Chapter 5
Big Data in Healthcare: New Methods of Analysis

Sarah N. Musy and Michael Simon

Abstract  With the ubiquitous availability of health-related data such as insurance claims, discharge abstracts, electronic health records, personal fitness devices or mobile phone applications, the amount of health data is increasing in size, but also in speed and in complexity. “Big data” provides new opportunities for nurse clinicians and researchers to improve patient health, health services and patient safety. Following this unprecedented amount and complexity of information available from different types of data sources, the processing and the analysis of big data challenges traditional analytical methods. For these reasons, a range of analytical approaches such as text mining and machine learning often developed in bioinformatics or engineering fields become of highest relevance to nurses wanting to work with big data. This chapter provides a brief overview of the main definitions and the analytical approaches of big data. The chapter gives two nursing research examples in the context of patient experience in cancer care and older people with dementia in nursing homes. In both cases the analytical approach (text mining and machine learning) is highly integrated into traditional research designs (a cross-sectional survey and a retrospective observational study), which highlights how traditional research designs become increasingly influenced by analytical strategies from big data or data science.

Keywords  Big data • Machine learning • Text mining • Predictive modeling • Data mining and knowledge discovery • Data visualization • Natural language processing
5.1 Introduction

In the last decade the term “big data” became widely used in the literature (Gandomi and Haider 2015). The term big data refers to large and complex data sets, which are difficult to analyze with traditional data processing methods (Frost and Sullivan 2015). Because technology and also analytical methods constantly develop, the threshold of what ‘difficult’ constitutes is moving too, making it hard to determine, which type or size of data can be described as big data. In this context often three dimensions of big data are discussed: volume, variety and velocity (Laney 2001).

Volume refers to the size of the data. A survey by IBM of 1144 participants found that half of respondents would consider a size of more than one terabyte as big data (see Fig. 5.1) (Schroeck et al. 2012). This answer depends also on today’s storage capacities, which are constantly increasing (Gandomi and Haider 2015). Variety refers to the structural heterogeneity of the data, which can be structured or unstructured. Only a small part of existing data (5%) is in a structured form, referring to data stored in the traditional row-columns database or spreadsheet, such as medication data (Cukier 2010). The rest is unstructured data, including text and multimedia (i.e. pictures, audio or video files) content, such as clinical notes (Gandomi and Haider 2015). Structured data is easier to work with, but using unstructured data is particularly challenging. Finally velocity refers to the speed of data generation and processing. Data is produced increasingly quickly, for example the micro-blogging service twitter serves about 350,000 tweets every minute. This allows analysis of data in real time or near-real time, but also requires capacity to process analyses accordingly (Shah 2015). In addition, other characteristics have emerged, such as veracity, variability or value, which are describes elsewhere (Normandeau 2012; Katal et al. 2013; Gandomi and Haider 2015).

The rapid digitization of health and health care data is leading to a dramatic growth of information on all levels of the healthcare system (Raghupathi and Raghupathi 2014; Larson 2013). In 2012 it was estimated that worldwide size of digital healthcare data reached 500 petabytes, corresponding to more than 13 years of HD video and is expected to be multiplied by 50 in 2020 (Hersh et al. 2011).

Alongside the question of what constitutes ‘big data’ and where big data is generated from, a range of analytical techniques, often labeled as “data science”, have

Fig. 5.1 Representation of the size of terabytes and petabytes with examples
emerged. Those techniques deal with difficulties from big data files, but also generate new opportunities with big data and data from traditional research designs. In the following sections, we will give a brief overview about the common sources of big data and typical analytical techniques and provide two examples where these techniques have been applied in nursing research. Finally, we will briefly discuss the challenges of big data and give a perspective of how data analytics from big data might influence nursing and nursing research.

5.2 Sources of Big Data

The large variety of sources and the rapid growth of healthcare data explain the interest in big data in healthcare. Figure 5.2 provides an overview of different sources. For a more detailed description, please refer to references by Weber et al. (2014) and Shah (2015). We divide data sources into two broad groups: routine data (e.g., automatically collected and readily available data) and research data (i.e., primary collected data).

Routine data is often administrative data, which is produced for financial management, such as insurance enrollment and provider claims, but might also contain medical information about the patients’ health or emotional status. Research data often is not considered to be big data, but some sources like genetic data produce datasets with thousands of variables, requiring analytical approaches often used in data science.

A common source of medical data in routine data is the electronic health records (EHRs). EHRs contain patient charts (e.g., vital signs), clinical notes (from physicians and nurses), procedure reports (e.g., catheters insertion), clinical assessments

![Diagram of data sources for healthcare](image-url)

**Fig. 5.2** Different types of data sources possible for the healthcare environment
(e.g., pain), care plans (e.g., nursing diagnosis), and medication information (e.g., type, dose, and time). EHRs are therefore a valuable source of information about patient demographics, diagnoses, procedures, symptoms, and medication.

Beside traditional data sources like the EHRs, new technologies coming from mobile phone applications or fitness trackers constantly collect data of users’ health or activities and become an important source of medical information. Data about the heart rate, weight, height, calories intake/expenditure, distance travelled, location information through GPS, and the number of steps provide a wealth of information about healthy people and their lifestyle. Data collected in hospitals concern patients with diseases and now data from healthy people can be collected outside the hospital environment. Using this type of data is still in its infancy, but health insurance companies are becoming interested in tracker data to provide incentives for people moving more.

Another relevant source for health research is social media data. Encompassing a variety of online platforms, social media allow users to create and exchange personal or professional content and are increasingly becoming a relevant source of data for healthcare purposes (Barbier and Liu 2011; Gundecha and Liu 2012). For instance does a Facebook’ friend influence you to take a certain drug? Or does the information on a blog about a drug have more impact than your physician’s prescription? These are examples of questions that can be answered with data from social media. However two problems currently inhibit the use and exploitation of social media data: most social media data is not accessible to researchers because of the proprietary nature of the systems capturing this data, but also because of data protection regulations. Again this type of data is often not really used for healthcare purposes, which is not the case for marketing companies that use them to track people preference or taste in order to provide them the products of interest.

On the other side, ‘big’ research data is considered to be important for future developments. With lowering prices, genetic data is becoming more and more accessible. Genetic data will become a part of personalized medicine to develop and provide personalized treatments based on the patient’s genome. Genomic data is in itself big data due to the large amount of variables (~3 billion base pairs in one genome).

Relatively new data sources in research are audio and video files. For instance audio analytics can support diagnosis, treatment or information about adults or children (Hirschberg et al. 2010). Patients with certain communication patterns, e.g., depression, schizophrenia, but also cancer, can benefit from audio recording support diagnosis and treatment. Another example is the analysis of infant cries, which has shown to give information about emotional status and health status (Patil 2010).

Still in its infancy compared to other data sources (Abraham and Das 2010) is real-time and pre-recorded video. Various techniques have been developed for its processing, but here big data is still a processing power challenge with one second of high-definition video being equivalent to over 2000 pages of text (Manyika et al. 2011).
5.3 Big Data Analytics

It is difficult to provide a coherent overview of the analytical techniques that are applied in the context of big data and data science because of the number of approaches, but also the disciplinary diversity ranging from informatics, engineering, statistics and others. Often these approaches are combined or overlap making it difficult to differentiate them. Hence we will introduce four areas of particular interest in the context of healthcare and nursing research: data mining, text mining, predictive modelling and machine learning.

5.3.1 Data Mining

Data mining is concerned with the detection of patterns in voluminous or complex data, where traditional methods failed to process and analyze them (Popowich 2005; Biafore 1999). Pattern recognition aims to identify potentially useful and understandable correlations in the data often to forecast or predict the likelihood of future events (Chung and Gray 1999). Predictive modeling is probably the most common application of data mining, which will be discussed in a following section. Data mining as most of the analytical tools in big data originates from database management, statistics and computer science, explaining its large panel of analytical tools.

The first step in the data mining approach is the understanding of the business at hand (Koh and Tan 2011, see Fig. 5.3). Understanding is crucial for any data mining analysis, since it aims to identify the objectives of the analysis, but also to understand how variables might be associated with each other. Understanding and preparing the data including description and visualization of the data, respectively sampling and data transformation, are important elements for any data modeling, but in particular for data mining approach. The analytical process consists of either traditional statistical methods (e.g., cluster analysis) (Copeland et al. 2009), discriminant analysis (Peterson et al. 2008), and regression analysis (Peterson et al. 2008) or non-traditional statistical methods (e.g., neural networks) (Azimi et al. 2015), decision

![Fig. 5.3 Different steps of the data mining approach according to Koh and Tan (2011)]
trees (Kang et al. 2016), link analysis (Nie et al. 2006) and machine learning (Russell et al. 1995, see Sect. 5.2.4). The evaluation stage compares models and results, and finally the deployment stage implements the data mining model.

So far data mining in health care has slowly been incorporated. But some applications do exist, such as medical insurance fraud and abuse detection. Through data mining insurers can establish norms and identify unusual claim patterns. This has led many insurers to use this information, resulting in a decrease of their losses and the costs of health care (Milley 2000).

### 5.3.2 Text Mining

Although great efforts have been made to underpin clinical data with structured terminologies a wide range of clinical information is still captured in unstructured format. Free-text, one possible format of data, is convenient and often used for clinical notes, procedure reports, emails, or online forums, since words contain useful information, not captured elsewhere. Text mining is a well-known method to deal with free-text. It is defined by Hearst as the process of detecting patterns and extracting knowledge from unstructured data into structured data (Hearst 1999; Popowich 2005). This transformation allows us to utilize such data for further analysis.

Figure 5.4 shows the different steps of the process. Information is retrieved to collect relevant texts, and then information is extracted (DeJong 1982). The structure of a text can be pre-processed and extracted in several ways using methods like stemming (only using the stem of a word), term document matrices (creating binary or weighted counts of words per document) or stop-word removal to make the corpora (the documents to be analyzed) accessible for further analysis (Meyer et al. 2008). Extracting information from text in a structured format is for instance useful to identify patients with certain diagnoses (Liao et al. 2015) or adverse events (Li et al. 2014) in electronic health records. The next step is a combination of semantic search (e.g., extracting sense from text and presenting it in a coherent manner) and data mining technique (e.g., by finding associations between the extracted pieces of information) (Meystre et al. 2008). Finally, in order to transform textual data into meaningful structured information, text mining is often combined with machine learning to classify text into certain categories (Gandomi and Haider 2015).

Text mining can synthesize information from many different sources and keep up-to-date with the large amounts of information. Text mining requires high levels of technical expertise and is widely used in fields like sociology, communication or bioinformatics, but so far has not gained much interest in healthcare (Raja et al. 2008).

![Text mining process](image-url)
However, some promising results for applications in healthcare research exist. For example in a single site study over a period of 6 months using the text mining approach electronic medical records from an emergency department, patients with shortness-of-breath were extracted (Cerrito and Cerrito 2006). They found that different physicians treated those complaints differently affecting care quality and costs.

5.3.3 Predictive Modelling

Predictive modelling is probably the best term representing big data analytics. Predictive modelling refers to the development of models to make accurate predictions and is often labeled as machine learning, artificial intelligence or pattern recognition depending on the discipline (Kuhn and Johnson 2013) and represents the “non-traditional” analytics part of data mining (Sect. 5.2.1). As Kuhn and Johnson (2013) describe, predictive modelling is primarily about prediction accuracy and less about the interpretation of the model. An example of this view is the identification of spam emails, where people are primarily interested in the effective trashing of spam emails as opposed to learning what features might be relevant or how an algorithm actually works. This view is based on Karl Popper’s criterion for judging a theory focusing more on its predictive power and less on its ability to explain a phenomenon (Dhar 2013). This focus on prediction accuracy is probably the key ingredient driving big data analytics and dominating the analytical landscape in this area. It is difficult to provide a coherent overview of the analytical techniques that are applied in the context of big data and data science because of the number of approaches, but we will introduce an area of particular interest in the context of healthcare and nursing research: machine learning.

5.3.4 Machine Learning

Machine learning refers to analytical approaches, which allow computers to learn from data. Machine learning addresses a range of problems with supervised, unsupervised, and reinforcement machine learning being the three main types of machine learning approaches (Russell et al. 1995).

The different steps of analysis with machine learning are shown in Fig. 5.5 (Kapitanova and Son 2012). Data collection refers to the extraction of data from one database or the combining of data from multiple databases. Considerable time may be needed to preprocess the data for missing, redundant, irrelevant and outlier data. Training a model requires the split of the data into a training set and a validation set. The training set is used to train and find the right algorithms (depending on the purpose). Once an algorithm is constructed a different set is used to validate the algorithm and test its performance (evaluating the model). The last step improves the performance of the algorithm by allowing the computer to refine it, in a stepwise fashion, with new variables.
Supervised machine learning refers to analytical approaches where the computer is presented with several inputs (e.g., patient records with structured and unstructured data) and outputs (e.g., a certain patient state) in order to develop an algorithm to classify future cases according to the given characteristics. Supervised machine learning typically starts with the manual annotation or classification of a set of records (e.g., determine a certain diagnosis in a patient record). The annotated records are then split into two (or more) sets: a training set and a test set. The training set is used to train the ‘machine’ (the computer) in order to identify records with a certain characteristic (e.g., a diagnosis) through one or more algorithms. Many algorithms like support vector machines (SVM) or decision trees have been developed in order to help with the classification task. The decision which algorithm works best is determined by the complexity, sample size and noise in the training data. Often several algorithms are tested for the task at hand and the decision about which algorithm or a combination of algorithms is used can be based on the performance of the algorithm(s) in the test set.

Unsupervised machine learning refers to situations when no label, structure or classification is available and the algorithm develops its own structure to describe the data provided. The most common approaches are k-means clustering, which uses the k-mean algorithm in order to classify n observations into k clusters. For example, this approach has been used to verify the information of web-based diabetes patient education material (Thakurdesai et al. 2004). Using a list of 53 sites for the study, K-mean cluster analysis was performed to classify the web-sites into four groups based on sum of scores obtained from core educational concepts (best, medium, good, and average). The results classify 12 websites in the best category, nine in medium, 24 in good, and eight in average.

5.4 Big Data Applications in Nursing

Although big data has received much attention in the science community, the uptake in nursing seems fairly limited so far. A search with the term “big data” in the nursing core journals on Pubmed in February 2016 revealed 30 hits, with fewer than three articles showing empirical research. This is partially the case because some of the research already applying big data methods does not necessarily refer to big data.
or the used analytical technique is only one part of a range of methodological features. We will provide two examples of research the senior author of this chapter has been involved with, which use text mining and machine learning methodology in the context of health services research in nursing. First we will describe a study which used text mining in combination with machine learning in order to describe written patients’ comments on their experiences of colorectal cancer care (Wagland et al. 2016) and, second, we will describe a study which used machine learning using a genetic algorithms to optimize covariate balance in a matched sample to explore case conferences in nursing homes (Palm et al. 2016).

Example 1 Text mining, machine learning and patient comments

Background. Surveys often contain open-ended questions providing potentially valuable and new information from survey participants. Unfortunately, data accumulated from these open-ended questions is difficult and particularly time-consuming to analyze. Traditionally this data is content analyzed in three consecutive steps, which are neither quantitative nor qualitative by nature (O’Cathain and Thomas 2004). In a first step responses are read and a coding frame is devised in order to describe the content of the comments. In a second step all comments are coded by raters, and reliability explored by double coding of a subset of comments. In the last step codes are described and the overall distribution of codes is described (O’Cathain and Thomas 2004). While this approach is sufficiently robust and reliable scaling-up to several hundreds or thousands of comments make this approach cumbersome and time-consuming task. Supervised machine learning combining text mining with machine learning allows us to train algorithms in order to detect certain types of patient comments, which then can further be analyzed.

Context of the study. The study by (Wagland et al. 2016) was challenged by the open-ended question of the national colorectal experience survey of 21,802 cancer patients in the UK, which contained 5634 responses with written comments. In a pilot study a small sample of comments was explored indicating a range of informative and relevant themes describing care experiences of colorectal cancer patients.

Methodology. In a first step a first random sample (rs1) of 400 comments was coded by three experienced qualitative researchers. The codes were developed in previous pilot study (Corner et al. 2013) and applied and adopted for this study. The framework coded comments as positive or negative experiences and whether specific forms of information to prepare patients were lacking. Cohen’s Kappa between the different raters ranged from substantial (0.64) to excellent (0.87). Inconsistencies between the data and the existing framework were discussed between researchers, with disagreements jointly resolved.

In a second step another random sample (rs2) of comments was coded by the qualitative researchers. Of the overall 800 coded comments (rs1 + rs2) 50% were used to train seven different machine learning algorithms. For training the algorithms a term document matrix is created, which counts the
occurrence of any used term of all included comments. In order to assess the performance of the algorithms the sample was randomly split into ten subsamples, conducting training in nine datasets, testing in one, and repeating this process ten times (tenfold cross validation). Algorithm performance is measured as sensitivity (true positives/(true positives + false negatives)), precision (true positives/(true positives + false positives)) and by the f-score ((2 × sensitivity × precision)/(sensitivity + precision)). While sensitivity and precision are common metrics in health research the f-score describes overall performance of an algorithm, representing the harmonic mean of precision and sensitivity.

For the third step the best four algorithms to identify either positive or negative patient experiences were combined. With this approach 1688 comments were identified and finally coded by the qualitative researchers. About 81% of the comments identified by the algorithms contained positive or negative patient experiences. Figure 5.6 provides an overview about the different steps in the analytical approach of the study.

**Conclusion.** In summary, this study showed that combining text mining and machine learning techniques was useful and practical to identify specific free-text comments within a large dataset, facilitating resource-efficient qualitative analysis. However this is only one example requiring more experiences with other datasets in order to fully appreciate the potential and limitations of such an approach.

![Analytical approach of the study by Wagland et al. (2016)](image_url)

**Fig. 5.6** Analytical approach of the study by Wagland et al. (2016)
Example 2 Machine learning for covariate balance using genetic algorithms

**Background.** Propensity score matching is a method for causal inference from observational studies (Dehejia and Wahba 2002). Causal inference in medicine is typically drawn from randomized controlled trials by assigning two groups of individuals to different treatments. The assignment can be part of regular care processes without randomization (as natural experiment) or as part of an experiment with randomization like in a randomized controlled trial to test a novel intervention. The randomization primarily serves the purpose to avoid selection bias, which occurs when individuals receiving the novel intervention are systematically different than individuals receiving usual care leading to biased estimates of the treatment effect. In a non-randomized study propensity score matching provides a mean to identify individuals in a treatment and control group sharing the same characteristics. A logistic regression model is calculated with the aim to determine the probability or ‘propensity’ of the participants to be in the treatment or control group given a set of observed variables. The propensity score is then used to match participants from the treatment to the control group (or vice versa). The core criterion whether the propensity score model works sufficiently well is the covariate balance between treatment and control group. For instance, whether the average age or the gender distribution in the treatment or the control group are similar, and is therefore comparable to how we would judge the randomization in a randomized controlled trial. Although conceptually the propensity score matching approach is strong, the practical application is obstructed by specifying the propensity model in a way that covariate balance is actually achieved (Diamond and Sekhon 2013).

**Context.** Case conferences are considered to be a key intervention to reduce challenging behavior in people with dementia in nursing homes. Dementia special care units (DSCUs) are units receiving additional funding in order to provide highly specialized care with more staff and are also expected to provide more frequent case conferences. The aim of the study was to determine whether case conferences are more frequently conducted in DSCUs than in traditional care units.

**Methodology.** In epidemiological terms the study is a cross-sectional observational study of 264 residents from 16 DSCUs and 48 TCUs. Information regarding case conferences was collected by the nurses using the Dementia Care Questionnaire. Several instruments were used to collect resident characteristics and challenging behavior including the Neuropsychiatric Inventory Questionnaire, the Physical Self-Maintenance Scale, Dementia Screening Scale, and sociodemographic information.

In this study propensity score matching was combined with the genetic algorithm (GenMatch). This approach solves the issue of iterative specification of the propensity score model in order to achieve covariate balance. Without the genetic algorithm, the propensity score model has to be manually...
5.5 Challenges of Big Data

Big data is challenging, but some analytical approaches either developed in the field of big data, or first applied in this context now help to address some of the challenges beyond big data itself in health services and clinical research. The common denominator of the different challenges in big data analytics is the difficulties in processing or analyzing the data. The nature and the type of these difficulties can vary widely depending on the structure, the analytical aim or the contextual factors of the data. We therefore only describe some of the challenges that arise for nurse clinicians and researchers analyzing big data.

Because most big data is not collected for a specific purpose, missing data is often an issue. Therefore, approaches to “link” several data sources in order to fill gaps in the data structure is an important strategy to overcome limitations of missing data (Weber et al. 2014). For instance if standardized patient identifiers are missing, algorithms taking patient characteristics into account might allow the linkage of different data sources. Because big data is necessarily observational analyses, we are often concerned with addressing selection bias and the estimation of causal effects (Rubin 2007). Although this is not unique to big data, analytical techniques such as propensity scores and mixed models are highly relevant in this context. Finally, privacy and security concerns exist in particular in the context of big data: with more data being linked to increase the depth of information or to compensate missing data, the more difficult it becomes to de-identify individuals (Sweeney 2000; Gymrek et al. 2013). Therefore the risks for patients and the legal or ethical aspects of using big data should always carefully considered (Kohane and Altman 2005).

Despite these challenges the described examples showcase how big data analytics help to solve some of the difficulties not only in big data, but also in more traditional research. For instance genetic or textual data might contain thousands of variables, making it almost impossible to select variables manually and guided by theory, requiring new analytical approaches in order to reduce the number of variables (Shah 2015). Example 1 is such an example, the text mining generates a 1800 × 4390 matrix, which does not pose a computational problem but a
conceptual—which variables of the 4390 should be used for predicting a “positive care experience”? Here supervised machine learning algorithms offer a solution, which is based on the performance of the algorithm providing a clear succinct criterion. Dealing with a different issue study 2 provides an example where the inherently difficult task of creating a sufficiently ‘balanced’ propensity score model is ‘automatized’ by the mean an evolutionary algorithm.

5.6 Conclusions

Big data coming from various sources (e.g., electronic health records, social media, audio or video data) in structured or unstructured form is growing at high speed. The processing and analysis of this mass of data is easily beyond the abilities of traditional methods. On one side, the amount of data exceeds the computing resources, and on the other side, and more importantly traditional analytical, but also conceptual means are overwhelmed by this amount of data.

Big data analytics and application in health care are still at an early stage of development and fairly slow in comparison to other industries. One reason explaining the late adoption of big data in health might be the contradiction between the population focus of big data and the practice of personalized medicine (Sacristán and Dilla 2015). On the other hand, fitness trackers being only one example are highly popular and applications using data from these devices are rapidly developing. Additionally, big data is expected to save money by preventing repeated testing and to provide better clinical decision support (Okun et al. 2013; Manyika et al. 2011; Feldman et al. 2012). Financial pressures on healthcare organizations will fuel the development and utilization of big data (Silver et al. 2001). For nursing practice, but also for nursing research big data provides great opportunities, however training and education need to keep up with this development to prepare nurses for utilizing big data for improving patient care.

References


Dhar V. Data science and prediction. Commun ACM. 2013;56:64–73.


Laney D. 3D data management: controlling data volume, velocity and variety. META Group Research Note 2001;6:70.
Schroeck M, Shockley R, Smart J, Romero-Morales D, Tufano P. Analytics: the real-world use of big data: how innovative enterprises extract value from uncertain data. IBM Institute for


Case Study 5.1: Value-Based Nursing Care Model Development

John M. Welton and Ellen Harper

Abstract This case study will describe the development of a national consensus model to measure patient-level nursing intensity and costs-per-patient in multiple care settings that support the continuum of care and produce objective measures of nursing value. Specifically, this case study will describe the creation of a common data dictionary that describes patient, nurse, and system-level data elements extracted from existing data sets to populate a conceptual model to measure nursing value.

Keywords Data dictionary • Data models • Value-based nursing care • Business intelligence • Nursing care quality and costs • Room and board • Diagnosis related group • Florence Nightingale

Jennifer hurried to finish her charting at the end of a busy night shift in the cardiac stepdown unit (CSU). The last item on her list was to complete her bill for each of the 4 patients she was assigned. This is a new change as her hospital recently implemented a value-based nursing care model. Billing for nursing care is a way to link individual nurses with each patient to identify the unique resources expended for each patient and use these data internally to allocate nursing time and costs. At 8:00 a.m. Jennifer attended her monthly practice council. Selected nurses from the CSU reviewed the overall patient care, adverse events, patient satisfaction and patient level nursing costs. In the past six months, the CSU was able to reduce length of stay and nursing care costs for congestive heart failure patients, their top diagnosis, by assigning a more experienced nurse on admission who often had a reduced assignment. While the first-day costs were higher than average, the nursing business intelligence analytics clearly demonstrated better outcomes by improving nursing care in the vulnerable first 24 hours after admission. Lastly, the nurses reviewed the real-time quality metrics assigned to each nurse. The CSU nursing performance metrics included medication administration delays, pain management, and glycemic control. Each nurse was rated by using time and event stamped data from the electronic health records: for example when a medication was due and the time difference for when it was administered. Jennifer smiled when she saw all her scores had improved from last month mostly from being better organized and “keeping her head in the game.”

This is a fictional story. However emerging new techniques for extracting data from the electronic health record (EHR) can identify the added value of nursing care as well as the individual contribution of each nurse. Nursing care value, in its simplest...
form, is the relationship between quality and costs, or quality and outcomes of care (Pappas 2013; Simpson 2013). When nursing care is appropriate and optimum, adverse events such as injuries, pressure ulcers, infections, and medication errors are reduced thereby decreasing the added costs associated with morbidity and mortality (Spetz et al. 2013; Staggs and Dunton 2014; Yakusheva et al. 2014).

In the current environment, quality and outcomes of care are measured at the individual patient level and aggregated across many patients within an identifiable entity such as a hospital inpatient unit or skilled nursing facility. The actual or true costs of nursing care are not directly measured for each patient and typically averaged across many nurses and many patients (Sanford 2010). Nursing care is rolled up to daily room and board charges (Thompson and Diers 1991), which hides the added value nurses bring to the bedside. Without a direct way to measure the actual or “true” cost and resources expended by nurses for each patient, it will be impossible to measure nursing care value (Welton 2010).

### 5.1.1 Value-Based Nursing Care and Big Data

In attempting to arrive at the truth, I have applied everywhere for information, but in scarcely an instance have I been able to obtain hospital records fit for any purposes of comparison. If they could be obtained, they would enable us to decide many other questions besides the one alluded to. They would show subscribers how their money was being spent, what amount of good was really being done with it, or whether the money was not doing mischief rather than good; they would tell us the exact sanitary state of every hospital and of every ward in it, where to seek for causes of insalubrity and their nature; and, if wisely used, these improved statistics would tell us more of the relative value of particular operations and modes of treatment than we have any means of ascertaining at present. They would enable us, besides, to ascertain the influence of the hospital with its numerous diseased inmates, its overcrowded and possibly ill-ventilated wards, its bad site, bad drainage, impure water, and want of cleanliness—or the reverse of all these—upon the general course of operations and diseases passing through its wards; and the truth thus ascertained would enable us to save life and suffering, and to improve the treatment and management of the sick and maimed poor.

Florence (Nightingale 1863, p. 176)

If we only had the data … Nightingale’s lament may be coming close to realization. A group of nurses and other professionals began meeting in June 2013 at the University Of Minnesota School Of Nursing to address the problem of growing amounts of healthcare and nursing data (Clancy et al. 2014; Westra et al. 2015a). An action plan committed nearly 100 attendees to address a wide range of issues related to data science, informatics, and how to leverage the burgeoning amounts of information contained with the EHRs to develop new approaches, methods, and analytics that ultimately will improve patient care outcomes and decrease costs (Westra et al. 2015b).

One expert workgroup was formed to address the issue of how to measure nursing value and develop new techniques that will provide real-time metrics to monitor quality, costs, performance, effectiveness, and efficiency of nursing care (Welton
2015). During an initial one-year interaction, members of the nursing value expert workgroup identified core issues needed to explicate nursing value (Pappas and Welton 2015; Welton and Harper 2015):

1. Identify individual nurses as providers of care
2. Define nursing care as the relationship between an individual nurse and patient, family, or community
3. Link nurses directly to patients within the EHRs and measure value at each unique nurse encounter
4. Identify nurse and nursing care performance at the patient, unit and hospital or business entity unit of analysis
5. Develop patient level nursing costs based on the direct care time and other measures of resources or services provided to patients (or families, communities)

The value-based nursing care model focuses on the individual nurse rather than nursing care as the basic unit of analysis. The model is software agnostic and setting neutral. The primary analytic approach is to use events and time-stamps to link nurses and patients. For example, in the vignette, Jennifer organized her care delivered to her four patients as distinct services, interventions, assessments, etc. Each patient has unique needs and these vary across the trajectory of an illness or episode of care such as a hospitalization. The ability to discern differences in time spent with individual patients as well as the associated dollars expended across a patient population provides much greater detail and more timely and actionable information about nursing care and added value that can be used for clinical and operational decision making.

Because the value-based nursing care model is focused on individual nurses, performance can be measured for each nurse using EHR data. In the vignette, the CSU is focusing on pain management. Each nurse conducts assessments, identifies problems (e.g., acute pain), provides interventions such as administering PRN opioid medication, and reassesses a patient’s response to interventions. Unit practice guidelines can be used to develop useful and objective information. For example, if the standard of care in the CSU is assessing for pain every 4 hours and follow up within 30 minutes after an intervention, extraction of nursing assessment documentation time data and pain acuity scores as well as the time and dose of PRN medications can be used to identify practice guideline adherence, patterns in using PRN opioids, and overall response to a nurse’s care for patients in pain. These data can be posted or used in value-based nursing performance metrics shared within the unit.

5.1.1.1 Extracting Nursing Data from the EHR

One of the vexing problems in building new patient and nurse level analytic models is the difficulty in finding and extracting key data from the many tables in a modern EHR and developing ways to do this across multiple software platforms.
Substantial resources are needed for even simple data inquiries and reports. If hospitals or other health care settings wish to compare results and information across multiple settings, a common method and model for extracting similar data is needed.

Part of the efforts of the value-based nursing care expert workgroup is to develop a common data model that can allow multiple healthcare settings to extract similar nursing related data. The model is a roadmap for information technology and business analyst professionals to develop extraction code and pull data into a common repository for planned or ad hoc analysis. A preliminary model has been proposed by the value-based nursing care expert workgroup (Welton & Harper 2016). The common data model allows extraction and collection of complex nursing related data across many different software platforms and settings (Fig. 5.1.1). Ultimately, this common data model provides a framework for using and analyzing data about nursing care in many different settings and across many different nurses.

5.1.1.2 Nursing Business Intelligence and Analytics (NBIA)

The data model provides a template as well as the key data that can be used in complex analysis and business intelligence efforts. For example, future systems will be able to monitor performance of nurses administering medications by deriving the time between when a medication was due and when administered using bar code technology (Welton 2013). Pattern detection algorithms will be able to detect when medication administration is becoming increasingly delayed due to high workload, which may be a precursor to late medication doses. Medication administration times can be used to analyze the relationship between unit churn, patient acuity and medication administration performance. Individual nurse performance can be monitored and specific questions addressed that may indicate difficulty in meeting clinical needs due to high complexity and the amount of drugs administered (Kalisch et al. 2011, 2014; Ausserhofer et al. 2014). Focused examination of high-risk drugs such as aminoglycoside antibiotics can be used to link operational aspects of nursing care such as staffing and assignment, with short term clinical goals such as avoidance of nephrotoxicity.

5.1.2 The Cost of Nursing Care

In the value-based nursing care model, the actual services delivered as well as the associated time are allocated to each patient. This overcomes a longstanding problem of using average time to identify costs of nursing care. What is an “average” patient? The ability to link nurses to patients and apply different nursing care resources accurately to each patient provides a means to detect differences in nursing intensity and costs across an episode of care such as a hospitalization as well as compare a similar patient within a specific diagnosis
located in a Diagnosis Related Group (DRG). Having the actual cost of nursing care and overhead costs such as management, benefits, and so forth, provides a way to estimate actual dollars expended for each patient and links to the billing and reimbursement system. In Fig. 5.1.1, components of the data model link patients, nurses, and charges.

This nursing value common data model provides a way to extract similar data across different EHRs and link nurses directly to patients. For example, if we were interested in examining the effects of young nurses (new graduates with less than...
1 year experience), the DateRN field in the Nurse table would be used as the date the nurse was first licensed and then calculate the difference in time in months for the current patient that nurse was caring for and then identify the cost of care (Wage) multiplied by the hours of care given to a specific patient. Data on nursing costs and experience could be summed for an individual patient across a hospitalization, and compared to other patients based on similar or different outcomes such as length of stay, total nursing time and costs, etc.

Patient-level costing for nursing care achieves the goal of better understanding cost drivers within the health care system (Kaplan and Porter 2011). Using nursing business intelligence and analytic tools previously described, nursing and healthcare finance leaders could identify nursing cost drivers and possible nursing intensity outlier patients. Using this new information, nursing care could be adjusted to achieve better matching of nurses to patients, and optimize assignments to achieve best clinical outcomes at the lowest costs of nursing care.

5.1.3 Summary

Value-based nursing care represents a new approach to using data to identify the added value nurses bring to patient care. It is a way to realize Nightingale’s vision. Collecting and analyzing data at the individual nurse-patient encounter provides greater granularity for identifying clinical and operational intelligence that can inform providers in near real-time to a range of important clinical and operational information needs.

In the vignette, Jennifer was an active participant in interpreting complex data derived from the clinical assessments, interventions, and outcomes identified in the nursing documentation of the EHR. In this fictional setting, nurses are individually and collectively accountable for their care. The goal is to achieve high value outcomes by optimizing nursing care at all the “touchpoints” where nurses interact with patients.

New analytic techniques will inform clinical and operational decision making in the here and now rather than waiting weeks or months. These new data tools will decrease the time between information and action. A value driven data environment will create new information to measure the nurse-patient encounter and share and compare data across the broad spectrum of healthcare that ultimately will achieve Nightingale’s vision to seek and realize excellence in everything we do.

References


Sanford KD. Nurse staffing. Finding the right number and mix. Healthc Financ Manage. 2010;64(9):38–9.


Welton JM. Nursing and the value proposition: how information can help transform the healthcare system. Minneapolis, MN, August 12–13: University of Minnesota, School of Nursing, Center for Nursing Informatics; 2013. Accessed 1 May 2016


