

Chapter 7

Risk Stratification and Prognosis Using Predictive Modelling and Big Data Approaches



Shyam Visweswaran and Gregory F. Cooper

Introduction to Predictive Models in Medicine

Prediction is critical to many activities in clinical medicine, such as assessing risk of developing disease in the future (risk assessment and stratification), determining the presence or absence of disease at the current time (diagnosis), forecasting the likely course of disease (prognosis), and predicting treatment response (therapeutics) [1]. In addition to clinical medicine, prediction plays a critical role in public health and in biomedical research. Predictive models that are derived from data can improve predictions and help guide decision-making in clinical medicine and in public health. Often predictive models are probabilistic models that compute the prediction as a probability, and such models are typically estimated from data using statistical and, more recently, machine-learning methods.

The better the predictive models, the better the decisions and the ensuing outcomes are likely to be for the individual and for the public at large. Even small improvements in predictive performance can have meaningful impact on individual and public health outcomes and costs. The burgeoning field of precision and personalized medicine aims to tailor risk assessment, diagnosis, prognosis and therapeutics to the characteristics of individuals that go beyond those measured during routine clinical care. The goal is to deliver the right treatment at the right time to the right patient based on complex patient characteristics that may be obtained from a range of molecular, clinical, and environmental measurements.

Traditionally, predictive models in medicine have been developed from data such as clinical findings, laboratory test results, and findings from clinical imag-

S. Visweswaran (✉) · G. F. Cooper
Department of Biomedical Informatics and of Intelligent Systems, University of Pittsburgh,
Pittsburgh, PA, USA
e-mail: shv3@pitt.edu; gfc@cbmi.pitt.edu

ing studies. Recent advances in two areas are making available big biomedical data at an unprecedented scale for use in clinical medicine, public health, and research. Electronic health records (EHRs) are widespread and are capturing ever more clinical data. EHR data coupled with administrative claims data are increasingly used for characterization of disease progression and outcomes, comparative effectiveness of treatments, and predictive and prognostic modeling. Such observational healthcare data sets contain data on millions to tens of millions of patients and hold the promise of enabling research into less frequent diseases and outcomes. Another advance is the burgeoning use of low cost omics technologies, which is producing a rich base of high-throughput molecular data, such as genomic variant, gene expression, proteomic, and metabolomic data. Omics data in conjunction with EHR data hold the promise of better prediction of diseases before their occurrence, increased accuracy of diagnosis of complex diseases, and more precisely targeted therapies.

Examples of Applications

Predictive models have applications in the domains of clinical practice, in public health, and in biomedical research. Table 7.1 gives illustrative examples of the application of predictive models for risk assessment across all three domains.

In *clinical medicine*, both predictive models and clinical decision rules are useful in assisting clinicians with decision making. Predictive models generate probabilities but do not recommend actions and the interpretation of the probabilities is left to the clinician. Clinical decision rules, in addition, suggest actions based on probabilities generated by a predictive model. Risk assessment models are useful in evaluating risk for developing disease that informs the initiation of preventive measures. An example of a risk assessment model is the Framingham Risk Score that predicts the 10-year risk of developing coronary heart disease from age, total and HDL cholesterol, blood pressure, diabetes, and smoking status [2]. This score is used clinically to identify those at high risk and initiate life style changes and cholesterol lowering pharmacotherapy.

Table 7.1 Illustrative examples of prediction that guide decision-making

Domain	Prediction task	Decision to make
Clinical medicine	Will an individual have a heart attack in the coming year?	Prescribe aspirin or not
Public health	How many residents of a county will have a heart attack in a year?	Determine the number of paramedics to be trained to perform electrocardiograms in ambulances
Biomedical research	How many heart attacks are likely to occur in the control arm of a clinical trial?	Enroll fewer or more participants

Predictive models are also useful for deciding whether or not to perform diagnostic testing. When the probability of the presence of disease is relatively high, diagnostic testing is indicated to confirm or rule out disease, while if the probability is low, no immediate testing is indicated. For example, sepsis is a relatively rare but a life-threatening cause of infection, and the definitive diagnostic test is a blood culture to detect bacteremia (presence of viable bacteria in circulating blood). A clinical decision rule has been described to selectively perform blood cultures in Emergency Department patients who are predicted to be at high risk of bacteremia. Features of the history, co-existing illnesses, physical examination, and laboratory testing were used to create a clinical decision rule that consists of major and minor criteria, and blood culture is indicated if at least one major criterion or two minor criteria are present [3, 4].

Furthermore, predictive models are useful for selecting treatment such that the anticipated benefit exceeds the risk of harm. For example, in patients with atrial fibrillation, antithrombotic agents are effective in reducing the risk of stroke while concurrently increasing the risk of serious bleeding. Predictive models that estimate a patient's stroke risk and bleeding risk are useful in identifying the appropriate antithrombotic agent for which the reduction in the risk of stroke most strongly outweighs the increased risk of bleeding [5].

In *public health*, predictive models are useful in surveillance and forecasting of epidemics like influenza. Traditional surveillance that is provided, for example, by the Centers for Disease Control and Prevention (CDC) relies on clinical findings, virology laboratory results, hospital admissions, and mortality data. Newer digital surveillance employs sources such as over-the-counter retail sales of medications, social network activity, and internet search engine queries [6]. Such surveillance produces forecasts that assist health officials to inform public health actions and allocate resources.

In *biomedical research*, predictive models may be useful in selection and stratification of participants in terms of baseline as well as predicted characteristics for a study such as a clinical trial. This allows enrollment of more refined subgroups and improves statistical analyses. For example, a trial in traumatic brain injury may exclude patients with high likelihood of a poor outcome. A prognostic model that predicts 6-month mortality in traumatic brain injury can be used to select patients who have a small probability of mortality [7].

Prognostic Versus Predictive Factors

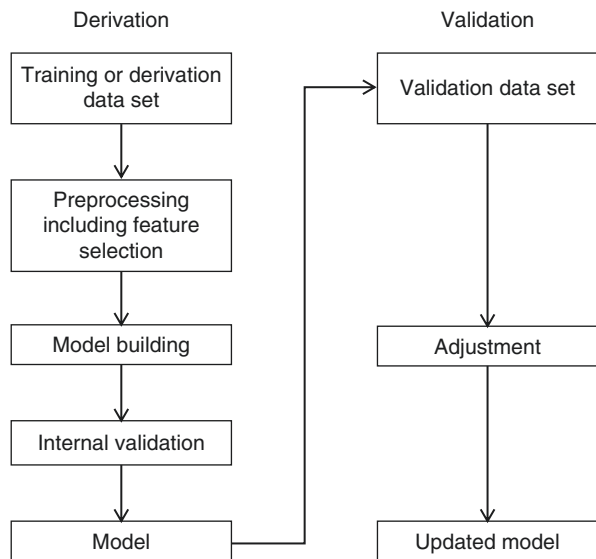
Some authors in the biomedical literature differentiate between prognostic and predictive factors or biomarkers. A prognostic factor is defined as a clinical or biological characteristic that is associated with a clinical outcome such as development or progression of disease, irrespective of the treatment. A predictive factor is defined as a characteristic that is associated specifically with response or lack of response to

a particular therapy [8, 9]. For example, a prognostic factor for primary breast cancer is any measurement available at the time of diagnosis or surgery that is associated with disease-free or overall survival in the absence of systemic adjuvant therapy, while a predictive factor is one that is associated with response or lack of response to systemic adjuvant therapy [8]. In this framework, a prognostic factor is predictive of a clinical outcome and a predictive factor is predictive of differences in response to a therapy. However, in this chapter, the terms prognostic and predictive are considered to be synonymous and denote the ability of a factor to predict outcomes.

Workflow of Development and Validation of Predictive Models

The development of predictive models in medicine consists of two phases, namely, derivation (or training) and validation (or external validation) [10, 11]. The workflow in the two phases is shown in Fig. 7.1. The derivation phase consists of collection of training or derivation data, preprocessing of the data that includes handling missing values and feature selection, building a multivariable model, and performing internal validation to assess the model’s predictive performance for discrimination and calibration. Internal validation is performed by splitting the data into training and test, by cross-fold validation or by leave-one-out validation. To perform cross-fold validation, data is partitioned into several equal parts; all parts except one are combined and the model is derived from it and evaluated on the left out part; this process is performed once for each part.

Fig. 7.1 Workflow of development of clinical predictive models. The two phases are derivation followed by validation. (Adapted from Fig. 1 in Ref. [10])



In the external validation phase, the predictive performance of the model is evaluated on data that is obtained independently of the derivation data. External validation is needed to assess generalizability of the model, such as temporal generalizability across data from different time periods, geographical generalizability across data obtained from different physical locations, and spectrum generalizability across data that possess differing disease severity or varying prevalence of the outcome. When external validation suggests that model performance needs to be improved, the model may have to be rebuilt or updated using the validation data. If the model is updated, the new model should be assessed for external validity on data that is obtained independently of the derivation and validation data before it is considered for deployment.

Emerging Informatics Methods

Challenges abound in the development of predictive models. This chapter focusses on four challenges and new approaches to surmount them. A critical challenge in the developing predictive models from big data is dimensionality reduction, which is the process of reducing the number of features in the data. Another challenge is the development of models that can not only adequately discriminate between individuals who will have an outcome and those who will not but also possess adequate calibration to predict accurately the actual risk of outcome [12]. For example, the European System for Cardiac Operative Risk Evaluation Score (EuroSCORE), a model to predict mortality from cardiac surgery, showed excellent discrimination but had poor calibration because it overestimated the risk of mortality in elderly patients (e.g., the model predicted mortality risk of 15% when the actual risk of dying after surgery was 8.8%) [13]. An updated version of the score called EuroSCORE II was developed to improve calibration [14]. A third challenge is developing models that perform well not only on the population as a whole but also perform well in the individual. Personalized modeling approaches can produce high performing and simpler models that are tailored to the individual. Finally, explanations for predictions produced by predictive models are necessary for real-world deployment.

Dimensionality Reduction

Two main approaches to dimensionality reduction are feature selection and feature extraction. Feature selection is the process of selecting a subset of the original set of features, to obtain a smaller subset, and feature extraction is the process of creating a new, smaller set of features from the original set of features. Thus, feature selection preserves a subset of the original features while feature extraction creates new ones.

In biomedical data sets where the number of features is in the tens of thousands or more, many of the potential predictor features are either redundant or irrelevant for predicting the target outcome. Predictive modeling techniques, including regression and classification methods, often perform poorly when all features are included, due to irrelevant features introducing noise. One approach is to preprocess the data set by selecting a reduced subset of features, and use that subset for predictive modeling. In addition to improving the predictive model's performance, feature selection reduces the computational cost and may provide better interpretability of the underlying processes that generated the data [15].

A good feature selection method identifies the smallest number of features that deliver maximal predictive performance. Feature selection methods can be broadly categorized into wrapper and filter methods. Wrapper methods evaluate feature subsets using the predictive model and select the best performing subset. Filter methods do not use the predictive model but instead apply statistical criteria to select the features and then construct the model with the selected features.

A filter-type feature selection approach that has been investigated extensively is based on identifying the Markov blanket of the outcome or target [16]. The Markov blanket of a target is defined as a minimal set conditioned on which all other measured features become independent of the target. A variety of Markov blanket discovery algorithms have been developed and evaluated on biomedical data [17].

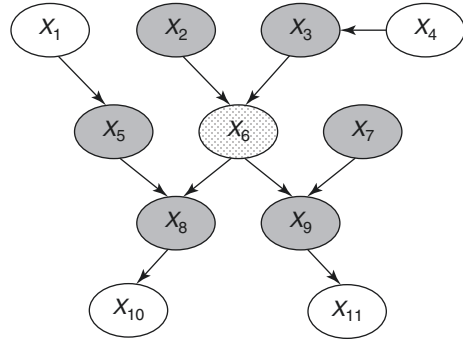
Markov Blanket Algorithms

A Bayesian network (BN) model is a graphical model that represents probabilistic relationships among a set of features X . A BN contains a graphical model structure that is a directed acyclic graph (DAG) that contains a node for every feature X_i and an arc between every pair of nodes if the corresponding features are directly probabilistically dependent. Conversely, the absence of an arc between a pair of nodes denotes probabilistic independence (often conditional) between the corresponding features. In addition, a BN contains a set of parameters θ that encode the probability distributions. In a BN, the immediate predecessors of a node X_i are called the parents of X_i , the immediate successors are called the children of X_i , and the remote successors are called the descendants of X_i . The joint probability distribution over X , represented by the parameters θ can be factored into a product of probability distributions defined on each node in the network.

The Markov blanket (MB) of a target X_i , is a set of features such that conditioned on the MB, X_i is conditionally independent of all other features. The MB consists of the parents, the children, and the parents of the children of X_i (see Fig. 7.2). The MB of a node X_i is noteworthy because it identifies a minimal set of features that are maximally predictive of X_i . A comprehensive review of the methods for the discovery of MBs from data is provided in [17, 18].

One of the earliest algorithms that discover MBs from data is the Grow-Shrink (GS) algorithm that works in two stages [19]. In the growing phase, it identifies

Fig. 7.2 An example MB. The MB of the node X_6 (shown stippled) consists of parents, X_2 and X_3 , children, X_8 and X_9 , and parents of the children, X_5 and X_7 . Nodes X_1 , X_4 , X_{10} and X_{11} are not part of the MB of X_6



features that are strongly associated with the target and in the shrinking phase, it reduces the estimated MB from the growing phase using conditional independence tests. The shrinking phase of the GS algorithm is not sound; this phase is improved in the Incremental Association Markov Blanket (IAMB) algorithm [20]. The growing phase of IAMB identifies all features that have a strong association with the target using a conditional mutual information test that conditions on the features in the MB so far. Falsely included features are removed in the shrinking phase that uses a conditional independence test between each feature in the MB and the target given the remaining features in the MB. The IAMB was shown to select MB features that when used in predictive models out performed classification algorithms when applied directly to the data without filtering. Moreover, though the MB itself can be used directly as a predictive model, it was out performed by other classification algorithms that used the features selected by IAMB [20]. Furthermore, IAMB and its variants were the first of the MB algorithms that were shown to scale to high-dimensional data sets.

More efficient and scalable algorithms that were introduced after IAMB include HITON and Max-Min Markov Blanket (MMMB) [21, 22]. These algorithms were shown to find MBs in a scalable and efficient manner. When HITON was evaluated in clinical, text, genomic, structural and proteomic data it was shown to have excellent performance in terms of parsimony and classification performance. Progress in developing scalable MB algorithms continues including the development of better conditional independence tests such as the kernel-based tests [23].

Biologically Motivated Feature Extraction

A commonly used technique of feature extraction is Principal Component Analysis that constructs a small set of discriminative features from the original features. Another technique is to use available knowledge to extract features. For example, features in gene expression data that contain measurements of individual genes are

combined to create pathway features based on current knowledge of known genes that are members of signaling and metabolic pathways. The pathway features are used to develop predictive models such as outcomes in cancer. This approach can also be viewed as automated, biologically inspired dimensionality reduction where the features are extracted automatically inspired by the types of pathways that are likely driving outcomes.

Model Averaging

When predictive models are estimated from data, multiple models often fit the data more or less equally well. It is usual, then, to select one of the models according to some criteria like model fit to the data or predictive performance of the model. The selection of one model over others that are almost as good can lead to overconfident predictions since it ignores the uncertainty in choosing one model to the exclusion of all others. Hence, it is desirable to model this source of uncertainty by appropriate selection and combination of multiple models. One coherent approach to dealing with the uncertainty in model selection is Bayesian model averaging (BMA) that is an extension of standard Bayesian inference. Typical Bayesian inference models parameter uncertainty through prior distributions, and BMA extends this approach to model uncertainty by estimating posterior distributions for both model parameters and the model structure [24].

BMA estimates the outcome as a weighted average of the outcome predictions of a set of models, with more probable models influencing the prediction more than less probable ones. In practical situations, the number of models to be considered may be enormous, and averaging the predictions over all of them by enumerating each model is infeasible. In selected model families, a closed form solution is available. The next section describes one such example where prediction using the naïve Bayes model can be performed efficiently by averaging over all naïve Bayes models. In most situations, a closed form solution will not be available. A pragmatic approach, then, is to average over a few good models, termed selective Bayesian model averaging, which serves to approximate the prediction obtained from averaging over all models.

Madigan and Raftery show that BMA is expected to have better predictive performance than any single model [25]. Empirically, the superior performance of BMA is supported by a range of case studies. Yeung et al. applied BMA to select genes from DNA microarray data to predict prognosis in breast cancer and showed that BMA identified smaller numbers of relevant genes that had comparable prediction accuracy to other methods that identified larger numbers of genes [26]. Wei et al. applied BMA to high-dimensional single nucleotide polymorphism (SNP) data and showed that it has better predictive performance than model selection [27]. A good overview of BMA is provided in [24] and a comprehensive review of the applications of BMA is described in [28].

Model Averaged Naïve Bayes

Bayesian model averaging of naïve Bayes (NB) models can improve predictions over a single NB model. The single NB model is widely used because of good discriminative performance and computational efficiency. However, on high-dimensional data sets, such as genome-wide single nucleotide polymorphisms (with features in the hundreds of thousands to millions), the predictions of NB tend to be poorly calibrated so that the predictions are too extreme with probabilities that are too close to 0 and 1. The model-averaged naïve Bayes (MANB) algorithm produces predictions by performing BMA over all possible NB models produced by feature selection on a given set of available predictors [29]. MANB averages over the predictions of these models, weighted by the posterior probability of each model. Compared to NB, MANB addresses the challenges of feature selection and tends to have better calibration than NB. MANB has almost the same computationally efficiency as NB. When evaluated on a genome-wide association dataset to predict late-onset Alzheimer's disease, MANB performed significantly better than NB, in terms of both discrimination and calibration [27].

Personalized Modeling

Much of predictive modeling in biomedicine has been based on the expected outcome of an average patient. Data from a population of patients with the same disease are pooled together for statistical analysis, and models derived from the analysis inform the management of future patients. In other words, the typical approach for modeling clinical outcomes is to derive a single predictive model from a dataset of individuals for whom the outcomes are known, and then to apply the model to predict outcomes for future individuals. Such a model is called a *population-wide model* since it is intended to be applied to an entire population of future individuals and is optimized to have good predictive performance on average on all members of that population. This approach has often been quite successful; however, it ignores important individual differences during model *construction*, such as differences in treatment response. Precision medicine aims to tailor clinical therapy to individual patients, with the goal of delivering the right treatments at the right time to the right patient [30]. An approach for better capturing individual differences during modeling is called *patient-specific modeling*, and it focuses on learning models that are tailored to the characteristics of the individual at hand for whom we wish to make a prediction. The basic notion is that patient-specific models that are optimized to perform well for a specific individual are likely to have better predictive performance for that patient than a population-wide model that is optimized to have good predictive performance on average on all future individuals [31, 32].

Personalized Decision Trees

An example of a patient-specific modeling method is the personalized decision tree model that takes advantage of the particular features of an individual [33]. The authors introduce several methods that derive personalized decision trees (a decision path, in fact). When compared to the Classification And Regression Tree (CART) population-wide decision-tree model, the personalized methods performed better in both discrimination and calibration.

Personalized Bayesian Model Averaging

Another example that combines personalized modeling with BMA is a patient-specific algorithm that uses MB models, carries out Bayesian averaging over a set of models to predict the outcome for an individual, and employs a patient-specific heuristic to locate a set of suitable models to average over [31, 34]. When compared to a range of population-wide models, the MB patient-specific models had better performance in both discrimination and calibration.

Explanations

With the increasing complexity of predictive models, a critical bottleneck in their widespread use is the availability of explanations that describe the basis of individual predictions [35]. For example, the insight that an explanation provides about why a particular patient is predicted with high probability to develop a disease, may lead a clinician receiving it to gain trust in that prediction. Such explanations may assist clinicians in making clinical decisions. Explanations differ from model interpretability that refers to understandability or intelligibility of the model in terms of structure and parameters. Some predictive models, such as logistic regression and decision trees, are easier to interpret. Most machine learning models are more opaque. Predictive explanation provides reasoning for the prediction that is made by a model for an individual. Good explanations are parsimonious so that they are readily and rapidly understood by the clinician user and use concepts that are understandable to the user, such as clinical features that are not modified or transformed [36]. Predictive explanations are potentially more useful than interpretable models in the context of clinical decision making, although they are complementary.

Predictive explanations may be based on the structure and parameters of the predictive model that yielded the prediction or may be based on an independent method that is applied after the predictive model has produced its prediction. The latter types of methods can be used with any type of predictive model and have wider applicability. A recently developed method is the Local Interpretable Model-Agnostic Explanations (LIME) that provides an explanation for a prediction by learning an interpretable model locally around the patient for whom we wish to make a prediction [37]. Figure 7.3 provides an example of the application of LIME to explain a clinical outcome prediction.

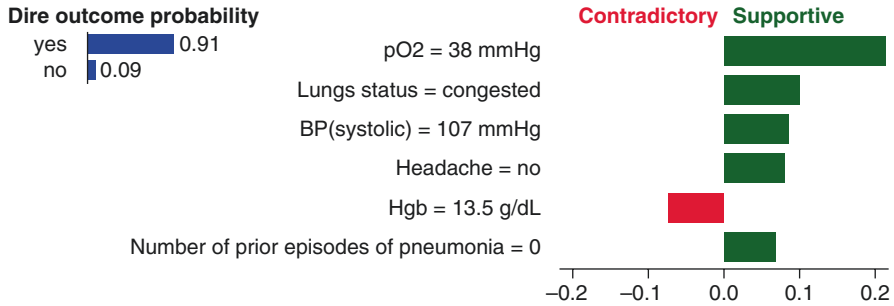


Fig. 7.3 An example explanation obtained from LIME for a patient with pneumonia who was predicted to have a very high probability of a dire outcome (i.e., death or severe complication). The plot at the top left shows the predicted probability distribution for dire outcome. The plot on the right shows the explanation for the prediction. The explanation is limited to six top ranked features by magnitude. The magnitude on the horizontal axis represents the weight of a feature. Green bars represent the magnitude of predictors that support the predicted outcome, while red bars represent the magnitude of contradictory features

Emerging Informatics Standards and Technologies

While many clinical predictive models are developed, few are validated externally, and even fewer are adopted in clinical practice. A key obstacle to the more widespread use of predictive models is the paucity of reporting standards, computable standards, and technologies. Moreover, while the workflow for development and validation of predictive models from research study data is well developed (see section “[Workflow of Development and Validation of Predictive Models](#)”), similar workflow for development of models from observational healthcare data is not yet mature.

Transparent Reporting of Predictive Models

One issue has been the poor quality and nonstandard reporting in published articles of descriptions of predictive models in medicine. The lack of a comprehensive, standard way of reporting the key details of studies that develop and validate models makes it difficult for the scientific and healthcare community to judge the validity and applicability of multivariable predictive models. To address this obstacle, a guideline for the Transparent Reporting of a multivariable predictive model for Individual Prognosis Or Diagnosis (TRIPOD) was introduced [38]. It provides a 22-item checklist that focuses on reporting how a predictive model study was designed, conducted, analyzed, and interpreted. This checklist provides guidance on reporting of items such as title, abstract, descriptions of predictors, outcomes and blinding, descriptions of development and validation data, model specification, development, performance and updating for both model development and external validation (see Table 7.2). A recent study showed that more than half of the items on the checklist were either absent or inadequately reported. Critical information for

Table 7.2 TRIPOD checklist for predictive model development and validation

Section/topic	Item	Checklist item
<i>Title and abstract</i>		
Title	1	D;V Identify the study as developing and/or validating a multivariable predictive model, the target population, and the outcome to be predicted.
Abstract	2	D;V Provide a summary of objectives, study design, setting, participants, sample size, predictors, outcome, statistical analysis, results, and conclusions.
<i>Introduction</i>		
Background and objectives	3a	D;V Explain the medical context (including whether diagnostic or prognostic) and rationale for developing or validating the multivariable predictive model, including references to existing models.
	3b	D;V Specify the objectives, including whether the study describes the development or validation of the model or both.
<i>Methods</i>		
Source of data	4a	D;V Describe the study design or source of data (e.g., randomized trial, cohort, or registry data), separately for the development and validation data sets, if applicable.
	4b	D;V Specify the key study dates, including start of accrual; end of accrual; and, if applicable, end of follow-up.
Participants	5a	D;V Specify key elements of the study setting (e.g., primary care, secondary care, general population) including number and location of centers.
	5b	D;V Describe eligibility criteria for participants.
	5c	D;V Give details of treatments received, if relevant.
Outcome	6a	D;V Clearly define the outcome that is predicted by the predictive model, including how and when assessed.
	6b	D;V Report any actions to blind assessment of the outcome to be predicted.
Predictors	7a	D;V Clearly define all predictors used in developing or validating the multivariable predictive model, including how and when they were measured.
	7b	D;V Report any actions to blind assessment of predictors for the outcome and other predictors.
Sample size	8	D;V Explain how the study size was arrived at.
Missing data	9	D;V Describe how missing data were handled (e.g., complete-case analysis, single imputation, multiple imputation) with details of any imputation method.
Statistical analysis methods	10a	D Describe how predictors were handled in the analyses.
	10b	D Specify type of model, all model-building procedures (including any predictor selection), and method for internal validation.
	10c	V For validation, describe how the predictions were calculated.
	10d	D;V Specify all measures used to assess model performance and, if relevant, to compare multiple models.
	10e	V Describe any model updating (e.g., recalibration) arising from the validation, if done.
Risk groups	11	D;V Provide details on how risk groups were created, if done.

Table 7.2 (continued)

Section/topic	Item		Checklist item
Development vs. validation	12	V	For validation, identify any differences from the development data in setting, eligibility criteria, outcome, and predictors.
<i>Results</i>			
Participants	13a	D;V	Describe the flow of participants through the study, including the number of participants with and without the outcome and, if applicable, a summary of the follow-up time. A diagram may be helpful.
	13b	D;V	Describe the characteristics of the participants (basic demographics, clinical features, available predictors), including the number of participants with missing data for predictors and outcome.
	13c	V	For validation, show a comparison with the development data of the distribution of important features (demographics, predictors and outcome).
Model development	14a	D	Specify the number of participants and outcome events in each analysis.
	14b	D	If done, report the unadjusted association between each candidate predictor and outcome.
Model specification	15a	D	Present the full prediction model to allow predictions for individuals (i.e., all regression coefficients, and model intercept or baseline survival at a given time point).
	15b	D	Explain how to use the predictive model.
Model performance	16	D;V	Report performance measures (with CIs) for the predictive model.
Model-updating	17	V	If done, report the results from any model updating (i.e., model specification, model performance).
<i>Discussion</i>			
Limitations	18	D;V	Discuss any limitations of the study (such as nonrepresentative sample, few events per predictor, missing data).
Interpretation	19a	V	For validation, discuss the results with reference to performance in the development data, and any other validation data.
	19b	D;V	Give an overall interpretation of the results, considering objectives, limitations, results from similar studies, and other relevant evidence.
Implications	20	D;V	Discuss the potential clinical use of the model and implications for future research.
<i>Other information</i>			
Supplementary information	21	D;V	Provide information about the availability of supplementary resources, such as study protocol, web calculator, and data sets.
Funding	22	D;V	Give the source of funding and the role of the funders for the present study.

Adapted from Table 1 in Ref. [38]

Items relevant only to the development of a predictive model are denoted by D, items relating solely to a validation of a predictive model are denoted by V, and items relating to both are denoted D;V. From *Annals of Internal Medicine*, Moons, Karel G.M.; Altman, Douglas G, Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): Explanation and Elaboration, 162(1), 1–73. Copyright © 2015 American College of Physicians. All Rights Reserved. Reprinted with the permission of American College of Physicians, Inc.

using the model, including model specification and performance, was inadequately reported for more than 80% of the models [39]. Increased adherence and further refinement of the TRIPOD checklist will enhance more transparent reporting of clinical predictive models.

Computable and Portable Predictive Models

Widespread use of predictive models in clinical medicine requires deployment of models in computable formats so that they can be applied to EHRs to automatically provide predictions and recommend actions in the context of a patient. Currently, well-described human-readable predictive models require manual translation to computable formats that is slow and resource-intensive. Rapid deployment of computable models will require development of new standards and technologies. These include the creation of standards for a computable representation of predictive models, development of tools to enable standards-based authoring of models, construction of infrastructure for execution of models in a variety of EHR systems, and digital libraries for collecting, storing, and sharing models.

In the domain of data, the FAIR Data Principles are a set of guiding principles that have been put forth to make data findable, accessible, interoperable, and reusable [40]. These principles facilitate the ability of computers to automatically find and use data and enable its reuse. A similar set of principles are needed for making computable predictive models findable, accessible, interoperable, and reusable. As an example, a computable phenotype is defined as a set of clinical features that can be determined from the data in EHRs, and efforts are ongoing to develop a set of standards for developing a computable phenotype representation that is easily authored, portable and executable. The recently described Knowledge Object Reference Ontology provides a framework to help make computable biomedical knowledge that includes computable phenotypes and predictive models findable, accessible, interoperable, and reusable [41].

Modeling Using Large Scale Observational Data

Observational healthcare data, that includes EHRs and administrative claims data, are increasingly available for secondary use and research through federated data networks. The PCORnet, funded by the PCORI, is a U.S.-wide federated network of EHR, claims, and patient reported outcome data on over 100 million patients [42]. The Accrual to Clinical Trials (ACT) network, funded by the NIH, is another U.S. federated network of EHR and claims data on over 40 million patients [43]. The Observational Health Data Sciences and Informatics (OHDSI) collaboration is a network of loosely collaborating sites with EHR and claims data on hundreds of millions of patients [44]. These networks have adopted similar data models that

specify standardized structure and content for observational data. In contrast to research study data that consist of specified measurements that are expressly measured for the study, observational healthcare data consists of clinical measurements across a range of domains (such as diagnoses, procedures, medications, and laboratory test values) that are captured during the process of care. Compared to study data, observational healthcare data are typically much larger with tens of thousands of measurements on tens of millions of individuals. For research use, healthcare data are standardized to common terminologies, such as ICD-9 and ICD-10 codes for diagnoses and procedures, RxNorm and National Drug Codes (NDC) for medications, and Logical Observation Identifiers Names and Codes (LOINC) for laboratory test results. Standardization of the data requires considerable time and resources to map the source data to standard terminologies and transform it in accordance to the data model specifications.

The use of healthcare data for predictive modeling is still in its infancy. It holds the promise of revolutionizing clinical predictive modeling on very large scales and across several different diagnoses, outcomes, and treatments simultaneously. The OHDSI community has introduced a framework for developing and validating predictive models using observational healthcare data. Moreover, open-source software is available that implements this framework for data that has been transformed to the Observational Medical Outcomes Partnership (OMOP) data model. This framework was applied to develop predictive models using several machine learning methods for 21 different outcomes in a population of pharmaceutically-treated depression patients across four observational data sets that contained a total of over 230 million patients. For some outcomes, high performing models were obtained while for other outcomes the models performed poorly, suggesting that observational data sets are likely to be useful for some outcomes but not for all, and that, healthcare data complement research study data [45].

Policy, Ethical, and Legal Challenges

The increasing availability of big biomedical data and the growing application of new statistical and machine learning methods for developing complex models from big data provide an opportunity for widespread development of clinical predictive models. When such models are deployed to provide targeted care, to improve outcomes, and to lower healthcare costs, several policy, ethical, and legal challenges arise. A comprehensive consideration of such issues is presented in a recent publication [46], and a few key issues are summarized in the next paragraph.

A primary consideration is that data used in model derivation and validation is representative of the whole population. Historically, members of certain racial and ethnic groups, people with disabilities, individuals in prison, and members of other vulnerable groups have been underrepresented in research studies. Such inequitable representation can lead to models that are not valid for parts of the population. In addition to extensive validation, models need to be evaluated in real-world settings before

deployment. A second consideration is that the models are developed both in human-readable and machine-readable forms using standards that are transparent and replicable. A third consideration is liability. Makers as well as users of predictive models may face liability if there are errors in the model or the model malfunctions. A fourth consideration is that population-wide models that are designed to improve outcomes in a population may produce a sub-optimal prediction for a specific patient. As a simple illustrative example, a population-wide model that predicts future morbidity may not include human immunodeficiency virus (HIV) status as a predictor because the proportion of HIV positive patients in the data is very small. Such a model will produce sub-optimal predictions for patients with positive HIV status, and a patient-specific model that includes HIV status as a predictor will provide better predictions. Ethical obligations of clinicians to act in the best interests of a patient may lead to increased use of patient-specific models over population-wide ones.

Conclusions

With increasing availability of big biomedical data, valid and high-performing predictive modeling methods are needed to leverage the data for clinical medicine, public health, and biomedical research. Several current trends indicate that biomedical data will become more readily available and that will accelerate the development of predictive models in medicine. For example, the National Institutes of Health's strategic plan for data science provides a roadmap for storing, managing, standardizing and publishing the vast amounts of data produced by biomedical research [47]. The Director of the National Library of Medicine at the National Institutes of Health anticipates an important role for a library of models that will identify, collect and archive biomedical models [48]. In addition to the expertise in academia, companies with expertise in artificial intelligence like Microsoft, Google, Baidu, and Apple are developing predictive models for healthcare [49]. The General Data Protection Regulation (GDPR) that was recently adopted by the European Union includes a "right to explanation" with regard to predictive models that seeks to enforce the availability of explanations for predictions made by models [50]. Thus, the coming decade will likely see increasing development and validation of predictive models from big biomedical data and will include advances in feature selection, high performance, personalization of models, and explanations of predictions.

References

1. Moons KGM, Altman DG, Vergouwe Y, Royston P. Prognosis and prognostic research: application and impact of prognostic models in clinical practice. *Br Med J*. 2009;338:b606.
2. Wilson PWF, D'Agostino RB, Levy D, Belanger AM, Silbershatz H, Kannel WB. Prediction of coronary heart disease using risk factor categories. *Circulation*. 1998;97:1837–47.

3. Jessen MK, Mackenhauer J, Hvass AMSW, Ellermann-Eriksen S, Skibsted S, Kirkegaard H, et al. Prediction of bacteremia in the emergency department: an external validation of a clinical decision rule. *Eur J Emerg Med.* 2016;23:44–9.
4. Shapiro NI, Wolfe RE, Wright SB, Moore R, Bates DW. Who needs a blood culture? A prospectively derived and validated prediction rule. *J Emerg Med.* 2008;35:255–64.
5. LaHaye SA, Gibbens SL, Ball DGA, Day AG, Olesen JB, Skanes AC. A clinical decision aid for the selection of antithrombotic therapy for the prevention of stroke due to atrial fibrillation. *Eur Heart J.* 2012;33:2163–71.
6. Brownstein JS, Freifeld CC, Chan EH, Keller M, Sonricker AL, Mekaru SR, et al. Information technology and global surveillance of cases of 2009 H1N1 influenza. *N Engl J Med.* 2010;362:1731–5.
7. Steyerberg EW. *Clinical prediction models: a practical approach to development, validation, and updating.* New York: Springer; 2008.
8. Clark GM, Zborowski DM, Culbertson JL, Whitehead M, Savoie M, Seymour L, et al. Clinical utility of epidermal growth factor receptor expression for selecting patients with advanced non-small cell lung cancer for treatment with erlotinib. *J Thorac Oncol.* 2006;1:837–46.
9. Sechidis K, Papangelou K, Metcalfe PD, Svensson D, Weatherall J, Brown G. Distinguishing prognostic and predictive biomarkers: an information theoretic approach. *Bioinformatics.* 2018;1:12.
10. Labarère J, Bertrand R, Fine MJ. How to derive and validate clinical prediction models for use in intensive care medicine. *Intensive Care Med.* 2014;40:513–27.
11. Hendriksen JMT, Geersing GJ, Moons KGM, De Groot JAH. Diagnostic and prognostic prediction models. *J Thromb Haemost.* 2013;11:129–41.
12. Alba AC, Agoritsas T, Walsh M, Hanna S, Iorio A, Devereaux PJ, et al. Discrimination and calibration of clinical prediction models: users' guides to the medical literature. *J Am Med Assoc.* 2017;318:1377–84.
13. Collart F, Feier H, Kerbaul F, Mouly-Bandini A, Riberi A, Mesana TG, et al. Valvular surgery in octogenarians: operative risks factors, evaluation of Euroscore and long term results. *Eur J Cardio Thoracic Surg.* 2005;27:276–80.
14. Nashef SAM, Roques F, Sharples LD, Nilsson J, Smith C, Goldstone AR, et al. Euroscore II. *Eur J Cardio Thoracic Surg.* 2012;41:734–45.
15. Guyon I, Elisseeff A. An introduction to variable and feature selection. *J Mach Learn Res.* 2003;3:1157–82.
16. Zeng Y, Luo J, Lin S. Classification using Markov blanket for feature selection. *IEEE International Conference on Granular Computing.* 2009. p. 743–7.
17. Aliferis CF, Statnikov A, Tsamardinos I, Mani S, Koutsoukos XD. Local causal and Markov blanket induction for causal discovery and feature selection for classification part I: algorithms and empirical evaluation. *J Mach Learn Res.* 2010;11:171–234.
18. Aliferis CF, Statnikov A, Tsamardinos I, Mani S, Koutsoukos XD. Local causal and Markov blanket induction for causal discovery and feature selection for classification part II: analysis and extensions. *J Mach Learn Res.* 2010;11:235–84.
19. Margaritis D, Thrun S. Bayesian network induction via local neighborhoods. *Proc Adv Neural Inf Process Syst.* 2000:505–11.
20. Tsamardinos I, Aliferis CF, Statnikov AR, Statnikov E. Algorithms for large scale Markov blanket discovery. *Proc Florida Artif Intell Res Soc.* 2003:376–80.
21. Aliferis CF, Tsamardinos I, Statnikov A. HITON: a novel Markov blanket algorithm for optimal variable selection. *AMIA Annu Symp Proc.* 2003;2003:21–5.
22. Tsamardinos I, Brown LE, Aliferis CF. The max-min hill-climbing Bayesian network structure learning algorithm. *Mach Learn.* 2006;65:31–78.
23. Strobl EV, Visweswaran S. Markov blanket ranking using kernel-based conditional dependence measures. *arXiv Prepr arXiv14020108.* 2014.
24. Hoeting JA, Madigan D, Raftery AE, Volinsky CT. Bayesian model averaging: a tutorial. *Stat Sci.* 1999;14:382–401.

25. Madigan D, Raftery AE. Model selection and accounting for model uncertainty in graphical models using Occam's window. *J Am Stat Assoc.* 1994;89:1535–46.
26. Yeung KY, Bumgarner RE, Raftery AE. Bayesian model averaging: development of an improved multi-class, gene selection and classification tool for microarray data. *Bioinformatics.* 2005;21:2394–402.
27. Wei W, Visweswaran S, Cooper GF. The application of naive Bayes model averaging to predict Alzheimer's disease from genome-wide data. *J Am Med Inform Assoc.* 2011;18:370–5.
28. Fragoso TM, Bertoli W, Louzada F. Bayesian model averaging: a systematic review and conceptual classification. *Int Stat Rev.* 2018;86:1–28.
29. Dash D, Cooper GF. Exact model averaging with naive Bayesian classifiers. *Proc Int Conf Int Conf Mach Learn.* 2002:91–8.
30. Collins FS, Varmus H. A new initiative on precision medicine. *N Engl J Med.* 2015;372:793–5. <https://doi.org/10.1056/NEJMp1500523>.
31. Visweswaran S, Angus DC, Hsieh M, Weissfeld L, Yealy D, Cooper GF. Learning patient-specific predictive models from clinical data. *J Biomed Inform.* 2010;43:669–85.
32. Visweswaran S, Cooper GF. Learning instance-specific predictive models. *J Mach Learn Res.* 2010;11:3333–69.
33. Visweswaran S, Ferreira A, Ribeiro GA, Oliveira AC, Cooper GF. Personalized modeling for prediction with decision-path models. *PLoS One.* 2015;10:e0131022.
34. Visweswaran S, Cooper GF. Patient-specific models for predicting the outcomes of patients with community acquired pneumonia. *AMIA Annu Symp Proc.* 2005;2005:759–63.
35. Suermondt HJ, Cooper GF. An evaluation of explanations of probabilistic inference. *Comput Biomed Res.* 1993;26:242–54.
36. Caruana R, Lou Y, Gehrke J, Koch P, Sturm M, Elhadad N. Intelligible models for healthcare: predicting pneumonia risk and hospital 30-day readmission. *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM;* 2015. p. 1721–30.
37. Ribeiro MT, Singh S, Guestrin C. Why should I trust you?: explaining the predictions of any classifier. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 2016.* p. 1135–44.
38. Moons KGM, Altman DG, Reitsma JB, Ioannidis JPA, Macaskill P, Steyerberg EW, et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): explanation and elaboration. *Ann Intern Med.* 2015;162:W1–73.
39. Heus P, Damen JAAG, Pajouheshnia R, Scholten RJPM, Reitsma JB, Collins GS, et al. Poor reporting of multivariable prediction model studies: towards a targeted implementation strategy of the TRIPOD statement. *BMC Med.* 2018;16:120.
40. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. The FAIR guiding principles for scientific data management and stewardship. *Sci Data.* 2016;3:160018.
41. Flynn AJ, Friedman CP, Boisvert P, Landis-Lewis Z, Lagoze C. The knowledge object reference ontology (KORO): a formalism to support management and sharing of computable biomedical knowledge for learning health systems. *Learn Heal Syst.* 2018;2:e10054.
42. Collins FS, Hudson KL, Briggs JP, Lauer MS. PCORnet: turning a dream into reality. *J Am Med Inform Assoc.* 2014;21:576–7.
43. Visweswaran S, Becich MJ, D'Itri VS, Sendro ER, MacFadden D, Anderson NR, et al. Accrual to clinical trials (ACT): a clinical and translational science award consortium network. *JAMIA Open.* 2018;1:147–52. <https://doi.org/10.1093/jamiaopen/ooy033>.
44. Hripcsak G, Duke JD, Shah NH, Reich CG, Huser V, Schuemie MJ, et al. Observational health data sciences and informatics (OHDSI): opportunities for observational researchers. *Stud Health Technol Inform.* 2015;216:574.
45. Reps JM, Schuemie MJ, Suchard MA, Ryan PB, Rijnbeek PR. Design and implementation of a standardized framework to generate and evaluate patient-level prediction models using observational healthcare data. *J Am Med Inform Assoc.* 2018;25:969–75.

46. Cohen IG, Amarasingham R, Shah A, Xie B, Lo B. The legal and ethical concerns that arise from using complex predictive analytics in health care. *Health Aff.* 2014;33:1139–47.
47. National Institutes of Health. NIH strategic plan for data science [Internet]. [cited 2018 Oct 21]. p. 1–26. <https://grants.nih.gov/grants/rfi/NIH-Strategic-Plan-for-Data-Science.pdf>.
48. Brennan PF. Models: the third leg in data-driven discovery – NLM musings from the mezzanine [internet]. 2017 [cited 2018 Oct 21]. <https://nlmdirector.nlm.nih.gov/2017/12/12/models-the-third-leg-in-data-driven-discovery/>.
49. Ravi D, Wong C, Deligianni F, Berthelot M, Andreu-Perez J, Lo B, et al. Deep learning for health informatics. *IEEE J Biomed Heal Informat.* 2017;21:4–21. <http://ieeexplore.ieee.org/document/7801947/>.
50. Voigt P, von dem Bussche A. The EU General Data Protection Regulation (GDPR): a practical guide. Cham: Springer; 2017.