Ben-Gurion University of the Negev

PhD Proposal

Generic Medical Risk Prediction for Prevention of Cardiovascular Disease

חיזוי סיכון גנרי ברפואה למניעת מחלה קרדיווסקולרית

Noam Barda, MD

December 21, 2018

Supervisor's Signature, Prof. Nadav Davidovich: ______________________

Supervisor's Signature, Prof. Eitan Bachmat: ______________________

Chairman of Departmental Graduate Studies's Signature: ______________________

Contents

1 Abstract

1.1 Background

Since the early 1990s, and more so in the last few years, multivariable risk models have been created to estimate patients' risk for different diseases over different time spans (e.g. [1, 2, 3]). These models are used to identify patients at risk and are capable of exact risk quantification over time[4]. Through their many variations, these risk models are included in different guidelines and occupy an important place in both primary prevention and diagnosis of different diseases[5, 4].

These risk models have been particularly used for the diagnosis and prevention of cardiovascular disease (CVD), which despite reduced incidence in the developed world in recent years[6, 7], remains a significant cause of morbidity and mortality[8].

Such multivariable risk models have traditionally required extensive input from domain experts (clinicians) for the choice of predictors and are based on classic biostatistical models, usually logistic and Cox regression.

1.2 Goals

In this thesis we intend to build a generic prediction framework, capable of constructing validated prediction models for any outcome with no domain expert input. We will use this framework to construct a predictive model for cardiovascular disease (CVD). We will externally validate existing CVD prediction models on the Israeli population. And we will compare the model we constructed to the current models, in essence simulating an application of said models for an EHR based population-wide prevention program.

1.3 Methods

This generic prediction framework will make use of a sparsity inducing algorithm fed a curated version of the majority of the variables in the Clalit's electronic health record. The algorithm will then choose the appropriate variables and construct a model.

To perform external validation and comparison, the three leading CVD models will be selected and recreated on the Clalit's database. The models will be compared on a common sample corresponding to the population on which we intend to simulate our intervention.

The simulated CVD intervention will examine a retrospective cohort of patients, and compare the decisions that would have been made for the patients using the existing models vs. the new model.

1.4 Importance

A generic prediction framework will allow easy construction of validated predictive models of high quality, with the variable selection portion affording possible biological insight.

External validation of CVD and mortality risk models, currently in wide use and integrated into guidelines, is of vital importance[9].

Simulating the intervention will allow us to examine the relative improvement in decision-making afforded by the new model. It will also allow us to gauge the potential effectiveness of interventions based on such risk models. Such interventions are becoming more and more possible with widespread electronic health record availability.

1.5 Keywords

Medical Prediction Models, Generic Prediction, Cardiovascular Disease, Electronic Health Records

2 תקציר בעברית

2.1 רקע

מתחילת שנות ה-90 החלה, ובשנים האחרונות גברה, יצירתם של מודלים רב-משתניים לחישוב סיכון למחלות שונות(לדוגמה [1, 2, 3]). מודלים אלו משמשים לזיהוי חולים בסיכון, ומאפשרים כימות מדויק של הסיכון לאורך שנים רבות[4]. כיום, בצורותיהן השונות, מודלים אלו כלולים בקווים המנחים של ארגונים מקצועיים רבים, ולהם מקום חשוב הן במניעה הראשונית והן באבחנה של מחלות [5, 4].

במודלים אלו נעשה שימוש בייחוד לחיזוי ולמניעה של מחלה קרדיו-וסקולרית, אשר חרף הירידה בהיארעותה בעולם המפותח בשנים האחרונות[6, 7], נותרה סיבה חשובה לתחלואה ולתמותה[8].

ככלל, מודלים אלו דרשו התייעצות נרחבת עם מומחי תוכן (קלינאים) על מנתשאלו יספקו את המידע הנדרש בנוגע למשתנים הנדרשים לחיזוי התוצאה, והתבססו על שיטות ביוסטטיסטיות מסורתיות, ובפרט על רגרסיה לוגיסטית ורגרסיית קוקס.

2.2 מטרות

בתזה זו אנו מתכוונים לבנות שלד גנרי לחיזוי סיכון, המסוגל לבנות מודלים מתוקפים לכל תוצא ללא כל מידע ממומחי תוכן. אנו נשתמש במנגנון זה לבנות מודל לחיזוי מחלה קרדיוסקולרית. אנו נבצע תיקוף חיצוני למודלים המובילים בספרות לחיזוי מחלה קרדיווסקולרית על האוכלוסיה הישראלת. ולבסוף, אנו נשווה את ביצועי המודל החדש למודלים הקיימים באופן שלמעשה יבצע סימולציה של תכנית התערבות מבוססת תיק רפואי ממוחשב על אוכלוסיית הקופה.

2.3 שיטות

השלד לחיזוי גנרי יעשה שימוש באלגוריתם המבצע בחירת משתנים כחלק מפעולתו. אלגוריתם זה יוזן עם מרבית המשתנים הזמינים בבסיס הנתונים של קופ"ח כללית, מהם יבחר המשתנים הרלוונטיים ובהם יעשה שימוש לבניית מודל.

על מנת לבצע תיקוף חיצוני והשוואה, שלושת המודלים המובילים לחיזוי CVD ייבחרו מהספרות וייבנו מחדש על גבי בסיס הנתונים של הכללית. המודלים יושוו בהרכב האוכלוסיה בו אנו עתידים לדמות התערבות.

סימולציית ההתערבות תבחן עקבה רטרוספקטיבית, ותשווה ההחלטות שהיו מתקבלות תוך שימוש במודלים הקיימים לעומת המודל החדש.

2.4 חשיבות

מסגרת לחיזוי גנרי תאפשר בניית מודלים מתוקפים ובאיכות גבוהה לחיזוי מחלות שונות, ועצם תהליך בחירת המשתנים יכול שיאפשר תובנות ביולוגיות.

תיקוף חיצוני של מודלים לחיזוי מחלה קרדיווסקולרית, המצויים בשימוש רחב ומשולבים בקווים המנחים, הוא בעל חשיבות מכרעת[10].

סימולציית ההתערבות תאפשר לנו לבחון את השיפור בקבלת ההחלטות שהיה מתקבל על-ידי המודל החדש, לו היה מיושם באופן גורף על אוכלוסיית הקופה כולה. הודות לזמינות הגוברת של תיקים רפואיים ממוחשבים, התערבויות שכאלה הופכות אפשריות יותר ויותר.

2.5 מילות מפתח

מודלי חיזוי רפואיים, חיזוי גנרי, מחלה קרדיווסקולרית, תיק רפואי ממוחשב.

3 Aim of the Thesis

The main aim of this thesis is to design, implement and evaluate an algorithm to develop multivariable predictive models based on Clalit Health Services' (CHS) electronic health record based database; to use it to generate a risk model for cardiovascular disease; to compare this model to existing risk models, externally validating them in the process; and to simulate an intervention on the Israeli population based on these risk models.

The aforementioned goal will require three steps:

Model Development 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.

Model Evaluation The above-mentioned approach will be used to construct a risk model for 10-year prediction of Cardiovascular Disease (CVD). The merits of this model will be tested by comparing it to the leading existing models from the literature. This evaluation will comprise a comprehensive test of these models' performance in both their original population composition and in a shared population with the characteristics we intend to use in our intervention.

Simulated Intervention The CVD models will be used to simulate an intervention based on a historic cohort of patients from the CHS' database. The true outcomes, the predicted outcomes based on existing models and the predicted outcomes based on our models will be compared.

Based on these aims, we hypothesize:

1. 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.
2. That a model developed in such fashion will outperform traditional risk models.

3. That existing CVD risk models will perform poorly on the ethnically distinct Israeli population.
4. That the advantages of such a model will have the potential to improve patient outcomes if used in a population-wide intervention based on EHR data.

4 Importance and Background

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

4.1 Part I

4.1.1 Methodology of Traditional Risk Models

For traditional medical risk models, two design decisions are ubiquitous[11]:

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.

4.1.2 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[12].

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.

4.1.3 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[13].

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 proportional hazards 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.

4.1.4 Parametric Vs. Non-Parametric Models

Parametric models, such as those described above, summarize the data with a set of parameters of fixed size that is independent of the number of training examples. This has the advantage of simplicity, interpretability and speed, but also leads to biases in prediction if the ”true” population model is different than the chosen model.

Non-parametric models make no such assumptions about the structure of the target function they seek to learn. This requires far more data for accurate training, and does not allow interpretation of coefficients using terms such as odds ratios, but does afford more predictive accuracy when sufficient data exists[14].

4.1.5 The Rise of AI and Machine Learning

In recent years the fields of machine and statistical learning have seen a tremendous rise[15]. this growth in machine learning, including predictive modeling, has occurred thanks to three main factors[16]:

- 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[15, 17].

4.1.6 Black-Box Vs. White-Box Models

While there are obstacles from many different domains to the integration of machine learning approaches in medicine: psychological, legal, regulatory and others, one overarching concern is the preeminence of black-box models in machine learning[18].

Broadly defined, black-box models are models whose results cannot be readily explained. For example, a logistic regression result can be fairly easily reasoned about: baseline risk was x%, and a certain combination of variables increased the risk by y% more. The same cannot be said for most models used in modern machine learning, including neural networks and tree-ensemble models. These models generate a result that is a complex non-linear function of their inputs, and one cannot easily explain why a specific patient got a risk of x%, while another got y%.

Beyond the legal and psychological difficulty this creates (how does one explain, to oneself and others, a decision based on unclear reasoning?), it also introduces the possibility of discrimination. The algorithm could choose to optimize for one (majority) population, while neglecting other (minority) populations[19]. This fascinating area of research falls under the more general notion of algorithmic fairness, more widely studied in other non-medical fields[20], and is beyond the scope of this thesis.

4.1.7 Electronic Health Record based Observational Studies

Most medical risk models in wide-use were developed based on specialized cohort studies[21]. 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[22], encompasses the full extent of a patient's encounters with the health system[23].

4.1.8 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. As this model will make use of large portions of the EHR data and will be purposely built on the Clalit's population, it is likely to perform well.

4.2 Part II

4.2.1 Epidemiology of Cardiovascular Disease and Stroke

In its usual definition, cardiovascular disease (CVD) includes several disease categories[24]:

1. Coronary Heart Disease
    (a) Myocardial Infarction
    (b) Angina Pectoris
    (c) Heart Failure
    (d) Coronary death
2. Cerebrovascular Disease
    (a) Stroke (Thrombotic and Hemorrhagic)
    (b) Transient Ischemic Attack
3. Peripheral Artery Disease
4. Aortic Disease
    (a) Atherosclerosis
    (b) Aneurysm
5. Rheumatic Heart Disease
6. Congenital Heart Disease
7. Venous Thromboembolism
    (a) Pulmonary Embolism
    (b) Deep Vein Thrombosis

CVD is very common. Lifetime risk for people aged 30 with no prior cardiovascular disease approaches 50 percent[25], with coronary heart disease being the most common specific diagnosis[26].

While the rates of cardiovascular disease have declined in developed countries over the last 30 years[6, 7], they remain significant public health problems, being the second most common cause of mortality and third most common cause of disability worldwide[27]. The statistics in Israel are similar[28].

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[29, 8]: Smoking, dyslipidemia, hypertension, diabetes, etc.

4.2.2 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[1, 30, 2, 31, 3, 32, 4]. Still the most notable of said risk models is the Framingham risk model family, developed on a US population in Massachusetts, Boston[1], and the SCORE risk model, developed in 2003 on a European population[2].

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[4] and their use in deciding on anti-platelet therapy[33], 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[34, 35], and have found use not only in prediction, but also in diagnosis[36]. This increasingly important place taken by risk models has brought about the publication of guidelines designed to regulate and improve their creation[10]. As estimating the probability for existing and future disease is a significant portion of the clinical process[37], 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.

4.2.3 Limitations of Risk Models

Naturally, risk models are developed on a specific population, whose data is available to the researchers developing the model. As patients differ in a variety of ways (both genetic and environmental), and even such basic things as lab methods and disease definitions differ in different areas, models tend to function better when used on the population on which they were developed[38, 39].

Recent models have tried to deal with this problem by including more ethnically varied populations[40] or recalibrating the model for each new population[34], but such efforts are limited to specific risk models, and even then have only been partially successful[41]. As one specific Israeli example, this phenomenon was observed in a recent publication that illustrated significant mis-calibration for osteoporosis prediction models that are in wide clinical use and incorporated into guidelines[41]. As the probabilities generated by the model eventually help determine the proper interventions to perform, according to respective guidelines, such mis-calibration could invalidate the use of the model, making external validation an important endeavor[9].

4.2.4 The Scientific Gap

Though the risk scores are currently used in common medical practice, external validation of international CVD risk models for the Israeli population has yet to be performed, and recommendations on which model to use are based on expert opinion[42].

We suggest, as a first effort, to externally validate widely used risk models for the prediction of CVD risk on the Israeli population. This could help decide which model has the best performance, and if all such models' performance is deemed unsatisfactory, this will have significant consequences for guidelines and practices based on said models. Immediately thereafter, these models will be compared to the internally developed model in part I.

4.3 Part III

4.3.1 Traditional Aim of Risk Models

Outside of the realm of medicine, risk models are used for great many purposes: deciding which customers are likely to default on loans, deciding which credit card deals are fraudulent, deciding which customers are likely to churn, etc.

Within the realm of medicine, the use of risk models is fairly consistent. When deciding on some intervention to lower some risk (e.g. statins for CVD), one has to always remember that interventions have risks themselves (e.g. rhabdomyolysis from statins). For any utility one mentally assigns lower CVD risk and higher rhabdomyolysis risk (in our example), the prescription of statins is more warranted if the baseline risk for CVD is higher. This is intuitive and simple - one does not walk around wearing a Hazmat suit if one is not in the immediate vicinity of hazardous materials (presumably because its hot within such suits).

With this logic in mind, risk models are constantly used, consciously and subconsciously, when deciding on diagnostic and therapeutic interventions. Consciously, for example, when deciding on aspirin and statins for CVD risk[4, 33], bisphosphonates for osteoporosis risk [43] or CT angiogram for pulmonary thromboembolism risk[44]. Subconsciously, for example, when deciding whether to refer a patient suspected of pneumonia to a chest x-ray.

4.3.2 The Way Risk Models are Used

For several reasons, utilizing risk models for these aims requires the direct involvement of a treating physician:

1. The different risk models require knowledge of a wide variety of clinical factors, including lab results that most patients are not expected to know themselves.
2. The decisions to be made can only be made by a physician. A patient cannot prescribe statins to himself.

And so the use of such model has mostly been limited to physicians. To make use of these risk models, the physician, usually the primary care physician, is required to the fill in the different covariates based on the patient's health record, communicate the results to the patient, and advise on whatever intervention is mandated to mitigate the risk. This entire time consuming act is expected to occur in an already time-strained primary care encounter[45].

4.3.3 The Scientific Gap

We suggest that the structure of the Israeli health care system is ideal for performing and evaluating an intervention based on EHR based prediction of cardiovascular risk.

Instead of the usual methodology, by which patients are identified as high-risk when they enter the physician's office, usually for other concerns and in severe time constraints, we will predict the risk at once for all patients in the database.

We will compare the decisions that would have been made had these different models been used on the retrospective cohort, in view of the actual outcomes of the different patients. This will allow us to judge whether the novel model would have indeed improved patient outcomes.

5 The Novelty of the Thesis

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

- 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.
- External validation of existing risk models is of utmost importance[9], as these models are used constantly as part of existing guidelines (e.g. the American Heart Association’s pooled risk model and Statin treatment[4], FRAX and Osteoporosis treatment[34]). This is especially true, as previous external validation studies have at times documented significant mis-calibration[39, 41], that would make treatment decisions based on the models problematic.
- Population wide EHR-data based interventions using predictive risk models have yet to be implemented, to the best of our information. Illustrating the advantages of such interventions, specifically with the rising availability of EHR data, is of significant importance.

6 Published Work

The epidemiological characteristics of CVD in general and of stroke in particular are well understood[6, 7], and the dominant risk factors in the population well mapped[29, 8]. This is true both in the developed and in the developing world[27]. It is also true in Israel[28].

The increasingly central role filled out by risk prediction models in medicine has been observed[37], as have the challenges of developing such models based on Electronic Health Record (EHR) data[21, 23]. 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[10].

Many CVD risk models have been developed in the last 30 years, most prominent of which are the Framingham[1, 30, 3, 4], SCORE[2] and Qrisk[31, 32] families of models. Two of these model families also offer a stroke-specific model[46, 47, 48].

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[30, 5, 4, 33].

Models’ tendency to under-perform when the target population is changed is widely recognized[38, 39, 40] and accordingly, the importance of external validation of models prior to their use in new population is recommended[9]. External validation of CVD models has been performed in several populations[38, 39, 40], though not in the Israeli population[42]. This is in contrast to, for example, Osteoporosis[41].

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[49].

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[15]. 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[50].

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.

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 4.4 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 a low 1%, 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[51, 52, 23]. 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. [53, 41]).

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[28] (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[54]. 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[55].
- 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

In part II, where external validation of international models is to take part, special care will be required to handle variables that are not perfect ”fits” for the Clalit's database, for example:

- UK socioeconomic status (”Townsend Deprivation Score”), which has different levels and is directed in the opposite direction (more means lower SES) than the Clalit's socioeconomic status.
- 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)[56] 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 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 Gradient Boosting

Gradient boosting[57] is an ensemble method that combines several weak learners (usually shallow trees) together using a weighted majority vote. Each consecutive learning phase focuses on those samples in the training set that were predicted wrong by the previous phases.

Gradient boosting induces sparsity by deciding on the important features at each split in each tree. The rules for these decisions are themselves parameters to the models, but all generally employ a version of Claude Shannon's information entropy[58]:

For a given variable x, the entropy is defined as

$$H(x) = -\sum_{i=1}^{n} P(x_i) \log P(x_i)$$

This entropy is maximized when the "doubt" about the value of a variable is maximal, and the different tree models strive to minimize it by choosing maximally informative variables for each split.

Hyper-parameter tuning, per each model's hyper-parameter lists, is conducted on the validation set using random search[59]. The best performing model with regard to area under the ROC curve is selected.

7.1.7 Learning Curves

Learning curves are plots of performance as a function of sample size. They provide simple visual indication that the model has a sufficiently large sample size so as not to over-fit or under-fit.

7.1.8 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 attemps to deal with this problem.

The framework suggested by Van Houwlingen et al[60] 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.

7.1.9 Imputation

As first explained by Rubin et al.[61], 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[61]. Multiple imputation with chained equations (MICE) is currently the standard way to impute data in biostatistical research[62].

7.1.10 The Bootstrap

The bootstrap, as developed by Efron et al.[63], 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[64].

7.1.11 Net Reclassification Improvement

Net Reclassification improvement (NRI) aims to measure the improvement in decisions made by using a different prediction model[65]. 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.

7.1.12 Decision Curves

Decision curves, who have recently gained popularity in the medical literature[66], 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 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, our 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, as illustrated in the following design diagram.

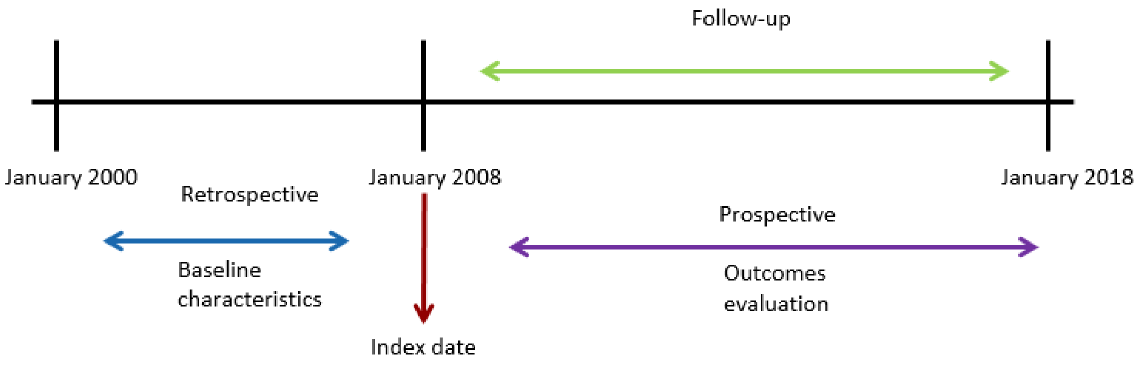

Figure 1: Study Design Timeline

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. Two models will be fit: LASSO and gradient boosting, as detailed above.
4. Cross-validation will be performed on the training set for hyper-parameter tuning.
5. Feature importance statistics will be described for each model. for LASSO - coefficients, for gradient boosting - total information gain.
6. 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.
7. Model performance on the test set will be reported for the following statistics [67, 68]
   - 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% and 10% risk threshold. These thresholds are chosen for their importance in existing guidelines[4, 33].

7.2.2 Part II

In this part, we will externally validate three models that are the most significant and widely used existing models designed to assess risk for cardiovascular disease (CVD).

The Models

The models to be compared are:

1. American Heart Association 2013 pooled cohort equations[4]
2. SCORE[2]
3. Qrisk[32]

Study Design

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

Study Population

The population is similar to the population in part I.

Study Timeline

The timeline is similar to the timeline in part I.

Variables

Each model uses its own variables, see appendix A for full model variable lists.

Descriptive Statistics

1. A population table (”table 1”) will be provided.
2. For each variable, a fitting description of its distribution will be provided (e.g. frequencies for categorical variables, mean and standard deviation for continuous variables).

Modeling

1. The published coefficients for each model will be used to recreate it accurately on the study population.
2. Linear recalibration, as detailed above, will be performed for each model using the training set.
3. Performance metrics, as in part I, will be stated for each model.
4. Confidence intervals, as above, will be generated using 500 bootstrap repetitions per each multiply imputed dataset.

7.2.3 Part III

In this part, we will compare the performance of the external models with that of the internal model, specifically noting the number of patients that would have received more appropriate care, had the recommendations per each model adhered to.

Study Design

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

Study Population

The population is similar to the population in part I.

Study Timeline

The timeline is similar to the timeline in part I.

Variables

1. The predictions from the previous parts.
2. CVD outcomes, as detailed above.

Model Performance Comparison

We will compare metrics that are related to the decisions that would have been made for each patient, in light of the outcomes we observed. The different metrics are:

1. Continuous Net Reclassification Improvement
2. Categorical Net Reclassification Improvement - Combined and separately for events and for non-events.
3. Decision Curves

All metrics are explained above. Metrics that rely on a threshold will be reported for 7.5% and 10% risk thresholds, as explained above.

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 ([69]), 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

This study requires an Institutional review board (IRB) approval. It has been submitted for such approval. Analysis will be commence after the approval is granted.

8 Preliminary Results

We will present preliminary results for the first two parts.

8.1 Part 1

Population Table for the AHA/ACC 2013 model[4]:

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

ROC Curve for the AHA/ACC 2013 Risk Score model is presented in appendix C.

8.2 Part 2

The population flow chart for the predictor is presented in appendix C.

9 References

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] Karel G M Moons et al. "Risk prediction models: II. External validation, model updating, and impact assessment." In: *Heart (British Cardiac Society)* 98 (9 May 2012), pp. 691–698. ISSN: 1468-201X. DOI: 10.1136/heartjnl-2011-301247.

[10] 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.

[11] 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.

[12] 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.

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

[14] Stuart Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach (2nd Edition)*. Prentice Hall, 2002. ISBN: 0137903952. URL: https://www.amazon.com/Artificial-Intelligence-Modern-Approach-2nd/dp/0137903952?SubscriptionId=0JYN1NVW651KCA56C102&tag=techkie-20&linkCode=xm2&camp=2025&creative=165953&creativeASIN=0137903952.

[15] 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.

[16] 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.

[17] 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.

[18] Nicholas Price. "Black-Box Medicine". In: *Harvard Journal of Law & Technology* 28.2 (2015), pp. 420–454.

[19] Úrsula Hébert-Johnson et al. "Calibration for the (Computationally-Identifiable) Masses". In: (Nov. 22, 2017). arXiv: `1711.08513v2 [cs.LG]`.

[20] Sam Corbett-Davies et al. "Algorithmic decision making and the cost of fairness". In: (Jan. 28, 2017). DOI: `10.1145/3097983.309809`. arXiv: `1701.08230v4 [cs.CY]`.

[21] 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`.

[22] 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`.

[23] 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`.

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

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

[26] 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`.

[27] 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`.

[28] 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`.

[29] 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`.

[30] 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.

[31] 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`.

[32] 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`.

[33] 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`.

[34] 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`.

[35] 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`.

[36] 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`.

[37] 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`.

[38] R B D'Agostino et al. "Validation of the Framingham coronary heart disease prediction scores: results of a multiple ethnic groups investigation." In: *JAMA* 286 (2 July 2001), pp. 180–187. ISSN: 0098-7484.

[39] Sylvie Bastuji-Garin et al. "The Framingham prediction rule is not valid in a European population of treated hypertensive patients." In: *Journal of hypertension* 20 (10 Oct. 2002), pp. 1973–1980. ISSN: 0263-6352.

[40] Andrew P DeFilippis et al. "An analysis of calibration and discrimination among multiple cardiovascular risk scores in a modern multiethnic cohort." In: *Annals of internal medicine* 162 (4 Feb. 2015), pp. 266–275. ISSN: 1539-3704. DOI: `10.7326/M14-1281`.

[41] 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`.

[42] Rafael Bitzur et al. "[ISRAELI GUIDELINES FOR THE TREATMENT OF HYPERLIPIDEMIA - 2014 UPDATE]." In: *Harefuah* 154 (5 May 2015), pp. 330–3, 337–8. ISSN: 0017-7768.

[43] Michael P Jeremiah et al. "Diagnosis and Management of Osteoporosis." In: *American family physician* 92 (4 Aug. 2015), pp. 261–268. ISSN: 1532-0650.

[44] P S Wells et al. “Excluding pulmonary embolism at the bedside without diagnostic imaging: management of patients with suspected pulmonary embolism presenting to the emergency department by using a simple clinical model and d-dimer.” In: *Annals of internal medicine* 135 (2 July 2001), pp. 98–107. ISSN: 0003-4819.

[45] Thomas R Konrad et al. “It’s about time: physicians’ perceptions of time constraints in primary care medical practice in three national healthcare systems.” In: *Medical care* 48 (2 Feb. 2010), pp. 95–100. ISSN: 1537-1948. DOI: 10.1097/MLR.0b013e3181c12e6a.

[46] 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.

[47] 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.

[48] 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.

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

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

[51] 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.

[52] 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.

[53] 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.

[54] 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.

[55] 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.

[56] 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.

[57] Yoav Freund and Robert E Schapire. “A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting”. In: *Journal of Computer and System Sciences* 55.1 (1997), pp. 119–139. DOI: 10.1006/jcss.1997.1504.

[58] C. E. Shannon. “A Mathematical Theory of Communication”. In: *Bell System Technical Journal* 27.3 (1948), pp. 379–423. DOI: 10.1002/j.1538-7305.1948.tb01338.x.

[59] James Bergstra and Yoshua Bengio. “Random Search for Hyper-parameter Optimization”. In: *J. Mach. Learn. Res.* 13 (Feb. 2012), pp. 281–305. ISSN: 1532-4435. URL: http://dl.acm.org/citation.cfm?id=2188385.2188395.

[60] 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.

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

[62] 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.

[63] 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.

[64] 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].

[65] 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.

[66] 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.

[67] Ewout W. Steyerberg. *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating (Statistics for Biology and Health).* Springer, 2008. ISBN: 978-0387772431. URL: https://www.amazon.com/Clinical-Prediction-Models-Development-Validation/dp/038777243X?SubscriptionId=0JYN1NVW651KCA56C102&tag=techkie-20&linkCode=xm2&camp=2025&creative=165953&creativeASIN=038777243X.

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

[69] Jacob Cohen. *Statistical Power Analysis for the Behavioral Sciences (2nd Edition).* Routledge, 1988. ISBN: 978-0805802832. URL: https://www.amazon.com/Statistical-Power-Analysis-Behavioral-Sciences/dp/0805802835?SubscriptionId=0JYN1NVW651KCA56C
tag=techkie-20&linkCode=xm2&camp=2025&creative=165953&creativeASIN=0805802835.

Appendices

A Model Variable Lists

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

1. American Heart Association 2013 pooled risk model
   - (a) Sex
   - (b) Age
   - (c) Total Cholesterol
   - (d) High Density Lipoprotein (HDL)
   - (e) Treated Systolic Blood Pressure
   - (f) Untreated Systolic Blood Pressure
   - (g) Smoking Status
   - (h) Diabetes
2. SCORE
   - (a) Sex
   - (b) Age
   - (c) Total Cholesterol
   - (d) Treated Systolic Blood Pressure
   - (e) Untreated Systolic Blood Pressure
   - (f) Smoking Status
3. QRisk2
   - (a) Ethnicity
   - (b) Age
   - (c) Sex
   - (d) Smoking Status
   - (e) Systolic Blood Pressure
   - (f) Total Cholesterol / High Density Lipoprotein Ratio
   - (g) Body Mass Index (BMI)
   - (h) Family History of Coronary Heart Disease (CHD)
   - (i) Townsend Deprivation Score
   - (j) Treated Hypertension
   - (k) Rheumatoid Arthritis
   - (l) Chronic Kidney Disease
   - (m) Type II Diabetes
   - (n) Atrial Fibrillation

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 Graphs and Drawings

C.1 AHA/ACC 2013 Risk Model Result Graph

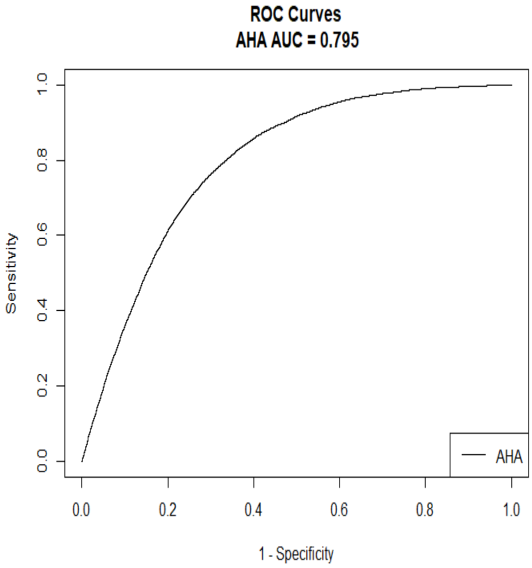

Figure 2: AHA/ACC 2013 ROC Curve, p-value < 0.001

C.2 Clalit Model Population Flow Chart

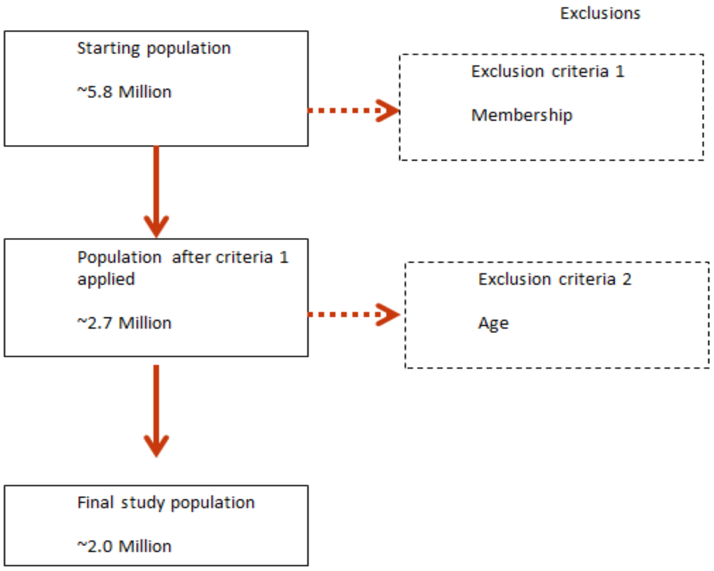

Figure 3: Population Flow Chart