Machine Learning Methods for Image-based Personalized Cancer Screening

by

Adam Yala

B.S., Massachusetts Institute of Technology (2016)
M.Eng, Massachusetts Institute of Technology (2017)

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Computer Science and Engineering

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

May 2022

Author . . . .
Department of Electrical Engineering and Computer Science
May 6, 2022

Certified by . . . .
Regina Barzilay
School of Engineering Distinguished Professor for AI and Health
Thesis Supervisor

Accepted by . . . .
Leslie A. Kolodziejski
Professor of Electrical Engineering and Computer Science
Chair, Department Committee on Graduate Students

Machine Learning Methods for Image-based Personalized Cancer Screening

by

Adam Yala

Submitted to the Department of Electrical Engineering and Computer Science
on May 6, 2022, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Computer Science and Engineering

Abstract

While AI has the potential to transform patient care, the development of equitable clinical AI models and their translation to hospitals remains difficult. From a computational perspective, these tools must deliver consistent performance across diverse populations and adapt to diverse clinical needs, while learning from biased and scarce data. Moreover, the development of tools relies on our capacity to balance clinical AI utility and patient privacy concerns. In this thesis, I will discuss our contributions in addressing the above challenges in three areas: 1) cancer risk assessment from imaging, 2) personalized screening policy design and 3) private data sharing through neural obfuscation. I have demonstrated that our clinical models offer significant improvements over the current standard of care across globally diverse patient populations. The models now underlie prospective clinical trails.

Thesis Supervisor: Regina Barzilay
Title: School of Engineering Distinguished Professor for AI and Health

Acknowledgments

First, I would like to thank my wonderful advisor, Regina Barzilay, who took a chance on me as an undergrad and fought countless battles to make this thesis work possible. I am perpetually inspired by her indomitable resilience, her boundless creativity and by her intense yet nurturing leadership. Intense, direct and empowering, Regina's mentorship enabled me to do my best work. Again and again (and across multiple new areas), Regina has shown me how exciting and life-changing research can be. Regina is a "boundary-pushing researcher crossed with Gal Gadot"[151]. Her shining example is what motivated me to pursue professorships, and I hope to build a research group as ambitious, impactful and empowering. Like my floofy dog Arya 0-1, I'm incredibly fortunate to have academically grown up in Regina's lab.

I also want to thank my committee members Tommi Jaakkola and Muriel Medard who have regularly provided me with brilliant insights and thoughtful feedback. Across projects in rationales, chemistry and privacy, I have been frequently awestruck at Tommi's technical clarity and his immediate insights; I'm lucky to have to have learned so much from him. I've also been incredibly fortunate to work with Muriel on privacy; she opened up this emerging area and I'm immensely lucky to have benefited from her theoretical insights, enthusiasm, rigor and thoughtful leadership. It has been so much fun to cross fields together. I am also forever grateful to Kevin Hughes, our first clinical collaborator at MGH and the one who opened the door for us to make an impact in cancer care. From our first pathology NLP paper[142] to our most recent JCO paper[145], we have been incredibly fortunate to benefit from Kevin's clinical depth, creativity, and eagerness to use new technologies.

I also want to thank my dear friends, lab-mates and collaborators at MIT for our many research discussions and collaborations; I'm lucky to have grown from the wisdom of many other students. I'm grateful to Jeremy Wohlwend, Peter Mikhael, Victor Quach, Tao Lei, Karthik Narasimhan, Tal Schuster, Tally Portnoi, Darsh Shah, Kyle Swanson, Homa Esfahanizadeh, Rafael D'Oliveira, Janice Yang, Ludvig Karstens, Yujia Bao and all the members of Regina's group. I also want to thank our labs wonderful admin assistant, Marcia Davidson, who has always been incredibly helpful in navigating the internals of MIT. I also want to thank my clinical collaborators outside of MIT, including Fredrik Strand, Gigin Lin, Lecia Sequist, Leslie Lamb, Florian Fintelmann, Hari Trivedi, Judy Chichoya, Michal Guindy, the Protea team, Silvia Sabino, Yung-Liang Wan and the Quanta team. This thesis, and the global clinical validation within, would not have been possible without these pioneering clinical collaborators. I also want to thank Casey Ross, my incredible collaborator from STAT news for our work together. I'm also grateful to the friends who've shared so many adventures with me along my journey and supported me through the crux sequences of my PhD, including Harry Bleyan, Cathie Yun, Sean Murphy, Nico Gomez, Elmer Guardado Sanchez, Kenneth Wong, Rachel Adler, Jeremy Wohlwend, Saloni Behl, Nicola Greco, Niccolo Pignatelli, Sebastien Velez and my brothers Nacim and Ryan. I also want to thank Zoe Xiong, who has made this last year so joyful and fulfilling. I also thank Arya, the best of all possible dogs, who has been alongside me for my entire PhD 0-1.

Finally, I would like to thank my parents, Nadia and Abdelhamid, for their endless support and inspiration. My parents have made countless sacrifices to bring me to the US, and to enable me to dream without restriction. This thesis is dedicated to them.

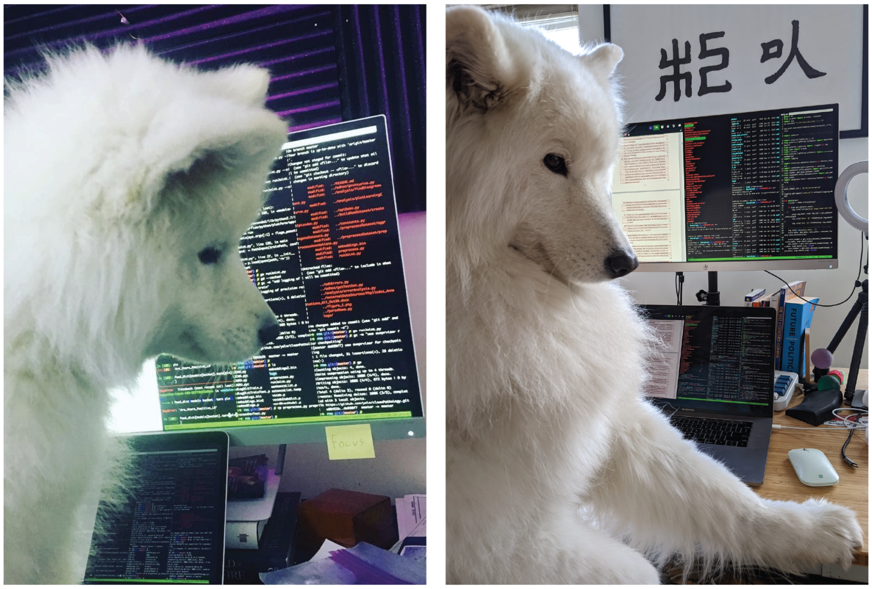

Figure 0-1: Illustration of growth over the course of my PhD, via Arya. Left: Arya when I was a master's student. Right: Arya when I was finishing this thesis.

Bibliographic Notes

Portions of thesis are based on previously published peer reviewed papers. Chapter 4 is currently under review. The list of reference publications by chapter is provided bellow:

- **Chapter 2: Towards robust mammography-based models for breast cancer risk.** Adam Yala, Peter G Mikhael, Fredrik Strand, Gigin Lin, Kevin Smith, Yung-Liang Wan, Leslie Lamb, Kevin Hughes, Constance Lehman, and Regina Barzilay. *Science Translational Medicine 13.578 (2021)* [146] and **Multi-Institutional Validation of a Mammography-based Breast Cancer Risk Model** Adam Yala, Peter Mikhael, Fredrik Strand, Gigin Lin, Siddharth Satuluru, et al. *Journal of Clinical Oncology 2021* [145]

- **Chapter 3: Optimizing risk-based breast cancer screening policies with reinforcement learning** Adam Yala, Peter Mikhael, Constance Lehman, Gigin Lin, Fredrik Strand, Yung-Liang Wang, Kevin Hughes et al. *Nature Medicine (2022).* [144]

- **Syfer: Neural Obfuscation for Private Data Release** Adam Yala, Victor

Quach, Homa Esfahanizadeh, Rafael GL D. Oliveira, Ken R. Duffy, Muriel Medard, Tommi S. Jaakkola, Regina Barzilay. *Preprint Arxiv (2022)* [147]

The code of the work presented is available at: github.com/yala .

Contents

List of Figures

List of Tables

Chapter 1

Introduction

While recent advances in machine learning have led to superhuman model performance on many tasks, developing equitable clinical AI tools that can readily be translated to clinical care remains difficult. From a computational perspective, these tools must deliver consistent performance across diverse populations while learning from biased and noisy data. Moreover, it is often not clear how to effectively translate clinical questions into AI models, or how to leverage them to benefit patient care given varying clinical constraints. Finally, the development of clinical AI tools is severely restricted by patient privacy considerations, resulting in a dearth of public datasets. To address these challenges, this thesis focuses on developing modeling approaches that are robust to data generation biases, can adapt to diverse clinical requirements and enable new privacy-utility trade-offs. I will present contributions in three areas:

1. Predicting future cancer risk

2. Designing personalized screening policies

3. Private data sharing

The clinical tools presented in this work offer significant improvements over the current clinical standard across globally diverse patient populations, and they are implemented at multiple hospitals.

1.1 Predicting Cancer Risk

Early detection key to improving cancer outcomes. Across multiple cancers, randomized clinical trails[33, 117, 118, 113, 39] have demonstrated that screening can significantly decrease cancer mortality by diagnosing the disease at earlier stages; for instance, mammography has been shown to decrease breast cancer mortality by 25% [39] and low-dose computed tomography was shown to decrease lung cancer mortality by 24%[33]. These findings have motivated considerable investments in population screening, with the United States spending an estimated 7.8 billion dollars on mammography screening per year[86] . While screening can improve patient cancer outcomes, the practice also carries significant harms including false positives, unnecessary biopsies, and patient anxiety. To balance the benefits of screening against the harms, screening programs rely upon *risk models*, which predict the probability of a patient developing cancer in the future, to determine how to allocate screening. Intuitively, patients at higher risk of breast cancer should be afforded more sensitive screening regimes, improving early detection, while patients at lower risk should receive less screening, minimizing over-treatment. Improvements in risk modeling would enable clinical guidelines to achieve further improve patient outcomes (through earlier detection or prevention) while reducing over treatment. However, despite decades of effort, the accuracy of risk models used in clinical practice remains modest. For instance, the Tyrer-Cuzick (TC) [123] and Gail [42] models achieved areas under the curve (AUCs) of 0.62 and 0.59, respectively, in a prospective UK screening cohort [19].

Since the first breast cancer risk model in 1989 [42], traditional risk assessment models have relied on a small number of categorical variables encoding patient demographics and clinical history combined with traditional statistical models to predict risk[42, 123, 27]. The improvement of these models has primarily relied on identifying new categorical variables to incorporate (i.e. feature engineering). For example, previous research [101, 27] explored multiple risk factors related to hormonal and genetic information and Brentnall et al [19] incorporated mammographic breast density, an expert defined imaging biomarker, into the Gail risk model and Tyrer-Cuzick model

(TC). The key limitation of expert defined risk markers, such as breast density, is that they cannot capture the full richness of a patient's phenotype, limiting the accuracy of the resulting risk models. Same-age patients who are assigned the same density score can have drastically different mammography with vastly different outcomes. Whereas previous studies [68, 16, 18] explored automated methods to assess breast density, these efforts reduced the mammographic input into a few statistics largely related to volume of glandular tissue that are not sufficient to distinguish patients who will and will not develop breast cancer.

We hypothesized that there are subtle but informative cues on mammograms that may not be discernible by human experts or simple volume-of-density measurements, and deep learning methods could leverage these cues to yield significantly improved risk models. However, developing clinically meaningful risk models directly from imaging requires addressing several unique computational challenges. Risk models must provide predictions at various time points in the future (e.g. one to five years) while learning from patient data with variable amounts of followup information. They should benefit from potentially missing non-image data, such as age and family history, and they must perform consistently across across heterogeneous mammography devices. Finally, to ensure equitable improvements in care, clinical AI tools must demonstrate robust performance across diverse screening populations.

To address these challenges, we developed Mirai, a mammography-based deep learning model designed to predict cancer risk. Mirai significantly outperformed the clinical standard of care, the Tyrer-Cuzick (TC) model, obtaining a five-year AUC of 0.76 compared to 0.62 by TC, and it maintained its accuracy across seven hospital systems from five countries[146, 145]. To achieve this performance, we designed Mirai to address the key requirements of cancer risk modeling. To estimate risk at multiple time points, we explicitly designed the model architecture to decompose the prediction of cumulative risk into multiple intermediate hazard predictions. This decomposition allows the model to leverage different features to predict long and short term risk, while producing self consistent predictions. To benefit from non-image risk factors (e.g. age, family history) that may be missing at test time, we trained Mirai in a multi-task fashion to predict non-image risk factors from the imaging. This auxiliary objective acts as an additional regularizer while enabling the model to conditionally impute non-image information if it isn't available. Finally, to ensure that Mirai obtained consistent accuracy and calibration across heterogeneous mammography machines, we proposed a conditional adversarial training scheme to eliminate hardware specific biases. Altogether, our work [145] demonstrates that image-based breast cancer model can offer broad and equitable improvements in care.

1.2 Designing Personalized Screening Policies

Effective population cancer screening programs must balance the benefits of early detection against the harms of over screening. This capacity relies on both our ability to predict future cancer risk, and our ability to design personalized risk-based screening policies. While recent deep learning advances have transformed cancer risk prediction (as discussed in 1.1), our ability to design risk-based screening policies still lags behind. Existing screening guidelines [32, 110] still rely on only a few features, such as a patient's age and smoking history, to decide who should get screened and how often, limiting their ability to personalize care. As a result, these programs generally assign all patients a single screening regime (e.g. annual screening) for decades. Novel AI-driven risk models[146, 143] offer us the opportunity to transform population screening. AI-based risk models can capture complex dependencies in a patients high dimensional data, yielding improved accuracy, and they offer a dynamic snapshot of a patient's risk as the patient's raw data evolves over time. To take full advantage of dynamic AI-based risk models, we propose computational framework, Tempo[144], to derive agile screening policies that can adjust the screening regime as the patients risk evolves. We hypothesized that by pairing AI-based risk models with AI-driven policy design, we could uncover significantly more efficient cancer screening policies, yielding both improved early detection and reduced over screening.

We can view breast cancer screening as a markov decision process, where we wish to develop a policy that can map a patient's risk (state) to a screening followup recommendation (action) in order to maximize that patient's chances of early detection while minimizing their screening harms (reward). With this formulation, reinforcement learning (RL) algorithms could train policies to make a sequence of screening-followup decisions to maximize the future reward. However, applying RL algorithms for screening policy design, a setting where only retrospective screening data is available, poses several unique challenges. First, the training data only includes patient risk assessments (states) when mammograms were taken; however, to simulate how a novel policy would have acted for a patient, we need also need to know the risk assessment at intermediate points. As result, we train an generative model that learns to interpolate a patient's risk at unobserved time points from observed screenings. This *risk progression model* allows us to leverage the retrospective screening trajectories as a full simulation environment to evaluate and train policies. Moreover, while we can leverage counterfactual reward metrics to evaluate the screening cost and early detection benefit of a policy, different hospital systems with disparate local resource constraints, have diverse desired weightings between different rewards (i.e. early detection benefits and screening costs). Furthermore, these hospital preferences are not known at training time. To enable our policies to support diverse and unknown clinical preferences, we condition our policies on the desired clinical trade-off and train our policies to generalize across possible preferences with multi-objective Q-learning[149].

We demonstrated the efficacy of Tempo in the context of breast cancer. We trained our risk-based screening policies on a large screening mammography dataset from Massachusetts General Hospital (MGH; USA) and validated this dataset in held-out patients from MGH and external datasets from Emory University (Emory; USA), Karolinska Institute (Karolinska; Sweden) and Chang Gung Memorial Hospital (CGMH; Taiwan). Across all test sets, we find that the Tempo policy combined with an image-based risk model is significantly more efficient than current regimens used in clinical practice in terms of simulated early detection per screen frequency. Moreover, we show that the same Tempo policy can be easily adapted to a wide range of possible screening preferences, allowing clinicians to select their desired trade-off between early detection and screening costs without training new policies. Finally, we demonstrate that Tempo policies based on AI-based risk models outperform Tempo policies based on less accurate clinical risk models. Altogether, our results show that pairing AI-based risk models with agile AI-designed screening policies has the potential to improve screening programs by advancing early detection while reducing over-screening.

1.3 Private Data Sharing

Data sharing remains a central challenge to the development for equitable clinical AI tools. The creation of public medical datasets is severely restricted by privacy regulations [54, 43] that aim to prevent the leakage of identifiable medical data We consider a scenario where a single large hospital wishes to release a large labeled medical imaging dataset (e.g. mammograms with cancer labels) to enable untrusted third parties to develop clinically meaningful AI models while preventing patient reidentification. To this end, we develop an encoding scheme for private data release.

An ideal encoding scheme would enable model development for arbitrary downstream tasks using standard machine learning tools while preventing raw data reidentification; moreover, this scheme should be computational efficient and not require data owners to train their own models (e.g. generative models[59, 141]) or to leverage expensive cryptographic primitives [44]. Designing such an encoding scheme has remained a long standing challenge for the community. For instance, differentially private approaches pursue this goal by adding independent random noise to limit the sensitivity of the encoding outputs to the input data. While these methods afford strong theoretical privacy guarantees, the magnitude of random noise to needed obtain privacy often results in too large of a utility loss in clinical tasks for practical use. More recently, several lightweight heuristic encoding schemes [55, 139] have been shown to achieve better modeling utility. However, these schemes only offer privacy in the best case, when the raw data distribution follows strong assumptions (e.g. the data is Gaussian [139]), and they do not offer privacy on real medical imaging datasets such as X-rays[147].

In dissertation, we propose *Syfer*, a neural obfuscation method to protect against re-identification attacks. In our framework, data owners encode their data with a random neural network (acting as their private key) for public release. While arbitrary random neural networks are not sufficient to achieve privacy on real word data (e.g. X-rays), we demonstrate how we how to shape the distribution of private keys by composing random layers with trained *obfuscator* layers. The obfuscator layers are trained on public data to precondition random transformations in order to achieve privacy against an estimated attacker while maintaining the invertability of the whole transform. In doing so, we learn a distribution of random encoders that tightly adapted to the characteristics of real world X-rays. To characterize the privacy utility trade-offs of complex encoding schemes on real world data, we introduce a flexible computational attacker and a realistic chest X-ray utility benchmark. We demonstrated that our scheme could obtain a 25 point AUC improvement over a differentially private baseline while maintaining high guesswork, a well-known metric in password security. Our results demonstrate that learned encoding schemes can offer significantly improved privacy utility trade-offs while supporting arbitrary and unknown downstream tasks.

1.4 Outline

The rest of this thesis is organized as follows:

- **Chapter 2** presents methods predicting cancer risk from medical imaging. The models significantly outperformed clinical standard in large-scale validation study across seven hospital systems from five countries.

- **Chapter 3** introduces a reinforcement-learning based framework for deriving personalized screening policies from longitudinal screening data and AI-based risk models (as explored in Chapter 2). We showed that this framework was significantly more efficient than existing clinical guidelines across diverse test sets from four hospitals, achieving earlier detection for lower screening costs.

- **Chapter 4** proposes a neural obfuscation method for encoding private data to protect against re-identification attacks. We demonstrated that X-ray classifiers built using this scheme obtained strong privacy and approached the performance of classifiers trained on raw images.

Chapter 2

Predicting Future Cancer Risk from Medical Images

2.1 Introduction

It is estimated that 39 million mammograms are performed in the United States every year [125, 67], with $1.1 billion dollars being spent by Medicare alone [49]. Despite the wide adoption of breast cancer screening, the practice is riddled with controversy. Proponents of more aggressive screening strategies aim to maximize the benefits of early detection [115, 50, 51, 77, 29, 114], whereas advocates of less frequent screening aim to reduce the false-positive assessments, anxiety, and costs for the patients who will never develop breast cancer [121, 103, 14, 20, 120]. As a result, in the United States, there are multiple guidelines with different recommendations about when to start screening, how often to get screened, and when supplemental screening is needed [132, 87, 109, 83, 84, 110]. We argue that both goals of earlier detection and reducing overtreatment can be achieved by leveraging more accurate risk models. With improved risk-based guidelines, we can offer more sensitive screening to patients who will develop cancer, achieving earlier detection while reducing unnecessary screening and overtreatment for the rest. Moreover, because of the scale of breast cancer screening, even modest improvements in screening guidelines have the potential to benefit a wide patient population.

All guidelines currently in clinical use leverage risk models. Some guidelines [84] use risk models as simple as a patient's age to determine whether, and how often, a woman should get screened, whereas others [87] combine multiple factors relating to age, hormonal factors, genetics, and mammographic breast density to determine whether supplemental imaging should be considered. However, despite decades of effort, the accuracy of risk models used in clinical practice remains modest. For instance, the Tyrer-Cuzick [123] and Gail [42] models achieved areas under the curve (AUCs) of 0.62 and 0.59, respectively, in a prospective UK screening cohort [19]. Recently, image-based deep learning models have shown considerable promise [143, 36], obtaining AUCs up to 0.70 for assessing 5-year risk and advancing the state of the art. However, to bring an image-based risk model to the clinic, we not only need to further improve its accuracy but must also validate its performance at scale across diverse populations and clinical settings. Furthermore, we need to demonstrate that it can identify more accurate high-risk cohorts. Here, we aimed to achieve all three of these goals by developing Mirai and studying its performance across seven hospital systems across the United States, Israel, Sweden, Taiwan, and Brazil.

2.2 Results

2.2.1 Overview of algorithm

In computational terms, risk assessment can be viewed as a prediction task, where the model is trained to associate features of mammograms with future cancer diagnoses. Although this setup, referred to as supervised learning, is commonly used for medical tasks [65, 94, 6, 31, 52], risk modeling also poses several unique requirements. It requires risk prediction at various time points, the ability to leverage potentially missing nonimage data (such as age and family history), and consistent performance across heterogeneous mammography devices. Inherent to risk modeling is learning from patients with variable amounts of follow-up and needing to assess risk at different time points. Although it is possible to train separate models to assess risk for each time point based on patients with the corresponding amount of follow-up (1 to 5 years), this approach can result in mutually inconsistent risk assessments. For instance, a model could predict that a patient has a higher risk of developing cancer within 2 years than within 5 years. Moreover, this approach does not leverage the inherent relationship between assessing risk at different time points. We address this by training a single model to predict risk at all time points and by explicitly designing the architecture to produce self-consistent predictions. This formulation also enables the model to learn from data with variable amounts of follow-up. Although our method primarily focuses on mammograms, we also wanted to leverage nonimage risk factors (for example, age and hormonal factors) if they were available. An obvious mechanism for incorporating nonimage risk factors is to add them as an input to the model jointly with the image. However, this design would prevent hospitals that do not collect this kind of information from using the model. Although we could impute this missing information by using a reference population, that would not take into account the relationship between the mammogram and the risk factors. To address this challenge, we trained our model to predict risk factor values from the mammogram, enriching our original objective with this new prediction task. This formulation enabled the model to benefit from available risk factor data while allowing it to impute the information if it is missing. To incorporate deep learning risk models into clinical guidelines, the models must be consistent across a range of mammography devices, in other words, they must predict the same risk for a patient regardless of the mammography device. We addressed this challenge by adopting a conditional-adversarial training scheme [153]. This training regime forces the model to induce image representation in a device-invariant fashion and to produce consistent risk assessments. Our full model, named Mirai, is depicted in Fig 2-1. It takes as input all standard views of a mammogram: left craniocaudal (L CC), left mediolateral-oblique (L MLO), right craniocaudal (R CC), and right mediolateral-oblique (R MLO). Mirai consists of four modules: an image encoder, an image aggregator, a risk factor predictor, and an additive-hazard layer. A run through the model works as follows: first, we pass each mammogram view independently through the image encoder. Next, we take each image representation as well as which view it came from (for example, L CC and R MLO), and pass it into the image aggregation module to combine information across views and obtain a representation of the entire mammogram. Given this rich representation of the mammogram, we then predict a patient's traditional risk factors as used in Tyrer-Cuzick (such as age, weight, and hormonal factors) and refer to this as our risk factor prediction module. If risk factor information is not available at inference time, we then use the predicted values. Next, we take the mammogram representation from our image aggregator, combined with our risk factor information (predicted or given), and predict a patient's risk with an additive-hazard layer. The additive-hazard layer predicts a patient's risk for each year over the next 5 years. Architectural details for each module are presented in the Methods, and all code is released.

2.2.2 Training and testing at MGH

We developed Mirai using the Massachusetts General Hospital (MGH) dataset, which consists of 210,819, 25,644, and 25,855 examinations from 56,786, 7020, and 7005 patients, for the training, validation, and test sets, respectively. This dataset contained detailed risk factor information, as used in Tyrer-Cuzick version 8 (TCv8), that was available at the time of mammography. The distribution of clinical risk factors in the MGH dataset, as used by TCv8, is shown in table B.2. A flowchart illustrating the construction on the MGH dataset is shown in Fig. 2-3.

To determine the impact of using predicted risk factors on Mirai's performance, we evaluated the model both when using the electronic health record-based and predicted risk factors, referring to the two scenarios as "Mirai with risk factors" and "Mirai without risk factors," respectively. We compared Mirai against three alternative risk models: Hybrid DL [143], Image-Only DL [143], and TCv8. Hybrid DL is a deep learning model based on both mammograms and traditional risk factors, and Image-Only DL is a deep learning model based only on mammograms. Hybrid DL requires traditional risk factors to predict risk, whereas Image-Only DL does not use

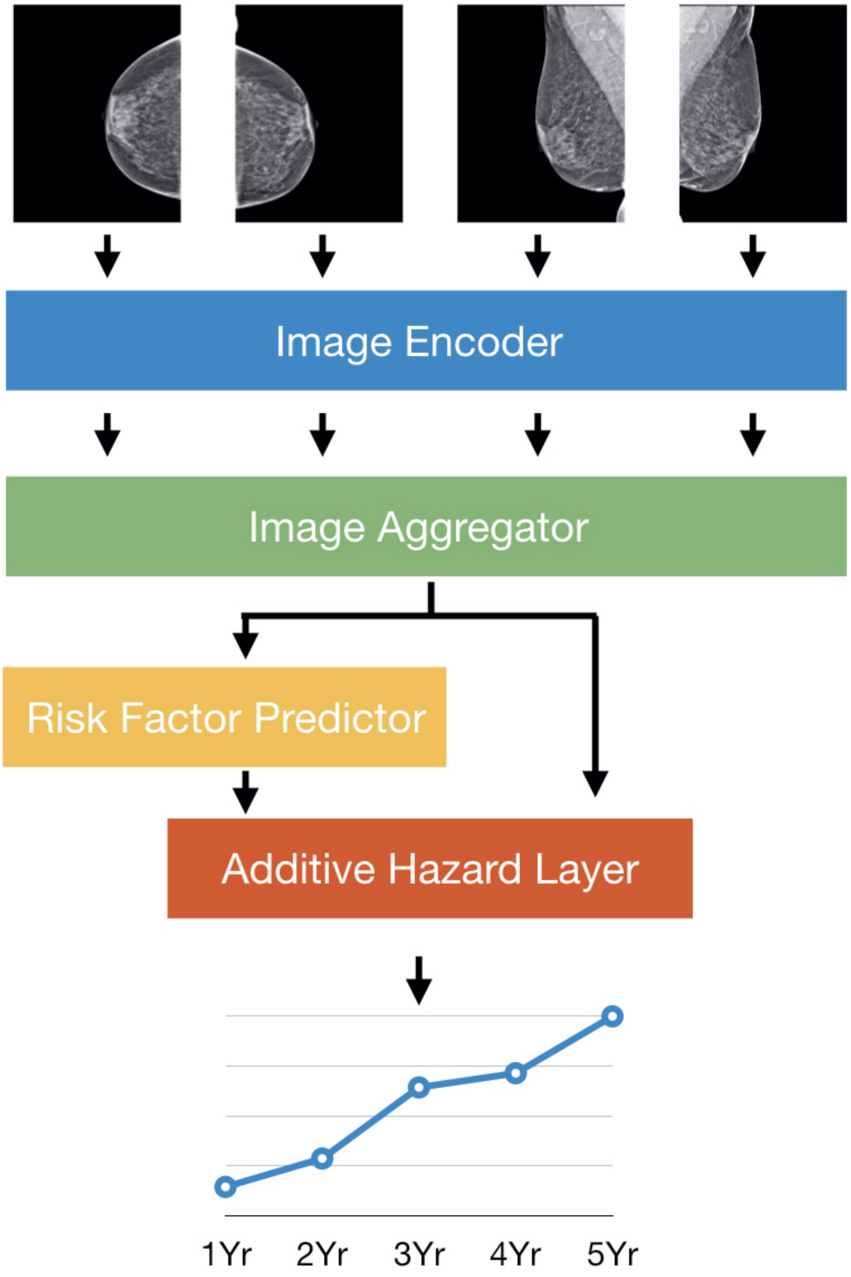

Figure 2-1: Schematic description of Mirai. The four standard views of an individual mammogram were fed into Mirai. The image encoder mapped each view to a vector. The image aggregator combined the four view vectors into a single vector for the mammogram. In this work, we used a single shared ResNet-18 as an image encoder, and a transformer as our image aggregator. The risk factor predictor module predicted all the risk factors used in the Tyrer-Cuzick model, including age, detailed family history and hormonal factors, from the mammogram vector. The additive hazard layer combined information from both the image aggregator and risk factors (predicted or given) to predict coherent risk assessments across five years.

such information. We note that Hybrid DL and Image-Only DL were both developed using the same MGH dataset as Mirai, and so, differences in performance can only be attributed to the algorithm design. Image-Only DL is equivalent to the image encoder component of Mirai trained by itself as a 5-year risk classifier. TCv8 is a traditional risk model that combines a variety of risk factors including age, family history, and hormonal factors and is a current clinical standard. We obtained TCv8 risk assessments using the Command-Line version of the IBIS Breast Cancer Risk Evaluation tool (version 8).

To better investigate the connection between risk estimation and cancer detection, we also compared Mirai with retrospective radiologist BI-RADS (Breast Imaging-Reporting and Data System) assessments and a recently proposed cancer detection model, Image-and-Heatmaps [135], on the MGH test set. Image-and-Heatmaps is a convolutional neural network trained on a large dataset from New York University (NYU) using both pixel-level and whole-image annotations to predict cancer within 120 days. We obtained Image-and-Heatmaps cancer predictions using their publicly available GitHub [45] and did not use test-time data augmentations or model ensembling.

On the 25,855 examinations (588 positive) in the MGH test set, Mirai with and without risk factors obtained C-indices of 0.76 (0.74 to 0.80) and 0.75 (0.72 to 0.78) compared with C-indices of 0.72 (0.69 to 0.75), 0.72 (0.69 to 0.75), and 0.64 (0.60 to 0.67) by Hybrid DL, Image-Only DL, and TCv8, respectively. The full results on the MGH dataset are summarized in Table 2.1, and receiver operating characteristic (ROC) curves for each time point are shown in Fig. 2-2. Mirai with risk factors had a significantly higher 5-year AUC than Hybrid DL, Image-Only DL, and TCv8 with P values of <0.001, <0.001, and <0.001, respectively. Mirai with risk factors did not have a significantly higher 5-year AUC than Mirai without risk factors (P = 0.27). We also present an analysis of model performance excluding cancers identified within 6 months of the screening mammogram, resulting in 25,708 examinations (441 positive) (table B.1). In this setting, Mirai with risk factors had a significantly higher 5-year AUC than Hybrid DL, Image-Only DL, and TCv8, with P values of <0.001, 0.02, and

<0.001, respectively, and did not have a significantly higher 5-year AUC than Mirai without risk factors (P = 0.27). We also evaluated the performance of radiologist BI-RADS assessments and Image-and-Heatmaps [135] in Table 2.1. Radiologists obtained ROC AUCs of 0.92 (0.90 to 0.95) and 0.75 (0.72 to 0.78) at 1 and 2 years, respectively, compared with 0.84 (0.81 to 0.88) and 0.80 (0.76 to 0.83) by Mirai. We found that Image-and-Heatmaps obtained a 1-year AUC of 0.78 (0.73 to 0.82) and a C-index of 0.68 (0.65 to 0.72).

We performed an ablation study of Mirai to investigate the effects of different design choices on overall performance and mammography device bias (table B.10 and fig. C-2). To evaluate the mammography device bias of a risk model, we trained a classifier to predict which machine was used to acquire a mammogram from the model's corresponding risk assessment and measured the AUC of this device-identity classifier on the MGH test set. We found that an ablation of Mirai without risk factors that removed conditional adversarial training obtained a device-identity AUC of 0.76 (0.75, 0.76), reflecting large device bias. With the addition of conditional adversarial training, Mirai without risk factors obtained a device-identity AUC of 0.50 (0.50, 0.50), effectively removing the bias. We evaluated the saliency of each risk factor in Mirai's predictions across the MGH test set in fig. C-4. The most important risk factors were a patient's BRCA status, if they had any family history (binary family history), and if they had had any children (parous), with average saliency scores of 0.07 (0.07, 0.07), 0.04 (0.04,0.04), and 0.03 (0.03, 0.03), respectively. In contrast, mammograms had an average saliency score of 2.19 (2.17, 2.22). We note that the mammogram obtained a 30-fold higher saliency score than the most important clinical factor, BRCA status. This finding is consistent both with the reported performance of Mirai with and without risk factors shown in Table 1 and the result that Mirai with risk factors did not obtain a significantly higher five-year AUC than Mirai without risk factors (p=0.27).

Model	Use RF	C-Index	1-Year AUC	2-Year AUC	3-Year AUC	4-Year AUC	5-Year AUC
TCv8	No	0.64 (0.60, 0.67)	0.66 (0.61, 0.71)	0.65 (0.61, 0.69)	0.64 (0.60, 0.68)	0.63 (0.59, 0.67)	0.62 (0.59, 0.66)
Radiologist BI-RADs	NA	0.67 (0.65, 0.70)	0.92 (0.90, 0.95)	0.75 (0.72, 0.78)	0.68 (0.65, 0.70)	0.64 (0.62, 0.67)	0.62 (0.60, 0.65)
Image-and-Heatmaps	No	0.68 (0.65, 0.72)	0.78 (0.73, 0.82)	0.73 (0.70, 0.77)	0.69 (0.66, 0.73)	0.67 (0.63, 0.70)	0.64 (0.60, 0.68)
Image-Only DL	No	0.72 (0.69, 0.75)	0.79 (0.75, 0.83)	0.75 (0.71, 0.78)	0.73 (0.70, 0.77)	0.73 (0.70, 0.76)	0.73 (0.70, 0.77)
Hybrid DL)	Yes	0.72 (0.69, 0.75)	0.78 (0.75, 0.82)	0.74 (0.71, 0.78)	0.72 (0.68, 0.75)	0.72 (0.68, 0.75)	0.72 (0.69, 0.76)
Mirai	No	0.75 (0.72, 0.78)	0.84 (0.80, 0.87)	0.78 (0.75, 0.82)	0.77 (0.74, 0.80)	0.76 (0.73, 0.79)	0.76 (0.73, 0.79)

Table 2.1: ROC AUCs and C-indices for Mirai and prior risk models on the MGH test set. We also evaluated Image-And-Heatmaps and radiologist BI-RADs assessments. RF refers to "risk factors". All metrics are followed by their 95% confidence interval.

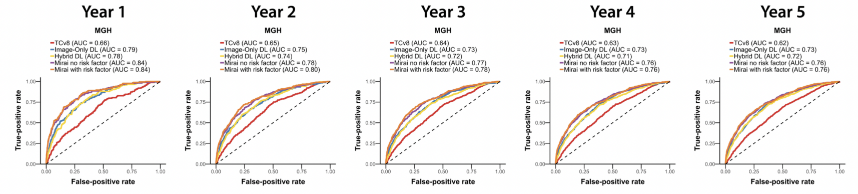

Figure 2-2: ROCs for model predictions on MGH test set.

2.2.3 Generalization to additional populations

For Mirai to be useful to the larger community, it must be validated in a diverse set of clinical environments and patient populations. To this end, we tested the model on a dataset from the Novant, Emory, Maccabi-Assuta, Karolinska, Chang Gung Memorial Hospital (CGMH) and Barretos which consisted of 14,157, 44,008, 6,189, 19,328, 13,356, and 5,900 examinations from 5,887, 16,495, 6,189, 7,353, 13,356, and 5,900 patients of which 235, 1,003, 186, 1,413, 244, and 146 examinations were followed by cancer within 5 years, respectively. Demographics of each dataset are shown in tables 2.2, B.2, B.3, B.4. A dataset construction flowchart for all datasets is shown in Fig. 2-3. Traditional risk factors were not available in either dataset. As a result, we tested Mirai without risk factors.

The performance of Mirai across all time-points and across all test sets is reported in Table 2.3. Mirai (without risk factors) performed similarly across all test sets, obtaining Uno's C-indices of 0.75 (0.70 to 0.80), 0.77 (0.75 to 0.79), 0.77 (0.73 to 0.81), 0.81 (0.79 to 0.82), 0.79 (0.76 to 0.83) and 0.84 (0.81 to 0.88) on the Novant, Emory, Maccabi-Assuta, Karolinska, CGMH and Barretos test sets, respectively. These results are similar to C-index of 0.75 (0.72 to 0.78) at MGH. Mirai obtained 1-year AUCs of 0.84 (0.80 to 0.87), 0.78 (0.73 to 0.84), 0.83 (0.81 to 0.86), 0.86 (0.81 to 0.91), 0.90

	MGH	Novant	Emory	Maccabi-Assuta	Karolinska	CGMH	Barretos
Unique patients	7,005 (233)	5887 (123)	16,495 (495)	6189 (186)	7,353 (799)	13,356 (244)	5,900 (146)
All exams	25855 (588)	14157 (235)	44008 (1003)	6189 (186)	19328 (1413)	13356 (244)	5900 (146)
Age at Exam							
<40	724 (7)	0 (0)	410 (11)	23 (1)	0 (0)	0 (0)	0 (0)
40-50	7025 (95)	3917 (53)	9047 (147)	1589 (47)	7814 (364)	4008 (74)	2810 (41)
50-60	7829 (188)	5368 (65)	12113 (235)	1232 (23)	5477 (387)	6301 (115)	2114 (59)
60-70	6708 (182)	4872 (117)	13182 (302)	2232 (57)	5174 (563)	3024 (55)	976 (46)
70-80	3001 (94)	0 (0)	7638 (285)	1038 (44)	863 (99)	0 (0)	0 (0)
80<	568 (22)	0 (0)	1495 (23)	75 (14)	0 (0)	0 (0)	0 (0)

Table 2.2: Demographics of Massachusetts General Hospital (MGH), Novant, Emory, Maccabi-Assuta, Karolinska, Chang Gung Memorial Hospital (CGMH), and Barretos test sets. Patient statistics are followed by the number of patients who were diagnosed with breast cancer within five years. Exam level statistics, including age demographics, are followed by the number of exams which were followed by a cancer diagnosis within five years.

Site	C-Index	1-Year AUC	2-Year AUC	3-Year AUC	4-Year AUC	5-Year AUC
MGH, USA	0.75 (0.72, 0.78)	0.84 (0.80, 0.87)	0.78 (0.75, 0.82)	0.77 (0.74, 0.80)	0.76 (0.73, 0.79)	0.76 (0.73, 0.79)
Novant, USA	0.75 (0.70, 0.80)	0.78 (0.73, 0.84)	0.76 (0.71, 0.81)	0.76 (0.71, 0.81)	0.75 (0.70, 0.80)	0.75 (0.70, 0.80)
Emory, USA	0.77 (0.75, 0.79)	0.83 (0.81, 0.86)	0.79 (0.77, 0.82)	0.77 (0.75, 0.80)	0.77 (0.75, 0.79)	0.76 (0.74, 0.79)
Maccabi-Assuta, Israel	0.77 (0.73, 0.81)	0.86 (0.81, 0.91)	0.81 (0.76, 0.87)	0.79 (0.75, 0.84)	0.77 (0.73, 0.81)	0.75 (0.71, 0.79)
Karolinska, Sweden	0.81 (0.79, 0.82)	0.90 (0.89, 0.92)	0.86 (0.84, 0.88)	0.82 (0.80, 0.84)	0.80 (0.79, 0.82)	0.78 (0.76, 0.80)
CGMH, Taiwan	0.79 (0.76, 0.83)	0.90 (0.87, 0.93)	0.86 (0.83, 0.90)	0.82 (0.78, 0.85)	0.80 (0.77, 0.84)	0.79 (0.75, 0.82)
Barretos, Brazil	0.84 (0.81, 0.88)	0.89 (0.86, 0.93)	0.87 (0.84, 0.91)	0.86 (0.83, 0.90)	0.85 (0.81, 0.89)	0.82 (0.78, 0.86)

Table 2.3: Area under the Receiver Operating Curve (AUCs) for predicting cancer in one to five years and Uno's C-index for Mirai on all test sets. All metrics are followed by their 95% confidence interval.

(0.89 to 0.92), 0.90 (0.87 to 0.93), and 0.89 (0.86 to 0.93) at MGH, Novant, Emory, Maccabi-Assuta, Karolinska, CGMH, and Barretos, respectively. Mirai obtained one higher 1-year AUC at Karolinska (0.90), CGMH (0.90), and Barretos (0.89), where screening is biennial, than at MGH (0.84), Novant (0.78), Emory (0.83), and Maccabi-Assuta (0.86), where screening is annual. The performance of Mirai when excluding cancers diagnosed within 6 months is shown in Table B.5. Here, Mirai obtained C-indices of 0.69 (0.66 to 0.73), 0.72 (0.66 to 0.79), 0.69 (0.66 to 0.72), 0.70 (0.64 to 0.76), 0.71 (0.69 to 0.74), 0.70 (0.66 to 0.75), and 0.78 (0.74 to 0.83) on the MGH, Novant, Emory,Maccabi-Assuta, Karolinska, CGMH, and Barretos test sets, respectively, compared with a C-index of 0.62 (0.58 to 0.67) obtained by TC on the MGH test set.

2.2.4 Identifying high-risk cohorts

To evaluate the clinical significance of Mirai's performance, we evaluated its ability to identify high-risk cohorts that may benefit from supplemental screening. To perform this analysis, we restricted our attention to patients who were initially screening negative and had at least 5 years of screening follow-up. We defined an examination as screening negative if it was not followed by a cancer diagnosis within 6 months. This resulted in cohorts of 9,284, 7,524, 8,640, 1,385, 7,194, 11,167, and 2,057 examinations from 3,957, 3,617, 5,774, 1,385, 5, 707, 11,167, and 2,057 patients of which 441, 140, 632, 107, 869, 139, and 70 were followed by cancer within 5 years from MGH, Novant, Emory, Maccabi-Assuta, Karolinska, CGMH, and Barretos, respectively. We defined the sensitivity of a guideline as the percentage of all patients who would develop cancer within 5 years included within the high-risk cohort and thus may benefit from supplemental screening. We defined the specificity of the guidelines as the percentage of all patients who do not develop cancer within 5 years not included in the high-risk cohort and thus may avoid overtreatment. We compared three guidelines for identifying high-risk patients: 20% lifetime risk by TC (TC guideline), Mirai at the specificity of the TC guideline, and Mirai at the sensitivity of the TC guideline. We studied Mirai at TC specificity and TC sensitivity to evaluate the potential of Mirai to improve early detection for a fixed cost (ie, specificity) and the potential to reduce costs for a fixed level of early detection (ie, sensitivity), respectively. The Mirai at TC specificity and Mirai at TC sensitivity guidelines were chosen to match the specificity and sensitivity of the TC guideline on the MGH development set. We only evaluated the TC model on the MGH test set as the necessary risk factors were not available at the other six institutions. We performed this analysis on all test sets and subgroups of the Emory data set by race. To illustrate the full spectrum of possible operating points for this use case, we also plot receiver operating curves for Mirai for each institution.

The performance of Mirai in selecting high risk cohorts across all tests sets is shown in Table 2.4. On the MGH test set, the Mirai at TC specificity guideline obtained a sensitivity of 39.7% (32.9 to 46.5) compared with a sensitivity of 22.9%(15.9 to

Method	Sensitivity	Specificity
MGH, USA: 9,284 exams from 3,957 patients. 441 exams followed by future cancer.		
Tyrer-Cuzick (TC) Lifetime Risk >20% [4]	22.9% (15.9, 29.6)	85.4% (84.1, 86.6)
Mirai at TC specificity [4]	39.7% (32.9, 46.5)	85.2% (84.1, 86.4)
Mirai at TC sensitivity	20.0% (14.2, 25.1)	94.2% (93.4, 94.9)
Novant, USA: 7,524 exams from 3,617 patients. 140 exams followed by future cancer.		
Mirai at TC specificity	50.0% (38.5, 61.4)	84.6% (83.3, 85.7)
Mirai at TC sensitivity	23.6% (14.1, 32.1)	95.4% (94.7, 96.0)
Emory, USA: 8,640 exams from 5,774 patients. 632 exams followed by future cancer.		
Mirai at TC specificity	36.7% (31.6, 41.8)	84.9% (84.0, 85.9)
Mirai at TC sensitivity	22.0% (17.4, 26.3)	91.5% (90.7, 92.2)
Maccabi-Assuta, Israel: 1,385 exams from 1,385 patients. 107 exams followed by future cancer		
Mirai at TC specificity	40.2% (30.9, 49.3)	84.8% (82.9, 86.8)
Mirai at TC sensitivity	22.4% (14.6, 30.2)	92.5% (91.1, 94.0)
Karolinska, Sweden: 7,194 exams from 5,707 patients. 869 exams followed by future cancer		
Mirai at TC specificity	42.9% (38.5, 47.0)	85.0% (84.0, 86.0)
Mirai at TC sensitivity	21.9% (18.4, 25.2)	94.3% (93.7, 95.0)
CGMH, Taiwan: 11,167 exams from 11,167 patients. 139 exams followed by future cancer		
Mirai at TC specificity	45.3% (36.7, 53.5)	84.6 (83.9, 85.2)
Mirai at TC sensitivity	23.0% (15.8, 29.8)	94.8% (94.4, 95.2)
Barretos, Brazil: 2,057 exams from 2,057 patients. 70 exams followed by future cancer.		
Mirai at TC specificity	37.1% (25.6, 48.1)	85.2% (83.6, 86.8)
Mirai at TC sensitivity	21.4% (11.1, 30.4)	91.7% (90.51, 92.92)

Table 2.4: High risk cohort analysis for all test sets. For each test set, we restricted our analysis to patients who were initially screening negative and had at least five-years of screening follow-up. We defined an exam as screening negative if it was not followed by a cancer diagnosis within six months. We defined a future c¬ancer as a pathology confirmed breast cancer diagnosis within five years of the mammogram. Mirai thresholds (i.e. "at TC sensitivity" and "at TC specificity") were chosen to match the performance of the Tyrer-Cuzick (TC) model on the development MGH set.

29.6) obtained by TC, yielding a significant improvement (P < .001). The Mirai at TC sensitivity obtained a specificity of 94.2% (93.4 to 94.9) compared with 85.4% (84.1 to 86.6) obtained by TC, yielding a significant improvement (P < .001). This performance was maintained across our other institutions. The Mirai at TC specificity guideline obtained sensitivities of 50.0% (38.5 to 61.4), 36.7% (31.6 to 41.8), 40.2% (30.9 to 49.3), 42.9% (38.5 to 47.0), 45.3% (36.7 to 53.5), and 37.1% (25.6 to 48.1) at Novant, Emory, Maccabi-Assuta, Karolinska, CGMH, and Barretos, respectively. The Mirai at TC sensitivity guideline obtained specificities of 95.4% (94.7 to 96.0), 91.5% (90.7 to 92.2), 92.5% (91.1 to 94.0), 94.3% (93.7 to 95.0), 94.8% (94.4 to 95.2), and 91.7% (90.51 to 92.92) at Novant, Emory, Maccabi-Assuta, Karolinska, CGMH, and Barretos, respectively. The Mirai receiver operating curves for selecting high-risk cohorts across all test sets are shown in Figure C-1.

2.2.5 Subgroup analysis

We also validated all risk models for different clinical subgroups of interest. In the MGH test set, we computed model C-indices for patients of different races (White, African American, and Asian American), different age groups, different density categories, and different mammography devices. We found that Mirai performed similarly across all groups. This information is available in table B.6. We note that the C-indices for Mirai with risk factors for White, Asian American, and African American patients were 0.75 (0.72 to 0.78), 0.80 (0.68 to 0.95), and 0.71 (0.55 to 0.90), respectively, compared with 0.64 (0.60 to 0.68), 0.54 (0.36 to 0.75), and 0.62 (0.44 to 0.84) for TCv8. In the Karolinska dataset, we computed Mirai C-indices by future cancer subtype (invasive, HER2 status, and so on) in table B.7. The distribution of cancer subtypes is reported in table B.8. We found that Mirai obtained similar C-indices across different subtypes, which is further supported by a t-SNE (t-distributed stochastic neighbor embedding) [74] analysis (Fig. C-3) showing that the model learns similar representations for mammograms regardless of the subtype of the future cancer.

As shown in Table B.9, we found that Mirai performed similarly across different race subgroups of the Emory test set. The Mirai at TC specificity guideline obtained sensitivities of 33.9% (26.3 to 41.0) and 40.0% (32.0 to 47.2) for African-American and White patients, respectively. The Mirai at TC sensitivity guideline obtained specificities of 90.7% (89.6 to 91.9) and 91.9% (90.8 to 93.0) for African-American and White patients, respectively.

2.3 Discussion

We developed a risk model, Mirai, to assess breast cancer risk from screening mammograms. Mirai demonstrated improved discriminatory capacity over the state-of-the-art clinically adopted Tyrer-Cuzick and prior deep learning approaches Hybrid DL and Image-Only DL. Moreover, we found that Mirai, which was trained at MGH, maintained its performance on datasets on test sets from seven hospitals across five countries. Externally validating our model across diverse clinical settings is especially important given recent negative findings for the generalization of other proposed mammography-based models for cancer risk [128]. We evaluated Mirai across races, ages, and breast density categories in the MGH test set, across races on the Emory dataset and across cancer subtypes on the Karolinska dataset and found that it performed similarly across all subgroups. We also demonstrated how Mirai could be implemented in current clinical pipelines focused on identifying high-risk patients and showed that it improved over existing risk models such as Tyrer-Cuzick lifetime risk.

Accurate short-term risk prediction (ie, within 5 years) is essential for early detection efforts in breast cancer. Traditional risk models, such as the TC model, are already widely implemented and support existing supplemental screening guidelines by the American Cancer Society, the American College of Radiology, and the National Comprehensive Cancer Network [110, 12, 83, 84, 9]. However, these models only provide a global risk prediction for large groups of patients, limiting their predictive accuracy for individuals and for specific time frames. Moreover, current guidelines for MRI eligibility[110, 12] leverage lifetime TC risk, which ignores a patient's short-term risk of breast cancer and further limits the model's predictive utility. Our retrospective analysis across multiple test sets suggests that Mirai has the potential to replace current risk models (eg, TC) in guidelines for MRI screening, improving early detection and reducing overtreatment. For instance, we found that Mirai could obtain 70% relative improvement in sensitivity over the TC-based guideline at MGH while maintaining the same specificity.

The performance of Mirai can be attributed to how its design captures unique characteristics of breast cancer risk estimation. Specifically, the model architecture jointly reasons over both different views of the mammogram and multiple time points of risk assessment. Moreover, we demonstrated how to incorporate nonimage risk factors such as age or hormonal factors to further refine accuracy, while enabling the model to impute this information if it is not provided. Last, we used a conditional adversarial training regime to learn image representations that are device invariant. Our work is also related to the large volume of work [135, 45, 72, 133, 106, 148, 4, 134, 99, 97, 41, 107, 64, 79] focused on developing deep learning models for breast cancer detection. Although the tasks of cancer detection and future cancer risk are distinct, we hypothesize that some of the technical lessons from the two tasks can be complementary. For instance, we hypothesize that aggressive model ensembling strategies used by [135, 79, 104] and the use of detailed cancer region annotations could be used to improve image-based risk models. Moreover, we hypothesize that our mechanisms for predicting risk at multiple time points, optionally using risk factors, and learning representations that are invariant to mammography machines could be used to improve the current state of the art in cancer detection systems.

Although Mirai can be tested as a cancer detection system, direct comparison to prior work in cancer detection is difficult due to a lack of publicly available code [79, 62] and the lack of common benchmarks. Not directly comparable, we note that Mirai obtained a 1-year AUC of 0.90 on the Karolinska test set, similar to the top single-model AUC 0.90 on a separate Karolinska test set reported by [104]. We also evaluated Image-and-Heatmaps [135], a recently proposed cancer detection model trained to predict cancer within 120 days, on a large dataset from NYU. Image-and-Heatmaps obtained a 120-day AUC of 0.89 on the NYU test set [135], and it obtained a 1-year AUC of 0.78 on the MGH test set. We note that it is difficult to compare this model with our own because of the difference in study objectives and training datasets. These results further highlight the importance of creating common benchmarks with standardized evaluation to enable direct comparison between models. We believe that sharing trained models is important for the continued development of cancer detection and risk assessment systems, and to this end, we are releasing our code and models for public research use.

There are multiple directions for future work that can further improve the accuracy and utilization of the imaging-based models for cancer risk. Although our model only considers a patient's current mammogram agnostic of previous imaging, it is known that changes in imaging over time contain a wealth of information. A natural next step is to develop methods that can effectively use a patient's full history of imaging. In a similar fashion, expanding the model to use tomosynthesis is likely to yield further performance improvements. Beyond work in improving accuracy, additional research is required to determine how to adapt image-based risk models to different mammography devices across multiple vendors. Although our conditional adversarial training scheme enabled us to obtain consistent risk assessments across mammography devices where we have training data, we did not evaluate whether our models can generalize to unseen mammography devices. In addition, although our own evaluation focused on defining high-risk cohorts, other methods are required to design more fine-grained risk-based guidelines.

Our study had limitations. Our analysis of the benefit of different screening guidelines was retrospective. Prospective clinical trials are needed to confirm the clinical benefit of identifying improved high-risk cohorts using Mirai and to establish Mirai guidelines. Moreover, Mirai was only developed and tested using Hologic mammograms. Future work will be needed to test and adapt this technology to more mammography vendors and to tomosynthesis images. Moreover, although Mirai provides a risk assessment for cancer in either breast, it does not provide a risk estimate for each breast.

In conclusion, Mirai, a mammography-based risk model, maintained its accuracy across globally diverse test sets from MGH, USA; Novant, USA; Emory, USA; Maccabi-Assuta, Israel; Karolinska, Sweden; CGMH, Taiwan; and Barretos, Brazil. Moreover, guidelines based on Mirai significantly outperformed the existing clinical guidelines based on the TC model at MGH and maintained their performance across all test sets. This is the broadest validation to date of an AI-based breast cancer model and demonstrates that the technology can offer broad and equitable improvements in care. Prospective clinical trials of this technology are warranted.

2.4 Methods

2.4.1 Study Design

The primary objectives of this study were to develop a model to assess breast cancer risk and to validate its performance across diverse populations and clinical settings. We designed and benchmarked our algorithm, Mirai, against the Tyrer-Cuzick model and other deep learning models trained on the same MGH dataset, namely, Image-Only DL and Hybrid DL, in predicting future risk. Although Mirai was trained to predict both first-time cancer cases and recurrences, we limited our analysis to patients without a prior history of breast cancer to enable a fair comparison against the Tyrer-Cuzick model. Our secondary objective was to demonstrate the ability of Mirai to identify high-risk cohorts and to compare it with the current Tyrer-Cuzick lifetime risk based guideline.

2.4.2 Description of Cohorts

Our retrospective study was approved by the institutional review board of each clinical institution with a waiver for written informed consent and was compliant with the Health Insurance Portability and Accountability Act. We collected data sets from Massachusetts General Hospital (MGH), USA; Novant, USA; Emory, USA; Maccabi-Assuta, Israel; Karolinska, Sweden; Chang Gung Memorial Hospital (CGMH), Taiwan; and Barretos, Brazil. Across all data sets, we collected mammograms from a large subset of patients and leveraged the mammograms to obtain Mirai risk assessments. Mirai was trained using Hologic images, and all mammograms included in this study were taken using a Hologic machine.

To collect the MGH data set, we collected consecutive screening mammograms from 80,134 patients screened between January 1, 2009, and December 31, 2016, at MGH. We obtained outcomes through linkage to a local five-hospital registry in the Massachusetts General Brigham healthcare system, alongside pathology findings from MGH's mammography electronic medical record. We excluded patients without at least 1 year of screening followup, who were diagnosed with other cancers (eg, sarcoma) in the breast or did not have all four views available, to identify 70,972 patients. Patients were randomly split into n = 56,786 for training, n = 7020 for development, and n = 7166 for testing. To enable fair comparison against the Tyrer-Cuzick model, we excluded 161 patients with prior history of breast cancer from the test set, leaving 7005 patients. Because each patient had multiple examinations, this resulted in 210,819, 25,644, and 25,855 examinations for training, development, and testing, respectively.

To collect the Novant data set, we selected 7,238 patients randomly from the cohort of all patients age 40-69 years screened at a Novant Health clinic between January 1, 2012, and December 31, 2016. We included all mammograms across this time period and obtained outcomes by querying both a local cancer registry and the Novant electronic medical record.We excluded patient examinations that did not have at least 1 year of screening follow-up with prior cancer or whose mammogram did not include all four standard views to identify 14,157 examinations from 5,887 patients.

To collect the Emory data set, we extracted 8 years of mammograms from an institutional database of all comers for screening mammography from 2013 to 2020 and randomly selected 30% of women from this database, totaling 75,010 examinations from 28,994 patients. We collected outcomes from pathology findings from Emory's institutional database using Magview software (Fulton, MD). As with other data sets, we excluded patients' examinations that did not have at least 1 year of screening follow-up, with prior cancer or whose mammogram did not contain all four standard views to identify 44,008 examinations from 16,495 patients.

To collect the Maccabi-Assuta data set, we selected all comers for screening mammography at Maccabi-Assuta during 2015 age 30 years or older, resulting in 9,775 examinations from 9,775 women. For each patient, we obtained dates of first breast cancer diagnosis from the Maccabi-Assuta electronic medical records and a regional registry. We excluded examinations from non-Hologic machines and patients with a history of breast cancer to identify 6,189 examinations from 6,189 patients.

The Karolinska data set was extracted from the Cohort of Screen-Aged Women[35]. All women age 40-74 years within the Karolinska University uptake area who had attended screening and were diagnosed with breast cancer, without implants and without prior breast cancer, from 2008 to 2016 were included, as well as a random sample of controls with at least 2 years of follow-up, from the same time period. The full Karolinska case-control data set included 11,303 women, and 70% of both cases and controls were randomly selected for inclusion in this study, resulting in 19,328 examinations from 7,353 patients.

To collect the CGMH data set, which consisted of 13,356 examinations from 13,356 patients, we selected random women undergoing screening mammography there between 2010 and 2011 who were age 45-70 years. Following local guidelines, we also included women age 40-44 years who had a family history of breast cancer. Cancer outcomes were obtained from the national cancer registry.

To collect the Barretos test set, we selected all women age 40 to 69 years who received screening mammograms at the Fernanopolis and Campo Grande units from January 2, 2014, to June 30, 2015, to obtain a cohort of 6,206 mammograms from 6,206 patients. Cancer outcomes were obtained from patient medical records at Barretos Cancer Hospital. We excluded mammograms without all four standard views, with prior cancer, and with insufficient follow-up to identify 5,900 examinations from 5,900 patients.

Across all data sets, we defined a cancer-positive outcome as a pathology-confirmed diagnosis of either invasive breast carcinoma or ductal carcinoma in situ. We used screening follow-up to define when patients were cancer-negative. For instance, we considered a patient negative for 3 years if they had screening follow-up for at least 3 years without a cancer diagnosis. For all data sets, except the CGMH data set, we excluded patients with prior cancer to enable fair comparison against the TC model, which does not assess risk for this population. We did not perform this exclusion for the CGMH data set because of difficulties in manual data curation.

2.4.3 Image Preprocessing

All of the mammograms used in this study were captured using either the Hologic Selenia or Selenia Dimensions mammography devices. We converted presentation view dicoms to PNG16 files using the DCMTK library. We used the dcmj2pnm program (v3.6.1, 2015) with +on2 and– min-max-window flags. We used torchvision (version 0.2.1) and Pillow (version 5.2.0) python libraries for image preprocessing and data augmentations. First, we resized each mammogram view to 1664 by 2048 pixels. Following standard practice [48], we normalized our images to have zero mean and unit variance. To this end, we calculated the pixel mean and standard deviation across the training set and normalized each image by this mean and standard deviation before feeding it into the model. We used the training set image mean and standard deviation for all images.

2.4.4 Architecture Details

We encoded each view of the mammogram independently using ResNet-18 [53], with a global max pooling layer at the end, to compress the image representation to a 512-dimensional vector, x. We refer to this as our Image Encoder. We note that this is akin to the Image-Only model from [143]. To aggregate the information from different views, we took the image representation from each view, and conditioned it on a learned view and laterality embedding, to obtain view-specific representations. To condition a vector x by an embedding e, we used the following expression:

$$h = (W_{scale}e) \times x + (W_{shift}e)$$

We then took these view-specific representations and passed them into a Transformer network [90] with attention-pooling to obtain a 512-dimensional mammogram level representation. We refer to this component as our Image Aggregator.

Given the mammogram-level representation, we trained the model to independently predict each risk factor as used in TCv8. We minimized the combined cross-entropy loss of predicting each risk factor, weighted by a hyperparameter lambda, and the log-likelihood loss of predicting future cancer. We note that the risk factor prediction module can be thought of as a generative model that uses the mammogram to impute missing risk factors, and thus allows the model to be run using the mammogram alone. We refer to this component as our Risk Factor Predictor.

The additive-hazard layer first took in a patient's features, m, from the mammogram representation and the traditional risk factors (predicted or given), and predicted a patient's baseline risk, $B(m)$ using a small network (in our case a linear layer). To predict risk at k years away from the mammogram, it separately predicted the positive 0-1 year marginal hazard (i.e., the additional risk of getting cancer in the next year) using network H_0, and the 1-2 year hazard using network H_1, etc. Each marginal hazard network, e.g H_1, is implemented as a linear layer followed by a ReLU. To obtain the overall risk at year k, the additive-hazard layer summed the baseline risk and the marginal hazards up to year k. This is summarized in equation 1, where $P(Y = 1, T = k|m)$ refers to a patient being diagnosed with cancer within k years. We note that this modeling objective follows seminal work [1] in linear additive-hazard survival models.

$$P(t_{cancer}) = k|m) = B(m) + \Sigma H_i(m)$$

The architecture of our additive-hazard layer ensured that risk predictions were always monotonic (that is, a patients two-year risk is always higher than their one-year risk) and enabled us to easily optimize our model by maximizing the log-likelihood of the observed data in our training set. For patients with less than five years of screening followup, we leveraged their data to supervise the prediction over the years for which we know their outcomes.

The device discriminator took as input the mammogram level representation from the Image Aggregator, as well as the predicted risk over time, and aimed to predict the identity of the device that took the mammogram, Hologic Selenia or Selenia Dimensions. This function is implemented as a two-layer multilayer perceptron with a batch-normalization [56] and ReLU nonlinearities.

2.4.5 Model Training

We trained Mirai in two phases; first, we trained the image encoder in conjunction with the risk factor predictor and additive hazard layer to predict breast cancer independently from each view without using conditional adversarial training. In this stage, we intialialized our image encoder with weights from ImageNet [102], and augmented our training set with random flips and rotations of the original images. We found that adding an adversarial loss at this stage or training the whole architecture end-to-end prevented the model from converging. In the second stage of training, we froze our image encoder, and trained the image aggregation module, the risk factor prediction module, the additive hazard layer, and the device discriminator in a conditional adversarial training regime [153]. We trained our adversary for three steps for every one step of training Mirai. In each stage, we performed small hyperparameter searches and chose the model that obtained the highest C-index on the development set.

2.4.6 Model Calibration

To obtain absolute probabilities of cancer, we utilized the Platt method [93] to calibrate the predicted probabilities of cancer on the development set. We calibrated each year's risk prediction separately. For instance, to calibrate our predictions for 5-year cancer risk, we restricted our calibrator to match the incidence seen for exams with at least five years of followup on the development set.

2.4.7 Saliency Analysis

Saliency scores for the model inputs were calculated with the integrated gradients method [111]. Specifically, the Image Aggregator of Mirai 'with risk factors' was passed the image representation from each view along with the patient risk factors from the MGH test set. The gradient of the 5-year logit score was then computed with respect to each individual input. The integral over the gradients was approximated using 150 steps and a baseline vector of all zeroes. Last, the saliency score was obtained by summing the attributions of each input, averaging over the entire test set, and taking the absolute value of the resulting mean.

2.4.8 Measuring Device Bias

To investigate the impact of different mammography devices on model calibration, we trained a device-identity classifier to recover which device an exam was taken from (Selenia Dimensions vs Lorad Selenia) from the risk assessment alone for each model, and report the ROC AUC of this classifier. Specifically, we trained a logistic regression model on the risk assessments of each model on the MGH validation set, and tested its ability to predict the correct mammography device on the MGH test set. If there exists a systematic bias in risk assessments by mammography device, then the device-identity classifier can leverage this signal to obtain a high AUC on the test set. For models that do not contain any device-related bias in their risk assessments, the device-identity classifier obtains an AUC of 0.50.

Both Hybrid DL [143] and ImageOnly DL [143] formulated five-year cancer risk prediction as a classification task and so they were trained on the 2009-2012 subset of the MGH dataset with five-years of followup. MGH only utilized one mammography machine during this time, Lorad Selenia, and as a result, ImageOnly DL and Hybrid DL did not learn device specific bias. In contrast, all Mirai ablation variants shown in table B.10 were able the full MGH training set because they used a survival formulation of cancer risk (additive hazard or Cox), thus were able to learn device-related bias.

2.4.9 T-SNE Analysis

For all t-SNE analysis, we used the final image hidden representation from Mirai and visualized it in two dimensions with the t-SNE function in sklearn.manifold module of scikit-learn 0.21.3 [91] with default parameters.

2.4.10 Ablation Analysis

To study the effect of our design decisions, we report a detailed ablation study of Mirai's components in table B.10. Moreover, to study the importance of our Additive Hazard formulation, we compare an Image Encoder with our Additive Hazard layer to an Image Encoder trained with a Cox Proportional Hazard's layer. The Cox proportional hazard layer predicted a single relative hazard per patient, analogous to $B(x)$, and this model was optimized to maximize the Cox partial likelihood objective, similar to prior work in deep Cox survival models [61].

2.4.11 Statistical Analysis

We evaluated all models by the AUC for 1- to 5-year outcomes. For instance, to compute the 3-year AUC, we considered a mammogram as positive if it was followed by a cancer diagnosis within 3 years and negative if it had at least 3 years of screening follow-up. We also calculated Uno's C-index[124], which offers a generalized AUC across all time points. To address that patients may have multiple examinations, we used a clustered bootstrap approach with 5000 samples to calculate confidence intervals. To assess the significance of the difference between two AUCs, we used the paired DeLong's test [34] as implemented in the pROC package in R [98]. To assess the significance of the difference between two ratios, we used a two-tailed t test as implemented in R[95]. For both tests, we used a predefined $P < 0.05$ for significance.

2.4.12 Data and materials availability

All code used for training and developing the models is available at learningtocure.csail.mit.edu (DOI: 10.5281/zenodo.4291202). The trained Mirai model is publicly available at github.com/yala/Mirai. All datasets were used under license to the respective hospital system for the current study and are not publicly available.

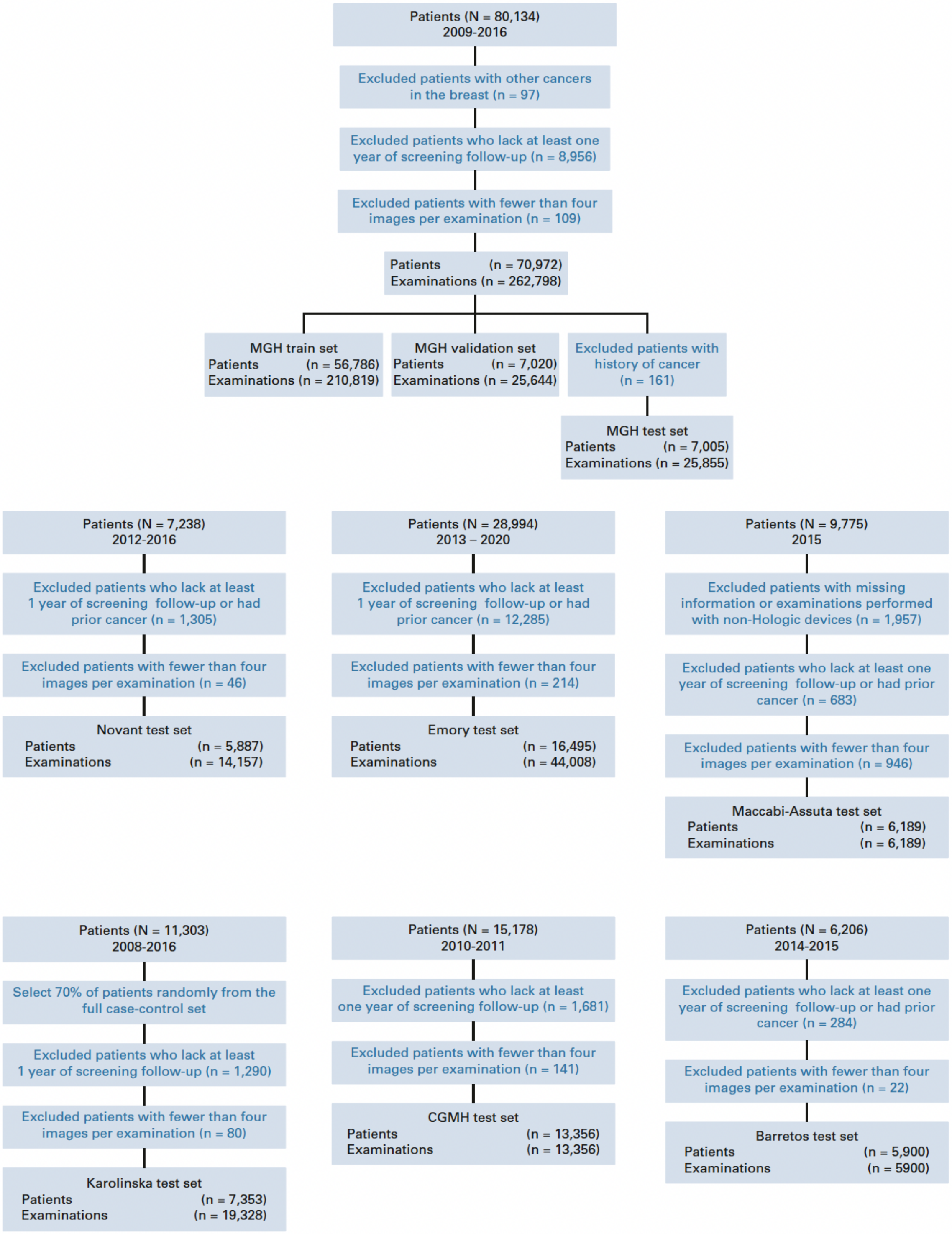
Patients (N = 80,134)
2009-2016
Excluded patients with other cancers in the breast (n = 97)
Excluded patients who lack at least one year of screening follow-up (n = 8,956)
Excluded patients with fewer than four images per examination (n = 109)
Patients (n = 70,972)
Examinations (n = 262,798)
MGH train set
Patients (n = 56,786)
Examinations (n = 210,819)
MGH validation set
Patients (n = 7,020)
Examinations (n = 25,644)
Excluded patients with history of cancer (n = 161)
MGH test set
Patients (n = 7,005)
Examinations (n = 25,855)
Patients (N = 7,238)
2012-2016
Excluded patients who lack at least 1 year of screening follow-up or had prior cancer (n = 1,305)
Excluded patients with fewer than four images per examination (n = 46)
Novant test set
Patients (n = 5,887)
Examinations (n = 14,157)
Patients (N = 28,994)
2013 – 2020
Excluded patients who lack at least 1 year of screening follow-up or had prior cancer (n = 12,285)
Excluded patients with fewer than four images per examination (n = 214)
Emory test set
Patients (n = 16,495)
Examinations (n = 44,008)
Patients (N = 9,775)
2015
Excluded patients with missing information or examinations performed with non-Hologic devices (n = 1,957)
Excluded patients who lack at least one year of screening follow-up or had prior cancer (n = 683)
Excluded patients with fewer than four images per examination (n = 946)
Maccabi-Assuta test set
Patients (n = 6,189)
Examinations (n = 6,189)
Patients (N = 11,303)
2008-2016
Select 70% of patients randomly from the full case-control set
Excluded patients who lack at least 1 year of screening follow-up (n = 1,290)
Excluded patients with fewer than four images per examination (n = 80)
Karolinska test set
Patients (n = 7,353)
Examinations (n = 19,328)
Patients (N = 15,178)
2010-2011
Excluded patients who lack at least one year of screening follow-up (n = 1,681)
Excluded patients with fewer than four images per examination (n = 141)
CGMH test set
Patients (n = 13,356)
Examinations (n = 13,356)
Patients (N = 6,206)
2014-2015
Excluded patients who lack at least one year of screening follow-up or had prior cancer (n = 284)
Excluded patients with fewer than four images per examination (n = 22)
Barretos test set
Patients (n = 5,900)
Examinations (n = 5900)

Figure 2-3: Dataset construction flow-charts

Chapter 3

Optimizing risk-based cancer screening policies with reinforcement learning

3.1 Introduction

For multiple diseases, early detection significantly improves patient outcomes[110, 131]. This motivates considerable investments in population-wide screening programs[30, 32] such as mammography for breast cancer. To be effective and economically viable, these programs must find the right balance between early detection and overscreening. This capacity builds on two complementary technologies: (1) the ability to accurately assess patient risk at a given time point and (2) the ability to design screening regimens based on this risk. With recent advances in deep learning, imaging and genetics, risk assessment technologies are rapidly improving[46, 146, 143]. However, our ability to utilize these predictions to personalize screening regimens lags behind. This deficiency is particularly apparent when the screening system has limited throughput.

In this paper, we focus on the design of screening regimens attuned to the increased capacity of the modern risk assessment models. The need for new methods to personalize screening is motivated by a substantial change in risk assessment algorithms. Traditional risk assessment models rely on a number of categorical variables encoding patient demographics and clinical history combined with traditional statistical models to predict risk[42, 123]. These scores are relatively static throughout a patient's lifetime, with changes typically driven by the patient's age. Moreover, the limited predictive capacity of these risk models restricts the scope of recommendations they support and, consequently, their impact on the screening regimen. Current guidelines divide the population into a few large groups, most often discriminating predicted high-risk patients from the rest, and recommend the same screening frequency to all the members of that cohort [13, 85, 109]. As a result, there remain large opportunities to further personalize care.

The power of novel, AI-driven risk models [143, 146, 35, 73] has given us an opportunity to fundamentally transform population screening. Deep learning algorithms enable these risk models to operate over raw patient data, such as imaging, in addition to traditional expert-specified categorical variables. Moreover, these models can detect highly complex dependencies, which further strengthens their predictive capacity relative to traditional methods. One distinctive feature of these risk models is that their predictions may fluctuate over time as the patient's raw data evolve. This feature suggests that screening regimens need to be flexibly adjusted with changes in risk and optimized over a patient's lifetime. We hypothesize that by pairing AI-based risk models and agile AI-based screening regimens, we can improve early detection while lowering the overall cost of screening. This article presents empirical findings that support this hypothesis in the area of breast cancer screening. The core methodology is applicable to other disease areas and other types of risk models beyond imaging.

3.2 Results

3.2.1 Overview of the algorithm

In computational terms, we can view breast cancer screening as a sequential decision task, where we wish to develop a policy (i.e., screening guideline) that predicts a follow-up recommendation for each patient to maximize their chance for early detection while minimizing screening costs. Intuitively, such a policy should recommend infrequent screenings for low-risk patients while prescribing a higher frequency of screenings for patients at increased risk. The question is how to personalize screening intervals based on a patient's risk profile. More formally, we can cast the screening problem as a Markov decision process, where a patient's state is their risk assessment, the possible actions are different follow-up recommendations (e.g., 6 months or 2 years) and rewards are a combination of expected early detection benefits and screening costs. This formulation enables us to find the best possible policy for this Markov decision process with reinforcement learning (RL) algorithms[100, 112]. RL algorithms train policies (i.e., machine learning models) to make a sequence of decisions that maximize future reward (e.g., early detection benefits) without explicit guidance on the right decision at intermediate steps. Policies are initialized randomly, and through a mix of random exploration and utilization of current knowledge, RL algorithms iteratively improve policies. We show how to leverage RL methods of determining effective cancer screening policies from retrospective screening data.

Applying RL in this context poses a unique challenge, namely the estimation of patient trajectories from retrospective data. The training data pertaining to individual patients only contain information about their risk at the time points when the mammogram was taken. However, to determine whether the algorithm makes the correct recommendation, we need to know the risk assessment at intermediate points. Therefore, we design an algorithm that learns to extrapolate a patient's risk at unobserved time points from the observed screenings. This estimation evolves as new mammograms of the patient become available. With access to these predictions, we can guide our reinforcement learner to adjust its actions according to the estimated risk. Using the retrospective trajectories as our simulation environment, we train screening policies to maximize the future reward given the patient's evolving risk assessments, as illustrated in Fig. 3-1. In doing so, our trained screening policies are specialized to the dynamics and subtleties of the underlying risk model.

Our full framework, named Tempo, is depicted in Fig. 3-2. As described above, we first train a risk progression neural network to predict future risk assessments given previous assessments. This model is then used to estimate patient risk at unobserved time points, and it enables us to simulate risk-based screening policies. Next, we

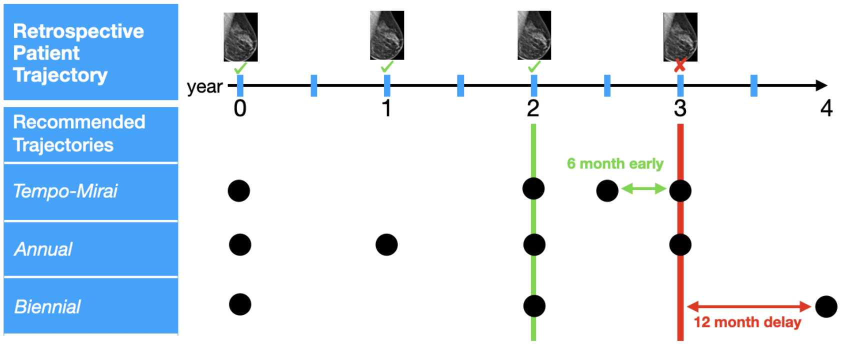

Figure 3-1: Retrospective patient trajectory from MGH test set compared to recommended trajectories by different guidelines. This patient was screened every year, from years zero to year three, and was diagnosed with breast cancer in year three. The red "x" and red line indicate the known time of cancer diagnosis. The green check marks indicate screening negative mammograms, and the green line indicates the last known negative time-point, i.e., year two. For each recommended trajectory, we can compute the screening cost and early detection benefit relative to the historical screening. We measure the early detection benefit of a policy, by comparing it's recommended screening dates to the last known negative date and the known cancer date. In our simulation, Tempo-Mirai, annual screening and biennial screening obtained an early detection benefit of 6.0 months, 0 months and -12.0 months respectively while recommending an average of 1.0, 1.0 and 0.5 mammograms per year for this patient.

train our screening policy, which is implemented as a neural network, to maximize the reward (i.e., a combination of early detection and screening costs) on our retrospective training set. We train our screening policy to support all possible early detection versus screening cost trade-offs using Envelope Q-learning[149], an RL algorithm designed to balance multiple objectives. The input of our screening policies is the patient's risk assessment and desired weighting between rewards (i.e., screening preference). The output of the policy is a recommendation for when to return for the next screen, ranging from 6 months to 3 years in the future, in multiples of 6 months. Our reward balances two contrasting aspects, one reflecting the imaging cost (i.e., the average number of mammograms per year recommended by the policy) and one modeling early detection benefit relative to the retrospective screening trajectory. Our early detection reward measures the time difference in months between each patient's recommended screening date, if it was after their last negative mammogram, and the actual diagnosis date. We evaluate screening policies by simulating the recommendations for held-out

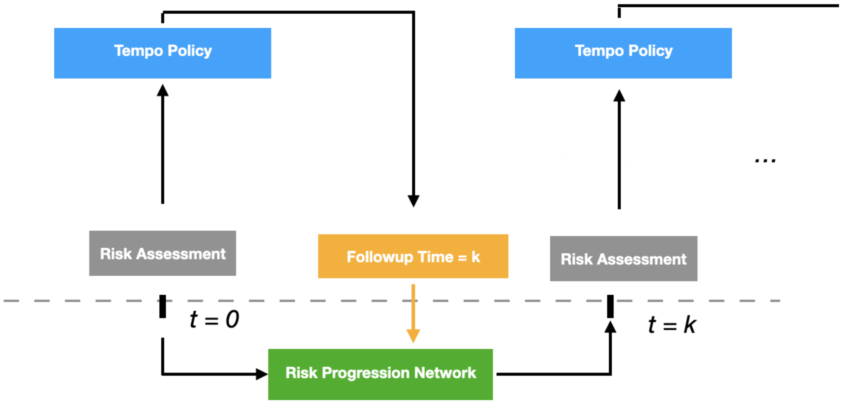

Figure 3-2: Overview of Tempo. Our Tempo Policy takes as input a risk assessment (e.g., from Mirai), and outputs a recommended followup time, such as k years into the future. If a risk assessment is not available at the time step k, we estimate the missing risk assessment using our Risk Progression Network.

patients. The exact reward details and the neural network architectures used are elaborated in Methods.

3.2.2 Experimental Setup

We developed Tempo using the MGH dataset, which consists of 137,682, 16,634 and 17,119 exams from 43,749, 5,399 and 5,525 patients for the training, validation and testing sets, respectively. For each exam, we had access to Mirai12 and Tyrer–Cuzick version 8 (TCv8; [123]) risk assessments. Mirai[146] is a recently proposed mammography-based AI risk model that predicts risk at multiple time points, and TCv8 [123] is a traditional risk model that combines a variety of risk factors, including age, family history and hormonal factors. For Tempo to be broadly applicable, its screening policies must be validated in new clinical environments and patient populations. To this end, we also validated Tempo on representative datasets from Emory (consisting of 22,094 exams from 10,369 patients), Karolinska (consisting of 14,356 exams from 7,191 patients) and CGMH (consisting of 12,280 exams from 12,280 patients). For each exam in the Emory, Karolinska and CGMH datasets, we obtained Mirai risk assessments. We note that the Emory, Karolinska and CGMH datasets

	MGH	Emory	Karolinska	CGMH
All exams	17119 (608)	22030 (723)	14362 (1768)	12280 (235)
Age				
<40	120 (2)	237 (7)	0 (0)	0 (0)
40-50	4710 (91)	4523 (114)	5921 (558)	3656 (74)
50-60	5271 (187)	6210 (162)	4200 (499)	5816 (109)
60-70	4728 (198)	7018 (231)	3903 (652)	2801 (52)
70-80	1997 (96)	3532 (195)	338 (59)	7 (0)
80<	313 (34)	510 (14)	0 (0)	0 (0)

Table 3.1: Demographics of Massachusetts General Hospital (MGH), Emory, Karolinska, and Chang Gung Memorial Hospital (CGMH) test sets. Each number is followed by the number of exams eventually followed by a cancer diagnosis.

were only used for held-out testing. The demographics for all test sets are reported in Table 3.1, and more detailed demographics for each dataset are shown in Tables B.11, B.12, B.13, B.14. Our dataset construction is shown in Fig. C-9. All datasets are described in detail in Methods.

Our primary objective was to develop personalized screening policies that would outperform current guidelines, improving early detection while reducing screening costs. To this end, we developed Tempo-Mirai, an RL-trained screening policy that operates on Mirai risk assessments. This policy takes as input a patient's Mirai risk assessment and outputs a follow-up recommendation as illustrated in Fig. 3-2. We implemented our risk progression model, which extrapolates unobserved Mirai risk assessments from prior risk assessments, as a recurrent neural network (RNN). This method is described in detail in Methods and validated in Table B.15. Sample risk progression predictions are shown in Fig. C-5.

We compared Tempo-Mirai with existing screening guidelines, including annual screening, biennial screening and a hybrid screening strategy recommended by the US Preventive Services Task Force (USPSTF)[109], which switches from annual screening to biennial screening at age 55 years. To assess the benefit of leveraging Mirai, a mammography-based AI risk model, over a traditional clinical risk model in the Tempo framework, we also developed Tempo-TCv8, a Tempo policy that operates on

TCv8 risk assessments. We utilized a deterministic model, static risk, to estimate risk progression for TCv8. This model is detailed in Methods. To quantify the benefit of using our RL approach to develop risk-based screening policies (i.e., Tempo) over a supervised learning approach, we also developed Supervised-Mirai and Supervised-TCv8. Instead of maximizing the overall reward with RL (without supervision for intermediate decisions), our supervised learning approach trains policies to predict the optimal follow-up recommendation at each time step. These baselines are detailed in Methods.

For each policy (e.g., Tempo-Mirai or annual screening), we measure its screening cost in terms of the average number of mammograms it recommends per year and its early detection benefit in months relative to historical screening. Our early detection metric assumed that early screening, following a patient's last negative mammogram, could offer a maximum early detection benefit of 18 months. We note that our early detection benefit metric is local and institution specific, as different institutions have different screening patterns. To directly compare policies that recommend differing numbers of mammograms, we also evaluated the efficiency of each policy, as measured by the early detection benefit in months divided by the number of mammograms per year recommended. Our efficiency metric is best suited to compare policies that obtain positive early detection benefits.

3.2.3 Evaluating personalized screening policies

The results of all screening policies across the MGH, Emory, Karolinska and CGMH test sets are illustrated in Table 3.2. We utilized the same Tempo-Mirai operating point across all test sets. We illustrate the performance of Tempo across different operating points (i.e., screening preferences) in all test sets in Fig. 3-3.

On the MGH test set, the annual and USPSTF guidelines obtained screening efficiencies (i.e., early detection benefit per screening cost) of 1.58 (95% confidence interval (CI), 0.54, 2.58) and 4.42 (95% CI, 5.83, 3.12). In contrast, Tempo-Mirai, Tempo-TCv8 and Supervised-Mirai obtained screening efficiencies of 4.29 (95% CI, 3.17, 5.25), 2.16 (95% CI, 1.18, 3.40) and 0.80 (95% CI, 0.58, 2.12), respectively. We found

Screening Policy	Risk model	Average number of Mammograms per Year	Earlier Detection in Months	Efficiency
MGH Test Set: 17,119 exams from 5,525 patients. 210 patients develop cancer.				
Annual	Age	1.0 (1.00, 1.00)	1.58 (0.54, 2.58)	1.58 (0.54, 2.58)
Biennial	Age	0.5 (0.50, 0.50)	-5.17 (-6.22, -4.13)	-10.34 (-12.44, -8.26)
USPSTF	Age	0.72 (0.71, 0.73)	-3.18 (-4.23, -2.22)	-4.42 (-5.83, -3.12)
Supervised	TCv8	1.66 (1.65, 1.69)	4.55 (3.51, 6.08)	2.74 (2.08, 3.70)
	Mirai	0.94 (0.92, 0.96)	0.75 (-0.55, 1.94)	0.80 (-0.58, 2.12)
Tempo	TCv8	0.96 (0.94, 0.97)	2.06 (1.14, 3.20)	2.16 (1.18, 3.40)
	Mirai	0.96 (0.94, 0.97)	4.10 (3.06, 4.96)	4.29 (3.17, 5.25)
Emory Test Set: 22,030 exams from 10,340 patients. 333 patients develop cancer.				
Annual	Age	1.0 (1.0, 1.0)	3.21 (2.37, 3.88)	3.21 (2.37, 3.88)
Biennial	Age	0.5 (0.5, 0.5)	-4.03 (–5.07, -3.22)	-8.07 (-10.14, -6.5)
USPSTF	Age	0.68 (0.67, 0.69)	-2.11 (-2.97, -1.40)	-3.12 (-4.36, -2.08)
Supervised	Mirai	1.16 (1.15, 1.18)	2.05 (0.48, 3.29)	1.76 (0.41, 2.86)
Tempo	Mirai	1.08 (1.07, 1.08)	6.39 (5.49, 6.99)	5.92 (5.06, 6.54)
Karolinska Test Set: 14,353 exams from 7,191 patients. 919 patients develop cancer.				
Annual	Age	1.0 (1.0, 1.0)	6.29 (5.76, 6.85)	6.29 (5.76, 6.85)
Biennial	Age	0.5 (0.5, 0.5)	-2.04 (-2.66, -1.41)	-4.07 (-5.32, -2.82)
USPSTF	Age	0.79 (0.79, 0.80)	1.02 (0.37, 1.63)	1.28 (0.46, 2.08)
Supervised	Mirai	0.60 (0.59, 0.61)	0.34 (-0.60, 1.24)	0.56 (-0.98, 2.11)
Tempo	Mirai	0.75 (0.74, 0.76)	7.23 (6.46, 7.97)	9.63 (8.53, 10.72)
CGMH Test Set: 12280 exams from 12280 patients. 235 patients develop cancer.				
Annual	Age	1.0 (1.0, 1.0)	11.00 (9.86, 12.28)	11.00 (9.86, 12.28)
Biennial	Age	0.5 (0.5, 0.5)	5.59 (4.11, 7.05)	11.18 (8.22, 14.09)
USPSTF	Age	0.78 (0.77, 0.78)	8.63 (7.15, 10.06)	11.10 (9.14, 13.01)
Supervised	Mirai	0.98 (0.97, 0.99)	8.02 (6.23, 10.21)	8.17 (6.30, 10.53)
Tempo	Mirai	0.88 (0.87, 0.89)	11.36 (10.21, 12.59)	12.92 (11.54, 14.41)

Table 3.2: Results for all screening policies on the MGH, Emory, Karolinska and CGMH test sets. For each policy, we report the average number of mammograms per year, the early detection benefit in months relative to historical screening (higher positive number means earlier), and the screening efficiency (higher positive number is better). We defined screening efficiency as the early detection benefit divided by the average number of mammograms per year. All metrics are followed by their 95% confidence interval.

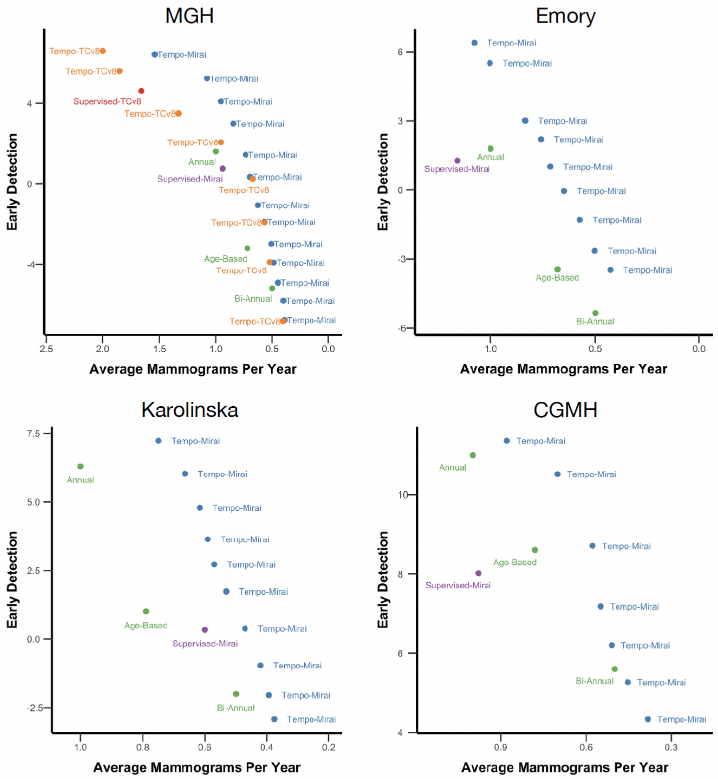

Figure 3-3: Early detection vs the number of mammograms per year at MGH (top left), Emory (top right), Karolinska (bottom left), CGMH (bottom right). Each point for a Tempo model (e.g., Tempo-Mirai) corresponds to an alternative preference in the trade-off between early detection and screening frequency. Tempo policies (i.e., Tempo-Mirai, Tempo-TCv8) are all trained using our reinforcement learning framework and Supervised policies (i.e., Supervised-Mirai, Supervised-TCv8) are trained using a supervised learning baseline. Mirai and TCv8 policies refer to policies that leverage Mirai, and Tyrer-Cuzick version 8 risk assessments respectively.

that Tempo-Mirai was significantly more efficient than Tempo-TCv8, Supervised-Mirai and annual screening ($P < 0.001$, $P < 0.001$ and $P < 0.001$), obtaining higher early detection per screening cost. Specifically, Tempo-Mirai obtained an early detection benefit of 4.10 (95% CI, 3.06, 4.96) months while recommending 0.96 (95% CI, 0.94, 0.96) mammograms per year, while the annual guideline obtained an early detection benefit of 1.58 (95% CI, 0.54, 2.58) months while recommending 1.0 mammograms per year.

In addition to overall performance on the test sets, we also studied the histogram of early detection benefits in Fig. C-6, and the histogram of recommended screening

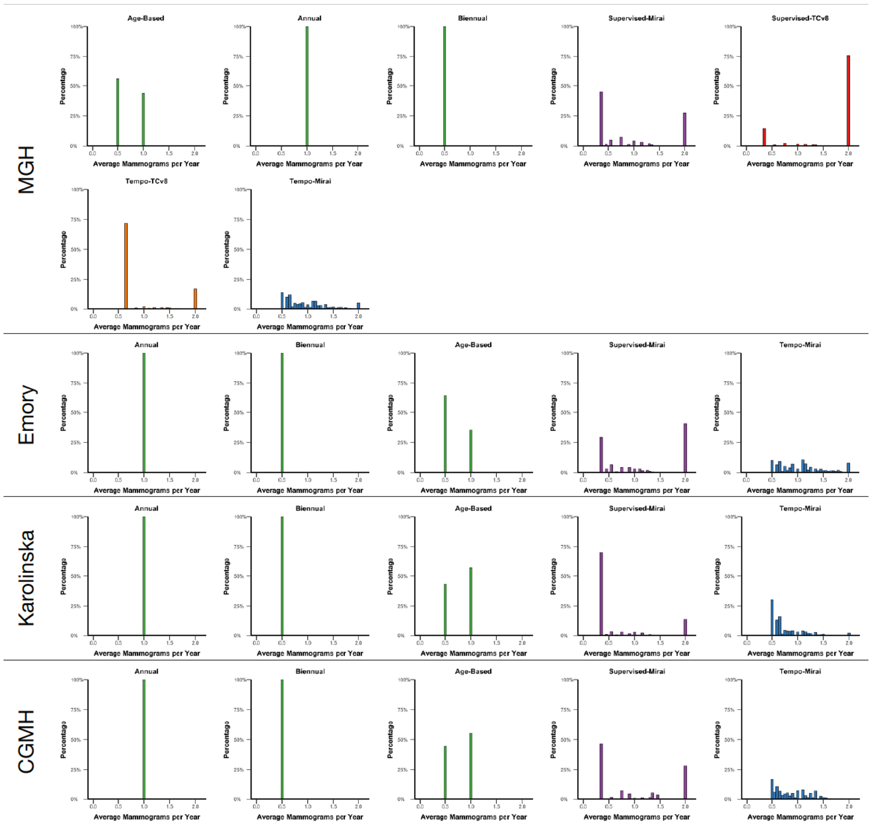

Figure 3-4: Histograms of screening frequency, i.e., average number of mammograms per year, as recommended by each screening policy for patients across the MGH (first row), Emory (second row), Karolinska (third row) and CGMH (fourth row) test sets.

frequencies in Fig. 3-4 and Fig. C-7. We note that all trained policies (for example, Tempo-Mirai, Supervised-Mirai) have the same set of possible recommendations ranging from a 6-month to 3-year screening follow-up, but we found that Supervised-Mirai only selected two options, recommending either 6 months or 3 years of follow-up. In contrast, Tempo-Mirai at our chosen operating point leveraged follow-up recommendations of 6 months, 1 year and 2 years. As shown on Fig. 3-4, we found that Tempo-Mirai offered a wider range of recommended frequencies than other methods, reflecting a larger degree for personalization. This reflects the optimization differences between the two policies. Tempo-Mirai is optimized to maximize overall reward across patient trajectories, as measured by early detection and screening cost, and does not receive any explicit guidance on the correct recommendation given a specific risk assessment. As a result, Tempo-Mirai has the flexibility to explore a wide range of possible recommendations during training to identify high-performing policies. In contrast, Supervised-Mirai has a rigid modeling objective; it is instead trained to predict the optimal (i.e., correct in hindsight) screening recommendation from each risk assessment, which is difficult given the uncertainty of real-world risk models.

To understand the flexibility of Tempo-based policies, we plotted the performance of each policy in Fig. 3-3 while varying the screening preference (i.e., operating point), which specifies the desired balance between early detection and screening cost. Across a wide range of possible operating points, Tempo-Mirai outperformed other policies in increasing early detection and reducing screening costs, demonstrating that the policy can be easily adapted to suit clinical requirements without retraining.

Next, we analyzed Tempo-Mirai's ability to generalize to new populations. To this end, we tested Tempo-Mirai, which was trained on MGH data, on test sets from Emory, Karolinska and CGMH. In the Emory test set, Tempo-Mirai, Supervised-Mirai and annual screening obtained efficiencies of 5.92 (95% CI, 5.06, 6.54), 1.76 (95% CI, 0.41, 2.86) and 3.21 (95% CI, 2.37, 3.88), respectively. In the Karolinska test set, Tempo-Mirai, Supervised-Mirai and annual screening obtained efficiencies of 9.63 (95% CI, 8.53, 10.72), 0.56 (95% CI, 0.98, 1.55) and 6.29 (95% CI, 5.76, 6.85), respectively. In the CGMH test set, Tempo-Mirai, Supervised-Mirai and annual screening obtained efficiencies of 12.92 (95% CI, 11.54, 14.41), 8.17 (95% CI, 6.30, 10.53) and 11.00 (95% CI, 9.86, 12.28), respectively. Tempo-Mirai was significantly more efficient than Supervised-Mirai and annual screening in all test sets, with $P < 0.001$ and $P < 0.001$ at Emory, $P < 0.001$ and $P < 0.001$ at Karolinska and $P < 0.001$ and $P = 0.02$ at CGMH.

While the above results show that Tempo-Mirai consistently improved over alternate policies in screening efficiency, we also observed that the absolute magnitude of early detection varied substantially across different datasets. For instance, annual screening obtaining early detection benefits of 1.58 (95% CI, 0.54, 2.58), 3.21 (95% CI, 2.37, 3.88), 6.29% (95% CI, 5.76, 6.85) and 11.0 (95% CI, 9.86, 12.28) months in the MGH, Emory, Karolinska and CGMH test sets, respectively. This difference can be attributed to the different rates of screening across the datasets; patients with future cancer at

MGH, Emory, Karolinska and CGMH obtained an average of 0.93, 0.94, 0.80 and 0.66 mammograms per year. These differences are further detailed in Tables B.11, B.12, B.13, B.14, which report detailed demographics of each dataset.

We also noted that Tempo-Mirai recommended different amounts of screening across the MGH, Emory, Karolinska and CGMH test sets, recommending an average of 0.96 (95% CI, 0.94, 0.97), 1.08 (95% CI, 1.07, 1.08), 0.75 (95% CI, 0.74, 0.76) and 0.88 (95% CI, 0.87, 0.89) mammograms per year, respectively. This difference can be attributed to differences in cancer incidence between the different centers. The 5-year cancer incidence at MGH, Emory, Karolinska and CGMH was 2.2%, 3.0%, 1.2% and 1.8%, respectively, and we expect Tempo to recommend higher rates of screening for higher risk populations. However, the model can offer a diverse set of possible operating points across all test sets, as illustrated in Fig. 3-3; our results indicate that different hospitals may need to input different operating points to obtain the same average screening volume

3.2.4 Subgroup analysis

We also investigated how our policies performed for different patient subgroups by race in the Emory test set in Table B.16 and by age and breast density in the MGH test set in Table B.17. We highlight the results of Tempo-Mirai, which obtained efficiencies of 5.92 (95% CI, 5.06, 6.54) and 4.29 (95% CI, 3.17, 5.25) on the Emory and MGH test sets. Tempo-Mirai obtained efficiencies of 5.89 (95% CI, 4.56, 7.02) and 6.06 (95% CI, 5.28, 7.04) for African American and white patients in the Emory test set. When grouping MGH patients by age, Tempo-Mirai obtained efficiencies of 3.41 (95% CI, 1.44, 5.53) or 4.45 (95% CI, 3.49, 6.03) for patients aged younger or older than 55 years, respectively. When grouping MGH patients by breast density category, Tempo-Mirai obtained efficiencies of 4.10 (95% CI, 2.85, 5.48) and 4.49 months (95% CI, 2.86, 6.35) for patients with nondense and dense breasts, respectively, where nondense refers to the Breast Imaging Reporting and Data System categories of almost entirely fatty or scattered areas of fibroglandular tissue and dense refers to the Breast Imaging Reporting and Data System categories of heterogeneously dense or extremely dense.

3.2.5 Robustness to assumptions

Our empirical results across the different test sets depend on the exact choice of assumptions of our early detection metric. As illustrated in Fig. 3-1, our early detection metric measured the time difference in months between each patient's recommended screening date and the diagnosis date. Our metric assumed that the maximum early detection benefit obtained through earlier screening was 18 months. To test our model's robustness to this assumption, we also evaluated Tempo-Mirai, Supervised-Mirai and annual screening across all test sets when setting our maximum early detection benefit assumption to 6, 12, 18 and 24 months. We note that we did not retrain Tempo-Mirai for this analysis and that Tempo-Mirai was originally trained using the 18-month assumption. For each policy, we measured its screening efficiency (i.e., the early detection benefit divided by the number of mammograms recommended per year) to enable a head-to-head comparison between policies that recommend different screening volumes. As shown in Extended Data Fig. C-8, Tempo-Mirai is more efficient than annual screening across all datasets and assumptions. This result is further supported by the histogram of early detection benefits shown in Fig. C-6.

3.3 Discussion

We developed an RL framework for personalized screening, Tempo, to predict follow-up recommendations from patient risk assessments. We demonstrated that a Tempo policy based on Mirai risk assessments was significantly more efficient than annual screening, achieving earlier detection per screening cost. Moreover, we showed that the same Tempo policy can be adapted to a wide range of possible screening preferences and that policies that leverage more accurate risk models (i.e., Mirai) outperform those based on less accurate risk models (i.e., Tyrer–Cuzick). We found that policies developed using data from MGH generalized to held-out test sets in Emory, Karolinska and CGMH and significantly outperformed both annual screening and our supervised learning baselines. Finally, we demonstrated our results were robust across a range of possible assumptions for our early detection metric.

Our screening policies can be easily implemented in any screening clinic where Mirai risk assessments are collected. Clinicians can retrospectively validate our trained screening policies on their own screening population and choose an operating point to achieve the desired balance between screening volume and early detection benefit. The installed policy can then offer clinicians suggested risk-based follow-up intervals immediately after a patient's risk assessment. Depending on clinical requirements, Tempo can be utilized to significantly reduce the volume of screening for a fixed early detection target or improve early detection for a fixed screening budget. For instance, we showed that Tempo-Mirai could obtain better early detection than annual screening at Karolinska while reducing screening by 25%. Given the scale and cost of breast cancer screening, even modest improvements in screening guidelines have the potential to benefit a wide patient population.

Our study is complementary to a rich body of work surrounding risk-based screening[42, 87, 84, 88]. Several guidelines already recommend supplemental imaging or chemoprevention based on risk assessments[84, 88, 127], and recent results from the DENSE trial[9] have shown that a breast density-based screening strategy could significantly reduce interval cancers compared to current screening. Our work is most closely related to the MyPeBS trial[28], which prospectively compares a personalized screening follow-up strategy based on either Tyrer–Cuzick[123] or MammoRisk[66] risk assessments with current national recommendations. These studies point to substantial clinical interest in risk-based screening; however, current methods for devising screening policies rely on categorizing patients into a few coarse categories (e.g., low and high risk), limiting personalization.

Our study provides a data-driven alternative for clinical decision-making and can be easily integrated into a screening trial or routine patient care. Our work is also complementary to ongoing efforts to improve mammography reading; Tempo screening policies can be deployed in tandem with new technologies aimed at improving breast cancer detection at the time of screening (i.e., computer-aided detection[79, 72] or triage systems[148, 99]. Our work is also related to a large volume of modeling studies focused on breast cancer[75, 8, 129, 76, 122, 105, 3]. Typically, these approaches operate over a model of disease progression that characterizes how patients transition between healthy and disease states. The transitions are informed by patient features and impact the likelihood of different observations, such as a palpable lump. Their probabilities can be estimated from retrospective data or retrieved from the literature. The approaches then work to identify the optimal screening policy under the specified disease progression model. While these approaches were the first to demonstrate the feasibility of developing personalized screening policies, they have several limitations that restrict their practical use in clinical settings. First, the postulated disease progression model does not capture the full complexity and uncertainty of cancer. Second, the methods generally assume that a patient's features are fixed and do not evolve over their lifetime[3]. This assumption does not hold in general and is not applicable to modern AI-based risk models that are sensitive to changes in patient health. In contrast, our framework does not assume a complete disease progression model; instead, it assumes access to a risk model (rather than discrete set of states) and a reward function that measures the performance of a screening trajectory given observational data. This relaxed assumption allows us to optimize screening policies directly on observed patient trajectories, which contain the full diversity of cancer diagnoses, as well as validate our policies on held-out patient populations, which may differ in their cancer characteristics, such as Emory, Karolinska and CGMH.

This study focuses on breast cancer screening using image-based risk models. However, our framework is flexible and can be readily utilized for other diseases, other forms of risk models and other definitions of early detection benefit. For instance, it can easily incorporate richer representations of the cancer outcomes. Recent work has highlighted concerns about the potential overtreatment of ductal carcinoma in situ[126]. Tempo policies can take these differences into account by leveraging separate reward metrics for the early detection of invasive and in situ cancers. In this scenario, Tempo policies would be trained using three reward metrics (early detection of invasive cancers, early detection of in situ cancers and screening cost), and clinicians would select a Tempo operating point (i.e., screening preference) that achieves the desired balance among the three metrics. In a similar fashion, our framework can be used to optimize more refined definitions of early detection benefit that account for properties of the cancer (e.g., tumor size and grade) at the time of diagnosis. For instance, given access to a patient's tumor properties, a cancer mortality model and a cancer growth model, a sophisticated early detection metric could directly estimate the reduced mortality risk if the cancer had been diagnosed at an earlier time point. Given a patient's age, this metric could also directly be tried to quality-adjusted life years. Similarly, more sophisticated measures of screening cost that take into account varying false-positive risks depending on patient characteristics (e.g., breast density) could be used to further refine screening policies. In this sense, prior work in modeling cancer mortality and screening benefits[75, 8, 129, 76, 122, 105] is complementary to our own. We expect that the utility of Tempo, which is agnostic to the underlying choice of screening metrics and risk model, will increase as risk models and outcomes metrics are further refined across more diseases.

There are multiple future directions that can further improve personalized screening algorithms. While our method focused on predicting follow-up recommendations given risk estimates from established risk models, one could instead directly input rich patient information, such as a patient's mammograms and family history, into the screening policy. Directly learning to interpret this information for the purpose of personalized screening in an end-to-end fashion may result in more accurate policies. Moreover, the action space of our method could be expanded to include different types of screening recommendations, such as leveraging magnetic resonance imaging or mammograms, and future work could separately model the costs and benefits of each modality. Finally, given improved screening policies, future work could also recalculate the earliest and latest age such that screening is still cost-effective for a patient.

This study has several limitations. Our early detection metric assumed that cancer is detectable up to a fixed time (18 months) before diagnosis. While we found that the trends reported in our study were robust to different values of this assumption (ranging from 6 to 24 months), none of these assumptions are individually correct across all cancers, as the early detection potential of a tumor depends on that tumor's characteristics at the time of diagnosis. Moreover, our screening cost metric, recommended mammography volume, does not provide a full analysis of screening cost; it does not quantify false-positive risks or additional screening harms.

Our simulations also did not account for the sensitivity of screening mammography or the probability of a patient entering the clinic with a palpable lump if their diagnosis is overly delayed. While our framework is agnostic to the specifics of how the rewards are formulated, further research using more refined early detection metrics, such as quality-adjusted life years, that explicitly model tumor characteristics at the time of detection and tumor growth is needed. While Tempo can be applied with any risk model, Tempo-Mirai inherits the limitations of Mirai. Mirai has only been validated using Hologic full-field digital mammograms, and future work is needed to adapt the risk model to more mammography vendors and tomosynthesis images. Finally, prospective trials are necessary to assess the efficacy of these models in clinical care before widespread adoption.

3.4 Methods

3.4.1 Study design

The primary objective of this study was to develop personalized screening policies that could improve early detection while reducing screening costs. To this end, we developed Tempo, an RL framework for personalized screening that can be paired with any risk model. As illustrated in Fig. 3-2, Tempo policies are neural networks that take as input a risk assessment and output a screening follow-up recommendation. In this study, we focused our attention on breast cancer screening, and we hypothesized that our Tempo policies could offer improved early detection benefits over annual screening without requiring more screening. Moreover, we hypothesized that these policies would generalize to new institutions. We developed Tempo-Mirai, an RL-based policy that operates on Mirai (version 0.4.0.) risk assessments, and compared this policy to existing guidelines, including annual and age-based screening. Mirai is a deep learning-based risk model that predicts a patient's future risk directly from their mammogram. To assess the benefit of leveraging Mirai risk assessments over a traditional risk assessment model (i.e., Tyrer–Cuzick), we also developed Tempo-TCv8, an RL-based policy that operates on Tyrer–Cuzick risk assessments. To evaluate the benefit of using our RL approach for creating personalized risk policies (i.e., Tempo), we also developed models based on a supervised learning approach, Supervised-Mirai and Supervised-TCv8. An RNN was used to estimate risk progression for Mirai, and a deterministic model, static risk, was used to estimate risk progression for TCv8. All models were trained on the MGH dataset and tested at MGH, Emory, Karolinska and CGMH.

3.4.2 Dataset Description

Dataset description. To develop Tempo, we collected consecutive full-field screening mammograms and detailed risk information at the time of mammography from 80,134 patients screened between 1 January 2009 and 31 December 2016 at MGH under approval of MGH's institutional review board (IRB) with a waiver for written informed consent and in compliance with the Health Portability and Accountability Act. We obtained outcomes through linkage to a local five-hospital registry in the Massachusetts General Brigham healthcare system, alongside pathology findings from MGH's mammography electronic medical record. We collected detailed risk factors, including those used by the Tyrer–Cuzick model, from provider-entered information and patient-entered questionnaires in the electronic medical record. We associated each mammogram with patient risk factors as present at the time of mammography. We excluded patients who were diagnosed with other cancers (e.g., sarcoma) in the breast or did not have all four views (left craniocaudal (CC), left mediolateral oblique (MLO), right CC and right MLO). For patients who developed cancer, we excluded exams made within 6 months of diagnosis. For patients who did not develop cancer, we excluded exams made within 3 years of the last follow-up screen. We note that 6 months and 3 years are the minimum and maximum follow-up recommendations for Tempo, so this exclusion enabled us to ensure that simulations always occur within the bounds of observed data. This exclusion resulted in 54,673 patients who were randomly split into groups for training (43,749), development (5,399) and testing (5,525). We note that this dataset was also used to develop Mirai12, so we used the same training, development and testing splits. Because each patient had multiple exams, this resulted in 137,682, 16,634 and 17,119 exams for training, development and testing, respectively. All mammograms were acquired on Hologic machines. For each exam, we obtained Mirai[146] risk assessments, as well as TCv8 risk assessments. Detailed demographic information for this dataset is available in Table B.11, and the dataset construction procedure is shown in Fig. C-9.

To evaluate the ability of Tempo policies to generalize to new populations, we collected the Emory, Karolinska and CGMH datasets under approval of the relevant IRBs with a waiver for written informed consent. To create the Emory test set, which contains a large representation of African American women, we extracted 8 years of full-field mammograms from an institutional database of all comers for screening mammography from 2013 to 2020 and randomly selected 30% of women (28,994 patients). All mammograms were acquired on Hologic machines. We collected outcomes from pathology findings from Emory's mammography electronic medical record. We obtained Mirai risk assessments for each exam. As with the MGH dataset, we excluded exams within 6 months of diagnosis. For patients who did not develop cancer, we excluded exams within 3 years of the last follow-up screen. This resulted in a total of 22,030 exams from 10,340 patients. Detailed demographics of this dataset are shown in Table B.12, and the dataset construction procedure is shown in Fig. C-9.

The Karolinska test set was extracted from the cohort of screen-aged women[35]. All women aged 40–74 years within the Karolinska University uptake area who had attended screening and were diagnosed with breast cancer, without implants or prior breast cancer, from 2008 to 2016 were included, as well as a random sample of controls with at least 2 years of follow-up data from the same time period. The full Karolinska case–control and validation datasets included 11,301 and 2,580 women, respectively. A random subset of 9,484 patients in total were selected for inclusion in this study. We included all full-field mammograms, acquired on Hologic machines, from 2008 to 2016 for the included women that contained all four views (left CC, left MLO, right CC and right MLO), resulting in 14,362 exams from 7,193 patients. We excluded exams within 6 months of a cancer diagnosis. For patients who did not develop breast cancer, we excluded exams within 3 years of the last screening follow-up. Because of the case-control dataset design, this dataset has a much higher ratio of patients who developed cancer, relative to the 1.9% incidence reported in the cohort of screen-aged women[35]. To take this into account, we randomly resampled patients who did not develop cancer from our cohort to produce a larger dataset with a 1.9% cancer incidence, resulting in a total of 93,052 exams from 7,193 patients. Detailed demographics are shown in Table B.13 with the dataset construction procedure in Fig. C-9. Given the 1.9% patient-level cancer rate and the length of the collection period, we estimated that the 5-year cancer incidence in the Karolinska population was 1.2%. For each exam, we obtained Mirai[146] risk assessments.

To create the CGMH test set, which consisted of 12,280 exams from 12,280 patients, we selected random women undergoing full-field screening mammography at CGMH between 2010 and 2011 who were aged 45–70 years. Women aged 40–44 years were also included if they had a family history of breast cancer, following local screening guidelines. All mammograms were acquired on Hologic machines. Cancer outcomes were obtained from the national cancer registry. Demographics for this dataset are available in Table B.14 and Fig. C-9. For all patients, we collected the date of last screening follow-up. We excluded patients with unknown age. For each patient who developed cancer, we also manually collected all the dates of their future screenings from 2010 to 2020 through chart review. This allowed us to estimate early detection benefits relative to historical screening. We did not collect all future screening dates for patients who did not develop cancer. For patients who developed cancer, we excluded exams within 6 months of diagnosis, while for patients who did not develop cancer, we excluded exams within 3 years of the last follow-up screen. For each exam, we obtained Mirai risk assessments. The CGMH test set only included one Mirai risk assessment per patient; as a result, our ability to estimate risk progression at CGMH is more limited compared with the other test sets, where the risk progression model benefited from multiple prior observations and made predictions across shorter time intervals. This limitation means it is more difficult to estimate the quality of Mirai-based policies (i.e., Tempo-Mirai and Supervised-Mirai) in this dataset.

For patients with multiple exams in a dataset, we considered each exam in their trajectory as a possible simulation starting point and evaluated screening policies across all starting points. For instance, consider a patient who was screened in years 1, 2 and 3. For training and evaluation, we consider the scenarios in which the patient started to follow the Tempo-Mirai policy at years 1, 2 and 3. Simulating policies from multiple starting points offers more information about the behavior of a policy. To account for these correlated simulations in computing our CIs, we used a clustered bootstrap procedure with 5,000 samples. We note that our risk progression model always had access to all prior observations and was not affected by the choice of simulation starting point.

For each trajectory, we considered its censor time as either the date of cancer diagnosis via biopsy or the date of last screening follow-up. We designed our screening policies to offer a minimum follow-up recommendation of 6 months and a maximum follow-up recommendation of 3 years. Because our follow-up intervals were in increments of 6 months, we discretized time across all trajectories into 6-month time steps. This was done by subtracting the first date in the trajectory from all dates and then dividing the date difference by 6 months using integer division (i.e., without rounding). As a result, an exam 9 months after time-step 0 was considered step 1. This design decision simplified our simulation code.

To ensure that our simulations always occurred within the time frame of the observed data, we excluded starting points where cancer was diagnosed in less time than the minimum action (6 months). For screening trajectories without a cancer diagnosis, we excluded starting points where the time to the last screening follow-up was less than the maximum action (3 years). To understand the latter exclusion, consider a patient with no known future cancer date who was screened at year 1 and had her last screening follow up at year 2. If a Tempo policy recommended follow-up in 3 years (e.g., return at year 4), then we could not assess whether that recommendation would result in a diagnosis delay as that time point (i.e., year 4) is unobserved. To avoid this scenario, we exclude exams where a Tempo policy cannot be evaluated (i.e., within 3 years of the last follow-up date if the patient does not develop cancer).

Mammograms were converted to the PNG16 format using the dcmj2pnm command of the DCMTK toolkit (version 3.6.1, 2015). Torchvision (version 0.2.1) and Pillow (version 5.2.0) Python libraries were used for image preprocessing and data augmentations.

3.4.3 Reward Design

We considered two rewards in our simulation environment: measuring imaging cost and early detection benefit. We modeled our imaging cost reward as the negative amount of mammograms per year recommended by a policy. To model early detection benefits, we measured the time difference in 6-month time steps between each patient's recommended screening date (if it was after their last negative mammogram) and the actual diagnosis date. We then converted this value into months. We defined a patient's diagnosis date as the date of their positive biopsy result. Negative values of this reward imply a delayed diagnosis, and positive values imply relative screening benefit over the retrospective trajectory. We capped maximum early detection benefit for any patient at 18 months and did not cap the possible screening delay. As a result, if a patient's last negative mammogram was 3 years before their cancer diagnosis and a screening policy recommended a mammogram 2 years and 1 year before a patient's cancer diagnosis, then we assigned this trajectory an early detection benefit of 18 months. We provide additional analysis for different possible assumptions for the maximum screening benefit in Fig.C-8. We also considered an alternative definition of early detection benefit, where a policy can only offer early detection if it recommends an additional screen within 18 months of the diagnosis date in Table B.18. In the above example where a patient is screened 2 years and 1 year before their diagnosis, this definition would yield an early detection benefit of 12 months instead of 18 months. Across both definitions (Table 3.2 and Table B.18), Tempo-Mirai obtains better efficiency than other guidelines (e.g., annual screening).

3.4.4 Risk progression models

As shown in Fig. 3-2, our risk progression models take as input a sequence of prior risk assessments and predict a risk assessment at the next time step. We considered two possible methods to estimate risk progression, namely Static Risk, which always predicted that a patient's risk at the next time step would be the same as at the last time step, and an RNN. Our RNN estimated risk progression in an iterative fashion; at each step, it took as input a single risk assessment and outputted a single risk assessment for the next time step. We implemented our RNN as a gated recurrent unit[25] with an additive hazard layer[146] and trained the model to minimize the Kullback– Leibler divergence between predicted risk assessments and the risk assessments observed in the MGH training set.

We experimented with different learning rates, hidden sizes, number of layers and dropout, and we chose the model that obtained the lowest validation Kullback–Leibler divergence in the MGH validation set. Our final risk progression RNN had two layers, a hidden dimension size of 100 and a dropout of 0.25, and it was trained for 30 epochs with a learning rate of 1×10^{-3} using the Adam optimizer. The outputs of our risk progression model for Tempo-Mirai are visualized in Fig. C-5. Given a trained risk progression model, we can now estimate unobserved risk assessments autoregressively. At each time step, the model takes as input the previous risk assessment, the prior hidden state, using the previous predicted assessment if the real one is not available and predicts the risk assessment at the next time step. We validated our risk progression network on the MGH, Emory and Karolinska test sets in Table B.15 and note that our RNN outperformed the static risk baseline in all datasets. Because we collected only one exam for each patient in the CGMH test set, we could not validate the risk progression network on that test set. Information regarding the implementation for each risk progression and hyperparameter search is available in our code release.

3.4.5 Personalized screening models

We implemented our personalized screening policy as a multiple-layer perceptron, which took as input a risk assessment and weighting between rewards and predicted the Q-value for each action (i.e., follow-up recommendation) across the rewards. This network was trained using Envelope Q-Learning[149]. Following recent work in deep RL[81, 5], we used an experience replay buffer to reduce correlation between our training batches and utilized a target Q-network[81] to stabilize training updates.

We experimented with different numbers of layers, hidden dimension sizes, learning rates, dropouts, exploration epsilons, target network reset rates and weight decay rates. We note that we conducted the same grid searches for Tempo-Mirai and Tempo-TCv8 and chose each model to maximize the average reward on the MGH validation set. Our final Tempo-Mirai model had six layers, each with 256 hidden units, followed by rectified linear unit (ReLU) nonlinearities. It was trained for 30 epochs using a learning rate of 1×10^{-3}, a dropout of 0.25 and a weight decay of 0.01 using the Adam optimizer, and the target network was reset every 1,000 batches. Our final Tempo-TCv8 model had four layers, each with 256 hidden units, followed by ReLU nonlinearities. It was trained for 30 epochs using a learning rate of 1×10^{-3}, a dropout of 0.25 and a weight decay of 0 using the Adam optimizer, and the target network was reset every 1,000 batches. Information regarding the implementation of each risk policy, the training code and our hyperparameter searches is available in our code release. For both Tempo-Mirai and Tempo-TCv8, we chose a reward weighting to approximately match the screening cost of annual screening on the MGH development set and used this reward weighting across all test sets. Tempo-Mirai used a reward weight of 0.5 and 3.0 for screening cost and early detection, respectively. Tempo-TCv8 used a reward weight of 0.77 and 3.0 for screening cost and early detection, respectively.

3.4.6 Supervised learning baseline

We implemented our supervised learning baselines, Supervised-Mirai and Supervised-TCv8, as a multiple layer perceptron, which took as input a risk assessment and predicted a probability distribution across follow-up recommendations. This network was trained to minimize the cross-entropy loss between its actions and the optimal sequence of actions. We computed optimal actions for each patient to maximize our rewards metrics. For patients who did not develop cancer within the time period of the maximum follow-up recommendation, the optimal action was the maximum follow-up recommendation of 3 years. For patients who developed cancer, the optimal action was to recommend a screening follow-up in the time step following the last negative mammogram. Unlike Tempo-Mirai, which is trained to maximize trajectory level rewards using RL, Supervised-Mirai is trained to maximize the likelihood of the optimal sequence of actions. As a result, Supervised-Mirai does not benefit from observing how its own errors compound across the trajectory at training time.

For each supervised model, we experimented with different numbers of layers, hidden dimension sizes, learning rates, dropouts and weight decays. To enable fair comparison against Tempo models, we searched the same space of hyperparameters and selected those that achieved the best average reward on the MGH validation set. Our final Supervised-Mirai model had eight layers, each with 512 hidden units, followed by ReLU nonlinearities. It was trained for 30 epochs using a learning rate of 1×10^{-3}, a dropout of 0.25, a weight decay of 0.1 and the Adam optimizer. Our final Supervised-TCv8 model also had eight layers, each with 512 hidden units, followed by ReLU nonlinearities. It was trained for 30 epochs using a learning rate of 1×10^{-4}, a dropout of 0.25, a weight decay of 0.1 and the Adam optimizer. Information regarding the implementation of each risk policy, the training code and our hyperparameter searches is available in our code release.

3.4.7 Statistical analysis

To calculate CIs while accounting for patients with multiple simulations, we used a clustered bootstrap approach with 5,000 samples. To assess significance in the difference between two metrics, we used a two-tailed t test with a predefined P value of 0.05 for significance.

3.4.8 Data and Code Availability

Data availability All datasets were used under license to the respective hospital system for the current study and are not publicly available. All models and code used for training, evaluating and developing Tempo are publicly available at learningtocure.csail.mit.edu and github.com/yala/Tempo (https://doi.org/10.5281/zenodo.5585318).

Chapter 4

Syfer: Neural Obfuscation for Private Data Release

4.1 Introduction

Data sharing is a key bottleneck for the development of equitable clinical AI algorithms. Public medical datasets are constrained by privacy regulations [54, 43], that aim to prevent leakage of identifiable patient data. We propose Syfer, an encoding scheme for private data release. In this framework, data owners encode their data with a random neural network (acting as their private key) for public release. The objective is to enable untrusted third parties to develop classifiers for the target task, while preventing attackers from re-identifying raw samples.

An ideal encoding scheme would enable model development for arbitrary (i.e unknown) downstream tasks using standard machine learning tools. Moreover, this scheme would not require data owners to train their own models (e.g. a generative model). Designing such an encoding scheme has remained a long-standing challenge for the community. For example, differentially private methods pursue this goal by leveraging random noise to limit the sensitivity of the encoding to the input data. However, this often results into too large of a utility loss. In this work, we propose to learn a keyed encoding scheme, which exploits the asymmetry between the tasks of model development and sample re-identification, to achieve improved privacy-utility trade-offs.

The relevant notion of privacy, as defined by HIPAA, is *de-identification*, i.e. preventing an attacker from identifying matching pairs between raw and encoded samples. We measure this risk using *guesswork*, i.e. the number of guesses an attacker requires to match a single raw image to its corresponding encoded sample. We consider an extreme setting where the attacker has access to the raw images, the released encoded data and the randomized encoding scheme, and only needs to predict the matching between corresponding pairs (raw image, encoded image). While the adversary can simulate the randomized encoding scheme, they do not have access to the data owner's private key. Our evaluation setup acts as a worst-case scenario for data privacy, compared to a real-world setting where the attacker's knowledge of the raw images is imperfect. To efficiently measure guesswork on real-world datasets, we leverage an model-based attacker trained to maximize the likelihood of re-identifying raw images across encodings.

While an arbitrary distribution of random neural networks is insufficient to achieve strong privacy (i.e. high guesswork) on real-world datasets, we can learn to shape this distribution to obtain privacy on real data by composing random layers with trained *obfuscator* layers. Syfer's obfuscator layers are optimized to maximize the re-identification loss of a model-based attacker on a public dataset while minimizing a reconstruction loss, maintaining the invertability of the whole encoding. To encode labels, we apply a random permutation to the label identities.

We trained Syfer on a public X-ray dataset from NIH, and evaluated the privacy and utility of the scheme on heldout dataset (MIMIC-CXR) across multiple attacker architectures and prediction tasks. We found that Syfer obtained strong privacy, with an expected guesswork of 8411, i.e. when presented with a grid of 10,000 raw samples by 10,000 encoded samples, it takes an attacker an average of 8411 guesses to correctly guess a correct (raw image, encoded image) correspondence. Moreover, models built on Syfer encodings approached the accuracy of models built on raw images, obtaining an average AUC of 0.78 across diagnosis tasks compared to 0.84 by a non-private baseline with the same architecture, and 0.86 by the best raw-image baseline. In contrast, prior encoding schemes, like InstaHide [55] and Dauntless [139], do not prevent re-identification, both achieving a guesswork of 1. While differential privacy schemes, such as DP-Image [69], can eventually meet our privacy standard with large enough noise, achieving a guesswork of 1379, this resulted in average AUC loss of 33 points relative to the raw-image baseline, i.e. an AUC of 0.53.

4.2 Related Work

Differentially Private Dataset Release Differential privacy [40] methods offer strong privacy guarantees by leveraging random noise to bound the maximum sensitivity of function outputs (e.g. dataset release algorithms) to changes in the underlying dataset. For instance, DP-Image [69] proposed to add laplacian noise to the latent space of an auto-encoder to produce differentially private instance encodings. Instead of directly releasing noisy data, [141, 119, 59] propose to leverage generative adversarial networks (GANs), trained in a differentially private manner (e.g. DP-SGD [2] or PATE [89]), to produce private synthetic data. However, differentially private GANs have been shown to significantly degrade image quality and result in large utility losses [23]. Instead of leveraging independent noise per sample to achieve privacy, Syfer obtains privacy through its keyed encoding scheme and thus enables improved privacy-utility trade-offs.

Cryptographic Techniques Cryptographic techniques, such as secure multiparty computation and fully homomorphic encryption [150, 47, 11, 21, 44, 17, 24] allow data owners to encrypt their data before providing them to third parties. These tools provide extremely strong privacy guarantees, making their encrypted data indistinguishable under chosen plaintext attacks (IND-CPA). However, building models with homomorphic encryption [82, 70, 60, 15] requires leveraging specialized cryptographic primitives and induces a large computational overhead (ranging from 100x-10,000x [71]) compared to standard model inference. As a result, these tools are still too slow for training modern deep learning models. In contrast, Syfer considers a weaker threat model, where attackers cannot query the data owner's private-encoder (i.e no plaintext attacks) and our scheme specifically defends against raw data re-identification (the privacy notion of HIPAA). Moreover, Syfer encodings can be directly leveraged by standard deep learning techniques, improving their applicability.

Lightweight Encoding Schemes Our work extends prior research in lightweight encoding schemes for dataset release. Previous approaches [63, 116, 108] have proposed tools to carefully distort images to reduce their recognition rate by humans while preserving the accuracy of image classification models. However, these methods do not offer privacy against machine learning based re-identification attacks. [136, 140, 137, 96] have proposed neural encoding schemes that aim to eliminate a particular private attribute (e.g. race) from the data while protecting the ability to predict other attributes (e.g. action) through adversarial training. These tools require labeled data for sensitive and preserved attributes, and cannot prevent general re-identification attacks while preserving the utility of unknown downstream tasks. Our work is most closely related to general purpose encoding schemes like InstaHide [55] and Dauntless [139, 138]. InstaHide encodes samples by randomly mixing images with MixUp [152] followed by a random bitwise flip. Dauntless encodes samples with random neural networks and proved that the scheme offers strong information theoretic privacy if the input data distribution is Gaussian. However, we show that neither InstaHide nor Dauntless meet our privacy standard on our real-world image datasets. In contrast, Syfer leverages a composition of trained obfuscator layers and random neural networks to achieve privacy on real word datasets while preserving downstream predictive utility.

Evaluating Privacy with Guesswork Our study builds on prior work leveraging guesswork to characterize the privacy of systems [78, 80, 7, 92, 10]. Guesswork quantifies the privacy of a system as the number of trials required for an adversary to guess private information, like a private key, when querying an oracle. In this framework, homomorphic encryption methods, which uniformly sample b-bit private keys, offer maximum privacy [38], as the average number of guesses to identify the

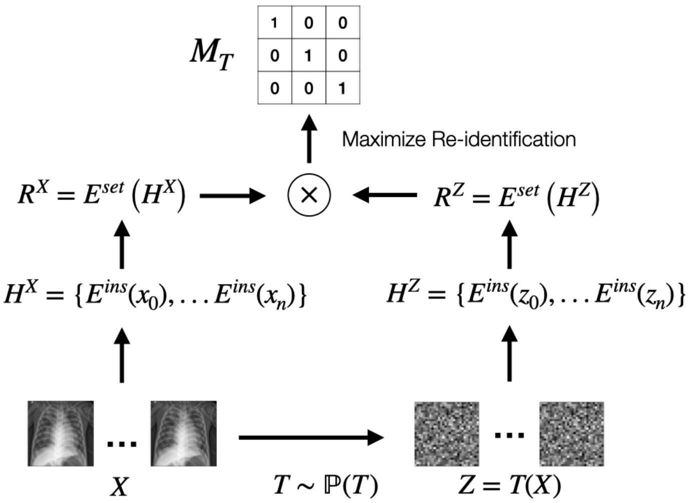
M_T
1 0 0
0 1 0
0 0 1
Maximize Re-identification
$R^X = E^{set}(H^X)$
$R^Z = E^{set}(H^Z)$
$H^X = \{E^{ins}(x_0), \dots E^{ins}(x_n)\}$
$H^Z = \{E^{ins}(z_0), \dots E^{ins}(z_n)\}$
X
$T \sim \mathbb{P}(T)$
$Z = T(X)$

Figure 4-1: Architecture of the model-based attacker. Given pairs of raw samples $(X, LF(X))$ and encoded samples (Z, Y), the attacker learns to recover matching pairs $(x, z) \in M_T$. In this figure, we omit label information for clarity.

correct key is 2^{b-1}. In the non-uniform guessing setting [26], guesswork offers a worst-case notion of privacy by capturing the situation where an attacker may only be confident on a single patient identity. Such privacy weaknesses are not measured by average case metrics, like Shannon entropy.

4.3 Method

We propose Syfer, an encoding scheme which uses a combination of learned *obfuscator* layers and random neural network layers to encode raw data. Syfer is trained to maximize the re-identification loss of an attacker while minimizing a reconstruction loss, which acts as a regularizer to preserve predictive utility for downstream tasks. To estimate the privacy of an encoding scheme on a given dataset, we use a model-based attacker trained to maximize the likelihood of re-identifying raw data. To encode the labels $LF(X)$, Syfer randomly chooses a permutation of label identities $\{1, ..., k\}$.

4.3.1 Privacy Estimation via Contrastive Learning

Before introducing Syfer, we adopt Eve's perspective and describe how to evaluate the privacy of encoding schemes. The attacker is given the candidate list $X_E = X_A$, and a fixed encoding scheme Γ, i.e. a fixed distribution $P(\boldsymbol{T})$. We propose an efficient contrastive algorithm to estimate $P\left((x,z) \in M_T | Z_A, Y_A, \Gamma, X_A\right)$. When the context allows it, we omit the conditional terms and use $P((x,z) \in M_T)$.

As shown in Figure 4-1, the attacker's model E is composed of an instance-level encoder E^{ins}, with parameters φ^{ins}, acting on individual images and their labels and a set-level encoder E^{set}, with parameters φ^{set}, taking a set of instance representations as input.

In each iteration, we sample a batch $X = (x_1, \ldots, x_b)$ of datapoints from $X_E = X_A$ and a transformation $T = (T^X, T^Y)$ according to the fixed distribution $P(\boldsymbol{T})$. Let $Z = T^X(X) = (z_1, \ldots, z_b)$ denote the transformed batch and $Y = T^Y(LF(X)) = (y_1, \ldots, y_b)$ the encoded labels. The hidden representations of the raw data are computed as a two-step process:

1. using E^{ins}, we compute $H^X = (h_1^X, \ldots, h_b^X)$ where each $h_i^X = E^{\text{ins}}\left(x_i, LF(x_i)\right)$;
2. using E^{set}, we compute $R^X = (r_1^X, ..., r_b^X)$ where each $r_i^X = E^{\text{set}}\left(h_i^X, H^X\right)$.

Similarly, for the encoded data, we form $H^Z = (h_1^Z, \ldots, h_b^Z)$ where $h_i^Z = E^{\text{ins}}\left(z_i, y_i\right)$ and $R^Z = (r_1^Z, ..., r_b^Z)$ where each $r_i^Z = E^{\text{set}}\left(h_i^Z, H^Z\right)$.

Following prior work on contrastive estimation [22], we use the cosine distance between hidden representations to measure similarity:

$$\text{sim}(r_i^X, r_j^Z) = \frac{\left(r_i^X\right)^\top r_j^Z}{\|r_i^X\| \|r_j^Z\|} \ .$$

Then, we estimate the quantity $P((x_i, z_j) \in M_T)$ as proportional to $\hat{p}(x_i, z_j)$:

$$\hat{p}(x_i, z_j) = \frac{\exp(\text{sim}(r_i^X, r_j^Z))}{\sum_{k,l}^{b} \exp(\text{sim}(r_k^X, r_l^Z))} \ .$$

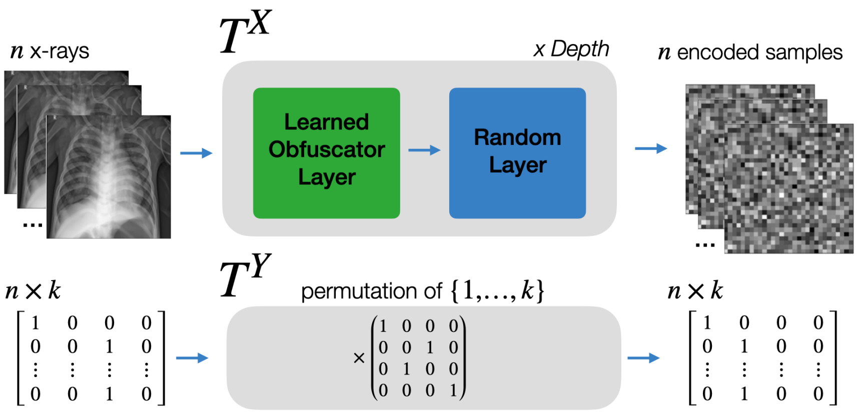

Figure 4-2: Proposed encoding scheme: Syfer uses repeating blocks of learned obfuscator layers and random neural network layers as T^X and samples a random permutation of $\{1, ..., k\}$ as T^Y.

The weights φ^{ins} and φ^{set} of the attacker's model E are trained to minimize the negative log-likelihood of re-identification across unknown T:

$$\mathcal{L}_{\text{reid}} = - \sum_{(x,z) \in M_T} \log \left(\hat{p}(x, z) \right) \ .$$

4.3.2 Syfer

Architecture As illustrated in Figure 4-2, we propose a new encoding scheme by learning to shape the distribution $P(\boldsymbol{T})$. Specifically, we parametrize a transformation T^X using a neural network that we decompose into blocks of learned *obfuscator* layers (weights θ_{Syfer}), and random layers (weights θ_{key}). The *obfuscator* layers are trained to leverage the randomness of the subsequent random layers and learn a distribution $P(\boldsymbol{T})$ that achieves privacy. In this framework, Alice constructs T^X by randomly sampling the weights θ_{key} and composing them with pre-trained obfuscator weights θ_{Syfer} to encode the raw data X. Alice chooses the label encoding T^Y by randomly sampling a permutation of the label identities $\{1, \ldots, k\}$, which is applied to $LF(X)$. We note that our T^Y assumes that Alice's dataset is class-balanced[1].

Alice's random choices of θ_{key} and T^Y act as her private key, and she can publish

[1] If Alice's data is not class-balanced, she can down-sample her dataset to a class-balanced subset before release.

the encoded data with diagnosis labels for model development while being protected from re-identification attacks. Given Bob's trained classifier to infer $T^Y(LF(x))$ from $T^X(x)$, Alice then uses $\left(T^Y\right)^{-1}$ to decode the predictions.

Algorithm 1 *Syfer* training

```
Initialize obfuscator parameters $\theta_{\text{Syfer}}$
Initialize attacker $E$ with parameters $\varphi = (\varphi^{\text{ins}}, \varphi^{\text{set}})$
Initialize decoders $D_1, \dots D_s$ with parameters $\beta_1, \dots, \beta_s$
For each decoder, sample random layer weights
    $\theta^1_{key}, \dots \theta^s_{key}$ (fixed throughout training)
Set flag *optimize_ estimators* $\leftarrow$ true
**repeat**
    Sample a batch of datapoints $X$ from $X^{\text{public}}$
    lightgray▷ Step 1: Compute re-identification loss
    Sample a set of random layer weights $\theta^{\text{batch}}_{key}$
    Using obfuscator parameters $\theta_{\text{Syfer}}$ and key $\theta^{\text{batch}}_{key}$:
    $T^{\text{batch}} \leftarrow f(\theta_{\text{Syfer}}, \theta^{\text{batch}}_{key})$
    $\left(Z^{\text{batch}}, Y^{\text{batch}}\right) \leftarrow T^{\text{batch}}(X, LF(X))$
    $R^Z \leftarrow E_\varphi\left(Z^{\text{batch}}, Y^{\text{batch}}\right)$
    $R^X \leftarrow E_\varphi(X, LF(X))$
    $\mathcal{L}_{\text{reid}} \leftarrow \text{contrastive\_loss}\left(R^X, R^Z\right)$
    lightgray▷ Step 2: Compute reconstruction loss
    $\mathcal{L}_{\text{rec}} \leftarrow 0$
    **for** $i \in \{1, \dots s\}$ **do**
        Using obfuscator parameters $\theta_{\text{Syfer}}$ and fixed key $\theta^i_{key}$:
        $T^i \leftarrow f(\theta_{\text{Syfer}}, \theta^i_{key})$
        $\left(Z^i, Y^i\right) \leftarrow T^i(X, LF(X))$
        $\mathcal{L}_{\text{rec}} \leftarrow \mathcal{L}_{\text{rec}} + \text{MSE}\left(D_i\left(Z^i\right), X\right)$
    **end for**
    lightgray▷ Step 3: Alternatively update parameters
    **if** *optimize_ estimators* **then**
        $\varphi \leftarrow \varphi - \nabla_\varphi \mathcal{L}_{\text{reid}}$
        $\beta_i \leftarrow \beta_i - \nabla_{\beta_i} \mathcal{L}_{\text{rec}}$    {for $i \in \{1, \dots s\}$}
        *optimize_ estimators* $\leftarrow$ false
    **else**
        $\theta_{\text{Syfer}} \leftarrow \theta_{\text{Syfer}} - \nabla_{\theta_{\text{Syfer}}}(\lambda_{\text{rec}} \cdot \mathcal{L}_{\text{rec}} - \lambda_{\text{reid}} \cdot \mathcal{L}_{\text{reid}})$
        *optimize_ estimators* $\leftarrow$ true
    **end if**
**until** convergence
```

Training Data owners may not have the computational capacity to train their own obfuscator layers, so we train Syfer without direct knowledge of X_A or LF. Instead, we rely on a public dataset X^{public} and use the null labeling function $LF(x) = 0$. To be successful, Syfer needs to generalize to held-out datasets, prediction tasks and attackers.

As shown in Algorithm 1, we train Syfer's obfuscator layers (parameters θ_{Syfer}) to maximize the loss of an attacker E (parameters $\varphi = (\varphi^{\text{ins}}, \varphi^{\text{set}})$) and to minimize the reconstruction loss of an ensemble of decoders $D_1, \ldots, D_s$ (parameters $\beta_1, \ldots \beta_s$).

At each step of training, we sample a transformation T^{batch} by choosing a new $\theta_{key}^{\text{batch}}$ to combine with the current θ_{Syfer} (Alg. 1, L.9-11). Using the current attacker weights φ, we then compute the re-identification loss (Alg. 1, L.13-15) as:

$$\mathcal{L}_{\text{reid}} = - \sum_{(x,z) \in M_T} \log\left(\hat{p}(x, z)\right)$$

Next, we estimate the overall invertability of the encoding scheme by measuring the reconstruction loss of an ensemble of decoders $D_1, \ldots, D_s$. For each each decoder D_i, we randomly sample a private key θ_{key}^{i}, which is fixed throughout the training algorithm. Each decoder D_i is trained to reconstruct X from $Z = T^i(X)$ where T^i is constructed by composing the current θ_{Syfer} with θ_{key}^{i}. We update β_i to minimize the reconstruction loss (Alg. 1, L.17-23):

$$\mathcal{L}_{\text{rec}} = \sum_{i=1}^{s} \left(E_X[||x - D_i \circ T^i(x)||^2] \right)$$

We train our attacker and decoders in alternating fashion with Syfer's obfuscator parameters. On even steps, Syfer's weights θ_{Syfer} are updated to minimize the loss:

$$\mathcal{L} = \lambda_{\text{rec}} \cdot \mathcal{L}_{\text{rec}} - \lambda_{\text{reid}} \cdot \mathcal{L}_{\text{reid}}$$

On odd steps, the attacker and decoders are updated to minimize $\mathcal{L}_{\text{reid}}$ and $\mathcal{L}_{\text{rec}}$ respectively (Alg. 1, L. 25-32).

In this optimization, the tasks of our attacker and decoders are asymmetric: the attacker is trained to generalize across transformations T (i.e. θ_{key}), while the decoders only need to generalize to unseen images, for a fixed key θ_{key}.

4.4 Experiments

Datasets For all experiments, we utilized two benchmark datasets of chest X-rays, NIH [130] and MIMIC-CXR [58], from the National Institutes of Health Clinical Center and Beth Israel Deaconess Medical Center respectively. Both datasets were randomly split into 60,20,20 for training, development and testing, and all images were downsampled to 64x64 pixels. We leveraged the NIH dataset to train all private encoding schemes (i.e. Syfer and baselines), and we evaluated the privacy and utility of all encoding schemes on the MIMIC-CXR dataset, for the binary classification tasks ($k = 2$) of predicting Edema, Consolidation, Cardiomegaly, and Atelectasis. This reflects the intended use of the tool, where a hospital leverages a pretrained Syfer for their heldout datasets.

For privacy and utility experiments considering a specific diagnosis task, we used a filtered version of the MIMIC-CXR data with balanced labels and explicit negatives. Specifically, for each diagnosis tasks, we followed common practice [57] and excluded exams with an uncertain disease label, i.e., the clinical diagnosis did not explicitly rule out or confirm the disease. Then, we selected one random negative control case for each positive case in order to create a balanced dataset. Our dataset statistics are shown in Appendix A.0.2.

Syfer Implementation Details As shown in Figure 4-2, Syfer consists of repeated blocks of trained obfuscator layers and random neural network layers. Following prior work in vision transformers [154], Syfer operates at the level of patches of images. We used a patch size of 16x16 pixels and 5 Syfer blocks for all experiments. We implemented our trained obfuscator layers as Simple Attention Units (SAU), a gated multi-head self attention module. We implemented our random neural networks as linear layers, followed by a SeLU nonlinearity and layer normalization. All random linear layers weights were sampled from a unit Gaussian, and we used separate random networks per patch. Our full Syfer architecture has 12.9M parameters, of which 6.6M are learned obfuscator parameters and 6.3M are random neural layer parameters. The SAU module is detailed in Appendix A.0.3.

We trained Syfer for 50,000 steps on the NIH training set to maximize the re-identification loss and minimize the reconstruction loss, with $\lambda_{\text{reid}} = 2$, $\lambda_{\text{rec}} = 20$. We trained our adversary and decoder for one step for each step of obfuscator training. We implemented the instance encoder E^{ins} and set encoder E^{set} of our adversary model as a depth 3 and depth 1 SAU respectively. We utilized separate E^{ins} networks to encode the raw data $(X, LF(X))$ and encoded data (Z, Y). We use a single decoder[2] D_1 (i.e. $s = 1$) and implement it as a depth 3 SAU. We used a batch size of 128, the Adam optimizer and a learning rate of 0.001 for training Syfer and our estimators. The training of Syfer is fully reproducible in our code release.

Privacy Estimators To evaluate the ability of Syfer to defend against re-identification attacks, we trained attackers to re-identify raw images from Syfer encodings on the MIMIC-CXR dataset. Since we cannot bound the prior knowledge the attacker may have over X_A, we consider the extreme case and train our attackers on their evaluation set, i.e. we only use MIMIC-CXR's training set for privacy evaluation. As a result, the attacker does not have to generalize to held-out images, but only to held-out private encoders T.

As described in Section 4.3.1, the attacker is trained to re-identify raw images from encoded images across new unobserved private keys using an image encoder E^{ins} and a set encoder E^{set}. This attacker estimates $P((x, z) \in M)$ for an encoding scheme Γ on a dataset X. Across our experiments, we implemented E^{ins} as either a ResNet-18 [53], a ViT [154], or a SAU. We implemented E^{set} as a depth 1 SAU. All attackers were trained for 500 epochs.

We computed the guesswork of each attacker by sorting the scores $\hat{p}(x, z)$ and identifying the index of the first correct correspondence. To measure the attackers average performance, we also evaluated the ROC AUC of the attacker attempting to predict an (x, z) matching as a binary classification task. A higher guesswork and lower re-identification AUC (ReID AUC) reflect a more private encoding scheme.

[2] Using an ensemble of $s = 5$ decoders did not significantly improve downstream utility.

Encoding	Guesswork	ReId AUC
Dauntless	1.0 $(1, 1)$	1.00 $(1.00, 1.00)$
InstaHide	1.0 $(1, 1)$	1.00 $(1.00, 1.00)$
DP-S, $b = 10$	1.2 $(1, 2)$	0.98 $(0.98, 0.98)$
DP-S, $b = 20$	7.2 $(1, 31)$	0.86 $(0.85, 0.86)$
DP-S, $b = 30$	68 $(1, 205)$	0.70 $(0.70, 0.70)$
DP-I, $b = 1$	5.0 $(1, 17)$	0.89 $(0.88, 0.89)$
DP-I, $b = 3$	77 $(3, 276)$	0.73 $(0.73, 0.73)$
DP-I, $b = 5$	1379 $(49, 4135)$	0.59 $(0.59, 0.60)$
Syfer-Random	1.7 $(1, 4)$	0.99 $(0.99, 0.99)$
Syfer (T^X only)	8476 $(1971, 20225)$	0.50 $(0.49, 0.52)$

Table 4.1: Privacy evaluation of different encoding schemes against an SAU based attacker on the unlabeled MIMIC-CXR dataset. For Syfer, only T^X is used. DP-S and DP-I stand for DP-Simple and DP-Image respectively. The scale parameter b characterizes the laplacian noise. Metrics are averages over 100 trials using 10,000 samples each, followed by 95% confidence intervals (CI).

Attacker	Guesswork	ReId AUC
SAU	8476 $(1971, 20225)$	0.50 $(0.49, 0.52)$
ViT	8411 $(5219, 12033)$	0.50 $(0.49, 0.51)$
Resnet-18	10070 $(9871, 10300)$	0.50 $(0.47, 0.53)$

Table 4.2: Privacy evaluation of Syfer across different attacker architectures on the unlabeled MIMIC-CXR dataset. Metrics are averages over 100 trials using 10,000 samples each, followed by 95% CI.

Diagnosis	Guesswork	ReId AUC
	Syfer	
Edema	3617 $(94, 11544)$	0.50 $(0.49, 0.51)$
Consolidation	1697 $(83, 5297)$	0.55 $(0.53, 0.57)$
Cardiomegaly	9834 $(2072, 15766)$	0.51 $(0.49, 0.53)$
Atelectasis	13189 $(2511, 28171)$	0.50 $(0.48, 0.52)$
Ablation: Syfer with no label encoding (T^X only)		
Edema	47 $(12, 83)$	0.76 $(0.76, 0.76)$
Consolidation	36 $(2, 104)$	0.76 $(0.76, 0.76)$
Cardiomegaly	42 $(17, 57)$	0.75 $(0.75, 0.75)$
Atelectasis	80 $(65, 98)$	0.75 $(0.75, 0.75)$

Table 4.3: Privacy evaluation of Syfer when released with different diagnoses in MIMIC-CXR dataset. Metrics are averages over 100 trials using 10,000 samples each, followed by 95% CI.

Generalized Privacy We first evaluated the guesswork and re-identification AUC (ReID AUC) of attackers trained using only encoded images (i.e. without labels) on

Encoding	E	Co	Ca	A	*Avg*
Using raw data	0.91	0.78	0.89	0.85	0.86
Using encoded data					
DP-S, $b = 10$	0.51	0.51	0.52	0.52	0.52
DP-S, $b = 20$	0.50	0.50	0.50	0.50	0.50
DP-S, $b = 30$	0.49	0.49	0.50	0.51	0.50
DP-I, $b = 1$	0.60	0.59	0.60	0.59	0.60
DP-I, $b = 2$	0.54	0.50	0.55	0.55	0.54
DP-I, $b = 5$	0.53	0.55	0.51	0.52	0.53
Syfer-Random	0.89	0.75	0.86	0.84	0.84
Syfer	0.82	0.69	0.81	0.78	0.78

Table 4.4: Utility for chest X-ray prediction tasks across different encoding schemes. All metrics are ROC AUCs across the MIMIC-CXR test set. Guides of abbreviations for medical diagnosis: (E)dema, (Co)nsolidation, (Ca)rdiomegaly and (A)telectasis.

the entire unfiltered MIMIC-CXR training set. For Syfer, this only requires using the neural encoder T^X. We compared Syfer to prior lightweight encoding schemes, including InstaHide [55] and Dauntless [139, 138]; and differential privacy methods, like DP-Image [69]. We now detail our baseline implementations.

- To assess the value of training Syfer's obfuscator layers, we compared Syfer to an ablation with randomly initialized obfuscator layers, Syfer-Random.
- InstaHide randomly mixes each private image with 2 other private images (i.e with MixUp [152]) and then randomly flips each pixel sign.
- Dauntless [138] applies a separate random linear layer to each 16x16 pixel patch of the images, with each random weight initialized as according to a standard Gaussian distribution.
- DP-Simple adds independent laplacian noise to each pixel of the image to obtain differential privacy. We evaluated using a scale (or diversity parameter) b of 10.0, 20.0 and 30.0.
- DP-Image [69] adds independent laplacian noise to the latent space of an auto-encoder to produce differentially private images. Our auto-encoder architecture is further detailed in Appendix A.0.4. We trained our auto-encoder on the NIH

dataset and applied it with laplacian noise on the MIMIC-CXR dataset. We evaluated using a scale b of 1.0, 2.0 and 5.0.

We report the expected guesswork and AUC for each attack as well as 95% confidence intervals (CI). To compute confidence intervals, we sampled 100 bootstrap samples of 10,000 images (all encoded by a single T) from the MIMIC-CXR training set. Our 100 bootstraps consisted of 10 random data samples (of 10,000) across 10 random T.

Privacy with Real Labeling Functions In practice, the encoded images are released with encoded labels to enable model development on tasks of interest. Using this additional knowledge, attackers may be able to better re-identify private data. To evaluate the privacy of Syfer encodings when released with public labels, we trained the attackers to re-identify raw images given access to (raw image, raw label) pairs and (obfuscated image, obfuscated label) pairs. To highlight the importance of Syfer's label encoding scheme T^Y in this scenario, we also train attackers on an ablation of Syfer which does not encode the labels and releases (obfuscated image, raw label). This corresponds to using only Syfer's neural encoder T^X.

We performed this attack independently per diagnosis. We implemented the instance encoder E^{ins} of our attacker as an SAU, our self-attention module, and represented the disease label an additional learned 256 dimensional input token for E^{ins}. As before, our attackers were trained for 500 epochs, and evaluated on the MIMIC-CXR training set. We report the expected guesswork and AUC for each attack as well as 95% confidence intervals. To compute confidence intervals, we sampled 100 random T and encoded the whole class-balanced MIMIC-CXR training set for each sampled T.

Utility Evaluation We evaluated the utility of an encoding scheme on the MIMIC-CXR dataset by measuring the ROC-AUC of diagnosis models trained using its encodings. We compared the utility of Syfer to a plaintext baseline (i.e. using raw data), which provides us with a utility upper bound. To isolate the impact of training

Syfer's obfuscator layers on utility, we also compared the utility of Syfer to Syfer-Random. We also computed the utility of our differential privacy baselines, DP-Simple with a scale parameter b of 10, 20 and 30 and DP-Image with a scale of 1, 2 and 5. For each encoding scheme, we experimented with different classifier architectures (e.g. SAU vs ResNet-18), dropout rates and weight decay, and selected the architecture that achieved the best validation AUC.

4.5 Results

Generalized Privacy We report our generalized privacy results, which consider re-identification attacks on the unlabeled MIMIC-CXR dataset, in Table B.19 and Table B.22, with higher guesswork and lower ReID AUC denoting increased privacy. While Syfer was trained to maintain privacy against an SAU-based attacker on the NIH training set, we found that its privacy generalized to a held-out dataset, MIMIC-CXR, and held-out attack architectures (e.g. ResNet-18 and ViT). Syfer obtained a guesswork of 8411 (95% CI 5219, 12033) and an ReId AUC of 0.50 (95% CI 0.49, 0.51) against a ViT attacker. We note that a guesswork of 10,000 corresponds to guessing randomly in this evaluation. In contrast, the InstaHide and Dauntless baselines could not defend against re-identification attacks obtaining both a guesswork of 1.0 (95% CI 1, 1). As illustrated in Appendix A.0.5, the differential privacy baselines can obtain privacy at the cost of significant image distortion. DP-Image with a laplacian noise scale of 5.0 obtained a guesswork of 1379 (95% CI 49, 4135) and an attacker AUC of 0.59 (95% CI 0.59, 0.60).

Privacy with Real Labeling Functions We evaluated the privacy of releasing Syfer encodings with different public labels in Table B.21. Releasing *raw labels* resulted in significant privacy leakage with guessworks ranging from 36 (95% CI 2, 104) to 80 (95% CI 65, 98) for Consolidation and Atelectasis respectively. In contrast, when labels are protected using Syfer's label encoding scheme and released alongside the image encodings, Syfer maintains privacy across all diagnoses tasks, with guessworks ranging from 1697 (95% CI 83, 5297) to 13189 (95% CI 2511, 28171) for Consolidation and Atelectasis respectively.

Utility Evaluation We report our results in predicting various medical diagnoses from X-rays in Table B.20. Models built on Syfer obtained an average AUC of 0.78, compared to 0.86 by the plaintext baseline and 0.84 by the Syfer-Random baseline. In contrast, the best differential privacy baseline, Image-DP, obtained average AUCs of 0.60, 0.54 and 0.53 when using a scale of 1 and 2 and 5 respectively. Syfer obtained a 25 point average AUC improvement over DP-Image while obtaining better privacy.

4.6 Discussion

We propose Syfer, an encoding scheme for releasing private data for machine learning model development while preventing raw data re-identification. Syfer uses trained obfuscator layers and random neural networks to minimize the likelihood of re-identification, while encouraging the invertability of the overall transformation. In experiments on MIMIC-CXR, a large chest X-ray benchmark, we show that Syfer obtains strong privacy across held-out attackers, obtaining an average guesswork of 8411, whereas prior encoding schemes like Dauntless [139], InstaHide [55] did not meet our privacy standard, obtaining guessworks of 1. While differential privacy baselines can achieve privacy with enough noise, we found this came with a massive loss of utility, with DP-Image obtaining an average AUC of 0.53 for a guesswork of 1379. In contrast, models built on Syfer encodings approached the utility of our plaintext baseline, obtaining an average AUC of 0.78 compared to 0.86 by the plaintext model.

Future Work While our threat model considers a computationally unbounded adversary, in practice, we rely on model-based attackers for both the development and evaluation of Syfer. More powerful models may result in more successful attacks on Syfer. As a result, continued research into re-identification algorithms is needed to offer stronger theoretical guarantees and develop more powerful encodings. Moreover, while we show that Syfer generalizes to an unseen datasets, this does not guarantee that it will generalize to arbitrary datasets. Additional research studying the privacy impact of domain shifts is also necessary.

Appendix A

Syfer: Supplementary Materials

A.0.1 Guesswork Supplementary Details

Recall that for an ordered list of mn correspondence guesses $(u_1, \ldots, u_{mn})$, where $u_i \in X_E \times Z_A$, the guesswork is defined as the rank of the first correct guess: $\mathcal{G} = \min_k\{k \text{ s.t. } u_k \in M_T\}$, where $M_T = \{(x, z) \in X_E \times Z_A \text{ s.t. } z = T^X(x)\}$. In the event of ties, the guesswork is computed as the expected value over permutations of the suitable subsets. In the paper, we use $X_E = X_A$ but the guesswork can be computed for an arbitrary superset $X_E \supseteq X_A$ of size m.

Guesswork Algorithm We propose the following algorithm to compute the guesswork for a given probability matrix and set of correct guesses.

Algorithm 2 Guesswork algorithm

```
Input Correct matching $M_T = \{(x_i, z_j) \text{ s.t. } z_j = T^X(x_i)\}$
Input Probability matrix $A$ where $A_{i,j} = P((x_i, z_j) \in M)$
Output Guesswork $\mathcal{G}$ for $A$
From $A$, extract
    $S = \{(i, j, A_{i,j}) \text{ for } 1 \leq i \leq m, 1 \leq j \leq n\}$
Partition $S$:
    $S = \bigcup_p S_p$ where $S_p = \{(i, j, A_{i,j}) \text{ s.t. } A_{i,j} = p\}$
Find the highest value of $p$ such that $A_p$ contains matches:
    $q = \max_p\{p \text{ s.t. } \exists(i, j, A_{i,j}) \in S_p \text{ s.t. } (x_i, z_j) \in M_T\}$
$\mathcal{G} \leftarrow 0$
for $p > q$ do
    $\mathcal{G} \leftarrow \mathcal{G} + |A_p|$
end for
$\mathcal{G} \leftarrow \mathcal{G} + \frac{1+|A_q|}{1+|A_q \cap M|}$
return $\mathcal{G}$
```

The expression $\frac{1+|A_q|}{1+|A_q \cap M|}$ is derived by computing the expected value of number of trials before success in the urn problem without replacement.

[guesswork calculation extended] Let $X_E = X_A = \{1, 2\}$ and disregard labels for now. Consider three transformations $X_A \to \{a, b, c, d\}$:

$$T_1 : \begin{cases} 1 & \mapsto a \\ 2 & \mapsto b \end{cases} \quad T_2 : \begin{cases} 1 & \mapsto b \\ 2 & \mapsto a \end{cases} \quad T_3 : \begin{cases} 1 & \mapsto c \\ 2 & \mapsto d \end{cases}$$

We evaluate the following encoding schemes, defined by the distribution used to sample T:

$\Gamma_1 : \quad P(T_1) = 2/3, \; P(T_2) = 1/3, \; P(T_3) = 0$

$\Gamma_2 : \quad P(T_1) = 1/2, \; P(T_2) = 1/2, \; P(T_3) = 0$

$\Gamma_3 : \quad P(T_1) = 1/3, \; P(T_2) = 1/3, \; P(T_3) = 1/3.$

For Γ_1, Eve observes $Z_A = \{a, b\}$ regardless of the choice of T_A. Given her knowledge of $P(\boldsymbol{T})$, she elects to rank $\{(1, a), (2, b)\}$ before $\{(1, b), (2, a))\}$ which gives guessworks $\mathcal{G}(T_1) = 1$ and $\mathcal{G}(T_2) = 3$. In expectation, the guesswork of Γ_1 is $5/3$ (with a variance of $8/27$).

For Γ_2, Eve observes $Z_A = \{a, b\}$ as well, but equally ranks all 4! orderings of the guesses $((1, a), (1, b), (2, a), (2, b))$, which leads to the same guesswork for both T:

$\mathcal{G}(T_1) = \mathcal{G}(T_2) = \frac{1}{2} \cdot 1 + \frac{1}{2} \cdot \frac{2}{3} \cdot 2 + \frac{1}{2} \cdot \frac{1}{3} \cdot 3 = \frac{5}{3}$ (no variance).

For Γ_3, whenever Eve observes $Z_A = \{c, d\}$, she deduces that $T_A = T_3$, which leads to a guesswork of 1. In the other cases, observing $Z_A = \{a, b\}$ means that T_1 and T_2 are equally likely, so the guesswork is 5/3. In expectation, the guesswork is 13/9, which is lower (and thus worse privacy) than the previous schemes.

Guessworks in Special Cases We discuss two special cases that arise when computing guesswork.

1. If all guesses in the bucket A_p of highest probability are correct guesses, then the guesswork is 1, characterizing a non-private scheme. Note that this does not depend on the cardinal of the bucket A_p of highest probability: regardless of whether the attacker is confidently correct about one matching pair or multiple matching pair, the guesswork will still be 1.

2. If the probability matrix is uniform (i.e. there is p for which $S = S_p$, such that all guesses are in the same bucket), then the guesswork is $\frac{mn+1}{n+1}$, i.e. $\mathcal{G} \approx |X_E|$. This characterizes an attacker that fails to capture any privacy leakage of the encoding scheme.

Note that $|X_E|$ is not an upper-bound of guesswork. An attacker that is confidently wrong can achieve a guesswork up to $mn - n + 1$.

Discussion on Eve's strategy In our definition of guesswork, Eve commits to a probability matrix $A_{i,j} = P((x_i, z_j) \in M)$, then enumerates her guesses in descending order of likeliness. This would not be the optimal strategy for an attacker who wishes to minimize the number of guesses required to identify a correct match. For instance, if $P((x_i, z_j)$ is uniform, Eve could commit to a single column (or row) and achieve an expected number of guesses of $m/2$ (or $n/2$). More generally, after Eve made her first guess u_1, she can assume the first guess was incorrect and recompute the new probability matrix $P((x_i, z_j) \in M | u_1 \notin M)$, then proceed with subsequent guesses. Such an auto-regressive strategy is costly to implement. In practice, Eve also would not have access to an oracle that notifies her when a guess is correct. Therefore, we adopt the definition of guesswork exposed in Section **??** as an efficient way to universally compare the privacy of different encoding schemes.

A.0.2 Dataset Statistics

We leveraged the NIH training set for training Syfer, and leveraged the unlabeled MIMIC-CXR training set for all generalized privacy evaluation. To evaluate utility and privacy with real labeling functions, we use the labeled subsets of the MIMIC-CXR dataset. The labeled MIMIC-CXR training and validation sets were filtered to be class balanced, by assigning random one negative control for each positive sample. The number of images per dataset is shown in Table A.0.2

Dataset	Train	Dev	Test
	Unlabeled		
NIH	40365	NA	NA
MIMIC-CXR	57696	NA	NA
	Labeled		
MIMIC-CXR E	3660	1182	12125
MIMIC-CXR Co	1120	375	11031
MIMIC-CXR Ca	11724	3876	12791
MIMIC-CXR A	2164	3992	12129

Table A.1: Dataset statistics for all datasets. The training and development sets of MIMIC CXR Edema, Consolidation, Cardiomegaly and Atelecatasis were filtered to contain one negative control for each positive sample. Guides of abbreviations for medical diagnosis: (E)dema, (Co)nsolidation, (Ca)rdiomegaly and (A)telectasis.

A.0.3 SAU: Simple Attention Unit

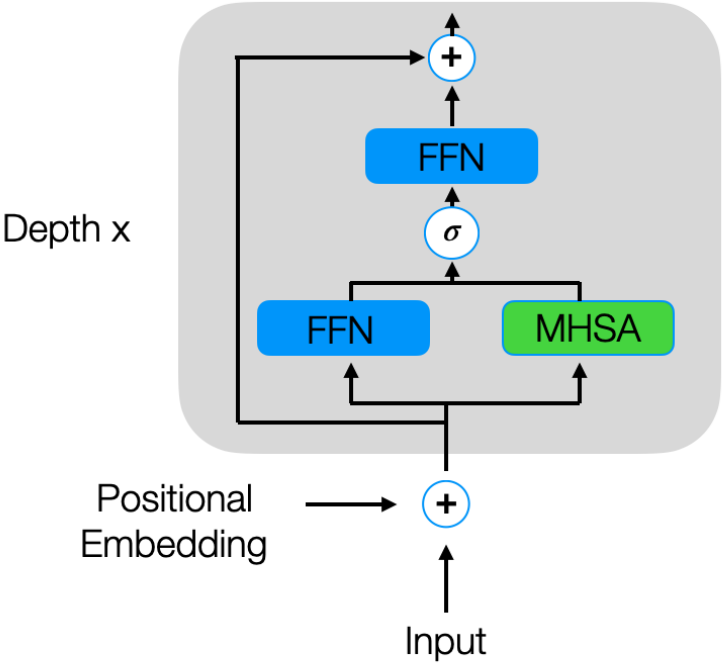

Figure A-1: Simple Attention Unit Architecture. The module uses a learnable gate at each layer to interpolate between leveraging behaving as a feed forward network (FFN) and a multi-headed self attention network (MHSA).

Our Simple Attention Unit (SAU), illustrated in Figure A-1, utilizes a learned gate, α, at each layer to interpolate between acting as a standard feed forward network (FFN) with no attention computation, and a multi-head self-attention (MHSA) network. We found that this allowed for faster and more stable training compared to ViTs[154, 37] in both privacy and utility experiments. To encode patch positions, we leverage a learned positional embedding for each location, following prior work [154, 37]. Each layer of the SAU is composed of the following operations:

$$x_{norm} = \text{BatchNorm}(x)$$

$$h_{ffn} = \text{SELU}(W_{in}x_{norm} + b_{in})$$

$$h_{attn} = \text{MHSA}(x_{norm})$$

$$h = \sigma(\alpha) \times h_{attn} + (1 - \sigma(\alpha)) \times h_{ffn}$$

$$o = \text{SELU}(W_o h + b_o) + x_{norm}$$

Where Multi-head self-attention (MHSA) is defined as:

$$K_i, Q_i, V_i = W_i x_{norm}$$

$$\text{head}_i = softmax(\frac{Q_i K_i^T}{\sqrt{d_k}})V^i$$

$$h_{attn} = \text{BatchNorm}(Concat(head_1, ..., head_h)W_{attn})$$

Where d_k is the dimension of each head, all W and b are learned parameters, and α is a learned gate. α is initialized at -2 for each layer.

A.0.4 DP-Image Baseline

DP-Image[69] is a differential privacy method based adding laplacian noise to the latent space of auto encoders to achieve differential privacy. We trained our auto-encoder on the NIH training set, with no noise, and apply it with noise on the MIMIC dataset. Our encoder, mapping each 64x64 pixel image x to a 256d latent code z, is composed of six convolutional layers, each followed by a leaky relu activation and batch normalization. Each convolutional layer had a kernel size of 3, a stride of 2. This was then reduced a single code z with global average pooling. Our decoder, which mapped z back to x, consisted of six transposed convolutional layers, each followed by a leaky relu activation and batch normalization. The auto encoder was trained to minimize the mean squared error between the decoded image and the original image.

A.0.5 Visualizations

In Figure A-2, we visualize the impact of Syfer and DP-Image encodings when using different amounts of noise (parametrized by the diversity parameter b). Each row represents a different image. The *raw_ x* column are raw images. The Syfer column shows encodings obtained when applying Syfer's neural encoder for a specific choice of private weights θ_{key}: those are representative of the released images. We then train a decoder D_T for a specific choice of private weights θ_{key}. During training, the decoder has access to parallel data (raw image, encoded Syfer image). We visualize the decoded images in the *Syfer decoded* column. Note that in our scenario, only Alice would be able to train such a decoder: Bob and Eve only have access to encoded images with labels. The *DP-image no noise* column is the reconstructed image obtained with the trained auto-encoder that is used for the DP-Image baseline. We also visualize the reconstructed images when varying amounts of noise are added.

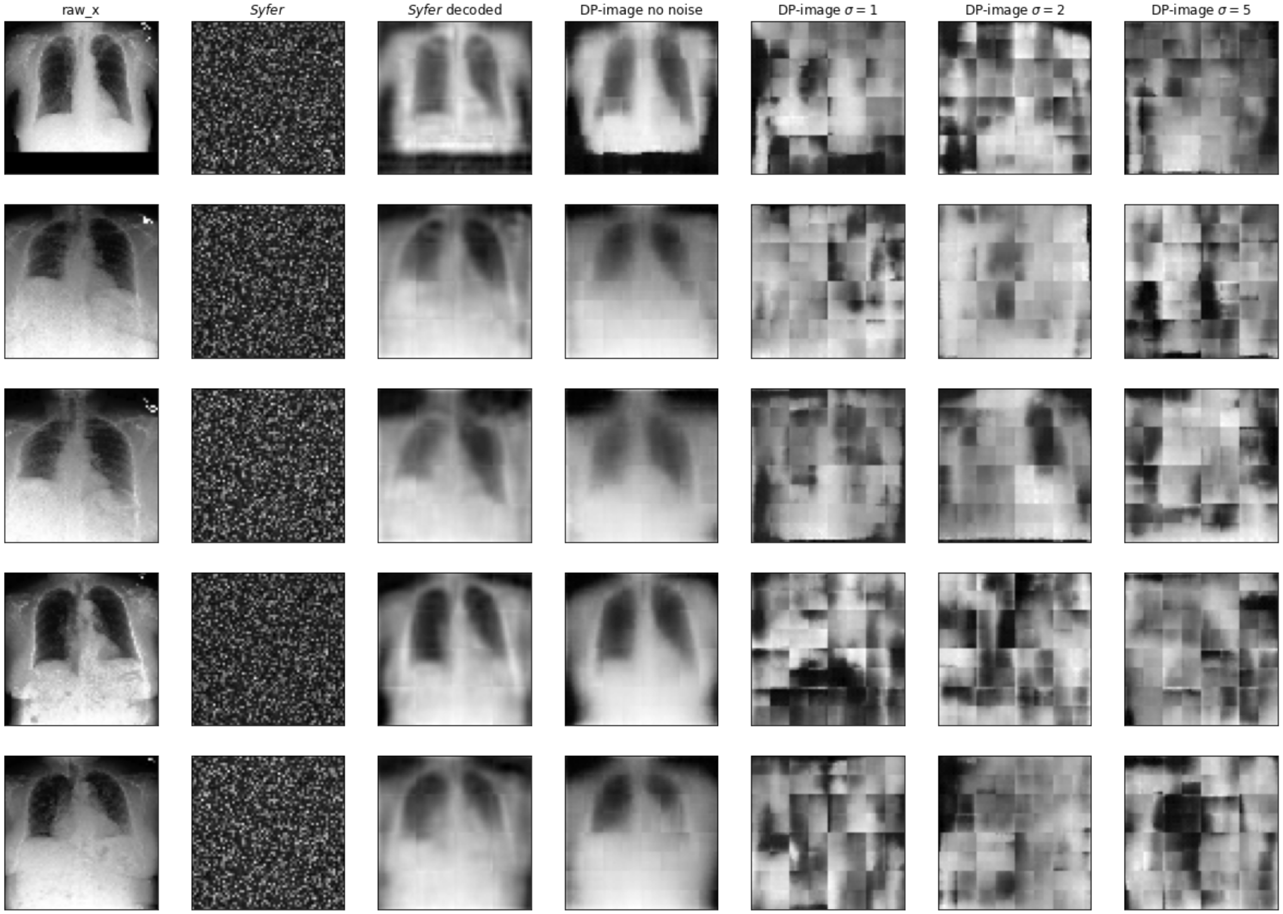

Figure A-2: Vizualisations of raw images, Syfer encodings, decoded Syfer encodings, and DP-Image encodings. Syfer encodings were obtained after applying the T^X part of Syfer to raw images. Decoded Syfer encodings are obtained by a model D_T trained on a set of parallel training data (plaintext attack). DP-Image encodings are shown with varying amount of noise.

A.0.6 Additional Privacy Analyses

Patch size	Guesswork	ReId AUC
	Syfer	
32px	102795 (25221, 235114)	50 (50, 50)
16px	12715 (4670, 31748)	50 (46, 54)

Table A.2: Privacy evaluation of Syfer against attackers attempting to re-identify patches of the encoded images. Guesswork was computed over a random subset of 10,000 samples. All metrics are followed by 95% confidence intervals.

A.0.7 Additional Utility Analyses

In Figure A-3, we plot the learning curves of Syfer, Syfer-Random and our plaintext baselines when training on fractions $\frac{1}{32}$, $\frac{1}{16}$, $\frac{1}{8}$, $\frac{1}{4}$, $\frac{1}{2}$ and 1 of the data.

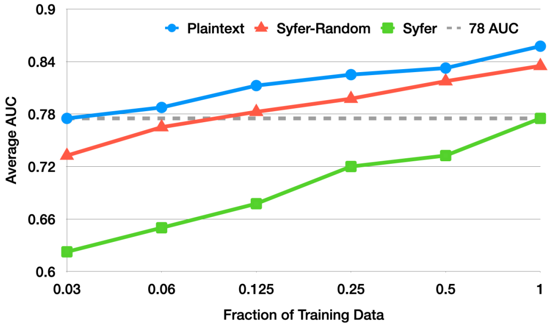

Figure A-3: Average AUC on MIMIC-Test set when training with different fractions of the data when using Syfer, Syfer-Random, and Plaintext encodings.

We find that it takes plaintext models $\frac{1}{2}$ of the training data to reach the full performance of Syfer-Random, indicating that using a random Syfer architecture harms sample complexity. Syfer, which achieves strong privacy, requires more data to achieve the same utility, with Syfer-Random achieving the same average AUC when using less than $\frac{1}{8}$ of the data and Plaintext achieving the same performance when using $\frac{1}{32}$.

Appendix B

Tables

Model	Use RF	C-Index	1-Year AUC	2-Year AUC	3-Year AUC	4-Year AUC	5-Year AUC
TCv8	Yes	0.62 (0.58, 0.67)	0.65 (0.51, 0.79)	0.64 (0.59, 0.70)	0.63 (0.58, 0.67)	0.62 (0.58, 0.66)	0.62 (0.57, 0.66)
Radiologist BI-RADs	NA	0.53 (0.50, 0.55)	0.74 (0.63, 0.86)	0.55 (0.51, 0.58)	0.53 (0.50, 0.55)	0.52 (0.50, 0.54)	0.52 (0.50, 0.53)
Image-And-Heathmaps	No	0.63 (0.59, 0.67)	0.71 (0.61, 0.84)	0.68 (0.63, 0.73)	0.64 (0.60, 0.69)	0.62 (0.58, 0.66)	0.59 (0.55, 0.63)
Image-Only DL	No	0.67 (0.64, 0.71)	0.64 (0.53, 0.76)	0.67 (0.62, 0.73)	0.68 (0.64, 0.72)	0.68 (0.65, 0.73)	0.70 (0.66, 0.73)
Hybrid DL	Yes	0.67 (0.63, 0.71)	0.63 (0.51, 0.76)	0.68 (0.63, 0.73)	0.67 (0.62, 0.71)	0.67 (0.63, 0.72)	0.69 (0.65, 0.73)
Mirai	No	0.69 (0.66, 0.73)	0.71 (0.60, 0.84)	0.71 (0.66, 0.76)	0.71 (0.67, 0.75)	0.71 (0.67, 0.75)	0.71 (0.68, 0.75)
	Yes	0.70 (0.66, 0.74)	0.72 (0.61, 0.84)	0.72 (0.67, 0.78)	0.72 (0.68, 0.76)	0.71 (0.68, 0.75)	0.72 (0.68, 0.76)

Table B.1: ROC AUCs and C-indices for Mirai and prior risk models on the MGH test set excluding cancers confirmed within six months of the screening mammogram. We also evaluated Image-And-Heatmaps and radiologist BI-RADs assessments. RF refers to "risk factors". All metrics are followed by their 95% confidence interval.

	MGH Training Set		MGH Validation Set		MGH Test Set	
Characteristics	All (%)	Cancer (%)	All (%)	Cancer (%)	All (%)	Cancer (%)
All exams	210819 (100%)	5379 (100%)	25644 (100%)	612 (100%)	25855 (100%)	588 (100%)
Age						
<40	5812 (2.8%)	84 (1.6%)	711 (2.8%)	7 (1.1%)	724 (2.8%)	7 (1.1%)
40-50	55905 (26.5%)	1113 (20.7%)	6821 (26.6%)	142 (23.2%)	7025 (27.2%)	95 (16.2%)
50-60	63314 (30.0%)	1348 (25.1%)	7762 (30.3%)	166 (27.1%)	7829 (30.3%)	188 (32.0%)
60-70	54925 (26.1%)	1770 (32.9%)	6674 (26.0%)	179 (29.3%)	6708 (25.9%)	182 (31.0%)
70-80	25401 (12.0%)	816 (15.2%)	3037 (11.8%)	102 (16.7%)	3001 (11.6%)	94 (16.0%)
80<	5461 (2.6%)	248 (4.5%)	639 (2.5%)	16 (2.6%)	568 (2.2%)	22 (3.7%)
Density						
Almost entirely fatty	20411 (9.7%)	315 (5.9%)	2429 (9.5%)	53 (8.7%)	2474 (9.6%)	31 (5.3%)
Scattered areas of fibroglandular tissue	102112 (48.4%)	2623 (48.8%)	12519 (48.8%)	261 (42.7%)	12490 (48.3%)	264 (44.9%)
Heterogeneously dense	78892 (37.4%)	2196 (40.8%)	9461 (36.9%)	263 (43.0%)	9751 (37.7%)	271 (46.1%)
Extremely dense	9293 (4.4%)	242 (4.5%)	1225 (4.8%)	35 (5.7%)	1129 (4.4%)	22 (3.7%)
BI-RADS						
0 - Additional imaging needed	13810 (6.6%)	1579 (29.4%)	1686 (6.6%)	164 (26.8%)	1785 (6.9%)	186 (31.6%)
1-Negative or 2-Benign	196797 (93.3%)	3786 (70.4%)	23932 (93.3%)	447 (73.0%)	24043 (93.0%)	400 (68.0%)
Other	47 (0.02%)	9 (0.2%)	3 (0.01%)	1 (0.2%)	4 (0.01%)	1 (0.2%)
Race						
White	171509 (81.4%)	4646 (86.4%)	20710 (80.8%)	518 (84.6%)	21006 (81.2%)	512 (87.1%)
African American	9883 (4.7%)	209 (3.9%)	1209 (4.7%)	26 (4.3%)	1204 (4.7%)	21 (3.6%)
Asian or Pacific Islander	9477 (4.5%)	160 (3.0%)	1231 (4.8%)	17 (2.8%)	1238 (4.8%)	26 (4.4%)
Hispanic	2266 (1.1%)	63 (1.2%)	260 (1.0%)	5 (0.8%)	225 (0.9%)	6 (1.0%)
Other Race	11423 (5.4%)	138 (2.6%)	1439 (5.6%)	20 (3.3%)	1486 (5.7%)	15 (2.6%)
Device						
Lorad Selenia	81106 (38.5%)	2009 (37.4%)	9850 (38.4%)	216 (35.29%)	9937 (38.4%)	241 (41.0%)
Selenia Dimensions	129493 (61.4%)	3150 (58.6%)	15767 (61.5%)	369 (60.29%)	15882 (61.4%)	311 (52.9%)
Unknown	220 (0.1)%	220 (4.1%)	27 (0.1%)	27 (4.4%)	36 (0.1%)	36 (6.1%)

Table B.2: Detailed demographics for Massachusetts General Hospital dataset. For each demographic, we report the number of corresponding mammography exams the percentage of they constitute of the total. All cancer counts reflect cancer within five-years.

	Novant Dataset	
Characteristics	All	Cancer
All exams	14157 (100.0)	235 (100.0)
Age		
40-50	3917 (27.67)	53 (22.55)
50-60	5368 (37.92)	65 (27.66)
60-70	4872 (34.41)	117 (49.79)
Race		
White	10555 (74.56)	185 (78.72)
African American	2687 (18.98)	44 (18.72)
Asian	220 (1.55)	0 (0.0)
Hispanic	391 (2.76)	5 (2.13)
American Indian or Alaskan Native	28 (0.2)	0 (0.0)
Time to Cancer		
0-1 year	95 (0.67)	95 (40.43)
1-2 years	52 (0.37)	52 (22.13)
2-3 years	48 (0.34)	48 (20.43)
3-4 years	31 (0.22)	31 (13.19)
4-5 years	9 (0.06)	9 (3.83)

Table B.3: Detailed demographics of Novant test set

	Novant Dataset	
Characteristics	All	Cancer
All exams	14157 (100.0)	235 (100.0)
Age		
40-50	3917 (27.67)	53 (22.55)
50-60	5368 (37.92)	65 (27.66)
60-70	4872 (34.41)	117 (49.79)
Race		
White	10555 (74.56)	185 (78.72)
African American	2687 (18.98)	44 (18.72)
Asian	220 (1.55)	0 (0.0)
Hispanic	391 (2.76)	5 (2.13)
American Indian or Alaskan Native	28 (0.2)	0 (0.0)
Time to Cancer		
0-1 year	95 (0.67)	95 (40.43)
1-2 years	52 (0.37)	52 (22.13)
2-3 years	48 (0.34)	48 (20.43)
3-4 years	31 (0.22)	31 (13.19)
4-5 years	9 (0.06)	9 (3.83)

Table B.4: Detailed demographics of Emory test set

Site	C-Index	1-Year AUC	2-Year AUC	3-Year AUC	4-Year AUC	5-Year AUC
MGH, USA	0.69 (0.66, 0.73)	0.71 (0.60, 0.84)	0.71 (0.66, 0.76)	0.71 (0.67, 0.75)	0.71 (0.67, 0.75)	0.71 (0.68, 0.75)
Novant, USA	0.72 (0.66, 0.79)	NA	0.71 (0.63, 0.80)	0.73 (0.66, 0.80)	0.72 (0.65, 0.79)	0.72 (0.66, 0.79)
Emory, USA	0.69 (0.66, 0.72)	0.74 (0.66, 0.84)	0.71 (0.68, 0.75)	0.70 (0.67, 0.73)	0.71 (0.68, 0.74)	0.71 (0.68, 0.74)
Maccabi-Assuta, Israel	0.70 (0.64, 0.76)	NA	0.67 (0.53, 0.83)	0.72 (0.66, 0.79)	0.70 (0.63, 0.76)	0.68 (0.62, 0.74)
Karolinska, Sweden	0.71 (0.69, 0.74)	NA	0.72 (0.67, 0.77)	0.73 (0.71, 0.76)	0.73 (0.70, 0.75)	0.71 (0.69, 0.73)
CGMH, Taiwan	0.70 (0.66, 0.75)	0.84 (0.72, 0.99)	0.76 (0.68, 0.84)	0.71 (0.64, 0.77)	0.71 (0.66, 0.76)	0.70 (0.66, 0.75)
Barretos, Brazil	0.78 (0.74, 0.83)	0.87 (0.80, 0.94)	0.82 (0.76, 0.89)	0.81 (0.76, 0.87)	0.79 (0.74, 0.84)	0.75 (0.70, 0.80)

Table B.5: Area under the Receiver Operating Curve (AUCs) for predicting cancer in one to five years and Uno's C-index for Mirai on all test sets excluding cancers diagnosed with six months of the mammogram. All metrics are followed by their 95% confidence interval.

Model	TCv8	ImageOnly	Hybrid DL	Mirai without Risk Factors	Mirai with Risk Factors
Race					
African American	0.62 (0.44, 0.84)	0.72 (0.61, 0.89)	0.73 (0.59, 0.88)	0.72 (0.56, 0.89)	0.71 (0.55, 0.90)
Asian	0.54 (0.36, 0.75)	0.68 (0.53, 0.85)	0.67 (0.50, 0.85)	0.77 (0.64, 0.92)	0.80 (0.68, 0.95)
White	0.64 (0.60, 0.68)	0.73 (0.69, 0.76)	0.72 (0.68, 0.75)	0.75 (0.71, 0.78)	0.75 (0.72, 0.78)
Age					
<50	0.63 (0.56, 0.71)	0.66 (0.59, 0.74)	0.68 (0.60, 0.77)	0.71 (0.63, 0.78)	0.71 (0.55, 0.90)
50-70	0.64 (0.60, 0.69)	0.71 (0.67, 0.74)	0.71 (0.68, 0.75)	0.74 (0.71, 0.78)	0.80 (0.68, 0.95)
>70	0.54 (0.46, 0.62)	0.76 (0.69, 0.83)	0.71 (0.63, 0.89)	0.74 (0.67, 0.82)	0.75 (0.72, 0.78)
Density					
Non-Dense	0.63 (0.58, 0.68)	0.71 (0.67, 0.76)	0.70 (0.66, 0.75)	0.74 (0.70, 0.78)	0.75 (0.71, 0.79)
Dense	0.64 (0.59, 0.69)	0.73 (0.69, 0.77)	0.73 (0.69, 0.78)	0.76 (0.72, 0.80)	0.76 (0.72, 0.80)
Mammography Device					
Lorad Selenia	0.65 (0.61, 0.70)	0.71 (0.67, 0.75)	0.71 (0.67, 0.76)	0.73 (0.69, 0.77)	0.74 (0.68, 0.78)
Selenia Dimensions	0.62 (0.57, 0.67)	0.74 (0.71, 0.78)	0.73 (0.69, 0.77)	0.77 (0.74, 0.81)	0.78 (0.74, 0.82)

Table B.6: C-Index for different models on different sub-populations in the MGH test set. All metrics are followed by their 95% confidence interval.

Subtype	C-Index	1-year AUC	2-year AUC	3-year AUC	4-year AUC	5-year AUC
Invasive	0.80 (0.78, 0.82)	0.90 (0.88, 0.92)	0.85 (0.83, 0.87)	0.81 (0.79, 0.83)	0.8 (0.78, 0.82)	0.77 (0.75, 0.79)
DCIS	0.81 (0.79, 0.84)	0.92 (0.9, 0.94)	0.88 (0.85, 0.91)	0.83 (0.81, 0.86)	0.81 (0.79, 0.84)	0.78 (0.76, 0.81)
ER+	0.81 (0.79, 0.83)	0.91 (0.89, 0.93)	0.87 (0.85, 0.89)	0.82 (0.81, 0.84)	0.81 (0.79, 0.83)	0.78 (0.76, 0.81)
ER-	0.75 (0.70, 0.80)	0.87 (0.82, 0.94)	0.79 (0.73, 0.85)	0.76 (0.71, 0.82)	0.75 (0.69, 0.80)	0.73 (0.68, 0.78)
PR+	0.80 (0.78, 0.82)	0.9 (0.88, 0.93)	0.86 (0.83, 0.88)	0.81 (0.79, 0.84)	0.8 (0.78, 0.82)	0.78 (0.75, 0.80)
PR-	0.81 (0.78, 0.84)	0.9 (0.87, 0.94)	0.86 (0.82, 0.90)	0.83 (0.79, 0.86)	0.81 (0.78, 0.85)	0.78 (0.74, 0.81)
HER2+	0.79 (0.75, 0.84)	0.92 (0.87, 0.97)	0.87 (0.82, 0.93)	0.83 (0.78, 0.88)	0.79 (0.74, 0.85)	0.75 (0.70, 0.81)
HER2-	0.81 (0.79, 0.83)	0.9 (0.88, 0.92)	0.86 (0.83, 0.88)	0.82 (0.80, 0.84)	0.81 (0.79, 0.83)	0.78 (0.76, 0.81)

Table B.7: C-Indices and ROC AUCs for Mirai in predicting cancers of different subtypes in the Karolinska test set. For each row in the table, we evaluate the ability of the model to discriminate between patients who developed the specific subtype of cancer (e.g., HER2-) from those who did not develop cancer. All metrics are followed by their 95% confidence interval.

Cancer Type	Number of Exams
Invasive	1243
DCIS	760
ER+	1093
ER-	183
PR+	934
PR-	341
HER2+	156
HER2-	884

Table B.8: Number of exams per cancer type in the Karolinska Dataset.

Method	Sensitivity	Specificity
African American: 4,311 exams from 2,831 patients. 301 exams followed by future cancer.		
Mirai at TC specificity	33.9% (26.3, 41.0)	83.7% (82.3, 85.1)
Mirai at TC sensitivity	22.9% (16.2%, 29.3)	90.7% (89.6, 91.9)
White: 3,728 exams from 2,501 patients. 306 exams followed by future cancer		
Mirai at TC specificity	40.0% (32.0, 47.2)	85.5% (84.0, 87.0)
Mirai at TC sensitivity	20.6% (14.6, 26.2)	91.9% (90.8, 93.0)

Table B.9: High risk cohort analysis for subgroups of the Emory dataset by race. We restricted our analysis to patients who were initially screening negative and had at least five-years of screening follow-up. We defined an exam as screening negative if it was not followed by a cancer diagnosis within six months. We defined a future cancer as a pathology confirmed breast cancer diagnosis within five years of the mammogram. Mirai thresholds (i.e. "at TC sensitivity" and "at TC specificity") were chosen to match the performance of the Tyrer-Cuzick (TC) model on the development MGH set

Model	Use Risk Factors	MGH Validation Set C-Index	MGH Test Set C-Index	Device-Identity Classifier AUC on MGH Test Set
TCv8	Yes	0.63 (0.59, 0.67)	0.64 (0.60, 0.67)	0.50 (0.50, 0.50)
ImageOnly DL	No	0.69 (0.66, 0.73)	0.72 (0.69, 0.75)	0.51 (0.50, 0.51)
Hybrid DL	Yes	0.71 (0.68, 0.75)	0.72 (0.69, 0.75)	0.50 (0.50, 0.50)
Image Encoder + Cox Proportional Hazard Layer	No	0.64 (0.60, 0.67)	0.63 (0.60, 0.67)	0.74 (0.73, 0.74)
Image Encoder + Additive Hazard Layer	No	0.71 (0.68, 0.75)	0.73 (0.70, 0.76)	0.77 (0.76, 0.77)
Image Encoder + Additive Hazard	No	0.73 (0.70, 0.76)	0.73 (0.70, 0.76)	0.68 (0.67, 0.69)
+ Predict Risk Factors	Yes	0.75 (0.72, 0.79)	0.74 (0.72, 0.77)	0.68 (0.67, 0.69)
Image Encoder + Additive Hazard + Image Aggregator	No	0.75 (0.72, 0.78)	0.75 (0.73, 0.78)	0.76 (0.75, 0.76)
+ Predict Risk Factors	Yes	0.77 (0.74, 0.80)	0.75 (0.72, 0.78)	0.74 (0.73, 0.74)
Mirai = Image Encoder + Additive Hazard + Image Aggregator	No	0.73 (0.70, 0.77)	0.75 (0.72, 0.78)	0.50 (0.50, 0.50)
+ Predict Risk Factors + Adversarial Training	Yes	0.76 (0.73, 0.80)	0.76 (0.74, 0.80)	0.50 (0.50, 0.50)

Table B.10: Ablation study of Mirai on the MGH datasets. We report the C-Index for each model on the MGH validation and test sets, as well as the AUC of the Device-Identity Classifier on the test set. All metrics are followed by 95% confidence intervals.

	MGH Train Set		MGH Validation Set		MGH Test Set	
Characteristics	All	Cancer	All	Cancer	All	Cancer
All exams	137682 (100.0)	5202 (3.8)	16634 (100.0)	613 (3.7)	17119 (100.0)	608 (3.6)
Age						
<40	948 (0.7)	28 (3.0)	114 (0.7)	2 (1.8)	120 (0.7)	2 (1.7)
40-50	36971 (26.9)	1051 (2.8)	4483 (27.0)	151 (3.4)	4710 (27.5)	91 (1.9)
50-60	42425 (30.8)	1331 (3.1)	5153 (31.0)	155 (3.0)	5271 (30.8)	187 (3.5)
60-70	37715 (27.4)	1763 (4.7)	4585 (27.6)	181 (3.9)	4728 (27.6)	198 (4.2)
70-80	16663 (12.1)	798 (4.8)	1958 (11.8)	107 (5.5)	1977 (11.5)	96 (4.9)
80<	2960 (2.1)	231 (7.8)	341 (2.1)	17 (5.0)	313 (1.8)	34 (10.9)
Density						
Almost entirely fatty	12639 (9.2)	294 (2.3)	1499 (9.0)	47 (3.1)	1569 (9.2)	30 (1.9)
Scattered areas of fibroglandular tissue	65353 (47.5)	2496 (3.8)	8007 (48.1)	250 (3.1)	8112 (47.4)	293 (3.6)
Heterogeneously dense	52991 (38.5)	2171 (4.1)	6255 (37.6)	276 (4.4)	6633 (38.7)	265 (4.0)
Extremely dense	6623 (4.8)	239 (3.6)	867 (5.2)	39 (4.5)	797 (4.7)	20 (2.5)
Race						
White	112055 (81.4)	4495 (4.0)	13432 (80.8)	518 (3.9)	13932 (81.4)	534 (3.8)
African American	6585 (4.8)	204 (3.1)	792 (4.8)	27 (3.4)	807 (4.7)	26 (3.2)
Asian or Pacific Islander	6055 (4.4)	136 (2.2)	779 (4.7)	16 (2.1)	817 (4.8)	21 (2.6)
Hispanic	1542 (1.1)	52 (3.4)	181 (1.1)	5 (2.8)	145 (0.8)	4 (2.8)
Other Race	11445 (8.3)	315 (2.8)	1450 (8.7)	47 (3.2)	1418 (8.3)	23 (1.6)
Time to Next Exam						
<1 year	1313 (1.0)	295 (22.5)	172 (1.0)	50 (29.1)	159 (0.9)	32 (20.1)
1-2 years	114503 (83.2)	4264 (3.7)	13791 (82.9)	486 (3.5)	14192 (82.9)	514 (3.6)
2-3 years	12377 (9.0)	429 (3.5)	1489 (9.0)	58 (3.9)	1536 (9.0)	41 (2.7)
>= 3 years	9489 (6.9)	214 (2.3)	1182 (7.1)	19 (1.6)	1232 (7.2)	21 (1.7)
Time to Cancer						
0-1 year	61 (0.0)	61 (100.0)	8 (0.0)	8 (100.0)	13 (0.1)	13 (100.0)
1-2 years	298 (0.2)	298 (100.0)	40 (0.2)	40 (100.0)	34 (0.2)	34 (100.0)
2-3 years	508 (0.4)	508 (100.0)	50 (0.3)	50 (100.0)	60 (0.4)	60 (100.0)
3-4 years	632 (0.5)	632 (100.0)	80 (0.5)	80 (100.0)	97 (0.6)	97 (100.0)
4-5 years	724 (0.5)	724 (100.0)	92 (0.6)	92 (100.0)	94 (0.5)	94 (100.0)
5-10 years	2979 (2.2)	2979 (100.0)	343 (2.1)	343 (100.0)	310 (1.8)	310 (100.0)

Table B.11: Detailed demographics of MGH dataset.

	Emory Dataset	
Characteristics	All	Cancer
All exams	22,030 (100.0)	723 (3.3)
Age		
<40	237 (1.1)	7 (3.0)
40-50	4,523 (20.5)	114 (2.5)
50-60	6,210 (28.2)	162 (2.6)
60-70	7,018 (31.9)	231 (3.3)
70-80	3,532 (16.0)	195 (5.5)
80<	510 (2.3)	14 (2.7)
Race		
White	9,780 (44.4)	348 (3.6)
African American	10,436 (47.4)	343 (3.3)
Asian	994 (4.5)	15 (1.5)
Native Hawaiian or Other Pacific Islander	122 (0.6)	9 (7.4)
American Indian or Alaskan Native	21 (0.1)	NA
Multiple	47 (0.2)	NA
Time to Next Exam		
<1 year	529 (2.4)	48 (9.1)
1-2 years	16,557 (75.2)	546 (3.3)
2-3 years	2,628 (11.9)	82 (3.1)
>= 3 years	2,316 (10.5)	47 (2.0)
Time to Cancer		
0-1 year	16 (0.1)	16 (100.0)
1-2 years	96 (0.4)	96 (100.0)
2-3 years	124 (0.6)	124 (100.0)
3-4 years	110 (0.5)	110 (100.0)
4-5 years	132 (0.6)	132 (100.0)
5-10 years	245 (1.1)	245 (100.0)

Table B.12: Detailed demographics of Emory test set

	Karolinska Dataset before resampling		Karolinska Dataset after resampling	
Characteristics	All	Cancer	All	Cancer
All exams	14362 (100.0)	1768 (12.3)	93052 (100.0)	1768 (1.9)
Age				
40-50	5921 (41.2)	558 (9.4)	39433 (42.4)	558 (1.4)
50-60	4200 (29.2)	499 (11.9)	27514 (29.6)	499 (1.8)
60-70	3903 (27.2)	652 (16.7)	24010 (25.8)	652 (2.7)
70-80	338 (2.4)	59 (17.5)	2095 (2.3)	59 (2.8)
Time to Next Exam				
<1 year	90 (0.6)	84 (93.3)	134 (0.1)	84 (62.7)
1-2 years	5421 (37.7)	618 (11.4)	35380 (38.0)	618 (1.7)
2-3 years	7087 (49.3)	912 (12.9)	45844 (49.3)	912 (2.0)
>= 3 years	1764 (12.3)	154 (8.7)	11694 (12.6)	154 (1.3)
Time to Cancer				
0-1 year	25 (0.2)	25 (100.0)	25 (0.0)	25 (100.0)
1-2 years	94 (0.7)	94 (100.0)	94 (0.1)	94 (100.0)
2-3 years	257 (1.8)	257 (100.0)	257 (0.3)	257 (100.0)
3-4 years	204 (1.4)	204 (100.0)	204 (0.2)	204 (100.0)
4-5 years	352 (2.5)	352 (100.0)	352 (0.4)	352 (100.0)
5-10 years	836 (5.8)	836 (100.0)	836 (0.9)	836 (100.0)

Table B.13: Detailed demographics of Karolinska test set. Because the Karolinska dataset was collected in a case-control design, it has a much higher cancer incidence than reported in the CSAW cohort. To take this into account, we randomly resampled this dataset to produce a larger dataset with 1.9% cancer incidence.

	CGMH Dataset	
Characteristics	All	Cancer
All exams	12280 (100.0)	235 (1.9)
Age		
40-50	3656 (29.8)	74 (2.0)
50-60	5816 (47.4)	109 (1.9)
60-70	2801 (22.8)	52 (1.9)
70-80	7 (0.1)	NA
Time to Next Exam		
<1 year	NA	13 (100.0)
1-2 years	NA	31 (100.0)
2-3 years	NA	50 (100.0)
>= 3 years	NA	141 (1.2)
Time to Cancer		
0-1 year	11 (0.1)	11 (100.0)
1-2 years	24 (0.2)	24 (100.0)
2-3 years	42 (0.3)	42 (100.0)
3-4 years	26 (0.2)	26 (100.0)
4-5 years	36 (0.3)	36 (100.0)
5-6 years	96 (0.8)	96 (100.0)

Table B.14: Detailed demographics of CGMH test set.

Risk Model	Progression Model	KL Divergence on Test Set (95% Confidence interval)
MGH Test Set: 17,119 exams from 5,525 patients. 210 patients develop cancer.		
Mirai	Static Risk	0.038 (0.036, 0.040)
	RNN	0.028 (0.026, 0.029)
Emory Test Set: 22,094 exams from 10,369 patients. 333 patients develop cancer.		
Mirai	Static Risk	0.035 (0.034, 0.036)
	RNN	0.029 (0.028, 0.030)
Karolinska Test Set: 14,353 exams from 7,191 patients. 919 patients develop cancer.		
Mirai	Static Risk	0.029 (0.027, 0.031)
	RNN	0.026 (0.025, 0.027)

Table B.15: Testing risk progression models on the MGH, Emory and Karolinska test sets. Static Risk assumes that patient risk does not change, i.e., risk assessments at future time steps will equal the last observed risk assessment. RNN is an auto-regressive recurrent neural network that was trained to predict future risk assessments from prior assessments on the MGH training set. For each model, we report the Kullback-Leibler (KL) divergence (lower is better), between the risk progression model predicted risk and the observed risk. All metrics are followed by their 95% confidence interval.

Screening Policy	Risk model	Average Mammograms per Year	Early Detection in Months	Efficiency
Race: African American. 10436 exams from 4716 patients. 158 patients develop cancer.				
Annual	Age	1.0 (1.0, 1.0)	3.10 (1.70. 4.22)	3.10 (1.70, 4.22)
Biennial	Age	0.5 (0.5, 0.5)	-3.97 (-5.73, -2.36)	-7.94 (-11.47, -4.72)
USPSTF	Age	0.68 (0.68, 0.69)	-1.96 (-3.57, -0.26)	-2.87 (-5.18, -0.39)
Supervised	Mirai	1.18 (1.16, 1.20)	2.34 (0.57, 4.45)	1.98 (0.48, 3.83)
Tempo	Mirai	1.10 (1.09, 1.11)	6.47 (5.05, 7.64)	5.89 (4.56, 7.02)
Race: White. 9780 exams from 4587 patients. 159 patients develop cancer.				
Annual	Age	1.0 (1.0, 1.0)	3.55 (2.16, 4.32)	3.55 (2.16, 4.32)
Biennial	Age	0.5 (0.5, 0.5)	-3.95 (-5.44, -2.80)	-7.90 (-10.89, -5.60)
USPSTF	Age	0.65 (0.64, 0.66)	-2.21 (-3.47, -0.86)	-3.40 (-5.26, -1.34)
Supervised	Mirai	1.16 (1.14, 1.28)	2.43 (0.86, 4.65)	2.09 (0.72, 4.07)
Tempo	Mirai	1.07 (1.06, 1.09)	6.50 (5.72, 7.46)	6.06 (5.28, 7.04)

Table B.16: Results for all screening policies on African American and White patients of the Emory test set. For each policy, we report the average number of mammograms per year, the early detection benefit in months relative to historical screening (higher positive number means earlier), and the screening efficiency (higher positive number is better). We defined screening efficiency as the early detection benefit divided by the average number of mammograms per year. All metrics are followed by their 95% confidence interval.

Screening Policy	Risk model	Average Mammograms per Year	Early Detection in Months	Efficiency
Age: <=55. 8,038 exams from 3,016 patients. 84 patients develop cancer.				
Annual	Age	1.0 (1.0, 1.0)	1.16 (-0.34, 2.62)	1.16 (-0.34, 2.62)
Biennial	Age	0.5 (0.5, 0.5)	-5.50 (-7.11, -3.90)	-11.00 (-14.22, -7.80)
USPSTF	Age	0.97 (0.96, 0.97)	0.848 (-0.76, 2.29)	0.88 (-0.78, 2.37)
Supervised	TCv8	1.45 (1.42, 1.48)	2.17 (0.39, 4.30)	1.50 (0.26, 3.03)
	Mirai	0.769 (0.75, 0.80)	-2.17 (-4.83, 1.01)	-2.82 (-6.08, 1.35)
Tempo	TCv8	0.96 (0.94, 0.99)	1.76 (0.21, 3.77)	1.83 (0.21, 4.01)
	Mirai	0.86 (0.84, 0.87)	2.92 (1.25, 4.63)	3.41 (1.44, 5.53)
Age: >55. 9081 exams from 2959 patients. 148 patients develop cancer.				
Annual	Age	1.0 (1.0, 1.0)	1.77 (0.68, 2.83)	1.77 (0.68, 2.83)
Biennial	Age	0.5 (0.5, 0.5)	-5.02 (-6.23, -3.89)	-10.04 (-12.47, -7.78)
USPSTF	Age	0.5 (0.5, 0.5)	-5.02 (-6.23, -3.88)	-10.04 (-12.47, -7.78)
Supervised	TCv8	1.85 (1.83, 1.87)	5.64 (4.28, 7.44)	3.06 (2.30, 4.07)
	Mirai	1.08 (1.06, 1.12)	2.09 (0.38, 3.93)	1.92 (0.34, 3.72)
Tempo	TCv8	1.72 (1.70, 1.74)	4.84 (3.20, 6.80)	2.82 (1.84, 4.02)
	Mirai	1.04 (1.03, 1.06)	4.63 (3.49, 6.03)	4.45 (3.29, 5.88)
Density: Non-dense. 9681 exams from 3370 patients. 120 patients develop cancer.				
Annual	Age	1.0 (1.0, 1.0)	1.54 (0.21, 2.65)	1.54 (0.21, 2.65)
Biennial	Age	0.5 (0.5, 0.5)	-5.26 (-6.64, -3.97)	-10.51 (-13.28, -7.95)
USPSTF	Age	0.67 (0.66, 0.67)	-4.05 (-5.43, -2.80)	-6.08 (-8.05, -4.27)
Supervised	TCv8	1.62 (1.6, 1.65)	4.29 (2.26, 6.31)	2.65 (1.38, 3.95)
	Mirai	0.87 (0.84, 0.90)	0.07 (-1.86, 2.03)	0.09 (-2.07, 2.42)
Tempo	TCv8	0.96 (0.93, 0.97)	1.56 (0.52, 2.48)	1.64 (0.54, 2.67)
	Mirai	0.94 (0.91, 0.95)	3.83 (2.70, 5.00)	4.10 (2.85, 5.48)
Density: Dense. 7430 exams from 2839 patients. 116 patients develop cancer.				
Annual	Age	1.0 (1.0, 1.0)	1.621 (0.28, 2.95)	1.621 (0.28, 2.95)
Biennial	Age	0.5 (0.5, 0.5)	-5.07 (-6.50, -3.61)	-10.15 (-12.98, -7.21)
USPSTF	Age	0.79 (0.78, 0.80)	-2.19 (-3.80, -0.69)	-2.78 (-4.77, -0.89)
Supervised	TCv8	1.72 (1.70, 1.74)	4.84 (3.20, 6.80)	2.82 (1.84, 4.02)
	Mirai	1.02 (0.98, 1.05)	1.52 (-0.17, 4.00)	1.49 (-0.15, 4.08)
Tempo	TCv8	0.96 (0.93, 0.98)	2.63 (1.34, 4.54)	2.77 (1.37, 4.89)
	Mirai	0.98 (0.96, 1.0)	4.4 (2.86, 6.08)	4.49 (2.86, 6.35)

Table B.17: Results for all screening policies on subgroups of the MGH test set. For each policy, we report the average number of mammograms per year, the early detection benefit in months relative to historical screening (higher positive number means earlier), and the screening efficiency (higher positive number is better). We defined screening efficiency as the early detection benefit divided by the average number of mammograms per year. All metrics are followed by their 95% confidence interval.

Screening Policy	Risk model	Average number of Mammograms per Year	Earlier Detection in Months	Efficiency
MGH Test Set: 17,119 exams from 5,525 patients. 210 patients develop cancer.				
Annual	Age	1.0 (1.00, 1.00)	1.37 (0.22, 2.41)	1.37 (0.22, 2.41)
Biennial	Age	0.5 (0.50, 0.50)	-5.54 (-6.71, -4.53)	-11.07 (-13.42, -9.07)
USPSTF	Age	0.72 (0.71, 0.73)	-3.52 (-4.58, -2.66)	-4.90 (-6.31, -3.60)
Supervised	TCv8	1.66 (1.65, 1.69)	4.51 (3.42, 6.02)	2.72 (2.03, 3.66)
	Mirai	0.94 (0.92, 0.96)	0.75 (-0.55, 1.94)	0.80 (-0.58, 2.1)
Tempo	TCv8	0.96 (0.94, 0.97)	1.84 (1.01, 2.88)	1.92 (1.04, 3.06)
	Mirai	0.96 (0.94, 0.97)	3.89 (2.82, 4.77)	4.07 (2.92, 5.05)
Emory Test Set: 22,030 exams from 10,340 patients. 333 patients develop cancer.				
Annual	Age	1.0 (1.0, 1.0)	2.85 (1.97, 3.51)	2.85 (1.97, 3.51)
Biennial	Age	0.5 (0.5, 0.5)	-4.61 (-5.59, -3.79)	-9.21 (-11.18, -7.59)
USPSTF	Age	0.68 (0.67, 0.69)	-2.63 (-3.39, -2.03)	-3.89 (-4.98, -3.03)
Supervised	Mirai	1.16 (1.15, 1.18)	1.97 (0.49, 3.27)	1.69 (0.42, 2.84)
Tempo	Mirai	1.08 (1.07, 1.08)	6.04 (5.24, 6.77)	5.61 (4.82, 6.32)
Karolinska Test Set: 14,353 exams from 7,191 patients. 919 patients develop cancer.				
Annual	Age	1.0 (1.0, 1.0)	5.94 (5.26, 6.56)	5.94 (5.26, 6.56)
Biennial	Age	0.5 (0.5, 0.5)	-2.76 (-3.36, -2.29)	-5.51 (-6.71, -4.58)
USPSTF	Age	0.79 (0.79, 0.80)	0.46 (-0.05, 1.02)	0.58 (-0.07, 1.30)
Supervised	Mirai	0.60 (0.59, 0.61)	0.07 (-0.78, 1.00)	0.12 (-1.28, 1.70)
Tempo	Mirai	0.75 (0.74, 0.76)	6.60 (5.85, 7.32)	8.79 (7.73, 9.86)
CGMH Test Set: 12280 exams from 12280 patients. 235 patients develop cancer.				
Annual	Age	1.0 (1.0, 1.0)	9.42 (8.32, 10.34)	9.42 (8.32, 10.34)
Biennial	Age	0.5 (0.5, 0.5)	1.81 (0.42, 3.17)	3.63 (0.87, 6.33)
USPSTF	Age	0.78 (0.77, 0.78)	5.90 (4.31, 6.97)	7.58 (5.52, 9.02)
Supervised	Mirai	0.98 (0.97, 0.99)	1.97 (0.49, 3.27)	1.69 (0.42, 2.84)
Tempo	Mirai	0.88 (0.87, 0.89)	8.68 (7.56, 9.80)	9.87 (8.52, 11.22)

Table B.18: Results for all screening policies on the MGH, Emory, Karolinska and CGMH test sets leveraging an alternative definition of early detection benefit. For each policy, we report the average number of mammograms per year, the early detection benefit in months relative to historical screening (higher positive number means earlier), and the screening efficiency (higher positive number is better). We defined screening efficiency as the early detection benefit divided by the average number of mammograms per year. All metrics are followed by their 95% confidence interval.

Encoding	Guesswork	ReId AUC
Dauntless	1.0 $(1, 1)$	1.00 $(1.00, 1.00)$
InstaHide	1.0 $(1, 1)$	1.00 $(1.00, 1.00)$
DP-S, $b = 10$	1.2 $(1, 2)$	0.98 $(0.98, 0.98)$
DP-S, $b = 20$	7.2 $(1, 31)$	0.86 $(0.85, 0.86)$
DP-S, $b = 30$	68 $(1, 205)$	0.70 $(0.70, 0.70)$
DP-I, $b = 1$	5.0 $(1, 17)$	0.89 $(0.88, 0.89)$
DP-I, $b = 3$	77 $(3, 276)$	0.73 $(0.73, 0.73)$
DP-I, $b = 5$	1379 $(49, 4135)$	0.59 $(0.59, 0.60)$
Syfer-Random	1.7 $(1, 4)$	0.99 $(0.99, 0.99)$
Syfer (T^X only)	8476 $(1971, 20225)$	0.50 $(0.49, 0.52)$

Table B.19: Privacy evaluation of different encoding schemes against an SAU based attacker on the unlabeled MIMIC-CXR dataset. For Syfer, only T^X is used. DP-S and DP-I stand for DP-Simple and DP-Image respectively. The scale parameter b characterizes the laplacian noise. Metrics are averages over 100 trials using 10,000 samples each, followed by 95% confidence intervals (CI).

Encoding	E	Co	Ca	A	*Avg*
Using raw data	0.91	0.78	0.89	0.85	0.86
Using encoded data					
DP-S, $b = 10$	0.51	0.51	0.52	0.52	0.52
DP-S, $b = 20$	0.50	0.50	0.50	0.50	0.50
DP-S, $b = 30$	0.49	0.49	0.50	0.51	0.50
DP-I, $b = 1$	0.60	0.59	0.60	0.59	0.60
DP-I, $b = 2$	0.54	0.50	0.55	0.55	0.54
DP-I, $b = 5$	0.53	0.55	0.51	0.52	0.53
Syfer-Random	0.89	0.75	0.86	0.84	0.84
Syfer	0.82	0.69	0.81	0.78	0.78

Table B.20: Utility for chest X-ray prediction tasks across different encoding schemes. All metrics are ROC AUCs across the MIMIC-CXR test set. Guides of abbreviations for medical diagnosis: (E)dema, (Co)nsolidation, (Ca)rdiomegaly and (A)telectasis.

Diagnosis	Guesswork	ReId AUC
	Syfer	
Edema	3617 $(94, 11544)$	0.50 $(0.49, 0.51)$
Consolidation	1697 $(83, 5297)$	0.55 $(0.53, 0.57)$
Cardiomegaly	9834 $(2072, 15766)$	0.51 $(0.49, 0.53)$
Atelectasis	13189 $(2511, 28171)$	0.50 $(0.48, 0.52)$
Ablation: Syfer with no label encoding (T^X only)		
Edema	47 $(12, 83)$	0.76 $(0.76, 0.76)$
Consolidation	36 $(2, 104)$	0.76 $(0.76, 0.76)$
Cardiomegaly	42 $(17, 57)$	0.75 $(0.75, 0.75)$
Atelectasis	80 $(65, 98)$	0.75 $(0.75, 0.75)$

Table B.21: Privacy evaluation of Syfer when released with different diagnoses in MIMIC-CXR dataset. Metrics are averages over 100 trials using 10,000 samples each, followed by 95% CI.

Attacker	Guesswork	ReId AUC
SAU	8476 $(1971, 20225)$	0.50 $(0.49, 0.52)$
ViT	8411 $(5219, 12033)$	0.50 $(0.49, 0.51)$
Resnet-18	10070 $(9871, 10300)$	0.50 $(0.47, 0.53)$

Table B.22: Privacy evaluation of Syfer across different attacker architectures on the unlabeled MIMIC-CXR dataset. Metrics are averages over 100 trials using 10,000 samples each, followed by 95% CI.

Appendix C

Figures

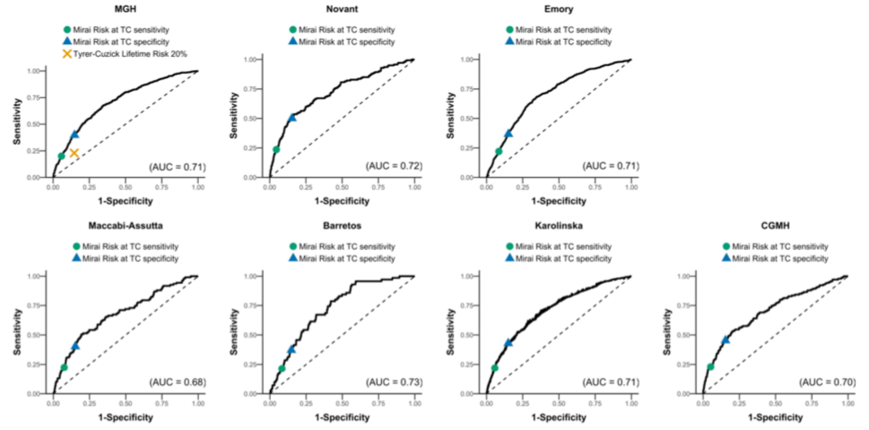

Figure C-1: Receiver operating curves for Mirai in selecting high cohorts across all test sets. These datasets are restricted to include patients who were screening negative and either had cancer within 5 years or 5 years of negative follow-up.

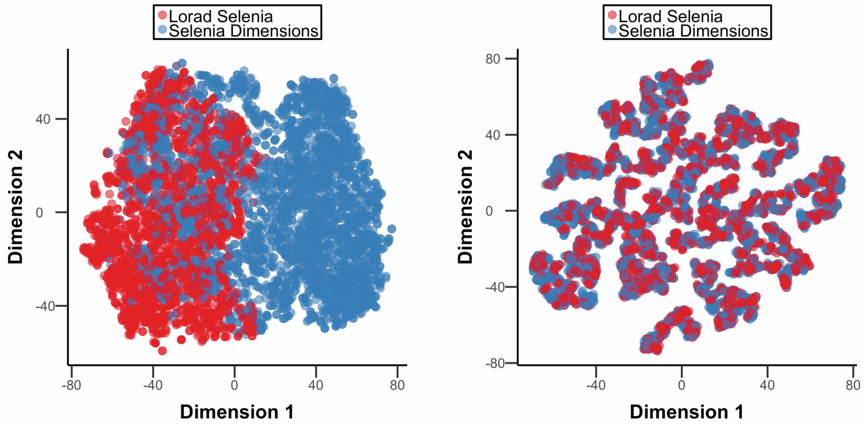

Figure C-2: T-SNE plots for Mirai's hidden representation colored by cancer subtypes factors on 1000 random positive exams from the Karolinska test set.

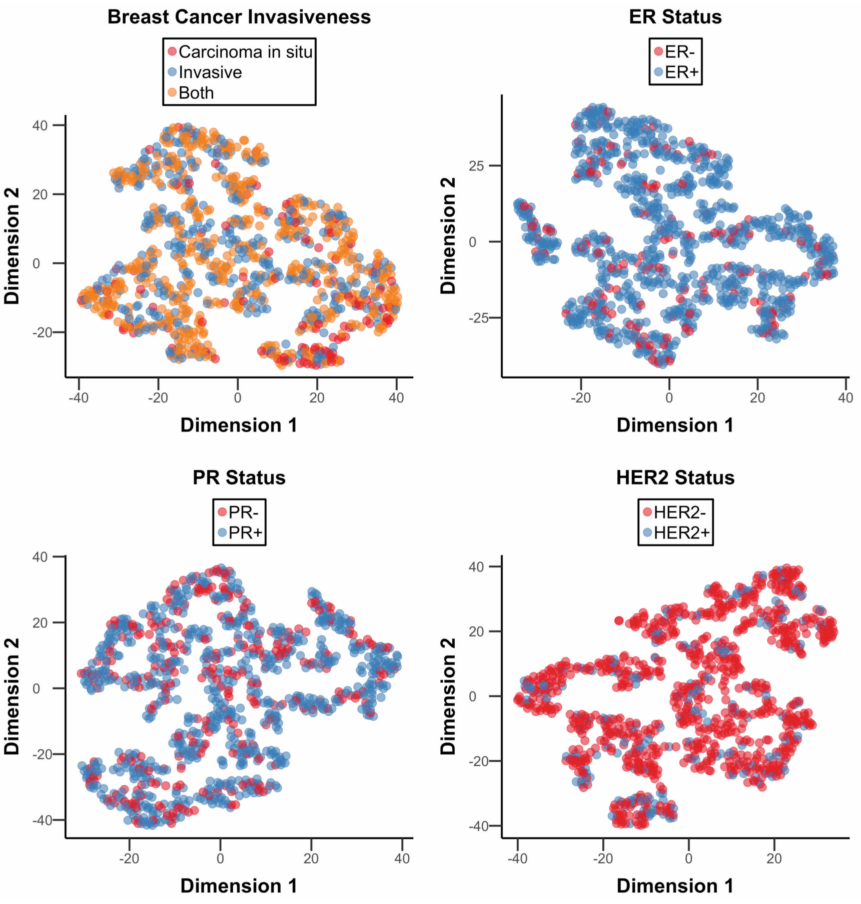

Figure C-3: T-SNE plots for Mirai's hidden representation colored by cancer subtypes factors on 1000 random positive exams from the Karolinska test set.

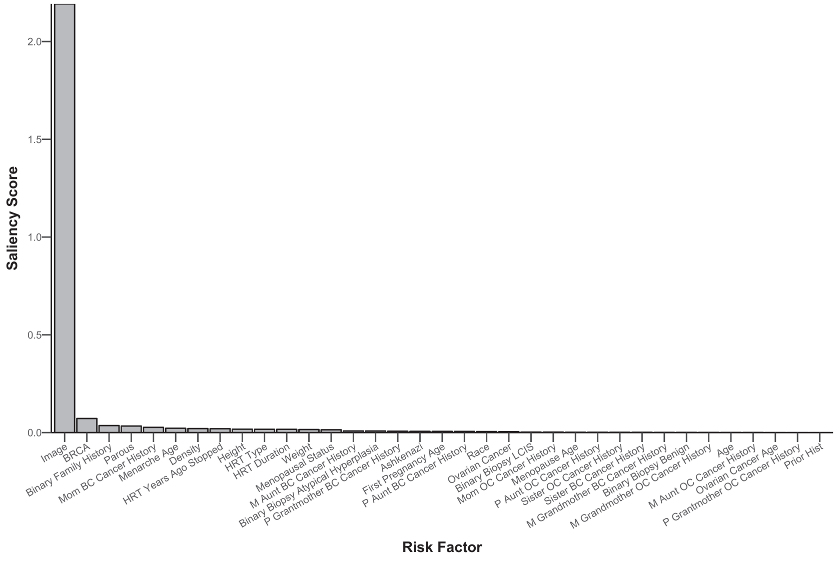

Figure C-4: Saliency scores of images and all clinical risk factors across the MGH test set.

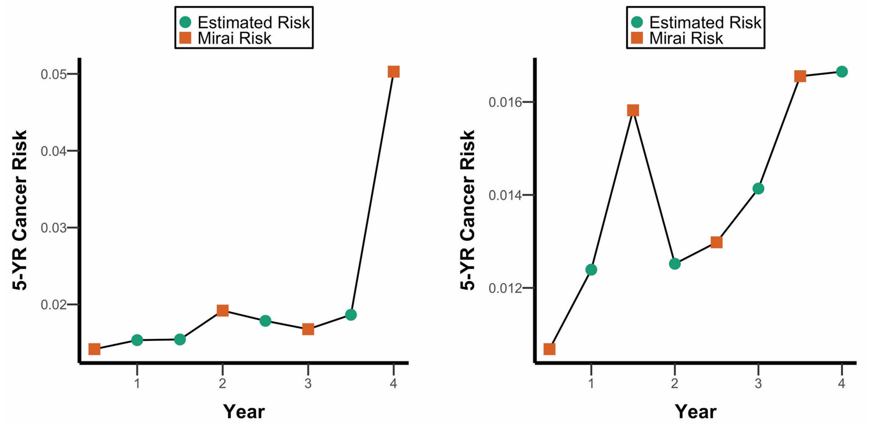

Figure C-5: Estimated (circle) and observed (square) Mirai five-year risk for two random patients in the MGH test set. We estimated unobserved risk observations using a recurrent neural network, which was optimized to predict future risk assessments from past risk assessments on the MGH training set.

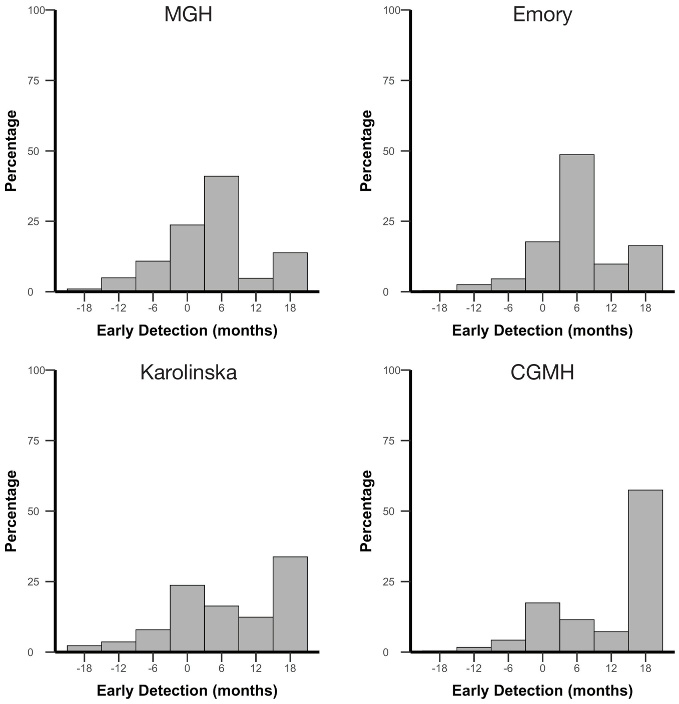

Figure C-6: Histogram of early detection benefit in months relative to historical screening for patients who developed cancer in the MGH (top left), Emory (top right), Karolinska (bottom left), and CGMH (bottom right) test sets.

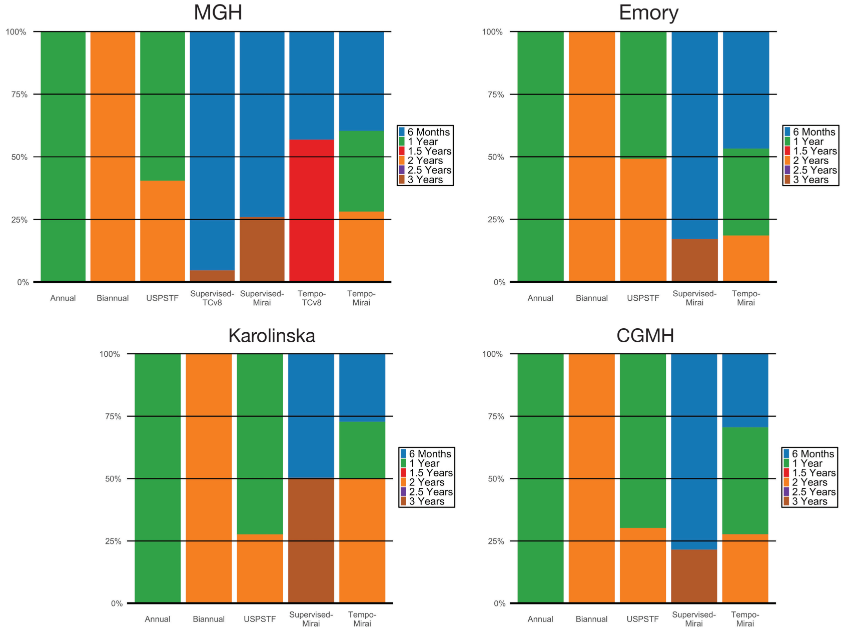

Figure C-7: Histogram of screening recommendations for each screening policy. MGH (top left), Emory (top right), Karolinska (bottom left), CGMH (bottom right).

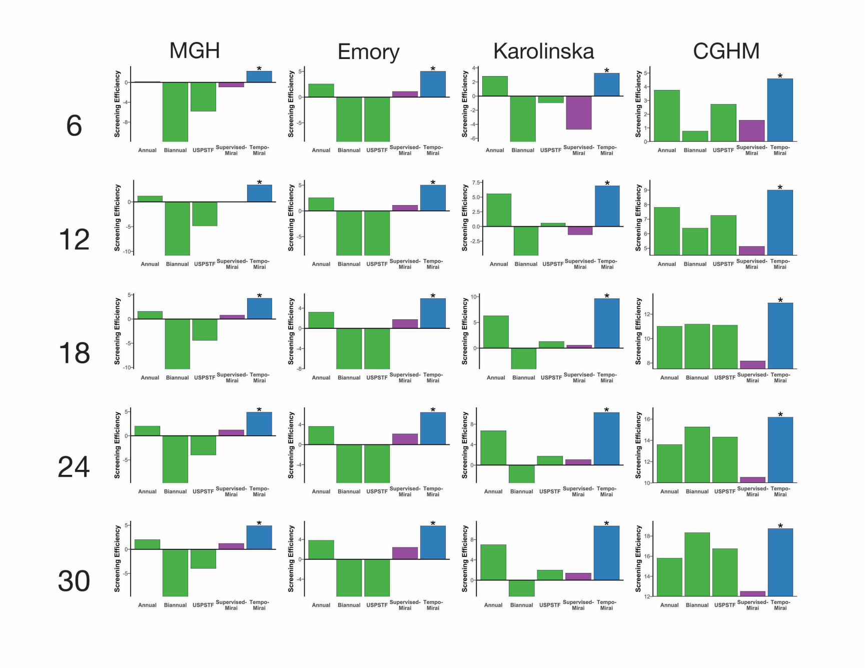

Figure C-8: Our early detection metric assumed that a cancer could be caught up to 18 months before diagnosis. To test the robustness of our results to this assumption, we also evaluated our screening policies when changing this assumption to six months, 12 months and 24 months. For each policy, we report its screening efficiency, which is defined as its early detection benefit in months divided by the amount of mammograms it recommends per year. We use a * to denote the policy with the highest screening efficiency.

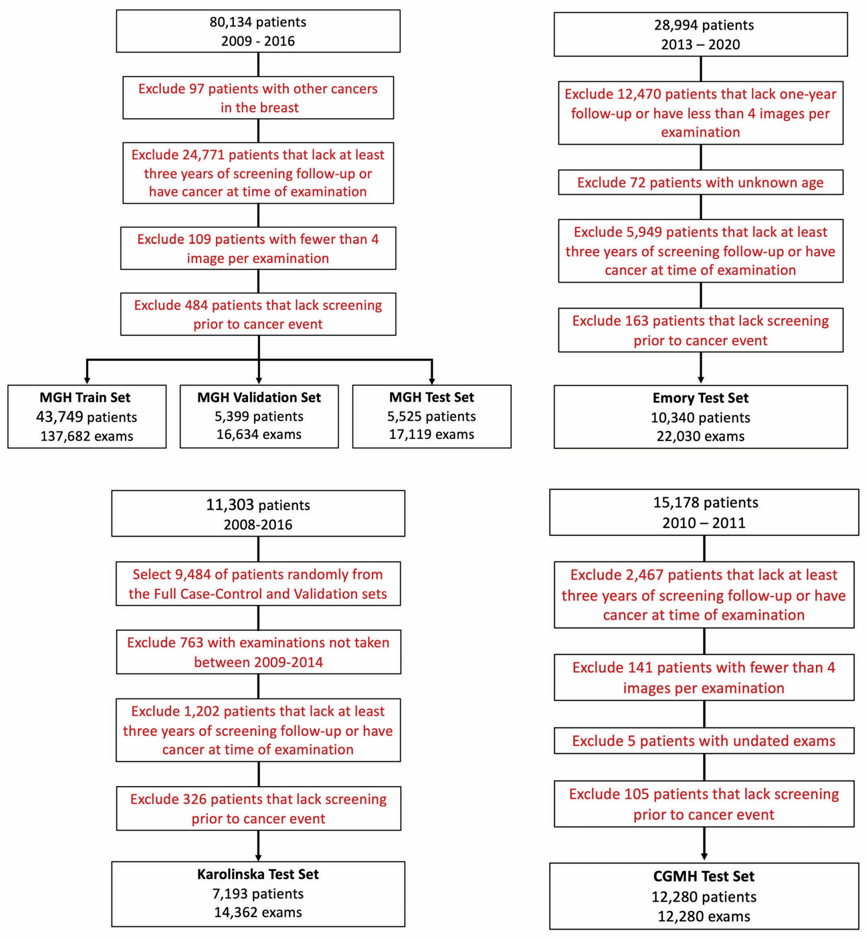
80,134 patients
2009 - 2016
Exclude 97 patients with other cancers in the breast
Exclude 24,771 patients that lack at least three years of screening follow-up or have cancer at time of examination
Exclude 109 patients with fewer than 4 image per examination
Exclude 484 patients that lack screening prior to cancer event
MGH Train Set
43,749 patients
137,682 exams
MGH Validation Set
5,399 patients
16,634 exams
MGH Test Set
5,525 patients
17,119 exams
28,994 patients
2013 – 2020
Exclude 12,470 patients that lack one-year follow-up or have less than 4 images per examination
Exclude 72 patients with unknown age
Exclude 5,949 patients that lack at least three years of screening follow-up or have cancer at time of examination
Exclude 163 patients that lack screening prior to cancer event
Emory Test Set
10,340 patients
22,030 exams
11,303 patients
2008-2016
Select 9,484 of patients randomly from the Full Case-Control and Validation sets
Exclude 763 with examinations not taken between 2009-2014
Exclude 1,202 patients that lack at least three years of screening follow-up or have cancer at time of examination
Exclude 326 patients that lack screening prior to cancer event
Karolinska Test Set
7,193 patients
14,362 exams
15,178 patients
2010 – 2011
Exclude 2,467 patients that lack at least three years of screening follow-up or have cancer at time of examination
Exclude 141 patients with fewer than 4 images per examination
Exclude 5 patients with undated exams
Exclude 105 patients that lack screening prior to cancer event
CGMH Test Set
12,280 patients
12,280 exams

Figure C-9: Dataset construction flow chart for the MGH dataset (top left), Emory (top right), Karolinska test set (bottom left), and CGMH test set (bottom right).

Bibliography

[1] Odd O Aalen and Thomas H Scheike. Aalen's additive regression model. *Wiley StatsRef: Statistics Reference Online*, 2014.

[2] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pages 308–318, 2016.

[3] Kartik Ahuja, William Zame, and Mihaela van der Schaar. Dpscreen: Dynamic personalized screening. In *Advances in Neural Information Processing Systems*, pages 1321–1332, 2017.

[4] Ayelet Akselrod-Ballin, Michal Chorev, Yoel Shoshan, Adam Spiro, Alon Hazan, Roie Melamed, Ella Barkan, Esma Herzel, Shaked Naor, Ehud Karavani, et al. Predicting breast cancer by applying deep learning to linked health records and mammograms. *Radiology*, 292(2):331–342, 2019.

[5] Marcin Andrychowicz, Filip Wolski, Alex Ray, Jonas Schneider, Rachel Fong, Peter Welinder, Bob McGrew, Josh Tobin, OpenAI Pieter Abbeel, and Wojciech Zaremba. Hindsight experience replay. In *Advances in neural information processing systems*, pages 5048–5058, 2017.

[6] Diego Ardila, Atilla P Kiraly, Sujeeth Bharadwaj, Bokyung Choi, Joshua J Reicher, Lily Peng, Daniel Tse, Mozziyar Etemadi, Wenxing Ye, Greg Corrado, et al. End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. *Nature medicine*, 25(6):954–961, 2019.

[7] E. Arikan. An inequality on guessing and its application to sequential decoding. In *IEEE International Symposium on Information Theory*, pages 322–, 1995.

[8] Turgay Ayer, Oguzhan Alagoz, and Natasha K Stout. Or forumÑa pomdp approach to personalize mammography screening decisions. *Operations Research*, 60(5):1019–1034, 2012.

[9] Marije F Bakker, Stéphanie V de Lange, Ruud M Pijnappel, Ritse M Mann, Petra HM Peeters, Evelyn M Monninkhof, Marleen J Emaus, Claudette E Loo, Robertus HC Bisschops, Marc BI Lobbes, et al. Supplemental mri screening for women with extremely dense breast tissue. *New England Journal of Medicine*, 381(22):2091–2102, 2019.

[10] Ahmad Beirami, Robert Calderbank, Mark M. Christiansen, Ken R. Duffy, and Muriel Médard. A characterization of guesswork on swiftly tilting curves. *IEEE Transactions on Information Theory*, 65(5):2850–2871, 2019.

[11] Michael Ben-Or, Shafi Goldwasser, and Avi Wigderson. Completeness theorems for non-cryptographic fault-tolerant distributed computation (extended abstract). In Janos Simon, editor, *Proceedings of the 20th Annual ACM Symposium on Theory of Computing, May 2-4, 1988, Chicago, Illinois, USA*, pages 1–10. ACM, 1988.

[12] Therese B Bevers, John H Ward, Banu K Arun, Graham A Colditz, Kenneth H Cowan, Mary B Daly, Judy E Garber, Mary L Gemignani, William J Gradishar, Judith A Jordan, et al. Breast cancer risk reduction, version 2.2015. *Journal of the National Comprehensive Cancer Network*, 13(7):880–915, 2015.

[13] Kirsten Bibbins-Domingo, David C Grossman, Susan J Curry, Karina W Davidson, John W Epling, Francisco AR García, Matthew W Gillman, Diane M Harper, Alex R Kemper, Alex H Krist, et al. Screening for colorectal cancer: Us preventive services task force recommendation statement. *Jama*, 315(23):2564–2575, 2016.

[14] M Bond, TG Pavey, K Welch, Chris Cooper, R Garside, Sarah Dean, and C Hyde. Systematic review of the psychological consequences of false-positive screening mammograms. 2013.

[15] Florian Bourse, Michele Minelli, Matthias Minihold, and Pascal Paillier. Fast homomorphic evaluation of deep discretized neural networks. In *Advances in Cryptology*, volume 10993, pages 483–512. Springer, 2018.

[16] Norman F Boyd, JW Byng, RA Jong, EK Fishell, LE Little, AB Miller, GA Lockwood, DL Tritchler, and Martin J Yaffe. Quantitative classification of mammographic densities and breast cancer risk: results from the canadian national breast screening study. *JNCI: Journal of the National Cancer Institute*, 87(9):670–675, 1995.

[17] Zvika Brakerski and Vinod Vaikuntanathan. Efficient fully homomorphic encryption from (standard) LWE. *SIAM J. Comput.*, 43(2):831–871, 2014.

[18] Kathleen R Brandt, Christopher G Scott, Lin Ma, Amir P Mahmoudzadeh, Matthew R Jensen, Dana H Whaley, Fang Fang Wu, Serghei Malkov, Carrie B Hruska, Aaron D Norman, et al. Comparison of clinical and automated breast density measurements: implications for risk prediction and supplemental screening. *Radiology*, 279(3):710–719, 2016.

[19] Adam R Brentnall, Elaine F Harkness, Susan M Astley, Louise S Donnelly, Paula Stavrinos, Sarah Sampson, Lynne Fox, Jamie C Sergeant, Michelle N Harvie, Mary Wilson, et al. Mammographic density adds accuracy to both the

tyrer-cuzick and gail breast cancer risk models in a prospective uk screening cohort. *Breast Cancer Research*, 17(1):147, 2015.

[20] John Brodersen and Volkert Dirk Siersma. Long-term psychosocial consequences of false-positive screening mammography. *The Annals of Family Medicine*, 11(2):106–115, 2013.

[21] David Chaum, Claude Crépeau, and Ivan Damgård. Multiparty unconditionally secure protocols (extended abstract). In Janos Simon, editor, *Proceedings of the 20th Annual ACM Symposium on Theory of Computing, May 2-4, 1988, Chicago, Illinois, USA*, pages 11–19. ACM, 1988.

[22] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[23] Victoria Cheng, Vinith M Suriyakumar, Natalie Dullerud, Shalmali Joshi, and Marzyeh Ghassemi. Can you fake it until you make it? impacts of differentially private synthetic data on downstream classification fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 149–160, 2021.

[24] Hyunghoon Cho, David J Wu, and Bonnie Berger. Secure genome-wide association analysis using multiparty computation. *Nature biotechnology*, 36(6):547–551, 2018.

[25] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

[26] Mark M. Christiansen, Ken R. Duffy, Flávio du Pin Calmon, and Muriel Médard. Brute force searching, the typical set and guesswork. In *IEEE International Symposium on Information Theory*, pages 1257–1261, 2013.

[27] Elizabeth B Claus, Neil Risch, and W Douglas Thompson. The calculation of breast cancer risk for women with a first degree family history of ovarian cancer. *Breast cancer research and treatment*, 28(2):115–120, 1993.

[28] ClinicalTrials.gov. National Library of Medicine (US). International randomized study comparing personalized, risk-stratified to standard breast cancer screening in women aged 40-70., 2019, Jul 18 -. Identifier NCT03672331. Retrieved from http://clinicaltrials.gov/ct/show/NCT03672331.

[29] Andrew Coldman, Norm Phillips, Christine Wilson, Kathleen Decker, Anna M Chiarelli, Jacques Brisson, Bin Zhang, Jennifer Payne, Gregory Doyle, and Rukshanda Ahmad. Pan-canadian study of mammography screening and mortality from breast cancer. *JNCI: Journal of the National Cancer Institute*, 106(11), 2014.

[30] Cathy Coleman. Early detection and screening for breast cancer. In *Seminars in oncology nursing*, volume 33, pages 141–155. Elsevier, 2017.

[31] Pierre Courtiol, Charles Maussion, Matahi Moarii, Elodie Pronier, Samuel Pilcer, Meriem Sefta, Pierre Manceron, Sylvain Toldo, Mikhail Zaslavskiy, Nolwenn Le Stang, et al. Deep learning-based classification of mesothelioma improves prediction of patient outcome. *Nature medicine*, 25(10):1519–1525, 2019.

[32] Susan J Curry, Alex H Krist, Douglas K Owens, Michael J Barry, Aaron B Caughey, Karina W Davidson, Chyke A Doubeni, John W Epling, Alex R Kemper, Martha Kubik, et al. Screening for cervical cancer: Us preventive services task force recommendation statement. *Jama*, 320(7):674–686, 2018.

[33] Harry J de Koning, Carlijn M van der Aalst, Pim A de Jong, Ernst T Scholten, Kristiaan Nackaerts, Marjolein A Heuvelmans, Jan-Willem J Lammers, Carla Weenink, Uraujh Yousaf-Khan, Nanda Horeweg, et al. Reduced lung-cancer mortality with volume ct screening in a randomized trial. *New England journal of medicine*, 382(6):503–513, 2020.

[34] Elizabeth R DeLong, David M DeLong, and Daniel L Clarke-Pearson. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, pages 837–845, 1988.

[35] Karin Dembrower, Peter Lindholm, and Fredrik Strand. A multi-million mammography image dataset and population-based screening cohort for the training and evaluation of deep neural networks-the cohort of screen-aged women (csaw). *Journal of digital imaging*, pages 1–6, 2019.

[36] Karin Dembrower, Yue Liu, Hossein Azizpour, Martin Eklund, Kevin Smith, Peter Lindholm, and Fredrik Strand. Comparison of a deep learning risk score and standard mammographic density score for breast cancer risk prediction. *Radiology*, 294(2):265–272, 2020.

[37] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[38] Flávio du Pin Calmon, Muriel Médard, Linda M. Zeger, João Barros, Mark M. Christiansen, and Ken R. Duffy. Lists that are smaller than their parts: A coding approach to tunable secrecy. In *Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 1387–1394, 2012.

[39] Stephen W Duffy, Daniel Vulkan, Howard Cuckle, Dharmishta Parmar, Shama Sheikh, Robert A Smith, Andrew Evans, Oleg Blyuss, Louise Johns, Ian O Ellis, et al. Effect of mammographic screening from age 40 years on breast cancer mortality (uk age trial): final results of a randomised, controlled trial. *The Lancet Oncology*, 21(9):1165–1172, 2020.

[40] Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3-4):211–407, 2014.

[41] Thibault Févry, Jason Phang, Nan Wu, S Kim, Linda Moy, Kyunghyun Cho, and Krzysztof J Geras. Improving localization-based approaches for breast cancer screening exam classification. *arXiv preprint arXiv:1908.00615*, 2019.

[42] Mitchell H Gail, Louise A Brinton, David P Byar, Donald K Corle, Sylvan B Green, Catherine Schairer, and John J Mulvihill. Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. *JNCI: Journal of the National Cancer Institute*, 81(24):1879–1886, 1989.

[43] GDPR. *EU General Data Protection Regulation of 2016.*

[44] Craig Gentry. Fully homomorphic encryption using ideal lattices. In Michael Mitzenmacher, editor, *Proceedings of the 41st Annual ACM Symposium on Theory of Computing, STOC 2009, Bethesda, MD, USA, May 31 - June 2, 2009*, pages 169–178. ACM, 2009.

[45] Krzysztof J Geras, Stacey Wolfson, Yiqiu Shen, Nan Wu, S Kim, Eric Kim, Laura Heacock, Ujas Parikh, Linda Moy, and Kyunghyun Cho. High-resolution breast cancer screening with multi-view deep convolutional neural networks. *arXiv preprint arXiv:1703.07047*, 2017.

[46] Rodrigo A Gier, Krista A Budinich, Niklaus H Evitt, Zhendong Cao, Elizabeth S Freilich, Qingzhou Chen, Jun Qi, Yemin Lan, Rahul M Kohli, and Junwei Shi. High-performance crispr-cas12a genome editing for combinatorial genetic screening. *Nature communications*, 11(1):1–9, 2020.

[47] Oded Goldreich, Silvio Micali, and Avi Wigderson. How to play any mental game or A completeness theorem for protocols with honest majority. In Alfred V. Aho, editor, *Proceedings of the 19th Annual ACM Symposium on Theory of Computing, 1987, New York, New York, USA*, pages 218–229. ACM, 1987.

[48] Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. Deep learning. vol. 1, 2016.

[49] Cary P Gross, Jessica B Long, Joseph S Ross, Maysa M Abu-Khalaf, Rong Wang, Brigid K Killelea, Heather T Gold, Anees B Chagpar, and Xiaomei Ma. The cost of breast cancer screening in the medicare population. *JAMA internal medicine*, 173(3):220–226, 2013.

[50] Swedish Organised Service Screening Evaluation Group et al. Reduction in breast cancer mortality from organized service screening with mammography: 1. further confirmation with extended data. *Cancer Epidemiology Biomarkers & Prevention*, 15(1):45, 2006.

[51] Allan Hackshaw. The benefits and harms of mammographic screening for breast cancer: building the evidence base using service screening programmes, 2012.

[52] Awni Y Hannun, Pranav Rajpurkar, Masoumeh Haghpanahi, Geoffrey H Tison, Codie Bourn, Mintu P Turakhia, and Andrew Y Ng. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nature medicine*, 25(1):65, 2019.

[53] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[54] HIPAA. *Health Insurance Portability and Accountability Act of 1996*.

[55] Yangsibo Huang, Zhao Song, Kai Li, and Sanjeev Arora. InstaHide: Instance-hiding schemes for private distributed learning. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4507–4518. PMLR, 13–18 Jul 2020.

[56] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

[57] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 590–597, 2019.

[58] Alistair EW Johnson, Tom J Pollard, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Yifan Peng, Zhiyong Lu, Roger G Mark, Seth J Berkowitz, and Steven Horng. Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*, 2019.

[59] James Jordon, Jinsung Yoon, and Mihaela Van Der Schaar. Pate-gan: Generating synthetic data with differential privacy guarantees. In *International conference on learning representations*, 2018.

[60] Chiraag Juvekar, Vinod Vaikuntanathan, and Anantha Chandrakasan. Gazelle: A low latency framework for secure neural network inference. In *Proceedings of the 27th USENIX Conference on Security Symposium*, SEC'18, page 1651–1668, USA, 2018. USENIX Association.

[61] Jared L Katzman, Uri Shaham, Alexander Cloninger, Jonathan Bates, Tingting Jiang, and Yuval Kluger. Deepsurv: personalized treatment recommender system using a cox proportional hazards deep neural network. *BMC medical research methodology*, 18(1):24, 2018.

[62] Hyo-Eun Kim, Hak Hee Kim, Boo-Kyung Han, Ki Hwan Kim, Kyunghwa Han, Hyeonseob Nam, Eun Hye Lee, and Eun-Kyung Kim. Changes in cancer detection and false-positive recall in mammography using artificial intelligence: a retrospective, multireader study. *The Lancet Digital Health*, 2(3):e138–e148, 2020.

[63] D. Ko, S. Choi, J. Shin, P. Liu, and Y. Choi. Structural image De-Identification for privacy-Preserving deep learning. *IEEE Access*, 8:119848–119862, 2020.

[64] Trent Kyono, Fiona J Gilbert, and Mihaela van der Schaar. Mammo: A deep learning solution for facilitating radiologist-machine collaboration in breast cancer diagnosis. *arXiv preprint arXiv:1811.02661*, 2018.

[65] Jiangwei Lao, Yinsheng Chen, Zhi-Cheng Li, Qihua Li, Ji Zhang, Jing Liu, and Guangtao Zhai. A deep learning-based radiomics model for prediction of survival in glioblastoma multiforme. *Scientific reports*, 7(1):1–8, 2017.

[66] M Le BoulcÕh, A Bekhouche, E Kermarrec, A Milon, C Abdel Wahab, S Zilberman, N Chabbert-Buffet, and I Thomassin-Naggara. Comparison of breast density assessment between human eye and automated software on digital and synthetic mammography: Impact on breast cancer risk. *Diagnostic and Interventional Imaging*, 2020.

[67] Constance D Lehman, Robert D Wellman, Diana SM Buist, Karla Kerlikowske, Anna NA Tosteson, and Diana L Miglioretti. Diagnostic accuracy of digital screening mammography with and without computer-aided detection. *JAMA internal medicine*, 175(11):1828–1837, 2015.

[68] Constance D Lehman, Adam Yala, Tal Schuster, Brian Dontchos, Manisha Bahl, Kyle Swanson, and Regina Barzilay. Mammographic breast density assessment using deep learning: clinical implementation. *Radiology*, 290(1):52–58, 2019.

[69] Bo Liu, Ming Ding, Hanyu Xue, Tianqing Zhu, Dayong Ye, Li Song, and Wanlei Zhou. Dp-image: Differential privacy for image data in feature space. *arXiv preprint arXiv:2103.07073*, 2021.

[70] Jian Liu, Mika Juuti, Yao Lu, and N. Asokan. Oblivious neural network predictions via minionn transformations. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, CCS '17, page 619–631, New York, NY, USA, 2017. Association for Computing Machinery.

[71] Guillermo Lloret-Talavera, Marc Jorda, Harald Servat, Fabian Boemer, Chetan Chauhan, Shigeki Tomishima, Nilesh N Shah, and Antonio J Pena. Enabling homomorphically encrypted inference for large dnn models. *IEEE Transactions on Computers*, 2021.

[72] William Lotter, Greg Sorensen, and David Cox. A multi-scale cnn and curriculum learning strategy for mammogram classification. In *Deep Learning in Medical*

Image Analysis and Multimodal Learning for Clinical Decision Support, pages 169–177. Springer, 2017.

[73] Michael T Lu, Vineet K Raghu, Thomas Mayrhofer, Hugo JWL Aerts, and Udo Hoffmann. Deep learning using chest radiographs to identify high-risk smokers for lung cancer screening computed tomography: development and validation of a prediction model. *Annals of Internal Medicine*, 2020.

[74] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

[75] Lisa M Maillart, Julie Simmons Ivy, Scott Ransom, and Kathleen Diehl. Assessing dynamic breast cancer screening policies. *Operations Research*, 56(6):1411–1427, 2008.

[76] Jeanne S Mandelblatt, Natasha K Stout, Clyde B Schechter, Jeroen J Van Den Broek, Diana L Miglioretti, Martin Krapcho, Amy Trentham-Dietz, Diego Munoz, Sandra J Lee, Donald A Berry, et al. Collaborative modeling of the benefits and harms associated with different us breast cancer screening strategies. *Annals of internal medicine*, 164(4):215–225, 2016.

[77] Michael G Marmot, DG Altman, DA Cameron, JA Dewar, SG Thompson, and Maggie Wilcox. The benefits and harms of breast cancer screening: an independent review. *British journal of cancer*, 108(11):2205–2240, 2013.

[78] J.L. Massey. Guessing and entropy. In *IEEE International Symposium on Information Theory*, page 204, 1994.

[79] Scott Mayer McKinney, Marcin Sieniek, Varun Godbole, Jonathan Godwin, Natasha Antropova, Hutan Ashrafian, Trevor Back, Mary Chesus, Greg C Corrado, Ara Darzi, et al. International evaluation of an ai system for breast cancer screening. *Nature*, 577(7788):89–94, 2020.

[80] Neri Merhav and Asaf Cohen. Universal randomized guessing with application to asynchronous decentralized brute—force attacks. In *IEEE International Symposium on Information Theory (ISIT)*, pages 485–489, 2019.

[81] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533, 2015.

[82] Payman Mohassel and Yupeng Zhang. Secureml: A system for scalable privacy-preserving machine learning. In *2017 IEEE Symposium on Security and Privacy, SP 2017, San Jose, CA, USA, May 22-26, 2017*, pages 19–38. IEEE Computer Society, 2017.

[83] Debra L Monticciolo, Mary S Newell, R Edward Hendrick, Mark A Helvie, Linda Moy, Barbara Monsees, Daniel B Kopans, Peter R Eby, and Edward A Sickles. Breast cancer screening for average-risk women: recommendations from the acr commission on breast imaging. *Journal of the American College of Radiology*, 14(9):1137–1143, 2017.

[84] Debra L Monticciolo, Mary S Newell, Linda Moy, Bethany Niell, Barbara Monsees, and Edward A Sickles. Breast cancer screening in women at higher-than-average risk: recommendations from the acr. *Journal of the American College of Radiology*, 15(3):408–414, 2018.

[85] Virginia A Moyer. Screening for lung cancer: Us preventive services task force recommendation statement. *Annals of internal medicine*, 160(5):330–338, 2014.

[86] Cristina O'Donoghue, Martin Eklund, Elissa M Ozanne, and Laura J Esserman. Aggregate cost of mammography screening in the united states: comparison of current practice and advocated guidelines. *Annals of internal medicine*, 160(3):145–153, 2014.

[87] Kevin C Oeffinger, Elizabeth TH Fontham, Ruth Etzioni, Abbe Herzig, James S Michaelson, Ya-Chen Tina Shih, Louise C Walter, Timothy R Church, Christopher R Flowers, Samuel J LaMonte, et al. Breast cancer screening for women at average risk: 2015 guideline update from the american cancer society. *Jama*, 314(15):1599–1614, 2015.

[88] Douglas K Owens, Karina W Davidson, Alex H Krist, Michael J Barry, Michael Cabana, Aaron B Caughey, Chyke A Doubeni, John W Epling, Martha Kubik, C Seth Landefeld, et al. Medication use to reduce risk of breast cancer: Us preventive services task force recommendation statement. *Jama*, 322(9):857–867, 2019.

[89] Nicolas Papernot, Martín Abadi, Ulfar Erlingsson, Ian Goodfellow, and Kunal Talwar. Semi-supervised knowledge transfer for deep learning from private training data. *arXiv preprint arXiv:1610.05755*, 2016.

[90] Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Łukasz Kaiser, Noam Shazeer, Alexander Ku, and Dustin Tran. Image transformer. *arXiv preprint arXiv:1802.05751*, 2018.

[91] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

[92] C.E. Pfister and W.G. Sullivan. Renyi entropy, guesswork moments, and large deviations. *IEEE Transactions on Information Theory*, 50(11):2794–2800, 2004.

[93] John Platt et al. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74, 1999.

[94] Ryan Poplin, Avinash V Varadarajan, Katy Blumer, Yun Liu, Michael V McConnell, Greg S Corrado, Lily Peng, and Dale R Webster. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. *Nature Biomedical Engineering*, 2(3):158, 2018.

[95] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2018.

[96] Nisarg Raval, Ashwin Machanavajjhala, and Jerry Pan. Olympus: Sensor privacy through utility aware obfuscation. *Proc. Priv. Enhancing Technol.*, 2019(1):5–25, 2019.

[97] Dezső Ribli, Anna Horváth, Zsuzsa Unger, Péter Pollner, and István Csabai. Detecting and classifying lesions in mammograms with deep learning. *Scientific reports*, 8(1):1–7, 2018.

[98] Xavier Robin, Natacha Turck, Alexandre Hainard, Natalia Tiberti, Frédérique Lisacek, Jean-Charles Sanchez, and Markus Müller. proc: an open-source package for r and s+ to analyze and compare roc curves. *BMC bioinformatics*, 12(1):1–8, 2011.

[99] Alejandro Rodriguez-Ruiz, Kristina Lång, Albert Gubern-Merida, Mireille Broeders, Gisella Gennaro, Paola Clauser, Thomas H Helbich, Margarita Chevalier, Tao Tan, Thomas Mertelmeier, et al. Stand-alone artificial intelligence for breast cancer detection in mammography: comparison with 101 radiologists. *JNCI: Journal of the National Cancer Institute*, 111(9):916–922, 2019.

[100] Diederik M Roijers, Peter Vamplew, Shimon Whiteson, and Richard Dazeley. A survey of multi-objective sequential decision-making. *Journal of Artificial Intelligence Research*, 48:67–113, 2013.

[101] Ronald K Ross, Annlia Paganini-Hill, Peggy C Wan, and Malcolm C Pike. Effect of hormone replacement therapy on breast cancer risk: estrogen versus estrogen plus progestin. *Journal of the National Cancer Institute*, 92(4):328–332, 2000.

[102] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

[103] Talya Salz, Alice R Richman, and Noel T Brewer. Meta-analyses of the effect of false-positive mammograms on generic and specific psychosocial outcomes. *Psycho-Oncology*, 19(10):1026–1034, 2010.

[104] Thomas Schaffter, Diana SM Buist, Christoph I Lee, Yaroslav Nikulin, Dezső Ribli, Yuanfang Guan, William Lotter, Zequn Jie, Hao Du, Sijia Wang, et al. Evaluation of combined artificial intelligence and radiologist assessment to interpret screening mammograms. *JAMA network open*, 3(3):e200265–e200265, 2020.

[105] John T Schousboe, Karla Kerlikowske, Andrew Loh, and Steven 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*, 155(1):10–20, 2011.

[106] Li Shen, Laurie R Margolies, Joseph H Rothstein, Eugene Fluder, Russell McBride, and Weiva Sieh. Deep learning to improve breast cancer detection on screening mammography. *Scientific reports*, 9(1):1–12, 2019.

[107] Yiqiu Shen, Nan Wu, Jason Phang, Jungkyu Park, Kangning Liu, Sudarshini Tyagi, Laura Heacock, S Kim, Linda Moy, Kyunghyun Cho, et al. An interpretable classifier for high-resolution breast cancer screening images utilizing weakly supervised localization. *arXiv preprint arXiv:2002.07613*, 2020.

[108] W. Sirichotedumrong, T. Maekawa, Y. Kinoshita, and H. Kiya. Privacy-preserving deep neural networks with pixel-based image encryption considering data augmentation in the encrypted domain. In *IEEE International Conference on Image Processing (ICIP)*, pages 674–678, 2019.

[109] Albert L Siu. Screening for breast cancer: Us preventive services task force recommendation statement. *Annals of internal medicine*, 164(4):279–296, 2016.

[110] Robert A Smith, Kimberly S Andrews, Durado Brooks, Stacey A Fedewa, Deana Manassaram-Baptiste, Debbie Saslow, and Richard C Wender. Cancer screening in the united states, 2019: A review of current american cancer society guidelines and current issues in cancer screening. *CA: a cancer journal for clinicians*, 69(3):184–210, 2019.

[111] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. *arXiv preprint arXiv:1703.01365*, 2017.

[112] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[113] László Tabár, Bedrich Vitak, Tony Hsiu-Hsi Chen, Amy Ming-Fang Yen, Anders Cohen, Tibor Tot, Sherry Yueh-Hsia Chiu, Sam Li-Sheng Chen, Jean Ching-Yuan Fann, Johan Rosell, et al. Swedish two-county trial: impact of mammographic screening on breast cancer mortality during 3 decades. *Radiology*, 260(3):658–663, 2011.

[114] László Tabár, Amy Ming-Fang Yen, Wendy Yi-Ying Wu, Sam Li-Sheng Chen, Sherry Yueh-Hsia Chiu, Jean Ching-Yuan Fann, May Mei-Sheng Ku, Robert A

Smith, Stephen W Duffy, and Tony Hsiu-Hsi Chen. Insights from the breast cancer screening trials: how screening affects the natural history of breast cancer and implications for evaluating service screening programs. *The breast journal*, 21(1):13–20, 2015.

[115] Laszlo Tabar, Ming-Fang Yen, Bedrich Vitak, Hsiu-Hsi Tony Chen, Robert A Smith, and Stephen W Duffy. Mammography service screening and mortality in breast cancer patients: 20-year follow-up before and after introduction of screening. *The Lancet*, 361(9367):1405–1410, 2003.

[116] M. Tanaka. Learnable image encryption. In *IEEE International Conference on Consumer Electronics-Taiwan (ICCE-TW)*, pages 1–2, 2018.

[117] National Lung Screening Trial Research Team. The national lung screening trial: overview and study design. *Radiology*, 258(1):243–253, 2011.

[118] National Lung Screening Trial Research Team. Reduced lung-cancer mortality with low-dose computed tomographic screening. *New England Journal of Medicine*, 365(5):395–409, 2011.

[119] Reihaneh Torkzadehmahani, Peter Kairouz, and Benedict Paten. Dp-cgan: Differentially private synthetic data and label generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.

[120] Anna NA Tosteson, Dennis G Fryback, Cristina S Hammond, Lucy G Hanna, Margaret R Grove, Mary Brown, Qianfei Wang, Karen Lindfors, and Etta D Pisano. Consequences of false-positive screening mammograms. *JAMA internal medicine*, 174(6):954–961, 2014.

[121] Anna NA Tosteson, Natasha K Stout, Dennis G Fryback, Suddhasatta Acharyya, Benjamin A Herman, Lucy G Hannah, and Etta D Pisano. Cost-effectiveness of digital mammography breast cancer screening. *Annals of internal medicine*, 148(1):1–10, 2008.

[122] Amy Trentham-Dietz, Karla Kerlikowske, Natasha K Stout, Diana L Miglioretti, Clyde B Schechter, Mehmet Ali Ergun, Jeroen J Van Den Broek, Oguzhan Alagoz, Brian L Sprague, Nicolien T Van Ravesteyn, et al. Tailoring breast cancer screening intervals by breast density and risk for women aged 50 years or older: collaborative modeling of screening outcomes. *Annals of internal medicine*, 165(10):700–712, 2016.

[123] Jonathan Tyrer, Stephen W Duffy, and Jack Cuzick. A breast cancer prediction model incorporating familial and personal risk factors. *Statistics in medicine*, 23(7):1111–1130, 2004.

[124] Hajime Uno, Tianxi Cai, Michael J Pencina, Ralph B D'Agostino, and LJ Wei. On the c-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Statistics in medicine*, 30(10):1105–1117, 2011.

[125] US Food and Drug Administration. Mammography quality standards act and program. https://www.fda.gov/radiation-emitting-products/mqsa-insights/mqsa-national-statistics.

[126] Maartje van Seijen, Esther H Lips, Alastair M Thompson, Serena Nik-Zainal, Andrew Futreal, E Shelley Hwang, Ellen Verschuur, Joanna Lane, Jos Jonkers, Daniel W Rea, et al. Ductal carcinoma in situ: to treat or not to treat, that is the question. *British journal of cancer*, page 1, 2019.

[127] Kala Visvanathan, Rowan T Chlebowski, Patricia Hurley, Nananda F Col, Mary Ropka, Deborah Collyar, Monica Morrow, Carolyn Runowicz, Kathleen I Pritchard, Karen Hagerty, et al. American society of clinical oncology clinical practice guideline update on the use of pharmacologic interventions including tamoxifen, raloxifene, and aromatase inhibition for breast cancer risk reduction. *Journal of clinical oncology*, 27(19):3235, 2009.

[128] Chao Wang, Adam R Brentnall, James G Mainprize, Martin Yaffe, Jack Cuzick, and Jennifer A Harvey. External validation of a mammographic texture marker for breast cancer risk in a case–control study. *Journal of Medical Imaging*, 7(1):014003, 2020.

[129] Wei Wang, Jinyang Gao, Meihui Zhang, Sheng Wang, Gang Chen, Teck Khim Ng, Beng Chin Ooi, Jie Shao, and Moaz Reyad. Rafiki: Machine learning as an analytics service system. *Proc. VLDB Endow.*, 12(2):128–140, Oct. 2018.

[130] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2097–2106, 2017.

[131] Karen J Wernli, Nora B Henrikson, Caitlin C Morrison, Matthew Nguyen, Gaia Pocobelli, and Paula R Blasi. Screening for skin cancer in adults: updated evidence report and systematic review for the us preventive services task force. *Jama*, 316(4):436–447, 2016.

[132] Timothy J Wilt, Russell P Harris, and Amir Qaseem. Screening for cancer: advice for high-value care from the american college of physicians. *Annals of internal medicine*, 162(10):718–725, 2015.

[133] Eric Wu, Kevin Wu, David Cox, and William Lotter. Conditional infilling gans for data augmentation in mammogram classification. In *Image Analysis for Moving Organ, Breast, and Thoracic Images*, pages 98–106. Springer, 2018.

[134] Kevin Wu, Eric Wu, Yaping Wu, Hongna Tan, Greg Sorensen, Meiyun Wang, and Bill Lotter. Validation of a deep learning mammography model in a population with low screening rates. *arXiv preprint arXiv:1911.00364*, 2019.

[135] Nan Wu, Jason Phang, Jungkyu Park, Yiqiu Shen, Zhe Huang, Masha Zorin, Stanisław Jastrzebski, Thibault Févry, Joe Katsnelson, Eric Kim, et al. Deep neural networks improve radiologists' performance in breast cancer screening. *IEEE transactions on medical imaging*, 39(4):1184–1194, 2019.

[136] Zhenyu Wu, Haotao Wang, Zhaowen Wang, Hailin Jin, and Zhangyang Wang. Privacy-preserving deep action recognition: An adversarial learning framework and a new dataset. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[137] Zhenyu Wu, Zhangyang Wang, Zhaowen Wang, and Hailin Jin. Towards privacy-preserving visual recognition via adversarial training: A pilot study. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 606–624, 2018.

[138] Hanshen Xiao and Srinivas Devadas. The art of labeling: Task augmentation for private (collaborative) learning on transformed data. *Cryptology ePrint Archive*, 2021.

[139] Hanshen Xiao and Srinivas Devadas. Dauntless: Data augmentation and uniform transformation for learning with scalability and security. Cryptology ePrint Archive, Report 2021/201, 2021. `https://eprint.iacr.org/2021/201`.

[140] Taihong Xiao, Yi-Hsuan Tsai, Kihyuk Sohn, Manmohan Chandraker, and Ming-Hsuan Yang. Adversarial learning of privacy-preserving and task-oriented representations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12434–12441, 2020.

[141] Liyang Xie, Kaixiang Lin, Shu Wang, Fei Wang, and Jiayu Zhou. Differentially private generative adversarial network. *arXiv preprint arXiv:1802.06739*, 2018.

[142] Adam Yala, Regina Barzilay, Laura Salama, Molly Griffin, Grace Sollender, Aditya Bardia, Constance Lehman, Julliette M Buckley, Suzanne B Coopey, Fernanda Polubriaginof, et al. Using machine learning to parse breast pathology reports. *Breast cancer research and treatment*, 161(2):203–211, 2017.

[143] Adam Yala, Constance Lehman, Tal Schuster, Tally Portnoi, and Regina Barzilay. A deep learning mammography-based model for improved breast cancer risk prediction. *Radiology*, 292(1):60–66, 2019.

[144] Adam Yala, Peter G Mikhael, Constance Lehman, Gigin Lin, Fredrik Strand, Yung-Liang Wan, Kevin Hughes, Siddharth Satuluru, Thomas Kim, Imon Banerjee, et al. Optimizing risk-based breast cancer screening policies with reinforcement learning. *Nature medicine*, pages 1–8, 2022.

[145] Adam Yala, Peter G Mikhael, Fredrik Strand, Gigin Lin, Siddharth Satuluru, Thomas Kim, Imon Banerjee, Judy Gichoya, Hari Trivedi, Constance D Lehman, et al. Multi-institutional validation of a mammography-based breast cancer risk model. *Journal of Clinical Oncology*, pages JCO–21, 2021.

[146] Adam Yala, Peter G Mikhael, Fredrik Strand, Gigin Lin, Kevin Smith, Yung-Liang Wan, Leslie Lamb, Kevin Hughes, Constance Lehman, and Regina Barzilay. Toward robust mammography-based models for breast cancer risk. *Science Translational Medicine*, 13(578):eaba4373, 2021.

[147] Adam Yala, Victor Quach, Homa Esfahanizadeh, Rafael GL D'Oliveira, Ken R Duffy, Muriel Médard, Tommi S Jaakkola, and Regina Barzilay. Syfer: Neural obfuscation for private data release. *arXiv preprint arXiv:2201.12406*, 2022.

[148] Adam Yala, Tal Schuster, Randy Miles, Regina Barzilay, and Constance Lehman. A deep learning model to triage screening mammograms: a simulation study. *Radiology*, 293(1):38–46, 2019.

[149] Runzhe Yang, Xingyuan Sun, and Karthik Narasimhan. A generalized algorithm for multi-objective reinforcement learning and policy adaptation. In *Advances in Neural Information Processing Systems*, pages 14636–14647, 2019.

[150] Andrew Chi-Chih Yao. How to generate and exchange secrets (extended abstract). In *27th Annual Symposium on Foundations of Computer Science, Toronto, Canada, 27-29 October 1986*, pages 162–167. IEEE Computer Society, 1986.

[151] Steven Zeitchik. Is artificial intelligence about to transform the mammogram? *The Washington Post*, 2021.

[152] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018.

[153] Mingmin Zhao, Shichao Yue, Dina Katabi, Tommi S Jaakkola, and Matt T Bianchi. Learning sleep stages from radio signals: A conditional adversarial architecture. In *International Conference on Machine Learning*, pages 4100–4109, 2017.

[154] Daquan Zhou, Bingyi Kang, Xiaojie Jin, Linjie Yang, Xiaochen Lian, Zihang Jiang, Qibin Hou, and Jiashi Feng. Deepvit: Towards deeper vision transformer. *arXiv preprint arXiv:2103.11886*, 2021.