Deep Science Publishing
https://doi.org/10.70593/978-93-49307-36-0_6

DeepScience
Open Access Books

Chapter 6: Predictive health modeling and risk stratification using longitudinal patient data

6.1. Introduction

Prognostic models from electronic medical records are widely used in medicine to manage and individualize the care of patients. Many of these models are specified to predict the risk of a well-defined event of interest in the future. This work focuses on the use of longitudinal patient data in the electronic health records (EHRs) to model and predict patient-specific health status over time, as well as using this predicted status to define patient risk groups based on future EHR data. The proposed methodology is demonstrated and evaluated on EHR historical data from a population of patients with advanced age simulated in a numerical experiment. The method and data for the experiment are provided, a predetermined list of analysis plans. The possible future developments and extensions of the work are also provided.

Healthcare organizations have been reconstructing digital records of their patients for several decades (Ganguli, 2024; Institute of Medicine, 2024; Toronto General Hospital, 2024). The data may include patient characteristics, diagnoses, prescribed medications and laboratory tests. The data is typically recorded in the form of multivariate time series. The use of patient data for research and personalized medicine offers the possibility of daily monitoring and immediate access to a large amount of data, which allows the usage of sensors to track patient activities. A substantial portion of current machine learning research is devoted to this goal through big data analytics for prognosis with time series, which aims to predict clinically relevant outcomes in the future from these high-frequency data. Statistical modeling with the use of patient data is a further step in understanding patients and treating them as individuals rather than members of the population group. On the basis of such data are built both simple prognostic models, and

deep models that allow for treatment advice to be made. However, the complexity and volume of patient data collected in the EHR can make modeling a challenging task.

In the first stage, the Cox regression model uses the patient's treatment progress features to estimate the risk weights for each patient in the training set, which means the risk of having chronic diseases in future periods. At the second stage, the convolutional neural network (CNN) used this information to determine whether it was classified as a high-risk chronic patient or not. The application of the proposed approach has been implemented for CareSet data from the hospital and obtained performance was evaluated. Moreover, a comparison study with commonly used algorithms demonstrates the effectiveness of proposed methods.

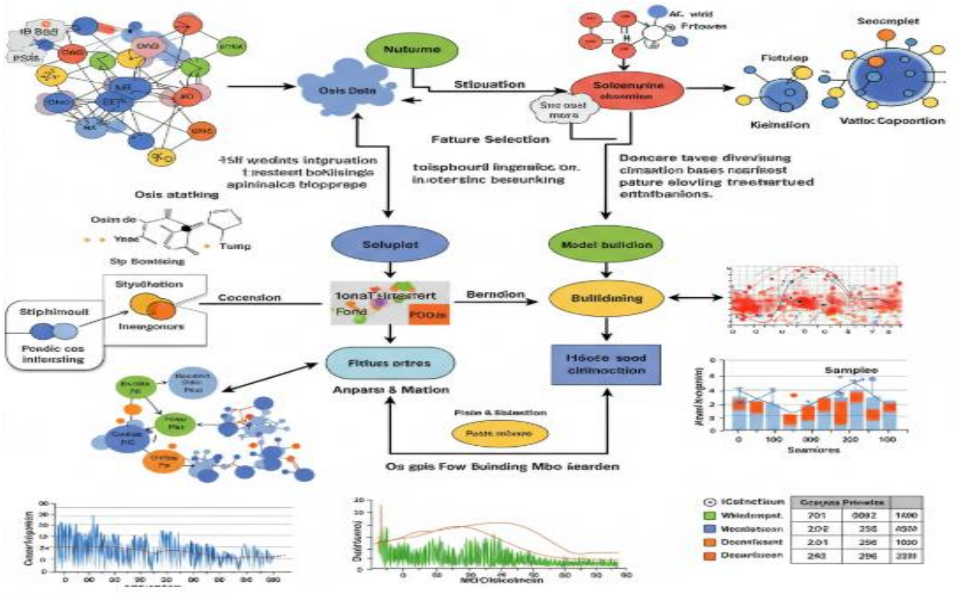

Fig 6.1: Patient Stratification Leveraging Omics Dataset and Clinical Biomarkers

6.1.1. Background and Significance

In today's data-driven society, many data analysis tasks are used in health informatics to predict risky events from health data. Chronic diseases have a continuing impact on the progress of the disease and are a risk for other diseases. This study proposes predictive health risk modelling and patient classification algorithm in patient medical records data with patient longitudinal data. Firstly, using a record number of patient data containing diagnosis, medication and operational data, a patient's longitudinal data format is proposed. Histogram gradient boosting model is proposed to model the risk of chronic disease as a time-dependent change in patient's treatment or medication treatment.

Proposed DRCO algorithm uses these returned models to determine the patient's risk of chronic diseases.

6.2. Literature Review

Prediction models are often developed to provide a continuous risk score to classify patients prior to the onset of the disease. However, clinical predictions can be more complex, such as prediction of binary events that occur at unknown follow-up times. This type of prediction in health has attracted researchers' attention to modeling the relationships between longitudinal measurements of predictor variables and the time-to-event outcome. A key problem in this context is how to extract useful information from the time course of longitudinal data.

There are recent medical studies and methodological advances on prediction of a health outcome of interest with a binary event using trajectories of predictor variables estimated from a linear mixed model so that unbalanced repeated measurements are up to a patient baseline time. The studies are mainly focused on development in two different broad areas. One area is the methods to assess morbidity over time in predicting health outcomes, using multiple types of repeated measurements which are formed prior to a study time. It has also developed methods to capture richness in longitudinal comorbidity measurements and their within-subject correlations in the framework of a Cox model. An alternative modeling is that patient variability in the rate of longitudinal measurements is incorporated in a time-to-event model. It is shown that the current application of the kidney failure risk equation to advanced chronic kidney disease can result in especially high individual patient risk prediction errors. In addition, the methods have been developed to investigate the incremental value of repeatedly measured risk factor trajectories over a decade for the prediction of cardiovascular disease in middle-aged adults using a joint model and an elastic net regression algorithm. The results suggest that a dynamic health index constructed by the regression coefficients stacked across multiple predictor variables from the joint model had the highest prediction power for cardiovascular disease incidence. An alternative application is in the predictive Alzheimer's disease domain, where a logistic model has been developed to estimate the risk prediction of mild cognitive impairment at six-monthly intervals. The study involving a longitudinal evaluation suggests that such a model, in general, can significantly improve the accuracy of predicting incipient Alzheimer's disease compared to simple standard tests or other recently developed biomarker-based models. Methodologically, the work generalized the joint modeling of longitudinal data and a discrete-time survival outcome to incomplete binary events and accommodated it with a simple reparametrization involving conditional probabilities. A dual-gradient method has been also developed to avoid estimating a high-dimensional covariance matrix and

the work involves patient-specific prediction of a binary event at which the data are pooled. The work also embodies a novel approach to test the assumptions of a given modeling strategy by fitting a likelihood ratio test to a modified model. The test can be used in the joint model, as presented, or in more general applications involving a time-to-event outcome. Only modeling strategies appropriate for the underlying events of interest should be considered because the tests can be performed in several different ways.

In view that random effects in a linear mixed model are difficult to interpret biologically, generally can only represent an average effect over the population, and can potentially carry undesirable correlation with the model residuals that may be used in the outcome equation; methods are developed about the potential binary event that can be collapsed into a hypothetical threshold or critical value of the fixed or random effects for each patient estimated from the model-generated trajectories. Miscellaneous statistical tests can then be applied to these hypothetical event times, establishing patients at higher or lower risk for an individual event, rather than estimating patient-specific event times directly. Given that a linear mixed model is known to consistently estimate patient-specific trajectories, under a special condition where the effect size of a particular predictor evaluated at the average time point for each group is equal to its binary regression coefficient in the segregation model, a modified logistic model has been proposed to predict a binary health event using a function of the observed fixed and random effects at the average follow-up time output by the linear mixed model. The function compared these two sets of estimates of fixed or random effects for a continuous predictor with a binary health event.

6.2.1. Historical Perspectives on Predictive Health Modeling

As a person changes throughout their life in age, lifestyle, socioeconomic status, and underlying diseases, a dynamic patient profiling should be seen by the medical monitoring systems to offer treatments much adapted to the patient's current status and evolution. Treatments adequate to the characteristics of the patients have a double positive impact, both enhancing the effectiveness of the care pathways and increasing the patient’s experience during the process. Hence, a blood test can inform the expected efficacy of different cancer treatments, cytokinetics can inform the patient’s capacity to metabolize a certain drug, and life habits can impact the benefits from a treatment. In the light of a revolution in the personalization of medical procedures, it is fundamental to find groups of patients sharing similar characteristics and behaviors regarding a certain condition; people suffering from a disease need to be segmented in a number of groups for a better care delivery. Precision medicine has been defined as the future paradigm that goes beyond genetic sequencing or risk models: personalized predictions that can

take advantage of the patient's health behavior, past conditions, mental health, social determinants, and even patient preferences. These kinds of predictions are expected to revolutionize the healthcare system, giving rise to proactive care pathways specifically designed for each patient. Any medical act generates data, and consequently, there are so many different ways in which medical monitoring can inform the expected onset of different health problems. However, the current medical systems are unable to take advantage of such a large volume of generated data. Clinical healthcare generates about 40 Zb of data, however, 90% of this data has been created just in the last 2 years. Allied with this, the scientific literature about the healthcare field doubles every 5 years, while it is impossible for clinicians to be up to date of new discoveries.

6.3. Methodology

Population health continues to be a major focus for many affordable care organizations (ACOs) and healthcare providers delivering Medicaid services. The number one goal or objective for all ACOs should be identifying the highest risk patients and then improving their overall health and lifestyle. With a data driven approach and predictive modeling, the Enhanced Patient Health Profile Algorithm can help achieve this objective. This algorithm will allow all customers to see a more detailed view into their population's health and risk factors associated with hospital utilization probabilities. But, a frequency distribution funnel views data analysis of overall health risk of the members of a physician group which provides an insight to the physician or analyst as to what kind of health risk score his physician's patients are. In population health, physician groups with Medicaid Accountable Care Entities provider contracts are required to monitor and manage their member populations' health risk status continuously. The proposed health risk analytics algorithm provides an overall patient health risk assessment that is a crucial analysis approach for physician groups to (1) identify their members' overall health risk status (2) further confirm the chosen provider group's risk score averages root causes by identifying contributing risk factors and (3) risk stratify members for appropriate case or disease management program selection. Finally, there is validation of calculated health risk scores of patient members using individual health risk scores available from state health authorities.

6.3.1. Data Collection Techniques

A patient's electronic health record (EHR) is a comprehensive record of the patient's clinical history, providing crucial information for healthcare professionals treating the patient. An analysis of electronic health record (EHR) data from a cohort of 5103 patients at a multi-hospital healthcare system in the United States captured longitudinal data that

span a seven-year period from 2010 to 2016. The target patient cohort in this research received medical care at one of ten different hospitals, with initial visits at various times in 2010. The collection and computation of comprehensive variables such as medical history, demographics, and lab test results from EHR data were based on an active or historical patient interval spanning a collection window configured around these initial visits, comprising a maximum of 1000 days in the past and 365 days in the future. This variable collection procedure was carried out for all patient records and for a number of different patient intervals and prediction time points. 54 different variables were computed, comprising 7 categories: demographics (DEM), medical history (MPS), family history (FHS), lab results (LAB), physical exam results (PER), prescriptions (PRE), and prior encounters (ENC). The number of variables per patient category is equal or fewer than 60 to avoid cluttering the model.

6.3.2. Data Preprocessing and Cleaning

Raw EHRs stored in hospitals' data warehouses cannot readily be used for developing clinical prediction models but must first be extracted, analyzed, and subjected to a series

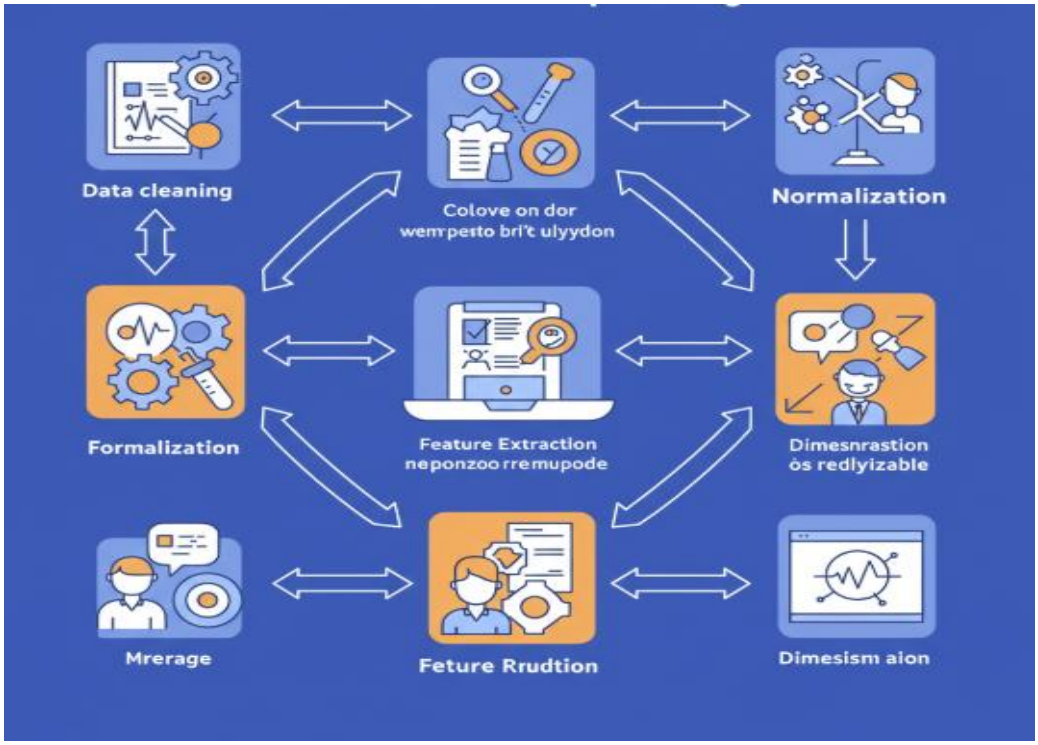

Fig 6.2: Medical data preprocessing

of preprocessing steps. These steps usually include validation (ensuring data accurately reflects reality), harmonization (establishing uniform representation), and transformation (bringing data into a form suitable for model development). Furthermore, it must be ensured that a sufficient number of data points are available and that clinically meaningful target labels can be extracted in the case of supervised learning. Nevertheless, assuming desired datasets are available, various opportunities arise when developing models in the context of hospitals through concrete examples, reflections, and open issues.

Discussion over choices made when pre-processing a large dissolved EHR dataset is closely associated with EHR model development research at hospital sites and is motivated by the perception of the multitude of obstacles brought about by the huge and often unruly nature of these data when planning to train machine learning models on them. Given a model derives predictive to all-cause morbidity on a large patient cohort following a general primary care dataset, the nature, manner, and outcomes of various data pre-processing steps are outlined. Key choices regarding the handling of index events, the method of calculating time-to-event sequences, and methods of dealing with missing data are discussed. Powerful tools built.

6.4. Predictive Modeling Techniques

Cardiovascular diseases are the biggest cause of mortality worldwide mainly due to strokes and myocardial infarctions. Hypertension, type 2 diabetes, cholesterol levels and obesity are typical risk factors of cardiovascular diseases. The capacity of the developed recipe to predict the onset of a cardiovascular disease is particularly useful for the health-care decision-makers because it may prevent severe personalized complications.

Risk prediction models of cardiovascular risk aid to characterize cardiovascular risk at the individual degree and hence could aid considerably on the clinical and public health decision processes. However, the development of personalized risk prediction systems for specific settings is highly involved and involves the estimation of large-scale demographic health data with both high quality and coverage. A novel data-driven model for six-month cardiovascular risk forecast based on longitudinal medical measurements from a large electronic health record dataset is presented and assessed on a test cohort. This method, with no need for additional non-temporal patient covariates, utilizes a novel approach to identify short-term risk profile changes as well as their interaction with the ensuing morbidity risk of the patient. The method designates patients with refined, time-varying personalized intervention strategies taking into account the particulars of the risk profile and temporal likeliness of the ensuing cardiac event.

6.4.1. Machine Learning Approaches

Risk prediction plays an essential role in clinical cardiology research. It enables the identification of at-risk patients for more intensive monitoring or therapy. Traditionally, most risk models have been based on regression models. While useful and robust, these statistical methods are limited in that they can only incorporate linear associations between predictors and the outcome, rely on a small number of predictors, and expect these predictors to be constant across observations. Machine-learning methods provide a means to address modeling challenges that are not well addressed by typical regression modeling approaches. Describe different challenges that arise in modeling patient data collected longitudinally, and then introduce different machine-learning approaches that can be taken to face these challenges. Finally, issues in the application of machine-learning methods are discussed including tuning parameters, loss functions, variable importance, and missing data, to serve as an introduction for those working on risk modeling to approach the field of machine learning. Risk modeling is critical in clinical cardiology research. Much effort has been expended to develop risk-calculators to inform clinical practice. Recently, several investigations have used machine-learning techniques, though they are not yet well recognized in clinical cardiology research. Risk prediction can serve an important role in the management and treatment of diseases. For example, it can be used in cardiology to identify individuals who may be at risk for developing cardiovascular disease. Moreover, risk prediction can be used in clinical practice to identify at-risk patients for more intensive monitoring or therapeutic intervention. In clinical cardiology research, the focus is typically on modeling risk to predict various adverse events, such as diagnostic events of a number of diseases, including acute myocardial infarction, hypertension, and heart failure. Most risk models are developed within the context of a Cox proportional hazards model, often with additional regularization to shrink coefficients. While useful and robust, a limitation of these statistical methods is that they can only model logistic associations between fixed predictors and the outcome. A method to handle temporal changes in predictors is to fit separate models at each time point; however, this approach requires a separate model to be fit for each time step.

6.4.2. Statistical Methods

Three approaches for predicting risk in longitudinal data were selected. One defines longitudinal data where each subject has vector-valued observations at equally-spaced time points, each with the same set of scalar components. This framework is known as the classical longitudinal data setting, panel data, or multilevel data. The outcome was when (time to a fixed event) a disease endpoint happened, what fraction of the area under the curve (with respect to some time of interest) had been covered by then. surv(q, time)

equals the probability of the failure variable q being less than or equal to time. In the four papers considered: First, the focus was on prediction rather than parameter estimation. The existing literature on this topic largely concerns parameter estimation-typically estimation of hazard ratios for a fixed set of regression coefficients. The predictNDE() function assumed a wide variety of loss functions could be used and instead loss was calculated using its negative log-transform. In this case, the loss was calculated in a way that would arise for either type of loss when fitting accelerations.

6.5. Risk Stratification Frameworks

To improve the predictive health and determine the increased protection probabilities (IPP), estimations of the disease occurrence are proposed within the framework of the unique relationship between the lifestyle risk factors and complex diseases on a person-by-person basis. The approach differs from the epidemiological studies as only an individual is considered and the objective is to estimate the awareness probabilities on a particular person's healthy bio-behaviours thereby increasing the chances of prevention of future diseases. A risk predictive framework (RPF) manages the risks over the lifetime ever since birth because they are derived from external conditions, which are usually changeable and easier to control.

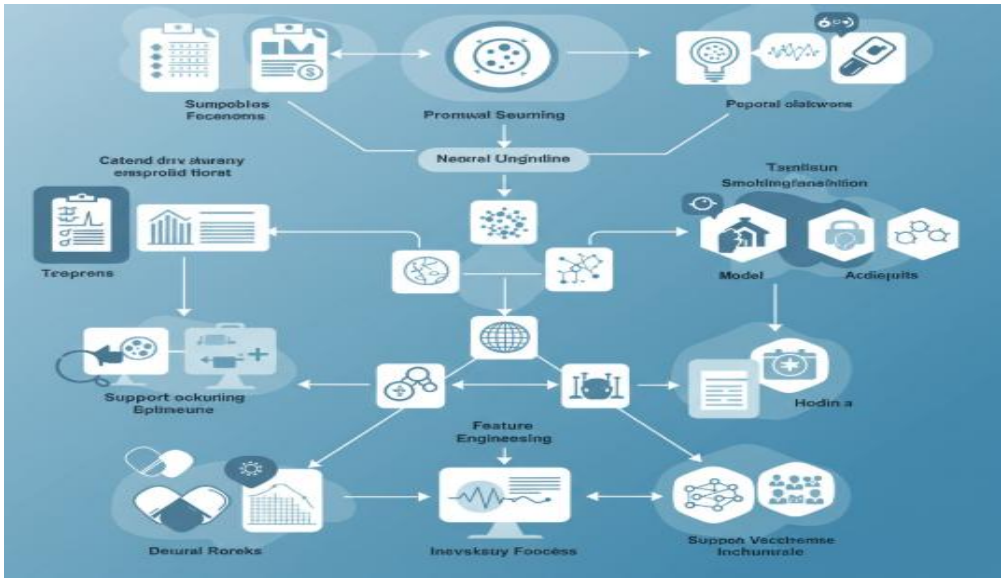

Fig : Health Risk Assessment Using Machine Learning

Retrieval of the longitudinal patient data is investigated, which is taken as the first step towards a better understanding of this type of temporal data in relation to complex disease occurrences. In this context, complex diseases are the common diseases and

prohibitive cost, due to the expensive and time-consuming experiments, to undertake and complete the validation experiments essentially reflects the current scenario described by the RoC. Therefore, a privacy-preserving biomarker design is pursued, which is essential to reduce the cost to gather the patient data before designing and validating diagnostic assays. Such assays estimate the risk of the occurrence of a disease in the future based on the patient data and can be used to validate hypotheses on novel disease risk factors or to identify healthy bio-behaviours that can prevent or decrease the risks of future diseases. The patient data that describe patients or healthy individuals are collected intensively over time. From multiple sources such as electronic healthcare records, doctor notes, laboratory results, and health insurance.

6.5.1. Defining Risk Stratification

This study used cost and utilization analysis of three years of claims data, statistical clustering, contrast mining, and logistic regression to identify patients within a managed population at risk for higher healthcare costs. Priority was given to clinically detailed detectable traits that could suit medium- and low-tech clinical practices as well as machine learning algorithms. Statistical analyses of 566,896 non-cancer patient-years of hospital outpatient cost were used to compare leading statistical clustering strategies for patient cost, rates of key care events, and gaps between the two. Statistical contrast mining analysis was extended to find detectable, low-variability non-concept-coded traits associated with higher hospital outpatient cost or care event rate. No clustering strategy found resembled a pyramidal tiers of steadily more costly patients as expected. In any identified tier, a minority of high-cost patients were found with long-range cost patterns. However, the continuity of rising mean cost and care event rate between fewer, more expensive tiers was well-preserved. Across methods, the two highest risk stratification tiers accounted for 42.5% of 27,867 prospective hospital outpatient charges. Some clustering strategies generated significant cost or utilization cost gaps at company-suggested membership thresholds proposed for likely implementation.

Patients in the most expensive (sub-)cluster of the most expensive (sub-)tier were on average 1.89 times as costly as all other ‘high-cost’ patients. Remaining (i.e. non high-cost) patients in costly (sub-)tiers who did not also belong to the most expensive cluster were no costlier than the overall population. The single hospital outpatient contrast having the high care event rate ratio threshold also combined with logistic regression predicted hospital outpatient resource utilization of the corresponding segment with at least 83.7% accuracy. Using logistic regression of individual company-developed patient typology, claimed patient type, cost, and care resource utilization predicted the most expensive 5% of managed population patients with 84.2% accuracy.

6.6. Conclusion

Driven by the need to restructure America's healthcare system, organizations have centralized resources and streamlined services. New products and services are being created every day for different health-related issues. Combining a continuous rise in medical healthcare costs with the demand of new products and services, society has turned to chronic condition management and risk stratification in order to analyze the risk of catastrophic levels of healthcare costs in different demographic groups. Chronic condition healthcare could significantly lower the overall expenditure on healthcare. Such services have been used in doctors' offices to determine the level of benefits provided to patients, and as a payment mechanism to calculate the risk of certain population groups in various areas. There is cost-effective access to certain specialized preventive treatments and health services by providing chronic health management services to the population. It allows proactively identifying high-risk patients in order to better plan resources and avoid costly health complications. On the other hand, problems arise when applying chronic risk and disease management to some service providers, as services do not necessarily accurately assess the risk level of a member or members to perform targeted actions.

Large healthcare providers in Illinois are using a company that provides risk scores for adults for a certain algorithm using administrative data. However, some physician groups choose members only by the level of this risk score. If we select members of the same age for a physician group and their calculated risk score comes out to be the same or members have minimal difference, we know that a particular physician is accepting members based on their good health conditions, called risk selection. An algorithm is proposed that uses structured query language techniques and predictive analysis to define the health care needs, best insurance practices, and the insurance production of a state-managed insurance agency. The produced model predicts health risk scores for each adult individual member in relation to essential health benefits insurance. To show that the proposed method is efficient and accurate, it was validated using a number of algorithms for multi-linear regression against available individual risk scores from the State of Illinois. For the second validation step, the original base model was run. The results show that the proposed chronic risk and disease management algorithm calculates health risk of state agency members. It is shown that the proposed algorithm is accurate in calculating a health score of an individual. The results confirm that the health score can be a vital factor in identifying members for the purposes of disease management. With the system proposed here, any healthcare provider or physician group can evaluate health risk status. By using the agency billing code, the billing code with the greatest difference in the calculated risk is retained. The final dataset with calculated health scores is requested for these codes. Examples that demonstrate the sensitivity,

specificity, accuracy, and positive and negative likelihood values of a dataset are further provided.

6.6.1. Emerging Technologies

This is the first attempt at bridging the gap between increasingly complex machine learning-based algorithms – such as random forests, gradient boosting, SVMs, etc. – and the relatively low-complexity, practical and intuitive understanding possessed in the realm of healthcare practice. Recent Random Forest results on a standard healthcare dataset, trained features of 1000 trees in ~20 min, and inference can be performed as quickly. A fact implies that off-the-shelf, powerful modeling techniques are within the grasp of the broader medical community today, not some sophisticated, next-generation hybrid deep network architecture. Moreover, public codebases granting free access to these sorts of machines have recently proliferated. We trust that this work will facilitate increased adoption of this technology in the healthcare delivery chain, further empowering both the medical community and patients to better understand the very nature of their risk and how to meaningfully address it.

Recent technological breakthroughs have enabled continuous and temporally rich collection of diverse data, which offer an opportunity for profoundly changing how care is delivered. The availability of sophisticated models has expanded our ability to convert this data into actionable clinical insights, resulting in the promise of a more accurate, timely, and efficient health care service. Some of the most exciting advancements have come from the application of modern deep learning techniques to a wide variety of medical data, from electronic health records (EHRs) to medical imaging. As such, others have seen many successes come from the application of less complex linear models and decision trees, which can provide more intuitive understanding of the underlying model dynamics and output. The effectiveness of these approaches is exacerbated by: 1) such models being easier to trust by healthcare professionals or regulators wary of “black box” AI systems, and 2) the ease in deploying them in low-resource environments. There is also the prevailing challenge of data scarcity in many medical applications, and the prowess of more complex models to overfit. This work presents a new human-computer interaction tool that brings healthcare professionals.

References

Ganguli, R. (2024). Machine learning algorithms detect high-risk pregnancies early. The Pitch: Patient Safety's Next Generation.
Ganguli, R. (2024). Machine learning algorithms detect high-risk pregnancies early. The Pitch: Patient Safety's Next Generation.

Institute of Medicine. (2024). Medical errors and patient safety. Journal of the American Medical Association.
Toronto General Hospital. (2024). AI command center improves patient safety. Journal of Hospital Administration.