Data science for nurses

Learn how to apply this skill to nursing projects.

By Alvin D. Jeffery, PhD, RN-BC, CCRN-K, FNP-BC

THE recent emergence of big data provides new insights into healthcare. Whether predicting clinical deterioration, adverse events, or even changes in patient census, data inform clinical care. Nurses play a major role in entering data into electronic health records (EHRs), and we should be using available tools to analyze and use these data to improve care delivery. Although significant advances have been made in collecting and analyzing the large amount of healthcare data now available, clinicians can be easily intimidated by some data science concepts, such as machine learning and artificial intelligence. Familiarizing yourself with data science and learning how others have implemented it may enhance your knowledge and confidence.

Data science defined

Numerous data science definitions exist, but one perspective emphasizes what it does rather than what it is. Donoho notes that data science focuses on gathering, preparing, exploring, representing, transforming, computing (with), modeling, visualizing, and presenting data. It lies at the intersection of domain knowledge, computer science, statistics, and data visualization/presentation. Data science began to emerge as a more distinct (and popular) field in the early 2000s as data sets grew in conjunction with increased access to affordable computing resources.

Various professions (biomedical, finance, law) use the same data science methods, but each field captures and processes data differently. The Berkeley School of Information divides the data science process into five steps: capture, maintain, process, analyze, and communicate.

Capture

Nurses and other clinicians play a critical role in this phase when they enter data into the EHR. They generate data via direct observation and documentation, automated capture from equipment (vital sign machines, ventilators), laboratory results, and external sources such as social determinants of health.

Maintain

Most large volumes of captured data are stored in databases or data warehouses. Important data should be duplicated and stored in multiple locations to create redundancy in the event of a natural disaster that destroys one of the physical locations or if a server unexpectedly malfunctions. Database analysts, data engineers, and computer scientists participate in data maintenance.

Process

The more exciting aspects of data science begin in the process phase where data scientists prepare (or clean) data for analysis. Common processing activities include identifying (and possibly inputting) missing data (not all patients will have laboratory values collected during a clinic visit), removing outliers (a heart rate erroneously entered as 450), and defining the outcome (using diagnostic codes to identify patients with heart failure). In the case of unstructured data (text

Nursing-relevant data science projects

The following table highlights several recently published studies that leveraged data science methods in nursing-relevant projects.

Authors	Outcome	Data science methods
Chang et al	Hospitalization	The authors validated an existing predictive model (specifically, a *logistic regression* model) for determining which patients are most likely to be hospitalized.
Nakatani et al	Falls	To add nursing data to a fall prediction model, the authors applied *natural language processing* to nurses' free-text documentation. The resulting embeddings (numerical representation of words) served as predictors.
Park et al	Mortality	The authors' primary purpose was to use *logistic regression* to predict mortality, but they also used *network analysis* methods (graphs to identify and represent relationships between people) to incorporate provider information as predictors in the model.
Patterson et al	Falls	The authors developed predictive models to identify which patients were likely to be readmitted within 6 months due to a fall-related injury. They used *regression* (a statistical method that specifies relationships—linear, logistic—between predictors and outcomes) and *tree-based algorithms* (machine learning methods—random forest, AdaBoost—that split predictors for more homogenous outcome groups).
Sandsdalen et al	Palliative care	To personalize care plans, the authors wanted to identify patient profiles that represent patient satisfaction. They applied a *cluster analysis* (algorithms that create groups in the data by maximizing within-group similarity and minimizing between-group similarity) to survey data.

notes, radiology images), data scientists take additional steps to convert them into numeric variables for analysis. Most data scientists agree that they spend at least 80% of their time on a data science project in the process phase.

Analyze

In this phase, processed data are fed into a statistical model (a mathematical representation of relationships between variables) or a machine learning algorithm (a procedure for applying steps to learn from data). In either case, this phase focuses on explanation, association, prediction, and pattern recognition.

The two most commonly used groups of analytical methods are supervised and unsupervised. Most of these methods require that data initially be represented as numeric or categorical variables. One exception is natural language processing, which focuses on a computer's ability to use language data.

Supervised methods learn from labeled example data to generalize findings to all possible input data. For example, imagine your hospital wants to predict which patients are most likely to be readmitted within 30 days. Data scientists select example data comprising a group of previously hospitalized patients with knowledge of which were readmitted within 30 days and which weren't. A patient's readmission status serves as a *label* that's used to make a prediction based on input data (comorbidity categories, discharge vital signs, insurance status). Some common supervised methods include regression (linear, logistic, survival), tree-based (decision trees, random forests), Naïve Bayes, and deep learning (neural networks).

Unsupervised methods learn from unlabeled

More information

To learn more about data science, develop specific skills, or find out how to become a data scientist, visit nursingdatascience.org.

To get involved in data science activities that support nursing care delivery, consider joining these organizations:

- American Medical Informatics Association (amia.org)
- American Nursing Informatics Association (ania.org)
- Nursing Knowledge: Big Data Science Conference (z.umn.edu/bigdata)

example data. They use characteristics to categorize or group data based on statistical criteria. For example, a public health nurse might have access to environmental and census tract data for a county. They could use an unsupervised method to create groups of neighborhoods that share similar features in their input data (similar rates of pollution exposure, education attainment, income level). Some common unsupervised methods include cluster analysis (k-means clustering, hierarchical) and dimensionality reduction (principal component analysis, linear discriminant analysis).

Natural language process techniques (bag-of-words, word embeddings) allow data scientists to convert unstructured text into numeric variables for use in other unsupervised methods or even supervised methods. (For more information about these algorithms, visit scikit-learn.org/stable and metacademy.org.)

Communicate

Reports generated in the final phase of the data science life cycle describe the project's process and results. Visual aids (figures or graphs) can help concisely represent the data and results. Model- or algorithm-specific metrics aid analysis—how well predictions compare to actual labels in supervised methods or how well groups were created in unsupervised methods. Those presenting the data should explicitly state any limitations within the previous four phases so that decision makers can interpret findings in light of assumptions made along the way. (See *Nursing-relevant data science projects*.)

Call to action

Data science continues to evolve as data increase in quantity and diversity. Most nurses don't actively participate in the maintain, process, and analyze phases of a data science project, but they do play a significant role in the capture and communicate phases. Nurses are essential team members in all healthcare data science projects because they can contribute to understanding nuances of data capture, which influences how data are processed and analyzed. In addition, many nurses in leadership and decision-making positions benefit from data science reports. Data science project findings have the potential to transform nursing practice. A knowledgeable nursing workforce that can interpret these findings has the potential to improve health and healthcare. AN

Alvin D. Jeffery is an assistant professor at Vanderbilt University School of Nursing and a nurse scientist at Tennessee Valley Healthcare System, U.S. Department of Veterans Affairs, in Nashville, Tennessee. **Dr. Jeffery received support for this work from the Agency for Healthcare Research and Quality (AHRQ) and the Patient-Centered Outcomes Research Institute (PCORI) under Award Number K12 HS026395, as well as the resources and use of facilities at the Department of Veterans Affairs, Tennessee Valley Healthcare System. The content is solely the responsibility of the author and does not necessarily represent the official views of AHRQ, PCORI, the Department of Veterans Affairs, or the United States Government.**

References

Bates DW, Saria S, Ohno-Machado L, Shah A, Escobar G. Big data in health care: Using analytics to identify and manage high-risk and high-cost patients. *Health Aff (Millwood)*. 2014;33(7):1123-31. doi:10.1377/hlthaff.2014.0041

Berkeley School of Information. What is data science? datascience.berkeley.edu/about/what-is-data-science

Chang ET, Piegari R, Wong ES, et al. Which patients are persistently high-risk for hospitalization? *Am J Manag Care*. 2019;25(9):e274-81.

Donoho D. 50 years of data science. *J Comput Graph Stat*. 2017;26(4):745-66. doi:10.1080/10618600.2017.1384734

Nakatani H, Nakao M, Uchiyama H, Toyoshiba H, Ochiai C. Predicting inpatient falls using natural language processing of nursing records obtained from Japanese electronic medical records: Case-control study. *JMIR Med Inform*. 2020;8(4):e16970. doi:10.2196/16970

Park Y, Karampourniotis PD, SyllaI I, Das AK. Hierarchical patient-centric caregiver network method for clinical outcomes study. *PloS One*. 2019;14(2):e0211218. doi:10.1371/journal.pone.0211218

Patterson BW, Engstrom CJ, Sah V, et al. Training and interpreting machine learning algorithms to evaluate fall risk after emergency department visits. *Med Care*. 2019;57(7):560-6. doi:10.1097/MLR.0000000000001140

Sandsdalen T, Wilde-Larsson B, Grøndahl VA. Patients' perceptions of the quality of palliative care and satisfaction—A cluster analysis. *J Multidiscip Healthc*. 2019;12:903-15. doi:10.2147/JMDH.S220656

Tideman S, Santillana M, Bickel J, Reis B. Internet search query data improve forecasts of daily emergency department volume. *J Am Med Inform Assoc*. 2019;26(12):1574-83. doi:10.1093/jamia/ocz154