PhD Proposal

Noam Barda

August 24, 2018

Contents

1 Abstract **4**

1.1 Background 4

1.2 Goals 4

1.3 Methods 5

1.4 Importance 6

2 Hebrew Abstract **6**

3 Aim of the Thesis **9**

4 Importance and Background **10**

4.1 Part I 11

4.1.1 Epidemiology of Cardiovascular Disease and Stroke . . 11

4.1.2 History of Multivariate Risk Models 12

4.1.3 Limitations of Risk Models 13

4.1.4 The Gap and our Thesis 13

4.2 Part II 14

4.2.1 Methodology of Traditional Risk Models 14
4.2.2 Generalized Linear Models 14
4.2.3 Cox Proportional Hazards 15
4.2.4 Parametric Vs. Non-Parametric Models 16
4.2.5 The Rise of AI and Machine Learning 16
4.2.6 Black-Box Vs. White-Box Models 17
4.2.7 Electronic Health Record based Observational Studies 17
4.2.8 The Gap and our Thesis 18
4.3 Part III 18
4.3.1 Traditional Aim of Risk Models 18
4.3.2 The Way Risk Models are Used 19
4.3.3 Empowering Patients to Control Risk Factors 20
4.3.4 Bridging Self-Care and Risk Modeling 20
4.3.5 The Gap and our Thesis 21

5 The Novelty of the Thesis 21

6 Published Work 22

7 Research Methodology 24
7.1 Planning 24
7.1.1 Part I 24
7.1.2 Part II 25
7.1.3 Part III 27
7.2 Data Extraction 29
7.3 Descriptive Statistics 31
7.4 Modeling and Inferential Statistics 32

7.4.1 Part I 32
7.4.2 Part II 34
7.4.3 Part III 36

8 Preliminary Results **37**
8.1 Part 1 37
8.2 Part 2 38
8.3 Part 3 38

9 References **38**

Appendices **49**

A Model Variable Lists **49**

B Extraction Protocol **51**
B.1 Outcome Diagnoses 51
B.2 Causes of Death 54
B.3 Background Diagnoses 54
B.4 Drugs 59

C Generic Predictor Variable List **60**

D Preliminary Result Graphs and Drawings **90**
D.1 Framingham Stroke Risk Score (FSRS) Result Graphs 90
D.2 Clalit Model Population Flow Chart 91
D.3 Application Mock Drafts 91

1 Abstract

1.1 Background

Despite reduced incidence in the developed world in recent years[43, 70], cardiovascular disease (CVD) remains a significant cause of morbidity and mortality[53].

Since the early 1990s, Multivariate risk models have been created to estimate patients' 10 year risk for cardiovascular events (e.g. [75, 14, 20]). These models are used to identify patients at risk and are capable of exact risk quantification over time[26]. Through their many variations, CVD risk models are included in different guidelines and occupy an important place in primary prevention of CVD[30, 26].

These models have several characteristics:

- Their performance is highest in the population used to develop them and is reduced on populations that are genetically or otherwise different[19, 3, 21].
- They are traditionally based on classic biostatistical models, usually logistic and Cox regression.
- Their use requires knowledge of test results (labs and otherwise) that patients are not generally familiar with, and the models themselves usually used to decide on treatments that require a physician to execute (e.g. statin treatment). This makes the traditional "costumers" of medical risk models the treating physicians, specifically those engaged in primary care.

1.2 Goals

We intend to pursue three goals in this thesis:

- Being ethnically distinct from the US and European populations used to develop existing models, CVD risk models are expected to perform sub-optimally in the Israeli population, but such external validation has yet to be performed on a population-wide scale[45]. We intend to perform external validation of CVD and stroke predictive models on a wide sample of the Israeli population.

- Recent advances in machine learning allow for a new approach to medical risk modeling[52]. Contrary to the classic approach that emphasizes domain knowledge for the pre-specification of risk factors, novel methods rely instead on sophisticated algorithms being presented with thousands of candidate variables and selecting the relevant ones by themselves. These variables are then used for the actual model building[73]. Such technologies allow a more standardized "one-size fits all" approach to risk modeling, utilizing a single comprehensive database with different outcomes[58]. We intend to establish a generic prediction framework allowing the standardized construction of validated predictive models for any outcome.

- Self-care is crucially important for disease prevention[54]. We intend to promote patient self-care by targeting a predictive model directly to patients through the Clalit's on-line website, in a way that encourages risk factor mitigation.

1.3 Methods

To perform external validation, four leading CVD models will be selected and recreated on the Clalit's database. The models will be compared in their original population composition and on a common sample corresponding to the population to which we intend to target our intervention.

The generic prediction framework will make use of a sparsity inducing algorithm fed with the majority of the variables in the Clalit's electronic health record. The algorithm will then choose the appropriate variables and con-

struct a model. The model will first be tuned on a validation set and then evaluated on a test set.

The direct-to-patient intervention will utilize the Clalit's website to present a risk calculator for stroke directly to patients. The calculator's variables will be fed directly from the patient's electronic health record. The patient will be encouraged to view the effect of changing specific risk factors and will be presented with personalized written recommendations for CVD risk mitigation. We will formally test the effect of such an intervention in a hypothesis testing framework. This will be done by matching patients who've used the calculator with age, sex and risk matched controls, and performing multivariate logistic regression to judge the effect of the intervention.

1.4 Importance

External validation of CVD risk models, currently in wide use and integrated into guidelines, is of vital importance[48].

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

Performing and evaluating the intervention will allow us to gauge the effectiveness of such self-care promoting interventions. These interventions are of particular importance in CVD[62] and are becoming more and more possible with widespread electronic health record availability.

2 Hebrew Abstract

למרות ירידה בהיארעותן בעשרות השנים האחרונות[43, 70], מחלות לב וכלי דם (קרדיו-וסקולריות) עודן גורם חשוב לתחלואה ולתמותה בעולם המפותח[53].

מאז תחילת שנות ה-90 החלה יצירתם של מודלים רב-משתניים לחישוב סיכון קרדיו-וסקולרי(לדוגמה [75, 14, 20]). מודלים אלו משמשים לזיהוי חולים בסיכון, ומאפשרים כימות מדויק של הסיכון לאורך שנים רבות[26]. כיום, בצורותיהן השונות, מודלים אלו כלולים בקווים המנחים של ארגונים מקצועיים רבים, ולהם מקום חשוב במניעה הראשונית של מחלות קרדיווסקולריות[30, 26].

למודלים אלו מספר מאפיינים:

- ביצועיהם של מודלים אלו מיטבי כאשר משתמשים בהם באוכלוסיות עליהם פותחו, ופוחת באוכלוסיות השונות מאוכלוסיות אלו מבחינה גנטית או אחרת[19, 3, 21].
- המודלים מבוססים ככלל על שיטות ביוסטטיסטיות מסורתיות, בפרט על רגרסיה לוגיסטית ורגרסיית קוקס.
- השימוש במודלים דורש ידע בדבר תוצאות בדיקות (מעבדה ואחרות) שבחלקו אינו ידוע לחולים. כמו כן, המודלים עצמם בד"כ משמשים על מנת להחליט בנוגע לטיפולים הדורשים מעורבות רופא (כגון רישום סטטינים). כתוצאה מכך, ה"לקוחות" המסורתיים של מודלי חיזוי הם רופאים מטפלים, ובפרט אלו העוסקים ברפואה ראשונית.

כוונתנו להשיג שלוש מטרות בתזה זו:

- בהיותה מובחנת אתנית מהאוכלוסיות האמריקאיות והאירופאיות עליהן פותחו, ביצועיהם של מודלי חיזוי קרדיו-וסקולריים באוכלוסיה הישראלית צפויים להיות ירודים. השערה זו טרם נבדקה על אוכלוסיה גדולה, המייצגת את האוכלוסיה הישראלית כולה[45]. אנו מתכוונים לבדוק השערה זו על מדגם נרחב של האוכלוסיה הישראלית, בהקשר של חיזוי שבץ.
- חידושים מודרניים בלמידה חישובית מאפשרים גישות חדשות למידול סיכון רפואי[52]. בניגוד לגישה הקיימת, המבוססת על שימוש בידע תחומי לפירוט-מראש של גורמי הסיכון, גישות מודרניות מתירות לאלגוריתם החישובי, המוזן עם מאות משתנים אפשריים, לברור מביניהם את המשתנים הרלוונטיים בכחות

עצמו[73]. לאחר מכן, משתנים אלו משמשים לבניית המודל עצמו. טכנולוגיות אלו מאפשרות גישה יותר אחידה למידול סיכון, המשתמשת בבסיס נתונים נרחב יחיד עם תוצאים שונים[58]. אנו מתכוונים ליצור בסיס נתונים שכזה על מנת לבנות מודל סיכון חדש לשבץ, אשר ישמש בעבודה זו.

- טיפול של הפרט בעצמו מהווה מרכיב קריטי במניעת מחלות[54]. אנו מתכוונים לקדם טיפול עצמי שכזה ע"י הצגת מודל חיזוי ישירות למטופלים באמצעות אתר האינטרנט של קופ"ח כללית, בצורה המעודדת הפחתת גורמי סיכון.

על מנת לבצע תיקוף חיצוני, ייבחרו ארבעה מודלים לחיזוי מחלות קרדיווסקולריות וייבנו מחדש על גבי אוכלוסיית כללית. המודלים יושוו הן בהרכב האוכלוסיה המקורי עליה נבנו והן על אוכלוסיה משותפת התואמת לאוכלוסיה עליה בכוונתנו לבצע את התערבות.

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 בהם יעשה שימוש לבניית מודל. המודל יכוונן על גבי אוכלוסיית פיתוח וייבדק מול אוכלוסיית בדיקה.

ההתערבות מוכוונת החולה תעשה שימוש באתר האינטרנט של קופ"ח כללית על מנת להציג ישירות לחולים מחשבון סיכון לשבץ. משתני המחשבון יוזנו באופן אוטומטי מהתיק הרפואי הממוחשב של המבוטחים. המבוטח יוכל לשנות את גורמי הסיכון השונים ולראות את אפקט השינוי, ובד בבד יוצגו לו המלצות מותאמות אישית להפחתת סיכון קרדיווסקולרי. אנו נבדוק באופו פורמלי את האפקט של התערבות זו. נעשה זאת ע"י התאמת אוכלוסיית המשתמשים לאוכלוסיית בקרה מבחינת גיל, מין וסיכון רקע, וביצוע רגרסיה לוגיסטית רב-משתנית לבחינת אפקט ההתערבות.

תיקוף חיצוני של מודלים לחיזוי מחלה קרדיווסקולרית, המצויים בשימוש רחב ומשולבים בקווים המנחים, הוא בעל חשיבות מכרעת[13].

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

ביצוע והערכת ההתערבות המתוכננת יאפשר לנו לשפוט את היעילות של התערבויות שכאלה, להם חשיבות מיוחדות במניעת מחלות קרדיווסקולריות[62], ואשר הופכות אפשריות יותר ויותר עם הזמינות הגוברת של תיקים רפואיים ממוחשבים.

3 Aim of the Thesis

The main aim of this thesis is to implement and evaluate an intervention in the Clalit Health Services' (CHS) population meant to reduce the incidence of stroke. This intervention will be targeted directly to the Clalit's patients via an online web-based risk model that will be developed in-house as part of this project.

The aforementioned goal will require three steps:

Model Evaluation The intervention requires a well calibrated and discriminative risk model for stroke, capable of correctly identifying patients at high risk and correctly identifying the specific risk factors comprising their risk. As a first option, we will evaluate leading, well-known, international risk models on our patient population. 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.

Model Development As international models are often mis-calibrated when applied to a new population, different than the population on which they were developed, we will develop a new model based on a modern and novel approach to developing risk models from Electronic Health Record (EHR) data. The full details of This approach will be detailed below, under "Research Methodology". This new model will be compared to existing models, and the best one chosen for the actual intervention.

Patient Intervention The chosen model will be embedded within an online web-based application accessible to the Clalit's population via its online

website ("Clalit Online"). The application will show the patients their risk, will allow them to interact with different variables to illustrate their effect on the risk, and will advise on ways to mitigate the risk. Once sufficient time has elapsed, the intervention will be evaluated for its effect, as will be detailed under "Research Methodology".

Based on these aims, we hypothesize that:

1. Existing risk models will have good discrimination, but poor calibration, when applied to the Israeli population. More Generally, we hypothesize that overall, their performance when validated on the Israeli population will be diminished compared to their performance as reported in the medical literature when tested on population more ethnically similar to their training population.
2. That a model developed locally on the Israeli population will outperform international models.
3. 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.
4. That a direct-to-patient risk calculator, designed to present patients with their risk and allowing them ways by which to lower that risk, will result in less overall risk, improvement in risk factors and more health-seeking behavior among patients who have used the calculator.

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 Epidemiology of Cardiovascular Disease and Stroke

In its usual definition, cardiovascular disease (CVD) includes several disease categories[74]:

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[59], with coronary heart disease being the most common specific diagnosis[4].

While the rates of cardiovascular disease in general, and stroke in particular, have declined in developed countries over the last 30 years[43, 70], they remain significant public health problems, being the second most common cause of mortality and third most common cause of disability worldwide[46]. The statistics in Israel are similar[23].

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[77, 53]: Smoking, dyslipidemia, hypertension, diabetes, etc.

4.1.2 History of Multivariate 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[75, 49, 14, 33, 20, 34, 26]. Still the most notable of said risk models is the Framingham risk model family, developed on a US population in Massachusetts, Boston[75], and the SCORE risk model, developed in 2003 on a European population[14].

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[26] and their use in deciding on anti-platelet therapy[6], 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[40, 41], and have found use not only in prediction, but also in diagnosis[69]. This increasingly important place taken by risk models has brought about the publication

of guidelines designed to regulate and improve their creation[13]. As estimating the probability for existing and future disease is a significant portion of the clinical process[47], 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.

4.1.3 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[19, 3].

Recent models have tried to deal with this problem by including more ethnically varied populations[21] or recalibrating the model for each new population[40], but such efforts are limited to specific risk models, and even then have only been partially successful[17]. 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[17]. 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[48].

4.1.4 The Gap and our Thesis

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[7].

We suggest, as a first effort, to externally validate widely used risk models for the prediction of stroke 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.

4.2 Part II

4.2.1 Methodology of Traditional Risk Models

For traditional medical risk models, two design decisions are ubiquitous[73]:

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.

4.2.2 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[50].

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 odds ratio of an exposure for a given outcome. The model can also be used for prediction, disregarding all such assumptions.

4.2.3 Cox Proportional Hazards

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[16].

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 odd 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. It should be mentioned that 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.

4.2.4 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 odd ratios, but does afford more predictive accuracy when sufficient data exists[64].

4.2.5 The Rise of AI and Machine Learning

In recent years the fields of machine and statistical learning have seen a tremendous rise[52]. this growth in machine learning, including predictive modeling, has occurred thanks to three main factors[65]:

- 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[52, 22].

4.2.6 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[57].

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[37]. This fascinating area of research falls under the more general notion of algorithmic fairness, more widely studied in other non-medical fields[15], and is beyond the scope of this thesis.

4.2.7 Electronic Health Record based Observational Studies

Most medical risk models in wide-use were developed based on specialized cohort studies[27]. 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 EHR data, 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[45], encompasses the full extent of a patient's encounters with the health system[28].

4.2.8 The Gap and our Thesis

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 stroke. 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.

4.3 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[26, 6], bisphosphonates for osteoporosis risk [39] or CT angiogram for pulmonary thromboembolism risk[72]. 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.

It should be said that this entire time consuming act is expected to occur in an already time-strained primary care encounter[42].

4.3.3 Empowering Patients to Control Risk Factors

This physician-led approach to health maintenance is by no means the only one. Complementary to it is the idea of self-care, the enlightening definition for which we'll copy from the WHO verbatim: **"Self-care in health refers to the activities individuals, families and communities undertake with the intention of enhancing health, preventing disease, limiting illness and restoring health. These activities are derived from knowledge and skills from the pool of both professional and lay experience. They are undertaken by lay people on their own behalf, either separately or in participative collaboration with professionals"**[54].

These ideas are of particular importance in the field of CVD management and prevention as[62]:

- CVD is very prevalent, and management of its many risk factors, signs and symptoms could "drown" modern primary care.
- CVD has many risk factors that can be addressed independently by patients with little-to-no physician involvement (e.g. by physical activity, proper diet, smoking cessation, etc.)

4.3.4 Bridging Self-Care and Risk Modeling

Naturally, knowledge of one's health status is paramount to one's management of it. Such plans as the American Heart Association's "know your numbers" are geared towards encouraging patients to question their family physicians regarding several basic clinical markers and lab results[62]. While this approach also requires the direct involvement of the treating physician, this requirement is brought about by the structure of the American health

care system and availability of information, and is not inherent in the problem.

Whether such interventions are effective is an open question. While the American health care system does not allow easy appraising of the effectiveness of interventions such as "know your numbers", other health care systems do. A study conducted in Australia found that patients advised regarding their own hypertension in incidental pharmacy visits, had improved knowledge of hypertension management and sought health-care more often[10].

4.3.5 The Gap and our Thesis

We suggest that the structure of the Israeli health care system is ideal for performing and evaluating an intervention based on promoting CVD prevention self-care.

In this thesis, the risk model will be presented directly to the patient, with the different covariates extracted seamlessly from the patient's electronic health record. The presentation will include visual queues regarding the normal range of risk factors and an ability to view the effect of altering specific risk factors.

It will also include personalized recommendations for the reduction of risk based on the patient's characteristics, suggesting self-care options when such options exist, and recommending physician contact when such contact is warranted.

5 The Novelty of the Thesis

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

- External validation of existing risk models is of utmost importance[48], as these models are used constantly as part of existing guidelines (e.g.

the American Heart Association's pooled risk model and Statin treatment[26], FRAX and Osteoporosis treatment[40]). This is especially true, as previous external validation studies have at times documented significant mis-calibration[3, 17], that would make treatment decisions based on the models problematic.

- While development of a new risk model for the prediction of stroke in the Israeli population is important and novel by itself, we propose that the methodology by which the model will be developed, and specifically its wide applicability, requiring little human intervention and pre-processing, is by itself even more meaningful. 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.

- Risk-model based interventions have usually taken one of two forms: Either physician guidelines based on the risk score(e.g. [34, 20]); or the simple presentation of the risk model to patients, allowing them to calculate their own risk, assuming they know and have measured all their covariates(e.g. [55]). We propose a third alternative, by which a risk model is shown to the patients directly, but is "wired" into the patient's electronic health record, presenting the patient his risk based on his recorded measurements. This model is novel, and to the best of our information no previous such attempts have been made and formally evaluated. It is also particularly useful for the Israeli population, owing to the depth of EHR data in Israel[45].

6 Published Work

The epidemiological characteristics of CVD in general and of stroke in particular are well understood[43, 70], and the dominant risk factors in the

population well mapped[77, 53]. This is true both in the developed and in the developing world[46]. It is also true in Israel[23].

The increasingly central role filled out by risk prediction models in medicine has been observed[47], as have the challenges of developing such models based on Electronic Health Record (EHR) data[27, 28]. 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[13].

Many CVD risk models have been developed in the last 30 years, most prominent of which are the Framingham[75, 49, 20, 26], SCORE[14] and Qrisk[33, 34] families of models. Two of these model families also offer a stroke-specific model[76, 18, 32].

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[49, 30, 26, 6].

Models' tendency to under-perform when the target population is changed is widely recognized[19, 3, 21] and accordingly, the importance of external validation of models prior to their use in new population is recommended[48]. External validation of CVD models has been performed in several populations[19, 3, 21], though not in the Israeli population[7]. This is in contrary to, for example, Osteoporosis[17].

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 in increasing force[51].

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[52]. 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[58].

Attempts to more proactively involve patients in their own care have existed

for a long time[1]. As AI and machine learning advance, it is only natural that these technologies will take part in such interventions. These attempts have taken varied forms: general diagnostic models based on provider statistics[2], smartphone apps designed to ease communications between providers and patients[67], wearable devices designed to encourage lifestyle change[29], etc.

More similar to this thesis are calculators presented directly to patients. These are many, and some deal directly with CVD and stroke[55]. But none of these calculators are based on locally developed models and none are able to access the patient's EHR, instead relying on him being knowledgeable about his risk factors. Also, being disconnected from the patient's EHR, these models cannot follow up on a patient's risk and health behavior, and accordingly cannot evaluate their own performance.

7 Research Methodology

We will elaborate on the following for each of the three parts:

- Planning, including population definition and variables.
- Data extraction.
- Descriptive statistics.
- Modeling and inferential statistics.

7.1 Planning

7.1.1 Part I

We will externally validate four models that encompass a large and representative sample of existing models, two designed to assess risk for cardiovascular disease (CVD) and two designed to assess risk specifically for stroke.

The models with their respective populations and outcomes are:

Name	Age	Model	Outcome
American Heart Association 2013 pooled risk model[26]	40-79	Cox proportional hazards	Myocardial infarction, coronary heart disease death, stroke and stroke death
Framingham general CVD risk score[20]	30-74	Cox proportional hazards	CHD death, non-fatal MI, coronary insufficiency, stroke (ischemic or hemorrhagic), TIA, peripheral vascular disease, heart failure
Modified Framingham Stroke Risk Score[76, 18]	55-84	Cox proportional hazards	All forms of stroke (transient ischemic event, thrombotic, hemorrhagic and subarachnoid hemorrhage)
QStroke[32]	25-84	Cox proportional hazards	Transient ischemic event or thrombotic stroke

Table 1: Models to be Externally Validated

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

7.1.2 Part II

Our model will be developed on the following population.

Inclusion:

- 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:

- Past stroke event.

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 1/6/2007, and follow up will persist until 1/6/2017, as illustrated in the following design diagram.

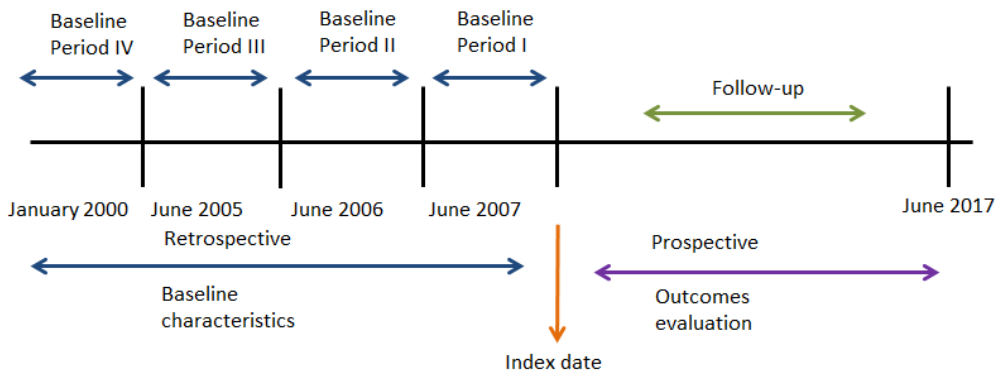

Figure 1: Study Design Timeline

Logically, model construction will encompass two steps: Using a sparsity inducing model to select among hundreds of variables, then building a model using the selected variables. In effect, algorithms will be chosen that perform both stages at once. See more details in the modeling section ahead.

The covariates supplied to the first step will be:

- Full demographic information, including age, sex, socioeconomic status, sector (arab/jew), ethnicity, etc.
- Clinical covariates, including blood pressure, height, weight, smoking status, etc. The data used will be the last result for each patient in the two years prior to the index date.

- Lab data, including all labs performed for each patient. Data will be extracted separately for the year before the index date, the year before that and the year before that.
- Chronic diagnoses, as defined by the Clalit's chronic registry[61], up to the index date.
- Drug dispensings, including all drug dispensed to the patient in ATC4 granularity[24]. Data will be extracted separately for the 3 years before the index date and all the years before that.

A full list can be found in appendix C.

7.1.3 Part III

The population for the application evaluation will include all members of the Clalit's insured population who will use the application from its launch date onwards. As the application will not display for patients with a prior stroke event, this exclusion will occur implicitly.

This will be a retrospective cohort design. The population will include all users of the application, with each user matched by age, sex and altogether 10-year risk with controls. The index date for each patient will be the date of his first recorded use of the application. The index date for each matched control will be the index date of his matched case. We'll collect covariates in the 3 years before the index date and outcomes in the 6 months following.

The following variables will be collected and used for adjustment:

1. Socioeconomic status
2. Sector (arab/jew)
3. Charlson score
4. Smoking status

5. Systolic and diastolic blood pressure
6. LDL
7. Weight
8. BMI
9. Drug adherence, as measured by the MPR and PDC indices[44]

We'll conduct a power analysis to calculate the required sample size using R's pwr package[11]. Assumptions:

- Model is a logistic regression model.
- Expected effect size is small, which for logistic regression is $R^2 = 0.02$[12].
- We'll have 11 confounders adjusted for, as just specified (the list plus use of the calculator, which is the main exposure).
- Requested power is 0.8, as usual.
- Significance level is 0.05.

We get a result of $n = 850$. As we assume we'll have at least 500 calculator users, we could settle for a 1-to-1 case-control ratio, but as controls are plentiful and the process automated, we'll opt for a 1-to-5 ratio.

Outcomes will be:

1. Change in overall risk.
2. Change in the four alterable risk factors:
   (a) Smoking status
   (b) Weight
   (c) LDL

(d) Systolic and diastolic blood pressure

3. Change in drug adherence.

4. Change in healthcare utilization (overall cost, number of physician visits and number of drug dispensings).

7.2 Data Extraction

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.

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[36, 38, 28]. 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. [60, 17]).

Data extraction principles for these studies are:

- 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[61]. 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[44].

- 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.

Information regarding patient's use of the calculator application will be collected directly from that application's database, which includes patients' information and access dates.

In part I, 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).

Stroke 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 stroke specialist. These definitions are very similar to those of the Israeli acute stroke registry[23] (active within the ICDC).

7.3 Descriptive Statistics

The specific population for each of the four models will be described in a dedicated population table ("Table 1") with appropriate statistics for each

variable: proportions for categorical variables, means and standard deviations for continuous variables.

The common population to be used for comparing the four external models and to construct the internal model will also be described in a population table. This table will include separate columns for the train and test populations (see ahead for modeling details), with the same appropriate statistics for each variable. Statistical tests will be used to compare these populations for differences in baseline variables that could affect model generalizability. The statistical tests to be used are Student's t-test for continuous variables and the Chi square goodness-of-fit test for categorical variables, once the basic assumptions (e.g. normality) are tested.

Lastly, the population in part III will also be displayed in a dedicated population table that will include columns for cases and controls. The same statistics and statistical tests as above will be used.

We will present basic descriptive graphs to visualize the relation between certain variables and the outcomes. For the internal predictive model in part II, this will obviously only be done using the training population, so as to avoid so-called "data leaks".

Missing data will be multiply imputed using chained equations. Specifically, continuous variables will be imputed using predictive mean matching, while categorical variables will utilize logistic regression[9]. Five datasets will be imputed, with the results combined as per Rubin's law[63].

7.4 Modeling and Inferential Statistics

7.4.1 Part I

All models will be evaluated twice:

1. Once on a population that exactly mirrors the population they were originally defined on.

2. Once on a common shared population that represents the population for which we intend to use the model in our thesis.

This design is similar to previously published work[17].

The first phase will employ the full population matching the model's inclusion and exclusion criteria, so as to mirror their development population as much as possible.

For the second phase we'll use only the inclusion and exclusion criteria detailed above. This will be a common, shared population so as to allow comparison of model's performance on a joint dataset.

The population will be separated into three sets for the sake of model development: Train, Validation and Test in a 72%/8%/20% ratio. The training and validation sets will be discussed in subsection "part II" ahead. The test set will be used for comparing model's performance.

The following performance statistics will be computed and reported for each model[66, 31]:

- 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[26, 6].

To calculate the risk scores, the exact coefficients as published by the model's authors will be used. If dedicated software is available, it will be used instead.

While the risk scores from the stroke models will be used as is, we will perform linear recalibration of the CVD models, as using their results as-is to predict stroke will necessarily incur mis-calibration due to "over-prediction". This

recalibration will be done using the framework suggested by Van Houwlingen et al[35]. Specifically, the model's linear predictor will be fit again as a sole predictor in a logistic regression model and the ensuing slope and intercept recorded. These will then be used to adjust all model predictions.

Mathematically:

$$\forall_i \mathrm{LP}_i = \sum_{j=1}^{p} \beta_j x_j$$

$$\hat{y}_i = \gamma \mathrm{LP}_i + \delta$$

Where LP_i is the linear predictor, β_{ij} is the coefficient for covariate j in patient i, x_{ij} is the covariate j in patient i, $\hat{y}$ is the recalibrated prediction, for which γ is the slope and δ the intercept.

Or in words: We take the linear predictor from the original model, but allow it a new slope and intercept, thus preserving the relative importance of each covariate in the model, with the freedom to reset the global risk.

7.4.2 Part II

To develop the new model, we will create a generic framework capable of generating models for any disease, given a fitting definition of the outcome.

The framework will serve two consecutive tasks. The first is to choose the relevant covariates from the long list of candidate covariates supplied to it. The second is to actually build the model.

It should be specifically noted that both parts carry independent significance. The covariate selection awards biological insight into the risk factors for a disease, while the model is the actual tool used for risk prediction.

The first step will involve applying a model to the training data that employs sparsity. That is, we will opt for models that include variable selection as

a part of the fitting process. The hyperparameters for these models will be tuned using the validation set.

The three sparsity inducing models we intend to fit are:

1. LASSO[68]
2. Gradient Boosting[25]
3. Random Forest[8]

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

$$\arg\min_{w} L(w) + \lambda \sum_{i} |\beta|_i$$

L being the likelihood function and lambda being a regularization parameter. 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, we will normalize the variables to have equal mean and standard deviation prior to model fitting.

Gradient boosting is an ensemble method that combines several weak learners (e.g. 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.

Random forest is also an ensemble method employing decision trees as the weak learners. It strives to induce variance among the trees by using bootstrapping to select the training set for each tree, and only using a randomly selected subset of features at each split in the tree.

Both gradient boosting and random forest induce 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:

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, will be conducted on the validation set using random search[5]. The best performing model with regard to area under the ROC curve will be selected.

The model as produced by the sparsity inducing algorithm will be compared to the existing models examined in phase I using the above mentioned performance measures. For the sake of demonstrating clinical utility, we will also compare the best model from phase I to our model for net reclassification improvement[56] and decision curves[71].

We will include a learning curve for our model so as to demonstrate lack of over-fitting.

While in essence we could use the models from phase I or the model as produced by our sparsity inducing algorithm as is, the requirement of using the model in a web application, including tight memory and responsiveness constrains, forces us to fit a simpler final model.

To account for this, we will take the covariates from the best model and use them construct a simple logistic regression model. This will be our final model to be used in the application.

7.4.3 Part III

Recall that this portion employs a cohort design and tests four outcomes: change in overall risk, change in specific risk factors, change in drug adherence and change in health-care utilization.

The modeling will be done using multivariate logistic regression for categorical outcomes and linear regression for continuous outcomes, adjusting for all the aforementioned variables. We'll report the respective coefficients, together with 95% confidence intervals and p-values.

8 Preliminary Results

We will present preliminary results for each of the 3 portions.

8.1 Part 1

Population Table for the Framingham Stroke Risk Score[18] model:

Variables	Categories	Population	censored	stroke	pval
Individuals	n	705703	665415	40288	
Sex	F	55.80%	56.10%	50.60%	
Sex	M	44.20%	43.90%	49.40%	<0.01
Age	Mean	66.4	66.2	69.7	
Age	SD	8.4	8.4	8.3	<0.01
Smoking	No	90.30%	90.30%	89.40%	
Smoking	Yes	9.70%	9.70%	10.60%	<0.01
Systolic BP	Mean	132.5	132.2	136.4	
Systolic BP	SD	17.4	17.4	18.6	<0.01
Diabetes Mellitus	Pos	17.10%	16.40%	28.30%	<0.01
Hypertension	Pos	47.90%	46.90%	63.40%	<0.01
CVD	Pos	24.20%	23.40%	36.70%	<0.01
AF	Pos	5.90%	5.60%	11.50%	<0.01
LVH	Pos	0.70%	0.60%	1.20%	<0.01

Calibration and ROC Curves for the Framingham Stroke Risk Score model are presented in appendix D.

8.2 Part 2

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

8.3 Part 3

Early mock drafts of the calculators for the Clalit's website, this specific one for chronic kidney disease, are presented in appendix D. Eventually, the finalized stroke calculator will look similar.

9 References

[1] Institute of Medicine (US) Committee on Quality of Health Care in America. "Crossing the Quality Chasm: A New Health System for the 21st Century". In: (2001).

[2] Meir Aurbach. *Instead of googling: A new app by Macabbi and Mhealth will make medical data available.* 2018. URL: https://www.calcalist.co.il/internet/articles/0,7340,L-3729191,00.html.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] Dominique A Cadilhac et al. "The Know Your Numbers (KYN) program 2008 to 2010: impact on knowledge and health promotion behavior among participants." In: *International journal of stroke : official journal of the International Stroke Society* 10 (1 Jan. 2015), pp. 110–116. ISSN: 1747-4949. DOI: 10.1111/ijs.12018.

[11] Stephane Champely. *pwr: Basic Functions for Power Analysis*. R package version 1.2-1. 2017. URL: https://CRAN.R-project.org/package=pwr.

[12] 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=0JYN1NVW651KCA56C102&tag=techkie-20&linkCode=xm2&camp=2025&creative=165953&creativeASIN=0805802835.

[13] Gary S Collins et al. "Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRI-

POD Statement." In: *European journal of clinical investigation* 45 (2 Feb. 2015), pp. 204–214. ISSN: 1365-2362. DOI: 10.1111/eci.12376.

[14] 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.

[15] 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].

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

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] WHO Collaborating Centre for Drug Statistics Methodology. *Guidelines for ATC classification and DDD assignment 2010*. Norwegian Institute of Public Health, 2010. ISBN: 978-8280823694. URL: https://www.amazon.com/Guidelines-ATC-classification-assignment-2010/dp/8280823697?SubscriptionId=0JYN1NVW651KCA56C102&tag=techkie-20&linkCode=xm2&camp=2025&creative=165953&creativeASIN=8280823697.

[25] 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.

[26] 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.

[27] 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.

[28] 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 As-*

sociation : JAMIA 24 (1 Jan. 2017), pp. 198–208. ISSN: 1527-974X. DOI: 10.1093/jamia/ocw042.

[29] Rebecca Gordon and Saul Bloxham. "Influence of the Fitbit Charge HR on physical activity, aerobic fitness and disability in non-specific back pain participants." In: *The Journal of sports medicine and physical fitness* 57 (12 Dec. 2017), pp. 1669–1675. ISSN: 1827-1928. DOI: 10.23736/S0022-4707.17.06688-9.

[30] 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.

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

[32] 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.

[33] 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.

[34] 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.

[35] 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.

[36] 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.

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

[38] 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.

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

[40] 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.

[41] 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.

[42] 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.

[43] 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.

[44] 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.

[45] 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.

[46] 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.

[47] 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.

[48] 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.

[49] 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.

[50] 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.

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

[52] 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.

[53] 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.

[54] World Health Organization et al. "Health education in self-care: Possibilities and limitations". In: (1984).

[55] Priya Parmar et al. "The Stroke Riskometer(TM) App: validation of a data collection tool and stroke risk predictor." In: *International journal of stroke : official journal of the International Stroke Society* 10 (2 Feb. 2015), pp. 231–244. ISSN: 1747-4949. DOI: 10.1111/ijs.12411.

[56] 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.

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

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

[59] 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.

[60] 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.

[61] 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.

[62] Barbara Riegel et al. "Self-Care for the Prevention and Management of Cardiovascular Disease and Stroke: A Scientific Statement for Healthcare Professionals From the American Heart Association." In: *Journal of the American Heart Association* 6 (9 Aug. 2017). ISSN: 2047-9980. DOI: 10.1161/JAHA.117.006997.

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

[64] 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.

[65] 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.

[66] 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.

[67] Kay Sundberg et al. "Early detection and management of symptoms using an interactive smartphone application (Interaktor) during radiotherapy for prostate cancer." In: *Supportive care in cancer : official journal of the Multinational Association of Supportive Care in Cancer* 25 (7 July 2017), pp. 2195–2204. ISSN: 1433-7339. DOI: 10.1007/s00520-017-3625-8.

[68] 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.

[69] 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.

[70] 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.

[71] 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.

[72] 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.

[73] 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.

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

[75] 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.

[76] 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.

[77] 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.

Appendices

A Model Variable Lists

1. American Heart Association 2013 pooled risk model
   (a) Sex
   (b) Age
   (c) Total Cholesterol
   (d) HDL
   (e) Treated Systolic Blood Pressure
   (f) Untreated Systolic Blood Pressure
   (g) Smoking Status
   (h) Diabetes
2. Framingham general CVD risk score
   (a) Sex
   (b) Age
   (c) Total Cholesterol
   (d) HDL
   (e) Treated Systolic Blood Pressure
   (f) Untreated Systolic Blood Pressure
   (g) Smoking Status
   (h) Diabetes
3. Modified Framingham Stroke Risk Score
   (a) Sex
   (b) Age

(c) Systolic Blood Pressure
(d) Hypertension Treatment
(e) Cardiovascular Disease
(f) Left Ventricle Hypertrophy
(g) Smoking Status
(h) Atrial Fibrillation
(i) Diabetes

4. QStroke
   (a) Sex
   (b) Age
   (c) Ethnicity
   (d) Smoking Status
   (e) Atrial Fibrillation
   (f) Systolic Blood Pressure
   (g) Total Cholesterol
   (h) HDL
   (i) BMI
   (j) Family History of Coronary Disease
   (k) Townsend Deprivation Score
   (l) Hypertension Treatment
   (m) Rheumatoid Arthritis
   (n) Chronic Kidney Disease
   (o) Type 1 Diabetes
   (p) Type 2 Diabetes
   (q) Coronary Heart Disease
   (r) Congestive Heart Failure
   (s) Valvular Heart Disease

B Extraction Protocol

B.1 Outcome Diagnoses

1. **Name** Intra-Cranial Hemorrhage

   ICD9 Codes 431%

   ICPC Codes NA

   CHR Codes NA

   Sources Admissions

   Free-Text Inclusion NA

   Free-Text Exclusion NA

   Comments Primary diagnosis only, not from rehabilitation

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

   Free-Text Inclusion NA

   Free-Text Exclusion NA

   Comments Primary diagnosis only, not from rehabilitation

3. **Name** CVA NOS

   ICD9 Codes 436%

   ICPC Codes NA

   CHR Codes NA

   Sources Admissions

   Free-Text Inclusion NA

   Free-Text Exclusion NA

Comments Primary diagnosis only, not from rehabilitation

4. **Name** Transient Ischemic Event

   ICD9 Codes 435%

   ICPC Codes NA

   CHR Codes NA

   Sources admissions, community, permanent, hospitals

   Free-Text Inclusion %transient%ischemic%attack%, %ischemic%attack%transient%, %transient%cerebral%ischemia%, %vertebral%artery%syndrome%, %ischemic%attack%transient%

   Free-Text Exclusion NA

   Comments Primary diagnoses only, not from rehabilitation, only community neurologis

5. **Name** Subarachnoid Hemorrhage

   ICD9 Codes 430%

   ICPC Codes NA

   CHR Codes NA

   Sources Admissions

   Free-Text Inclusion NA

   Free-Text Exclusion NA

   Comments Primary diagnosis only, not from rehabilitation

6. **Name** Myocardial Infarction

   ICD9 Codes 410%

   ICPC Codes NA

   CHR Codes NA

   Sources Admissions

   Free-Text Inclusion NA

Free-Text Exclusion NA

Comments Primary diagnosis only, not from rehabilitation

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

   Free-Text Inclusion %angina%, %prectoris%, %heart%attack%, %myocardial%inf%, %ischemic%heart%, %ischaemic%heart%, %coronary%atherosclerosis%, %arterioscl%cardiovascular%, %post%coronary%bypass%, %coronary%insuf%, %atheroscl%cardiovasc%, %acute%coronary%, %cardial%ischemia%, %intermediate%coronary%, %dyspnea%effort%, infarction%myocardial%, %infarction%subendocardial%, %subendocardial%infarction%

   Free-Text Exclusion %fear%, %gynecologic%, %no%disease%, %us%examination%, %normal%, %breast%, %medical%examination%, %herp%angina%, %hearing%

   Comments NA

8. **Name** Congestive Heart Failure

   ICD9 Codes 428%

   ICPC Codes NA

   CHR Codes 112%

   Sources community, admissions, permanent

   Free-Text Inclusion %congestive%heart%, %heart%failure%, %systolic%dysfunction%, %diastolic%dysfunction%, %ventricular%failure%, %CHF%, %ventricular%d[yi]sfunction%

   Free-Text Exclusion NA

   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

   Free-Text Inclusion %peripheral%vascular%, %PVD%, %claudication%, %buerger%, %thromboangiitis%obliterans%

   Free-Text Exclusion %neurogenic%, %spinal%, , %dissection%, %acute%, %vitreous%, %floater%, %eye%, %detachment%, %PVD%BE%, %BE%PVD%, %OD%PVD%, %PVD%OD%, %PVD%LE%, %LE%PVD%, %raynaud%

   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

   Free-Text Inclusion %cerebrovascular%accident%, %transient%ischemic%attack%, %intracerebral%hemorrhage%, %CVA%, %cerebelar%hemorrhage%,

%cerebral%hemorrhage%, %cerebral%vasospasm%, %cerebrovascular%disease%, %stroke%, %cerebral%ischemia%, %subarachnoid%hemorrhage%, %ischemic%attack%transient%, %aneurysm%berry%ruptured%, %intracranial%hemorrhage%, %hemorrhage%brain%nontraumic%

Free-Text Exclusion %extradural%

Comments NA

2. **Name** Left Ventricular Hypertrophy

   ICD9 Codes 429.3%

   ICPC Codes NA

   CHR Codes NA

   Sources community, admissions, permanent

   Free-Text Inclusion %cardiomegaly%, %ventricular%, %hypertrophy%

   Free-Text Exclusion NA

   Comments Primary diagnosis NA

3. **Name** Congestive Heart Failure

   ICD9 Codes 428%

   ICPC Codes NA

   CHR Codes 112%

   Sources community, admissions, permanent

   Free-Text Inclusion %congestive%heart%, %heart%failure%, %systolic%dysfunction%, %diastolic%dysfunction%, %ventricular%failure%, %CHF%, %ventricular%d[yi]sfunction%

   Free-Text Exclusion NA

   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

Free-Text Inclusion %angina%, %prectoris%, %heart%attack%, %myocardial%inf%, %ischemic%heart%, %ischaemic%heart%, %coronary%atherosclerosis%, %arterioscl%cardiovascular%, %post%coronary%bypass%, %coronary%insuf%, %atheroscl%cardiovasc%, %acute%coronary%, %cardial%ischemia%, %intermediate%coronary%, %dyspnea%effort%, infarction%myocardial%, %infarction%subendocardial%, %subendocardial%infarction%

Free-Text Exclusion %fear%, %gynecologic%, %no%disease%, %us%examination%, %normal%, %breast%, %medical%examination%, %herp%angina%, %hearing%

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

   Free-Text Inclusion %peripheral%vascular%, %PVD%, %claudication%, %buerger%, %thromboangiitis%obliterans%

   Free-Text Exclusion %neurogenic%, %spinal%, , %dissection%, %acute%, %vitreous%, %floater%, %eye%, %detachment%, %PVD%BE%, %BE%PVD%, %OD%PVD%, %PVD%OD%, %PVD%LE%, %LE%PVD%, %raynaud%

   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

Free-Text Inclusion %hypertension%, %hypertensive%, %hypert%with%, %nephrosclerosis%, %hypert%, %essential%hypert%, %hypertesion%, %hypertention%

Free-Text Exclusion %low%, %w/o%, %pulmonary%, %pulmoanry%, %ocular%, %portal%, %holter%, %no%hypert%, %no%retino%, %pre%hyper%, %borderline%, %prostat%, %hyperthy%, %hypertrig%, %ventricular%, %tonsil%, %hypertroph%, %hypertg%, %hyperton%, %cranial%, %endomet%, %adenoid%

Comments NA

7. **Name** Rheumatoid Arthritis

   ICD9 Codes 714.0%, 714.2%

   ICPC Codes L88%

   CHR Codes 231%

   Sources community, permanent, chronic registry, hospitals

   Free-Text Inclusion % rheumatoid% arthritis%, % arthritis% atrophic%

   Free-Text Exclusion NA

   Comments NA

8. **Name** Chronic Kidney Disease

   ICD9 Codes 585%

   ICPC Codes NA

   CHR Codes 177%

   Sources community, permanent, chronic registry, hospitals

   Free-Text Inclusion %chronic%kidney%, %chronic%renal%, %renal% failure%chronic%, %uremia%

Free-Text Exclusion NA

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

   Free-Text Inclusion %valv%, %stenosis%, %regurgitation%, %incompetence%, %insufficiency%, %ebstein%, %tricuspid%atresia%, %pulmonary%atresia%

   Free-Text Exclusion NA

   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 Generic Predictor Variable List

1. age
2. birth_area_desc
3. bmi_last_v
4. charlson
5. DBP_last_v
6. date_of_birth
7. date_of_death
8. ethnicity
9. GFR
10. immigration_date
11. LEUCOCYTES-SED
12. ERYTHROCYTES-SED
13. EPITHELIAL-SED
14. BACTERIA-SED
15. ESTRADIOL (E-2)
16. PROGESTERONE
17. 17-OH-PROGESTERONE
18. LH
19. Throat culture p
20. FSH
21. WBC
22. PROLACTIN
23. TESTOSTERONE- TOTAL
24. Aerobic blood cult.

25. Anaerobic blood cult
26. Bacterial culture
27. Body fluid culture
28. Ear culture 1
29. Fungal culture
30. DHEA SULPHATE
31. MRSA culture
32. CORTISOL-BLOOD
33. Pediatric blood cul
34. Sputum culture
35. Stool culture
36. Throat culture
37. Urine dipsl culture
38. Urine culture
39. Urine plating cult.P
40. Wound culture
41. RBC
42. Wound and Sec.Aer+An
43. TSH
44. First isolate
45. Second isolate
46. PRELIMINARY
47. T3-TOTAL
48. T3- FREE
49. CPE culture result
50. T4- FREE

51. HB
52. PTH
53. HCT
54. PLT
55. MCV
56. MCH
57. MCHC
58. RDW
59. MPV
60. BAB- Blood agar base
61. MacConkey agar
62. Sabour dextrose agar
63. PCT
64. ESBL test
65. PDW
66. MID abs.
67. MID %
68. VAR-F
69. LI
70. HDW
71. MICRO-F
72. ABO conf final (ad)
73. Rh confirmation
74. RH
75. BLOOD TYPE
76. Antibody screen Fin

77. BLAST-F
78. LUC%
79. LUC abs
80. MPXI
81. LEFT SHIFT
82. MACRO%
83. MICRO %
84. HYPER%
85. HYPO %
86. ANISO-F
87. LYMP.abs
88. LYM%
89. NEUT.abs
90. NEUT%
91. MONO.abs
92. MONO%
93. EOS.abs
94. EOS %
95. BASO abs
96. BASO %
97. HCT/HGB Ratio
98. Gram stain direct
99. Parasites microscopy
100. RDW-SD
101. RDW-CV
102. RETICUL. COUNT abs

103. RETICULOCYTES COUNT%
104. ALY%
105. ALY
106. LIC%
107. LIC
108. P-LCR
109. CHr
110. CH
111. PLATLATE CLUMPS
112. MICRO%/HYPO%
113. NORMOBLAST.abs
114. NORMOBLAST.%
115. TRANSGLUTAMINASE_IgA
116. C13 UREA BREATH CALC
117. PT-INR
118. PT-SEC
119. PT %
120. APTT-sec
121. APTT-R
122. FIBRINOGEN CALCU
123. FIBRINOGEN
124. CONTROL PT
125. CONTROL PTT
126. OCCULT BLOOD STOOL
127. LYMPHOCYTES %-DIF
128. GLUCOSE

129. OCCULT BLOOD SCREEN
130. NEUTROPHILS%-DIF
131. UREA
132. MONOCYTES%-DIF
133. CREATININE
134. EOSINOPHILS%-DIF
135. URIC ACID
136. BASOPHILS%-DIF
137. SODIUM
138. POTASSIUM
139. CHLORIDE
140. CALCIUM
141. lab_208512
142. PHOSPHORUS
143. CHOCOLATE
144. STABS %-DIF
145. PROTEIN-TOTAL
146. ALBUMIN
147. ATYPICAL LYMPH.%-DIF
148. CHOLESTEROL
149. LYMPHOCYTES abs-DIF
150. TRIGLYCERIDES
151. NEUTROPHILS abs-DIF
152. CHOLESTEROL- HDL
153. MONOCYTES abs-DIF
154. CHOLESTEROL-LDL calc

155. EOSINOPHILS abs-DIF
156. BASOPHILS abs-DIF
157. ALK. PHOSPHATASE
158. GOT (AST)
159. GPT (ALT)
160. STABS abs-DIF
161. GGT
162. LDH
163. ATYPICAL LYMPH-DIF
164. CK-CREAT.KINASE(CPK)
165. ANISOCYTOSIS-DIF
166. AMYLASE
167. IRON
168. TRANSFERRIN
169. BILIRUBIN TOTAL
170. BILIRUBIN-DIRECT
171. ALBUMIN (BY EP)
172. GLOBULIN ALPHA-1
173. GLOBULIN ALPHA-2
174. GLOBULIN GAMA
175. REMARK-MANU-DIF
176. MAGNESIUM
177. HEMOGLOBIN A1C %
178. FRUCTOSAMINE
179. SODIUM
180. POTASSIUM

181. CHLORIDE
182. CALCIUM IONIZED
183. pH BLOOD
184. pCO2
185. TOTAL HEMOGLOBIN
186. HCO3
187. TCO2
188. pO2
189. O2 SATURATION
190. Hct- OXIMETRY
191. CARBOXYHEMOGLO-OXIME
192. DEOXYHEMOGLOBIN-OXIM
193. METHEMOGLOBIN-OXIMET
194. OXYHEMOGLOBIN-OXIMET
195. STD BASE EXC (SBE)
196. TOTAL OXYGEN CONTENT
197. O2 TEN. AT 50% SAT
198. ACTUAL BASE EXCBSS
199. STANDARD BICARBONATE
200. LACTATE
201. VITAMIN B12
202. VITAMIN D (25-OH)
203. LIPASE
204. BILIRUBIN-NEONATAL
205. ZINC
206. REMARK-MAN-DIF

207. TROPONIN I
208. TROPONIN T
209. FOLIC ACID
210. CHOLESTEROL/ HDL
211. TRANSFERRIN SATURATI
212. MICROALBUMIN/CREAT
213. GLUCOSE I
214. BILIRUBIN INDIRECT
215. IRON SATURATION
216. GLOBULIN
217. ICTERIC
218. HEMOLYTIC
219. LIPEMIC
220. VLDL
221. ESR
222. CREATININE- U 24h
223. PROTEIN- U SAMPLE
224. CREATININE- U SAMPLE
225. SODIUM-URINE SAMPLE
226. POTASSIUM- U SAMPLE
227. CALCIUM- URINE 24h
228. PROTEIN- URINE 24h
229. MICROALBUMIN- U 24h
230. CALCIUM- U SAMPLE
231. MICROALBUMIN-U SAMP
232. HbF

233. HbA2
234. CARBAMAZEPINE
235. DIGOXIN
236. VALPROIC ACID
237. GLUCOSE 50g
238. CREATININE ENZ.CHILD
239. GLOM.FILTR.RATE
240. NON-HDL_CHOLESTEROL
241. ANA PATTERN
242. ANA TITER
243. CELIAC SCREEN
244. ANTISTREPTOLYSIN O
245. C-REACTIVE PROTEIN
246. RHEUMATOID FACTOR
247. ANTINUCLEAR Ab_(ANA)
248. ANTI CARDIOLIPIN IgM
249. THYROGLOBULIN Ab
250. DNA (ds) Ab
251. RNP
252. Sm (anti Smith Ab)
253. COMPLEMENT C3
254. COMPLEMENT C4
255. TOXOPLASMA IgG
256. TOXOPLASMA IgM
257. HELICO PYLORI IgG
258. REMARKS (GENERAL)

259. REMARKS MICRO
260. HEPATITIS Bs Ag
261. HEPATITIS C Ab
262. HEPATITIS Bs Ab
263. HEPATITIS A IgM
264. CMV IgG
265. CMV IgM
266. EBV VCA_IgG
267. EBV IgG-EBNA
268. EBV VCA IgM
269. RUBELLA Ab IgG
270. Anti Jo-1 Ab
271. Scl-70 Ab
272. ANTI CARDIOLIPIN IgG
273. HEPATITIS Bc Ab TOT.
274. Ig-E TOTAL
275. ANTI THYROID PEROXID
276. COOMBS INDIRECT
277. IgG
278. IgM
279. IgA
280. ANTICENTROMERE Ab
281. ALPHA FETOPROTEIN TM
282. CA-125
283. CA-15-3
284. CA-19-9

285. CEA
286. FERRITIN
287. PSA
288. GLUCOSE - U STRIP
289. BILIRUBIN- U STRIP
290. KETONES- U STRIP
291. SPECIFIC GRAV-U STRI
292. PROTEIN- U STRIP
293. NITRITE- U STRIP
294. LEUCOCYTES - U STRIP
295. ERYTHROCYTES-U STRIP
296. UROBILINOGEN-U STRIP
297. PH- U STRIP
298. SBP_last_v
299. sector
300. SES
301. sex
302. smoking_last_v
303. sw_confined
304. sw_immigrant
305. sw_malig_active
306. sw_malig_ever
307. sw_nursing_care
308. Antiinfectives for Local Oral Treatment
309. Corticosteroids for Local Oral Treatment
310. Calcium Compounds

311. Combination of Complexes of Calcium, Magnesium and Aluminum-Compounds
312. H2 Receptor Antagonists
313. Proton Pump Inhibitors
314. Synthetic Anticholinergics, Esters with Tertiary Amino Groups
315. Papaverine and Derivatives
316. Other drugs for func. gastro. disorders
317. Other Antispasmodics in Combination with Analgesics
318. Propulsives
319. Serotonin (5HT3) Antagonists
320. Other Antiemetics
321. Bile Acid and derivatives
322. Softeners, Emollients
323. Contact Laxatives
324. Osmotically Acting Laxatives
325. Enemas
326. Other drugs for constipation
327. Charcoal Preparations
328. Bismuth Preparations
329. Oral Rehydrating Salt Formulations
330. Antipropulsives
331. Aminosalicylic Acid and Similar Agents
332. Antidiarrheal Microorganisms
333. Enzyme Preparations
334. Insulins and Analogues, for Injection, Fast Acting
335. Insulin and Analogues, for Injection, Long Acting

336. Biguanides
337. Sufonylureas
338. Combinations or Oral Blood Glucose Lowering Drugs
339. Alpha Glucosidase Inhibitors
340. Dipeptidyl peptidase 4 (DPP-4) Iinhibi.
341. Glucagon-like peptide-1(GLP-1) analogues
342. Sodium-glucose co-transfer 2(SGLT2) Inhibitors
343. Other Blood Glucose Lowering Agents, Excluding Insulins
344. Multiple Vitamins with Minerals
345. Vitamin D and Analogues
346. Vitamin B1, in Combination with Vitamin B6 and/or B12
347. Ascorbic Acid (Vitamin C), Plain
348. Ascorbic Acid (Vitamin C), Combinations
349. Other Plain Vitamin Combinations
350. Vitamins, Other Combinations
351. Calcium
352. Calc.,Comb.with vit.D and/or other drugs
353. Potassium
354. Magnesium
355. Varoius Alimentary Tract and Metabolism Products
356. Vitamin K Antagonists
357. Heparin Group
358. Platelet Aggregation Inhibitors, Excluding Heparin
359. Direct Thrombin Inhibitors
360. Direct factor Xa inhibitors
361. Iron Bivalent, Oral Preparations

362. Iron Trivalent, Oral Preparations
363. Iron , Parenteral Preparations
364. Iron in Combination with Folic Acid
365. Iron in Other Combinations
366. Vitamin B12 (Cyanocobalamine and Derivatives)
367. Folic Acid and Derivatives
368. Other Antianemic Preparations
369. Digitalis Glycosides
370. Antiarrhythmics, Class IC
371. Antiarrhythmics, Class III
372. Organic Nitrates
373. Imidazoline Receptor Agonists
374. Alpha Adrenergic Blocking Agents
375. Thiazides, Plain
376. Sulfonamides, Plain
377. Aldosterone Antagonists
378. Other Antihemorrhoidals for Topical Use
379. Beta Blocking Agents, Non Selective
380. Beta Blocking Agents, Selective
381. Alpha and Beta Blocking Agents
382. Dihydropyridine Derivatives
383. Phenylalkalylamine Derivatives
384. ACE Inhibitors, Plain
385. ACE Inhibitors and Diuretics
386. ACE Inhibitors and Calcium Channel Blockers
387. Angiotensin II Antagonists, Plain

388. Angiotensin II Antagonists and Diuretics
389. Angiotensin II Antagonists and Calcium Channel Blockers
390. HMG CoA Reductase Inhibitors
391. Fibrates
392. Other lipid modifying agents
393. Imidazole Derivatives
394. Other Antifungals for Topical Use
395. Antifungals for Systemic Use
396. Zinc Products
397. Soft Paraffin and Fat Products
398. Carbamide Products
399. Salicylic Acid Products
400. Other Emollients and Protectives
401. Protectives Against UV-Radiations for Topical Use
402. Cod Liver Oil Ointments
403. Other Cicatrizants
404. Antihistamines for Topical Use
405. Anesthetic for Topical Use
406. Other Antipsoriatics for Topical Use
407. Other Antibiotics for Topical Use
408. Sulfonamides
409. Antivirals
410. Corticosteroids, Potent (Group III)
411. Corticosteroids, Very Potent (Group IV)
412. CORTICOSTEROIDS, WEAK, COMBINATIONS WITH ANTIBIOTICS

413. CORTICOSTEROIDS, MODERATELY POTENT, COMBINATIONS WITH ANTIBITICS
414. Corticosteroids, Potent, Combinations with Antibiotics
415. Corticosteroids, Potent, Other Combinations
416. Biguanides and Amidines
417. Phenol and Derivatives
418. Iodine Products
419. Other Antiseptics and Disinfectants
420. Retinoids for Topical Use in Acne
421. Peroxides
422. Antiinfectives for Treatment of Acne
423. Retinoids for Treatment of Acne
424. Medicated Shampoos
425. Wart and Anti-Corn Preparations
426. Other Dermatologicals
427. Imidazole Derivatives
428. Progestogens and Estrogens, Fixed Combinations
429. Progestogens
430. 3-Oxoandrosten (3) Derivatives
431. Natural and Semisynthetic Estrogens, Plain
432. Pregnen (4) Derivatives
433. Estren Derivatives
434. Progestogedns and Estrogens in Fixed Combination
435. Gonadotrophins
436. Antiandrogens and Estrogens
437. Selective Estrogen Receptor Modulator

438. Acidifiers
439. Urinary Concrement Solvents
440. Drugs for urinary frequency and incontinence
441. Drugs Used in Erectile Dysfunction
442. Other Urologicals
443. Alpha-Adrenoreceptor Antagonists
444. Testerone-5 Alpha Reductase Inhibitors
445. Other Drugs Used to Treat Benign Prostatic Hypertrophy
446. Somatotrophin and Somatrophin Agonists
447. Vasopressin and Analogues
448. Glucocorticoids
449. Thyroid Hormones
450. Other anti-parathyroid agents
451. Tetracyclines
452. Penicillins with Extended Spectrum
453. Beta Lactamase Sensitive Penicillins
454. Combinations of Penicillins, Including Beta-Lactamase Inhibitors
455. First-Generation Cephalosporins
456. Second-Generation Cephalosporins
457. Combinations of Sulfonamides and Trimethoprim, Including Derivatives
458. Macrolides
459. LINCOSAMIDES
460. Fluoroquinolones
461. Nitrofuran Derivatives
462. Other Antibacterials

463. Triazole Derivatives
464. Nucleosides and Nucleotides (excl. Reverse Transcriptase Inhibitors)
465. Antivirals for treatment of HIV infections, combinations
466. Influenza Vaccines
467. Hepatitis Vaccines
468. Bacterial and Viral Vaccines, Combined
469. Pyrimidine Analogues
470. Monoclonal Antibodies
471. Other Antineoplastic Agents
472. Gonadotrophin Releasing Hormone Analogues
473. Antiestrogens
474. Aromatase inhibitors
475. SELECTIVE IMMUNOSUPPRESANTS
476. Tumor Necrosis Factor Alpha (TNBF-a) Inhibitors
477. Calcineurin Inhibitors
478. Other Immunosuppressants
479. Acetic Acid Derivatives and Related Substances
480. Oxicams
481. Propionic Acid Derivatives
482. Coxibs
483. Other Antiinflammatory and Antirheumatic Products, Non Steroids
484. Antiinflammatory Preparations, Non Steroid for Topical Use
485. Capsicum and Similar Agents
486. Preparations with Salicylic Acid Derivatives
487. Ethers, Chemically Close to Antihistamines
488. Other Centrally Acting Agents

489. Preparations Inhibiting Uric Acid Production
490. Preparations with No Effect on Uric Acid Metabolism
491. Biphosphonates
492. Biphosphonates and Calcium, Sequential Preparations
493. Amides
494. Natural Opium Alkaloids
495. Phenylpiperidine Derivatives
496. Oripavine Derivatives
497. Opioids in combin. with non-opioid analg.
498. Other Opioids
499. Pyrazolones
500. Anilides
501. SELECTIVE 5HT-RECEPTOR AGONISTS
502. Barbiturates and Derivatives
503. Hydantoins and Derivatives
504. Benzodiazepine Derivatives
505. Carboxamide Derivatives
506. Fatty Acid Derivatives
507. Other Antiepileptics
508. Tertiary Amines
509. DOPA and DOPA Derivatives
510. Amantane Derivatives
511. Dopamine Agonists
512. Monoamine Oxidase Type B Inhibitors
513. Phenothiazines with Aliphatic Side Chain
514. Phenothiazines with Piperazine Structure

515. Phenothiazines with Piperidine Structure
516. Butyrophenone Derivatives
517. Thioxanthine Derivatives
518. Dibenzepines, Oxazepines, thiazepines, and oxepines
519. Benzamides
520. Lithium
521. Other Antipsychotics
522. Benzodiazepine Derivatives
523. Benzodiazepine Derivatives
524. Benzodiazepine Related Drugs
525. Other Hypnotics and Sedatives
526. Nonselective Monoamine Reuptake Inhibitors
527. Selective Serotonin Reuptake Inhibitors
528. Other Antidepressants
529. Centrally Acting Sympathomimetics
530. Anticholinesterases
531. Other Anti-Dementia Drugs
532. Drugs Used in Nicoitine Dependence
533. Antivertigo Preparations
534. Nitroimidazole Derivatives
535. Aminoquinolines
536. Benzimidzole Derivatives
537. Other Ectoparasiticides, Including Scabies
538. Other Insecticides and Repellants
539. Sympathomimetics, Plain
540. Sympathomimetics, Combinations Excluding Corticosteroids

541. Corticosteroids
542. Other Nasal Preparations
543. Sympathomimetics
544. Antiseptics
545. Selective Beta-2 Adrenoreceptor Agonists
546. Adrenergics in combinations with corticosteroids or other drugs, excl. anticholinergics
547. Corticosteroids
548. Anticholinergics
549. Leukotriene Receptor Antagonists
550. Expectorants
551. Mucolytics
552. Opium Alkaloids and Derivatives
553. Other Cough Suppressants
554. Opium Derivatives and Expectorants
555. Substituted Alkylamines
556. Phenothiazine Derivatives
557. Piperazine Derivatives
558. Other Antihistamines for Systemic Use
559. ANTIBIOTICS
560. Fluoroquinolones
561. Corticosteroids, Plain
562. Corticosteroids and Antiinfectives in Combination
563. Sympathomimetics in Glaucoma Therapy
564. Carbonic Anhydrase Inhibitors
565. Beta Blocking Agents

566. Prostaglandin Analogues
567. Sympathomimetics Used as Decongestants
568. Other Antiallergics
569. Other Ophthalmologicals
570. Corticosteroids and Antiinfectives in Combination
571. Analgesics and Anesthetics
572. Indifferent Preparations
573. Antiinfectives
574. Drugs for Treatment of Hyperkalemia
575. Nutrients with Low Calcium Content
576. Other Infant Formulas
577. Fat/Carbohydrate/Protein/Minerals/Vitamins, Combinations
578. Solvents and Diluting Agents, Including Irrigating Solutions
579. Cosmetics-C
580. Cosmetics-E
581. Hearing loss and Deafness
582. S - IHD (s/p MI)
583. S - Ischemic Heart Disease
584. Valvular Cardiac Dis (excl. MVP)
585. CHF-systolic w/o selected medications
586. CHF-systolic with selected medications
587. CHF-non systolic
588. CHF NOS with diuretics
589. CHF NOS
590. Cardiomyopathy
591. IHSS

592. Atrial fibrilation
593. Arrhythmia other
594. S - Hypertension / Diet Treatment
595. S - Hypertension / Drug Treatment
596. S - Hypertension / Unknown Treatment
597. Pulmonary Hypertension
598. s/p CVA
599. Carotid Artery Disease
600. S - Diabetes PVD
601. S - PVD
602. Aortic Aneurism
603. S - Amputation of Limb (Diabetic)
604. Amputation of Limb (Non-Diabetic)
605. S - Amputation of Limb
606. COPD
607. S - Asthma
608. s/p Pulmonary Embolism
609. s/p Pneumothorax
610. Chronic Bronchitis
611. Bronchiectasis
612. Hepatitis B Carrier
613. Celiac Disease
614. Peptic Ulcer
615. Reflux Esophagitis / Gastritis / Deudenitis
616. Irritable Bowel Syndrome
617. Addisons Disease

618. Chronic Act/Per Hepatitis
619. Cirrhosis
620. Wilsons Disease
621. Other Liver Disease
622. Ulcerative Colitis
623. Crohns Disease
624. H - Dialysis
625. Kidney Transplant
626. Prostatic Hypertrophy
627. Liver Transplant
628. Heart Transplant
629. Pancreas Transplant
630. Lung Transplant
631. Bone Marrow Transplant
632. Other Transplant
633. Unknown Transplant
634. Chronic Renal Failure
635. s/p splenectomy
636. Infertility Male/Female
637. S - Diabetic Nephropathy
638. S - Other Kidney Disease
639. Hepatitis C Carrier
640. Pemphigus Vulgaris
641. Psoriasis
642. Hidradenitis Suppurativa
643. Arthropathy

644. SLE
645. Rheumatoid Arthritis
646. Osteoporosis
647. Joint Replacement
648. Sarcoidosis
649. Scleroderma
650. Behcets Disease
651. Other Rheumatic / Autoimmune
652. Polymyalgia Rheumatica
653. Gout
654. s/p Head of Femur Fracture
655. Congenital Anomalies
656. H - AIDS Patient
657. HIV Carrier
658. Familial Mediteranean Fever
659. G-6-P-D Deficiency
660. Amyloidosis
661. S - Breast Cancer
662. S - Malignancy of Colon or Rectum
663. S - Malignancy of Prostate
664. S - Malignancy of Lung
665. S - Malignancy of Bladder
666. S - Malignancy of Ovary
667. S - Malignancy of Uterus
668. S - Malignancy of Pancreas
669. S - Malignancy of Brain / CNS

670. S - Stomach Cancer
671. S - Melanoma
672. S - Hodgkins Lymphoma
673. S - Non Hodgkin Lymphoma / Mycosis Fungoides
674. S - Acute Leukemia
675. S - Chronic Leukemia
676. S - Malignancy of Kidney
677. S - Malignancy of Larynx
678. S - Malignancy of Cervix Uteri
679. S - Malignancy of Pharynx
680. S - Malignancy of Esophagus
681. S - Malignancy of Liver / Bile Ducts
682. S - Malignancy of Thyroid
683. S - Malignancy of Bone
684. S - Malignancy of Connective Tissue / Sarcoma
685. S - Malignancy of Other Male/Female Genital Organs
686. S - Multiple Myeloma
687. S - Polycytemia Vera
688. S - Myelodysplastic Syndrom
689. S - Myelo/Lymphoproliferative Syndrom
690. S - Neurofibromatosis
691. S - Malignancy of Other Sites
692. S - Malignancy of Unknown Site
693. Tuberculosis
694. Benign Brain Tumor
695. Disability / Bedbound

696. Disability / Homebound
697. Disability / Requires assistance with ambulation
698. Breast Family History
699. Colon Family History
700. Hyperthyroidism
701. Hypothyroidism
702. Tuberculosis s/p
703. S - Diabetes / Diet Treatment
704. S - Diabetes / Oral Treatment
705. S - Diabetes / Insulin Treatment
706. S - Diabetes / Insulin + Oral Treatment
707. S - Diabetes / Unknown Treatment
708. H - Gaucher Disease
709. Hypo/Hyperparathyroidism
710. Acromegaly
711. Obesity
712. S - Hyperlipidemia / No Treatment
713. S - Hyperlipidemia / Treatment
714. S - Hyperlipidemia / Unknown Treatment
715. Cystic Fibrosis
716. Hyperprolactinemia
717. Other Endocrine and Metabolic Disease
718. Syphilis / Gonorrhea
719. Cushings Disease
720. Diabetes Insipidus
721. Hypophysary Adenoma

722. H - Hemophilia
723. H - Thalassemia Minor
724. H - Thalassemia Intermedia
725. H - Thalassemia Major
726. H - Thalassemia NOS
727. Other Hematologic Dis (excl. Iron Def Anemia)
728. Sickle Cell Anemia
729. Pernicious Anemia
730. ITP
731. Psychoses
732. Schizophrenia
733. Bipolar Disease (Manic Depressive)
734. Autism
735. Neuroses
736. Depression
737. Anxiety
738. Eating Disorders
739. Alcohol Abuse
740. Current Smoker
741. Former Smoker
742. Smoker
743. Drug Abuse
744. Mental Retardation (incl. Down)
745. Dementia / Alzheimers / OMS
746. Myasthenia Gravis
747. Parkinsons Disease

748. Epilepsy
749. Multiple Sclerosis
750. Cerebral Palsy
751. Huntingtons Chorea
752. Familial Dysautonomia
753. Hereditary Neurological Disease
754. Muscular Dystrophy
755. Motor Neuron Disease
756. S - Diabetic Neuropathy
757. s/p TIA
758. S - Other Neurological Disease
759. S - Diabetic Retinopathy
760. Non-Diabetic Retinopathy
761. S - Retinopathy
762. Glaucoma
763. Blindness
764. Retinitis Pigmentosum
765. Chronic Medication User

D Preliminary Result Graphs and Drawings

D.1 Framingham Stroke Risk Score (FSRS) Result Graphs

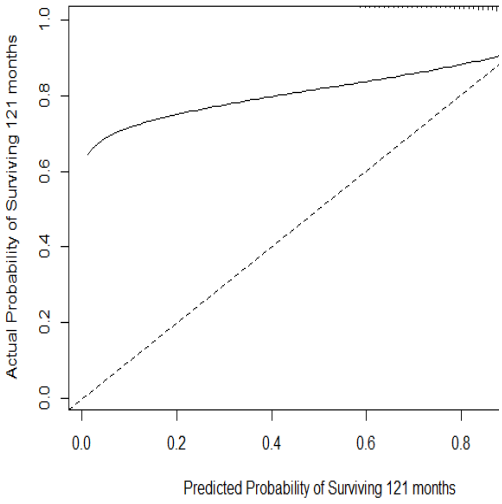

Figure 2: FSRS Calibration Curve

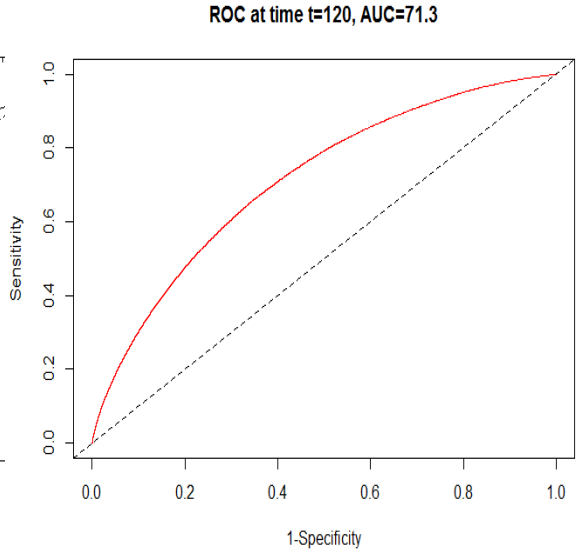

Figure 3: FSRS ROC Curve

D.2 Clalit Model Population Flow Chart

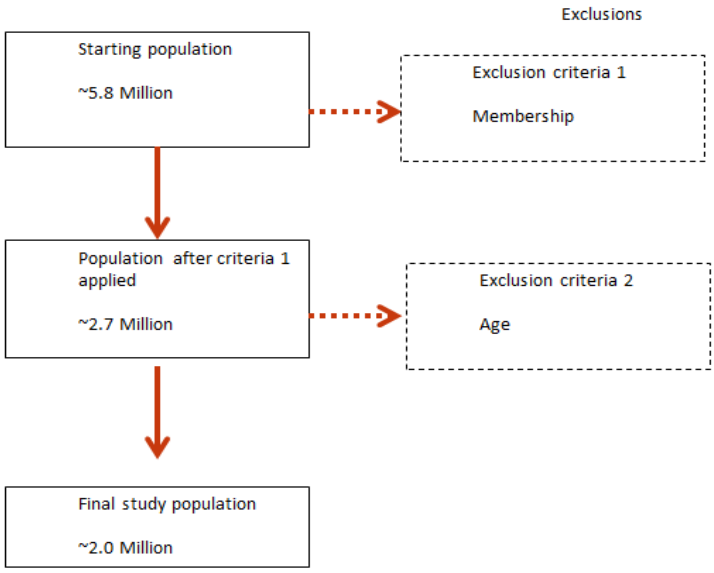

Figure 4: Population Flow Chart

D.3 Application Mock Drafts