothesis set is advantageous since such a classifier is widely used and easily interpretable for medical applications [42]-[45]. As shown in Fig. 3, ConfidentCare will associate a decision tree active classifier with every partition of the personal feature space. Such a tree represents the policy to follow with patients who belong to that partition; what tests to recommend and how to map the BI-RADS scores resulting from one test to a new test recommendation or a diagnostic decision.

Learning the optimal decision tree $h^* \in \mathcal{H}$ is known to be an NP-hard problem [49]. Thus, we resort to greedy algorithm $\mathcal{A}$, which we call the confidence-based Cost-sensitive decision tree induction algorithm ($ConfidentTree$). The main idea

Algorithm 1: $Split(\mathcal{X}_d, d_x, \tau, \Delta)$.

Input: A set N training vectors $\bar{\mathcal{X}}_d$, $K > M$, a distance metric d_x, a Gail model parameter τ, and a precision level Δ.
Output: Two centroids μ_1 and μ_2;
Initialize $D_{-1} = 1$, $D_0 = 0$, $k = 0$, and $\mu_1 = \mathbf{x}_d^{i_*}, i_* = \arg\min_i \mathbf{G}(\mathbf{x}_d^i, \tau)$, ;
$\mu_2 = \mathbf{x}_d^{i^*}, i^* = \arg\max_i \mathbf{G}(\mathbf{x}_d^i, \tau)$;
$\mathcal{C}_1 = \emptyset, \mathcal{C}_2 = \emptyset$;
while $\frac{D_{k-1} - D_k}{D_k} > \Delta$ **do**
  $\mathcal{C}_1 = \left\{ \mathbf{x}_d^i \middle| \forall \mathbf{x}_d^i \in \mathcal{X}_d, d_x(\mathbf{x}_d^i, \mu_1) < d_x(\mathbf{x}_d^i, \mu_2) \right\}$;
  $\mathcal{C}_2 = \mathcal{X}_d / \mathcal{C}_1$;
  $\mu_1 = \frac{1}{|\mathcal{C}_1|} \sum_{i=1}^{N} \mathbb{I}_{\mathbf{x}_d^i \in \mathcal{C}_1} \mathbf{x}_d^i$;
  $\mu_2 = \frac{1}{|\mathcal{C}_2|} \sum_{i=1}^{N} \mathbb{I}_{\mathbf{x}_d^i \in \mathcal{C}_2} \mathbf{x}_d^i$;
  Set $k \leftarrow k + 1$;
  Compute the *2-means objective function* $D_k = \frac{1}{N} \sum_{j=1}^{2} \sum_{i=1}^{N} \mathbb{I}_{\mathbf{x}_d^i \in \mathcal{C}_j} d_x(\mathbf{x}_d^i, \mu_j)$;
end

of $ConfidentTree$ is to select tests (nodes of the tree) in a greedy manner by using a splitting rule that operates as follows: in each step, label the leaves that come out of each possible test such that the pessimistic estimate for the FNR (given the confidence level $1-\delta$) is less than η, and then pick the test that maximizes the ratio between the information gain and the test cost. After growing such a tree, we apply post-pruning based on confidence intervals of error estimates [50]. If there is no possible labeling of the tree leaves that satisfy the FNR requirements, the algorithm reports the infeasibility of the FNR and confidence levels set by the clinician given the training set provided to the program. More precisely, the algorithm $ConfidentTree(S_m, \pi_M(\mathcal{X}_d, d_x), j, \eta, 1-\delta)$ takes the following inputs:

- the size-m training set S_m,
- the personal feature space partitioning $\pi_M(\mathcal{X}_d, d_x)$,
- the index j of the partition for which we are designing the active classifier,
- the FNR constraint η, and
- the confidence level $1 - \delta$.

Given these inputs, the algorithm then executes the following steps:

- Extract the training instances that belong to partition $\mathcal{C}_j$.
- Grow a decision tree with the nodes being the screening tests in $\mathcal{T}$. The edges are the BI-RADS scores with the following thresholds: BI-RADS < 3, BI-RADS $\in \{3, 4\}$, and BI-RADS > 4. This classification is based on domain knowledge [28]; the first category corresponds to a probably negative diagnosis, the second corresponds to a suspicious outcome, whereas the third corresponds to a probably malignant tumor.
- Split the tree attributes as follows: for each test, label the leaves such that the pessimistic estimate (see [50] for confidence interval and error estimates in the C4.5 algorithm) for the FNR is equal to η, and then compute the cost function for each test, and select the test that maximizes the ratio between the information gain and the cost function.
- Apply post-pruning based on confidence intervals of the error estimates as in the C4.5 algorithm [50]. This step is carried out in order to avoid overfitting.
- Report the infeasibility of constructing a decision tree with the given FNR and confidence requirements if the pessimistic estimate for the FNR exceeds η.

A detailed pseudocode for $ConfidentTree$ is given in Algorithm 2. ConfidentCare invokes this algorithm whenever the personal feature space is partitioned, and the active classifiers need to be constructed.

Algorithm 2: $ConfidentTree(S_m, \pi_M(\mathcal{X}_d, d_x), j, \eta, 1-\delta)$

Input: A set of training instances S_m, a partitioning $\pi_M(\mathcal{X}_d, d_x)$, a partition index j, Maximum tolerable FNR η, and confidence level $1 - \delta$.
Output: A cost-sensitive decision-tree h_j that can be used as an active classifier for partition $\mathcal{C}_j$.;
Let B_1 be the event that BI-RADS < 3, B_2 be that BI-RADS $\in \{3, 4\}$, and B_3 be BI-RADS > 4 ;
Extract the training set that belong to the targeted partition $S_m^j = \{(\mathbf{x}_i, y_i) \,|\forall i \in [m], \mathbf{x}_{i,d} \in \mathcal{C}_j \}$;
For each test, label the leaves attached to edges B_1, B_2, and B_3 such that the empirical FNR is less than the solution of the following equation for $\hat{F}$

$$\eta = \frac{\hat{F} + \frac{Q^{-1}(\delta)}{2n} + Q^{-1}(\delta)\sqrt{\frac{\hat{F}}{n} - \frac{\hat{F}^2}{n} + \frac{Q^{-1}(\delta)^2}{4n^2}}}{1 + \frac{Q^{-1}(\delta)^2}{n}},$$

where $Q(.)$ is the Q-function and n is the number of training instances covered by the leaf for which the classification is 1. ;
Given this labeling, let $\hat{F}_p$ be the empirical value of the false positive rate, then pick the test $s \in \mathcal{T}$ that maximizes $\frac{I(s;S_m^j)}{\gamma \hat{F}_p + (1-\gamma)\bar{c}_s}$, where $I(x;y)$ is the mutual information between x and y. ;
Apply post-pruning using confidence intervals for error estimates: a node is pruned if the error estimate of its induced sub-tree i s lower than error estimate of the node.

ConfidentCare uses the modules $ConfidentTree$ and $Split$ in order to iteratively partition the feature space and construct active classifiers for each partition. ConfidentCare runs in two stages: the offline policy computation stage, and the policy execution stage. In the offline policy computation stage, the following steps are carried out:

1) Use the $Split$ function to split all current partitions of the personal feature space.
2) Use the $ConfidentTree$ to create new active classifiers for the split partitions, if constructing a decision tree for a specific partition is infeasible, stop splitting this partition, otherwise go to step (1).

After computing the policy, ConfidentCare handles the incoming patients in the policy execution stage as follows:

1) Observe the personal features of the incoming patient, measure the distance between her feature vector and the centroids of the learned partitions, and associate her with the closest partition and the associated active classifier.
2) Apply active classification to the patient. After each test outcome, ConfidentCare prompts a recommended test (the next node in the decision tree), and an intermediate diagnosis together with an associated confidence interval. The clinician and the patient will then decide whether or not to proceed and take the next test.

The pseudocode for ConfidentCare in both the offline and online modes is given in Algorithm 3. In the following Theorem, we show that the greedy ConfidentCare algorithm can guarantee a reasonable performance.

Algorithm 3: $ConfidentCare\,(S_m, \delta, \eta)$.

Input: A training set S_m, required confidence level δ, and FNR constraint η.
Output: A sequence of recommendations, intermediate diagnoses with confidence intervals, and a final diagnosis;
Offline policy computation stage: ;
Initialize $M = \infty, q = 0$;
Initialize $\mu = \emptyset$ (set of centroids of the personal feature space) ;
Hyper-parameters τ, γ, and Δ can be tuned through a validation set;
while $q \neq M$ **do**
  $M = |\mu|$;
  Create a partitioning $Part(\mathcal{X}_d, d_x)$ based on the centroids in μ ;
  For $j = 1$ to M ;
  $\mu \rightarrow Split(\mathcal{X}_d, d_x, \tau, \Delta)$;
  $h_j = ConfidentTree(S_m, \pi_M(\mathcal{X}_d, d_x), j, \eta, 1-\delta)$;
  If h_j is infeasible: $q \leftarrow q + 1$;
  EndFor
end
Policy execution stage: ;
For the incoming patient i, find the partition it belongs to by computing the distance $d_x(\mathbf{x}_{i,d}, \mu_j)$ for every partition $\mathcal{C}_j$, and associate it with the partition j^* that gives the minimum distance ;
Use classifier h_{j^*} to recommend tests and issue diagnoses

Fig. 4 demonstrates the operation of the iterative algorithm; in each iteration, partitions are split as long as a decision tree for the new partitions are feasible, and the corresponding decision trees are learned. The end result is a set of decision trees for the different partitions, representing different policies to be followed for every class of patients. Following the CPGs correspond to having a single decision tree for the entire personal feature space, which may consistently perform poorly over specific partitions of the feature space, i.e. specific subgroups of patients.

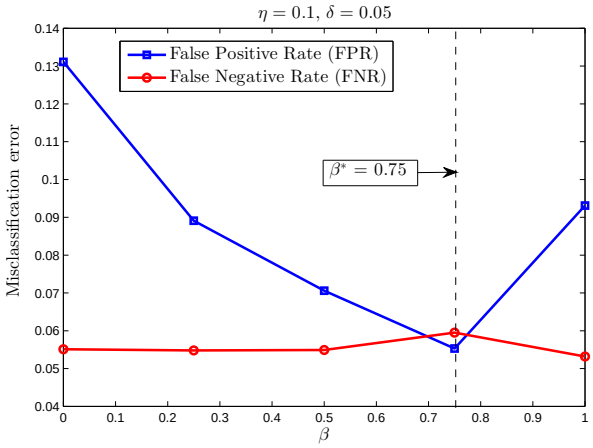

Fig. 5: Optimal selection for the distance metric parameter β for $\eta = 0.1$ and $\eta = 0.1$.

V. ConfidentCare in UCLA Medical Center

In this section, we use real-world data to illustrate the performance gains achievable with the use of ConfidentCare for breast cancer treatment. Moreover, we evaluate the performance of ConfidentCare and the added value of personalization by comparing it with CPGs, and policies that are designed in a "one-size-fits-all" fashion.

A. Real-World Dataset for Breast Cancer Patients

A de-identified dataset of 25,594 individuals who underwent screening via mammograms (MG), magnetic resonance imaging (MRI) and ultrasound (US) at the UCLA medical center is utilized to gain insight into the performance of CondfidentCare. The features associated with each individual are: age, breast density, ethnicity, gender, family history, age at menarche, and age at the first child birth. Each individual has underwent at least one of three screening tests: a MG, an MRI, an US, or a combination of those. With each test taken, a BI-RADS score is associated. Table IV shows the entries of the dataset and the features associated with every patient. The dataset is labeled by 0 for patients who have a negative diagnosis, and 1 for patients who have a positive diagnosis (malignant tumor). The dataset is imbalanced (or biased) as there are significantly more instances of MG compared with US or MRI. Moreover, most patients exhibited negative test results. Table V lists the percentages of patients who took each screening test, and the percentage of patients with positive diagnoses. All features were converted into numerical values and normalized. The normalized monetary costs for MG, US, and MRI where set to 0.1, 0.2 and 0.7 respectively, and γ is set to 0.5. In the following subsection, we demonstrate the operation of ConfidentCare.

B. Learning the distance metric

Recall from Section IV that clustering of the patients' personal feature space was carried out using a distance metric that combines both the feature values and the risk assessments as computed by the Gail risk model using the parameter β.

TABLE IV: De-identified breast cancer screening tests dataset

Patient ID	Age	Breast density	Ethnicity	Gender	Family history	Mammogram BI-RADS score	MRI BI-RADS score	Ultrasound BI-RADS score
1	71	Almost entirely fat (<25%)	U	F	Maternal Aunt	1	-	-
2	72	Almost entirely fat (<25%)	U	F	Maternal Cousin	2	1	-
3	60	Heterogeneously dense (51% - 75%)	B	F	-	2	-	-
4	66	Almost entirely fat (<25%)	W	F	Sister	1	-	-
5	56	Heterogeneously dense (51% - 75%)	W	F	-	1	-	-
. .	. .	. .	. .	. .	. .	. .	. .	. .
11,733	39	Heterogeneously dense (51% - 75%)	A	F	-	2	1	-
. .	. .	. .	. .	. .	. .	. .	. .	. .
25,594	67	Heterogeneously dense (51% - 75%)	W	F	Mother	2	2	1

TABLE V: Statistics for the dataset involved in the experiments

Category	Percentage
MG BI-RADS	93.39%.
MRI BI-RADS	2.75%.
US BI-RADS	9.21%.
Patients with malignant tumor	8.33%.

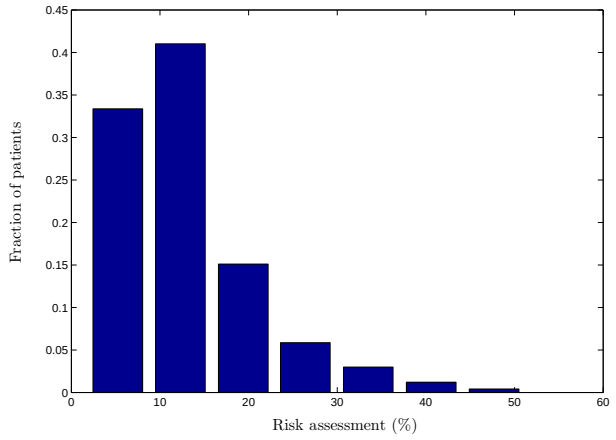

Fig. 6: Histogram for the Gail risk assessments for patients in the dataset.

Setting the parameter $\beta = 0$ corresponds to risk stratification, whereas setting $\beta = 1$ corresponds to stratifying the personal feature space while disregarding the prior information provided by the Gail model. Since the Gail model does not incorporate all the patients features (e.g. family history), one expects that the best choice of β will be between 0 and 1, for that both the personal features and the risk assessments of the patients contains (non-redundant) information about patients' similarity. As shown in Fig. 5, for an FNR constraint of $\eta = 0.1$ and confidence parameter of $\delta = 0.05$, we find that $\beta = 0.75$ is the best choice of the distance metric

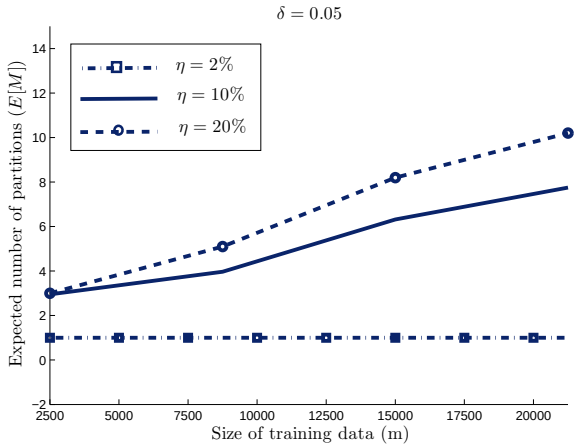

Fig. 7: The expected number of partitions (clusters) of the personal feature space versus the size of the training set .

since it maximizes the system's accuracy (FNR and FPR). This means that for $\eta = 0.1$ and $\delta = 0.05$, it is better to incorporate more information from the personal features than from the risk assessment. Our interpretation for such a result is that since most of the patients in the dataset have a low to average risks, as shown in the histogram plotted in Fig. 6, the information contained in the Gail risk assessment is not enough to differentiate between patients and bundle them into clusters. Therefore, we use the value $\beta = 0.75$ when running ConfidentCare in all the experiments. All average performance measures in this paper where obtained via 50-fold cross validation.

C. ConfidentCare Performance Evaluation

In this subsection, we investigate the operation and performance of ConfidentCare in terms of clustering and policy construction, endured monetary costs, and accuracy. As we can see in Fig. 7, ConfidentCare can (on average) discover more

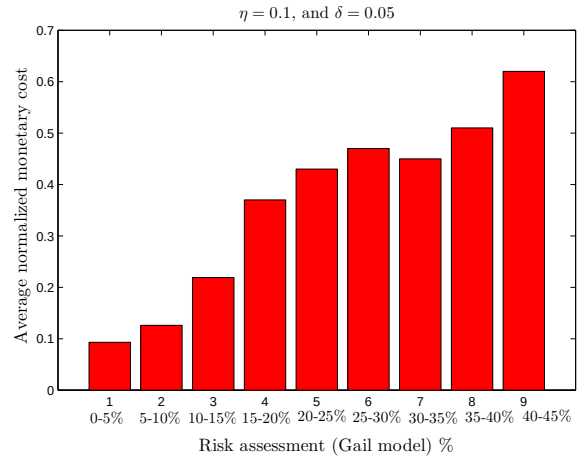

Fig. 8: Average normalized monetary cost endured by ConfidentCare for patients with different risk assessments.

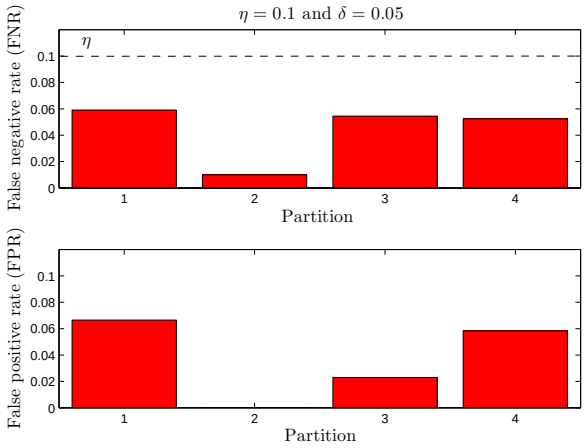

Fig. 9: FNR and FPR of ConfidentCare for different partitions of the personal feature space.

subgroups of patients for whom it can construct a screening policy with the desired confidence level as the size of the training data increases. In agreement with our expectation that the more training examples provided to ConfidentCare, the higher the number of clusters can be constructed with guaranteed performance bounds. Note that for different settings for the constraint η, the possible levels of stratification are different. For a fixed size of the training data, as the FNR constraint becomes tighter, the level of personalization decreases. For instance, we can see in Fig. 7 that the expected number of partitions for $\eta = 0.2$ is greater than that for $\eta = 0.1$, whereas for $\eta = 0.02$ the system can never find any feasible partitioning of the feature space regardless of the size of the training data.

Fig. 8 shows the average (normalized) monetary costs endured by ConfidentCare for patients with different risk assessments. As the risk level increases, the costs increase consequently since ConfidentCaare would recommend more tests (including the expensive MRI test) to patients with high level of risk for developing breast cancer. As seen, the personalized screening policy is different for each cluster.

In Fig. 9, we plot the FNR and FPR with respect to every partition constructed by the algorithm in a specific realization of ConfidentCare which was able to discover 4 partitions. It is clear that the FNR satisfies the constraint of $\eta = 0.1$ for all partitions. The FPR for different partitions, for instance we can see that partition 2 has a FPR of 0, whereas other partitions have a non-zero FPR. In Fig. 10, we show the partitions (in a 2D subspace of the original personal feature space) and the constructed policy corresponding to each cluster. It can be seen that patients who are young in age and have low breast density are recommended to take no tests, whereas other subgroups are recommended to take a MG test. We also note that the policy is more "aggressive" for patients with high breast density, i.e. for partition 3, a relatively low BI-RADS score from a MG can still lead to a recommendation for an addition US or an MRI, whereas for other subgroups the policy is more conservative in terms of recommending additional screening tests. This results as higher breast densities lead to more difficult tumor detection.

Note that Fig. 9 represents just a single realization of ConfidentCare, and thus it does not reveal the amount of confidence we have in the algorithm satisfying the FNR constraint with a high probability. In order to verify the confidence level in the policy constructed by ConfidentCare, we run the algorithm for 100 runs and see the fraction of time where the FNR in the testing set for any partition exceeds the threshold η. It can be seen that this is bounded by the specified confidence level δ.

D. ConfidentCare and Standard CPGs

We compare the performance of ConfidentCare with that of the current clinical guidelines in order to assess the value of personalization in terms of cost-efficiency. We compare the monetary cost of ConfidentCare with that of the American Cancer Society (ACS) screening guidelines issued in 2015 [51]. The reason for selecting this specific CPG is that it already applies a coarse form of risk stratification: low, average and high risk women are recommended to take different sets of tests. In Fig. 12, we plot the distribution of the normalized monetary cost of ConfidentCare together with that of the ACS over different levels of risk. ConfidentCare is expected to reduce screening costs since it supports a finer stratification of the patients, and thus recommends screening tests only to patients who need them based on thier features and previous test results. The comparison in Fig. 12 is indeed subject to the selection of η and δ by clinicians (or institutions). The more we relax the FNR and confidence constraints, the more savings we attain in terms of the monetary costs.

Finally, we compare the accuracy of ConfidentCare with that of a single decision tree of tests that is designed in a "one-size-fits-all" fashion. In particular, we build a tree of tests using the well-known C4.5 algorithm [50], and then compare its performance with that of ConfidentCare with respect to every partition found by ConfidentCare. From Fig. 13, we can see that for the same realization illustrated in Fig. 9 and 10, both approaches have a comparable FNR, but ConfidentCare outperforms a single decision tree in terms of

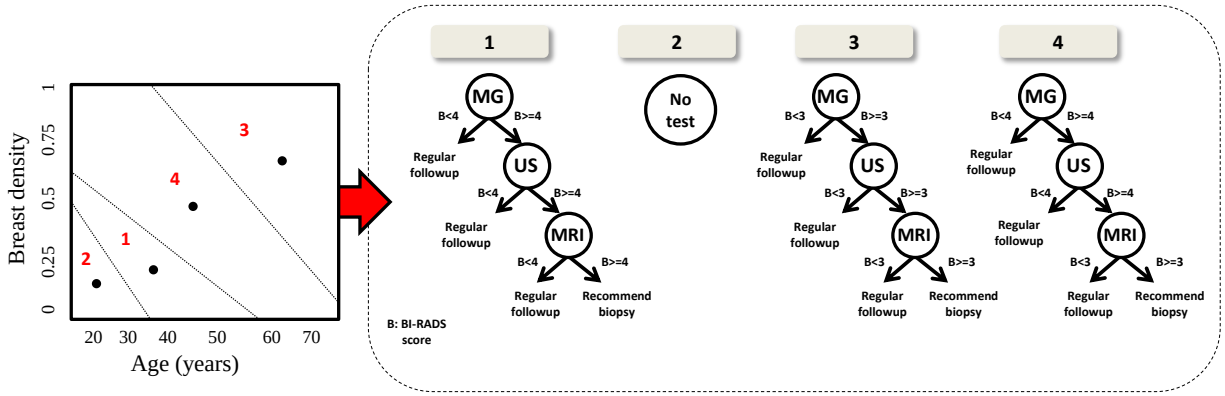

Fig. 10: The personal feature space partitions and the corresponding screening policy.

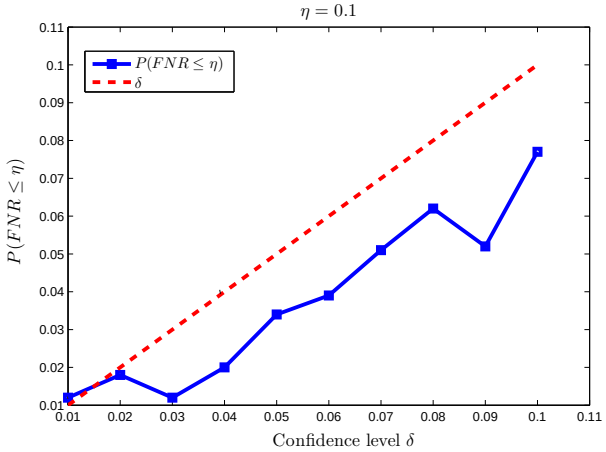

Fig. 11: The probability that the FNR of ConfidentCare is below η versus the confidence parameter δ.

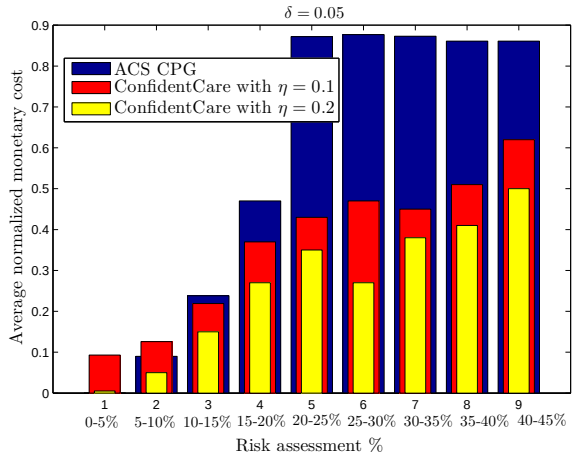

Fig. 12: Average normalized monetary cost versus risk assessment for ConfidentCare and the ACS guidelines.

the FPR for all the 4 partitions. This is because ConfidentCare deals differently with women belonging to different subgroups as shown in Fig. 10, i.e. for instance women in partition 2 are not recommended to take any tests. In other words,

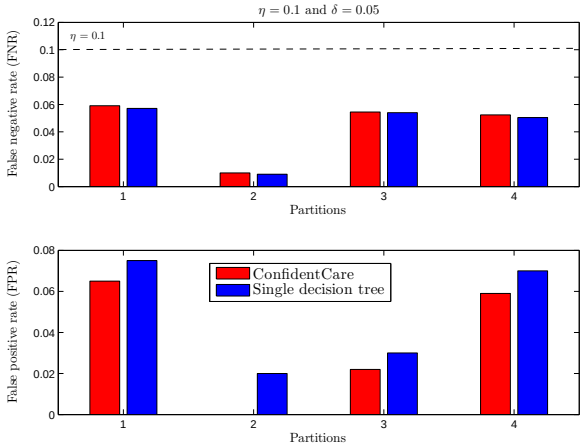

Fig. 13: FNR and FPR of ConfidentCare and a single decision tree of screening tests.

TABLE VI: FNR and FPR for ConfidentCare (with $\eta = 0.1$ and $\delta = 0.05$) and a single C4.5 decision tree

Algorithm	FNR	FPR
Single C4.5 decision tree	0.0501	0.0488.
ConfidentCare	0.0512	0.037.

ConfidentCare avoids recommending unnecessary tests, which reduces the rate of false positives. The average values of the FNR and FPR for 50 runs of ConfidentCare and a single decision tree are reported in Table VI, where a gain of 31.91% with respect to the FPR is reported.

E. Discussion and future work

The screening policy which we developed aimed to managing the short-term screening procedure, i.e. the policy was recommending a sequence of screening tests for the patient based on the outcomes of those tests, and such tests are expected to be taken in a relatively short time interval. Our framework can be extended to design policies that are concerned with the long-term patient outcomes, and are capable of not only recommend tests to the patient, but also recommend

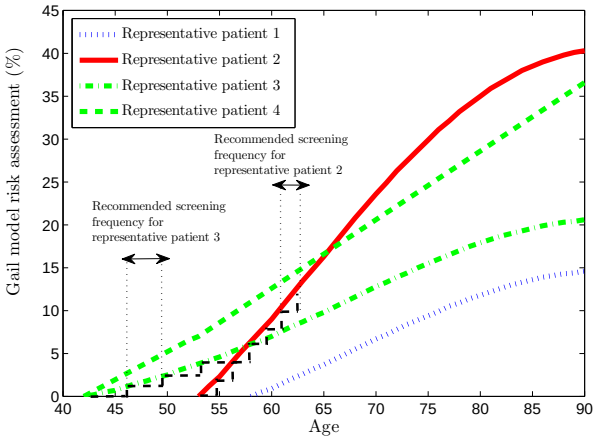

Fig. 14: Risk assessment over time for the representative patients (centroids) constructed by ConfidentCare.

the frequency with which screening should be carried out for different subgroups of patients. To see how our framework can be extended to handle such a setting, we plot the risk of developing breast cancer over time for the representative agents (centroids) of 4 clusters constructed in one realization of the algorithm in Fig. 14. Each cluster exhibits a different rate of risk growth over time, i.e. for instance while clusters 3 and 4 in Fig. 14 comprise women of almost the same age, patients in cluster 3 develop a risk for breast cancer more quickly than patients in cluster 4 due to other factors (e.g. family history). Thus, ConfidentCare can be modified to not only recommend a sequence of tests to patients in different clusters, but also to compute the optimal frequency of screening (steps over time for which the patient need to be regularly screened) that would maximize a long-term objective function. Intuitively, the frequency of screening would depend on the slope of the risk assessment over time, i.e. clusters with steeper slopes would demand more frequent screening. Our framework is well suited to capture such a setting, and the ConfidentCare algorithm can be modified to construct a screening policy that maximizes long-term outcomes with high levels of confidence.

VI. Conclusions

In this paper, we developed ConfidentCare: a clinical decision support system that learns a personalized screening policy from electronic health record data. ConfidentCare operates by stratifying the space of a womans features and learning cost-effective and accurate personalized screening policies with guaranteed performance bounds. ConfidentCare algorithm iteratively stratifies the patients' feature space into disjoint clusters and learns active classifiers associated with each cluster. We have shown that the proposed algorithm improves the cost efficiency and accuracy of the screening process compared to current clinical practice guidelines, and state-of-the-art algorithms that do not consider personalization.

Acknowledgment

We would like to thank Dr. Camelia Davtyan (Ronald Reagan UCLA Medical Center) for her valuable help and precious comments on the medical aspects of the paper. We also thank Dr. William Hoiles (UCLA) for the valuable discussions that we had with him on this paper.

Appendix A
Proof of Theorem 1

Recall that learning the optimal hypothesis for every partition requires the same sample complexity of $N^*_{\mathcal{H}}(\delta, \epsilon, \epsilon_c)$. Since the training set has m samples, then the maximum number of partitions is achieved if we can find a partitioning of the personal feature space $\pi_{M^*}(\mathcal{X}_d, d_x)$ such that:

1) The hypothesis set $\mathcal{H}$ is PAC-learnable for every partition, i.e. $\inf_{h\in\mathcal{H}} \text{FNR}_j(h) \leq \eta$ for every partition $\mathcal{C}_j$ in $\pi_{M^*}(\mathcal{X}_d, d_x)$, and $\mathcal{H}$ has a finite VC-dimension [47].
2) There are $N^*_{\mathcal{H}}(\delta, \epsilon, \epsilon_c)$ training samples in every partition.

Condition (1) has to be satisfied for any hypothesis class that will lead to the optimal level of personalization, since otherwise problem (4) will be infeasible, whereas condition (2) that we can have at most $M^* = \left\lfloor \frac{m}{N^*_{\mathcal{H}}(\delta,\epsilon,\epsilon_c)} \right\rfloor$ partitions in $\pi_{M^*}(\mathcal{X}_d, d_x)$.

Appendix B
Proof of Theorem 2

The Theorem states that for all finite hypothesis classes $\mathcal{H}$ with $|\mathcal{H}| < \infty$, a sufficient condition for PAO-learnability is that $\inf_{h\in\mathcal{H}} \text{FNR}(h) \leq \eta$. If $\mathcal{H}$, then we know that problem (4) will have a feasible solution only if $\inf_{h\in\mathcal{H}} \text{FNR}(h) \leq \eta$, thus for any PAO-learnable finite hypothesis class $\mathcal{H}$, the condition $\inf_{h\in\mathcal{H}} \text{FNR}(h) \leq \eta$ has to be satisfied.

Now we prove the converse, and show that for every finite hypothesis set $\mathcal{H}$, the condition $\inf_{h\in\mathcal{H}} \text{FNR}(h) \leq \eta$ implies PAO-learnability. Recall from the definition of PAO-learnability that $\mathcal{H}$ is learnable if

1) $\mathcal{H}^* = \{h : \forall h \in \mathcal{H}, \text{FNR}_j(h) \leq \eta\} \neq \emptyset$, with $h^* = \arg\inf_{h\in\mathcal{H}^*} C(h)$.
2) For every $(\epsilon_c, \epsilon, \delta) \in [0,1]^3$, there exists a polynomial function $N^*_{\mathcal{H}}(\epsilon, \epsilon_c, \delta) = poly(\frac{1}{\epsilon_c}, \frac{1}{\epsilon}, \frac{1}{\delta})$, such that for any $m \geq N^*_{\mathcal{H}}(\epsilon, \epsilon_c, \delta)$, there is a learning algorithm $\mathcal{A}$ for which

$$\mathbb{P}_{S_m\sim\mathcal{D}^{\otimes m}}\left(C\left(\mathcal{A}\left(S_m\right)\right) \geq C(h^*) + \epsilon_c\right) \leq \delta,$$

$$\mathbb{P}_{S_m\sim\mathcal{D}^{\otimes m}}\left(\text{FNR}(\mathcal{A}\left(S_m\right)) \geq \text{FNR}(h^*) + \epsilon\right) \leq \delta.$$

Condition (1) is already satisfied since $\inf_{h\in\mathcal{H}} \text{FNR}(h) \leq \eta$. Thus, it remains to show that the finiteness of the hypothesis set implies that condition (2) will indeed be satisfied. Note that given that $\inf_{h\in\mathcal{H}} \text{FNR}(h) \leq \eta$, which implies the feasibility of the learning problem, it suffices to prove that the functions $\text{FNR}_j(h)$ and $C_j(h)$ are Glivenko-Cantelli classes (i.e. classes that exhibit uniform convergence for any distribution $\mathcal{D}$) with respect to partition $\mathcal{C}_j$ and the hypothesis set $\mathcal{H}$ in order to

prove learnability. Note that the FNR and the cost functions are Glivenko-Cantelli classes if

$$\lim_{m\downarrow\infty} \mathbb{P}\left(\sup_{h\in\mathcal{H}}\left|\text{FNR}_j(h) - \hat{\text{FNR}}_j(h, S_m^j)\right| = 0\right) = 1,$$

$$\lim_{m\downarrow\infty} \mathbb{P}\left(\sup_{h\in\mathcal{H}}\left|C_j(h) - \hat{C}_j(h, S_m^j)\right| = 0\right) = 1,$$

where $\hat{\text{FNR}}_j(h, S_m^j) = \frac{\sum_{(\mathbf{x},y)\in S_m^j} \mathbb{I}_{\{h_j(\mathbf{x}_s)\neq y, y=1\}}}{\sum_{(\mathbf{x},y)\in S_m^j} \mathbb{I}_{\{y=1\}}}$ is the empirical FNR of hypothesis h measured based on the sample S_m^j, and similarly, $\hat{C}_j(h, S_m^j)$ is the empirical cost. In the following, we prove that both the FNR and cost functions exhibit uniform convergence. Note that in order for uniform convergence to materialize, the following conditions must be satisfied

$$\mathbb{P}_{S_m\sim\mathcal{D}^{\otimes m}}\left(\bigwedge_{h\in\mathcal{H}}\left(\left|\hat{\text{FNR}}_j(h, S_m^j) - \text{FNR}_j(h)\right| \leq \epsilon\right)\right) \geq 1-\delta,$$

$$\mathbb{P}_{S_m\sim\mathcal{D}^{\otimes m}}\left(\bigwedge_{h\in\mathcal{H}}\left(\left|\hat{C}_j(h, S_m^j) - C_j(h)\right| \leq \epsilon_c\right)\right) \geq 1-\delta,$$

which can be combined as shown in (B.2), and rewritten as a union of events as shown in (B.3), and then upper-bounded using a union bound followed by an application of Hoeffding's inequality as shown in (B.6). From (B.6), it can be seen that condition (2) is satisfied for any finite hypothesis set with $|\mathcal{H}| < \infty$, for which $N_{\mathcal{H}}^*(\epsilon, \epsilon_c, \delta) = \frac{\log(4|\mathcal{H}|/\delta)}{2\min\{\epsilon^2, \epsilon_c^2\}}$.

REFERENCES

[1] M. A. Hamburg, F. S. Collins, "The path to personalized medicine," *New England Journal of Medicine,* vol. 363, no. 4, pp. 301-304, Jul. 2010.

[2] L. Chin, J. N. Andersen, and P. A. Futreal, "Cancer genomics: from discovery science to personalized medicine," *Nature medicine,* vol. 17, no. 3, pp. 297-303, 2011.

[3] L. Hood, S. H. Friend, "Predictive, personalized, preventive, participatory (P4) cancer medicine," *Nature Reviews Clinical Oncology,* vo. 8, no. 3, pp. 184-187, Mar. 2011.

[4] https://www.whitehouse.gov/precision-medicine.

[5] D. Patil, "The White House Precision Medicine Initiative: Technology Turned Innovative Solutions," *AAAS Annual Meeting 2016*, AAAS, Feb. 11-15, 2016.

[6] M. X. Ribeiro, A. J. M. Traina, C. Traina, P. M. Azevedo-Marques, "An Association Rule-Based Method to Support Medical Image Diagnosis With Efficiency," *IEEE Trans. Multimedia*, vol. 10, no. 2, pp. 277-285, Feb. 2008.

[7] F. Liu, Y. Zhang, S. Liu, B. Zhang, Q. Liu, Y. Yang, B. Zhang, J. Luo, B. Shan, and J. Bai, "Monitoring of Tumor Response to Au Nanorod-Indocyanine Green Conjugates Mediated Therapy With Fluorescence Imaging and Positron Emission Tomography," *IEEE Trans. Multimedia*, vol. 15, no. 5, pp. 1025-1030, Aug. 2013.

[8] L. Tabar, et al., "Reduction in mortality from breast cancer after mass screening with mammography: randomised trial from the Breast Cancer Screening Working Group of the Swedish National Board of Health and Welfare," *The Lancet*, vol. 325, no. 8433, pp. 829-832, 1985.

[9] H. Weedon-Fekjaer, R. R. Pal, and J. V. Lars, "Modern mammography screening and breast cancer mortality: population study," *BMJ*, vol. 348, no. 3701, pp. 1-8, 2014.

[10] C. Harding, F. Pompei, D. Burmistrov, H. G. Welch, R. Abebe, R. Wilson, "Breast cancer screening, incidence, and mortality across US counties", *JAMA internal medicine.,* vol. 175, no. 9, pp. 1483-1489, Sep. 2015.

[11] N. J. Wald, "Guidance on terminology," *Journal of Medical Screening*, vol. 15, no. 1, pp. 50-50, 2008.

[12] B. B. Spear, M. Heath-Chiozzi, and J. Huff, "Clinical application of pharmacogenetics," *Trends Mol. Med.*, vol. 7, no. 5, pp. 201-204, May 2001.

[13] J. A. Tice, et al., "Mammographic breast density and the Gail model for breast cancer risk prediction in a screening population," *Breast cancer research and treatment*, vol. 94, no. 2, pp. 115-122, 2005.

[14] M. H. Gail, L. A. Brinton, D. P. Byar, D. K. Corle, S. B. Green, C. Schairer, J. J. Mulvihill, "Projecting individualized probabilities of developing breast cancer for white females who are being examined annually," *J. Natl. Cancer Inst.,* vol. 81, no. 24, pp. 1879-1886, Dec. 1989.

[15] J. P. Costantino, M. H. Gail, D. Pee, S. Anderson, C. K. Redmond, J. Benichou, and H. S. Wieand, "Validation studies for models projecting the risk of invasive and total breast cancer incidence," *J. Natl. Cancer Inst.,* vol. 15, no. 91, pp. 1541-1548, Sep. 1999.

[16] M. H. Gail, J. P. Costantino, J. Bryant, R. Croyle, L. Freedman, K. Helzlsouer, and V. Vogel, "Weighing the Risks and Benefits of Tamoxifen Treatment for Preventing Breast Cancern," *Journal of the National Cancer Institute*, vol. 91, no. 21, pp. 1829-1846, 1999.

[17] J. T. Schousboe, K. Kerlikowske, A. Loh, S. R. Cummings, "Personalizing mammography by breast density and other risk factors for breast cancer: analysis of health benefits and cost-effectiveness," *Annals of internal medicine,* vol. 155, no. 1m pp. 10-20, Jul. 2011.

[18] F. Cardoso, et al., "Locally recurrent or metastatic breast cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up," *Annals of oncology*, vol. 23, no. 7, 2012.

[19] S. Aebi, et al., "Primary breast cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up," *Annals of oncology*,vol. 22, no. 6, 2011.

[20] R. A. Smith, V. Cokkinides, A. C. von Eschenbach, B. Levin, C. Cohen, C. D. Runowicz, S. Sener, D. Saslow, and H. J. Eyre, "American Cancer Society guidelines for the early detection of cancer," *CA: a cancer journal for clinicians,* vol. 52, no. 1, pp. 8-22, Jan. 2002.

[21] A. C. von Eschenbach, "NCI remains committed to current mammography guidelines," *The oncologist*, vol. 7, no. 3, pp. 170-171, 2002.

[22] T. Onega, et al., "Breast cancer screening in an era of personalized regimens: A conceptual model and National Cancer Institute initiative for risk?based and preference?based approaches at a population level," *Cancer*, vol. 120, no. 19, pp. 2955-2964, 2014.

[23] S. Molinaro, S. Pieroni, F. Mariani, M. N. Liebman, "Personalized medicine: Moving from correlation to causality in breast cancer," *New Horizons in Translational Medicine,* vol. 2, no. 2, Jan. 2015.

[24] S. A. Feig, "Personalized Screening for Breast Cancer: A Wolf in Sheep's Clothing?", *American Journal of Roentgenology*, vol. 205, no. 6, pp. 1365-1371, Dec. 2015.

[25] S.-H. Cho, J. Jeon, and S. I. Kim, "Personalized Medicine in Breast Cancer: A Systematic Review," *J. Breast Cancer,* vol. 15, no. 3, pp. 265-272, Sep. 2012.

[26] J. S. Mandelblatt, N. Stout, and A. Trentham-Dietz, "To screen or not to screen women in their 40s for breast cancer: Is personalized risk-based screening the answer?" *Annals of internal medicine*, vol. 155, no. 1, pp. 58-60, 2011.

[27] J. D. Keen, "Analysis of health benefits and cost-effectiveness of mammography for breast cancer," *Annals of internal medicine*, vol. 155, no. 8, 2011.

[28] "ACR BI-RADS breast imaging and reporting data system: breast imaging Atlas 5^{th} Edition," *American College of Radiology*, 2013.

[29] L. Liberman and J. H. Menell, "Breast imaging reporting and data system (BI-RADS)," *Radiologic Clinics of North America*, vol. 40, no. 3, pp. 409-430, 2002.

[30] R. D. Rosenberg, et al., "Effects of age, breast density, ethnicity, and estrogen replacement therapy on screening mammographic sensitivity and cancer stage at diagnosis: review of 183,134 screening mammograms in Albuquerque, New Mexico," *Radiology*, vol. 209, no. 2, pp. 511-518, 1998.

[31] D. B. Rubin, M. J. van der Laan, "Statistical issues and limitations in personalized medicine research with clinical trials," *The international journal of biostatistics,* vol. 8, no. 1, Jul. 2012.

[32] K. Kerlikowske, et al., "Likelihood ratios for modern screening mammography: risk of breast cancer based on age and mammographic interpretation," *JAMA*, vol. 276, no. 1, pp. 39-43, 1996.

[33] S. A. Murphy, "Optimal dynamic treatment regimes," *Journal of the Royal Statistical Society: Series B (Statistical Methodology),* vol. 65, no. 2, pp. 331-355, May 2003.

[34] E. B. Laber, D. J. Lizotte, M. Qian, W. E. Pelham, and S. A. Murphy, "Dynamic treatment regimes: Technical challenges and applications," *Electronic journal of statistics,* vol. 8, no. 1, Jan. 2014.

$$\mathbb{P}_{S_m\sim\mathcal{D}^{\otimes m}}\left(\bigwedge_{h\in\mathcal{H}}\left(\left|\hat{C}_j(h,S_m^j)-C_j(h)\right|\leq\epsilon_c\bigwedge\left|\hat{\text{FNR}}_j(h,S_m^j)-\text{FNR}_j(h)\right|\leq\epsilon\right)\right)\geq 1-\delta \tag{B.2}$$

$$\mathbb{P}_{S_m\sim\mathcal{D}^{\otimes m}}\left(\bigvee_{h\in\mathcal{H}}\left(\left|\hat{C}_j(h,S_m^j)-C_j(h)\right|\geq\epsilon_c\bigvee\left|\hat{\text{FNR}}_j(h,S_m^j)-\text{FNR}_j(h)\right|\geq\epsilon\right)\right)\leq\delta \tag{B.3}$$

$$\mathbb{P}_{S_m\sim\mathcal{D}^{\otimes m}}\left(\bigvee_{h\in\mathcal{H}}\left(\left|\hat{C}_j(h,S_m^j)-C_j(h)\right|\geq\epsilon_c\bigvee\left|\hat{\text{FNR}}_j(h,S_m^j)-\text{FNR}_j(h)\right|\geq\epsilon\right)\right)$$

$$\leq\sum_{h\in\mathcal{H}}\mathbb{P}_{S_m\sim\mathcal{D}^{\otimes m}}\left(\left(\left|\hat{C}_j(h,S_m^j)-C_j(h)\right|\geq\epsilon_c\bigvee\left|\hat{\text{FNR}}_j(h,S_m^j)-\text{FNR}_j(h)\right|\geq\epsilon\right)\right) \tag{B.4}$$

$$\leq 2\left|\mathcal{H}\right|\left(\exp(-2m_j\epsilon^2)+\exp(-2m_j\epsilon_c^2)\right) \tag{B.5}$$

$$\leq 4\left|\mathcal{H}\right|\exp(-2m_j\min\{\epsilon_c^2,\epsilon^2\}). \tag{B.6}$$

[35] B. Chakraborty, and S. A. Murphy, "Dynamic treatment regimes," *Annual review of statistics and its application*, p. 447, 2014.

[36] E. E. Moodie, T. S. Richardson, and D. A. Stephens, "Demystifying optimal dynamic treatment regimes," *Biometrics,* vol. 63, no. 2, pp. 447-455, Jun. 2007.

[37] F. J. Diaz, M. R. Cogollo, E. Spina, V. Santoro, D. M. Rendon, and J. de Leon, "Drug dosage individualization based on a random-effects linear model," *Journal of biopharmaceutical statistics,* vol. 22, no. 3, pp. 463-484, May 2012.

[38] R. S. Kulkarni, K. M. Sanjoy, and J. N. Tsitsiklis, "Active learning using arbitrary binary valued queries," *Machine Learning,* vol. 11, no. 1, pp. 23-35, 1993.

[39] S. Tong and D. Koller, "Support vector machine active learning with applications to text classification," *The Journal of Machine Learning Research*, vol. 2, pp. 45-66, 2002.

[40] C. Persello and L. Bruzzone, "Active and semisupervised learning for the classification of remote sensing images," *IEEE Trans. Geosci. and Remote Sens.,* vol. 52, no. 11, pp. 6937-6956, Nov. 2014.

[41] R. Greiner, A. J. Grove, D. Roth, "Learning cost-sensitive active classifiers," *Artificial Intelligence,* vol. 139, no. 2, pp. 137-174, Aug. 2002.

[42] A. Freitas, A. Costa-Pereira, P. Brazdil, "Cost-sensitive decision trees applied to medical data," *in Data Warehousing and Knowledge Discovery, Springer Berlin Heidelberg*, pp. 303-312, Jan. 2007.

[43] C. X. Ling, Q. Yang, J. Wang, S. Zhang, "Decision trees with minimal costs," *in Proc. of ICML*, p. 69, Jul. 2004.

[44] S. Yu, B. Krishnapuram, R. Rosales, and R. B. Rao, "Active sensing," *International Conference on Artificial Intelligence and Statistics,* 2009.

[45] S. Lomax, and S. Vadera, "A survey of cost-sensitive decision tree induction algorithms," *ACM Computing Surveys (CSUR),* vol. 45, no. 2, 2013.

[46] V. Bryant, "Metric Spaces: Iteration and Application," *Cambridge University Press*, 1985.

[47] S. S.-Shwartz, S. B.-David, "Understanding Machine Learning: From Theory to Algorithms," *Cambridge University Press*, 2014.

[48] P. S. Lloyd, "Least squares quantization in PCM," *IEEE Trans. Info. Theory,* vol. 28, no. 2, pp. 129-137, 1982.

[49] L. Hyafil and Ronald L. Rivest, "Constructing optimal binary decision trees is NP-complete," *Information Processing Letters*, vol. 5, no. 1, pp. 15-17, 1976.

[50] J. R. Quinlan, "C4.5: programs for machine learning," *Elsevier*, 2014.

[51] http://www.cancer.org/cancer/breastcancer/moreinformation/breastcancerearlydetection/breast-cancer-early-detection-acs-recs.