Nonhypothesis-Driven Research: Data Mining and Knowledge Discovery

16

Mollie R. Cummins

Abstract

Clinical information, stored over time, is a potentially rich source of data for clinical research. Knowledge discovery in databases (KDD), commonly known as data mining, is a process for pattern discovery and predictive modeling in large databases. KDD makes extensive use of data mining methods, automated processes, and algorithms that enable pattern recognition. Characteristically, data mining involves the use of machine learning methods developed in the domain of artificial intelligence. These methods have been applied to healthcare and biomedical data for a variety of purposes with good success and potential or realized clinical translation. Herein, the Fayyad model of knowledge discovery in databases is introduced. The steps of the process are described with select examples from clinical research informatics. These steps range from initial data selection to interpretation and evaluation. Commonly used data mining methods are surveyed: artificial neural networks, decision tree induction, support vector machines (kernel methods), association rule induction, and k-nearest neighbor. Methods for evaluating the models that result from the KDD process are closely linked to methods used in diagnostic medicine. These include the use of measures derived from a confusion matrix and receiver operating characteristic curve analysis. Data partitioning and model validation are critical aspects of evaluation. International efforts to develop and refine clinical data repositories are critically linked to the potential of these methods for developing new knowledge.

Keywords

Knowledge discovery in databases · Data mining · Artificial neural networks Support vector machines · Decision trees · *k*-Nearest neighbor classification Clinical data repositories

M. R. Cummins, PhD, RN (✉)
College of Nursing, University of Utah, Salt Lake City, UT, USA
e-mail: mollie.cummins@utah.edu

R. L. Richesson, J. E. Andrews (eds.), *Clinical Research Informatics*,
Health Informatics, https://doi.org/10.1007/978-3-319-98779-8_16

Clinical information, stored over time, is a potentially rich source of data for clinical research. Many of the concepts that would be measured in a prospective study are already collected in the course of routine healthcare. Based on comparisons of treatment effects, some believe well-designed case-control or cohort studies produce results equally rigorous to that of randomized controlled trials, with lower cost and with broader applicability [1]. While this potential has not yet been fully realized, the rich potential of clinical data repositories for building knowledge is undeniable. Minimally, analysis of routinely collected data can aid in hypothesis generation and refinement and partially replace expensive prospective data collection.

While smaller samples of data can be extracted for observational studies of clinical phenomena, there is also an opportunity to learn from the much larger, accumulated mass of data. The availability of so many instances of disease states, health behaviors, and other clinical phenomena bears an opportunity to find novel patterns and relationships. In an exploratory approach, the data itself can be used to fuel hypothesis development and subsequent research. Importantly, one can induce executable knowledge models directly from clinical data, predictive models that can be implemented in computerized decision support systems [2, 3]. However, the statistical approaches used in cohort and case-control studies of small samples are not appropriate for large-scale pattern discovery and predictive modeling, where bias can figure more prominently, data can fail to satisfy key assumptions, and p values can become misleading.

Knowledge discovery in databases (KDD), also commonly known as data mining, is the process for pattern discovery and predictive modeling in large databases. An iterative, exploratory process distinctly differs from traditional statistical analysis in that it involves a great deal of interaction and subjective decision-making by the analyst. KDD also makes extensive use of data mining methods, which are automated processes and algorithms that enable pattern recognition and are characteristically machine learning methods developed in the domain of artificial intelligence. These methods have been applied to healthcare and biomedical data for a variety of purposes with good success and potential or realized clinical translation.

The Knowledge Discovery in Databases Process

Casual use of the term *data mining* to describe everything from routine statistical analysis of small data sets to large-scale enterprise data mining projects is pervasive. This broad application of the term causes semantic difficulties when attempting to communicate about KDD-relevant concepts and tools. Though multiple models and definitions have been proposed, the terms and definitions used in this chapter will be those given by Fayyad and colleagues in their seminal overview of data mining and knowledge discovery. The Fayyad model encompasses other leading models. Fayyad and colleagues define data mining as the use of machine learning, statistical, and visualization techniques algorithms to enumerate patterns, usually in an automated fashion, over a set of data. They clarify that data mining is one step in a larger knowledge discovery in databases (KDD) process that includes

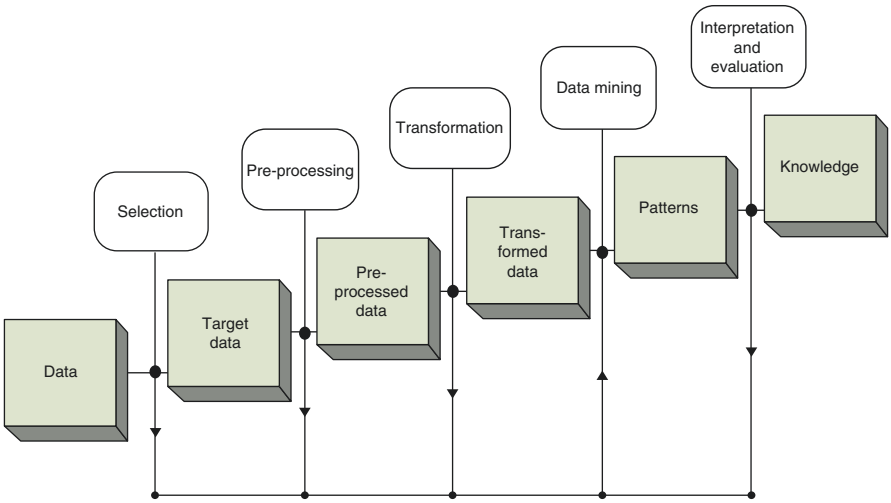

Fig. 16.1 Fayyad's knowledge discovery in databases process

data mining, along with any necessary data preparation, sampling, transformation, and evaluation/model refinement [4]. The encompassing process, the KDD process, is iterative and consists of multiple steps, depicted in Fig. 16.1. Data mining is not helpful or productive in inducing clinical knowledge models outside of this larger, essential process. Unless data mining methods are applied within a process that ensures validity, the results may prove invalid, misleading, and poorly integrated with current knowledge. As Fig. 16.1 depicts, the steps of KDD are iterative, not deterministic. While engaging in KDD, findings at any specific step may warrant a return to previous steps. The process is not sequential, as in a classic hypothetical-deductive scientific approach.

Data Selection

KDD projects are typically incepted when there is a clinical or operational decision requiring a clear and accurate knowledge model or in order to generate promising hypotheses for scientific study. These projects develop around a need to build knowledge or provide some guidance for clinical decision-making. Or lacking a particular clinical dilemma, a set of data particularly rich in content and size relevant to a particular clinical question may present itself. However, the relevant data is usually not readily available in a single flat file, ready for analysis. Typically, a data warehouse must be queried to return the subset of instances and attributes containing potentially relevant information. In some cases, clinical data will be partially warehoused, and some data will also need to be obtained from the source information system(s).

Just 20 years ago, data storage was sufficiently expensive, and methods for analysis of large data sets sufficiently immature, that clinical data was not routinely stored

apart from clinical information systems. However, there has been constant innovation and improvement in data storage and processing technology, approximating or exceeding that predicted by Moore's law. The current availability of inexpensive, high-capacity hard drives and inexpensive processing power is unprecedented. Data warehousing, the long-term storage of data from information systems, is now common. Transactional data, clinical data, radiological data, and laboratory data are now routinely stored in warehouses, structured to better facilitate secondary analysis and layered with analytic tools that enable queries and online analytic processing (OLAP).

Since clinical data is collected and structured to facilitate healthcare delivery and not necessarily analysis, key concepts may be unrepresented in the data or may be coarsely measured. For example, a coded field may indicate the presence or absence of pain, rather than a pain score. Proxies, other data attributes that correlate with unrepresented concepts, may be identified and included. For example, if a diagnosis of insulin-dependent diabetes is not coded, one might use insulin prescription (in combination with other attributes found in a set of data) as a proxy for Type I diabetes diagnosis. The use of proxy data and the triangulation of multiple data sources are often necessary to optimally represent concepts and identify specific populations within clinical data repositories [5]. A relevant subset of all available data is then extracted for further analysis.

Preprocessing

It is often said that preprocessing constitutes 90% of the effort in a knowledge discovery project. While the source and basis for that adage is unclear, it does seem accurate. Preprocessing is the KDD step that encompasses data cleaning and preparation. The values and distribution of values for each attribute must be closely examined, and with a large number of attributes, the process is time-consuming. It is sometimes appropriate or advantageous to recode values, adjust granularity, ignore infrequently encountered values, replace missing values, or to reduce data by representing data in different ways. For example, ordinality may be inherent in categorical values of an attribute and enable data reduction. An example exists in National Health Interview Survey data, wherein type of milk consumed is a categorical attribute. However, the different types of milk are characterized by different levels of fat content, and so the categorical values can be ordered by % fat content [6]. Each categorical attribute with n possible values constitutes n binary inputs for the knowledge discovery process. By restructuring a categorical attribute like type of milk consumed as an ordinal attribute, the values can be represented by a single attribute, and the number of inputs is reduced by $n - 1$. If attributes are duplicative or highly correlated, they are removed.

The distribution of values is also important because highly skewed distributions do not behave well mathematically with certain data mining methods. Attributes with highly skewed distributions can be adjusted to improve results, typically through normalization. The distribution of values is also important so that the investigator(s) is familiar with the representation of different concepts in the data set and can determine whether there are adequate instances for each attribute-value pair.

Transformation

Transformation is the process of altering the coded representation of data as input in order to reduce dimensionality or the number of rows and columns. Dimensionality reduction is often necessary in order to avoid combinatorial explosion or simply to improve computational efficiency during knowledge discovery. Combinatorial explosion is the vast increase in the number of possible patterns/solutions to a classification problem that occur with increases in the number of attributes. If a data set contains n input attributes, the number of possible combinations of attribute-value pairs that could be used to predict an outcome is 2^n. For a mere 16 inputs ($n = 16$), the number of possible combinations is 65,536. Every additional input results in increased computational demand. For knowledge discovery involving very large data sets, it is often necessary to create an alternate representation of the original input data, a representation that is computationally more manageable. Methods of transformation include wavelet transformation, principal components analysis, and automated binning (discretization) of interval attributes.

Data Mining

Data mining is the actual application of statistical and machine learning methods to enumerate patterns in a set of data [4]. It can be approached in several different ways, best characterized by the type of learning task specified. Artificial intelligence pioneer Marvin Minsky [7] defined learning as "making useful changes in our minds." Data mining methods "learn" to predict values or class membership by making useful, incremental model adjustments to best accomplish a task for a set of training instances. In unsupervised learning, data mining methods are used to find patterns of any kind, without relationship to a particular target output. In supervised learning, data mining methods are used to predict the value of an interval or ordinal attribute or the class membership of a class attribute (categorical variable).

Examples of unsupervised learning tasks:

- Perform cluster analysis to identify subgroups of patients with similar demographic characteristics.
- Induce association rules that detect novel relationships among attribute-value pairs in a pediatric injury database.

Examples of supervised learning tasks:

- Predict the blood concentration of an anesthetic given the patient's body weight, gender, and amount of anesthetic infused.
- Predict smoking cessation status based on health interview survey data.
- Predict the severity of medical outcome for a poison exposure, based on patient and exposure characteristics documented at the time of initial call to a poison control center.

Data mining methods are numerous, and it is important to understand enough about each method to use it appropriately. Some methods are highly flexible, capable of modeling very complex decision boundaries (artificial neural networks, support vector machines), while other methods are advantageous because they can be readily understood (classification and regression trees, association rules). Bayesian methods are distinctive in modeling dependencies among data. A comprehensive description of data mining methods is beyond the scope of this chapter but can be found in any data mining textbook. This chapter includes only a brief description of several important methods.

Artificial Neural Networks

Artificial neural networks constitute one of the oldest and perpetually useful data mining methods. The most fundamental form of an artificial neural network, the threshold logic unit, was incepted by McCulloch and Pitts at the University of Chicago during the 1930s and 1940s as a mathematical representation of frog neuron [8]. Contemporary artificial neural networks are multilayer networks composed of processing elements, variations of McCulloch and Pitt's original TLUs (Fig. 16.2). Weighted inputs to each processing element are summed, and if they meet or exceed a certain threshold value, they produce an output. The sum of the weighted inputs is a probability of class membership, and when deployed, the threshold of artificial neural networks can be adjusted for sensitivity or specificity.

Artificial neural networks make incremental adjustments to the weights according to feedback of training instances during a procedure for weight adjustment. Weight settings are initialized with random values, and the weighted inputs feed a network of processing elements, resulting in a probability of class membership and a

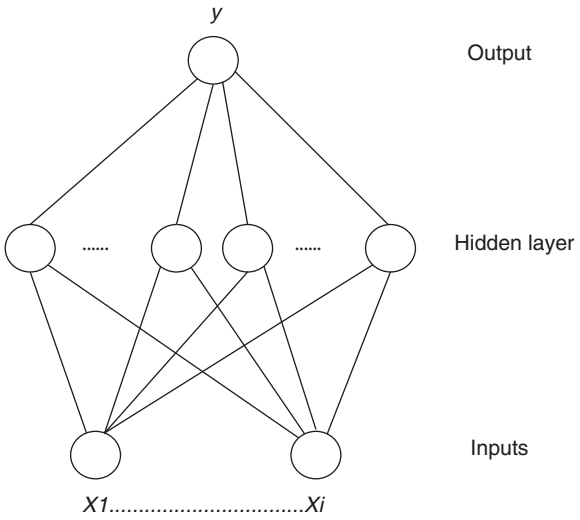

Fig. 16.2 Multilayer artificial neural network

prediction of class membership for each instance. The predicted class membership is then compared to the actual class membership for each instance. The model is incrementally adjusted, in a method specific to one of many possible training algorithms, until all instances are correctly classified or until the training algorithm is stopped. Because artificial neural networks incrementally adjust until error is minimized, they are prone to overtraining, modeling nuances, and noise in the training data set, in addition to valid patterns. In order to avoid overtraining, predictions are also incrementally made for a portion of data that has been set aside, not used for training. Each successive iteration of weights is used to predict class membership for the holdout data. Initially, successive iterations of weight configurations will result in decreased error for both the training data and the holdout data. As the artificial neural network becomes overtrained, error will increase for the holdout data and continue to decrease for the training data. This transition point is also the stopping point and is used to determine the optimal weight configuration (Fig. 16.3). Over multiple experiments, artificial neural networks can assume very different weight configurations but with varied configurations demonstrating equivalent performance.

Deep learning, [9] a powerful method for knowledge discovery used when very large amounts of data and training examples are available, is based upon artificial neural networks. In deep learning, the networks may have numerous layers and inputs, including multiple representation layers; the representation layers are refined in a "pre-training" step. This approach allows for effective, automatic identification of features, and so it effectively eliminates the need for more laborious forms of feature selection. Deep learning has led to extraordinary breakthroughs in image and language processing [10]. Its utility in modeling human behavior and health outcomes is not yet well characterized.

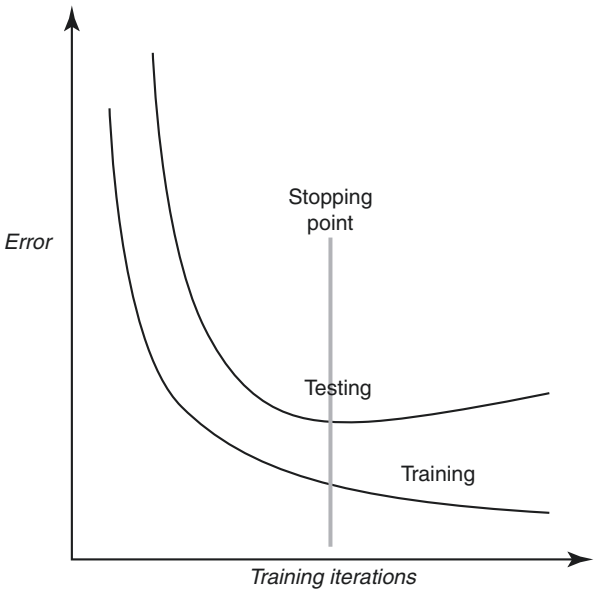

Fig. 16.3 Training/testing curves

Decision Trees

Decision trees, methods including classification and regression trees (CART) and an almost identical method known as C4.5, developed in parallel by Quinlan and others in the early 1980s [11]. These methods are used for supervised learning tasks and induce tree-like models that can be used to predict the output values for new cases. In this family of decision tree methods, the data is recursively partitioned based on attribute values, either nominal values or groupings of numeric values. A criterion, usually the information gain ratio of the attributes, is used to determine the order of the attributes in the resulting tree. Unless otherwise specified, these methods will induce a tree that classifies every instance in the training data set, resulting in an overtrained model. However, models can be post-pruned, eliminating leaves and nodes that handle very few instances and improving the generalizability of the model.

Decision trees are readily comprehensible and can be used to understand the basic structure of a pattern in data. They are sometimes used in the preprocessing stage of data mining to enhance data cleaning and feature subset selection. The use of decision tree induction methods early in the KDD process can help identify the persistence of rogue variables highly correlated with the output that are inappropriate for inclusion. However, ensembles of multiple decision trees, such as those utilized in random forest methods, tend to outperform single decision trees.

Support Vector Machines

Support vector machine methods were developed by Vapnik and others in the 1970s through the 1990s [12–14]. Support vector machines, like artificial neural networks, can be used to model highly complex, nonlinear solutions; however, they require the adjustment of fewer parameters and are less prone to overtraining. The method implements a kernel transformation of the feature space (attributes and their values) and then learns a linear solution to the classification problem (or by extension, regression) in the transformed feature space. The linear solution is made possible because the original feature space has been transformed to a higher-dimensional space. Overtraining is avoided through the use of maximal margins, margins that parallel the optimal linear solution and that simultaneously minimize error and maximize the margin of separation.

k-Nearest Neighbor

The *k*-nearest neighbor classification method (a common classification method and so-called “hot deck” method in missing value imputation) infers binary class membership on the basis of known class membership for similar instances. The output is inferred based on the majority class value for similar instances. This is a relatively simple algorithmic approach to classification. It has been shown robust in the

presence of missing values and with large numbers of attributes [15]. It is a case-based reasoning method that learns pattern in the training data only when it is required to classify each new testing instance.

Association Rules

Association rule induction is a method used for unsupervised learning. This method is used to identify if-then relationships among attribute-value pairs of any kind. For example, a pattern this algorithm could learn from a data set would be If COLOR = red, then FRUIT = apple. Higher-order relationships can also be found using this algorithm. For example, If COLOR = red and SKIN = smooth, then FRUIT = apple. Relationships among any and all attribute-value combinations will be described, regardless of importance. Many spurious relationships will typically be described, in addition to meaningful and informative relationships. The analyst must set criteria and limits for the order of relationships described, the minimum number of instances (evidence), and percentage of instances for which the relationship is true (coverage).

Bayesian Methods

Bayesian networks (in general) are networks of variables that describe the conditional probability of class membership based on the values of other attributes in the data. For example, a Bayesian network to predict the presence or absence of a disease would model P (disease symptoms). That conditional probability is then used to infer class membership for new instances. The structure and probabilities of the network can be directly induced from data, and the structure can be specified by domain experts with probabilities derived from actual data. These models become complex as join probability distributions become necessary to model dependencies among input data. Naïve Bayes is the most fundamental form of these methods, in which conditional independence between the input variables is assumed (thus the descriptor "naïve").

Interpretation and Evaluation

For supervised learning tasks, an output is specified, and a predictive model is induced. The error of induced models in predicting the output, whether the output is a real number or class membership, is used to evaluate the models. These metrics can be calculated by applying the model to predict outputs for data where actual output is known and comparing the predicted outputs to the actual outputs. For real number outputs, the error is the difference between the actual and predicted outputs. Error terms, including LMS error and RMSE, are used to quantify error.

For class variable outputs, error is misclassification. Each prediction constitutes a true positive, true negative, false positive, or false negative, and a confusion matrix is constructed from which various accuracy metrics are derived. Many data mining methods produce models that calculate a probability of class membership, to which a threshold is applied. At any given threshold, the confusion matrix may change. A higher threshold will result in fewer false positives, while a lower threshold will maximize sensitivity. This is advantageous in that the threshold can be adjusted in order to optimize these parameters for clinical applications. However, the predictive performance of the model cannot be adequately represented by metrics calculated with a single threshold confusion matrix. Instead, receiver operating curve (ROC) analysis is used.

An ROC curve is derived from the confusion matrix, by plotting the true-positive fraction vs. the false-positive fraction. Hanley and McNeil [16] define the index known as the area under the ROC curve as the probability that a randomly chosen subject of a given class will be predicted to belong to that class versus a randomly chosen subject that does not belong to that class [16]. ROC analysis originated in Great Britain during World War II, as a method of quantifying the ability of submarine sonar operators to distinguish signal indicating the presence of enemy ships. It was later adopted in radiology to quantify diagnostic accuracy. A detailed discussion of ROC analysis, specific to knowledge discovery and data mining in biomedical informatics, is found in Lasko et al. [17].

In order to obtain unbiased estimates of accuracy, it is necessary to calculate accuracy of model performance on a set of data that has not been used in training, testing, or model selection. This validation data set must be set aside before data mining methods are applied. Validation data sets differ from testing data sets. While validation data sets are not used during the data mining step, testing data sets are used in an interactive fashion to select model parameters and architecture. When cross validation is used, each testing instance also serves as a training instance. Even if cross validation is not used, and testing data sets do not contribute training instances, testing data sets are certainly used to compare and make choices about model parameters during the data mining step of the KDD process, so any estimates of accuracy calculated using testing data are biased. It is necessary to calculate accuracy using an entirely separate body of data, the validation set. Data partitioning, the assignment of available instances to training, testing, and validation data sets, is critical to interpretation and evaluation in KDD.

Applications of Knowledge Discovery and Data Mining in Clinical Research

Knowledge discovery and data mining methods have been used in numerous ways to generate hypotheses for clinical research.

Knowledge discovery and data mining methods are especially important in genomics, a field rich in data but immature in knowledge. In this area of biomedical research, exploratory approaches to hypothesis generation are accepted, even

necessary, in order to accelerate knowledge development. Data mining methods are often used to identify genetic markers of disease and genotype-phenotype associations for closer examination. For example, microarray analysis employs automated machine learning and statistical methods to identify patterns and associations in gene expression relevant for genetic epidemiology, pharmacogenomics, and drug development [18].

While KDD and data mining methods have demonstrated their ability to discern patterns in large, complex data, their usefulness in identifying patterns across biomedical, behavioral, social, and clinical domains is tempered by the disparate ways in which data is represented across research databases. It is difficult to aggregate clinical and genomic data, for instance, from diverse sources because of differences in coding and a lack of syntactic and semantic interoperability. Currently, a great deal of effort is being devoted to development of systems and infrastructure to facilitate sharing and aggregation of data.

Commonly Encountered Challenges in Data Mining

Rare Instances

Rare instances pose difficulty for knowledge discovery with data mining methods. In order for automated pattern search algorithms to learn differences that distinguish rare instances, there must be adequate instances. Also, during the data mining step of the KDD process, rare instances must be balanced with no instances for pattern recognition. If only 1 out of every 100 patients in a healthcare system has a fall incident, a sample of instances would be composed of 1% fall and 99% no-fall patients. Any classification algorithm applied to this data could achieve 99% accuracy by universally predicting that patients do not fall. If the sample is altered so that it is composed of 50% fall and 50% no-fall patients or if weights are applied, true patterns that distinguish fall patients from no-fall patients will be recognized. Afterwards, the models can be adjusted to account for the actual prior probability of a fall. In cases where inadequate instances exist, rare instances can be replicated, weighted, or simulated.

Sources of Bias

Mitigation of bias is a continual challenge when using clinical data. Many diverse sources of bias are possible in secondary analysis of clinical data. Verification bias is a type of bias commonly encountered when inducing predictive models using diagnostic test results. Because patients are selected for diagnostic testing on the basis of their presentation, the available data does not reflect a random sample of patients. Instead, it reflects a sample of patients heavily biased toward presence of a disease state. Another troublesome source of bias relates to inadequate reference standards (gold standards). Machine learning algorithms are trained on sets of

instances for which the output is known, the reference standard. However, clinical data may not include a coded, sufficiently granular representation of a given disease or condition. Even then, the quality of routinely collected clinical data can vary dramatically [6]. Diagnoses may also be incorrect, and source data, such as lab and radiology results, may require review by experts in order to establish the reference standard. If this additional step is necessary to adequately establish the reference standard, the time and effort necessary to prepare an adequate sample of data may be substantial. For an extended discussion of these and other sources of bias, the reader is referred to Pepe [19].

Many concepts in medicine and healthcare are not precisely defined or consistently measured across studies or clinical sites. Changes in information systems certainly influence the measurement of concepts and the coding of the data that represents those concepts. When selecting a subset of retrospective clinical data for analysis, it is wise to consult with institutional information technology personnel who are knowledgeable about changes in systems and databases over time. They may also be aware of documents and files describing clinical data collected using legacy systems, information that could be crucially important.

Limitations

The limitations in using repositories of clinical data for research are related to data availability, data quality, representation and coding of clinical concepts, and available methods of analysis. Since clinical information systems only contain data describing patients served by a particular healthcare organization, clinic, or hospital, the data represent only the population served by that organization. Any analysis of data from a single healthcare organization is, in effect, a convenience sample and may not have been drawn from the population of interest.

Data quality can vary widely and is strongly related to the role of data entry in workflow. For example, one preliminary study of data describing smoking status revealed that the coded fields describing intensity and duration of smoking habit were completed by minimally educated medical assistants, instead of nurse practitioners or physicians. Data describing intensity and duration of smoking habit were also plagued by absurdly large values. These values may have been entered by medical assistants when the units of measurement enforced by the clinical information system did not fit descriptions provided by patients. For example, there are 20 cigarettes in a pack. When documenting the intensity of the smoking habit, a medical assistant may have incorrectly entered “10” instead of “0.5” into a field with the unit of measurement “packs per day,” not “number of cigarettes per day” [6].

Infrastructure for Knowledge Discovery

The power of the KDD process, and of data mining methods, to enable large-scale knowledge discovery lies in their singular capacity to identify previously unknown

patterns, in data sets too large and complex for human pattern recognition. However, in order to identify true and complete patterns, all the relevant concepts must be represented in the data. Representations of key concepts, whether gene expression, environmental exposure, or treatment, often exist. However, they exist in siloed data repositories, owned by different scientific groups. Development of systems and infrastructure to support sharing and aggregation of scientific data is essential for understanding complex multifactorial relationships in biomedicine. The potential of KDD for advancing biomedical knowledge will not be fully realized until these systems and infrastructure are in place.

One earlier and influential infrastructure project in the United States was caBIG®, the cancer biomedical informatics grid. This project addressed barriers posed by lack of interoperability and siloed data by promoting fundamental change in the way clinical research is conducted. caBIG® collaborators developed open-source tools and architecture that enable federated sharing of interoperable data, using an object-oriented data model and standard data definitions. In early 2009, the University of Edinburgh became the first European university to deploy a caBIG application, caTISSUE repository [20]. However, in 2012, caBIG in the United States was reassessed.[1] The activities of the cancer Biomedical Informatics Grid (caBIG) program of the National Cancer Institute (NCI) were integrated into the Institute's new National Cancer Informatics Program (NCIP). NCIP provides many biomedical informatics resources for the cancer research community.

Another major approach to facilitating biomedical knowledge discovery has been that of the semantic web [21]. The semantic web is an extension of current web-based information retrieval that enables navigation and retrieval of resources using semantics (meaning) in addition to syntax (specific words or representations). Development of the semantic web is broadly important for information retrieval and use but specifically valuable for biomedical research because of its ability to make scientific data retrievable and usable across disciplines and scientific groups. In a recent methodological review, Ruttenberg and colleagues emphasized the importance of scientific ontology, standards, and tools development for the semantic web in order for biomedical research to realize the benefits. All-purpose semantic web schema languages RDFS and OWL can be used to manage relationships among data elements in information systems used to manage clinical studies. "Middle" ontologies are being developed to specifically address data relationships in scientific work [21].

Enterprise data warehouses (EDW) are repositories of clinical and operational data, populated by source systems but completely separate from those systems. EDWs facilitate secondary analysis by integrating data from diverse systems in a single location. The data is not used to support patient care or operations. It exists in a stand-alone repository optimized for secondary analysis. Typically, a layer of analytic tools is used to support queries and OLAP (online analytic processing). In some healthcare organizations, all clinical data may be warehoused. In other organizations, data collected by certain systems may be excluded, or certain types of

[1] Kush R. Where is caBIG Going? [Internet]. CDISC Website. 2012. Available from: http://www.cdisc.org/where-cabig-going?

data may be excluded. In these cases, data extracted from the EDW may need to be aggregated with data stored only in source systems. It is crucially important that data warehouses be optimized to facilitate scientific analytics. The way in which the data is stored and the development of powerful tools for examining and extracting the data directly influence the feasibility and quality of knowledge discovery using the data.

Success in aggregating data from diverse sources representing the spectrum of factors that affect human health, such as genomics, geography and community characteristics, social and behavioral determinants of health, environmental exposures, and healthcare, could enable unprecedented system-level insight into human health, using methods of knowledge discovery and data mining. In fact, the National Institutes of Health has launched a large initiative, the Child Health Outcomes (ECHO) Program, to create the infrastructure to support large cohort studies that can accomplish these types of analyses [22]. Pediatric asthma is an example of a disease thought to be influenced by multiple factors, including genomics, social and behavioral determinants of health, healthcare, and environmental air quality. In recent years, the NIH National Institute for Biomedical Imaging and Bioengineering funded PRISMS (Pediatric Research Using Integrated Sensor Monitoring Systems), a large scientific project aimed at achieving system-level insight in pediatric asthma. The PRISMS project is advancing the development of air quality sensors, both personal and environmental, optimized for use in research. However, it is also devoting resources to the development of informatics centers such as University of Utah's Utah PRISMS Center. The Utah PRISMS Center along with a partner informatics center located at the University of California, Los Angeles, is developing an informatics platform capable of receiving, processing, and storing the large quantities of data generated by sensors and producing data sets for analysis. A data coordinating center, currently based at the University of Southern California, then facilitates data integration and analysis. This project will enable exposomic research related to pediatric asthma, at varied spatiotemporal scale [23, 24].

Conclusion

Knowledge discovery and data mining methods are important for informatics because they link innovations in data management and storage to knowledge development. The sheer volume and complexity of modern data stores overwhelms statistical methods applied in a more traditional fashion. In the past, the inductive approach of data mining and knowledge discovery has been criticized by the statistical community as unsound. However, these methods are increasingly recognized as necessary and powerful for hypothesis generation, given the current data deluge. Hypotheses generated through the use of these methods, and unknown without these methods, can then be tested using more traditional statistical approaches. As the statistical community increasingly recognizes the advantages of machine learning methods and engages in knowledge discovery, the line between the statistical and machine learning worlds becomes increasingly blurred [25].

Much criticism is tied to the iterative and interactive nature of the knowledge discovery process, which is not consistent with the very sequential scientific method. Indeed, it is very important that data mining studies be replicable. In order for studies to be replicable, it is important that the analyst keep detailed records, particularly as data is transformed and sampled. It is also crucial that domain experts be involved in decision-making about data selection and feature selection and transformation, as well as the iterative evaluation of models. The quality of resultant models is evidenced by performance on new data, and models should be validated on unseen data whenever possible. Models also must be calibrated for the target population with which they are being used. Uncalibrated models will certainly lead to increased error [26].

While the data deluge is very real, our technology for optimally managing and structuring that data lags behind. In clinical research, data mining and knowledge discovery awaits the further development of high-quality clinical data repositories. Many data mining application studies in the biomedical literature find that model performance is limited by the concepts represented in the available data. For optimal use of these methods, all relevant concepts in a particular area of interest must be represented. The old adage "garbage in, garbage out" applies. If a health behavior (i.e., smoking) is believed to be related to biological, social, behavioral, and environmental factors, a data set composed of only biological data will not suffice. Additionally, much of the data being accumulated in data warehouses is of varied quality and is not collected according to the more rigorous standards employed in clinical research. As more sophisticated systems for coding and sharing data are devised, we find ourselves increasingly positioned to apply data mining and knowledge discovery methods to high-quality data repositories that include most known and possibly relevant concepts in a given domain.

In the ever-intensifying data deluge, knowledge discovery methods represent one of several pivotal tools that may determine whether human welfare is advanced or diminished. It is important for scientists engaged in clinical research to develop familiarity with these methods and to understand how they can be leveraged to advance scientific knowledge. It is also critical that clinical scientists recognize the dependence of these methods upon high-quality data, well-structured clinical data repositories, and data sharing initiatives.

References

1. Benson K, Hartz AJ. A comparison of observational studies and randomized, controlled trials. Am J Ophthalmol. 2000;130(5):688.
2. Aronsky D, Fiszman M, Chapman WW, Haug PJ. Combining decision support methodologies to diagnose pneumonia. In: Proceedings of the AMIA symposium; 2001. p. 12–6.
3. Lagor C, Aronsky D, Fiszman M, Haug PJ. Automatic identification of patients eligible for a pneumonia guideline: comparing the diagnostic accuracy of two decision support models. Stud Health Technol Inform. 2001;84(Pt 1):493–7.
4. Fayyad U, Piatetsky-Shapiro G, Smyth P. From data mining to knowledge discovery in databases. AI Mag. 1996;17(3):37–54.

5. Aronsky D, Haug PJ, Lagor C, Dean NC. Accuracy of administrative data for identifying patients with pneumonia. Am J Med Qual. 2005;20(6):319–28. https://doi.org/10.1177/1062860605280358.
6. Poynton MR, Frey L, Freg H. Representation of smoking-related concepts in an electronic health record. In: Medinfo 2007: Proceedings of the 12th world congress on health (medical) informatics; building sustainable health systems; 2007. p. 2255.
7. Minsky M. The society of mind. New York: Simon & Schuster; 1986.
8. McCulloch WS, Pitts W. A logical calculus of the ideas immanent in nervous activity. Bull Math Biophys. 1943;5(4):115–33. https://doi.org/10.1007/BF02478259.
9. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521(7553):436.
10. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. Commun ACM. 2017;60(6):84–90. https://doi.org/10.1145/3065386.
11. Quinlan JR. C4. 5: programs for machine learning. Oxford: Elsevier; 2014.
12. Cristianini N, Shawe-Taylor J. An introduction to support vector machines and other kernel-based learning methods. Cambridge, UK: Cambridge University Press; 2000.
13. Vapnik VN. The nature of statistical learning theory. New York: Springer; 1995. p. 188.
14. Vapnik VN. Statistical learning theory. New York: Wiley; 1998. p. 736.
15. Jonsson P, Wohlin C. Benchmarking k-nearest neighbour imputation with homogeneous likert data. Empir Softw Eng. 2006;11(3):463–89. https://doi.org/10.1007/s10664-006-9001-9.
16. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (roc) curve. Radiology. 1982;143(1):29–36. https://doi.org/10.1148/radiology.143.1.7063747.
17. Lasko TA, Bhagwat JG, Zou KH, Ohno-Machado L. The use of receiver operating characteristic curves in biomedical informatics. J Biomed Inform. 2005;38(5):404–15. https://doi.org/10.1016/j.jbi.2005.02.008.
18. Cordero F, Botta M, Calogero RA. Microarray data analysis and mining approaches. Brief Funct Genomics. 2007;6(4):265–81. https://doi.org/10.1093/bfgp/elm034.
19. Pepe MS. The statistical evaluation of medical tests for classification and prediction. Oxford: Oxford University Press; 2003. ISBN 9780198509844.
20. Breiman L. Statistical modeling: the two cultures (with comments and a rejoinder by the author). Stat Sci. 2001;16(3):199–231. https://doi.org/10.1214/ss/1009213726.
21. Genomeweb. Persistent systems helps first european deploy cabig's catissue repository. 2009.
22. Ruttenberg A, Clark T, Bug W, Samwald M, Bodenreider O, Chen H, Doherty D, Forsberg K, Gao Y, Kashyap V, Kinoshita J, Luciano J, Marshall MS, Ogbuji C, Rees J, Stephens S, Wong GT, Wu E, Zaccagnini D, Hongsermeier T, Neumann E, Herman I, Cheung K-H. Advancing translational research with the semantic web. BMC Bioinforma. 2007;8(3):S2. https://doi.org/10.1186/1471-2105-8-s3-s2.
23. Program E. Environmental influences on child health outcomes (echo) program. 1/30/2018), ECHO supports multiple longitudinal studies using existing study populations to investigate environmental exposures on child health and development.
24. Burnett N. Harmonization of sensor measurement to support health research. In: Proceedings of the national conference of undergraduate research 2017. 2017.
25. Kelly KE, Whitaker J, Petty A, Widmer C, Dybwad A, Sleeth D, Martin R, Butterfield A. Ambient and laboratory evaluation of a low-cost particulate matter sensor. Environ Pollut. 2017;221:491–500. https://doi.org/10.1016/j.envpol.2016.12.039.
26. Matheny ME, Ohno-Machado L, Resnic FS. Discrimination and calibration of mortality risk prediction models in interventional cardiology. J Biomed Inform. 2005;38(5):367–75.