Longitudinal Biomarker Trajectories and Transition State Modelling: Machine Learning Approaches to Disease Progression Prediction and Clinical Staging

Dr. Daniela Ramos, Associate Professor of Computer Science, University of São Paulo, Brazil

1. Introduction to Disease Progression Prediction

Introduction Disease progression prediction is a crucial part of the arsenal of tools available to healthcare providers. By annotating the future states of an individual according to a disease's natural history, diagnosis and treatment decisions become much more efficient and effective. Biologically, diseases are typically driven by a complex interplay of genetic, lifestyle, and environmental factors that interact according to the disease context to determine disease initiation and progression. For instance, in cancer, a multistep carcinogenesis model is often invoked in which accumulated genetic and epigenetic changes progressively cause cells to become invasive relative to their normal neighbors and to acquire the complex phenotype of "malignancy." Another prime driver of disease progression is the "social determinants" of the disease that underlie the progressive divergence of health states. From a clinical standpoint, accurate forecasting is necessary in order to inform diagnosis, predict treatment responses, and guide other clinical decisions. Yet disease progression is a complex probabilistic process driven by a multitude of interrelated factors and thus faces large prediction uncertainty. For chronic diseases, the ability to precisely determine how the disease evolves over time would have major consequences for healthcare systems and the affected individuals. The ability to anticipate disease progression far into the future presents new treatment opportunities at earlier stages of the treatment trajectory. Furthermore, states can be predicted in the emerging science of personalized medicine, thus enhancing the outcomes for individuals affected by the disease. Traditionally, various statistical methods and mechanistic models have been used to tackle length-scale forecasting of disease progression. However, these commonly utilized tools suffer from several difficulties, including the need for patient-specific prior details, frequent data

This work is licensed under CC BY-NC-SA 4.0.

measurement, and the relatively poor results of vital outcomes. It has now become possible to forecast changes over the course of a disease trajectory by using machine learning.

1.1. Significance of Predicting Disease Progression

Predicting disease progression is critical to individualize care. Clinicians count on subtle or explicit predictions. In the former scenario, predictions could trigger confirmatory diagnostic procedures, are used to assess the risk of developing a specific disease by carriers of risk alleles, and can be used to identify the duration of subclinical and atypical lung carcinomas in a follow-up program. In the medical care of established diseases, it is used to guide clinical decisions and/or justify healthcare expenditures. Accurate disease progression predictions may help authorities distribute resources according to the estimated future healthcare needs. Again, individuals can use that information to better manage and cope with their illness: caregivers, for example, can use it to enhance their knowledge, and the patient can use it to adhere to therapeutic strategies, follow a diet, or engage in physical activity. Thus, predictions influence many stakeholders and can be used as a unifying paradigm to foster patient-tailored interventions and prompt engagement and education.

Health trajectories are often characterized by alternating disease images. The ability to forecast which goes further may be of great value in determining advanced therapy strategies. Hence, predicting disease progression reflects a rephrasing of one of the goals of predictive medicine, which aims at combining the health of many and the disease of one, to capture the complexity of chronic diseases. It is essential for public health planning, along with the identification of disease onset, treatment allocation, and drug prescribing. Furthermore, predicting the chronicity of a disease is of immense relevance in the development of preventive strategies. Some pioneering works have tried, with limited success, to predict adverse outcome processes in chronic disease, such as the rapid progression of liver disease. The successful model is a combination of the insights of eight specialists about patients included in a recent intervention trial. Each of the experts spent one day a week for eight weeks on the project. Other examples of predictive modeling in such a patient population are lacking.

This work is licensed under CC BY-NC-SA 4.0.

2. Foundations of Machine Learning in Healthcare

In its practice, machine learning utilizes large medical datasets - electronic health records, omics data, imaging, wearable device data, and many other sources of information that are combined and analyzed using algorithms. In this chapter, we discuss the processes by which complex, multivariable medical datasets are used to create ML models that can make individualized predictions. Machine learning is built on statistical principles, using models and algorithms to learn important features of the input that are relevant to the output. This process is agnostic to the type of data being used and operates on the principles of feature quality improvement and redundancy reduction.

The advent of machine learning methodologies in healthcare has brought with it new opportunities and new concerns. One reason for the proliferation of machine learning methods in healthcare is the sensing of big data that underpins precision medicine. One of the most imminently deliverable approaches to the use of machine learning in healthcare is the use of ML-derived models for the identification of which patients are at risk of future health events. However, despite such promise, the infallibility of ML models and methods is not self-evident. Problems surrounding missing and unmeasured data and competing risks and longitudinal data structures are yet to be fully resolved. Incorrect reliance on ML using poor quality data can inadvertently prolong sections, exacerbate health disparities, and reinforce patterns of care that are not in patients' best interests. When promoting the use of AI technologies in healthcare, the historic and current impacts of ethically dubious practices cannot be ignored. Technology should not be exploited for profit with disregard to patient safety.

2.1. Key Concepts and Terminology

Over the past decade, there has been both increasing interest and rapid growth in medical applications of machine learning. This topic has broad multidisciplinary relevance: from developing core methods applicable to a range of common types of data, data sources, and clinical questions to tailoring new treatments for complex chronic conditions, making predictions relevant to personalized approaches, and providing tools for shared decision-making for individual patients, their families and caregivers, and other important stakeholders. This paper focuses on the application of machine learning models that can be utilized for forecasting disease progression.

This work is licensed under CC BY-NC-SA 4.0.

Datasets usually include one or more files, with one row per subject-timepoint combination. Each row typically includes at least the following three types of information—features, labels, and covariates: Features are the 'predictor' variables used to train the model and can contain a variety of different data types (e.g., continuous, categorical, or imaging data). Labels refer to the outcome of interest; e.g., the target variable that the machine learning model should aim to predict, or changes in the target variable that the model should aim to forecast. Covariates are not used directly in making prognostic predictions but may include important information about subjects that should not be used 'post hoc' to artificially enrich the performance of a model (e.g., diagnosis, whether a subject is receiving ventilation, performing extreme sports, or using an investigational intervention). One key aspect we consider is whether models will be trained to perform unsupervised or supervised tasks.

In a supervised learning problem, models are trained to make predictions quantitatively close to actual labeled outcomes. In our context, this may involve predicting disease status or symptom scores for neurological examination, healthcare resource utilization, patient-reported outcomes, and motor or cognitive scales often used in clinical trial design. There are multiple examples of this being done for a range of different diseases; indeed, predictive disease progression modeling is a well-established area in the statistical sciences, especially in the context of neurodegeneration. Despite making progress in the accuracy of predictions by leveraging supervised machine learning models, much current application has only been shown to function as relative improvements by assessing models on the same dataset used to train the models. It is also a major challenge to generalize many supervised learning models because solutions often involve relationships that are peculiar to the training dataset and are harder to generalize to different disease cohorts, thus potentially requiring new machine learning-trained software to be developed for each new cohort. Overall, application and regulatory acceptance will also require generalizable models.

3. Types of Machine Learning Models for Disease Progression Prediction

Many machine learning models have been specifically developed for disease progression prediction. Here, we categorize different prediction models on a different basis. A suitable categorization is on the basis of the learning paradigm used. Based on this, forecast models can be broadly categorized as models based on (1) supervised

This work is licensed under CC BY-NC-SA 4.0.

learning, (2) unsupervised learning, and prediction models that can be formulated as problems in (3) reinforcement learning.

Supervised learning models are further categorized based on the underlying decision mechanism into methods like decision trees, neural networks, ensemble methods, kernel methods, and feature selection-based methods. There are different clinically relevant models where one or more of the models based on different learning paradigms are suitable. The choice of a particular algorithm is influenced by several case-specific aspects which might be governed by the clinical question the modeling is trying to address (e.g., presence of heterogeneity in the patient population, type of attributes in the dataset—continuous, categorical, ordinal, volume of data, data being original or preprocessed, etc.). Similarly, the choice of the objective, which can be maximizing overall accuracy or predicting outcomes of a rare subpopulation of interest, is also made based on a careful consideration of the clinical scenario. Therefore, an appropriate algorithm for a clinical scenario is best chosen after having addressed these aspects.

3.1. Supervised Learning Algorithms

Supervised learning is used to build predictive models as we train the models using a "labeled" dataset, consisting of input-output pairs, so that the model can learn the target function. The trained model captures the relationship between the input features and target labels, and then we can use the model to predict new instances' target labels for which only input features are available. There are a variety of algorithms available to train a supervised learning model, such as linear regression, logistic regression, decision trees, rule induction, support vector machines, k-nearest neighbors, Gaussian processes, random forests, gradient boosting, and artificial neural networks.

Supervised learning is widely studied in the healthcare domain for numerous applications, such as disease prognosis, predictive modeling, automated diagnosis support systems, and treatment strategies. The advantages of using supervised algorithms in the healthcare domain include their high accuracy and interpretability, while there are also some challenges that need to be solved, including the need for high-quality labeled data for training, avoiding overfitting the models to the training data, as well as developing models with good generalization to be applied to unseen testing data. Therefore, feature selection and engineering have been studied to develop unsupervised and supervised algorithms to facilitate disease progression modeling.

This work is licensed under CC BY-NC-SA 4.0.

Overall, it is evident that supervised algorithms are widely used in developing disease progression models.

4. Applications of Machine Learning in Clinical Outcomes Forecasting

Applications of Machine Learning in Clinical Outcomes Forecasting Machine learning is transforming numerous aspects of healthcare. The utility of ML models in forecasting clinical outcomes constitutes a transformative new technology in the management of patients. Forecasting patient values in multiple electronic health record measurements over time serves to summarize the patient course in a particular dimension or dimensions. This forecasted trajectory can then serve as an input into a range of predictive modeling applications, such as predicting all-cause mortality or predicting complications.

Numerous specific longitudinal outcome forecasting applications exist across a range of clinical domains, including oncology, where ML forecasts of tumor progression can significantly affect multiple levels of healthcare, cardiology, and infections. In most forecast models, precisely what is estimated is some aspect of patient disease progression. This could be of an oncologic disease, a time delay from the end of therapy until "local failure," prediction of certain complications, such as a disease relapse, or even a disease-free interval, if the measurement is something like serum prostate-specific antigen levels. Models predict "endpoints," typically defined in clinical guidelines. Results of ML forecasting models should be used to support clinical decisions, not replace them. In general, ML models indicate malignancy close to the endpoint; therefore, move to a treatment change if available, and it is advisable to try to move straight to the probable next best therapy, not wait for overt confirmation.

The area of ML forecasting is still largely a research area, with some key clinical barriers to wider application needing resolution. Most model-building efforts select patient populations that are less likely to reflect the general patient cohort to which an ML model will be applied, which hinders generalization. Many of the oncology studies are based on a clinical trial database that may also limit generalization because trial patients are more likely to be heavier and more rigorous healthcare consumers. Several studies have assessed whether offering clinical decision support in the form of ML patient trajectories, such as assessing patient mortality risks and also estimating likely 5-year best outcomes for a range of common cancer types, alters patient outcomes. In some of

This work is licensed under CC BY-NC-SA 4.0.

these, for individual patients, findings indicate that estimated mortality is a significant prognostic factor in treatment decision-making. Many of these ML trajectory patient management studies are currently ongoing. Application of ML to EHR measurements may offer a powerful means of improving a range of healthcare by being able to detect significant changes in the pattern of disease early or encouraging more refined and discrete disease stratification for effective personalized approaches to medicine.

4.1. Case Studies and Success Stories

In this subsection, we share cases where machine learning models have a concrete application and have been used in real-world scenarios, illustrating how their use can support the forecasting of clinical outcomes and also provide valuable insights. The ten cases cover different applications in the medical fields, such as being used not only to predict how long a patient will survive but also how well, to determine the likelihood of recurrence, or how severe a disease might develop based on the individual patient's clinical information. For each case study, we present the scenario in which it is applied, apply the SWOT analysis to detail its challenges, both methodological and practical, its significance within the medical field, and the results and metrics obtained from the machine learning models used in each case.

Finally, we conclude by briefly reflecting on the future and expanding research themes in these clinical follow-up scenarios and provide insights regarding business and translation that can stimulate directions for further reading. All cases are discussed in chronological order of the real scenarios they are based on.

5. Personalized Treatment Strategies Using Machine Learning

Treatment strategies personalized for each patient

Personalized treatment strategies, tailored to an individual's illness, are a major new focus for healthcare. In contrast to the traditional one-size-fits-all approach, treatment personalization aims to tailor healthcare decisions to the specific needs and characteristics of each patient. The individualized insights unlocked by machine learning can provide new solutions to enable personalization in healthcare. The potential for personalization made possible through machine learning has led to growing interest and investment in this space. The vast volume of healthcare big data, generated from electronic health records, clinical registries, molecular and genetic data,

This work is licensed under CC BY-NC-SA 4.0.

and detailed patient reports, provides the fuel that machine learning techniques need to define different and new treatment strategies.

Predictive analytics can estimate how an individual patient will do and can compare the likely outcomes of any treatment options, all based on the specific characteristics of the patient. This type of analysis is known as predictive modeling. Another type of strategy is to group people who share similar characteristics together, using techniques known as patient stratification, and to then base different recommended intervention strategies for each subcategory of patients. Predictive and stratification modeling techniques often use combinations of predictor variables, often derived from processing large, heterogeneous datasets, to make these types of personalized recommendations. For example, combining patient demographic, genetic, and clinical information can enable machine learning models to predict disease progression, treatment outcomes, or a patient's response to therapies. However, models must be appropriately validated before they can yield meaningful results. Before any machine learning model is used as a basis for decision-making in patient care, it will be important to assess whether the new decisions improve patient health and reduce the risk of any adverse events. Clinical trials in a real-world setting can measure the impact of these new decisions on treatment effects with more confidence. Such trials are also able to identify potential negative consequences of poor-quality predictions and retraining management strategies and model development.

Treatment personalization implies that many different intervention strategies may need to be identified to ensure that each patient can be treated in a way that helps them in the best way, with minimized harm. Different patient groups with unique disease trajectories may emerge from large emergency department patient cohorts, and many patient stratification algorithms can be based on these findings. While population disease progression tends to be described with large cohorts, sometimes epitomized by big data, for machine learning models to be tailored to individual patients, these findings must be robust and accurate; employing large enriched heterogeneity cohorts will echo precision medicine. Model-based strategies in medicine make the promise of the potential reality of making large amounts of data digestible and actionable. They have the potential of improving the efficacy of intervention without substantially increasing adverse effects of an intervention. Robust data, robust models, and thorough testing in clinical settings of an intervention strategy derived by AI have the potential to

This work is licensed under CC BY-NC-SA 4.0.

be a real game changer in healthcare with the potential to save even more lives. It is essential that as we move into this exciting era, research in AI and in health evaluation is non-stagnant, as without this fundamental research, the potential of AI will remain just that.

5.1. Patient Stratification Techniques

The group of patients who will respond to a particular therapy comprises the patient population that will benefit. To develop personalized treatment strategies to address the specific needs of different patient subpopulations, clinical trials must be able to identify which patients will respond to a given therapy or therapy combination. The stratification process involves a subgroup or subgroups of patients who are grouped based on one or more shared characteristics, including demographic characteristics, lifestyle, genetics, clinical data, etc. Several patient stratification methodologies have been proposed, for example, based on patient clustering algorithms, decision trees, logistic regression, decision-analytic models, etc. The purpose of stratification is to identify sets of subpopulations of patients who differ from one another in the expected scale of the disease, either with or without a specific treatment. By considering this and identifying patients who would not otherwise have been eligible for clinical trials, there is potential for these sets of subpopulations to be the subject of research, leading to the development of additional new and innovative therapies.

In addition to identifying treatment-responsive subpopulations, patient stratification has the potential to aid in the identification of subpopulations unlikely to benefit from a treatment or for whom adverse outcomes may be severe. For example, in the development of checkpoint inhibitor cancer therapies, it was quickly recognized that only a small subset of patients was likely to benefit from the treatment. A primary challenge associated with the proliferation of patient stratification models, mapping of these approaches onto different patient populations, and the use of diverse biomarker technologies is the ability to generalize the model prediction.

Early real-world examples of successful patient stratification include the correlation of high HER2 receptor expression with response to trastuzumab in patients with breast cancer and hepatitis B virus infection status with the response to lamivudine in patients with chronic hepatitis. In both of these cases, the publication of randomized controlled study results to identify patients by the status of these biomarkers and patient outcomes

This work is licensed under CC BY-NC-SA 4.0.

in subgroups that tested negative versus positive preceded much of the actual use of this test-treat strategy. This is a similar trend to the regularization of patient stratification for COVID-19, with the reorientation of therapeutic clinical trials to regimes correlated with oxidized glutathione ratio and patient mortality before saturation kinetics had been empirically determined. Continuous learning from real-world data and the modification of subpopulation definitions over time are key drivers and feedback loops for population stratification.

6. Future Direction

Artificial intelligence has great potential to translate data science into the clinic. In the years to come, we anticipate continued advancements in data pre-processing, algorithm development, and new data type integrations as sources of variance and bias. Wide-ranging influences on decision-making to improve individuals' health exist. In practice, future success will come from interdisciplinary collaborations between data scientists and healthcare providers who can adopt the technology and shape policy around its use. Privacy and bias issues must be addressed proactively. Databases of electronic medical records, which can include claims and social surveillance, are being analyzed with AI to predict and influence a wide range of health effects above and beyond clinical outcomes. Indeed, real-time data from wearable technologies are a rapidly emerging feature of 21st century healthcare and enable monitoring in supervised and unmonitored settings. As AI continues to make inroads into the clinic, the potential to use models for patients to directly and personally tailor care pathways to 'individualized patient journeys' is exciting. There are many remaining challenges to AI enhancing healthcare, including scale-up and implementation within health systems. The willingness and energy of patients, volunteers, clinicians, and data scientists to work together is essential to meet these challenges.

7. Conclusion

In this essay, we have used six cases to demonstrate how machine learning techniques have been utilized in the construction of predictive models. These models are designed to bridge the gap between clinical patient data and the creation of novel knowledge. Our potential is that such models will, in the medium and long terms, contribute to improvements in two long-term clinical decisions: the decision on intervention in patients whose progression probability has crossed a predefined threshold, and the

This work is licensed under CC BY-NC-SA 4.0.

stratification decision between possible interventions based on their predicted impact. Both developments have the potential to contribute to important improvements in patient outcomes and cost savings when they are part of hospital workflows.

The key to progress in digital health domains such as predictive machine learning models is not only to improve performance, generalizability, and interpretability of models, but also to integrate the usage of models in the workflow, monitor usability, and demonstrate clinical end-user uptake. Therefore, it is important to think about usability and uptake early in the process of developing machine learning models. This makes it reasonable to collect the right information in a meaningful way, to explain to clinical partners what a machine learning model is and what it can predict, and to involve clinical partners in the development of the model in order to maximize its clinical acceptance.

In conclusion, it is clear that the field of disease progression prediction is rapidly growing and technologically developing. Many recent methods already perform better than conventional regression analyses, even if they are not yet considered ready for translation to clinical healthcare settings. In the future, medical records will increasingly contain data for training a person-specific prediction model that incorporates both physiological, morphological, behavioral, and genetic data. In order for the model to be user- and clinician-friendly, hybrid user interfaces that are discreet, non-interfering, and adaptive could be used. In this direction, we foresee that clinical software, based on machine learning models and containing continual machine learning capacity, will contribute to improvements in patient management.

This work is licensed under CC BY-NC-SA 4.0.