attributes of interest, or containing only aggregate data), noisy (containing errors or outliers), and inconsistent (containing discrepancies in codes or names). The nature of the data and the issues related to preprocessing of data for analytics are explained in detail in Chapter 2.

Step 4: Model Building

In this step, various modeling techniques are selected and applied to an already prepared data set to address the specific business need. The model-building step also encompasses the assessment and comparative analysis of the various models built. Because there is not a universally known *best* method or algorithm for a data mining task, one should use a variety of viable model types along with a well-defined experimentation and assessment strategy to identify the "best" method for a given purpose. Even for a single method or algorithm, a number of parameters need to be calibrated to obtain optimal results. Some methods may have specific requirements in the way that the data is to be formatted; thus, stepping back to the data preparation step is often necessary. Application Case 4.4 presents a research study where a number of model types are developed and compared to each other.

Application Case 4.4

Data Mining Helps in Cancer Research

According to the American Cancer Society, half of all men and one-third of all women in the United States will develop cancer during their lifetimes; approximately 1.5 million new cancer cases were expected to be diagnosed in 2013. Cancer is the second-most-common cause of death in the United States and in the world, exceeded only by cardiovascular disease. This year, over 500,000 Americans are expected to die of cancer—more than 1,300 people a day—accounting for nearly one of every four deaths.

Cancer is a group of diseases generally characterized by uncontrolled growth and spread of abnormal cells. If the growth and/or spread are not controlled, it can result in death. Even though the exact reasons are not known, cancer is believed to be caused by both external factors (e.g., tobacco, infectious organisms, chemicals, and radiation) and internal factors (e.g., inherited mutations, hormones, immune conditions, and mutations that occur from metabolism). These causal factors may act together or in sequence to initiate or promote carcinogenesis. Cancer is treated with surgery, radiation, chemotherapy, hormone therapy, biological therapy, and targeted therapy. Survival statistics vary greatly by cancer type and stage at diagnosis.

The 5-year relative survival rate for all cancers is improving, and decline in cancer mortality had reached 20% in 2013, translating into the avoidance of about 1.2 million deaths from cancer since 1991. That's more than 400 lives saved per day! The improvement in survival reflects progress in diagnosing certain cancers at an earlier stage and improvements in treatment. Further improvements are needed to prevent and treat cancer.

Even though cancer research has traditionally been clinical and biological in nature, in recent years data-driven analytic studies have become a common complement. In medical domains where data- and analytics-driven research have been applied successfully, novel research directions have been identified to further advance the clinical and biological studies. Using various types of data, including molecular, clinical, literature-based, and clinical trial data, along with suitable data mining tools and techniques, researchers have been able to identify novel patterns, paving the road toward a cancer-free society.

In one study, Delen (2009) used three popular data mining techniques (decision trees, artificial neural networks, and SVMs) in conjunction with logistic regression to develop prediction models for prostate cancer survivability. The data set contained around 120,000 records and 77 variables. A *k*-fold cross-validation methodology was used in model building, evaluation, and comparison. The results showed

(*Continued*)

Application Case 4.4 (Continued)

that support vector models are the most accurate predictor (with a test set accuracy of 92.85%) for this domain, followed by artificial neural networks and decision trees. Furthermore, using a sensitivity–analysis-based evaluation method, the study also revealed novel patterns related to prognostic factors of prostate cancer.

In a related study, Delen, Walker, and Kadam (2005) used two data mining algorithms (artificial neural networks and decision trees) and logistic regression to develop prediction models for breast cancer survival using a large data set (more than 200,000 cases). Using a 10-fold cross-validation method to measure the unbiased estimate of the prediction models for performance comparison purposes, the results indicated that the decision tree (C5 algorithm) was the best predictor, with 93.6% accuracy on the holdout sample (which was the best prediction accuracy reported in the literature), followed by artificial neural networks, with 91.2% accuracy, and logistic regression, with 89.2% accuracy. Further analysis of prediction models revealed prioritized importance of the prognostic factors, which can then be used as a basis for further clinical and biological research studies.

In the most recent study, Zolbanin, Delen, and Zadeh (2015) studied the impact of comorbidity in cancer survivability. Although prior research has shown that diagnostic and treatment recommendations might be altered based on the severity of comorbidities, chronic diseases are still being investigated in isolation from one another in most cases. To illustrate the significance of concurrent chronic diseases in the course of treatment, their study used the Surveillance, Epidemiology, and End Results (SEER) Program's cancer data to create two comorbid data sets: one for breast and female genital cancers and another for prostate and urinal cancers. Several popular machine-learning techniques are then applied to the resultant data sets to build predictive models (see Figure 4.4). Comparison of the results has shown that having more information about comorbid conditions of patients can improve models' predictive power, which in turn can help practitioners make better diagnostic and treatment decisions. Therefore, the study suggested that proper identification, recording, and use of patients' comorbidity status can potentially lower treatment costs and ease the healthcare-related economic challenges.

These examples (among many others in the medical literature) show that advanced data mining techniques can be used to develop models that possess a high degree of predictive as well as explanatory power. Although data mining methods are capable of extracting patterns and relationships hidden deep in large and complex medical databases, without the cooperation and feedback from the medical experts, their results are not of much use. The patterns found via data mining methods should be evaluated by medical professionals who have years of experience in the problem domain to decide whether they are logical, actionable, and novel enough to warrant new research directions. In short, data mining is not meant to replace medical professionals and researchers, but to complement their invaluable efforts to provide data-driven new research directions and to ultimately save more human lives.

Questions for Discussion

1. How can data mining be used for ultimately curing illnesses like cancer?
2. What do you think are the promises and major challenges for data miners in contributing to medical and biological research endeavors?

Sources: Zolbanin, H. M., Delen, D., & Zadeh, A. H. (2015). Predicting overall survivability in comorbidity of cancers: A data mining approach. *Decision Support Systems, 74,* 150–161; Delen, D. (2009). Analysis of cancer data: A data mining approach. *Expert Systems, 26*(1), 100–112; Thongkam, J., Xu, G., Zhang, Y., & Huang, F. (2009). Toward breast cancer survivability prediction models through improving training space. *Expert Systems with Applications, 36*(10), 12200–12209; Delen, D., Walker, G., & Kadam, A. (2005). Predicting breast cancer survivability: A comparison of three data mining methods. *Artificial Intelligence in Medicine, 34*(2), 113–127.