AI-Driven Big Data Analytics for Personalized Treatment Planning

Yannis Ioannidis
Asst Professor

Abstract

Big data and AI are poised to transform healthcare due to their capacity for providing personalized predictions and treatment decisions for individuals. Such predictions, often termed risk scores, are commonly used in predictive modeling for risk stratification. Nevertheless, risk scores are only part of a comprehensive predictive analytics strategy for personalized health. Data-driven risk scores must inform the causal-inference models that determine the optimal treatment to maximize the long-term well-being of individuals. Together, these two components can augment clinical decision support for health providers.

Despite these advances, the clinical translation of AI-augmented analytics relies critically on real-world evidence derived from their prospective application in clinical practice. Nevertheless, few of the aforementioned AI efforts require evaluation through pilot studies that collect evidence of their practical value. Future work, therefore, focuses on developing a supporting architecture to maintain a real-time repository of prospectively collected evidence. Such work is informed by discussions with health domain experts and ethics approval from an institutional review board.

Keywords: Big Data in Healthcare, Artificial Intelligence in Healthcare, Personalized Medicine, Predictive Analytics, Risk Stratification Models, Clinical Risk Scores, Causal Inference in Healthcare, Treatment Optimization, Clinical Decision Support Systems, Data-Driven Healthcare Models, Real-World Evidence (RWE), Prospective Clinical Studies, Healthcare Data Repositories, AI Model Evaluation, Evidence-Based Medicine, Health Data Analytics, Patient-Centric Care, Ethical AI in Healthcare, Institutional Review Board (IRB) Approval, Translational Healthcare AI.

1. Introduction

Data-driven analytics and machine-learning algorithms are revolutionizing business, medicine, science, and nearly all spheres of social life. In health care, advances in predictive analytics and clinical decision support based on big data and AI hold great promise. Consequently, augmented analytical methods will enable risk assessment, causal inference,

effective treatment selection, and personalised healthcare treatment planning. These methodologies are expected to become essential paradigms for understanding complex real-world business systems and designing real-business solutions. Yet, personalized treatment selection has remained an unresolved problem within the health care industry framework. Ultimately, no health care decision can be more critical than determining the optimal treatment procedures for a specific individual—a decision that substantially influences the length and quality of life.

Big-data-driven healthcare analytics, supported by AI technology, are predicted to become important tools in determining such personalised treatment solutions and medical-planning systems. Implementation success will depend, however, on the quality of data acquisition and the data-source–technology integration framework. An adequate data acquisition structure and methodology remain unfinished business in the intelligent systems development area. Research is therefore warranted in establishing the necessary data-technology integration design and infrastructure to facilitate standard-setting pilot studies.

1.1. Overview and Significance

Recent advances in predictive analytics and machine learning provide unprecedented opportunities for delivering personalized healthcare services at scale. Novel approaches leverage population-level datasets to predict disease risk, provide clinical decision support, and assess the impact of interventions for treatment optimization. These predictive capabilities can improve healthcare outcomes through early intervention, more effective treatment selection, reduced care fragmentation, and better resource allocation.

AI-augmented big data analytics provides an effective framework for building scalable personalized healthcare services. Although soon-to-appear annotated clinical databases will enable directly supervised techniques for addressing practically relevant questions, heterogeneity in complex systems limits accuracy in prognosis, diagnosis, and treatment selection. Unsupervised and reinforcement learning techniques, however, enable tackling higher-level challenges through risk stratification, predictive modeling of time-to-events, and causal inference.

2. Background and Theoretical Foundations

The underlying conceptual framework for AI and big data-augmented personalized healthcare treatment planning comprises predictive, diagnostic, prescriptive and perspective analytics. Medical knowledge serves as the key asset for extraction of the algorithms and models needed to formulate hypotheses and rules. Ethical concerns are addressed using the issues of privacy, bias and discrimination, data origin and algorithmic decisions as examples. AI and big-data-augmented predictive analytics enables health risk stratification to allow early and targeted medical actions on the at-risk patients.

The present research summarizes several practicable, non-technical advances in these fields of knowledge including proposals for the augmentation of predictive analytics and clinical decision support, reinforcement learning for treatment recommendation, causal inference for treatment effect modelling, medical care planning based on multi-party health care cost contributory planning, supporting patients' decision-making, supporting the planning of preventive programs via semi-Markov decision processes, incorporating health care coverage decision-making processes, the assessment of privacy protective techniques and the exposition of perturbed data for risk and preparedness analysis.

2.1. Conceptual Framework and Key Principles

Response optimization relies on a conceptual framework that integrates Augmented Intelligence, Deep Learning, and Big Data. Augmented Intelligence enhances human cognition via Deep Learning-powered pattern recognition. The synergy of AI and Big Data optimally transforms historical data into actionable knowledge. The analysis of tailored patient datasets yields risk models and treatment predictions that strengthen clinical decision support systems. Predictive models identify high-risk patients requiring intervention. Treatment assignment predictions stratify patients by best-expected treatment outcome. Transferable predictive models developed on large, heterogeneous Big Data cohorts facilitate response modeling in new, smaller patient sets. Comprehensive harnessing of medical Big Data sets a foundation for Precision Medicine.

ISSN: 3067-4190
JOURNAL OF ARTIFICIAL INTELLIGENCE AND BIG DATA DISCIPLINES
Volume: 01 Issue: 01
RECEIVED: OCTOBER 22 REVISED: NOVEMBER 09 ACCEPTED: NOVEMBER 24 PUBLISHED: DECEMBER 20

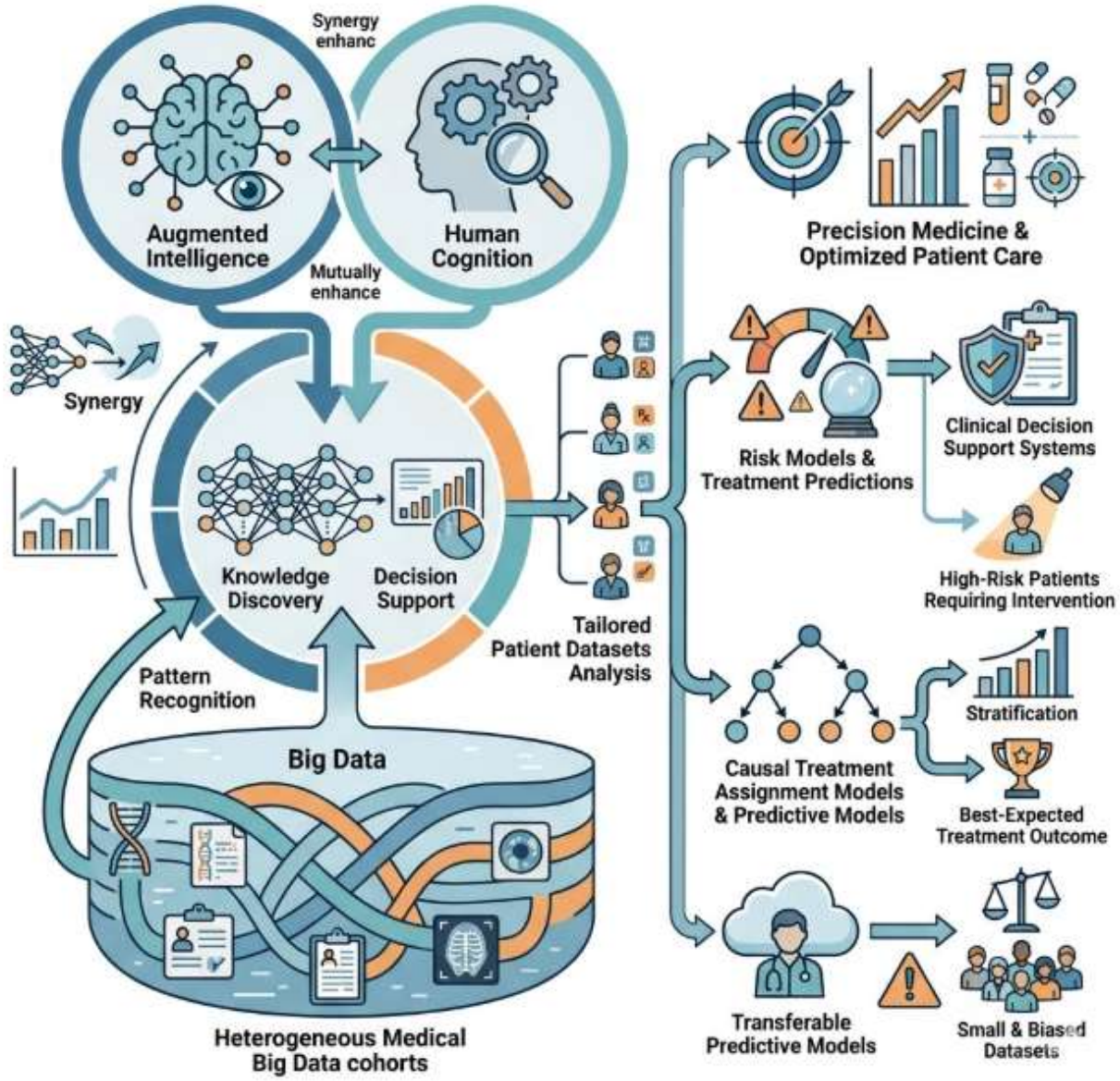


Fig 1: Navigating Complexity: An Integrated Framework for Augmented Intelligence and Deep Learning-Powered Precision Medicine Using Heterogeneous Big Data

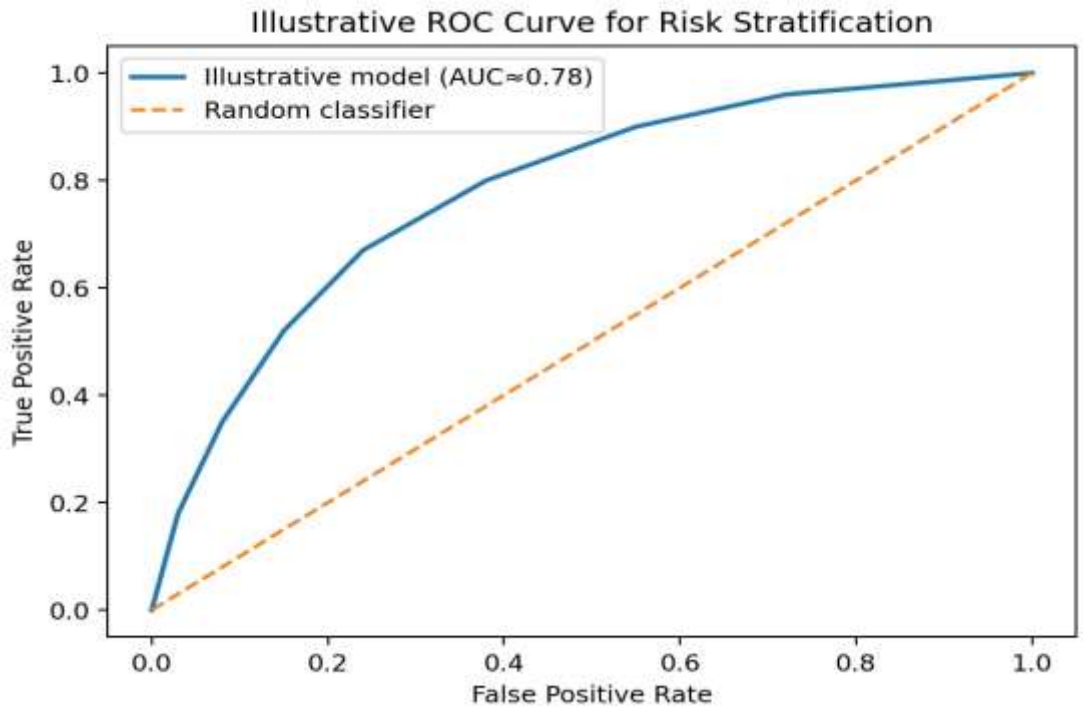
Augmented Intelligence enhances human intelligence with automatic pattern recognition of deep learning. Artificial Intelligence and Big Data mutually augment each other when knowledge discovery of historical data is translated to decision support for new patients. If the clinical dataset of interest is small and biased toward positive outcomes, machine learning may not yield optimal models. The complement of established AI models with causal treatment assignment models can mitigate this risk. Predictive models not only identify high-risk patients likely to benefit from intervention but also suggest the optimal treatment when more than one option is offered.

3. Methodology

The success of personalized healthcare relies on how well the applied methods help clinicians and healthcare practitioners make better predictions, recommendations, and treatment plans relative to previous standard-of-care methods. Advances in predictive analytics, clinical decision support, and treatment optimization for personalized healthcare dependence upon augmenting big data analytics with AI methods addressing increasingly granular details of predictive modeling and risk stratification, personalization of treatment selection, and assessment of potential treatment impacts. The resulting collection of meta-analytic methods can be scaled and adapted to different types of healthcare data sources.

Activation of patients' genetic, molecular, electronic medical records, and longitudinal lifestyle natural history repositories for novel discovery analytics presents new opportunities and challenges for personalized healthcare, particularly how to incorporate traditional demographic, general disorder category, and clinical specialists' expertise with other patient information for optimized AI-augmented decision-making. The combination of different stratification features into individualized risk prediction models enhances their predictive power and reliability. The increasing availability of data and information integration capabilities enables analogy-supported predictive modeling, which does not require an exhaustive database of all possible input-output combinations. The goal is an inclusive family of methods, progressively addressing limitations of predictive modeling in areas such as clinical advice, preventive exposure decision-support, lifestyle-switch suggestion, or active intervention treatment adoption.

ISSN: 3067-4190
JOURNAL OF ARTIFICIAL INTELLIGENCE AND BIG DATA DISCIPLINES
Volume: 01 Issue: 01
RECEIVED: OCTOBER 22 REVISED: NOVEMBER 09 ACCEPTED: NOVEMBER 24 PUBLISHED: DECEMBER 20



Equation 1) Counterfactual Risk Attribution: direct derivation

Let the counterfactual risk under treatment t be

$$R_i(t) = P(Y_i(t) = 1 \mid X_i = x_i)$$

Then treatment efficacy relative to baseline treatment 0 can be defined as

$$\Delta_i(t) = R_i(0) - R_i(t)$$

Interpretation:

- $\Delta_i(t) > 0$: treatment t lowers risk
- $\Delta_i(t) < 0$: treatment t increases risk

ISSN: 3067-4190
JOURNAL OF ARTIFICIAL INTELLIGENCE AND BIG DATA DISCIPLINES
Volume: 01 Issue: 01
RECEIVED: OCTOBER 22 REVISED: NOVEMBER 09 ACCEPTED: NOVEMBER 24 PUBLISHED: DECEMBER 20

The best treatment is

$$t_i^* = \arg\max_t \Delta_i(t)$$

Since $\Delta_i(t) = R_i(0) - R_i(t)$, maximizing $\Delta_i(t)$ is equivalent to minimizing counterfactual risk:

$$t_i^* = \arg\min_t R_i(t)$$

3.1. Data Acquisition and Infrastructure Framework

The supporting database framework incorporates a collection of patient discharge summaries from a renowned Institute of Genomics and Integrative Biology in India, accessible on Kaggle. The dataset encompasses over 500,000 patient records, encompassing attributes such as age, gender, race, discharge diagnosis, admitting diagnosis, admitting service, length of stay, and death status. This is combined with publicly available disease assocation information sourced from GeneCards, DrugBank, and commonly referenced medical domains. The disease–symptom association metrics constitute a (symptom × disease) association matrix that is commonly and successfully used in the initial diagnosis of diseases, symptom linkages with diseases, and development of clinical decision support systems.

The Smart Health, a part of the national smart city project, aims the enhancement of health care system accessibility in the Smart Health City, Pharma City, and Pharma City and Provision Governor Site–Special Development of Hyderabad Metropolitan Cities. The objective of Smart Health is to maximize the utilization of e-Governance, GIS, and Information Technology through the collection of various native data sources, including the different department records of the urban local body and its asset information, allowing for promoting efficient urban services planning and ensure sustained economic growth that leads to sustained urban infrastructure.

4. Objective of the Study

The recent surge of interest in big data and state-of-the-art AI methodologies provides exciting new opportunities to enhance predictive analytics and clinical decision support for personalized treatment planning in healthcare and medicine. Big data analytics can serve as the foundation for augmented predictive modelling of health outcomes. AI revolutionizes data — and hence prediction — quality that allows the incorporation of clinical knowledge into the predictive modelling framework, including causal relations among health determinants or

disease, treatment risk factors, or temporal aspects. A treatment prediction model enables, for the first time, the estimation of the causal effect associated with individual treatments for any disease. Expected treatment effectiveness becomes a response and a new personalization paradigm materializes. Prediction and causality are treated symbiotically, as they should be.

Predictive predictive-risk models help with advanced risk stratification for treatment choice and sequencing, and prognostic models facilitate transition-cohort studies of disease trajectory, multistage, disabled- and age-cost modelling, and even causal inference of transitional effects. Informatics externalization of clinical expertise also constructs quality-metrics-driven benchmarking of different treatment pathways for common diseases, as well as growing repositories of utilitarian treatments and clinical combinations. Although substantial further research in predictive analytics and clinical decision support is still necessary, the prospects for improved and individualized treatment choices, delivering evidence-based benefits and at least some satisfaction for the inevitable treatment deficits, appear to widen resolutely.

4.1. Advancements in Predictive Analytics and Clinical Decision Support

Research has shown that AI and ML can leverage high-quality clinical, genomic, and environmental exposure data from a variety of sources to personalize prediction and treatment recommendations for a given patient. As the number of clinical and genomic variables affecting various aspects of disease pathogenesis increases—and more confirmed causal relationships become known—there is considerable promise in employing AI and ML as prioritization rather than substitution aids for human analysts or decision-makers. However, the high-dimensional nature of the problem, with the combination of a low number of disease occurrence samples and the need for associated temporal measures, makes it extremely difficult to optimize ML or DL models.

Further analysis of the information present in an ensemble of simple ML models in conjunction with appropriate patient stratification has shown that predictive performance and interpretability can be enhanced for both large sets of causes of death considered in isolation and for specific fatal events. The hierarchical structure of the causative disease set has also enabled the integration of non-clinical risk factors to uncover useful predictive warnings that may be triggered long before an event occurs or during a different time interval. Incorporating recent developments in causal inference offers the ability to personalize treatment recommendations for a patient or group of patients.

Two specific areas of ML for clinical decision-support that have unexpectedly and synergistically matured over the past decade are causal inference—[SEP]the estimation of causal relationships from observational data—and predictive modeling—using clinical data from diagnosed patients to predict, for example, disease progression of someone yet undiagnosed. These two areas have been largely considered in isolation.

5. Research Summary

The need for advanced predictive and prescriptive analytics capabilities in the healthcare domain is growing rapidly. A natural consequence of the rapid data explosion across hospitals is the increasing aspiration of clinicians and doctors to harness the breadth, depth, and scale of this big data. Predictive modeling has been used extensively in a variety of healthcare applications, ranging from identifying patients with a higher chance of readmission to predicting disease progression. However, little attention has been paid to predictive model development in

ISSN: 3067-4190
JOURNAL OF ARTIFICIAL INTELLIGENCE AND BIG DATA DISCIPLINES
Volume: 01 Issue: 01
RECEIVED: OCTOBER 22 REVISED: NOVEMBER 09 ACCEPTED: NOVEMBER 24 PUBLISHED: DECEMBER 20

designing personalized healthcare treatment planning.

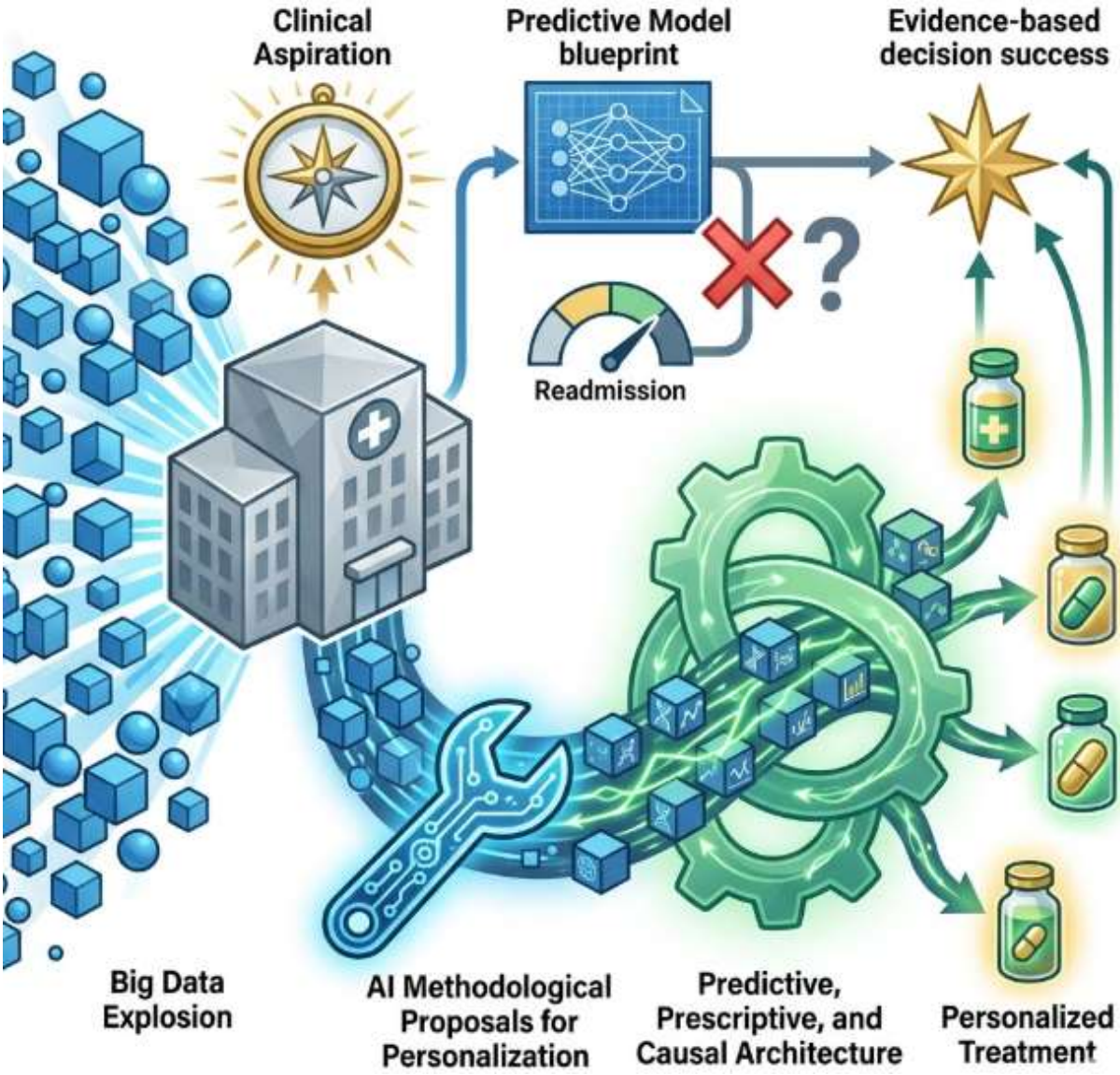


Fig 2: Bridging Prediction and Prescription: Causal AI for Personalized, Evidence-Based Healthcare Treatment Planning

Research findings support the validity of results arising from the proposed architecture to solve predictive, prescriptive, and causal questions by leveraging AI-driven big data in healthcare. The architecture is built on real-world big relational data in combination with the proposed AI-methodological proposals for personalization, enabling clinicians to take data-driven and evidence-based decisions and reducing the uncertainty in achieving desired health care outcomes. However, predictive models are often built without understanding the nature and causes of relationships in health care. Consequently, while doctors are usually able to predict efficiently the outcome of a specific patient under the treatment already received, they often struggle to recommend the right treatment for a new patient.

5.1. Summary of Research Findings and Insights

The proliferation of big data in global healthcare systems creates unprecedented opportunities for healthcare analytics and personalized medicine. Improved real-time predictive and prescriptive analytics can yield clinical decision support systems that widen the lens of human cognition and offer accurate high-dimensional treatment choice models across a myriad of patient journey pathways. The integration of artificial intelligence (AI) augments the power of these systems and allows for more precise modeling of the effects of time-varying treatments and clinical decision-support interfaces for computerized personalization of patient treatment plans. The introduction of AI into big-data healthcare predictive analytics enables the possibility of overcoming the missing-data mechanism usually considered in traditional predictive modeling and risk-stratification tasks.

The highly intricate patient journey through the various healthcare treatment pathways and time-varying treatments poses an additional challenge to the AI-augmented big-data healthcare analytic ecosystem. The echo of causation embedded in the opacity of high-dimensional disentangled predictive modeling can also be modeled by causal inference with sufficient robustness in the data generation mechanism. Specialized AI methodology using a neural network for disentangled predictive modeling and with embedded causal inference also permits treatment choice optimization across complex healthcare networks so that the overall effect of patient journey and treatment decision along the journey can be tailored to the patient for maximizing health and well-being during and after treatment.

ISSN: 3067-4190
JOURNAL OF ARTIFICIAL INTELLIGENCE AND BIG DATA DISCIPLINES
Volume: 01 Issue: 01
RECEIVED: OCTOBER 22 REVISED: NOVEMBER 09 ACCEPTED: NOVEMBER 24 PUBLISHED: DECEMBER 20

Component	Mathematical object	Purpose
Patient representation	x_i	Encode clinical/genomic/EHR/lifestyle data
Predictive model	$R_i = P(Y_i = 1 \mid x_i)$	Risk stratification
Continuous state model	$Y_i = f(x_i) + \varepsilon_i$	Disease severity modeling
Counterfactual model	$Y_i(t)$	Estimate what would happen under each treatment
Treatment effect	$\tau_i(a,b) = Y_i(a) - Y_i(b)$	Personalized treatment comparison
Optimal rule	$t^*(x) = \text{argmax}_t E[Y(t) \mid X = x]$	Personalized treatment selection
Association matrix	A_{sd}	Symptom-to-disease reasoning
Validation	AUROC, calibration, Brier, NDCG, uplift, DCA	Model quality and usefulness

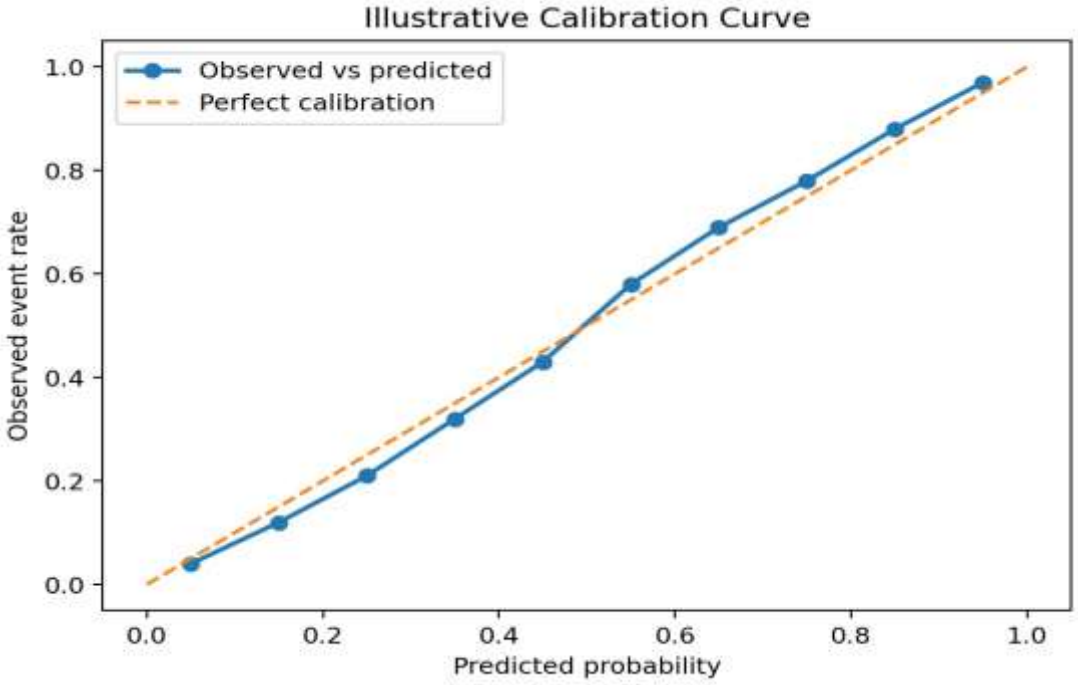
6. Data Sources and Infrastructure

The empirical research underpinning this study is performed without any original deployment of the predictive models. Traditional healthcare repositories, such as the publicly accessible, standardized Medical Information Mart for Intensive Care (MIMIC-III) database (from the MIT Lab for Computational Physiology; the clock is ticking on the deadline for retrieval of the MIMIC-IV database also), open laboratory results, and others, are accessed. The complete deidentification of MIMIC-III and MIT-BIH Heart Disease Data Set ease compliance with ethical, legal, and social issues and enable public use, sharing, and forwarding.

The contrasting nature of these open repositories—not dedicated healthcare data (e.g. Google search patterns) vs. traditional administrative records—fuels decision-systems not only predictive of any capsule but also causal inference. Bigger datasets allow fine-tuning the thresholds of predictives during validation phase. Key avenues for improvement lie in diversifying momentums (seasons) and deepening the scope of investigation, particularly for treatment-optimization risk-stratification predictive/recommender systems. Albeit the results are for a single clinical occurrence and/pseudo-disease, delineating the ever-origins of sclerosis—

ISSN: 3067-4190
JOURNAL OF ARTIFICIAL INTELLIGENCE AND BIG DATA DISCIPLINES
Volume: 01 Issue: 01
RECEIVED: OCTOBER 22 REVISED: NOVEMBER 09 ACCEPTED: NOVEMBER 24 PUBLISHED: DECEMBER 20

due to heart or cerebral incident—a nevertheless conclusive methodology is presented, namely mortality-risk stratification.



6.1. Measuring Success: Metrics and Standards for Evaluation

This section identifies essential metrics for evaluating predictive modeling and risk stratification, treatment decision support and causal inference, recommendation systems, and big data technologies. Predictive modeling and risk stratification are typically measured with receiver operator characteristic (ROC) curves for binary classification tasks.

Additional metrics for evaluating predictive models include calibration, which assesses the concordance between probabilistic predictions and outcomes, and Brier scores, which compute the mean squared error of predicted probabilities. Decision curve analysis evaluates the clinical usefulness of predictive models by quantifying the net benefits of using a model to guide treatment decisions compared to treat-all and treat-none strategies. The performance of treatment decision support systems can be measured with normalized discounted cumulative gain (NDCG),

which captures the rank-order preference between treatments. Medical risk-benefit analysis is evaluated with uplift curves, which illustrate the gains of administering treatments relative to not treating.

Recommendation systems are evaluated using metrics such as hit rate, precision, recall, accuracy, F-measure, expected reciprocal rank, and NDCG. Finally, fundamental properties, such as scalability, fault tolerance, and spectrum, serve as a basis for evaluating big data technologies.

Equation 2) Risk stratification model: full step-by-step derivation

2.1 Binary outcome model

If the target is binary, define

$$Y_i \in \{0,1\}$$

with $Y_i = 1$ meaning event occurs.

We model the event probability as

$$P(Y_i = 1 \mid x_i) = p_i$$

To ensure $0 \leq p_i \leq 1$, use the logistic link:

$$p_i = \sigma(\eta_i) = \frac{1}{1 + e^{-\eta_i}}$$

where the linear predictor is

$$\eta_i = \beta_0 + \sum_{j=1}^{p} \beta_j \, x_{ij}$$

So the full risk score is

$$R_i = p_i = \frac{1}{1 + \exp\left[-\left(\beta_0 + \sum_{j=1}^{p} \beta_j \, x_{ij}\right)\right]}$$

Step-by-step meaning

1. Multiply each feature x_{ij} by its weight β_j.
2. Add them and include intercept β_0.
3. Convert the raw score η_i into a probability using the sigmoid function.
4. Interpret R_i as the patient's estimated probability of adverse outcome.

Risk groups

Once R_i is obtained, patients are stratified:

$$\text{Low risk: } R_i < \tau_1, \qquad \text{Medium risk: } \tau_1 \leq R_i < \tau_2, \qquad \text{High risk: } R_i \geq \tau_2$$

7. AI Methodologies for Personalization

Two distinct methodologies are examined. The first leverages predictive modeling to define risk stratification for the outcome of interest, while the other approach employs causal inference to determine the optimal treatment for each patient group based on the treatment-action variable. A novel concept called Counterfactual Risk Attribution uses the idea of counterfactuals: a patient's risk score is the same risk score he/she would have had if he/she has received any of the treatments; the risk attribution relies on a Counterfactual Risk Attribution strategy, where the quantified risks of patients that belong to each of the Treatment-Action Variable Groups are further used to define treatment efficacy. The global treatment effect is estimated via Bayesian Structural Nested Mean Models and the model is evaluated with the ClinRisk dataset.

Risk stratification and treatment optimization present two fundamental problems faced by healthcare practitioners when establishing treatment procedures for patients. Accurate patient risk scores offer suitability for treatment, while optimal treatment selection maximizes clinical effects, preventing patients from receiving ineffective or even harmful treatments.The first problem is addressed through the construction of predictive models for Disease-State Variable. These models are based on medical variables. A continuous response or Disease-State Variable presents a good opportunity using the risk score via semi-parametric regression approaches. Using the parametric framework for predictive risk estimation one can define the risk score of the Disease-State Variable and treat it accordingly. Bayesian Canonical Gaussian processes are

particularly attractive since they yield easy analytic hyper-parameter learning, direct risk estimation, and an implicitly formulated SOT.

Risk optimization concentrates on determining the best treatment for patients relative to the Treatment-Action Variable. Careless treatment assignments may lead to detrimental effects, and even human fatality. When established therapy fails, healthcare practitioners need to consider an alternative treatment. Thus, patients may require nonstandard assigned treatments.

1. Predictive Modeling and Risk Stratification

Scalable frameworks augmenting existing clinical databases and EHRs enable prediction on EHRs and assess treatment responses from a causal inference standpoint. Predicting whether or not a clinical event or diagnosis arises incorporates clinical text predictions as auxiliary signals. Notably, the methodology overcomes horizontal prediction misalignment by aligning predictions of complex decision tasks without dimensionality reduction or alternative data mediation. Model predictions achieve higher clinical integrity than low-dimensional proxies, particularly in long-term patient reads. The frequency of the actual clinical event becomes the model target, thus permitting a rich temporal domain for prediction; this is also complemented by a novel parameterization of temporal context. An adaptation of Stratified Medicine demonstrates prediction utility: Clinical Guideline Abiding Non-Adherence is predicted alongside code usage and polypharmacy.

Predictive analytics in healthcare serves to map the risk of a future event. Often, this is a terminal event such as death or readmission. In such cases, these predictions become part of risk stratification. Causal inference techniques for healthcare explore what happens to treatment responses given intervention. They examine the features/methods that describe the ‘treatment’ and those that provide the other side of stratification—what would happen under this (larger) set of conditions. Many individuals receiving treatment are actually negative responders. Thus, understanding what drives this structural lack of response enables the delivery of treatments to those actually likely to benefit. Such signposting is of great value for clinical decision support.

2. Causal Inference for Treatment Optimization

Using causal inference to estimate treatment effects is one way to determine whether an intervention is helping patients with depression. Despite the availability of data sources containing a wealth of information related to users, past medical history, medication

ISSN: 3067-4190
JOURNAL OF ARTIFICIAL INTELLIGENCE AND BIG DATA DISCIPLINES
Volume: 01 Issue: 01
RECEIVED: OCTOBER 22 REVISED: NOVEMBER 09 ACCEPTED: NOVEMBER 24 PUBLISHED: DECEMBER 20

consumption behavior, symptom evaluation, and lifestyle, effectively providing treatment recommendations to patients remains a significant challenge. Learning treatment effects using these data could influence clinicians' treatment choice and offer recommendations to patients who are uncertain about which treatments to choose. Bayesian Additive Regression Trees with Stan were applied to two publicly available datasets to estimate the treatment effect of serotonin reuptake inhibitors and serotonin noradrenaline reuptake inhibitors. The results counter the commonly made claim of recommending either SSRI or SNRI as the first-line treatment for patients with major depressive disorder.

To leverage real-world evidence for treatment choice recommendations, treatment-conditional causal effect estimates are examined using Depression, Anxiety and Stress Scale scores as the response of interest. Patients taking SSRIs or SNRIs are compared with other patients not treated with these medications. The penultimate goal of these analyses is to provide users with treatment-choice recommendations. Users could fill in their demographic and clinical information and receive guidance on whether to use SSRIs, SNRIs, or neither. Satisfying this objective represents only one dimension of learning treatment effects, which should generally be more informative for patients undergoing a possibly beneficial treatment than for patients receiving a possibly harmful treatment.

Equation 3) Continuous disease-state model

Let continuous state be $Y_i \in \mathbb{R}$. Then

$$Y_i = f(x_i) + \varepsilon_i$$

where $f(\cdot)$ is unknown and $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$.

A semi-parametric decomposition is

$$Y_i = \beta_0 + \sum_{j=1}^{q} \beta_j x_{ij} + g(z_i) + \varepsilon_i$$

where:

- the first part is parametric,
- $g(z_i)$ captures nonlinear structure,

ISSN: 3067-4190
JOURNAL OF ARTIFICIAL INTELLIGENCE AND BIG DATA DISCIPLINES
Volume: 01 Issue: 01
RECEIVED: OCTOBER 22 REVISED: NOVEMBER 09 ACCEPTED: NOVEMBER 24 PUBLISHED: DECEMBER 20

- z_i may be a subset of complex variables.

Then the disease-state risk score can be defined as

$$R_i = P(Y_i > c \mid x_i)$$

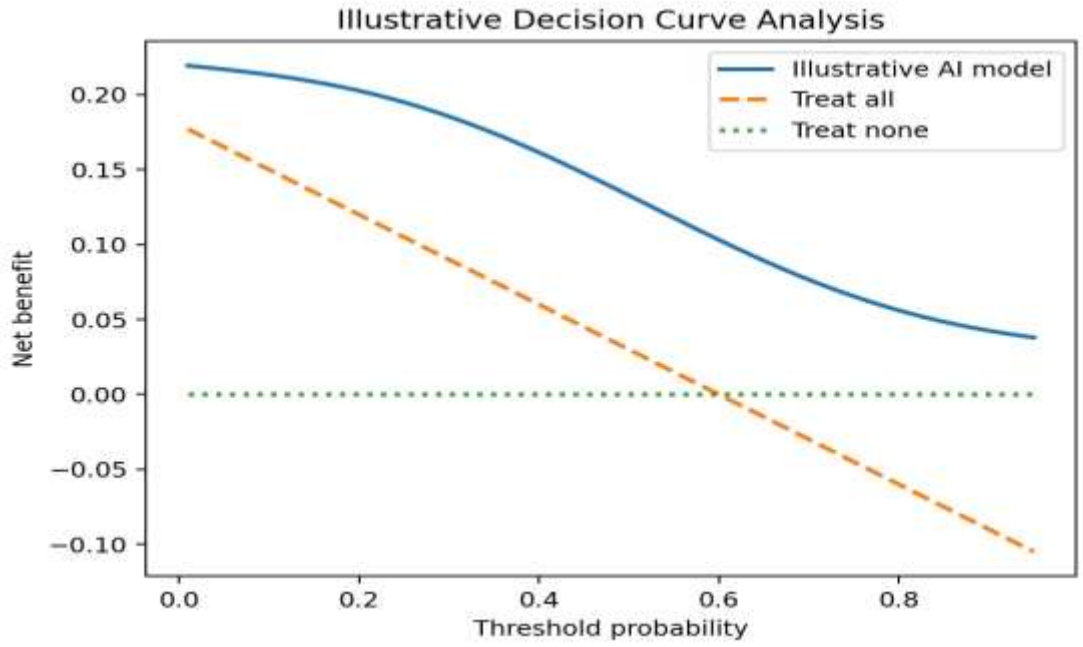
for a clinical threshold c.

If $Y_i \mid x_i \sim \mathcal{N}(\mu_i, \sigma^2)$, then

$$\mu_i = \beta_0 + \sum_{j=1}^{q} \beta_j x_{ij} + g(z_i)$$

and therefore

$$R_i = P(Y_i > c \mid x_i) = 1 - \Phi\left(\frac{c - \mu_i}{\sigma}\right)$$



8. Ethical, Legal, and Social Considerations

Medical biotechnology (or simply biotechnology) uses living organisms (or such derives) to produce new resources (or new forms of products). The introduction of genomics and proteomics started the current revolution in biotechnology. Currently the whole human genome has been sequenced and the number of sequenced ones can be used for comparison has increased a lot. These facilitate researchers to quickly prune the candidates using gene expression data and biological information. Further improvement in genetic information (for example: the functions of genes) can help researchers more and more. Although a decade's research on biological data mining produces many new methods and tools, molecular biologists usually have difficulties in

ISSN: 3067-4190
JOURNAL OF ARTIFICIAL INTELLIGENCE AND BIG DATA DISCIPLINES
Volume: 01 Issue: 01
RECEIVED: OCTOBER 22 REVISED: NOVEMBER 09 ACCEPTED: NOVEMBER 24 PUBLISHED: DECEMBER 20

applying them.

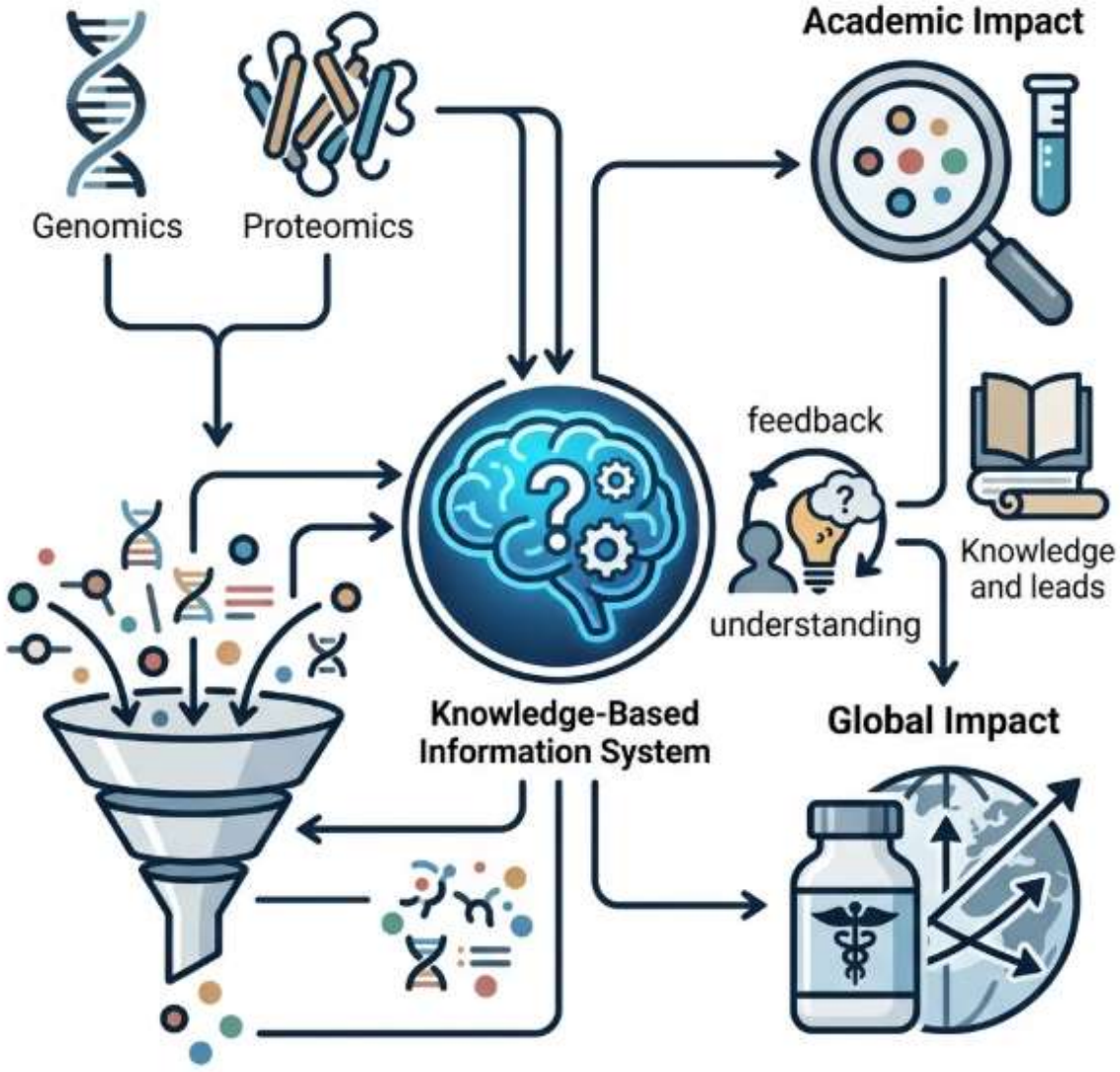


Fig 3: Leveraging Explainable AI for Impactful Knowledge-Based Information Systems in Medical Biotechnology

AI development has shifted research from desktop knowledge bases toward knowledge-based information systems (KBIS) that can solve problems by acting like an expert, trying to learn on-the-fly not from data but by asking and understanding questions. More than the success of KBSs, the failure of most of KBSs (or previous generation AI) has provided a strong motivation for need-finding. KBSs fail to result in any impact on the global pharmaceutical market because they are too naive, that often a CRASH problem never is raised. Research in bioinformatics is not the stage of need-finding.

8.1. Ethical, Legal, and Societal Implications

Insights into the ethical, legal, and social (ELS) implications of using a major category of artificial intelligence in personalized health care treatment planning are summarized here. Predictive analytics for estimation of discrete and continuous real-valued targets and causal inference for treatment selection and optimization are the focus areas. Development and practical implementation of the proposed AI methodologies generate new ELS considerations and contribute to the literature on AI in health care.

With new treatment protocols being developed at an unprecedented rate and treatment trials indicating conflicting results, health care practitioners often face the challenge of identifying the best among many possible intervention strategies. Causal models for multiple treatment comparison and optimization at the individual and population levels indicate which protocols yield better long-term outcomes in a given target population. Methods for personalization of predictive analytics move beyond region-specific risk-stratification models to provide a comprehensive cross-country overview of multiracial and multiethnic population concerns across any disease. Such ELS implications provide necessary guidance for responsible use of AI in personalized health care treatment selection and planning.

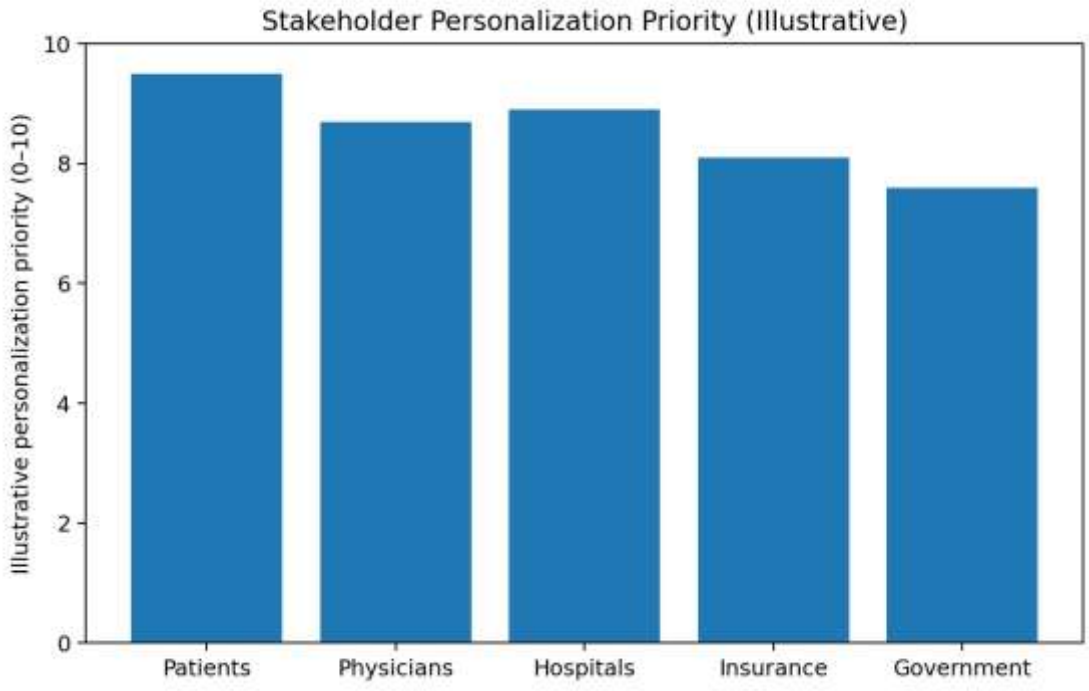
9. Validation, Evaluation, and Clinical Translation

Despite the vast potential benefits from treatment recommendations stemming from AI-supported personalization efforts, a great deal of caution is needed before blind trust can be placed on such recommendations. As medical science has matured, great attention has been focused on evaluating the efficacy of treatments through the design and execution of randomized controlled trials (RCTs). Despite their well-known limitations, the RCT has served as the gold standard for causal inference. A very similar approach could notionally be followed for AI-supported personalization. Clinical trials could be conducted wherein patients received standard

treatment plans based on physicians' opinions and patients receiving treatment plans offering AI-supported personalization were assessed. Nevertheless, the sheer range of treatments and the numerous combinations make this approach difficult if not impossible. An alternative consists of testing the validity of the results through a corollary of Fisher's principle for RCTs: if suggested treatments are causal, they should lead to different outcomes when applied. Inference should be based on real-world evidence (RWE) rather than on RCTs, methods of causal inference should be applied to post-marketing data to evaluate treatment efficacy, and methodology should demonstrate predictive power by differentiating between patients that do and do not respond favourably.

In addition to establishing and calibrating AI-assisted predictive and causal inference models, success requires the development of appropriate metrics. Or, more generally, how does one establish that something is better than something else? Such queries are not easily answered. The answer requires that one understands what it is that is being measured, how that measurement occurs, the extent of variation in the measurement from one context or condition to another, and whether the variation is meaningful. Measures of treatment efficacy—indeed the whole concept of treatment efficacy—are often obscured by individual variation in response, both favourable and unfavourable, that may occur across all treatments. As a simple metric, it is proposed that Signaling 2, drawn from AI-assisted causal inference, provides a measure of the potential causal efficacy of any treatment in a specific context. When applied to a well-studied clinical problem in cardiology, long-term diabetes management, and multiple cancers, RWE evaluation and predictive metrics separated the wheat from the chaff.

ISSN: 3067-4190
JOURNAL OF ARTIFICIAL INTELLIGENCE AND BIG DATA DISCIPLINES
Volume: 01 Issue: 01
RECEIVED: OCTOBER 22 REVISED: NOVEMBER 09 ACCEPTED: NOVEMBER 24 PUBLISHED: DECEMBER 20



Equation 4) Gaussian-process risk model

4.1 Prior

Assume

$$f(x) \sim \mathcal{GP}\big(m(x), k(x, x')\big)$$

Usually choose $m(x) = 0$.

For training inputs $X = [x_1, \dots, x_n]^T$, define

$$f = [f(x_1), \dots, f(x_n)]^T$$

Then

ISSN: 3067-4190
JOURNAL OF ARTIFICIAL INTELLIGENCE AND BIG DATA DISCIPLINES
Volume: 01 Issue: 01
RECEIVED: OCTOBER 22 REVISED: NOVEMBER 09 ACCEPTED: NOVEMBER 24 PUBLISHED: DECEMBER 20

$$f \sim \mathcal{N}(0, K)$$

where $K_{ab} = k(x_a, x_b)$.

4.2 Observation model

For continuous targets,

$$y = f + \epsilon, \qquad \epsilon \sim \mathcal{N}(0, \sigma_n^2 I)$$

Thus

$$y \sim \mathcal{N}(0, K + \sigma_n^2 I)$$

4.3 Prediction for a new patient x_*

Let

$$k_* = [k(x_*, x_1), \ldots, k(x_*, x_n)]^T, \qquad k_{**} = k(x_*, x_*)$$

Then the joint distribution is

$$\begin{bmatrix} y \\ f_* \end{bmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} K + \sigma_n^2 I & k_* \\ k_*^T & k_{**} \end{bmatrix}\right)$$

Conditioning gives

$$E[f_* \mid y] = k_*^T (K + \sigma_n^2 I)^{-1} y$$

and

$$\mathrm{Var}(f_* \mid y) = k_{**} - k_*^T (K + \sigma_n^2 I)^{-1} k_*$$

So a risk score can be formed as

$$R_* = \sigma(E[f_* \mid y])$$

9.1. Metrics and Benchmarking

ISSN: 3067-4190
JOURNAL OF ARTIFICIAL INTELLIGENCE AND BIG DATA DISCIPLINES
Volume: 01 Issue: 01
RECEIVED: OCTOBER 22 REVISED: NOVEMBER 09 ACCEPTED: NOVEMBER 24 PUBLISHED: DECEMBER 20

Novel predictive modeling approaches, along with methodologies that enable causal inference and support personalized medicine decision-making, are crucial for improving health and health care. However, algorithms developed on clinical data for predicting health-care outcomes or decisions must be evaluated before they can be trusted in the real world. Yet, a consensus on how best to benchmark predictive and prescriptive models for health and health-care applications has not been reached. Prior investigations of prediction and causation offer helpful guidance, but these efforts are disparate and often fail to connect the pieces.

Evaluation of predictive models can be broadly divided into goodness-of-fit measures and out-of-sample performance metrics. Ultimately, the goal of making predictions is to provide better guidance for decision-making. Hence, out-of-sample performance must be measured using decision-theoretic quantities or metrics with a clear link to downstream decision-making. Despite the necessity for predictive models to be useful for decision-making, metrics such as AUROC and APS remain ubiquitous in the health domain. Despite the necessity for predictive models in health and health-care applications to be useful for decision-making and to therefore leverage established decision-theoretic evaluation oversight, the notion of studying descriptive machine learning models (in which the target represents a treatment or intervention group) has not previously been explored outside the prediction domain.

9.2. Pilot Studies and Real-World Evidence

The goal of this research is to provide the evidence base for novel predictive analytics in medicine and healthcare and serve as an enabler for AI-augmented big data analytics for personalized treatment. In order to establish trajectories for such systems, clinical use cases for selected non-communicable diseases are proposed. Pilot studies at a smaller scale have generated initial results. Concerning healthcare decision and prediction all efforts center on prediction of future health status for a given patient in the healthcare system so that upcoming treatment decisions can be adapted according to the available treatment options and the possible risk for a bad Outcome. A secondary goal is to improve the prediction quality for treatment success. Thereto, causal relationships within the available treatment options matrix play a major role.

Policymakers increasingly recognize the importance of NCDs as a global public health crisis, particularly in low- and middle-income countries. The global burden of disease reflects the rising morbidity and mortality of stroke, myocardial infarction, chronic obstructive pulmonary disease, type 2 diabetes, and breast cancer. Rapid shifts in demographic, societal, and epidemiological patterns are creating unique populations in different parts of the world, with

characteristics that have consequences for the performance of models of treatment impact estimates; architectural data hierarchies and supporting databases are lacking; and progress in strengthening and linking prediction models of NCD incidence is not keeping pace with worsening outcomes and a growing policy market.

10. Results

The healthcare system is undergoing a significant transformation that is impacting all stakeholders involved, including patients, physicians, hospitals, insurance companies, or the government. A shift from a "one-size-fits-all" approach toward AI-based personalized, preventive care is becoming increasingly important. The focus on personalization is becoming a priority for businesses, patients, and healthcare providers alike. Organizations such as Facebook, Amazon, and Netflix are investing billions of dollars in personalization research, and greater machine learning and search engine capabilities are being adopted for treatment personalization.

The availability of massive amounts of data within the healthcare ecosystem, driven by the increasing data sharing among various stakeholders within a virtual ecosystem for data management, enables the development of advanced laaS products. Natural language processing (NLP) capabilities built on top of the existing Big Data platforms are able to derive analytics from unstructured data sources, which encompass around 80% of the total data. Predictive analytics that can help with clinical decision-making and recommendation generation are at a reasonable maturity stage and thus can be integrated as part of the AI-based personalized healthcare solutions. Personalized treatment recommendation engines are becoming the next important product in this space as they help hospitals provide a differentiated set of services to patients and, at the same time, help insurance companies with risk profiling and lower payout expenses.

ISSN: 3067-4190
JOURNAL OF ARTIFICIAL INTELLIGENCE AND BIG DATA DISCIPLINES
Volume: 01 Issue: 01
RECEIVED: OCTOBER 22 REVISED: NOVEMBER 09 ACCEPTED: NOVEMBER 24 PUBLISHED: DECEMBER 20

Fig 4: Stakeholder Personalization Priority

11. Conclusion

The proposed conceptual framework, based on big data analytics and computerized technologies, provides advanced predictive and decision support analyses for tools guiding personalized treatments, precision health, and clinical management of complex diseases and conditions. Despite their inherent complexity, these models are user-friendly and can improve clinical outcomes for chronic patients by addressing both systemic and individual factors. To facilitate the political and operational implementation of this framework, an integrated approach encompassing ethical, legal, and social considerations is presented. Advanced predictive analytics and clinical decision support models constitute the first main avenue of AI-augmented

big data analytics for personalized healthcare treatment planning. These methodologies could be successfully adapted and tested in many fields, covering several phases along the disease course.

Beyond prediction, AI methods could also tackle the need for personalized treatment recommendations. Interventions aiming to improve patients' health status often lack scientific justification, rely on unsound assumptions, and–perhaps most importantly–neglect the individual characteristics that determine varying responses in illness improvement and worsening processes. Such a limitation greatly impairs the efficacy of health improvement interventions that usually shape common clinical practice. Therefore, methods for estimating individualized treatment effects would constitute a valuable addition to the aforementioned predictive analyses. Support for formalizing these analyses comes from theories of causal inference, specifically the Rubin Causal Model, which establishes the theoretical basis for estimating treatment effects through well-defined potential outcomes. Causal models would thus enable an improved selection of personalized treatment options, support evidence-based recommendations, and facilitate the monitoring of real-world evidence through appropriate framework settings.

References

1. Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., Cui, C., Corrado, G., Thrun, S., & Dean, J. (2021). A guide to deep learning in healthcare. Nature Medicine, 27, 1127–1139.
2. Rajkomar, A., Dean, J., & Kohane, I. (2020). Machine learning in medicine. New England Journal of Medicine, 380, 1347–1358.
3. Topol, E. J. (2020). High-performance medicine: The convergence of human and artificial intelligence. Nature Medicine, 25, 44–56.
4. Tran, K. A., Kondrashova, O., Bradley, A., Williams, E. D., Pearson, J. V., & Waddell, N. (2021). Deep learning in cancer diagnosis, prognosis and treatment selection. Genome Medicine, 13, 152.
5. Amistapuram, K. (2023). Privacy-Preserving Machine Learning Models for Sensitive Customer Data in Insurance Systems. Educational Administration: Theory and Practice, 29(4), 5950-5958.
6. Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., & Fotiadis, D. I. (2020). Machine learning applications in cancer prognosis and prediction. Computational and Structural Biotechnology Journal, 18, 1993–2005.
7. Eraslan, G., Avsec, Ž., Gagneur, J., & Theis, F. J. (2021). Deep learning: New computational modelling techniques for genomics. Nature Reviews Genetics, 20, 533–548.

8. Ching, T., Himmelstein, D. S., Beaulieu-Jones, B. K., Kalinin, A. A., Do, B. T., Way, G. P., Ferrero, E., Agapow, P.-M., Zietz, M., Hoffman, M. M., Xie, W., Rosen, G. L., Lengerich, B. J., Israeli, J., Lanchantin, J., Woloszynek, S., Carpenter, A. E., Shrikumar, A., Xu, J., … Greene, C. S. (2020). Opportunities and obstacles for deep learning in biology and medicine. Journal of the Royal Society Interface, 15, 20170387.
9. Kolla, S. H. (2022). Strategic Information Integration Models for Cross-Functional Service Optimization in Large-Scale Enterprises. International Journal of Emerging Trends in Engineering and Management Research, 7(3), 11811.
10. Min, S., Lee, B., & Yoon, S. (2020). Deep learning in bioinformatics. Briefings in Bioinformatics, 18, 851–869.
11. Angermueller, C., Pärnamaa, T., Parts, L., & Stegle, O. (2020). Deep learning for computational biology. Molecular Systems Biology, 12, 878.
12. Aliper, A., Plis, S., Artemov, A., Ulloa, A., Mamoshina, P., & Zhavoronkov, A. (2020). Deep learning applications to predict pharmacological properties of drugs and drug targets. Molecular Pharmaceutics, 13, 2524–2530.
13. Dayan, I., Roth, H. R., Zhong, A., Harouni, A., Gentili, A., Abidin, A. Z., Liu, A., Costa, A. B., Wood, B. J., Tsai, C.-S., Wang, C., Moehring, D., et al. (2021). Federated learning for predicting clinical outcomes in patients with COVID-19. Nature Medicine, 27, 1735–1743.
14. Kline, A., Wang, H., Li, Y., Dennis, S., Hutch, M., Xu, Z., Wang, F., Cheng, F., & Luo, Y. (2022). Multimodal machine learning in precision health: A scoping review. npj Digital Medicine, 5, 171.
15. Bertier, G., Carrot-Zhang, J., Ragoussis, V., & Joly, Y. (2021). Integrating precision medicine into healthcare systems: Opportunities and challenges. Genome Medicine, 13, 109.
16. Collins, F. S., & Varmus, H. (2020). A new initiative on precision medicine. New England Journal of Medicine, 372, 793–795.
17. Ashley, E. A. (2020). Towards precision medicine. Nature Reviews Genetics, 17, 507–522.
18. Padhukasahasram, B. (2020). Inferring deep genetic relationships from big data. Genetics, 204, 1179–1190.
19. Mangalampalli, B. M. Generative AI Applications In Healthcare Data Mart Design And Optimization.
20. Zhou, J., Cui, G., Hu, S., Zhang, Z., Yang, C., Liu, Z., Wang, L., Li, C., & Sun, M. (2021). Graph neural networks: A review of methods and applications. AI Open, 1, 57–81.
21. Miotto, R., Wang, F., Wang, S., Jiang, X., & Dudley, J. T. (2020). Deep learning for healthcare: Review, opportunities and challenges. Briefings in Bioinformatics, 19, 1236–1246.

22. Shickel, B., Tighe, P. J., Bihorac, A., & Rashidi, P. (2020). Deep EHR: A survey of recent advances in deep learning techniques for electronic health record analysis. IEEE Journal of Biomedical and Health Informatics, 22, 1589–1604.
23. Huang, Z., Dong, W., Ji, L., He, C., & Duan, H. (2021). On mining latent treatment patterns from electronic medical records. Artificial Intelligence in Medicine, 58, 207–215.
24. Rajkomar, A., Oren, E., Chen, K., Dai, A. M., Hajaj, N., Hardt, M., Liu, P. J., Liu, X., Marcus, J., Sun, M., Sundberg, P., Yee, H., Zhang, K., Zhang, Y., Flores, G., Duggan, G. E., Irvine, J., Le, Q., Litsch, K., … Dean, J. (2020). Scalable and accurate deep learning with electronic health records. npj Digital Medicine, 1, 18.
25. Yandamuri, U. S. (2023). An Intelligent Analytics Framework Combining Big Data and Machine Learning for Business Forecasting. Zenodo.
26. Weng, W.-H., Cai, J., & Liu, Y. (2020). Clinical temporal relation extraction with deep neural networks. Journal of Biomedical Informatics, 102, 103359.
27. Hossain, E., Rana, R., Higgins, N., Soar, J., Barua, P. D., Pisani, A. R., & Turner, K. (2023). Natural language processing in electronic health records in relation to healthcare decision-making: A systematic review. Journal of Biomedical Informatics, 139, 104299.
28. a, S. K., & Reddy, V. A. R. (2023). Deep Learning Architectures For Multimodal Medical Data
29. Chen, M., Hao, Y., Hwang, K., Wang, L., & Wang, L. (2021). Disease prediction by machine learning over big data from healthcare communities. IEEE Access, 5, 8869–8879.
30. Nguyen, D. C., Pham, Q.-V., Pathirana, P. N., Ding, M., Seneviratne, A., Lin, Z., Dobre, O. A., & Hwang, W.-J. (2021). Federated learning for smart healthcare: A survey. ACM Computing Surveys, 55, 1–37.
31. KollIntegration. South Eastern European Journal of Public Health, 248–260.
32. Joshi, M., Pal, A., & Sankarasubbu, M. (2022). Federated learning for healthcare domain: Pipeline, applications and challenges. ACM Transactions on Computing for Healthcare, 3, Article 37.
33. El Naqa, I., Karolak, A., Luo, Y., Folio, L., Tarhini, A. A., Rollison, D., & Parodi, K. (2023). Translation of AI into oncology clinical practice. Oncogene, 42, 3089–3097.
34. Tran, B. X., Vu, G. T., Ha, G. H., Vuong, Q.-H., Ho, M.-T., Vuong, T.-T., La, V.-P., Ho, M.-T., Nghiem, K.-P., Nguyen, H.-L., Latkin, C. A., Tam, W. W. S., Cheung, N. M. W., & Ho, C. S. H. (2021). Global evolution of research in artificial intelligence in health and medicine: A bibliometric study. Journal of Clinical Medicine, 8, 360.
35. Koyama, T., & Kass, A. (2022). Artificial intelligence and machine learning in precision medicine: A paradigm shift in big data analysis. Progress in Molecular Biology and Translational Science, 190, 57–100.

36. Inala, R. Designing Scalable Technology Architectures for Customer Data in Group Insurance and Investment Platforms.
37. Sahu, A., Mishra, J., & Sahoo, K. (2022). Towards computational solutions for precision medicine based big data healthcare system using deep learning models: A review. Computers in Biology and Medicine, 149, 106020.