Ben-Gurion University of the Negev

PhD Proposal

Improving Cardiovascular Disease Prediction

שיפור חיזוי מחלה קרדיווסקולרית

Noam Barda, MD

January 3, 2019

Supervisor's Signature, Prof. Nadav Davidovich: _____

Supervisor's Signature, Prof. Eitan Bachmat: _____

Chairman of Departmental Graduate Studies's Signature: _____

Contents

1 Abstract **3**

1.1 Background . . . 3

1.2 Goals . . . 3

1.3 Methods . . . 3

1.4 Importance . . . 3

1.5 Keywords . . . 3

2 תקציר בעברית **4**

2.1 רקע . . . 4

2.2 מטרות . . . 4

2.3 שיטות . . . 4

2.4 חשיבות . . . 4

2.5 מילות מפתח . . . 4

3 Aim of the Thesis **4**

4 Importance and Background **5**

4.1 General Background . . . 5

4.1.1 Epidemiology of Cardiovascular Disease and Stroke . . . 5

4.1.2 History of multivariable Risk Models . . . 6

4.2 Part I . . . 6

4.2.1 Coronary Artery Calcium . . . 6

4.2.2 Integration of CAC Scores into CVD Prediction Models . . . 7

4.2.3 Medical Computer Vision . . . 7

4.2.4 Automatic CAC Scoring . . . 7

4.2.5 The Scientific Gap . . . 7

4.3 Part II . . . 7

4.3.1 Methodology of Traditional Risk Models . . . 7

4.3.2 Generalized Linear Models . . . 8

4.3.3 Cox Proportional Hazards Model . . . 8

4.3.4 The Rise of AI and Machine Learning . . . 8

4.3.5 Electronic Health Record based Observational Studies . . . 9

4.3.6 The Scientific Gap . . . 9

4.4 Part III . . . 9

4.4.1 Aim of Risk Models . . . 9

4.4.2 Correlation is not Causation . . . 10

4.4.3 Causal Inference . . . 10

4.4.4 The Scientific Gap . . . 10

5 The Novelty of the Thesis **11**

6 Published Work **11**

7 Research Methodology **12**

7.1 Important Concepts . . . 12

7.1.1 Source of Data for Study . . . 12

7.1.2 Issues with EHR Data . . . 12

7.1.3 Data Extraction Principles . . . 12

7.1.4 Migration of Foreign-defined Variables . . . 13

7.1.5 LASSO Regression . . . 13

7.1.6 Causal Forests . . . 14

7.1.7 Linear Recalibration . . . 14

7.1.8 Imputation . . . 14

7.1.9 The Bootstrap . . . 15

7.1.10 Net Reclassification Improvement . . . 15

7.1.11 Decision Curves . . . 15

7.2 Analysis Plan . . . 15

7.2.1 Part I . . . 15

7.2.2 Part II . . . 17

7.2.3 Part III . . . 19

7.3 Ethics . . . 20

8 Preliminary Results **20**

8.1 AHA/ACC 2013 model . . . 20

9 References **21**

Appendices **27**

A Model Variable Lists **27**

B Extraction Protocol **28**

B.1 Outcome Diagnoses . . . 28

B.2 Causes of Death . . . 29

B.3 Background Diagnoses . . . 29

B.4 Drugs . . . 31

C Preliminary Result Tables, Graphs and Drawings **32**

C.1 AHA/ACC 2013 Population Table . . . 32

C.2 AHA/ACC 2013 Risk Model Result Graph . . . 32

C.3 Clalit Model Population Flow Chart . . . 33

1 Abstract

1.1 Background

Cardiovascular disease (CVD), including myocardial infarction (MI) and cerebrovascular accident (CVA), remains a significant cause of morbidity and mortality[1] worldwide, despite reduced incidence in the developed world in recent years[2, 3].

Since the early 1990s, and more so in the last few years, multivariable risk prediction models have been created to estimate patients' risk for cardiovascular disease (e.g. [4, 5, 6]). These models are used to identify patients at risk and are capable of exact risk quantification over time[7]. Through their many variations, these risk models are included in different guidelines and occupy an important place in prevention of CVD[7].

1.2 Goals

The goal of this thesis is to examine different methods by which to improve CVD prediction. We will examine ways to utilize novel biomarkers, to generate better and more generic prediction models and to combine prediction with causal inference.

1.3 Methods

The baseline model used will be the 2013 American Heart Association/American College of Cardiology (AHA/ACC) pooled cohort equations[7]. This model is of great importance, and is used in the AHA/ACC guidelines and the United States Preventive Services Task Force (USPSTF) guidelines for CVD prevention[8]. The population of all studies is the general population of Clalit Health Services (CHS), the largest sick fund in Israel.

We will first employ automated coronary calcium scores in an attempt to improve the prediction. These scores will be generated automatically from existing chest CT scans via a computer vision machine learning algorithm. Next, We will construct an automatic generic algorithm to generate prediction models using little-to-none human intervention. This algorithm will extract variables from the database, choose the most relevant ones and construct a model while performing hyperparameter tuning. Lastly, We will use causal inference techniques for estimating the Average Treatment Effect (ATE) and Individual Treatment Effect (ITE) to estimate the effect of intervening on different patient characteristics.

1.4 Importance

CVD prediction is crucial for CVD prevention. All methods suggested in this thesis hold a promise for improving CVD prediction and allowing better tailoring of intervention techniques.

1.5 Keywords

Medical Prediction Models, Generic Prediction, Computer Vision, Cardiovascular Disease, Electronic Health Records, Causal Inference

2 תקציר בעברית

2.1 רקע

מחלה קרדיווסקולרית, הכוללת אוטם לבבי ושבץ מוחי, עודה סיבה מובילה לחולי ותמותה[1] ברחבי העולם, וזאת חרף ירידה בהיארעותה בעולם המערבי בשנים האחרונות[2, 3].

מתחילת שנות ה-90 החלה, ובשנים האחרונות גברה, יצירתם של מודלים רב-משתניים לחישוב סיכון למחלות שונות(לדוגמה [4, 5, 6]). מודלים אלו משמשים לזיהוי חולים בסיכון, ומאפשרים כימות מדויק של הסיכון לאורך שנים רבות[7]. כיום, בצורותיהן השונות, מודלים אלו כלולים בקווים המנחים של ארגונים מקצועיים רבים, ולהם מקום חשוב הן במניעה הראשונית והן באבחנה של מחלות [9, 7].

2.2 מטרות

will We prediction. CVD improve to which by methods different examine to is thesis this of goal The to and models prediction generic more and better generate to biomarkers, novel utilize to ways examine inference. causal with prediction combine

מטרתה של תזה זו היא לבחון דרכים שונות לשפר חיזוי מחלה קרדיווסקולרית. אנו נבחן אמצעים לשלב ביומרקרים חדשניים, לייצר מודלי חיזוי טובים וגנריים יותר ולשלב חיזוי עם הסקה סיבתית.

2.3 שיטות

המודל הבסיסי בו נעשה שימוש בעבודה זו הוא מודל האגודה האמריקאית למחלה לב משנת 2013[7]. למודל זה חשיבות רבה, והוא משמש בקווים המנחים למניעת מחלה קרדיווסקולרית של האגודה האמריקאית למחלות לב ושל הצוות האמריקאי לשירותים מניעתיים[8]. האוכלוסיה בכל המחקרים תהיה אוכלוסיית המבוטחים של קופת החולים כללית, הקופה הגדולה בישראל.

בתחילה נשתמש בניקוד לכמות הסידן בעורקי הלב על מנת לשפר החיזוי. ניקוד זה יופק באופן אוטומטי ב-CT חזה קיימים באמצעות אלגוריתמי למידת מכונה לראיה ממוחשבת. לאחר מכן אנו נבנה אלגוריתם אוטומטי אשר תפקידו לייצר מודלי חיזוי עם מינימום התערבות אנושית. אלגוריתם זה ישלוף המשתנים מבסיס הנתונים, יבחר הרלוונטיים מהם ביותר ויבנה מודל תוך ביצוע טיוב לפרמטרים. לבסוף, אנו ניעזר בטכניקות של הסקה סיבתית על מנת להעריך האפקט הממוצע והאישי של התערבויות שונות בפרט להפחתת סיכון קרדיווסקולרי.

2.4 חשיבות

חיזוי מחלה קרדיווסקולרית חיונית למניעתה. לכל השיטות המתוארות בעבודה זו היכולת לשפר חיזוי מחלה קרדיווסקולרית ולאפשר החלטות מושכלות יותר בנוגע למניעה.

2.5 מילות מפתח

מודלי חיזוי רפואיים, חיזוי גנרי, ראיה ממוחשבת, מחלה קרדיווסקולרית, תיק רפואי ממוחשב, הסקה סיבתית.

3 Aim of the Thesis

The main aim of this thesis is to examine different methods of improving cardiovascular disease risk prediction.

Three methods will be explored:

Use of Automatically Generated Biomarkers We will examine the effect of incorporating into the model a novel biomarker based on the coronary calcium score. This biomarker will be automatically generated using a machine learning algorithm from existing chest CTs.

Generic Risk Prediction A modern and novel approach to develop risk models based on Electronic Health Record (EHR) data will be developed. The full details of this approach will be detailed below, under "Research Methodology", but briefly, it will require no preliminary domain expertise, instead utilizing modern methods to simultaneously choose variables and create the model based on them.

Determining Intervention Effects CVD Prediction models are often used to decide on interventions designed for disease prevention. But by themselves, these models give no information regarding the relative efficacy of such interventions. We will utilize methods designed for causal inference to estimate the expected effect of such interventions at the patient level.

Based on these aims, we hypothesize:

1. That automatically generated biomarkers from existing scans will improve CVD prediction with no patient harms.
2. That using less pre-specification of risk factors, and allowing a computerized algorithm to select risk factors in an autonomous fashion, will enable detection of novel risk factors, whose inclusion in future risk models will improve their performance.
3. That causal inference methods can be used to estimate not only a patient's risk, but also the change in risk given certain interventions.

4 Importance and Background

We will survey the pertinent background in general and then for each step in turn, highlighting the gap in existing knowledge to which we seek to contribute.

4.1 General Background

4.1.1 Epidemiology of Cardiovascular Disease and Stroke

In its usual definition, cardiovascular disease (CVD) includes several disease categories[10]:

1. Coronary Heart Disease
2. Cerebrovascular Disease
3. Peripheral Artery Disease
4. Aortic Disease
5. Rheumatic Heart Disease

6. Congenital Heart Disease
7. Venous Thromboembolism

CVD is very common. Lifetime risk for people aged 30 with no prior cardiovascular disease approaches 50 percent[11], with coronary heart disease being the most common specific diagnosis[12].

While the rates of cardiovascular disease have declined in developed countries over the last 30 years[2, 3], they remain significant public health problems, being the second most common cause of mortality and third most common cause of disability worldwide[13]. The statistics in Israel are similar[14].

Among diseases with such a significant public health impact, cardiovascular disease stands out in two ways. First, its risk factors are well understood, with 90% of its population-attributable-risk caused by nine risk factors. It's also a very preventable disease, as these risk factors are mostly preventable[15, 1]: Smoking, dyslipidemia, hypertension, diabetes, etc.

4.1.2 History of multivariable Risk Models

These unique characteristics have made CVD the main outcome in risk models, when such models began to enter clinical practice in the 1990s[4, 16, 5, 17, 6, 18, 7]. Still the most notable of said risk models is the Framingham risk model family, developed on a US population in Massachusetts, Boston[4], and the SCORE risk model, developed in 2003 on a European population[5].

Perhaps more important that their mere existence, is that these models have made their way into widely-accepted international guidelines, with their use mandated in routine clinical care. Two examples we'll cite are the use of these risk models in deciding on Statin therapy[7] and their use in deciding on anti-platelet therapy[8], both for primary prevention of CVD.

While CVD prediction was the bedrock for clinical risk models, they have since spread to encompass a large variety of diseases categories[19, 20], and have found use not only in prediction, but also in diagnosis[21]. This increasingly important place taken by risk models has brought about the publication of guidelines designed to regulate and improve their creation[22]. As estimating the probability for existing and future disease is a significant portion of the clinical process[23], and as this task can in large parts be automated, it seems likely that risk models will gain an increasingly important place in the medical practice.

4.2 Part I

4.2.1 Coronary Artery Calcium

Atherosclerotic disease is the main cause of CVD. As sclerotic plaques in the arteries bind calcium, which in turn is visible on imaging studies, quantification of coronary artery calcium (CAC) via CT scans can be used to quantify the extent of atherosclerotic disease in the coronary arteries. It has been demonstrated that such CAC quantification can help guide medical therapy and lifestyle modification choices[24, 25, 26, 27, 28].

Traditionally, CAC scoring has required a dedicated electrocardiogram (ECG)-gated cardiac computed tomography (CT) scan performed with and without intravenous contrast at significant financial costs[29]. More recently, several studies have demonstrated good inter-technique concordance for CAC scoring on low-dose, non-gated chest CTs as compared to dedicated, contrast-enhanced and ECG-gated cardiac CTs[30, 31].

4.2.2 Integration of CAC Scores into CVD Prediction Models

On the one hand, it has been shown that integrating CAC scores into existing CVD prediction models yields more accurate predictions[32, 33]. On the other hand, obtaining a CAC score requires human radiologist time for scoring and dedicated CT scans, with their associated radiation exposure and monetary costs. Weighting these benefits and harms, a recent USPSTF statement recommends against the routine incorporation of CAC scores into CVD prediction models[34].

4.2.3 Medical Computer Vision

With the fairly recent rise of deep neural networks, computer performance in vision tasks has risen enormously, and can rival human performance on many tasks. Specifically in the medical field, such neural network-based algorithms are now used to diagnose fractures[35], to review pathology slides[36] and to assist in pneumonia diagnosis[37], among other things. To be trained, these algorithms require annotated data, usually annotated by a human specialist; but once trained, these algorithms require no further human involvement.

4.2.4 Automatic CAC Scoring

It has recently been shown that reliable CAC scoring can be obtained algorithmically from low dose chest CT data[38, 39] and that such measurements were correlated with cardiovascular events in a cohort of individuals undergoing CT screening for lung cancer[40]. It has not been demonstrated if such automated scores are capable of improving on existing multivariable risk prediction models[34].

4.2.5 The Scientific Gap

CAC scores have proven utility for CVD prediction, but obtaining them involves harms that are often unacceptable. Such harms are non-existent when the CAC score is obtain automatically from existing CT scans. We will evaluate the improvement to existing prediction algorithms when such automatically obtained CAC scores are integrated into the model.

4.3 Part II

4.3.1 Methodology of Traditional Risk Models

For traditional medical risk models, two design decisions are ubiquitous[41]:

1. They are based on traditional biostatistical methodology such as generalized linear and cox models.
2. They rely heavily on the use of domain expertise to identify relevant risk factors.

Informally described, we could say that the model is tasked to estimate the relative weights of risk factors, themselves independently pre-identified by domain experts.

4.3.2 Generalized Linear Models

Generalized linear models (GLMs) are parametric models that are generalizations of ordinary linear regression, allowing outcome variables to have non-normal error distributions[42].

While classic linear regression follows the form:

$$E[Y] = x^t\beta$$

GLMs have the form:

$$E[Y] = g^{-1}(x^t\beta)$$

With g being the link function connecting the linear predictor space with the outcome space.

For example, logistic regression uses the logit function as the link, $\mu = \frac{\exp(x^t\beta)}{1+\exp(x^t\beta)}$, while linear regression uses the identity function.

The model then uses a loss function, usually maximum likelihood, to estimate the coefficients of the model. Under certain assumptions, these coefficients can have epidemiological interpretations, such as the coefficients of logistic regression being interpreted as the odds ratio of an exposure for a given outcome. The model can also be used for prediction, disregarding all such assumptions.

4.3.3 Cox Proportional Hazards Model

The cox model is a survival analysis model (that is, it uses a compound outcome of time-to-event data) that is semi-parametric. A baseline hazard (λ_0) is estimated non-parametrically from the data, while a parametric linear hazard model is estimated in parallel[43].

The overall hazard model is thus $\lambda(t) = \lambda_0(t) \cdot x^t\beta$. The hazard itself is a somewhat elusive term rooted in calculus, representing the probability of death at a certain infinitesimal time window assuming survival up to that point. Survival is then one minus the integral of the hazard over time.

Similar to GLMs, the coefficients are estimated using a process of maximum likelihood (dubbed partial likelihood in the context of Cox regression), and under strict assumptions have the interpretation of hazard ratios, similar to odds ratios.

The assumptions for cox regression warrant special mention. While the assumption of linearity is similar to GLMs, cox regression also assumes proportionality - that is, that the hazard ratio between risk factors remains constant over time. This is a very strong assumption that does not always hold. Some models circumvent this assumption at the cost of complexity and loss of interpretability. Just as before, the model can also be used for prediction, disregarding all assumptions.

4.3.4 The Rise of AI and Machine Learning

In recent years the fields of machine and statistical learning have seen a tremendous rise[44]. this growth in machine learning, including predictive modeling, has occurred thanks to three main factors[45]:

- A large increase in the amount of accessible data.

- The development of new algorithms and methods.
- An increase in computation power.

These new methods have several defining characteristics, including:

- The use of a wider range of algorithms, not limited to generalized linear models.
- Less reliance on domain expertise, in essence allowing the algorithm to both find the main risk factors and to estimate their respective weights.
- The need for larger sample sizes, to allow the more complex modeling to occur successfully.

To date, these methods have yet to gain wide-acceptance in medical practice[44, 46].

4.3.5 Electronic Health Record based Observational Studies

Most medical risk models in wide-use were developed based on specialized cohort studies[47]. This has the known advantages of cohort studies, most notably the accurate definition of exposures and outcomes, but is expensive and time-consuming, and by definition only allows inclusion of risk factors that were decided on in advance and measured as part of the study. On the other hand, with the larger availability of EHRs, risk models developed on such data have risen in amount. These models have the known disadvantages of EHR data (first of which are the non-standardized definitions), but offer a wealth of information that in certain cases, including the case in Israel[48], encompasses the full extent of a patient's encounters with the health system[49].

4.3.6 The Scientific Gap

We suggest using the unique availability of widely encompassing EHR data with large historic depth, coupled with modern statistical learning methods, to develop a generic method for generation of risk models based on the Clalit's EHR.

This method will make use of most available EHR data, and will require no pre-specification of risk factors, instead allowing the algorithm to ascertain the relative importance of the different factors by itself. Not only will this allow the creation of accurate risk models, it will also provide a way to automatically identify associations that exist in the EHR and could represent novel risk factors and biological pathways.

We will then use this method to develop a specific model to predict cardiovascular disease.

4.4 Part III

4.4.1 Aim of Risk Models

Invariably, the aim of prediction is prevention. CVD carries a high burden of mortality and morbidity, but as its risk factors are well mapped[15], steps can be taken to mitigate the risk. But - deciding on proper interventions requires knowledge of not only the absolute risk, but also the relative effectiveness of the different possible interventions. Coupled with the harms, this allows proper decision making.

4.4.2 Correlation is not Causation

By themselves, predictive models do not give any information regarding the expected effect of any change. Predictive models model the probability of the outcome given the covariates, and contain no direct information regarding the effect of changing any such covariate.

Mathematically, given outcome Y, covariate vector X and treatment T, predictive models model

$$E[Y|X,T]$$

And subtracting the predictions with the treatment in different states yields us

$$E[Y|T=1] - E[Y|T=0]$$

This is, in general, not equal to the causal effect of T on Y, which is defined as

$$E[Y^1] - E[Y^0]$$

Where Y^i is the counterfactual for $T = i$.

That is to say - correlation is not causation. Regardless, knowledge of the absolute risk without knowledge of the relative effects of different interventions is not sufficient for deciding on interventions.

4.4.3 Causal Inference

In general, the individual treatment effect (ITE) for a specific patient, $Y_i^1 - Y_i^0$ can never be known, as one of the counterfactuals will always be hidden. But, under these assumptions, the average causal effect is identifiable from observational data:

Ignorability Given the covariates, the outcome is independent of the treatment: $Y \perp\!\!\!\perp T|X$

Positivity The probability of treatment is positive for any vector of covariate values: $P(T = 1|X) > 0$

Consistency The intervention is well defined, and exposes the matching counterfactual value: $P(Y^i) = P(Y|T = i)$

Under these assumptions, the causal effect can be identified using a variety of methods well known to epidemiologists (i.e. covariate adjustment, inverse probability weighting, matching). When this effect is estimated conditional on certain values for X, we derive the Conditional Average Treatment Effect (CATE, $E[Y_1|X] - E[Y_0|X]$), which is a useful surrogate for the ITE.

Knowledge of a patient's absolute prediction for CVD, together with the individualized expected benefits, would allow optimal decision making.

4.4.4 The Scientific Gap

We suggest that a predictive model coupled with conditional average treatment effects is a far more powerful tool than just predictions. Such a tool would allow both better decisions by the physician and would motivate the patient to reduce his risks, with full knowledge of the expected benefit. We are not aware of any such tools in existence.

5 The Novelty of the Thesis

All aforementioned aspects of the thesis contain measures of novelty to them:

Use of Automatically Generated Biomarkers Use of existing studies (imaging or otherwise) to derive biomarkers with predictive power could be of great use to patients, who would benefit from more accurate prediction. The automation of the process would also render it feasible for implementation.

Generic Risk Prediction We propose that the methodology by which the model will be developed, and specifically its wide applicability, requiring little human intervention and pre-processing, offers significant advantages. The ability to identify risk factors and construct models for a wide variety of pathologies, some of which "unmapped" in regard to their primary risk factors, offers a promise of better understanding and more focused interventions to prevent these diseases.

Determining Intervention Effects To date, predictive models have focused on prediction alone. While this is useful, it is not the entire picture, and modeling the treatment effects is of paramount importance for better physician and patient decision making.

6 Published Work

The epidemiological characteristics of CVD in general and of stroke in particular are well understood[2, 3], and the dominant risk factors in the population well mapped[15, 1]. This is true both in the developed and in the developing world[13]. It is also true in Israel[14].

The increasingly central role filled out by risk prediction models in medicine has been observed[23], as have the challenges of developing such models based on Electronic Health Record (EHR) data[47, 49]. This rapid rise in the number of risk prediction models has led to the writing of specific guidelines on how to develop such risk models and report their results[22].

Many CVD risk models have been developed in the last 30 years, most prominent of which are the Framingham[4, 16, 6, 7], SCORE[5] and Qrisk[17, 18] families of models. Two of these model families also offer a stroke-specific model[50, 51, 52].

Risk models have been incorporated into guidelines for the prevention, diagnosis and treatment of varying conditions. Specifically for CVD prediction, these risks help decide on cholesterol lowering treatment, anti-platelet treatment and more generally, the intensity of follow-up[16, 9, 7, 8].

The utility of CAC is well known[26], and the debate regarding its use ongoing[53]. Attempts to automatically derive it have been published [38, 39] and tested for correlations with the outcome [24], but never assessed for yield beyond the the existing state-of-the-art model[34].

Much has been written on the advent of AI in general and machine learning in particular. In a relatively short time span, these technologies have penetrated large parts of the domains of modern life, and continue to do so with increasing force[54].

That this process has been relatively slow in medicine is also widely recognized, and many efforts now exist to better incorporate such technologies in health-care[44]. Specifically for risk prediction models, recent literature has emerged that details attempts at developing more generic risk models, though different than the idea proposed here both in method and in goal[55].

Causal inference is not a new methodology [56, 57], but has been garnering much attention lately[58]. While many calculators are published online, allowing patients to modify their risk factors (i.e. [59]), none that we know of incorporates causal inference.

7 Research Methodology

We will divide this section into two parts. In the first part, we'll review important concepts that will be used in the analysis. In the second part, we'll detail the exact research plan for this thesis per each part.

7.1 Important Concepts

7.1.1 Source of Data for Study

The general population for all different parts of the study is the population of patients insured by Clalit Health Services (CHS). CHS is the largest sick fund in Israel, with an insured population of 4.5 active members. Clalit is both an insurer and a provider, directly providing primary care, specialist care, lab, imaging and pharamacy services. Additionally, clalit directly operates several large hospitals. The "attrition rate" (the percentage of patients leaving the sick fund each year) stands on less than 2%, allowing long term follow-up of patients.

The data will be collected using the CHS's electronic health record (EHR). CHS has maintained a comprehensive electronic health record since the year 2000, and has continued to improve it with time. This EHR contains, among others, demographic data, medical data (including clinical covariates, lab results, imaging studies, etc.) and claims data for both services rendered as part of the mandatory health insurance and for services rendered as part of the additive insurance ("Mashlim"). On top of the internal Clalit data, the database also contains external information such as the ministry of interior's causes of death listings and the ministry of health's cancer registry. This comprehensive database, combining both medical and claims data, covers large facets of a person's health.

7.1.2 Issues with EHR Data

The difficulties that arise in conducting observational studies on EHR data are many and well documented: Data inaccuracy, missing data, cohort effects, selection biases, myriad ontologies, etc[60, 61, 49]. Some of these issues, such as missing data, can be partially dealt with using statistical methods (see ahead), while some require in-depth expertise and know-how regarding the data's structure and collection methods, knowledge that can only be acquired through rigorous analysis of it. The Clalit's research institute's (CRI) is the research body for Clalit Health Services, and is thus the main consumer of the clalit's EHR data. This grants the CRI intimate knowledge of the data, as is evidenced by the many studies published in major journals based on the Clalit's database and on the CRI's methods in extracting its information (e.g. [62, 63]).

7.1.3 Data Extraction Principles

- CVD definitions, that are used as the outcome in the different models, will be based on those defined by a consensus committee organized by the CRI and headed by a cardiology

and neurology specialists. These definitions similar to those used outside the CRI, such as by the Israeli acute stroke registry[14] (active within the ICDC).

- Demographic characteristics will be extracted from the Clalit's demographic database. Those that are time-dependent (e.g. age) will be extracted current to the index dates, those that are constantly overridden will be extracted to their latest value (e.g. SES).
- Cause of death will be collected directly from the ministry of interior's causes of death table.
- Clinical covariates will be extracted from their dedicated database. The latest value prior to the index date will be used. Tests that can be used as-is (e.g. systolic blood pressure) will be used as-is. Weights and heights measured within a 3-month span will be joined for the calculation of BMI. Smoking status will be "flattened" to never/present/past to account for partial "pack-years" reporting.
- Lab data will be extracted from the dedicated lab results database, using the latest lab values prior to the index date.
- Diagnoses will be collected from the community (both session and permanent diagnoses), from hospitalizations and from the Clalit's chronic registry[64]. Diagnoses will be extracted based on ICD9 codes, ICPC codes and chronic registry codes. Community diagnoses will be corroborated using free text validation so as to exclude suspicions, etc.
- Drug dispensings will be evaluated using the dedicated pharmacy database. Actual dispensings will be counted (as opposed to prescriptions). Drug adherence will be calculated using drug prescriptions and drug dispensings, with PDC and MPR as the actual statistics[65].
- Health care utilization will be calculated by simply counting and summing the patient's encounters and actual cost, both in the community and in hospitals.

7.1.4 Migration of Foreign-defined Variables

When using models developed abroad, special care will be required to handle variables that are not perfect "fits" for the Clalit's database, for example diagnoses, that are collected based on dedicated physician visits in cohort studies and on ICD codes in EHR based studies, will be collected using a mixture of ICD codes, free text validation and validation using lab measurements (e.g glucose for diabetes) and drug dispensings (e.g. diuretics, ACE inhibitors, beta blockers and calcium channel blockers for hypertension).

7.1.5 LASSO Regression

Least absolute shrinkage and selection operator (LASSO)[66] is a variant of logistic regression that adds a regularization term based on the sum of the absolute values of the coefficients (L_1 norm) to the normal loss function to be optimized. Namely, the model minimizes:

$$\arg\min_{\beta} \sum_i y_i \cdot \hat{y}_i + (1 - y_i) \cdot (1 - \hat{y}_i) + \lambda \sum_i |\beta|_i$$

β being the vector of coefficients and lambda being a regularization parameter. This is the normal logistic regression loss function, summed with a regularization term based on the L_1 norm. Owing to the geometric structure of the L_1 norm, this has the effect of setting many covariates to 0, inducing sparsity. The parameter lambda is selected using cross-validation on the validation set, with predictive performance (e.g. AUROC) as the goal.

As the regularization portion of the loss is dependent on variable scales, variables are normalized to have equal means and standard deviations prior to model fitting.

7.1.6 Causal Forests

Causal Forests[67] are a variant of the widely used prediction oriented random forests[68], meant for estimating conditional average treatment effects. While both methods are based on a "forest" of decision trees, where random forests optimizes prediction accuracy of the outcome in each leaf, causal forests instead maximize treatment effect heterogeneity. By first splitting the sample into a "tree construction sample" and an "effect estimation" sample, this allows estimation of the conditional average treatment effects.

7.1.7 Linear Recalibration

Calibration is the agreement between predicted and observed probabilities. When a model is applied to a population different than its original training set, it tends to be mis-calibrated. Recalibration attempts to deal with this problem.

The framework suggested by Van Houwlingen et al[69] performs linear recalibration for binary prediction models. It uses the predictions from the original model as a sole covariate in a new logistic regression model. The predictions from this new model are then the recalibrated predictions to be used in subsequent phases.

Mathematically, the new model being fit is:

$$\log(\frac{y_i}{1-y_i}) = \gamma\hat{p} + \delta$$

Where $\hat{p}$ are the predictions from the original model, y_i are the outcomes, γ is the slope and δ the intercept.

Conceptually, we take the predictions from the original model, but allow them a new slope and intercept, thus preserving the relative importance of each covariate in the model, with the freedom to reset the global risk.

7.1.8 Imputation

As first explained by Rubin et al.[70], missing data can be one of three types

1. Missing completely at random (MCAR) - the data is missing in a pattern that is unrelated to other variables and to the outcome.
2. Missing at random (MAR) - The data is missing in a pattern that is related to other measured variables.
3. Missing not at random (MNAR) - The data is missing in a pattern that is related to unmeasured variables or to its own value.

The first type can be ignored. The third type can not, in theory, be dealt with. The second type requires handling to avoid bias. Multiple imputation can be used to fill in the missing values while still retaining a measure of the variance created by the act of imputation[70]. Multiple imputation with chained equations (MICE) is currently the standard way to impute data in biostatistical research[71].

7.1.9 The Bootstrap

The bootstrap, as developed by Efron et al.[72], is a resampling technique whereby an original sample of size n is resampled n times with replacement. This process is repeated k times. The resulting k samples of size n can be used to estimate the variance of different statistics in relation to their corresponding population parameters. This method allows determination of standard errors for statistics that lack a theoretical distribution.

Combining multiple imputation and the bootstrap is an open problem in biostatistics. Multiple ways have been suggested and compared empirically on simulated data[73].

7.1.10 Net Reclassification Improvement

Net Reclassification improvement (NRI) aims to measure the improvement in decisions made by using a different prediction model[74]. Continuous sums the total number of patients who have been reclassified in the correct direction (i.e. were given a higher score if they did eventually experience an event, and a lower score otherwise). Categorical NRI does the same thing, but only considers movement around a decision threshold.

7.1.11 Decision Curves

Decision curves, who have recently gained popularity in the medical literature[75], aim to compare different models in relation to the decision made using them on different thresholds. The basic measure used is "Net Benefit", measured as

$$\text{Net Benefit} = \frac{\text{True Positives}}{N} - \frac{\text{False Positives}}{N} \cdot \frac{p_t}{1 - p_t}$$

Where p_t is the threshold used for decision making. Conceptually, net benefit measures a weighted average of true positives and false positives, weighted by the relative importance the clinician and patient give each result. For example, is the decision threshold is 25%, this implies that a true positive is 3 times more important that a false positive. Decision curves are then a plot of net benefit for different thresholds.

7.2 Analysis Plan

We detail each part in turn.

7.2.1 Part I

In this part we will examine the predictive gain of a novel biomarker extracted using automated machine learning algorithms.

Study Design

This is a retrospective cohort study based on electronic health record data.

Study Population

Inclusion Criteria:

- Ages 40-79.
- At least 1 year of continuous membership in the Clalit prior to the index date.
- Continuous membership until the study end date or until death.
- A non-contrast chest CT compatible with CAC estimation.

Exclusion Criteria:

- Past CVD event (as detailed above).

Study Timeline

For a sufficiently large sample size,, the model will predict disease for 5 years after the index date. The index date will be set at 1/6/2012, and follow up will persist until 1/6/2017.

Variables

The variables for this study are the variables used for the AHA/ACC model, as detailed in appendix A, together with the novel biomarker.

Biomarker Derivation

The CAC scores will be derived by a machine learning algorithm similar to that described by Shadmi et al[39]. Briefly described, general non-contrast chest CTs that are not ECG-gated are run through a fully convolutional neural network. The network segments the image, giving each voxel a probability of belonging to a coronary calcium patch. Using a validation set, a threshold is then found that best approximates a "ground truth" Agatston score manually annotated by radiologists. This finalized algorithm is then used to determine the score in the CTs used in this study.

Modeling

1. CAC scores will be generated as detailed above.
2. The population will be divided into a training and test set. The division will be, as customary, 80% to the training set and 20% to the test set.
3. The baseline model will be reconstructed on the training set, and linearly recalibrated as described above.

4. The recalibrated model will then be compared to the same model with the novel biomarker added.
5. We will compare:
    - Area under the receiver operating characteristics (AUROC) curve, or c-statistic.
    - Calibration plots.
    - Brier score, as a combined measure of prediction accuracy.
    - Net Benefit, Sensitivity, Specificity, PPV and NPV for the 3.75% risk threshold.
    - Net Reclassification Improvement, as detailed above.
6. Missing data and confidence intervals (CIs) will be dealt with as detailed by Schomaker et al[73]: for each statistic that requires a CI, the data will be bootstrapped 500 times; for each bootstrap sample, five complete datasets will be generated using multiple imputation performed independently on the bootstrapped sample and the out-of-bag sample. The model will then be fit to the bootstrapped sample and tested on the out-of-bag sample. From these 2,500 samples, the mean will be used as the point estimate, and the 2.5 and 97.5 percentiles used for the CI. For comparative performance, the difference between the two models will be used. Calibration plots will be drawn on a single bootstrap sample.

7.2.2 Part II

In this part we will develop a framework for the generic generation of prediction models. This framework will perform feature selection as part of the modeling process.

Study Design

This is a retrospective cohort study based on electronic health record data.

Study Population

Inclusion Criteria:

- Ages 30-90.
- At least 1 year of continuous membership in the Clalit prior to the index date.
- Continuous membership until the study end date or until death.

Exclusion Criteria:

- Past CVD event (as detailed above).

Study Timeline

As is the standard for cardiovascular disease risk models, the model will predict disease for 10 years after the index date. The index date will be set at 1/1/2008, and follow up will persist until 1/1/2018.

Variables

- Demographics - Taken at index date
  - Sex
  - Age
  - Socioeconomic Status by clinic and by address
  - Country of birth (coalesced into regions when necessary) and immigration date
  - Ethnicity by country of individual's or parents' birth
  - Sector (clinic level data - predominantly Arab / Jewish)
- Clinical Markers - Last result before index date
  - Body Mass Index
  - Glomerular Filtration Rate
  - Blood Pressure
  - Smoking Status
  - Charlson Co-morbidity Index
  - Past Malignancy Status (ever diagnosed with cancer)
- Diagnoses - Before index Date
  - Chronic Diagnosis - From chronic registry
  - Community Diagnosis - At ICD9 level
  - Hospitalization Diagnoses - At ICD9 level
- Medication Prescriptions and Dispensings - Before Index Date
  - All drugs at ATC5 level
- Labs - Last result before Index Date
  - All labs per CHS' coding
- Procedures - Before Index Date
  - All procedures at ministry of health code level

Modeling

1. The population will be divided into a training and test set, to allow for accurate estimation of predictive performance. The division will be, as customary, 80% to the training set and 20% to the test set.
2. Missing data will be imputed once for the training set and 5 times for the test set, as detailed above.
3. a LASSO model will be fit, as detailed above.
4. Cross-validation will be performed on the training set for hyper-parameter tuning.

5. Confidence intervals for the different statistics will be derived by bootstrapping the test set, as detailed above. 500 bootstraps per imputed dataset will be performed. the 2.5% and 97.5% percentiles of the resulting results will be reported as the 95% confidence interval.

6. Model performance on the test set will be reported for the following statistics [76, 77]

    - Area under the receiver operating characteristics (AUROC) curve, or c-statistic, as a measure of discrimination.
    - Calibration slope as a measure of calibration.
    - Brier score, as a combined measure of prediction accuracy.
    - Sensitivity, Specificity, PPV and NPV for the 7.5% risk threshold.
    - Net Reclassification Improvement, as detailed above.

7.2.3 Part III

In this part, we will create a prediction model coupled with conditional average treatment effects.

Study Design

This is a retrospective cohort study based on electronic health record data.

Study Population

Inclusion Criteria:

- Ages 40-79.
- At least 1 year of continuous membership in the Clalit prior to the index date.
- Continuous membership until the study end date or until death.
- A non-contrast chest CT compatible with CAC estimation.

Exclusion Criteria:

- Past CVD event (as detailed above).

Study Timeline

As is the standard for cardiovascular disease risk models, the model will predict disease for 10 years after the index date. The index date will be set at 1/1/2008, and follow up will persist until 1/1/2018.

Variables

The variables for this study are the variables used for the AHA/ACC model, as detailed in appendix A.

Modeling

1. The baseline prediction will be the the linearly recalibrated AHA/ACC prediction used as in part I.
2. The following list of interventions will be gauged for causal effect:
   - Changes in smoking status (starting or stopping).
   - Losing or gaining weight
   - Changing LDL values
   - Changing Glucose values
3. For each such intervention, a propensity score will be constructed using multivariable logistic regression.
4. This score will then be used to examine the population for positivity (as defined above), with the population trimmed as needed to ensure that assumption holds.
5. A model will then be fit to estimate relative risk conditional average treatment effects. If population size allow, we will use causal forests, if not, we will use doubly robust[56] logistic regression with interactions terms.
6. These risks will then be multiplied by the absolute prediction to evaluate for treatment effects.

Power Analysis

Assuming a test set of over 300,000 patients, of which 10% will have any CVD event over 5 years, our study is sufficiently powered to detect even small differences.

For the sake of completeness, using the calculation described by Cohen ([78]), the power of this study to detect a small effect via logistic regression is >0.99, assuming an eventual 30 variables in the model, and a significance level (alpha) of 0.05.

7.3 Ethics

Parts I and III have already been given IRB approval. Part II has been submitted for approval.

8 Preliminary Results

8.1 AHA/ACC 2013 model

- Population Table for the AHA/ACC 2013 model is presented in appendix C.
- ROC Curve for the AHA/ACC 2013 Risk Score model is presented in appendix C.
- The population flow chart for the predictor is presented in appendix C.

9 References

[1] Martin J O'Donnell et al. "Global and regional effects of potentially modifiable risk factors associated with acute stroke in 32 countries (INTERSTROKE): a case-control study." In: *Lancet (London, England)* 388 (10046 Aug. 2016), pp. 761–775. ISSN: 1474-547X. DOI: 10.1016/S0140-6736(16)30506-2.

[2] Silvia Koton et al. "Stroke incidence and mortality trends in US communities, 1987 to 2011." In: *JAMA* 312 (3 July 2014), pp. 259–268. ISSN: 1538-3598. DOI: 10.1001/jama.2014.7692.

[3] Anne M Vangen-Lønne et al. "Declining Incidence of Ischemic Stroke: What Is the Impact of Changing Risk Factors? The Tromsø Study 1995 to 2012." In: *Stroke* 48 (3 Mar. 2017), pp. 544–550. ISSN: 1524-4628. DOI: 10.1161/STROKEAHA.116.014377.

[4] P W Wilson et al. "Prediction of coronary heart disease using risk factor categories." In: *Circulation* 97 (18 May 1998), pp. 1837–1847. ISSN: 0009-7322.

[5] R M Conroy et al. "Estimation of ten-year risk of fatal cardiovascular disease in Europe: the SCORE project." In: *European heart journal* 24 (11 June 2003), pp. 987–1003. ISSN: 0195-668X.

[6] Ralph B D'Agostino et al. "General cardiovascular risk profile for use in primary care: the Framingham Heart Study." In: *Circulation* 117 (6 Feb. 2008), pp. 743–753. ISSN: 1524-4539. DOI: 10.1161/CIRCULATIONAHA.107.699579.

[7] David C Goff et al. "2013 ACC/AHA guideline on the assessment of cardiovascular risk: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines." In: *Circulation* 129 (25 Suppl 2 June 2014), S49–S73. ISSN: 1524-4539. DOI: 10.1161/01.cir.0000437741.48606.98.

[8] Kirsten Bibbins-Domingo and U.S. Preventive Services Task Force. "Aspirin Use for the Primary Prevention of Cardiovascular Disease and Colorectal Cancer: U.S. Preventive Services Task Force Recommendation Statement." In: *Annals of internal medicine* 164 (12 June 2016), pp. 836–845. ISSN: 1539-3704. DOI: 10.7326/M16-0577.

[9] Ian Graham et al. "European guidelines on cardiovascular disease prevention in clinical practice: executive summary: Fourth Joint Task Force of the European Society of Cardiology and Other Societies on Cardiovascular Disease Prevention in Clinical Practice (Constituted by representatives of nine societies and by invited experts)." In: *European heart journal* 28 (19 Oct. 2007), pp. 2375–2414. ISSN: 0195-668X. DOI: 10.1093/eurheartj/ehm316.

[10] WHO. *Cardiovascular Disease fact sheet*. 2017. URL: http://www.who.int/mediacentre/factsheets/fs317/en/.

[11] Eleni Rapsomaniki et al. "Blood pressure and incidence of twelve cardiovascular diseases: lifetime risks, healthy life-years lost, and age-specific associations in 1 · 25 million people." In: *Lancet (London, England)* 383 (9932 May 2014), pp. 1899–1911. ISSN: 1474-547X. DOI: 10.1016/S0140-6736(14)60685-1.

[12] Emelia J Benjamin et al. "Heart Disease and Stroke Statistics-2017 Update: A Report From the American Heart Association." In: *Circulation* 135 (10 Mar. 2017), e146–e603. ISSN: 1524-4539. DOI: 10.1161/CIR.0000000000000485.

[13] Rafael Lozano et al. "Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: a systematic analysis for the Global Burden of Disease Study 2010." In: *Lancet (London, England)* 380 (9859 Dec. 2012), pp. 2095–2128. ISSN: 1474-547X. DOI: 10.1016/S0140-6736(12)61728-0.

[14] Israeli Center for Disease Control. *National Stroke Registry in Israel, 2014-2015*. Ed. by Inbar Zucker. 2017. URL: https://www.health.gov.il/publicationsfiles/stroke_registry_report_2014-2015.pdf.

[15] Salim Yusuf et al. "Effect of potentially modifiable risk factors associated with myocardial infarction in 52 countries (the INTERHEART study): case-control study." In: *Lancet (London, England)* 364 (9438 2004), pp. 937–952. ISSN: 1474-547X. DOI: 10.1016/S0140-6736(04)17018-9.

[16] National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel III). "Third Report of the National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel III) final report." In: *Circulation* 106 (25 Dec. 2002), pp. 3143–3421. ISSN: 1524-4539.

[17] Julia Hippisley-Cox et al. "Derivation and validation of QRISK, a new cardiovascular disease risk score for the United Kingdom: prospective open cohort study." In: *BMJ (Clinical research ed.)* 335 (7611 July 2007), p. 136. ISSN: 1756-1833. DOI: 10.1136/bmj.39261.471806.55.

[18] Julia Hippisley-Cox et al. "Predicting cardiovascular risk in England and Wales: prospective derivation and validation of QRISK2." In: *BMJ (Clinical research ed.)* 336 (7659 June 2008), pp. 1475–1482. ISSN: 1756-1833. DOI: 10.1136/bmj.39609.449676.25.

[19] J A Kanis et al. "FRAX and the assessment of fracture probability in men and women from the UK." In: *Osteoporosis international : a journal established as result of cooperation between the European Foundation for Osteoporosis and the National Osteoporosis Foundation of the USA* 19 (4 Apr. 2008), pp. 385–397. ISSN: 0937-941X. DOI: 10.1007/s00198-007-0543-5.

[20] Devan Kansagara et al. "Risk prediction models for hospital readmission: a systematic review." In: *JAMA* 306 (15 Oct. 2011), pp. 1688–1698. ISSN: 1538-3598. DOI: 10.1001/jama.2011.1515.

[21] Juliet A Usher-Smith et al. "Risk Prediction Models for Colorectal Cancer: A Systematic Review." In: *Cancer prevention research (Philadelphia, Pa.)* 9 (1 Jan. 2016), pp. 13–26. ISSN: 1940-6215. DOI: 10.1158/1940-6207.CAPR-15-0274.

[22] Gary S Collins et al. "Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD Statement." In: *European journal of clinical investigation* 45 (2 Feb. 2015), pp. 204–214. ISSN: 1365-2362. DOI: 10.1111/eci.12376.

[23] Karel G M Moons et al. "Prognosis and prognostic research: what, why, and how?" In: *BMJ (Clinical research ed.)* 338 (Feb. 2009), b375. ISSN: 1756-1833. DOI: 10.1136/bmj.b375.

[24] Richard A P Takx et al. "Pulmonary function and CT biomarkers as risk factors for cardiovascular events in male lung cancer screening participants: the NELSON study." In: *European radiology* 25 (1 Jan. 2015), pp. 65–71. ISSN: 1432-1084. DOI: 10.1007/s00330-014-3384-6.

[25] Suzette E Elias-Smale et al. "Coronary calcium score improves classification of coronary heart disease risk in the elderly: the Rotterdam study." In: *Journal of the American College of Cardiology* 56 (17 Oct. 2010), pp. 1407–1414. ISSN: 1558-3597. DOI: 10.1016/j.jacc.2010.06.029.

[26] Raimund Erbel et al. "Coronary risk stratification, discrimination, and reclassification improvement based on quantification of subclinical coronary atherosclerosis: the Heinz Nixdorf Recall study." In: *Journal of the American College of Cardiology* 56 (17 Oct. 2010), pp. 1397–1406. ISSN: 1558-3597. DOI: 10.1016/j.jacc.2010.06.030.

[27] Rozemarijn Vliegenthart et al. "Coronary calcification improves cardiovascular risk prediction in the elderly." In: *Circulation* 112 (4 July 2005), pp. 572–577. ISSN: 1524-4539. DOI: 10.1161/CIRCULATIONAHA.104.488916.

[28] Christoph R Becker. "Estimation of cardiac event risk by MDCT." In: *European radiology* 15 Suppl 2 (Feb. 2005), B17–B22. ISSN: 0938-7994.

[29] Ethan J Halpern et al. "Cost-effectiveness of coronary CT angiography in evaluation of patients without symptoms who have positive stress test results." In: *AJR. American journal of roentgenology* 194 (5 May 2010), pp. 1257–1262. ISSN: 1546-3141. DOI: 10.2214/AJR.09.3209.

[30] Ming-Ting Wu et al. "Coronary arterial calcification on low-dose ungated MDCT for lung cancer screening: concordance study with dedicated cardiac CT." In: *AJR. American journal of roentgenology* 190 (4 Apr. 2008), pp. 923–928. ISSN: 1546-3141. DOI: 10.2214/AJR.07.2974.

[31] Matthew J Budoff et al. "Coronary artery and thoracic calcium on noncontrast thoracic CT scans: comparison of ungated and gated examinations in patients from the COPD Gene cohort." In: *Journal of cardiovascular computed tomography* 5 (2 2011), pp. 113–118. ISSN: 1876-861X. DOI: 10.1016/j.jcct.2010.11.002.

[32] Tamar S Polonsky et al. "Coronary artery calcium score and risk classification for coronary heart disease prediction." In: *JAMA* 303 (16 Apr. 2010), pp. 1610–1616. ISSN: 1538-3598. DOI: 10.1001/jama.2010.461.

[33] Philip Greenland et al. "Coronary artery calcium score combined with Framingham score for risk prediction in asymptomatic individuals." In: *JAMA* 291 (2 Jan. 2004), pp. 210–215. ISSN: 1538-3598. DOI: 10.1001/jama.291.2.210.

[34] US Preventive Services Task Force et al. "Risk Assessment for Cardiovascular Disease With Nontraditional Risk Factors: US Preventive Services Task Force Recommendation Statement." In: *JAMA* 320 (3 July 2018), pp. 272–280. ISSN: 1538-3598. DOI: 10.1001/jama.2018.8359.

[35] Robert Lindsey et al. "Deep neural network improves fracture detection by clinicians." In: *Proceedings of the National Academy of Sciences of the United States of America* 115 (45 Nov. 2018), pp. 11591–11596. ISSN: 1091-6490. DOI: 10.1073/pnas.1806905115.

[36] Babak Ehteshami Bejnordi et al. "Using deep convolutional neural networks to identify and classify tumor-associated stroma in diagnostic breast biopsies." In: *Modern pathology : an official journal of the United States and Canadian Academy of Pathology, Inc* 31 (10 Oct. 2018), pp. 1502–1512. ISSN: 1530-0285. DOI: 10.1038/s41379-018-0073-z.

[37] Rahib H Abiyev and Mohammad Khaleel Sallam Ma'aitah. "Deep Convolutional Neural Networks for Chest Diseases Detection." In: *Journal of healthcare engineering* 2018 (2018), p. 4168538. ISSN: 2040-2295. DOI: 10.1155/2018/4168538.

[38] Ivana Isgum et al. "Automatic coronary calcium scoring in low-dose chest computed tomography." In: *IEEE transactions on medical imaging* 31 (12 Dec. 2012), pp. 2322–2334. ISSN: 1558-254X. DOI: 10.1109/TMI.2012.2216889.

[39] R. Shadmi et al. "Fully-convolutional deep-learning based system for coronary calcium score prediction from non-contrast chest CT". In: *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)* (2018), pp. 24–28. ISSN: 1945-8452. DOI: 10.1109/ISBI.2018.8363515.

[40] Richard A P Takx et al. "Quantification of coronary artery calcium in nongated CT to predict cardiovascular events in male lung cancer screening participants: results of the NELSON study." In: *Journal of cardiovascular computed tomography* 9 (1 2015), pp. 50–57. ISSN: 1876-861X. DOI: 10.1016/j.jcct.2014.11.006.

[41] Stephen F. Weng et al. "Can machine-learning improve cardiovascular risk prediction using routine clinical data?" In: *PLOS ONE* 12.4 (2017). Ed. by Bin Liu, e0174944. DOI: 10.1371/journal.pone.0174944.

[42] J. A. Nelder and R. W. M. Wedderburn. "Generalized Linear Models". In: *Journal of the Royal Statistical Society. Series A (General)* 135.3 (1972), p. 370. DOI: 10.2307/2344614.

[43] David Cox. "Regression Models and Life-Tables". In: *Journal of the royal statistical society* (1972).

[44] Ziad Obermeyer and Ezekiel J Emanuel. "Predicting the Future - Big Data, Machine Learning, and Clinical Medicine." In: *The New England journal of medicine* 375 (13 Sept. 2016), pp. 1216–1219. ISSN: 1533-4406. DOI: 10.1056/NEJMp1606181.

[45] Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014. ISBN: 978-1107057135. URL: https://www.amazon.com/Understanding-Machine-Learning-Theory-Algorithms/dp/1107057132?SubscriptionId=0JYN1NVW651KCA56C102&tag=techkie-20&linkCode=xm2&camp=2025&creative=165953&creativeASIN=1107057132.

[46] Rahul C Deo. "Machine Learning in Medicine." In: *Circulation* 132 (20 Nov. 2015), pp. 1920–1930. ISSN: 1524-4539. DOI: 10.1161/CIRCULATIONAHA.115.001593.

[47] Benjamin A Goldstein, Ann Marie Navar, and Michael J Pencina. "Risk Prediction With Electronic Health Records: The Importance of Model Validation and Clinical Context." In: *JAMA cardiology* 1 (9 Dec. 2016), pp. 976–977. ISSN: 2380-6591. DOI: 10.1001/jamacardio.2016.3826.

[48] Christian Lovis and Ronni Gamzu. "Big Data in Israeli healthcare: hopes and challenges report of an international workshop". In: *Israel Journal of Health Policy Research* 4.1 (2015). DOI: 10.1186/s13584-015-0057-0.

[49] Benjamin A Goldstein et al. "Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review." In: *Journal of the American Medical Informatics Association : JAMIA* 24 (1 Jan. 2017), pp. 198–208. ISSN: 1527-974X. DOI: 10.1093/jamia/ocw042.

[50] P A Wolf et al. "Probability of stroke: a risk profile from the Framingham Study." In: *Stroke* 22 (3 Mar. 1991), pp. 312–318. ISSN: 0039-2499.

[51] R B D'Agostino et al. "Stroke risk profile: adjustment for antihypertensive medication. The Framingham Study." In: *Stroke* 25 (1 Jan. 1994), pp. 40–43. ISSN: 0039-2499.

[52] Julia Hippisley-Cox, Carol Coupland, and Peter Brindle. "Derivation and validation of QStroke score for predicting risk of ischaemic stroke in primary care and comparison with other risk scores: a prospective open cohort study." In: *BMJ (Clinical research ed.)* 346 (May 2013), f2573. ISSN: 1756-1833. DOI: 10.1136/bmj.f2573.

[53] Philip Greenland et al. "Coronary Calcium Score and Cardiovascular Risk." In: *Journal of the American College of Cardiology* 72 (4 July 2018), pp. 434–447. ISSN: 1558-3597. DOI: 10.1016/j.jacc.2018.05.027.

[54] Andrew Ng. *AI is the new electricity*. 2017. URL: https://www.gsb.stanford.edu/insights/andrew-ng-why-ai-new-electricity.

[55] Alvin Rajkomar et al. "Scalable and accurate deep learning for electronic health records". In: *arxiv* (Jan. 24, 2018). arXiv: 1801.07860v2 [cs.CY].

[56] Donald B. Rubin Guido W. Imbens. *Causal Inference for Statistics, Social, and Biomedical Sciences*. Cambridge University Pr., May 31, 2015. 644 pp. ISBN: 0521885884. URL: https://www.ebook.de/de/product/22646309/guido_w_imbens_donald_b_rubin_causal_inference_for_statistics_social_and_biomedical_sciences.html.

[57] Christopher Winship Stephen L. Morgan. *Counterfactuals and Causal Inference*. Cambridge University Pr., Oct. 11, 2015. 328 pp. ISBN: 1107694167. URL: https://www.ebook.de/de/product/22837200/stephen_l_morgan_christopher_winship_counterfactuals_and_causal_inference.html.

[58] Dana Mackenzie Judea Pearl. *The Book of Why*. Hachette Book Group USA, May 15, 2018. 432 pp. ISBN: 046509760X. URL: https://www.ebook.de/de/product/30501615/judea_pearl_dana_mackenzie_the_book_of_why.html.

[59] NHS. *Check Your Heart Age*. URL: https://www.nhs.uk/conditions/nhs-health-check/check-your-heart-age-tool/.

[60] George Hripcsak et al. "Bias associated with mining electronic health records." In: *Journal of biomedical discovery and collaboration* 6 (June 2011), pp. 48–52. ISSN: 1747-5333. DOI: 10.5210/disco.v6i0.3581.

[61] Peter B Jensen, Lars J Jensen, and Søren Brunak. "Mining electronic health records: towards better research applications and clinical care." In: *Nature reviews. Genetics* 13 (6 May 2012), pp. 395–405. ISSN: 1471-0064. DOI: 10.1038/nrg3208.

[62] Orna Reges et al. "Association of Bariatric Surgery Using Laparoscopic Banding, Roux-en-Y Gastric Bypass, or Laparoscopic Sleeve Gastrectomy vs Usual Care Obesity Management With All-Cause Mortality." In: *JAMA* 319 (3 Jan. 2018), pp. 279–290. ISSN: 1538-3598. DOI: 10.1001/jama.2017.20513.

[63] Noa Dagan et al. "External validation and comparison of three prediction tools for risk of osteoporotic fractures using data from population based electronic health records: retrospective cohort study." In: *BMJ (Clinical research ed.)* 356 (Jan. 2017), p. i6755. ISSN: 1756-1833. DOI: 10.1136/bmj.i6755.

[64] G Rennert and Y Peterburg. "Prevalence of selected chronic diseases in Israel." In: *The Israel Medical Association journal : IMAJ* 3 (6 June 2001), pp. 404–408. ISSN: 1565-1088.

[65] Wai Yin Lam and Paula Fresco. "Medication Adherence Measures: An Overview." In: *BioMed research international* 2015 (2015), p. 217047. ISSN: 2314-6141. DOI: 10.1155/2015/217047.

[66] Robert Tibshirani. "Regression shrinkage and selection via the lasso: a retrospective". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 73.3 (2011), pp. 273–282. DOI: 10.1111/j.1467-9868.2011.00771.x.

[67] Stefan Wager and Susan Athey. "Estimation and Inference of Heterogeneous Treatment Effects using Random Forests". In: (Oct. 14, 2015). arXiv: http://arxiv.org/abs/1510.04342v4 [stat.ME].

[68] Leo Breiman. "Random Forests". In: *Machine Learning* 45.1 (2001), pp. 5–32. ISSN: 1573-0565. DOI: 10.1023/A:1010933404324. URL: https://doi.org/10.1023/A:1010933404324.

[69] H C van Houwelingen. "Validation, calibration, revision and combination of prognostic survival models." In: *Statistics in medicine* 19 (24 Dec. 2000), pp. 3401–3415. ISSN: 0277-6715.

[70] Donald B. Rubin, ed. *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons, Inc., 1987. DOI: 10.1002/9780470316696.

[71] Stef van Buuren and Karin Groothuis-Oudshoorn. "mice: Multivariate Imputation by Chained Equations inR". In: *Journal of Statistical Software* 45.3 (2011). DOI: 10.18637/jss.v045.i03.

[72] Robert Tibshirani Bradley Efron. *An Introduction to the Bootstrap*. Taylor & Francis Ltd, May 15, 1994. 456 pp. ISBN: 0412042312. URL: https://www.ebook.de/de/product/3596896/bradley_efron_robert_tibshirani_an_introduction_to_the_bootstrap.html.

[73] Michael Schomaker and Christian Heumann. "Bootstrap Inference when Using Multiple Imputation". In: (Feb. 25, 2016). DOI: 10.1002/sim.7654. arXiv: http://arxiv.org/abs/1602.07933v6 [stat.ME].

[74] Michael J Pencina et al. "Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond." In: *Statistics in medicine* 27 (2 Jan. 2008), 157–72; discussion 207–12. ISSN: 0277-6715. DOI: 10.1002/sim.2929.

[75] Andrew J Vickers, Ben Van Calster, and Ewout W Steyerberg. "Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests." In: *BMJ (Clinical research ed.)* 352 (Jan. 2016), p. i6. ISSN: 1756-1833. DOI: 10.1136/bmj.i6.

[76] Ewout W. Steyerberg. *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating (Statistics for Biology and Health)*. Springer, 2008. ISBN: 978-0387772431.

[77] Frank E. Harrell. *Regression Modeling Strategies*. Springer International Publishing, 2015. DOI: 10.1007/978-3-319-19425-7.

[78] Jacob Cohen. *Statistical Power Analysis for the Behavioral Sciences (2nd Edition)*. Routledge, 1988. ISBN: 978-0805802832.

Appendices

A Model Variable Lists

In the following section, % and _ are wildcards meaning, respectively, any string and any character.

American Heart Association 2013 pooled risk model

1. Sex
2. Age
3. Total Cholesterol
4. High Density Lipoprotein (HDL)
5. Treated Systolic Blood Pressure
6. Untreated Systolic Blood Pressure
7. Smoking Status
8. Diabetes

B Extraction Protocol

B.1 Outcome Diagnoses

1. **Name** Intra-Cranial Hemorrhage
   ICD9 Codes 431%
   ICPC Codes NA
   CHR Codes NA
   Sources Admissions
   Comments Primary diagnosis only, not from a rehabilitation ward
2. **Name** Ischemic CVA
   ICD9 Codes 433, 433._, 433._1, 434%, 362.3[1-3], 362.4%
   ICPC Codes NA
   CHR Codes NA
   Sources Admissions
   Comments Primary diagnosis only, not from a rehabilitation ward
3. **Name** CVA NOS
   ICD9 Codes 436%
   ICPC Codes NA
   CHR Codes NA
   Sources Admissions
   Comments Primary diagnosis only, not from a rehabilitation ward
4. **Name** Transient Ischemic Event
   ICD9 Codes 435%
   ICPC Codes NA
   CHR Codes NA
   Sources admissions, community, permanent, hospitals
   Comments Primary diagnoses only, not from a rehabilitation ward, only community neurologist
5. **Name** Subarachnoid Hemorrhage
   ICD9 Codes 430%
   ICPC Codes NA
   CHR Codes NA
   Sources Admissions
   Comments Primary diagnosis only, not from a rehabilitation ward
6. **Name** Myocardial Infarction
   ICD9 Codes 410%
   ICPC Codes NA

CHR Codes NA

Sources Admissions

Comments Primary diagnosis only, not from a rehabilitation ward

7. **Name** Non-MI Coronary Heart Disease

   ICD9 Codes 41[01234]%

   ICPC Codes K75, K76

   CHR Codes 110.1, 110.9

   Sources admissions, permanent, diagnoses and hospitals

   Comments NA

8. **Name** Congestive Heart Failure

   ICD9 Codes 428%

   ICPC Codes NA

   CHR Codes 112%

   Sources community, admissions, permanent

   Comments NA

9. **Name** Peripheral Vascular Disease

   ICD9 Codes 443%, 440.[23489]%, 250.7%, 444.2%

   ICPC Codes K92

   CHR Codes 126%

   Sources community, permanent, chronic registry, hospitals

   Comments Exclude ophtalmologist diagnoses

B.2 Causes of Death

1. **Name** Coronary Death

   ICD10 Codes (I11% OR I13% OR I21% OR I24% OR I25% OR I20% OR I44% OR I47% OR I50% OR I51%) AND (NOT I456%) AND (NOT I514%)

B.3 Background Diagnoses

1. **Name** Stroke (all kinds)

   ICD9 Codes 43[0-8]%

   ICPC Codes K90

   CHR Codes 95.2, 124

   Sources community, admissions, permanent, chronic registry

   Comments NA

2. **Name** Left Ventricular Hypertrophy

   ICD9 Codes 429.3%

ICPC Codes NA

CHR Codes NA

Sources community, admissions, permanent

Comments Primary diagnosis NA

3. **Name** Congestive Heart Failure

   ICD9 Codes 428%

   ICPC Codes NA

   CHR Codes 112%

   Sources community, admissions, permanent

   Comments NA

4. **Name** Coronary Heart Disease

   ICD9 Codes 41[012-34]%

   ICPC Codes K75, K76

   CHR Codes 110.1, 110.9

   Sources community, permanent, chronic registry, hospitals

   Comments NA

5. **Name** Peripheral Vascular Disease

   ICD9 Codes 443%, 440.[23489]%, 250.7%, 444.2%

   ICPC Codes K92

   CHR Codes 126%

   Sources community, permanent, chronic registry, hospitals

   Comments Exclude ophtalmologist diagnoses

6. **Name** Hypertension

   ICD9 Codes 40[12345]

   ICPC Codes K85, K86, K87

   CHR Codes 120%

   Sources community, permanent, chronic registry, hospitals

   Comments NA

7. **Name** Rheumatoid Arthritis

   ICD9 Codes 714.0%, 714.2%

   ICPC Codes L88%

   CHR Codes 231%

   Sources community, permanent, chronic registry, hospitals

   Comments NA

8. **Name** Chronic Kidney Disease

   ICD9 Codes 585%

ICPC Codes NA

CHR Codes 177%

Sources community, permanent, chronic registry, hospitals

Comments NA

9. **Name** Valvular Heart Disease

   ICD9 Codes 424.0%, 424.1%, 424.2%, 424.3%, 394%, 395%, 396%, 397%, 093.2%, 746.0%, 746.1%, 746.2%, 746.3%, 746.4%, 746.5%, 746.6%

   ICPC Codes K83%

   CHR Codes 111%

   Sources community, permanent, chronic registry, hospitals

   Comments NA

10. **Name** Diabetes Mellitus

    ICD9 Codes Use internal CRI registry

    ICPC Codes Use internal CRI registry

    CHR Codes Use internal CRI registry

    Sources NA

    Free-Text Inclusion NA

    Free-Text Exclusion NA

    Comments NA

11. **Name** Atrial Fibrillation

    ICD9 Codes Use internal CRI registry

    ICPC Codes Use internal CRI registry

    CHR Codes Use internal CRI registry

    Sources NA

    Free-Text Inclusion NA

    Free-Text Exclusion NA

    Comments NA

B.4 Drugs

1. **Name** Hypertension

   ATC Codes C09, C07AB03, C07FB03, C07CB03, C07CB53, C07BB03, C07DB01, C07DB01, C07AB02, C07FX03, C07FB13, C07FB02, C07FX05, C07CB02, C07BB02, C07BB52, C08C, C08G, C03A, C02AC01

2. **Name** Diabetes Mellitus

   ATC Codes A10

3. **Name** Anti-coagulants

   ATC Codes B01AA03, B01AA07, B01AA02, B01AE07, B01AF01, B01AF02

C Preliminary Result Tables, Graphs and Drawings

C.1 AHA/ACC 2013 Population Table

Population Table for the AHA/ACC 2013 model[7]:

Variables	Categories	0	1	pval
Individuals	n	1758405	38356	
Age	Mean (SD)	51.8 (15.0)	66.6 (12.8)	<0.01
Age	Median (IQR)	50.0 (39.0-62.0)	68.0 (57.0-77.0)	
SES	Mean (SD)	9.9 (4.1)	9.6 (3.9)	<0.01
SES	Median (IQR)	10.0 (6.0-13.0)	10.0 (6.0-12.0)	
BMI	Mean (SD)	27.7 (5.4)	28.8 (5.4)	<0.01
BMI	Median (IQR)	27.0 (24.0-30.7)	28.1 (25.1-31.7)	
SBP	Mean (SD)	124.9 (17.1)	135.9 (19.2)	<0.01
SBP	Median (IQR)	120.0 (113.0-134.0)	132.0 (120.0-146.0)	
DBP	Mean (SD)	76.3 (9.4)	78.3 (10.1)	<0.01
DBP	Median (IQR)	78.0 (70.0-80.0)	80.0 (70.0-83.0)	
GFR	Mean (SD)	92.3 (20.3)	78.2 (20.7)	<0.01
GFR	Median (IQR)	94.3 (79.6-107.2)	80.3 (64.3-93.1)	
Glucose	Mean (SD)	98.1 (24.9)	114.2 (35.9)	<0.01
Glucose	Median (IQR)	92.0 (84.0-103.0)	102.0 (90.0-128.0)	
LDL	Mean (SD)	117.6 (30.9)	116.6 (32.5)	<0.01
LDL	Median (IQR)	116.0 (96.0-138.0)	114.8 (93.0-138.6)	
HDL	Mean (SD)	47.9 (12.2)	46.5 (12.1)	<0.01
HDL	Median (IQR)	46.0 (39.0-55.0)	45.0 (38.0-53.0)	
Triglycerides	Mean (SD)	193.6 (38.5)	194.7 (41.0)	
Triglycerides	Median (IQR)	191.0 (167.0-217.0)	191.0 (166.0-220.0)	<0.01

C.2 AHA/ACC 2013 Risk Model Result Graph

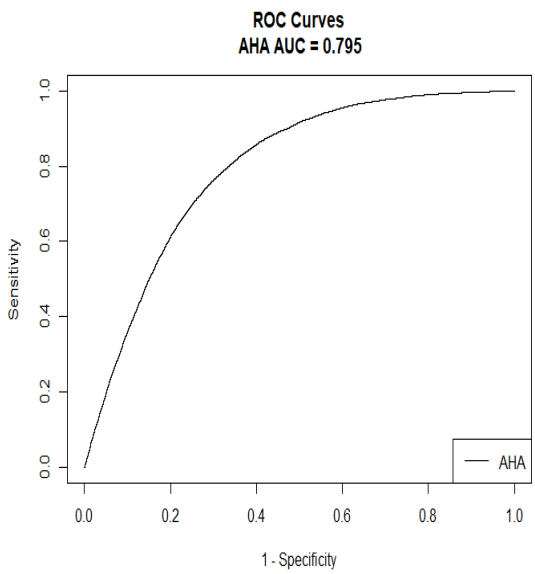

Figure 1: AHA/ACC 2013 ROC Curve, p-value < 0.001

C.3 Clalit Model Population Flow Chart

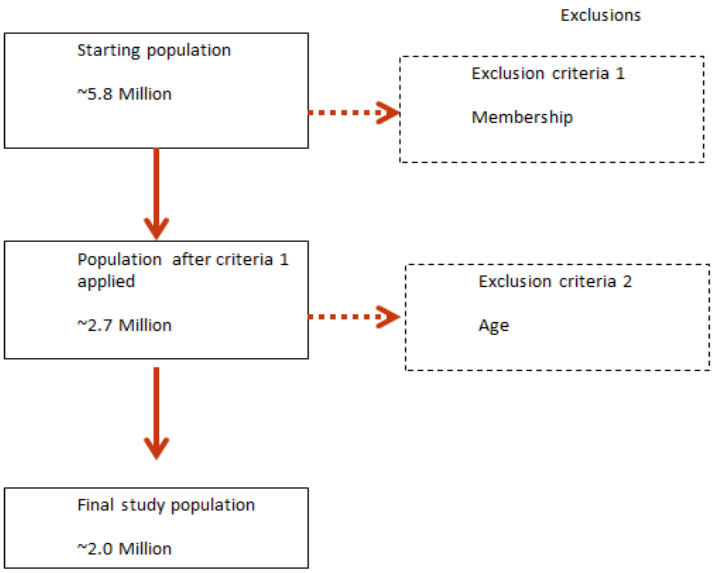

Figure 2: Population Flow Chart