# Clinical Feature Engineering and Outcome Prediction in Electronic Health Records: Machine Learning Approaches to Enhanced Predictive Analytics in Healthcare

*Dr. Olga Petrova, Professor of Applied Mathematics, National Research University Higher School of Economics (HSE), Russia*

## 1. Introduction to Predictive Analytics in Healthcare

The article is devoted to the application of models of machine learning and artificial intelligence for data analysis of medical information and the definition of the state of the human body. Many questions that are directly or indirectly connected with healthcare deserve special attention. These include the problems of diagnostics of diseases and anomalies, methods for predicting the course of complex diseases, as well as tasks aimed at determining optimal treatment strategies. This work addresses the issue of predicting the timing of the outbreak of a pandemic of infectious diseases and, in general, the reasons for the recorded changes in the rates of requests from patients and users in services directly related to solving health-related problems. This is a fundamentally important issue that is particularly acute in the context of the deterioration of the epidemiological situation in the world, new, previously unknown, or underestimated infectious diseases.

The solution to this task requires a detailed analysis of data arising directly during this period and accumulated over a long time. The development of short-, medium-, and long-term forecasts for the distribution of the situation, being known in advance, will allow for anticipating a sharp increase in the frequency of such requests in the healthcare system and implementing a number of recommended measures to develop preventive strategies to prevent the spread of infectious pathogens. The task considered in this article is to predict patient visits to the dispensary with complaints about the presence of various diseases at several intervals of time within several months, both the required actions of patients and doctors and the optimality of further maintenance of such a dispensary on the territory can be controlled in the future.

### 1.1. Definition and Importance

Advances in machine learning techniques, as well as increasing the availability of healthcare data, are enabling the development of AI models that can predict events of critical importance in healthcare, such as outbreaks of infectious diseases or patients' needs. This paper provides a discussion of methodological frameworks involved in two key healthcare predictive analytics areas: forecasting infectious disease cases and forecasting patient-level or population-level need for healthcare resources. These discussions aim to contribute to fulfilling the need for increased awareness of healthcare application-specific predictive analytics methodological details for the benefit of healthcare professionals and predictive model developers alike.

Predictive analytics may be employed to enhance healthcare by having AI models help forecast events of critical importance, enabling their controlled and timely management. In particular, predictive analytics can play a key role in two healthcare-related areas: public health and patient care at healthcare facilities. Predictive models can provide knowledge that helps public health professionals deal more effectively with infectious diseases, as they allow for the advocacy of more effective interventions in pandemic planning and response and provide information around which advanced preparedness plans can be based. At healthcare facilities, predictive models can assist in the design of more appropriate and timely patient care interventions that match patient demands with resource management efforts and manage patient flow and hospital surge capacity during critical events, which is one of the continuous challenges that emergency departments need to address.

### 1.2. Applications in Disease Outbreaks and Patient Needs Forecasting

Another application in disease outbreaks and patient needs forecasting. Early warnings and predictions by these ML models can support authorities in allocating resources and making policies, with significant improvement in the control of the disease, patient care, and health system performance. Model examples include forecasting of Ebola hemorrhagic fever, epidemics for two common surgical complications, influenza-like illness, hospital readmissions, mortality among patients with chronic obstructive pulmonary disease, healthcare-associated infections, requiring interventions for attending to patients with trauma, staffing needs, and performance on exam questions, mortality in intensive care units, the likelihood of requiring immense medical resources

due to deterioration of a patient's condition, patient admission flow, patient screening rates, adverse events, risk of circulatory system diseases, patient appointments, and pneumonia during the flu season.

Unlike the complexity of EHR data used by most time-wasting ML models, those considered in this study only require easy-to-access data, which include historical pneumonia activity volume within the USA and a vector of external factors. With a large domain set suitably abstracted from complex EHR data and a long enough time, the method will elegantly discover the most predictive domain that should include the decision support objective being tested. Iterative recursive feature elimination analysis generates time-domain metrics that indicate the lag associated with the highest predictive capability. The sequential predictors are trained on decoupled domain features that were pre-analytically transformed and classified by the event qualifiers of increasing severity, as determined by a tiered weighted relative toxicity within the domain classifiers. Furthermore, the extracted patient data and transformed external variables are used to inform the time-domain preclassifier, which is a script that applies the domain qualifier and formulates the resultant training set.

**2. Fundamentals of Machine Learning in Healthcare**

The application of a learning model that accurately predicts future patients' requirements could provide benefits for both patients and providers. From the point of view of the care receivers, the model contributes to maintaining the highest level of treatment and prevention, instead of cure only. Care providers can minimize treatment expenses and prevent patient discomfort by forecasting individual patient needs accurately. It is important to recognize which model would provide the most accurate forecasting for a particular patient before investing in the acquisition and implementation efforts. This paper addresses how three different data learning models performed under different circumstances in predicting upcoming patient needs. These prognostic learning models have been particularly popular in nursing because they assist in constructing projections on different health outcomes.

Uncertainty in the prognosis, on the other hand, puts a strain on how predictive models can be put to practical use. The emotional and financial challenges involved in confronting an illness have transformed health care from treating acute diseases to forecasting and preventing chronic conditions earlier. In the last decade, steady

developments of new study methodologies such as long-term studies, sickness management models, and learning techniques for predictive analytics have emerged, facilitating comprehensive analysis to preempt health problems. As part of patient care and services, healthcare professionals are now able to use extra time and resources in outreach services and workshops to serve those communities and individuals at greatest risk of disease outbreaks or negative health events.

### 2.1. Supervised, Unsupervised, and Reinforcement Learning

Machine learning is a method or solution that enables a system to learn using data. ML can be classified into three categories: supervised learning, unsupervised learning, and reinforcement learning. Supervised learning consists of learning tasks that map input data to output data. In the mapping, there may exist a function that the mapping model is trying to adjust. During the training process, the model tries to develop the function that maps the training data to the output. The difference between the output of the model and the training data will be compared by a cost function, which will help to minimize the difference between the model output and the training data. The cost function will assist the model in adjusting the function parameters during the learning process. The process of using data to enable the system to learn is known as training. To evaluate the model's performance, a set of test data can be used to assess and produce metrics that show the quality of the model.

On the other hand, unsupervised learning is a class of ML research related to learning from data that involves only input data and no corresponding output data. Instead of allowing the model to learn to map the input data to an output, the model tries to identify the groups associated with the input data. These groups can be seen as clusters in the space of input vectors, where the input data are grouped together based on their similarity in the input space. The simple case of clustering can be used to enhance the model's capabilities to learn from data and represent the input data in an unsupervised way. Data are loaded into the model, and the result will provide clusters based on the similarity. The system is not informed of what the result will be, only that it should determine how the input data fit into one of the output clusters.

### 2.2. Feature Engineering and Selection

Feature engineering and selection are at the core of performing predictive analytics effectively. Feature engineering requires selecting a small and useful set of features,

including demographic and clinical information. This information is frequently observational or real-time and can be collected through regular healthcare and public health systems. The healthcare domain is data-rich, but the high dimension, low sample size, and use of diverse data types and formats make feature acquisition, preparation, and construction all challenging, problem-based tasks. The feature-selection process is useful for selecting only those features of interest by filtering out irrelevant and redundant features. The three primary groups of methods, whose utilization depends on the data's structure, are the wrapper, embedded, and filter techniques. The first two options especially benefit from having a hyperparameter for the base ML model, aimed at improving the performance of the model. Popular methods for feature engineering and selection using ML for healthcare applications have utilized features derived from general statistics, predictive or descriptive models' non-clinical features, clinical-level generalization, transfer learning on clinical data, and multi-source data fusion.

Candidate healthcare features can be derived from analytics from various domains, including epidemiology and biology, clinical medicine and systems, public and environmental health, social medicine sciences, and global and individual views. Clinical systems' non-clinical features help describe the context around the patient's healthcare experience or help understand the demand for patient services. There are typically more non-clinical features than clinical ones, and these are aggregated at different levels. Clinical levels of generalization aggregate clinically relevant features that can be reduced by applying different methods. These direct strategies, including using standard machine learning techniques or more recent deep learning algorithms, directly modify the original feature points. An example of a non-direct strategy is transfer learning, which requires little contextual information. For example, a feature vector is trained using a different dataset and applied directly by making slight modifications to the model's architecture. Finally, multi-source data fusions, including well-defined procedures identified in the data preprocessing phase, are employed during data aggregation from different feature sources. These tasks typically seek data interdependencies based on semantics, correlation, and other exploratory features.

### 3. AI Models for Disease Outbreak Prediction

Following the epidemics in recent times, such as SARS, MERS, and, above all, the most novel Zika and Ebola outbreaks, there is no doubt that accurate and early prediction of

epidemic waves or viral latency periods is among the most important and challenging systemic problems that scientists and governments have to deal with. In general, the prevention and treatment of emerging diseases is a goal that should be written in capital letters in the agenda of any responsible government. Thus, formally oriented, our study here addresses this extremely important element. The final model of our proposal, denominated Reservoir Model, aims to characterize the dynamics of disease outbreaks through a set of variables that relate the vulnerable population, movement and contact rate between the susceptible and exposed populations, and demographic vulnerability factors. Our methodology, built on information and knowledge from cutting-edge machine learning and simulation models suitable for these purposes, promoted models that can be considered a group of the best algorithms.

MLP, being the backbone of the Reservoir Model, is technically robust and is suitable for a variety of problems, including classification and regression, due to the simplicity of its structure and the ease with which the weights of certain attributes of importance in curve fitting and decision making can be found. OLS, in turn, is the most intuitive and robust ML model available today for solving the problem of sum of squares and parameter estimation in the ordinary least squares context. Although our statistical approach is not entirely based on machine learning algorithms, we note that the proposed model, able to estimate a series of factors related to a disease outbreak, was chosen in favor of its flexibility and the methodological solutions that arise. In this sense, the steps we follow and the final model obtained provide assistance to its decision maker when estimating theoretical questions related to events in the dynamics of epidemic waves in a more general sense.

### 3.1. Time Series Forecasting

Time series forecasting is one of the areas in machine learning and neural networks that requires time series data to model and forecast disease outbreaks and other relevant healthcare scenarios that require time-wise prediction. Both statistical and an emerging AI class of models, like recurrent neural network models and convolutional neural network models, provide a variety of options for effective and efficient modeling of the time series patterns. These models have developed into refined versions of basic feedforward neural network models. The modeling of the dynamic patterns in time series data with a recurrent system requires proper modeling of the recurrent system's

internal architecture, since the recurrent connections are unrolled back through time in recurrent system models. The forward-backward process and laying the results of the forward movement a few time steps closer to present time steps is an error-based global optimization problem called learning of an RNN model. Unlike the case of feedforward neural networks, this learning problem of RNN models is not optimization of scalar weight parameters, but rather optimization of feedback parameters in the recurrent layers. Because this feedback parameter optimization is needed to handle both short-range and long-range dependencies present in time series data, the predictive accuracy of the model decreases with longer time series ranges. A type of recurrent neural network model that has shown effective learning of unique dynamic behaviors in time series data comes in the form of Long Short Term Memory models, which have inherited recurrent feedback from a previously unrolled state so many times that the vanishing gradient is reduced. Because of its promising performance for sequence modeling, the LSTM model has become a very popular choice for time series forecasting and sequence tagging problems today. A similar construction to the LSTM model with an extra self-gating structure called reset and update gates in place of or in addition to the input gate took off in parallel, known as the GRU model.

### 3.2. Classification Models

Classification models, also known as classifiers, are a subset of supervised learning models, and they primarily focus on predicting the category of a given variable. Gini impurity, entropy, chi-square, information gain, and support vector machines are some of the most commonly used supervised learning models. Logistic regression and decision tree models are other popular classification models used for disease prediction in the healthcare domain. All these models have been widely applied in disease prediction with high accuracy. Among these classification techniques, the support vector machine approach has the upper hand in cases of a higher degree of pathology separation.

The support vector machine algorithm has shown very good performance, with high predictability accuracy when working with a smaller training set. This makes the support vector machine approach very suitable for health predictions in medical studies. A study into classification and regression tree-based random forest decision-making for determination of rapid test sensitivity on dengue fever subjects also used the support

vector machine technique. The results report distinct and well-determined separation of the samples from dengue hemorrhagic fever and healthy samples. It is particularly suggested that support vector machine be used for disease differentiation studies. High accuracy in the diagnosis of prostate cancer has been achieved using the support vector machine machine learning technique. Other users of the support vector machine technique have obtained very good results in other pathologies such as cancer, T2DM, coronary, kidney, pulmonary, and oral disorders.

**4. AI Models for Patient Needs Forecasting**

Enhancing the overall patient outcomes is one of the primary objectives of healthcare organizations. With the increasing focus on value-based care, the patient health monitoring model moves from reactive treatment to proactive healthcare. This has resulted in healthcare management technology undergoing a fundamental shift over the last decade, harnessing smart health data to improve health outcomes. As the complexity of care needs continues to grow, healthcare systems are constantly looking for ways to help patients be informed, in control, and proactive in managing their health. To assist healthcare organizations and caregivers in proactively anticipating patient needs, an AI model is proposed whose immediate scope is predicting the next unavoidable emergency hospital admission or an emergency room visit, so that caregivers can prevent the visit.

Numerous studies demonstrate and validate the fact that hospital readmission risk can be forecasted quite accurately from structured EHR data. However, with the increasing shift in healthcare delivery from being facility-centric to patient-centric, the newly developed patient-centered models are becoming more important. Additionally, along with providing a patient-centered healthcare delivery model, it is equally important to provide a quick response and decisional support for personalized care, especially in the comfort of home. The AI-based patient needs forecasting model aims to provide a healthcare organization with all the required information to make informed decisions.

**4.1. Recommendation Systems**

The prior sections in Part II, Machine Learning Models, were aimed at supervised learning: we had a labeled data set. We also described a number of unsupervised learning algorithms. In this section, we describe a class of algorithms called collective prediction or recommendation systems, which exploit annotations. These algorithms

were designed specifically for scenarios where a fair amount of data is available, but labels might be missing. To us, such scenarios are fairly common in healthcare; for example, we often know where there are free beds in the hospital, but what we do not know is which patient should take up that bed, as any one of those waiting for beds might take a sudden turn for the worse, becoming our top medical priority at the time. People find this problem intuitively easy; it is difficult for a machine. The alert monitors use a similar algorithm to wake you early if your flight is likely to be delayed. Personal devices also provide gentle recommendations for extended activity. One hospital uses crowd-sourced wisdom about dos and don'ts in emergency situations routed to local ERs. Finally, most people use film or show recommendations daily with great satisfaction from good examples.

One simple approach to recommendation is the low-rank factorization that outlines, which is similar in many ways to basic collaborative filtering methods. These methods look for the rows and columns in a data matrix that are missing many entries and try to fill in the gaps using linear algebraic methods. Measuring performance here is a delicate issue, as we have known ratings for many matrices, and new methods may not always point to the highest ranked record.

**4.2. Natural Language Processing**

With the proliferation and continued digitization of healthcare data due to advancements in electronic health records and other health information systems, there is increasing interest in leveraging AI technologies to extract knowledge from this data. In particular, models that can effectively process the large volume of free text data are becoming increasingly popular. This is evident from the significant number of clinical NLP models recently reported, which have been applied in the processing and analysis of clinical and biomedical records. These clinical NLP models inherit textual data challenges from general NLP, yet are designed for specialized functions that respond to unique healthcare use cases, such as entity recognition, concept normalization, assertion detection, named entity recognition, entity concept identification, and consequently support a range of analytic tasks, such as classification, clustering, and prediction.

Distinct from more traditional healthcare data analysis that relies heavily on structured and strictly defined clinical data, such as laboratory tests, physician orders, and structured notes, clinical NLP offers the capacity to incorporate, in an automated way,

unstructured and free-text data into predictive models, improving model performance. Such clinical NLP makes it possible to convert difficult-to-extract free text data into an easy-to-analyze tabular structure, enabled by neural models' potential to represent and capture complex n-gram patterns and common linguistics in clinical language. Importantly, clinical NLP has the ability to seamlessly and comprehensively complement all forms of structured data with free-text data from electronic health records, facilitating enhanced predictive modeling and, consequently, better resolution of patient- and population-level health needs.

**5. Challenges and Ethical Considerations in Implementing AI in Healthcare**

The application of AI models in healthcare is not without its challenges. There can be accuracy and quality issues related to the training, maintenance, and validation of AI models. Training predictive models effectively on a large-scale data set can present a significant challenge to processing power and lengthy analysis time, and incomplete, biased, and inaccurate data limit the efficacy of predictive models. Data privacy, data security, and trust are considered major barriers when working with patients and enabling the clinical acceptability and utility of predictive models. There are concerns that data mandated by the rules, laws, and ethics can be misused, leaked, or compromised. These concerns can originate from previous incidents where patient data have been compromised as a result of security problems, data sharing externalities, and cloud vulnerabilities.

Patient health data are among the most sensitive forms of information, and patients seek to protect it as best they can. Data sharing between institutions is vital if predictive models are trained on a full range of patient concentration patterns. However, institutions are prone to sharing data due to concerns regarding regulatory constraints and the institutional risks inherent in sharing information. In recognizing the potential for patients to become more protective of their information, data are inherently stigmatized and often kept as private as possible. As healthcare leaders know, failing to encourage patients to offer information prevents the accuracy and effectiveness of care from increasing. This practical distribution of information burdens sacrifices the ability to improve the quality of services for the desirability of preventing patient acceptability policies from collecting patient information in the first place. Data quality is dependent on the collection of a complete dataset, including all relevant information, to be analyzed

as patient data is often collected piecemeal and from patients with various degrees of health. The personal beliefs of the patient about who may see the material can influence the amount of data released, the depth of the detail published, and the accuracy of the valid data. Data sharing policies remain a sensitive challenge as legislative incentives are increasingly in place to demand more open data.

### 5.1. Data Privacy and Security

Data in the healthcare domain is not just big; it is vital and of utmost privacy and has serious implications if misused or breached. Patient, administrative, and other data come together to fuel improved decisions on recent disease outbreaks such as the latest Ebola virus disease epidemic, the global threats of human influenza and avian influenza, and even other health threats such as autoimmune disease, obesity, diabetes, and hypertension. The United States healthcare and public health sector has moved to cloud-based EMR systems. Parents' delegation of health records for educational purposes in public and private schools for children's school entrance health service requirements frequently underpin this move. With public fear of the leaks on sensitive and potentially embarrassing personal information supplied to these institutions, the data privacies of these digital repositories need proactive protection through predictive analytics for legitimate data use.

Information privacy addresses how personal data can be applied by data collectors who provide the means for conducting this mining. It extends into healthcare information technology, electronic medical records, and health infrastructure platforms. To reduce systemic risks, effective cybersecurity protocols can be formulated and implemented using modern AI methods in predictive health threat modeling. Developed algorithms can be applied to cyber surveillance of healthcare incidents in the interdependent systems of a U.S. city, one of the most complex and dynamic environments that can be modeled, to reduce systemic risks with electronic health record systems and other interdependencies. Machine learning-based diagnostic applications can save lives. The promise of Portable-Assisted Lifestyles can be realized through advanced sensor technology that performs workflow mining—monitoring and collating commonly occurring physical and social activities within established regimen and procedurally designed models—can be analyzed to infer the health status of recipients, providing

ease of independent living in familiar surroundings, all the while protecting user privacy through activity-based model abstraction.

### 5.2. Bias and Fairness

Before deploying a prediction model, one might also want to consider examining biases or other moral and ethical concerns about the model. For health care prediction models, there are many groups and attributes to consider. Common attributes include, for example, gender, race, class, or disability, which might have an association with the outcome variable and could result in errors in prediction. Sometimes, the differences in probabilities of the predicted classes reflect actual societal bias, such as the lack of equality of education or health care access. In our application, fairness is not only an issue with underrepresented groups, but a lack of fairness could have substantial personal, financial, and life-threatening consequences.

Algorithms, which take raw data as input, are often influenced by the biased underlying training data and therefore produce biased predictions. When predictions from biased data are used to make decisions, they are likely to advantage specific groups of people or industries while being unfair toward others. In disparate treatment, specific groups are underprivileged and, in the worst case, are totally rejected. In our case, with our underfitted machine learning model for forecasting the utilization of specialized care, the consequences wouldn't be that dire, but they would include financial losses, long waiting times to adopt designed benefit programs, trust in expensive health care plans, and administrative burdens. For both our other use cases, the unfair treatment of underprivileged people in a city would not only annoy the citizens but increase their frustration and reduce local cooperation, as not all underrepresented groups would be able to enjoy all the benefits of the smart city.

## 6. Future Direction

Developing AI models to accurately predict disease outbreaks before they occur and effectively identify patients likely to face disease complications should continue as a priority area for future research. This will require developing additional disease-specific datasets. Sharing such datasets will be beneficial to all, as small numbers of researchers typically struggle to accrue the number of cases necessary to develop robust prediction models. Moreover, models will likely generalize better and be more robust if they are trained on data from multiple reviewers and settings. Transferability increases value.

The rapid pace of EHR digitization and the scale of big data present enormous opportunities but also challenges. As with all AI models, the accuracy of those developed for forecasting disease outbreaks and patient needs is conditioned by the quality of data inputs. We will need to invest additional time and resources to clean EHR data and understand the strengths and limitations of digital records and machine learning. AI and human cooperation enhance predictive analytics. We can often achieve better performance and accuracy by combining AI approaches with those based on human knowledge and understanding. AI models help by processing big data, thereby detecting associations not previously recognized. What is new is the scale of the relationships that can be discovered. Physicians contribute by providing the intuition and experience to understand if these associations are plausible and may have the potential to positively impact the health of their patients. The best decisions for patient care will often draw upon both sources of information. Interdisciplinary research is key. Successfully applying AI to challenging research problems, such as forecasting disease outbreaks, is often more effective when accomplished by research teams made up of individuals from diverse scientific backgrounds. Researchers specializing in areas like statistics and machine learning are able to effectively implement novel algorithmic and computational strategies. They can often do more harm than good when designing and executing these strategies without the help of healthcare experts. Success depends on the ability of data scientists and infectious disease experts to share knowledge and collaborate in both the design and interpretation of models.

**7. Conclusion**

In this chapter, we described a showcase of predictive analytics in healthcare by taking the example of two useful recurring applications: the modeling of infectious disease outbreaks at the community level and the modeling of noncommunicable disease events or needs at the individual level for patients with chronic conditions. We provided a brief overview of the healthcare domain and described multiple healthcare stakeholders, followed by characterizing challenges and gaps in healthcare predictive analytics. We covered the concept of features to be used as predictors, social and environmental determinants, and predictors of healthcare utilization. We also discussed a few candidate machine learning predictive models and their associated evaluation metrics. Subsequently, an overview of representing the models for public health and clinical needs in healthcare was also provided. Finally, we discussed a realized impact of

machine learning in healthcare, highlighting the deployment considerations, ethical considerations, and specialist user considerations in healthcare predictive analytics. We then wrap up our discussions with a summary of the future direction of the work.

In this chapter, we focused on two special classes of predictive modeling in healthcare (i.e., infectious disease oral function modeling and chronic disease event and function modeling). However, many other more traditional predictive models have also been developed in healthcare, such as mortality and in-hospital readmission prediction, readmission for specific emergency department patients, early hospital readmission prediction, elective hospital readmission prediction, risk adjustment for measures of healthcare network quality, predicting future healthcare cost, and length of hospital stay prediction. In the future, we plan to provide additional illustrations of these more general non-special class predictive healthcare models, as well as to provide a summary of an AI model catalogue for healthcare that can be used by healthcare personnel from all fields interested in using AI support for patient care or health policy decisions.