CHAPTER

11 Generation of Knowledge for Clinical Decision Support

Statistical and Machine Learning Techniques

Michael E. Matheny and Lucila Ohno-Machado

11.1 Introduction

Clinical decision support (CDS) systems must rely on knowledge that originates from a variety of sources, including domain literature, expert knowledge, and statistical analysis of data. However, selecting the sources and integrating this knowledge into a functional system is not a trivial task. The earliest systems for medical diagnosis, in the 1960s, used Bayesian probability models (as discussed in Chapter 2), relying on either databases of patient data or subjective estimates of prior and conditional probabilities from human experts. In most early CDS systems developed in the 1970s and 1980s, because of the dearth of available data and the interest at the time in the burgeoning field of artificial intelligence, knowledge was acquired directly from medical experts (as discussed further in Chapter 10). In some systems, pioneered by an antibiotic treatment advisor program known as MYCIN (Shortliffe et al., 1975), knowledge was encoded in the form of rules that were triggered and chained according to an embedded or an external inference engine (Shortliffe, 1976). In MYCIN, rules were expert-derived and had associated certainty factors, a mathematical formulation of a quasi-statistical representation of degree of belief, developed by Shortliffe et al. (1975) Various approaches to medical diagnostic reasoning from this period such as the Present Illness Program (Pauker et al., 1976), Internist/QMR (Miller et al., 1982; Miller et al., 1986), DXplain (Barnett et al., 1987) (still in use) and others that were developed subsequently were constructed from a set of physician-based assessments of (a) the strength with which clinical findings evoke a certain diagnosis, (b) prevalence of diseases, and (c) related indices. As noted by Heckerman and colleagues, the formal mathematical definitions of these indices in terms of probabilities have not been fully elucidated Heckerman and Miller (1986).

Even when newer knowledge representation strategies incorporating statistical data, such as Bayesian networks (Pearl, 1988), were proposed by some researchers

Clinical Decision Support.

in the late 1980s (Heckerman, 1990; Beinlich et al., 1989), definition of the graph structure and probabilities involved in the model were usually still assessed by experts. For example, Shwe and colleagues (Swhe et al., 1991) "translated" the QMR representation into Bayesian networks, and showed that the same diagnostic quality could be achieved with a representation that made explicit important modeling assumptions. However, the popularity of Bayesian networks in the medical community did not grow as expected, and this type of model is still primarily used in the medical domain primarily for research purposes, with very few exceptions. Reasons for this limitation may include the need for severe model simplifications in order to make these models practical for clinical use. These simplifications may in turn reduce the main advantages of using Bayesian networks, which include their explicit knowledge representation combining a sound probabilistic modeling of dependencies with a visually appealing display. Algorithms for learning Bayesian networks from data have evolved in the past two decades but are also primarily used in research applications (Cooper and Herskovits, 1992; Buntine, 1996; Moore and Lee, 1998).

Most clinical decision support systems in current use do not learn from data and still rely on the rule-based paradigm, mainly in the form of single IF/THEN rules (see Chapter 15) or as computer-interpretable guidelines (see Chapter 16) which chain together steps in a care process using branching decision rules. In both cases, although probabilistic considerations have usually gone into constructing the rules or guidelines, using evidence-based medicine techniques (see also Chapter 12), the rules and guidelines tend to be stated in a deterministic fashion, without associated probabilities.

There are at least two factors that contribute to this predominant reliance on expert assessments for the construction of CDS systems:

- Data are either not available or not structured enough to allow knowledge to be "learned" from them.
- Techniques to discover patterns in data are not well disseminated or not well evaluated in the biomedical community.

A potential third factor may be that systems derived from human knowledge in which nonprobabilistic rules are defined by experts may be more clearly understandable by clinicians. For example, if an expert can articulate all the rules that were used to make a diagnosis and how they were chained, then a system based on these rules can potentially explain its reasoning in a way that clinicians would be able to understand and accept (Clancey, 1983). Whether understanding and agreement by clinicians is necessary for the underlying logic in CDS systems to be useful remains a controversial issue. Although, as noted above, decision support applications currently in use in clinical environments rely in large part on deterministic rules for their "logic," this should not necessarily mean that other approaches are not as good or perhaps even better. For domains in which structured data are abundant, and the decisions are made at times in which a snapshot of these data could help identify specific patterns, pattern recognition algorithms from the

fields of statistical and machine learning can be of great value. This is, of course, becoming especially true in the era of genomics and of "big data," as we discuss in Chapter 2, in which it is both important and increasingly feasible to base decisions about a patient on a comparison with the experiences of a subset of maximally similar patients. The number of factors to be considered and their combinatorics make the task of developing a rule set to cover all of the variations in genomic risk factors, presence of disease, stage of disease, comorbidities, treatments, and complications of treatments more and more intractable. Chapter 13 discusses the growing opportunities for harnessing population health data for decision support.

There have been extensive new developments in statistical and machine learning research in the past few decades. These advances have coincided with improvements in data quality and quantity from the implementation of large repositories of structured electronic data, some of which are based on domain-specific data element standards (Cannon et al., 2001; Wattigney et al., 2003; Pollock et al., 1998). Increased availability of data has allowed the further development of several models that can detect patterns in biomedical data and generalize well to previously unseen cases. Clinical decision support systems that rely on patterns that are recognized in these data are now available in virtually every medical specialty (Knaus et al., 1985; Grundy et al., 1999; Shaw et al., 2002; Goldman et al., 1982; Baxt 1991; O'Leary et al., 1998). Just as in the foregoing discussion relating rule-based systems and more sophisticated knowledge representation paradigms, simple understandable models (e.g. linear and logistic regression and linear score systems) have far outweighed in number and utilization the more sophisticated machine learning models (e.g. support vector machines, neural networks, and recursive partitioning algorithms), many of which remain limited to research applications.

Although many studies have shown the efficacy of CDS systems, several recent studies have found that implementation of CDS may not improve quality of care (Romano and Stafford, 2011) and, in some cases, may result in adverse outcomes (Han et al., 2005) or experience less than optimal performance (Saverno et al., 2011). These cautionary tales, coupled with the release in the US of the Stage 1 (2010) and 2 (2012) Meaningful Use Final Rules, which include functional requirements regarding the implementation and expansion of CDS capacities, coupled to reimbursement, have led to the publication of several guidelines on "best practices" for CDS implementation, by the Institutes of Medicine, AHRQ (Das and Eichner, 2010), and numerous domain experts. Among these recommendations is the improvement of specificity and sensitivity in CDS systems for personalized medicine and reduction of alert fatigue. Probabilistic modeling-based approaches to CDS have been shown to achieve some of these goals in research settings, and there is a slowly increasing uptake in the use of probabilistic modeling methods embedded in CDS systems.

In this chapter, we will review the methodologies of the most commonly used diagnostic and prognostic models in the medical domain, and discuss specific strengths and weaknesses of alternative modeling methods. Popular examples of some modeling methods will be discussed. Since our focus is on models that have

been utilized in practice, the discussion will concentrate on logistic regression models, classification trees, and artificial neural networks. We conclude with a discussion on current directions for the field.

Note the absence of sections dedicated to other topics that have received wide coverage in the computer science literature, but that in fact have limited representation in clinical informatics applications and are beyond the scope of an introductory chapter. For example, although rule-induction algorithms and kernel-based classifiers such as support vector machines (Boser et al., 1992) have often been utilized in research applications, few actual applications are used in clinical practice, and therefore we elected not to cover these models in this chapter. Refer to statistical and machine learning textbooks for a review of these topics (Duda et al., 2001; Hastie et al., 2001).

Another omission is the discussion of optimization techniques such as genetic algorithms and evolutionary computing (Koza, 1992), and formalism extensions such as fuzzy logic (Zadeh, 1994) and rough sets Pawlak (1982). Elements of these techniques can be used in conjunction with the classifiers discussed here in a number of different ways, but they do not constitute classifiers themselves. Furthermore, there are no examples of practical use of these techniques in clinical decision support.

11.2 Learning from data

Statistical and machine learning pattern recognition algorithms have been in existence for several decades. These algorithms recognize regularities in data and construct a model that can be utilized in new cases. Interest in this type of method has increased in the past two decades, with the addition of new algorithms such as neural networks and support vector machines (Vapnik, 1995). A myriad of publications in the scientific and lay literature can now be found under the rubric of "data mining." Data mining techniques are pattern recognition techniques intended to find correlations and relationships in the plethora of data. The term is intriguing, but also somewhat misleading. Most pattern recognition or predictive models used in clinical domains are confirmatory rather than exploratory in nature. The distinction between unsupervised and supervised learning models is directly related to this issue.

Unsupervised learning models are not based on predefined classifications and are used frequently for exploratory data analyses in domains in which knowledge is sparse. For example, high-throughput data are often subject to unsupervised learning modeling so that "clusters" of variables or objects can be revealed without guidance from the users or the existing literature. The objective is to unveil hidden patterns in the data that were not previously anticipated, and label these patterns "a posteriori." This is in sharp contrast with *supervised learning* models, in which the objective is to determine how to best classify objects with predefined labels representing classes of interest (e.g. malignant versus benign cases) using the data at

hand. As expected, unsupervised learning models have not been applied in clinical decision support systems and have a limited role in this area. All models in current use for clinical decision support have been based either on expert knowledge or supervised learning models. The latter is the subject of this chapter.

In order to understand how a model can be derived from data, it is useful to construct an artificial example. Suppose a researcher does not know the range of normal values for a new diagnostic test, but she does have a large data set indicating, for a set of patients, the value of the test and the actual diagnosis for each patient. Also suppose that there are missing and noisy data in the data set. The task is to determine the range of normal values for the test, so that when anyone examines the value for a new patient, it would be possible to declare, with a certain level of confidence, whether the result pointed to an abnormality or not. While one might not need a sophisticated model to answer this simple question, it would be necessary to review all labeled data to determine optimal thresholds to label a result as "normal" or "abnormal."

This analysis can extend to several tests and clinical findings, and multiple possible diagnoses, in which case the task is to find optimal combinations of values that are most frequently associated with particular diagnoses, since a single test or clinical finding in isolation may not suffice. Researchers would have to examine several thousands of records containing dozens of attributes for each patient to determine which combinations of variable values seem to be most likely to be associated with each diagnosis. Given time and memory limitations, it might be difficult to build this type of classifier. For this type of problems, utilizing multivariate techniques that "learn" from data can be very helpful.

11.3 Overview of logistic regression

The first step towards the construction of a predictive model is the selection of the variables that are going to be considered, from a data set containing large numbers of cases. The number of cases needs to exceed the number of variables; a well-known heuristic is that the number of variables utilized in a model should not exceed one tenth of the number of cases. The type of modeling technique needs also to be selected. Logistic regression is by far the most popular method for constructing predictive models in medicine (Lemeshow and Le Gall, 1994). This type of classification model usually deals with binary outcomes such as diagnosis of a certain disease or condition (e.g. myocardial infarction), or prognosis within a certain period of time (e.g. death while in hospital). Using a large number of training cases, it is possible to estimate the parameters of a logistic regression model with a certain level of confidence and estimate the future performance of the model in previously unseen cases. The level of confidence will depend on the number and quality of cases (e.g. presence of outliers and noise), as well as how well the model fits the training data.

The logistic function links i predictors, or independent variables, each denoted by x_i and collectively represented by the vector $\mathbf{x}$, to the dependent variable being predicted, represented by Y using the logistic function as in the equation below:

$$Y = \frac{1}{1 + e^{-(\beta \mathbf{x} + c)}}$$

This function tries to model a step function with two possible values for Y, and it is therefore used to classify binary outcomes. The resulting function is a continuous value from 0 to 1 along a sigmoid curve. Figure 11.1 illustrates a logistic regression model (a sigmoidal function), and this function is also one of the possible functions used within the nodes of artificial neural networks, which we will describe in Section 11.4. In most models, Y is a binary variable representing patient status as having a certain disease or condition ($Y = 1$) or not ($Y = 0$), or prognostic class, and the vector $\mathbf{x}$ represents the clinical, laboratory, and demographic predictors (e.g. x_1 may represent *age*, x_2 may represent *TSH*, and so on). The vector β represents the coefficients that are estimated for each predictor and c is a constant. The parameters of the logistic function are usually obtained by maximum likelihood estimation using iterative algorithms (Hosmer and Lemeshow, 1989). The coefficients correspond directly to the log of the odds ratio associated with each variable. The parameter c calibrates the model for the baseline rate of the outcome of interest. These features make the model somewhat easy to interpret, since the sign and magnitude of the coefficients (when standardized) can provide a direct indication of how much each particular predictor is associated with an increased risk of a certain

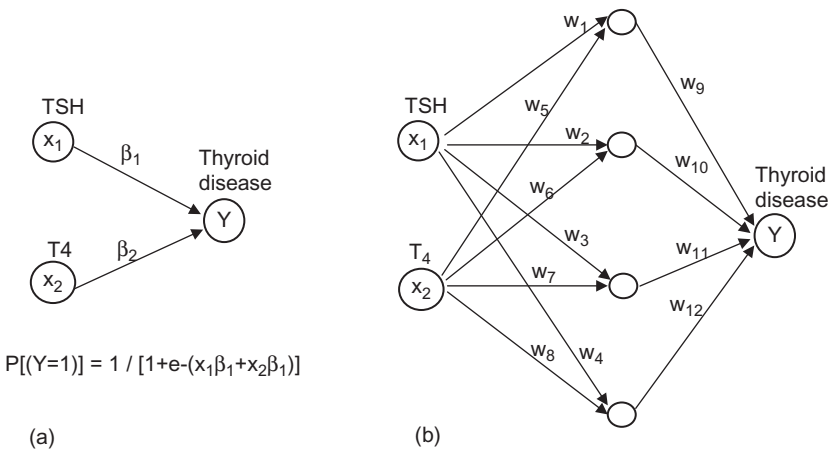

FIGURE 11.1

(a) Example of a simple bivariate logistic regression model (no intercept is included for simplicity). (b) Example of an artificial neural network constructed for the same purpose.

outcome (e.g. large positive coefficients will usually increase the probability of Y = 1 for variables such as those representing most laboratory assays).

For certain data, predictors may need to be combined in interaction terms or transformed so that a good fit to the data can be obtained. Consider the example in Figure 11.2: a laboratory test value that is considered normal within a certain range (e.g. *TSH* within 0.4–6 μU/mL), and abnormal otherwise. Even in this simple

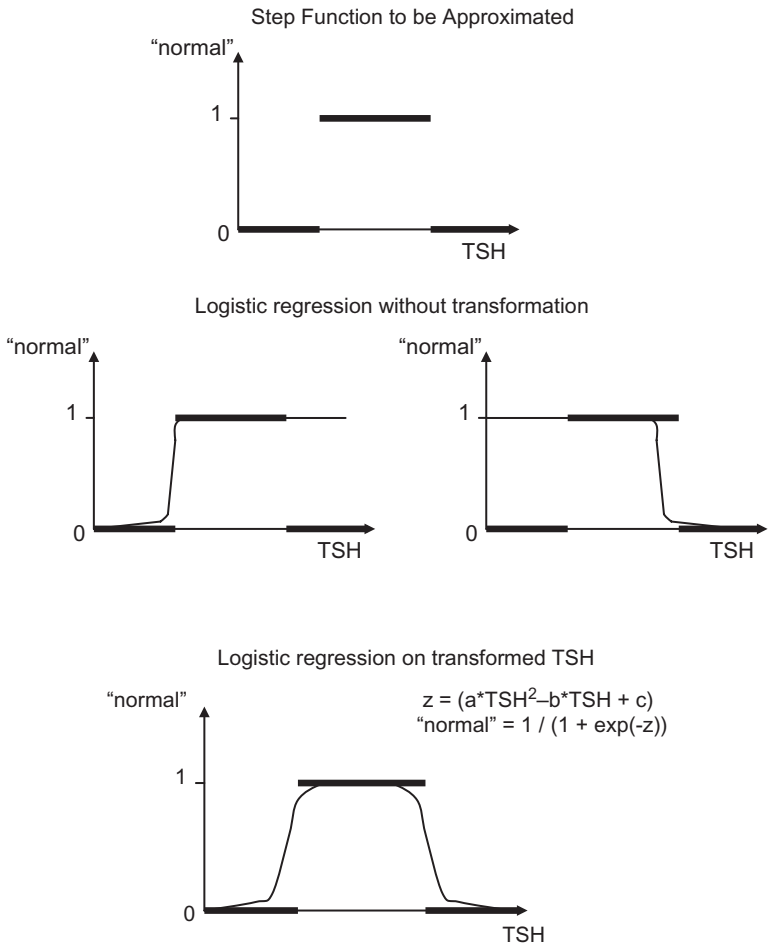

FIGURE 11.2

A step function (bold) indicating "normal" laboratory values within a certain range. The step function is overlaid with logistic functions for illustration purposes. Without variable transformation, the logistic regression function will always miss one of the extremes of values, misclassifying values within that range as not "normal." A simple quadratic transformation can make logistic regression work for this example.

univariate problem of classifying the values into normal and abnormal, a logistic regression model in which variables are not transformed will not be able to correctly classify all cases, even in the absence of noise. The reason is simple: the logistic regression function is monotonic and would necessarily classify a portion of the abnormal cases (either the low or high values) as being normal. However, a simple transformation of the variable, in this case, a quadratic, might allow the logistic regression model to correctly classify all cases.

Figure 11.3 illustrates a bivariate problem in which values for two laboratory tests have to be within a certain range for the patient to be considered healthy. In this example, both free T_4 and *TSH* need to be within normal limits for the classification "euthyroid" to be made. It is easy to see that no single line would separate the shaded area from the rest, which means that no linear model can produce correct classifications for all cases. Variable transformations or interaction terms are necessary.

Problems that are not solvable by linear or semilinear models such as logistic regression without variable transformations or addition of interaction terms are known as nonlinearly separable problems. Although logistic regression models can be used in linearly nonseparable problems, predetermining which transformations or interactions are necessary is a laborious ad hoc process that is computationally intractable unless the number of variables is very small. Furthermore, the interpretation of a model that uses transformed variables or interaction terms is difficult. Therefore, most models that are used in practice do not make use of interaction or transformed terms.

Techniques that originated in the computer science community have addressed nonlinearly separable problems in different ways. We review next some of these techniques.

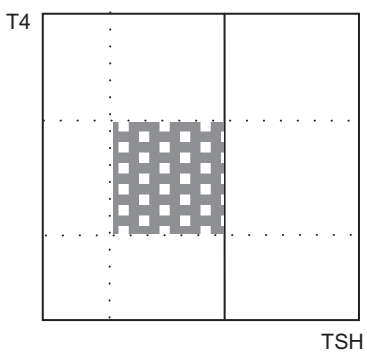

FIGURE 11.3

Simplified bivariate example. For a case to be considered "euthyroid" (shaded area), values for both tests have to be within a certain range. Without variable transformation, logistic regression will not work for all cases because the problem is not linearly separable.

11.4 Overview of some machine learning models

Artificial intelligence techniques such as those commonly referred to as *machine learning* techniques have been explored to address some potential limitations of standard modeling techniques. Among these techniques, classification trees and artificial neural networks have been the most popular in the medical domain.

11.4.1 Classification trees

Classification trees recursively and univariately partition cases into two subgroups (Breiman et al., 1984). At each branch in an upside-down tree, as illustrated in Figure 11.4, the attribute-value pair that best partitions the cases into the categories of interest (e.g. "euthyroid" or not) is chosen. A simple step function assigns "yes" or "no" to the criterion in question (e.g. TSH > 6 = yes). This is repeated until the partitions that represent the "leaves" of the tree have only cases from a single category. Figure 11.4 illustrates the simplified bivariate example from Figure 11.3. The first attribute-value pair to be chosen is (TSH, 6 μU/mL). Cases in the right branch/leaf (TSH > 6) are not euthyroid. Cases in the left branch (TSH ≤ 6) may be euthyroid or not. The next attribute is T_4 at 12.5. Cases in the right branch/leaf (T_4 > 12.5) are not euthyroid. Cases in the left branch may be euthyroid. The pair (TSH, 0.4 μU/mL) is then chosen, and cases are classified into "Not euthyroid" if TSH ≤ 0.4. Otherwise, (T_4, 6.5 is chosen) and those cases with T_4 > 6.5 are classified as "euthyroid."

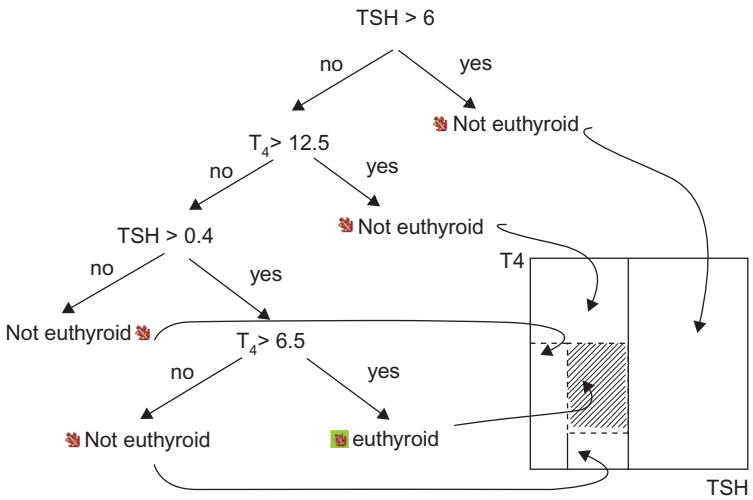

FIGURE 11.4

Classification tree for the bivariate outcome problem illustrated in Figure 11.3. Cases are recursively partitioned according to the attribute-value pair that best divides the cases into "euthyroid" or not. The resulting partitions can be easily visualized in this simplified two-dimensional problem.

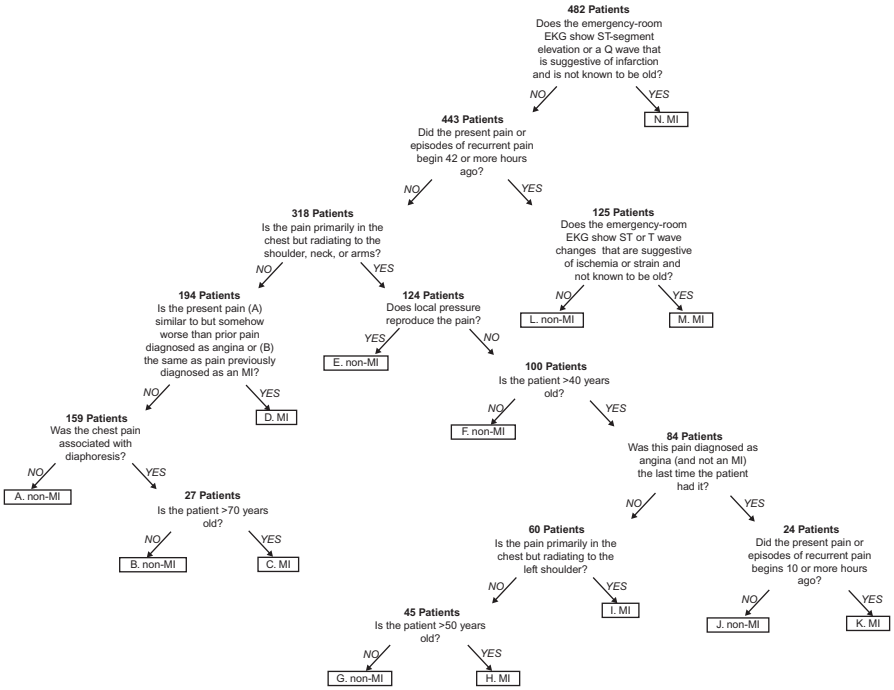

FIGURE 11.5

Computer-derived decision tree for the classification of patients with acute chest pain. Reproduced (with permission) from Goldman and colleagues.(Goldman et al., 1982) "Each of the 14 letters (A through N) identifies a terminal branch of the tree." In the Goldman study, seven terminal branches (C, D, H, I, K, M, and N) contained all the patients with acute myocardial infarction, along with a portion of the patients with other diagnoses.

Note that classification trees can solve nonlinearly separable problems, since the number of branches is not limited. However, given their limitation of using only univariate cuts at each branching point, there may be too many branches for the tree to be easy to interpret. Pruning algorithms have been developed to address this issue (Gelfand et al., 1991). The Goldman tree (shown in Figure 11.5) for deciding whether a patient with chest pain should be admitted to the emergency department is the prime example of an application of a classification tree (Goldman et al., 1982). This study identified nine important clinical factors that enabled the system to correctly categorize all (60) patients with myocardial infarction (MI) in the sample (482). Sensitivity was an absolute priority in this model, and a portion (71) of patients without MI were categorized as false positives. The clinical factors were: age, duration of pain, chest pain ± radiation, presence of diaphoresis, history of angina (and severity of pain) or prior MI, local pressure causes reproduction of pain, EKG ST-segment changes, Q waves, or T-wave changes not known to be old.

11.4.2 Artificial neural networks

The use of artificial neural networks (ANNs) has been reported in several medical domains, particularly in critical care (Frize et al., 2001; Dybowski et al., 1996; Tu and Guerriere, 1993; Kayaalp et al., 2000; Fraser and Turney, 1990) ANNs are highly flexible models composed of several processing units. Each of these units processes incoming information and may propagate information forward if warranted by their activation function. The most common activation function is the logistic, which has been presented in Section 11.3. The logistic function tries to model a step function that has been widely used to represent the electrical conduction in real neurons, which only propagate electric impulses if a certain threshold value is achieved. Although it is possible to build ANNs without utilizing an intermediate "hidden" layer of neurons, their flexibility comes from the inclusion of more than one nonlinear "hidden" node in this layer. In fact, the limitation of perceptrons, which were precursors to ANNs and were subject of much interest in the mid 1950s, was noted by several authors (Minsky and Papert, 1969). The same authors noted that multilayered perceptrons did not suffer from this limitation, but at that time there were no algorithms for estimating weights of multilayered perceptrons. The field was stagnant until Rumelhart et al. (1986) published the backpropagation algorithm in the mid 1980s. In the following two decades, a plethora of successful applications were reported in and out of the medical literature, but many of these research models did not translate into real clinical applications. Some, however, have been evaluated in real applications, such as the automated analysis of Pap smears (Baxt, 1991; O'Leary et al., 1998). There is no theoretical advantage of using ANNs over logistic regression in binary classification problems unless the ANNs have a hidden layer of nonlinear neurons. Hence, we will limit our discussion to this type of ANNs.

Figure 11.1 illustrated the similarities and differences between binary logistic regression and commonly used ANNs with a single output unit. ANNs and logistic regression models have several differences: (1) the activation function of the output unit needs not be the logistic in ANNs; (2) ANNs have intermediate processing units, often called hidden units or hidden nodes; and (3) ANNs can have multiple output units, so different classification problems can be modeled with a single network (although one could argue that polytomous logistic regression also allows for multiple outcomes to be modeled).

The hidden units in ANNs operate between the inputs and the outputs to process information to be sent to the output unit; Figure 11.6 illustrates how an ANN might solve the linearly nonseparable problem of classifying cases into "euthyroid" or not based on values of two laboratory tests, as illustrated in Figure 11.3. In this example, the activation functions of the intermediate layer of neurons correspond to the branching points that define the partitions of the classification tree presented in Section 11.4.1, but this will often not be the case. Furthermore, in the example we used step functions in the hidden layer. Step functions are not linear, and their combination offers a potential advantage over logistic regression, which does not have

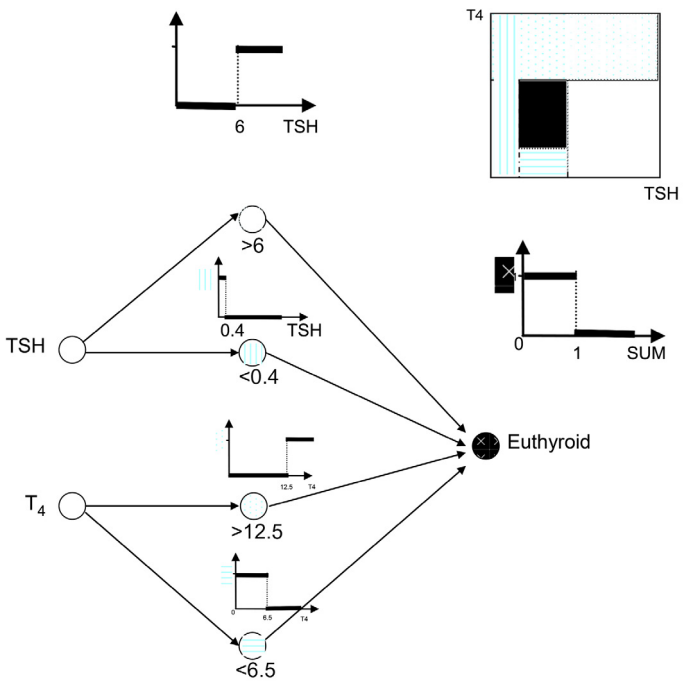

FIGURE 11.6

Artificial neural network with a hidden layer of nodes. For didactic purposes, activation functions in this example correspond to step functions that define partitions similar to the ones in the classification tree. Corresponding sigmoid (logistic) functions would be used in practice. As opposed to the classification tree example, the partitions here are overlapping. The outputs of the step functions are multiplied by their respective weights and combined as inputs to the output unit. The output unit has a step function that determines whether a case is "euthyroid" or not.

a hidden layer.[1] We use them here for illustration purposes, although we remind the readers that the most commonly used function in ANNs is the sigmoid function.

The outputs of these hidden layer functions are multiplied by the weights that lead into the output node, and summed to serve as input to the output node, which classifies cases into "euthyroid" or not. We do not represent every possible weight between the input layer and the hidden layer, so as to allow better visualization in the picture, but the reader can, for purposes of simplicity, assume here that the non-displayed connections are associated with null weights.

[1] If linear functions are used in the hidden layer, there is no advantage of ANN over logistic regression.

11.5 Prediction models in medicine

Many of the published medical classification and prediction models are research tools with limited utilization in clinical care. Even though early CDS systems addressed mostly diagnostic, test sequencing, and treatment choice aspects of clinical care, many of the most popular statistical and machine learning classification models in medicine used today address *prognostic* aspects of care.

In this section, we will discuss some applications of the modeling techniques described above, although each clinical example will not include all methods. In almost all cases, logistic regression modeling techniques are the most commonly reported and used in clinical practice for a variety of prognostication and classification purposes. The other techniques presented were generally compared to the standard of logistic regression, and rarely outperformed it. The lack of widespread use of a number of these models can be attributed to the lack of general knowledge of the methods and the greater complexity of the techniques. In many cases, the amount of data available does not allow the construction of complex models with many parameters, since there is a tendency of these models to overfit the data and not generalize well to new cases. As data becomes more abundant, this limitation is likely to play a smaller role.

The most common indices of model performance are *discrimination* and *calibration*. Discrimination assesses how well the models can potentially discriminate positive and negative cases in general. Models that estimate higher probabilities of outcome "1" for cases that had that outcome have high discrimination, which is usually measured as the area under the ROC curve (Hanley and McNeil, 1982). Calibration assesses how close the model's estimated probability is to the "true" underlying probability of the outcome of interest. Calibration in logistic regression models is usually assessed by a plot of the average estimate within a group to the expected ratio of the outcomes in that group. Typically, deciles are constructed based on the model's sorted estimates. The Hosmer-Lemeshow goodness-of-fit (HL-GOF) test is still the most widely used statistical test to assess calibration, although there are concerns that it can be biased when, for example, there are several ties among the estimates (Lemeshow and Hosmer, 1982).

Even models with high areas under the ROC curve can estimate probabilities that are very far from reality, or "uncalibrated." For example, the AUC remains unchanged with any monotonic transformation of the estimates (e.g. dividing all estimates by 10, which would decrease an estimate of 80% to 8%, and an estimate of 15% to 1.5%, but preserve the order of estimates, would result in no change in the AUC). A good example of this can be seen in the application of the MPM-II ICU mortality risk model on the California Intensive Care Outcomes Project (CALICO), where the AUC of the model on statewide data was 0.803 (0.790–0.815), but calibration was poor by both Observed/Expected Ratio, shown in Figure 11.7, as well as HL-GOF testing (CALICO. Final Report, 2007).

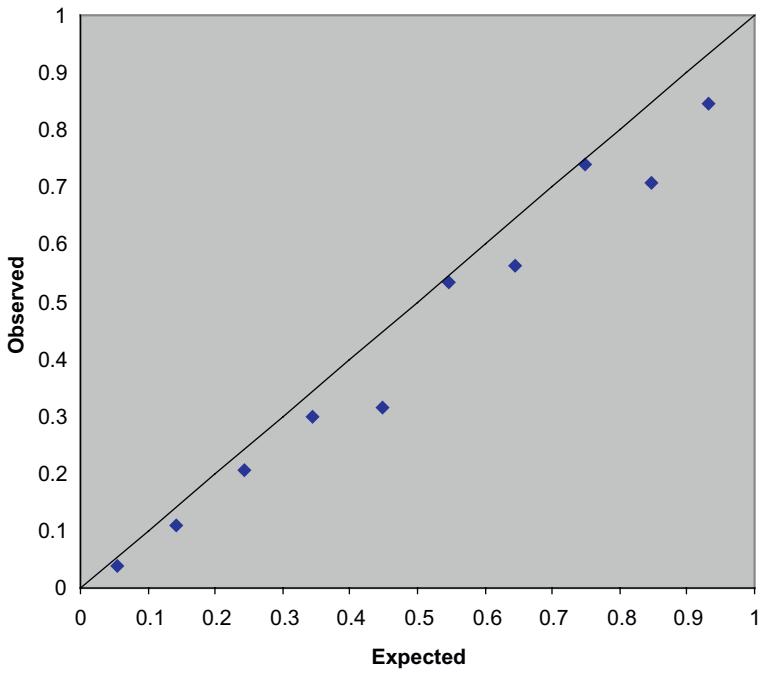

FIGURE 11.7

Observed/Expected (O/E) Ratio plot using the HL-GOF h test decile categories to show the divergence from the ideal (where the expected proportion of outcomes equals the observed proportion of events for each decile), particularly for higher risk patients.

11.5.1 Prognosis of ICU mortality

The Acute Physiology and Chronic Health Evaluation series of models (APACHE-II (Knaus et al., 1991) and APACHE-IV (Zimmerman et al., 2006)) constitute some of the most widely used logistic regression-based predictive models. These tools are used in intensive care units (ICUs) to predict in-hospital mortality based on a variety of physiologically-based variables. The initial version of APACHE (Knaus et al., 1981) was notable as the first clinical predictive model to exclusively use objective physiological parameters to predict outcome, and was an expert-based scoring system using these parameters to estimate the risk of outcome.

Both APACHE-II and APACHE-IV remain in use today for research, quality control, and clinical applications. APACHE-II was published in 1985 using a much larger development data set (5,815 admissions from 13 hospitals) than APACHE, and improved upon the expert-based scoring system with the inclusion of a logistic regression model using a patient's expert-based physiology score, emergency status, and adjustments for certain diagnostic categories. The model showed good discrimination on different independent evaluation sets (Jacobs et al., 1987;

Giangiuliani et al., 1989; Chisakuta and Alexander, 1990; Turner et al., 1991; Teskey et al., 1991; Wong et al., 1995), but its calibration was found to be highly variable. Since the model was made publicly available, it was used in many different validation studies.

APACHE-III was published in 1991, having been developed in response to criticisms regarding the case-mix and generalizability of APACHE-II. The system was developed from a database of 17,440 patients across 40 ICUs in the United States. APACHE-III was a commercial product, and was not made as easily available to the medical community at large as APACHE-II, but external evaluations conducted were similar to APACHE-II, indicating good discrimination and highly variable calibration (Zimmerman et al., 1998; Pappachan et al., 1999; Carneiro et al., 1997; von Bierbrauer et al., 1998; Bastos et al., 1996; Ihnsook et al., 2003; Rivera-Fernandez et al., 1998; Cook, 2000). APACHE-IV was introduced in 2006 as a large scale remodeling of APACHE-III and is also a commercial product. This remodeling effort included remodeling 42 of the 72 underlying APACHE III equations and the removal of 11 equations that were no longer appropriate, or no longer reflected in clinical practice (Zimmerman et al., 2006).

These models remain useful in research, but limitations in calibration and across disparate patient populations have restricted their use in some clinical situations (particularly with respect to application to individual patients). Other prognostic systems for the adult ICU, more common in Europe, are the Simplified Acute Physiologic Score SAPS-3, and the Mortality Prediction Model MPM-III. The Sequential Organ Failure Assessment SOFA model has also been used to assess organ function over time. These models or their earlier versions have been extensively compared all over the world in disparate patient populations. Several reviews and comparisons among these models have been published to date (Vincent et al., 1996; Ohno-Machado et al., 2006; Castella et al., 1991; Rowan et al., 1994; Wilairatana et al., 1995; Del Bufalo et al., 1995; Castella et al., 1995; Moreno et al., 1998; Nouira et al., 1998; Tan, 1998; Patel and Grant, 1999; Vassar et al., 1999; Katsaragakis et al., 2000; Livingston et al., 2000; Capuzzo et al., 2000; Markgraf et al., 2000; Beck et al., 2003; Keegan et al., 2012; Vasilevskis et al., 2009; Hwang et al., 2012; Costa e Silva et al., 2011; Shrope-Mok et al., 2010).

Multiple studies have compared logistic regression to artificial neural networks in this domain. Clermont and colleagues (Clermont et al., 2001) found that with a development data set of sufficient size (1,200), locally developed logistic regression and artificial neural networks performed equivalently in terms of both calibration (adequate) and discrimination (AUCs ranging from 0.80 to 0.84). However, both models experienced performance degradations as the development sample size decreased. Another smaller study with a development set of 168 undertaken by Dybowski and colleagues (Dybowski et al., 1996) showed superior discrimination of the ANN compared to LR (0.863 vs. 0.753 AUC, respectively).

Some studies have compared the APACHE-II LR model to ANNs. Nimgaonkar and colleagues (Nimgaonkar et al., 2004) found, after developing an ANN on 1,962 patients in an Indian ICU with the 22 APACHE-II variables, that the ANN had

superior discrimination to APACHE-II (0.87 vs. 0.77 AUC, respectively). Wong and colleagues (Wong and Young, 1999) performed a similar comparison with a development data set of 2,932 patients in the UK, and found that the two methods had equivalent discrimination (0.82 vs. 0.83 AUC for ANN and APACHE, respectively).

Comparisons of calibration were also done in some of the studies, but they were problematic because the LR model was developed on external patient populations disparate from the locally-derived UK and Indian populations utilized for the ANN models. Comparisons of discrimination do not suffer from this problem in the same way.

11.5.2 Cardiovascular disease risk

Another category of extremely well-known prediction tools in medicine provides estimates to patients of the risk of developing future heart disease. Although over 100 risk prediction models have been developed for this purpose, US medical practice has almost exclusively used the family of 10 year heart disease risk models developed from patients in one of the most famous patient cohorts who have been followed in the community of Framingham, Massachusetts, since the early 1950s. From oldest to most recent, these include models developed from the initial examination of the Framingham Offspring Study (FOS) (Kannel et al., 1979), the 11th examination of the original Framingham cohort (FC) (Anderson et al., 1991), or 1st examination of the FOS and 11th examination of the FC (Wilson et al., 1998), or the 1st or 3rd examination of the FOS and the 11th examination of the FC (D'Agostino et al., 2008). Worldwide, other well-known models have been developed from the PROCAM (Assmann et al., 2002) (Germany), UKPDS (Stevens et al., 2001) (United Kingdom), and QRESEARCH (Simmons et al., 2008; Hippisley-Cox et al., 2007) (United Kingdom) cohorts. All of the well-known models developed in this domain are based on (Cox and Oakes, 1984) proportional hazards and logistic regression methods.

The widespread use of these models is related to a number of key factors that influence the utility and generalizability of the prediction model. First, the modeled outcome is of paramount importance, since heart disease is the number one cause of mortality in the US, accounting for 23.8% of all deaths (Hoyert and Xu, 2012). Effective treatments exist for many of the outcome predictors, such as hypertension, hyperlipidemia, and smoking. Second, the patient population that was used in model development is in many ways representative of the American population. The Framingham cohort was an excellent source of data because the longitudinal nature of the cohort allowed reliable discrimination of patients at higher risk but who had not yet presented any sign or symptom of heart disease. One of the primary limitations of the cohort was the lack of racial diversity.

External validation of these models has shown good discrimination and moderate calibration, with some limitations when applied to populations with significantly different demographics and specific comorbidities (such as diabetes) (Guzder et al., 2005; Stephens et al., 2004; Lenz and Muhlhauser, 2004; Song and

Brown, 2004). Recalibration strategies have been used to remediate this problem (Ridker et al., 2007; Paynter et al., 2009). Perhaps more concerning in this domain is the tendency to externally validate the models with different outcome definitions, most commonly seen between more stringent primary modeling outcomes and more relaxed external validation outcomes (Ridker et al., 2007; Berry et al., 2007; Denes et al., 2007). For example, in some cases the models were developed on Hard CHD, which included sudden CHD death or myocardial infarction only, but then externally validated on Total CHD, which included Hard CHD outcomes as well as unstable angina and angina pectoris. In almost all cases, experts recommend remodeling rather than recalibration when outcome definitions need to be changed, even when the changes are minor.

These models have also clearly delineated the relative magnitude of various risk factors associated with heart disease, and have been used by a number of medical associates to establish guidelines of care (Grundy et al., 1998). In addition, the models are distributed as simple equations that can be quickly scanned by clinicians and patients, or embedded in calculators or computer-based software (Hingorani and Vallance, 1999).

11.5.3 Prognosis in interventional cardiology

Another widely studied area of risk prediction and stratification has been for the outcomes of death (Peterson et al., 2010) and significant morbidity (such as bleeding (Mehta et al., 2009), acute kidney injury (Brown et al., 2008), or unplanned 30 day readmission) in interventional cardiology. Risk modeling in this domain has been particularly popular for a number of reasons. First, most of the treatments (balloon angioplasty, coronary artery stenting, or atherectomy) are therapies directed at preventing myocardial infarction in patients who have developed significant heart disease. All of the remarks on the importance of this disease process in the prior section apply; these therapies can become necessary when prevention strategies fail.

Perhaps even more importantly, since interventional treatment is relatively intensive and done within a medical center, detailed data collection has been able to provide high-quality source data. The data quality has been also facilitated by the establishment of a national standard for the collection and storage of interventional cardiology data (Cannon et al., 2001), which has since undergone additional iterations in response to more detailed device identification needs and changing clinical practice. In addition, a number of the adverse outcomes associated with the treatment (or lack of treatment) are realized quickly. This is important because, in general, a model's performance is inversely related to the distance in time of the prediction from the occurrence of the outcome of interest. These factors have allowed the resulting models for this domain to attain high levels of discrimination.

Development of logistic regression prediction models for postprocedural mortality following angioplasty has followed a path similar to modeling in other medical domains. In general, the development populations were initially small and originated from a single center, which resulted in low generalizability

(Hannan et al., 1992; Resnic et al., 2001). These were followed by regional, multi-institutional models (Ellis et al., 1997; O'Connor et al., 1999; Hannan et al., 1997; Moscucci et al., 2001). Finally, the American College of Cardiology (ACC) aggregated data from centers across the United States to generate a mortality risk prediction model (Shaw et al., 2002).

These models have been externally validated on a number of independent data sets. In general, discrimination has remained excellent across disparate patient populations and over more than a decade of changing clinical care. However, as noted in many modeling domains, calibration is a problem and seems to be directly related to both the size of the development data, and how far in the past they were collected (Holmes et al., 2003; Holmes et al., 2000; Moscucci et al., 1999; Rihal et al., 2000; Singh et al., 2003; Matheny et al., 2005). For example, large changes in procedural care or the types of devices implanted, such as the introduction of heart valve replacement through cardiac catheterization (Tang et al., 2012; Durand et al., 2013), are likely to disrupt risk modeling performance and require development of updated models through remodeling or recalibration techniques. As awareness of the need for periodic updating of risk models has become more established, the ACC has pursued a remodeling strategy by conducting periodic updates to the risk models over the last decade, which has shown that some variables increase in significance while others fade to insignificance when predicting mortality (Peterson et al., 2010; Shaw et al., 2003).

11.5.4 Pneumonia severity-of-illness index

Finally, another logistic regression risk model example that has had a significant impact in the emergency department for both work flow (documentation requirements) and treatment is the Pneumonia Severity Index (PSI) developed from the Pneumonia Patient Outcomes Research Team (PORT) (Fine et al., 1997).

The team developed a prediction rule for the risk of death within 30 days for adult patients with community-acquired pneumonia. This disease is diagnosed in approximately four million adults each year in the US, and over 600,000 of the diagnosed patients are hospitalized (Garibaldi, 1985). The aggregate cost of hospitalization for this disease was estimated at four billion dollars per year (Dans et al. 1984; La Force 1985). The results of the PORT study suggested that, if the risk model had been used to treat patients based on the risk categories suggested, 26–31% of patients who had been hospitalized for care could have been treated safely as outpatients, and an additional 13–19% could have been hospitalized only for brief observation (Fine et al., 1997).

The key factors that led to the widespread use of this risk prediction tool were a combination of coinciding interest in evidence-based medical practice and in cost containment, as well as the high quality of the risk prediction tool. The model was validated on over 50,000 patients in 275 US and Canadian hospitals in the PORT study. Prior pneumonia risk prediction tools had suffered from small development population sizes (Daley et al., 1988; Keeler et al., 1990; Kurashi et al., 1992; Fine

et al., 1995) and limited external validation (Kurashi et al., 1992; Fine et al., 1995; Marrie et al., 1989).

The model has been widely used, and incorporated in both paper (Dean et al., 2000) and electronic (Aronsky et al., 2001) decision support tools for use in determining hospital admission from an emergency department. A number of subsequent multicenter randomized prospective studies have supported the use of the PSI as an appropriate admission tool (Marrie et al., 2000; Atlas et al., 1998). It was incorporated into the American Thoracic Society's (ATS) Community-Acquired Pneumonia guidelines (Niederman et al., 2001), although the society emphasized the limitations of the model in populations that were not well represented in the development data set (such as outpatient clinic patients), echoing findings from a few studies (Marras et al., 2000). PSI was incorporated into the Infectious Diseases Society of America/ATS consensus guidelines (Mandell et al., 2007) in 2007. Subsequent meta-analyses showed that the PSI has similar performance to CURB65 and CRB65, which are alternative tools (Chalmers et al., 2010; Loke et al., 2010; Chalmers et al., 2011). In addition to these tools, there are a number of factors that physicians must take into account. such as the presence of coexisting conditions, patients' preferences, and inadequate home support (Halm et al., 2000). Cooper and colleagues (Cooper et al., 2005) reported that several types of classifiers can achieve similar performance in this domain.

11.6 Conclusions

The utilization of statistical and machine learning techniques to discover knowledge from existing clinical data has become an integral component of biomedical informatics. The techniques for constructing and evaluating classification and prediction models are constantly evolving, and there are few theoretical justifications for preferring one learning technique over another. Some models, notably those constructed using logistic regression techniques, have been popularized in the medical domain, especially for research. These models span a limited number of specialties, and are for the most part concerned with prognostication. To our knowledge, there have been no formal large-scale studies documenting the utilization of these models by nonacademic clinicians at large. Even though some models are widely available on the web, there is currently no information on how many times they have actually been used in the provision of care. Several questions still remain:

- What types of data repositories can reasonably be used for medical pattern discoveries? Can data collected during clinical care be used to build decision support models? If so, what types of learning methods are adequate for sparse and noisy data?
- When can models originated from data of a single population be generalized to other populations? How can researchers assess the generalizability of such models?
- How can knowledge acquired from experts be integrated with knowledge discovered from real data?

None of these questions has been fully answered by the medical informatics community, but research in this area is encouraging. The popularity of some data-derived classification and prediction models, and their endorsement by health care institutions (an online model for assessing the risk of breast cancer is available at the NCI web site, for example), indicate that there is increasing interest in their use as diagnostic or prognostic tools. The availability of such models on the web also contributes to their utilization by the public at large.

It is important that clinicians utilize classification and prediction models. However, the integration of any computer system in the process of care is challenging. The electronic medical record is still not a reality in some settings in which medicine is practiced. The effective utilization of CDS systems depends on their seamless integration in a computer environment that is effectively used by practicing clinicians; hence it is premature to expect that predictive models will be largely utilized until this barrier is completely removed. The absence of a suitable computer environment is the first obstacle, but other issues also need further consideration. In order to provide counseling at the individual level, predictive models have to improve to the point that the uncertainty and imprecision of the estimates are acceptable from a clinical perspective. The poor calibration of estimates can be caused by limited representation, at the model construction phase, of the subpopulation to which models will be applied. Yet, collecting proper data from the institutions in which models are expected to be applied is not a trivial task. Until these types of deficiencies are properly acknowledged and fixed, and studies show that the predictive models perform at least at the same level as humans, the utilization of predictive models for individual care may remain limited. However, given the rapid pace of technological advances in biomedicine and the increasing utilization of computers by health care providers, it is expected that better models will continue to be developed which may soon be incorporated as additional tools in the provision of individualized care.

References

Anderson, K.M., Wilson, P.W., Odell, P.M., Kannel, W.B., 1991. An updated coronary risk profile. A statement for health professionals. Circulation 83 (1), 356–362.

Aronsky, D., Chan, K.J., Haug, P.J., 2001. Evaluation of a computerized diagnostic decision support system for patients with pneumonia: study design considerations. J. Am. Med. Inf. Assoc. 8 (5), 473–485.

Assmann, G., Cullen, P., Schulte, H., 2002. Simple scoring scheme for calculating the risk of acute coronary events based on the 10-year follow-up of the prospective cardiovascular munster (PROCAM) study. Circulation 105 (3), 310–315.

Atlas, S.J., Benzer, T.I., Borowsky, L.H., et al., 1998. Safely increasing the proportion of patients with community-acquired pneumonia treated as outpatients: an interventional trial. Arch. Internal Med. 158 (12), 1350–1356.

Barnett, G.O., Cimino, J.J., Hupp, J.A., Hoffer, E.P., 1987. DXplain. An evolving diagnostic decision-support system. JAMA 258 (1), 67–74.

Bastos, P.G., Sun, X., Wagner, D.P., Knaus, W.A., Zimmerman, J.E., 1996. Application of the APACHE III prognostic system in brazilian intensive care units: a prospective multicenter study. Intensive Care Med. 22 (6), 564–570.

Baxt, W.G., 1991. Use of an artificial neural network for the diagnosis of myocardial infarction. Ann. Intern. Med. 115, 845–848.

Beck, D.H., Smith, G.B., Pappachan, J.V., Millar, B., 2003. External validation of the SAPS II, APACHE II and APACHE III prognostic models in South England: a multicentre study. Intensive Care Med. 29 (2), 249–256.

Beinlich, I.A., Suermondt, H.J., Chavez, R.M., Cooper, G.F., 1989. The ALARM monitoring system: A case study with two probabilistic inference techniques for belief networks. Proc. Second Eur. Conf. Artif. Intell. Med., 247–256.

Berry, J.D., Lloyd-Jones, D.M., Garside, D.B., Greenland, P., 2007. Framingham risk score and prediction of coronary heart disease death in young men. Am. Heart J. 154 (1), 80–86.

von Bierbrauer, A., Riedel, S., Cassel, W., von Wichert, P., 1998. Validation of the acute physiology and chronic health evaluation (APACHE) III scoring system and comparison with APACHE II in german intensive care units]. Anaesthesist 47 (1), 30–38.

Boser, B., Guyon, I., Vapnik, V.A., 1992. Training algorithm for optimal margin classifiers. In: Proceedings of the Fifth Annual Workshop on Computational Learning Theory. July 27–29, Pittsburgh, PA, USA.

Breiman, L., Friedman, J., Olshen, R., Stone, C., 1984. Classification and Regression Trees. Wadsworth and Brooks.

Brown, J.R., DeVries, J.T., Piper, W.D., et al., 2008. Serious renal dysfunction after percutaneous coronary interventions can be predicted. Am. Heart J. 155 (2), 260–266.

Buntine, W.L., 1996. A guide to the literature on learning probabilistic networks from data. IEEE T Knowl. Data En. 8, 195–210.

CALICO. Final Report: California Intensive Care Outcomes Project (CALICO), 2007. <http://www.oshpd.ca.gov/HID/Products/PatDischargeData/ICUDataCALICO/CALICO_05-07.pdf> (accessed October 18.10.13.).

Cannon, C.P., Battler, A., Brindis, R.G., et al., 2001. American college of cardiology key data elements and definitions for measuring the clinical management and outcomes of patients with acute coronary syndromes. A report of the american college of cardiology task force on clinical data standards (Acute Coronary Syndromes Writing Committee). J. Am. Coll. Cardiol. 38 (7), 2114–2130.

Capuzzo, M., Valpondi, V., Sgarbi, A., et al., 2000. Validation of severity scoring systems SAPS II and APACHE II in a single-center population. Intensive Care Med. 26 (12), 1779–1785.

Carneiro, A.V., Leitao, M.P., Lopes, M.G., De Padua, F., 1997. Risk stratification and prognosis in critical surgical patients using the acute physiology, age and chronic health III system (APACHE III)]. Acta Med. Portuguesa 10 (11), 751–760.

Castella, X., Gilabert, J., Torner, F., Torres, C., 1991. Mortality prediction models in intensive care: acute physiology and chronic health evaluation II and mortality prediction model compared. Crit. Care Med. 19 (2), 191–197.

Castella, X., Artigas, A., Bion, J., Kari, A., 1995. A comparison of severity of illness scoring systems for intensive care unit patients: results of a multicenter, multinational study. The European/North American Severity Study Group. Crit. Care Med. 23 (8), 1327–1335.

Chalmers, J.D., Singanayagam, A., Akram, A.R., et al., 2010. Severity assessment tools for predicting mortality in hospitalised patients with community-acquired pneumonia. Systematic review and meta-analysis. Thorax 65 (10), 878–883.

Chalmers, J.D., Mandal, P., Singanayagam, A., et al., 2011. Severity assessment tools to guide ICU admission in community-acquired pneumonia: systematic review and meta-analysis. Intensive Care Med. 37 (9), 1409–1420.

Chisakuta, A.M., Alexander, J.P., 1990. Audit in intensive care. The APACHE II classification of severity of disease. Ulster Med. J. 59 (2), 161–167.

Clancey, W.J., 1983. The epistemology of a rule-based expert system: a framework for explanation. Artif. Intell. 20, 215–251.

Clermont, G., Angus, D.C., DiRusso, S.M., Griffin, M., Linde-Zwirble, W.T., 2001. Predicting hospital mortality for patients in the intensive care unit: a comparison of artificial neural networks with logistic regression models. Crit. Care Med. 29 (2), 291–296.

Cook, D.A., 2000. Performance of APACHE III models in an Australian ICU. Chest 118 (6), 1732–1738.

Cooper, G., Herskovits, E.A., 1992. Bayesian method for the induction of probabilistic networks from data. Mach. Learn. 9, 309–347.

Cooper, G.F., Abraham, V., Aliferis, C.F., et al., 2005. Predicting dire outcomes of patients with community acquired pneumonia. J Biomed. Inform. 38, 347–366.

Costa e Silva, V.T., de Castro, I., Liano, F., Muriel, A., Rodriguez-Palomares, J.R., Yu, L., 2011. Performance of the third-generation models of severity scoring systems (APACHE IV, SAPS 3 and MPM-III) in acute kidney injury critically ill patients. Nephrol. Dial. Transplant. 26 (12), 3894–3901.

Cox, D.R., Oakes, D., 1984. Analysis of Survival Data. Chapman & Hall, New York.

D'Agostino Sr., R.B., Vasan, R.S., Pencina, M.J., et al., 2008. General cardiovascular risk profile for use in primary care: the framingham heart study. Circulation 117 (6), 743–753.

Daley, J., Jencks, S., Draper, D., Lenhart, G., Thomas, N., Walker, J., 1988. Predicting hospital-associated mortality for Medicare patients. A method for patients with stroke, pneumonia, acute myocardial infarction, and congestive heart failure. JAMA 260 (24), 3617–3624.

Dans, P.E., Charache, P., Fahey, M., Otter, S.E., 1984. Management of pneumonia in the prospective payment era. A need for more clinician and support service interaction. Arch. Internal Med. 144 (7), 1392–1397.

Das, M., Eichner, J., March 2010. Challenges and Barriers to Clinical Decision Support (CDS) Design and Implementation Experienced in the Agency for Healthcare Research and Quality CDS Demonstrations (Prepared for the AHRQ National Resource Center for Health Information Technology under Contract No. 290-04-0016.) AHRQ Publication No. 10-0064-EF. Rockville, MD: Agency for Healthcare Research and Quality.

Dean, N.C., Suchyta, M.R., Bateman, K.A., Aronsky, D., Hadlock, C.J., 2000. Implementation of admission decision support for community-acquired pneumonia. Chest 117 (5), 1368–1377.

Del Bufalo, C., Morelli, A., Bassein, L., et al., 1995. Severity scores in respiratory intensive care: APACHE II predicted mortality better than SAPS II. Respir. Care 40 (10), 1042–1047.

Denes, P., Larson, J.C., Lloyd-Jones, D.M., Prineas, R.J., Greenland, P., 2007. Major and minor ECG abnormalities in asymptomatic women and risk of cardiovascular events and mortality. JAMA 297 (9), 978–985.

Duda, R., Hart, P., Stork, D., 2001. Pattern Classification, second ed.. Wiley Interscience.

Durand, E., Borz, B., Godin, M., et al., 2013. Performance analysis of EuroSCORE II compared to the original logistic EuroSCORE and STS scores for predicting 30-Day mortality after transcatheter aortic valve replacement. Am. J. Cardiol. 111 (6), 891–897.

Dybowski, R., Weller, P., Chang, R., Gant, V., 1996. Prediction of outcome in critically ill patients using artificial neural network synthesised by genetic algorithm. Lancet 347 (9009), 1146–1150.

Ellis, S.G., Weintraub, W., Holmes, D., Shaw, R., Block, P.C., King III, S.B., 1997. Relation of operator volume and experience to procedural outcome of percutaneous coronary revascularization at hospitals with high interventional volumes. Circulation 95 (11), 2479–2484.

Fine, M.J., Hanusa, B.H., Lave, J.R., et al., 1995. Comparison of a disease-specific and a generic severity of illness measure for patients with community-acquired pneumonia. J. Gen.Internal Med. 10 (7), 359–368.

Fine, M.J., Auble, T.E., Yealy, D.M., et al., 1997. A prediction rule to identify low-risk patients with community-acquired pneumonia. N. Engl. J. Med. 336 (4), 243–250.

Fraser, R.B., Turney, S.Z., 1990. An expert system for the nutritional management of the critically ill. Comput. Methods Prog. Biomed. 33 (3), 175–180.

Frize, M., Ennett, C.M., Stevenson, M., Trigg, H.C., 2001. Clinical decision support systems for intensive care units: using artificial neural networks. Med. Eng. Phys. 23 (3), 217–225.

Garibaldi, R.A., 1985. Epidemiology of community-acquired respiratory tract infections in adults. Incidence, etiology, and impact. Am. J. Med. 78 (6B), 32–37.

Gelfand, S.B., Ravishankar, C.S., Delp, E.J., 1991. An iterative growing and pruning algorithm for classification tree design. IEEE Trans. Pattern Anal. Mach. Intell. 13, 163–174.

Giangiuliani, G., Mancini, A., Gui, D., 1989. Validation of a severity of illness score (APACHE II) in a surgical intensive care unit. Intensive Care Med. 15 (8), 519–522.

Goldman, L., Weinberg, M., Weisberg, M., et al., 1982. A computer-derived protocol to aid in the diagnosis of emergency room patients with acute chest pain. NEJM 307, 588–596.

Grundy, S.M., Balady, G.J., Criqui, M.H., et al., 1998. Primary prevention of coronary heart disease: guidance from Framingham: a statement for healthcare professionals from the AHA task force on risk reduction. Am. Heart Assoc. Circulation 97 (18), 1876–1887.

Grundy, S.M., Pasternak, R., Greenland, P., Smith Jr., S., Fuster, V., 1999. Assessment of cardiovascular risk by use of multiple-risk-factor assessment equations: a statement for healthcare professionals from the american heart association and the american college of cardiology. Circulation 100 (13), 1481–1492.

Guzder, R.N., Gatling, W., Mullee, M.A., Mehta, R.L., Byrne, C.D., 2005. Prognostic value of the Framingham cardiovascular risk equation and the UKPDS risk engine for coronary heart disease in newly diagnosed Type 2 diabetes: results from a united kingdom study. Diabetic Med. 22 (5), 554–562.

Halm, E.A., Atlas, S.J., Borowsky, L.H., et al., 2000. Understanding physician adherence with a pneumonia practice guideline: effects of patient, system, and physician factors. Arch. Internal Med. 160 (1), 98–104.

Han, Y.Y., Carcillo, J.A., Venkataraman, S.T., et al., 2005. Unexpected increased mortality after implementation of a commercially sold computerized physician order entry system. Pediatrics 116 (6), 1506–1512.

Hanley, J.A., McNeil, B.J., 1982. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 143 (1), 29–36.

Hannan, E.L., Arani, D.T., Johnson, L.W., Kemp Jr., H.G., Lukacik, G., 1992. Percutaneous transluminal coronary angioplasty in new york state. Risk factors and outcomes. JAMA 268 (21), 3092–3097.

Hannan, E.L., Racz, M., Ryan, T.J., et al., 1997. Coronary angioplasty volume-outcome relationships for hospitals and cardiologists. JAMA 277 (11), 892–898.

Hastie, T., Tibshirani, R., Friedman, J., 2001. The Elements of Statistical Learning. Springer, New York, NY.

Heckerman, D., 1990. A tractable inference algorithm for diagnosing multiple diseases. Proc. Fifth Conf. Uncertainty Artif. Intell., 163–171.

Heckerman, D., Miller, R.A., 1986. Towards a better understanding of the INTERNIST-1 knowledge base. Medinfo 86.

Hingorani, A.D., Vallance, P., 1999. A simple computer program for guiding management of cardiovascular risk factors and prescribing. BMJ 318 (7176), 101–105.

Hippisley-Cox, J., Coupland, C., Vinogradova, Y., Robson, J., May, M., Brindle, P., 2007. Derivation and validation of QRISK, a new cardiovascular disease risk score for the united kingdom: prospective open cohort study. BMJ 335 (7611), 136.

Holmes Jr., D.R., Berger, P.B., Garratt, K.N., et al., 2000. Application of the new york state PTCA mortality model in patients undergoing stent implantation. Circulation 102 (5), 517–522.

Holmes, D.R., Selzer, F., Johnston, J.M., et al., 2003. Modeling and risk prediction in the current era of interventional cardiology: a report from the national heart, lung, and blood institute dynamic registry. Circulation 107 (14), 1871–1876.

Hosmer, D.W., Lemeshow, S., 1989. Applied Logistic Regression. Wiley, New York.

Hoyert, D.L., Xu, J., 2012. Deaths: Preliminary Data for 2011. CDC/NCHS, Nat. Vital Stat. Syst. 61 (6), 52.

Hwang, S.Y., Lee, J.H., Lee, Y.H., Hong, C.K., Sung, A.J., Choi, Y.C., 2012. Comparison of the sequential organ failure assessment, acute physiology and chronic health evaluation II scoring system, and trauma and injury severity score method for predicting the outcomes of intensive care unit trauma patients. Am. J. Emerg. Med. 30 (5), 749–753.

Ihnsook, J., Myunghee, K., Jungsoon, K., 2003. Predictive accuracy of severity scoring system: a prospective cohort study using APACHE III in a Korean intensive care unit. Int. J. Nursing Stud. 40 (3), 219–226.

Jacobs, S., Chang, R.W., Lee, B., 1987. One year's experience with the APACHE II severity of disease classification system in a general intensive care unit. Anaesthesia 42 (7), 738–744.

Kannel, W.B., Feinleib, M., McNamara, P.M., Garrison, R.J., Castelli, W.P., 1979. An investigation of coronary heart disease in families. The framingham offspring study. Am. J. Epidemiol. 110 (3), 281–290.

Katsaragakis, S., Papadimitropoulos, K., Antonakis, P., Strergiopoulos, S., Konstadoulakis, M.M., Androulakis, G., 2000. Comparison of acute physiology and chronic health evaluation II (APACHE II) and simplified acute physiology score II (SAPS II) scoring systems in a single Greek intensive care unit. Crit. Care Med. 28 (2), 426–432.

Kayaalp, M., Cooper, G.F., Clermont, G., 2000. Predicting ICU mortality: a comparison of stationary and nonstationary temporal models. Proc. AMIA., 418–422. Annual Symposium.

Keegan, M.T., Gajic, O., Afessa, B., 2012. Comparison of APACHE III and IV, SAPS 3 and MPM-III, and Influence of Resuscitation Status on Model Performance. Chest 142 (4), 851–858.

Keeler, E.B., Kahn, K.L., Draper, D., et al., 1990. Changes in sickness at admission following the introduction of the prospective payment system. JAMA 264 (15), 1962–1968.

Knaus, W.A., Zimmerman, J.E., Wagner, D.P., Draper, E.A., Lawrence, D.E., 1981. APACHE-acute physiology and chronic health evaluation: a physiologically based classification system. Crit. Care Med. 9 (8), 591–597.

Knaus, W.A., Draper, E.A., Wagner, D.P., Zimmerman, J.E., 1985. APACHE II: a severity of disease classification system. Crit. Care Med. 13 (10), 818–829.

Knaus, W.A., Wagner, D.P., Draper, E.A., et al., 1991. The APACHE III prognostic system. Risk prediction of hospital mortality for critically ill hospitalized adults. Chest 100 (6), 1619–1636.

Koza, J.R., 1992. Genetic Programming: on the Programming of Computers by Means of Natural Selection. MIT Press, Cambridge, MA.

Kurashi, N.Y., al-Hamdan, A., Ibrahim, E.M., al-Idrissi, H.Y., al-Bayari, T.H., 1992. Community acquired acute bacterial and atypical pneumonia in Saudi Arabia. Thorax 47 (2), 115–118.

La Force, F.M., 1985. Community-acquired lower respiratory tract infections. Prevention and cost-control strategies. Am. J. Med. 78 (6B), 52–57.

Lemeshow, S., Hosmer Jr., D.W., 1982. A review of goodness of fit statistics for use in the development of logistic regression models. Am. J. Epidemiol. 115 (1), 92–106.

Lemeshow, S., Le Gall, J.R., 1994. Modeling the severity of illness of ICU patients. a systems update. JAMA 272 (13), 1049–1055.

Lenz, M., Muhlhauser, I., 2004. Cardiovascular risk assessment for informed decision making. Validity of prediction tools. Med. Klinik. 99 (11), 651–661.

Livingston, B.M., MacKirdy, F.N., Howie, J.C., Jones, R., Norrie, J.D., 2000. Assessment of the performance of five intensive care scoring models within a large Scottish database. Crit. Care Med. 28 (6), 1820–1827.

Loke, Y.K., Kwok, C.S., Niruban, A., Myint, P.K., 2010. Value of severity scales in predicting mortality from community-acquired pneumonia: systematic review and meta-analysis. Thorax 65 (10), 884–890.

Mandell, L.A., Wunderink, R.G., Anzueto, A., et al., 2007. Infectious Diseases Society of America/American Thoracic Society consensus guidelines on the management of community-acquired pneumonia in adults. Clin. Infect Dis. 44 (Suppl. 2), S27–S72.

Markgraf, R., Deutschinoff, G., Pientka, L., Scholten, T., 2000. Comparison of acute physiology and chronic health evaluations II and III and simplified acute physiology score II: a prospective cohort study evaluating these methods to predict outcome in a german interdisciplinary intensive care unit. Crit. Care Med. 28 (1), 26–33.

Marras, T.K., Gutierrez, C., Chan, C.K., 2000. Applying a prediction rule to identify low-risk patients with community-acquired pneumonia. Chest 118 (5), 1339–1343.

Marrie, T.J., Durant, H., Yates, L., 1989. Community-acquired pneumonia requiring hospitalization: 5-year prospective study. Rev. Infect. Dis. 11 (4), 586–599.

Marrie, T.J., Lau, C.Y., Wheeler, S.L., Wong, C.J., Vandervoort, M.K., Feagan, B.G., 2000. A controlled trial of a critical pathway for treatment of community-acquired pneumonia. CAPITAL Study Investigators. Community-Acquired pneumonia intervention Trial assessing levofloxacin.(see comment). JAMA 283 (6), 749–755.

Matheny, M.E., Ohno-Machado, L., Resnic, F.S., 2005. Discrimination and calibration of mortality risk prediction models in interventional cardiology. J. Biomed. Inform. 38 (5), 367–375.

Mehta, S.K., Frutkin, A.D., Lindsey, J.B., et al., 2009. Bleeding in patients undergoing percutaneous coronary intervention: the development of a clinical risk algorithm from the national cardiovascular data registry. Circ. Cardiovasc. Interv. 2 (3), 222–229.

Miller, R., Masarie, F.E., Myers, J.D., 1986. Quick medical reference (QMR) for diagnostic assistance. MD Comput. 3 (5), 34–48.

Miller, R.A., Pople Jr., H.E., Myers, J.D., 1982. Internist-1, an experimental computer-based diagnostic consultant for general internal medicine. N. Engl. J. Med. 307 (8), 468–476.

Minsky, M.L., Papert, S., 1969. Perceptrons. MIT Press, Cambridge, MA.

Moore, A., Lee, M.S., 1998. Cached sufficient statistics for efficient machine learning with large datasets. JAIR 8, 67–91.

Moreno, R., Apolone, G., Miranda, D.R., 1998. Evaluation of the uniformity of fit of general outcome prediction models. Intensive Care Med. 24 (1), 40–47.

Moscucci, M., O'Connor, G.T., Ellis, S.G., et al., 1999. Validation of risk adjustment models for in-hospital percutaneous transluminal coronary angioplasty mortality on an independent data set. J. Am. Coll. Cardiol. 34 (3), 692–697.

Moscucci, M., Kline-Rogers, E., Share, D., et al., 2001. Simple bedside additive tool for prediction of in-hospital mortality after percutaneous coronary interventions. Circulation 104 (3), 263–268.

Niederman, M.S., Mandell, L.A., Anzueto, A., et al., 2001. Guidelines for the management of adults with community-acquired pneumonia. Diagnosis, assessment of severity, antimicrobial therapy, and prevention. Am. J. Respir. Crit. Care Med. 163 (7), 1730–1754.

Nimgaonkar, A., Karnad, D.R., Sudarshan, S., Ohno-Machado, L., Kohane, I., 2004. Prediction of mortality in an Indian intensive care unit. Comparison between APACHE II and artificial neural networks. Intensive Care Med. 30 (2), 248–253.

Nouira, S., Belghith, M., Elatrous, S., et al., 1998. Predictive value of severity scoring systems: comparison of four models in tunisian adult intensive care units. Crit. Care Med. 26 (5), 852–859.

O'Connor, G.T., Malenka, D.J., Quinton, H., et al., 1999. Multivariate prediction of in-hospital mortality after percutaneous coronary interventions in 1994–1996. Northern New England Cardiovascular Disease Study Group. J. Am. Coll. Cardiol. 34 (3), 681–691.

O'Leary, T.J., Tellado, M., Buckner, S.B., Ali, I.S., Stevens, A., Ollayos, C.W., 1998. PAPNET-assisted rescreening of cervical smears: cost and accuracy compared with a 100% manual rescreening strategy. JAMA 279, 235–237.

Ohno-Machado, L., Resnic, F.S., Matheny, M.E., 2006. Prognosis in Critical Care In: Yarmush, M.L. Diller, K.R. (Eds.), Annual Review of Biomedical Engineering, Vol 8 Nonprofit Publisher of the Annual Review of TM Series, Palo Alto, CA.

Pappachan, J.V., Millar, B., Bennett, E.D., Smith, G.B., 1999. Comparison of outcome from intensive care admission after adjustment for case mix by the APACHE III prognostic system. Chest 115 (3), 802–810.

Patel, P.A., Grant, B.J., 1999. Application of mortality prediction systems to individual intensive care units. Intensive Care Med. 25 (9), 977–982.

Pauker, S.G., Gorry, G.A., Kassirer, J.P., Schwartz, W.B., 1976. Towards the simulation of clinical cognition. Taking a present illness by computer. Am. J. Med. 60 (7), 981–996.

Pawlak, Z., 1982. Rough sets. Int. J. Inf. Comput. Sci. 11, 341–356.

Paynter, N.P., Chasman, D.I., Buring, J.E., Shiffman, D., Cook, N.R., Ridker, P.M., 2009. Cardiovascular disease risk prediction with and without knowledge of genetic variation at chromosome 9p21.3. Ann. Intern. Med. 150 (2), 65–72.

Pearl, J., 1988. Probabilistic Reasoning in Intelligent Systems. Morgan-Kaufmann, San Mateo, CA.

Peterson, E.D., Dai, D., DeLong, E.R., et al., 2010. Contemporary mortality risk prediction for percutaneous coronary intervention results from 588,398 procedures in the national cardiovascular data registry. J. Am. Coll. Cardiol. 55 (18), 1923–1932.

Pollock, D.A., Adams, D.L., Bernardo, L.M., et al., 1998. Data elements for emergency department systems, release 1.0 (DEEDS): a summary report. DEEDS writing committee. J. Emergency Nursing. 24 (1), 35–44.

Resnic, F.S., Ohno-Machado, L., Selwyn, A., Simon, D.I., Popma, J.J., 2001. Simplified risk score models accurately predict the risk of major in-hospital complications following percutaneous coronary intervention. Am. J. Cardiol. 88 (1), 5–9.

Ridker, P.M., Buring, J.E., Rifai, N., Cook, N.R., 2007. Development and validation of improved algorithms for the assessment of global cardiovascular risk in women: the Reynolds Risk Score. JAMA 297 (6), 611–619.

Rihal, C.S., Grill, D.E., Bell, M.R., Berger, P.B., Garratt, K.N., Holmes Jr., D.R., 2000. Prediction of death after percutaneous coronary interventional procedures. Am. Heart J. 139 (6), 1032–1038.

Rivera-Fernandez, R., Vazquez-Mata, G., Bravo, M., et al., 1998. The Apache III prognostic system: customized mortality predictions for Spanish ICU patients. Intensive Care Med. 24 (6), 574–581.

Romano, M.J., Stafford, R.S., 2011. Electronic health records and clinical decision support systems: impact on national ambulatory care quality. Arch. Intern. Med. 171 (10), 897–903.

Rowan, K.M., Kerr, J.H., Major, E., McPherson, K., Short, A., Vessey, M.P., 1994. Intensive Care society's acute physiology and chronic health evaluation (APACHE II) study in britain and Ireland: a prospective, multicenter, cohort study comparing two methods for predicting outcome for adult intensive care patients. Crit. Care Med. 22 (9), 1392–1401.

Rumelhart, D.E., Hinton, G.E., Williams, R.J., 1986. Learning representations by back-propagating errors. Nature 323, 533–536.

Saverno, K.R., Hines, L.E., Warholak, T.L., et al., 2011. Ability of pharmacy clinical decision-support software to alert users about clinically important drug-drug interactions. J. Am. Med. Inform Assoc. 18 (1), 32–37.

Shaw, R.E., Anderson, H.V., Brindis, R.G., et al., 2002. Development of a risk adjustment mortality model using the american college of cardiology-national cardiovascular data registry (ACC-NCDR) experience: 1998–2000. J. Am. Coll. Cardiol. 39 (7), 1104–1112.

Shaw, R.E., Anderson, H.V., Brindis, R.G., et al., 2003. Updated risk adjustment mortality model using the complete 1.1 dataset from the American College of Cardiology National Cardiovascular Data Registry (ACC-NCDR). J. Invasive Cardiol. 15 (10), 578–580.

Shortliffe, E.H., 1976. Computer-Based Medical Consultations, MYCIN. Elsevier, New York, NY.

Shortliffe, EH, Davis, R, Axline, SG, Buchanan, BG, Green, CC, Cohen, SN., 1975. Computer-based consultations in clinical therapeutics: explanation and rule acquisition capabilities of the MYCIN system. Comput. Biomed. Res. 8 (4), 303–320.

Shrope-Mok, S.R., Propst, K.A., Iyengar, R., 2010. APACHE IV versus PPI for predicting community hospital ICU mortality. Am. J. Hosp. Palliat Care 27 (4), 243–247.

Simmons, R.K., Sharp, S., Boekholdt, S.M., et al., 2008. Evaluation of the Framingham risk score in the European Prospective Investigation of Cancer-Norfolk cohort: does adding glycated hemoglobin improve the prediction of coronary heart disease events? Arch. Intern Med. 168 (11), 1209–1216.

Singh, M., Rihal, C.S., Selzer, F., Kip, K.E., Detre, K., Holmes, D.R., 2003. Validation of mayo clinic risk adjustment model for in-hospital complications after percutaneous

coronary interventions, using the national heart, Lung, and blood institute dynamic registry. J. Am. Coll. Cardiol. 42 (10), 1722–1728.
Song, S.H., Brown, P.M., 2004. Coronary heart disease risk assessment in diabetes mellitus: comparison of UKPDS risk engine with Framingham risk assessment function and its clinical implications. Diabetic Med. 21 (3), 238–245.
Stephens, J.W., Ambler, G., Vallance, P., Betteridge, D.J., Humphries, S.E., Hurel, S.J., 2004. Cardiovascular risk and diabetes. Are the methods of risk prediction satisfactory? Eur. J. Cardiovasc. Prev. Rehabil. 11 (6), 521–528.
Stevens, R.J., Kothari, V., Adler, A.I., Stratton, I.M., 2001. The UKPDS risk engine: a model for the risk of coronary heart disease in Type II diabetes (UKPDS 56). Clin. Sci. (Lond) 101 (6), 671–679.
Swhe, M., Middleton, B., Heckerman, D., Henrion, M., Horvitz, E., Lehmann, H., 1991. Probabilistic diagnosis using a reformulation of the INTERNIST-1/QMR knowledge base I: the probabilistic model and inference algorithms. Methods Inf. Med. 30, 241–255.
Tan, I.K., 1998. APACHE II and SAPS II are poorly calibrated in a Hong Kong intensive care unit. Ann. Acad. Med., Singapore 27 (3), 318–322.
Tang, G.H., Lansman, S.L., Cohen, M., et al., 2012. Transcatheter aortic valve implantation: current developments, ongoing issues, future outlook. Cardiol Rev. 21 (2), 55–76.
Teskey, R.J., Calvin, J.E., McPhail, I., 1991. Disease severity in the coronary care unit. Chest 100 (6), 1637–1642.
Tu, J.V., Guerriere, M.R., 1993. Use of a neural network as a predictive instrument for length of stay in the intensive care unit following cardiac surgery. Comput. Biomed. Res. 26 (3), 220–229.
Turner, J.S., Mudaliar, Y.M., Chang, R.W., Morgan, C.J., 1991. Acute physiology and chronic health evaluation (APACHE II) scoring in a cardiothoracic intensive care unit. Crit. Care Med. 19 (10), 1266–1269.
Vapnik, V.N., 1995. The Nature of Statistical Learning Theory. Springer-Verlag, New York, NY.
Vasilevskis, E.E., Kuzniewicz, M.W., Cason, B.A., et al., 2009. Mortality probability model III and simplified acute physiology score II: assessing their value in predicting length of stay and comparison to APACHE IV. Chest 136 (1), 89–101.
Vassar, M.J., Lewis Jr., F.R., Chambers, J.A., et al., 1999. Prediction of outcome in intensive care unit trauma patients: a multicenter study of acute physiology and chronic health evaluation (APACHE), trauma and injury severity score (TRISS), and a 24-hour intensive care unit (ICU) point system. J. Trauma-Inj. Infection Crit. Care 47 (2), 324–329.
Vincent, J.L., Moreno, R., Takala, J., et al., 1996. The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure. On behalf of the working group on sepsis-related problems of the european society of intensive care medicine. Intensive Care Med. 22 (7), 707–710.
Wattigney, W.A., Croft, J.B., Mensah, G.A., et al., 2003. Establishing data elements for the paul coverdell national acute stroke registry: part 1: proceedings of an expert panel. Stroke 34 (1), 151–156.
Wilairatana, P., Noan, N.S., Chinprasatsak, S., Prodeengam, K., Kityaporn, D., Looareesuwan, S., 1995. Scoring systems for predicting outcomes of critically ill patients in northeastern thailand. Southeast Asian J. Trop. Med. Public Health 26 (1), 66–72.

Wilson, P.W., D'Agostino, R.B., Levy, D., Belanger, A.M., Silbershatz, H., Kannel, W.B., 1998. Prediction of coronary heart disease using risk factor categories. Circulation 97 (18), 1837–1847.

Wong, D.T., Crofts, S.L., Gomez, M., McGuire, G.P., Byrick, R.J., 1995. Evaluation of predictive ability of APACHE II system and hospital outcome in Canadian intensive care unit patients. Crit. Care Med. 23 (7), 1177–1183.

Wong, L.S., Young, J.D., 1999. A comparison of ICU mortality prediction using the APACHE II scoring system and artificial neural networks. Anaesthesia 54 (11), 1048–1054.

Zadeh, L.A., 1994. Fuzzy logic, neural networks, and soft computing. Commun. ACM. 37, 77–84.

Zimmerman, J.E., Wagner, D.P., Draper, E.A., Wright, L., Alzola, C., Knaus, W.A., 1998. Evaluation of acute physiology and chronic health evaluation III predictions of hospital mortality in an independent database. Crit. Care Med. 26 (8), 1317–1326.

Zimmerman, J.E., Kramer, A.A., McNair, D.S., Malila, F.M., 2006. Acute physiology and chronic health evaluation (APACHE) IV: hospital mortality assessment for today's critically ill patients. Crit. Care Med. 34 (5), 1297–1310.