ISSN: 2577-2996

PROGRESS IN MEDICAL SCIENCES, 2025
VOL 9, NO. 1, PAGE 1 – 4
DOI: doi.org/10.47363/PMS/2025(9)E173

REVIEW ARTICLE

Open Access

Risk Scoring Models in Healthcare Data Analytics: A Practitioner's Perspective on Machine Learning, Interpretability, and Regulatory Compliance

Simran Sethi

USA

ABSTRACT

The development of risk scoring models has significantly contributed to the field of healthcare analytics because of its capabilities in predicting diseases, managing healthcare resources, and even affecting policy formulation. Advanced machine learning (ML) techniques have been developed over time that are outpacing the ability of traditional statistical models to achieve their goals. However, concerns regarding model transparency, real-world validation, and compliance with regulatory bodies, chiefly in relation to HIPAA and HCC in the US, persist. This paper presents consolidated comparisons of testimonies acquired during the rollout of an enterprise scale risk scoring module leveraging patient health records and claims documents focusing on effective model design, interpretability, and compliance imperatives. The objective of this paper is to provide other healthcare practitioners and researchers who wish to employ sophisticated risk scoring systems which can function seamlessly within the bounds of legal compliance and sufficient explanation afterwards with both compliance and practical aspects in mind.

ARTICLE HISTORY

Received January 15, 2025
Accepted January 22, 2025
Published January 29, 2025

KEYWORDS

Healthcare Data Analytics, Risk Scoring Models, Machine Learning, Interpretability, Regulatory Compliance, HIPAA, Hierarchical Condition Categories (HCC), Feature Engineering, EHR Data, Model Validation

Introduction

In the last 10 years, the deployment of risk scoring models in healthcare has increased due to the features of EHR systems, the growing sophistication of health data analytics, and the increased medical data. Simple or complex risk scores aim to estimate the probability of any adverse health outcome, event utilization, or expenditure incurred over a certain period.[1] These models have proven invaluable in providing decision support both at the single unit level (for example, patient care planning) and at the aggregate(unit) level (for example, insurance claims payment, resource allocation).

Modern risk models go beyond the historical focus of clinical practice which utilizes regression-based methods. Random forests, gradient boosting machines, and neural networks are some of the Machine Learning (ML) methods that have been proposed for the extraction of complex non-linear relationships from large dimensional data [1, 2]. Although these ML models have the opportunity for better performance, their "black-box" nature makes them challenging to interpret, which is needed to earn the trust of the clinicians for responsible clinical deployment [2-4]. In the United States, regulatory matters are further complicated by the presence of the Health Insurance Portability and Accountability Act (HIPAA), which controls the privacy and security of patient's data [5]. Furthermore, the HCC (Hierarchical Condition Category) models are often used by healthcare payers and providers to determine reimbursements based on the risk profiles of patients, which necessitates validation and clear justification [6].

This paper provides a practitioner's view on designing, deploying, and maintaining risk scoring models in healthcare analytics by leveraging extensive experience in patient risk module construction in a large scale setting. We were involved in the creation of a Patient Risk Score/module (CMS risk score) that is implemented in Scala and uses over 100gb of electronic medical and health records. The discussion is framed around a literature review of existing risk scoring techniques and their interpretability, as well as understanding compliance issues with HIPAA and HCC.

Organization

Section II reviews the state of the art of risk scoring models where the primary focus is on ML compared to classical statistics. Section III proposes a number of key design issues such as feature engineering, interpretability, and model building validation. Section IV discusses the aspects of regulatory compliance with particular reference to HIPAA and HCC. Lessons from scaling risk scoring modules are discussed in Section V with our closing remarks and next steps provided in Section VI.

Landscape of Risk Scoring Models

Traditional Statistical Models

For the past decades risk prediction in healthcare has been done using regression type techniques. Logistic regression is particularly common because of its interpretability, ease of use, and relatively good performance on tabular data. Accepted clinically scores, which include the Framingham Risk Score for cardiovascular disease and the Gail model for breast cancer risk, use a limited

Contact: Simran Sethi, USA.

set of features that are selected using domain knowledge [1]. Such models translate risk into manageable formats through either regression coefficients or point-based systems which quantify risk and hence make clinician acceptance easier.

Never the less some researchers have found serious problems with these linear methods, especially with more complicated feature spaces and interactions [1, 2]. Despite these issues, logistic regression remains an important reference point, and in many comparatives well designs approaches, logistic regression has performed equally or even slightly better when compared to machine learned methods in smaller datasets or where regularization was appropriately applied [1].

B. Machine Learning Approaches

The widespread use of EHRs, claims data and various types of imaging has exponentially expanded the application of machine learning in risk prediction [3]. Improved performance has been recorded in tree-based ensemble techniques like random forests and gradient boosting which are common with large feature spaces and non-linear interactions [3]. In particular, Imaging based diagnostics and genomics are witnessing the rise of deep learning methods such as feed-forward neural networks and convolutional neural networks [7].

However, as with most advanced ML methods, the most significant drawback is the "black-box" perspective where one cannot understand the internal workings of the model [2]. In healthcare and other high stakes industries, the clinicians and regulators set the bar for not only high performance but also how the predictions are arrived at. Additionally, ML model deployment is not made easy as robust hyperparameter tuning and the dangers of overfitting require extensive validation, generalizability, and even governance [1, 2].

Key Design Considerations

Feature Engineering

Risk scoring models are built using an array of steps, and feature engineering stands out as one of the most important. In EHR, features would contain raw components such as diagnosis, medication, lab tests, and demographic information, among others. Our experience at Innovaccer highlighted encoding ICD codes as both acute and chronic diagnosis conditions. Grouping singular diagnoses into condition categories like HCC clusters is helpful in summarizing patient morbidity profiles [6].

Risk estimates can be further refined by incorporating temporal features, including recent admissions into the hospital, changes in medication prescription, or increased windows of disease exacerbation [1]. For example, severe inpatient episodes could dramatically increase readmission risk in the short term. Often, categorizing these temporal patterns with rolling windows or time decay factors proved useful. Natural language processing (NLP) can also help with unstructured data (such as clinician notes) which is not available in structured fields [2].

Interpretability Methods

Interpretability serves as the backbone for healthcare risk modeling, as clinicians expect transparency in this area. While less advanced models are regress-and-simplistic based, and thus easy to understand, more complex models utilize ML algorithms for a multitude of reasons: explainability, interpretability, or even something else entirely.:

[1] Inherently Interpretable ML: Generalized additive models with pairwise interactions (GA2M) are able to learn black box models and provide predictive transparency at the same time [2]. Sparse rule-based classifiers can also have transparent logic [2].

[2] Posthoc Explanations: Certain methods like Shapley Additive Explanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME) offer customized feature importance which aids stakeholders in understanding the reasoning behind derived predictions [3]. Although there is greater transparency at the end, these techniques add a layer of complexity to the model and require skillful data science knowledge to be utilized appropriately.

[3] Hybrid Approaches: These approaches are a fusion of interpretable rule sets for the major features of interest, and an ML model for the remaining complexities. Such methods are more transparent than traditional black-box methods [2].

Through interpretability, clinical trust is enhanced which makes it easier for other users to adopt this and ethical guidelines for medical AI. It also enables data scientists to identify bias or artifacts that may have been trained [3].

Model Validation and Performance Metrics

For dependable risk scores, model validation is critical. The TRIPOD guidelines advise on the division of the data into training, validation, and test sets while performing internal validation through cross-validation [1]. In addition, generalizability is verified through external validation on a different population or dataset.

Metrics that have Gained Popularity Include:

- **Discrimination:** Area under the Receiver Operating Characteristic curve (AUC) or Precision-Recall curves.
- **Calibration:** Comparing predicted probabilities to observed event rates, especially through calibration plots.
- **Reclassification measures:** Net Reclassification Improvement (NRI) or Integrated Discrimination Improvement (IDI).

Reliability calls for periodic recalibration whenever patient demographics or clinical protocols change; this is also referred to as dataset drift [1], [3]. Another best practice is tracking model performance over the duration of use and instituting automated detection of performance decline [3].

Regulatory Compliance in the US Context

HIPAA Considerations

As with any US healthcare analytics, abiding by HIPAA regulations is a must. HIPAA provides privacy and security measures for protected health information (PHI) by direct identifiers such as the patient's name or social security number and indirect identifiers such as specific dates and geographical details [5].

To perform analytics, an organization will often use a de-identified or limited dataset which can be processed or shared. De-identification removes all of the 18 specific identifiers, which allows for further usage of the data without being subject to HIPAA's most stringent provisions [5]. On the other hand, if the data is identified for real-time clinical purposes, secure environments and Business Associate Agreements must be put in place to meet compliance. Common security measures include encryption, audit trails, and role-based access control [5].

For risk scoring solutions that embed themselves into clinical workflows (such as with hospital EHR systems), HIPAA-compliant deployment usually encompasses separate production environments, thorough tracking, and controlled access. PHI in themselves may be considered Model predictions if they are identifiable to specific patients [5]. Hence, powerful PHI and administrative controls must always be in place.

HCC Risk Adjustments

Comorbidities of patients are taken into consideration in the adjusted cohort treatment file model, or the Hierarchical Condition Category (HCC) model, and is the basis for risk adjustment in the Medicare Advantage and other USA insurance programs since it is a bidirectional reimbursement model [6]. HCCs bundle together related ICD codes. Each category is assigned a risk score, which when aggregated is used to calculate the patient's reimbursement. Patients with multiple chronic conditions are within the higher reimbursement range categories in order to anticipate healthcare expenses.

The area of HCC calculations has always rested on regression-derived based expense claims information, but newer strategies are looking to incorporate machine learning. ML can assist in uncovering undocumented hidden risks which can potentially improve cost estimation [6]. Still, complex algorithms pose a problem. Regulators require clarity and freedom in evaluating accounts and that is difficult with the intricacy of ML. As such, HCC practitioners bear the burden of justifying model outputs, ensuring that ICD coding is compliant with clinical documents.

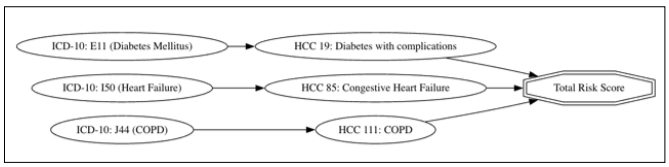

Figure: HCC Risk Adjustment Model Overview

Lessons Learned from Real-World Implementation

Model Building and Deployment in Scala

At my previous Healthcare-startup company, we developed a risk scoring platform (including CMS risk scores) on a Scala-based microservices architecture. Scala's strong typing, concurrency framework (Akka), and functional programming paradigms provided advantages in handling streaming data from large EHR pipelines. Key lessons included:

[1] **Efficient data ingestion:** Pulling data with a large multi-dimensional cube: Application of Apache Spark for other 100 GB + datasets made it possible to distribute processing. The system was effective with concurrent uptakes of data with little lag.

[2] **Feature engineering pipelines:** Underlining the importance of reusable diagnosis driven coding modules. The code could be altered easily.

[3] **Containerization and orchestration:** Providing clients with environment independent services using Docker and Kubernetes.

[4] **Version control and reproducibility:** Rendering and keeping an audit trail of meticulous versioning of data transformations vital for HIPAA compliance as well as quality assurance.

Balancing Complexity with Interpretability

The application of risk scores using random forests and gradient boosting needed to have increased predictive accuracy tes revalidation focused especially in certain subpopulations like diabetic patients or patients with multiple comorbidities. However, outputs were always demanded to be more transparent from clinicians and clients. Explaining patient specific risk factors using SHAP based explanations managed to close the trust gap on the phong by [3]. Still, for some scenarios such as compliance audits and real-time feedback loops given to the physicians, more interpretable regularized logistic regression models were needed as a relief option.

Ensuring Ongoing Validation

Continuous performance monitoring is vital for real-world implementations. We periodically recalibrated the models upon shifts in coding practices, as changes in ICD usage or improvements in HCC documentation can inflate or deflate risk scores unexpectedly [6]. Automating these monitoring processes minimized manual overhead and ensured timely detection of anomalies.

Conclusion and Future Directions

The healthcare vertical, both analytically, PPC based and MCM, relies on clinical risk scoring models. Our experience outlines how ML can be utilized in an explainable manner with tips for ensuring compliance with legal regulations. The deep learning revolution has only started and with the associated growth of healthcare data, innovations will focus on deep learning, techniques like federated learning to solve the privacy problem, and more sophisticated clinical workflow compatible explainable AI.

In particular, we expect more unstructured clinical note and imaging text data features to be used in clinical risk scoring. Systems which learn continuously—where model parameters change in almost real time with the evolving patient population—will also be gaining importance. In parallel, laws will try to catch up, which will possibly cause regulations for AI based systems to be more stringent in regards to auditing and visibility.

Last but not least, there is a need to prepare for compliance and explainability which is a prerequisite for convincing stakeholders in healthcare and partnered industries for the maintenance of ethical business practices guarantees to provide the replaceable quality of the predictive analysis results. The main question is: what to guarantee that the AI model continuously achieves better patient outcomes versus deteriorating healthcare quality.

References

[1] E Christodoulou, S Ma, G Collins, Ewout W Steyerberg, Jan Y Verbakel, et al. "A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models". Journal of Clinical Epidemiology. 2019; 110: 12-22.

[2] R Caruana, Y Lou, J Gehrke, Paul Koch, Marc Sturm, et al. "Intelligible Models for Healthcare: Predicting Pneumonia Risk and Hospital 30-day Readmission". in Proc. 21th ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining, Sydney, Australia. 2015; 1721-1730.

[3] C Rudin. "Stop explaining black-box machine learning models for high stakes decisions and use interpretable models instead,". Nature Machine Intelligence. 2019; 1: 206-215.

[4] B Ambale-Venkatesh, C Yang, Y Wu, Kiang Liu, W Gregory Hundley, et al. "Machine Learning for cardiovascular risk prediction in the Multi-Ethnic Study of Atherosclerosis". Circulation Research. 2017; 121: 1092-1101.

[5] I G Cohen, W Nicholson Price II. "Privacy in the age of medical big data". Nature Medicine. 2018; 25: 37-43.

[6] G Weissman. "Hierarchical Condition Categories for Pulmonary Diseases: Population Health Management and Policy Opportunities". Chest. 2019; 155: 868-873.

[7] S Hussain, M Akhund, R Bano. "Breast cancer risk prediction using machine learning: a systematic review". *Frontiers in Oncology*. 2024;14: 1-12.