Published in final edited form as:

West J Nurs Res. 2017 January ; 39(1): 20–41. doi:10.1177/0193945916672828.

Framework for Mining and Analysis of Standardized Nursing Care Plan Data

Ashfaq Khokhar, PhD[§] [Professor and Department Chair], Muhammad Kamran Lodhi[#] [Doctoral Student], Yingwei Yao, PhD^{*} [Research Associate Professor], Rashid Ansari, PhD[#] [Professor and Department Head], Gail Keenan, PhD, RN FAAN^{*} [Professor], and Diana J. Wilkie, PhD, RN, FAAN^{*} [Professor and Director, Academic Center of Excellence for Palliative Care Research and Education]

[§]Illinois Institute of Technology, Chicago, IL, USA

[#]University of Illinois at Chicago, Chicago, IL, USA

^{*}University of Florida, Gainesville, FL, USA

Abstract

Despite an unprecedented amount of health related data being amassed from various technological innovations, our ability to process this data and extract hidden knowledge has yet to catch up with this explosive growth. Although nursing care plans can be an effective tool to support the achievement of desired patient outcomes, their online collection, storage, and processing is lagging far behind. As a result, the impact of nursing care is not well understood from qualitative as well as quantitative perspectives. In this paper we first outline a complete life cycle of nursing care data, and present a knowledge discovery and analysis framework for such data sets. We also highlight big data issues pertaining to the analysis of nursing care data. Using an exemplar data set, we demonstrate the broad applicability of the proposed framework by showing knowledge discovery results for different outcomes related to patients, nursing staff, and administrators.

Due to rapid advances in digital technologies, the use of electronic media to capture, process, and store information is witnessing growth at an unprecedented pace (Hilbert & López, 2011). While the stored information is reaching zettabytes (Mearian, 2011), our ability to process and analyze such large amounts of data, however, is lagging far behind. This dichotomy is especially true for the healthcare sector, which was mandated by the American Recovery and Reinvestment Act of 2009 to adopt and demonstrate “meaningful use” of electronic health records (EHR). Within EHR systems, the collection and meaningful reuse of data related to nursing care of hospitalized patients is even further lagging.

The “Meaningful reuse” of EHR data, as defined by HealthIT.gov, consists of using digital medical and health records to achieve the following:

Corresponding author: Ashfaq Khokhar, PhD, Professor and Department Chair, Illinois Institute of Technology, Chicago, IL, USA; phone 312.567.5780, Fax 312.567.8976, ashfaq@iit.edu.

Disclosure. The HANDS software that was used in this study is now owned and distributed by HealthTeam IQ, LLC. Dr. Gail Keenan is currently the President and CEO of this company and has a current conflict of interest statement of explanation and management plan in place with the University of Florida.

- Improve quality, safety, efficiency, and reduce health disparities
- Engage patients and families
- Improve care coordination, and population and public health
- Maintain privacy and security of patient health information.

Unfortunately, there are a number of factors that contribute to limited meaningful reuse of EHR data. One of the biggest factors is the complexity and veracity of the data. The heterogeneity of the data sources, and multiple data collection standards are additional factors that make EHRs a prime Big Data example. A number of new tools and techniques are emerging to enable the discovery of knowledge from the prolific amount of health-related data being amassed in various technological innovations (e.g., the Internet, patient devices, and EHR). In fact, combining and analyzing different sources and formats of data are at the heart of the big data science discovery movement. One way for nursing to participate meaningfully in this effort is to generate data that are a valid representation of nursing care and amenable to efficient processing and analysis. In this paper, we report our research team's strategy for uncovering knowledge in a high-quality source of structured nursing data gathered in 9 nursing units (4 acute care hospitals) in which the same tool was used for 12 to 36 months. These efforts are a first step toward intelligent merging of nursing data with other sources to further examine and improve the impact of nursing within a wider context.

Below we describe our knowledge discovery framework and discuss big data issues related to our analysis and results. As shown in Figure 1, the knowledge discovery framework explicates the analytical component of the Nursing Plan of Care Data Life Cycle (Ecosystem). In summary, the Ecosystem visually depicts the cycle our team follows to generate and improve the production of standardized nursing data that were used in the knowledge discovery exemplars presented (Keenan et. al, 2012). Our analysis specifically required grappling with big data issues such as high dimensionality, sparseness, feature extraction, and feature selection to produce findings relevant to patients, care providers, and healthcare organizations. Each of these big data issues is discussed within the context of an exemplar from the data gathered with a nursing EHR system, Hands-on Automated Nursing Data System (HANDS) (Keenan, Yakel, & Marriott, 2005).

Background

Our knowledge discovery framework (Figure 2), presented in greater detail below, utilizes valid and reliable standardized nursing care plan data extracted from the HANDS, a well-tested EHR system designed for collecting such data (Keenan et. al, 2012). Though our exemplars are based on one system's data, the framework can be applied to nursing data collected by EHRs with other care planning attributes and features (e.g., user interface, database structure, standardized terminologies, and decision support). Further, it is important to note that standardized nursing care plan data collected in an EHR can be made available for analysis by itself or merged with other EHR data (e.g., lab, radiology, medical) ultimately enabling a comprehensive evaluation of factors that influence patient outcomes. The knowledge discovery framework in this paper intentionally focuses only on nursing data

since this is considered a basic building block for effectively merging and analyzing it with other healthcare data.

The Case for Standardized Data Collection

When data stored in digital form are not structured or coded using standardized terminologies and linked with other related data items, its processing and analysis pose formidable challenges. Unstructured data refers to information that does not have a pre-defined model or has not been organized in any pre-defined structure linking different data elements. Without standardization it is nearly impossible to determine if different words/terms used to describe the same process are actually comparable. This is particularly problematic when trying to compare words across units and across settings where standards of care and cultures can vary tremendously. Unstructured documentation of care not only influences the validity of the data collected but also can be a source of unintended communication errors resulting from misunderstanding and misinterpretation. Another challenge is the transformation of the raw data into usable information. Without this transformation, the current data mining techniques often miss a substantial amount of the information because of different ambiguities. The processing of unstructured data is also much more labor-intensive and therefore more costly, even with advances in natural language processing. In the absence of standardization, the impact of nursing care has remained fairly invisible due to the infeasibility of evaluating non-structured nursing care data except on a very small scale. Recently, there have been a number of attempts to capture robust structured data sets using standardized nursing terminologies (Tastan et al., 2014).

In our work we have used standardized nursing care plan data in which the nursing elements of diagnosis, outcome, and intervention are coded respectively with NANDA-I (Herdman & Kamitsuru, 2014), Nursing Outcome Classification (NOC) (Moorhead, 2013), and Nursing Intervention Classification (NIC) (Bulechek, Butcher, Dochterman, & Wagner, 2013). For the sake of completeness, we next provide a brief description of the exemplar data set.

The Data Set

Our data set was gathered with HANDS in a multisite study in 4 unique Midwest hospitals (n=8 units) for 12 or 24 months (Keenan et al., 2012) and a 9th unit for 36 months (Keenan & Yakel, 2005). The database stores care planning information (nursing diagnoses, interventions, outcomes, and goals) for each shift with the nurse (experience, education) and patient (age, race, sex) demographics (Figure 3). The information in each care plan is structured using standardized terminologies, as depicted in Figure 4. In Figure 5, we show an example plan of care initiated at the time of admission that lists the current outcome levels, and also sets the expected goals for all the outcomes tracked for the patient.

The database is extremely high-dimensional with 747 attributes (variables), but relatively few are active for any given patient since a patient is unlikely to have hundreds of diagnoses, outcomes, and interventions. On average, a patient has four different diagnoses, five different outcomes and ten different interventions on his/her plan of care, resulting in a highly sparse dataset with 2% density. Across a large group of patients, all 747 attributes, however, would have some data. The size of this data set, its high dimensionality, and

sparseness make it an ideal vehicle to study Big Data issues relating to nursing care. The dataset used in our study consists of 357,522 nursing care plans entered by 787 nurses corresponding to 42,403 episodes from 34,927 unique patients. Each patient episode, defined as a continuous stay at one hospital unit, spans 4 days and has 8.5 care plans on average. As reported previously, the validity and reliability of the nurses' use of the diagnoses, intervention, and outcome terms was excellent as was the fidelity of their use of the system at each handoff (Keenan et al., 2012).

Knowledge Discovery and Analysis Framework

For any given data set, the knowledge discovery involves a series of steps that must be executed to extract the hidden information. In this section we present our framework for knowledge discovery and analysis in nursing care data in greater detail. Figure 2 depicts the flow and relationship of different modules and processes in this framework. At a very abstract level, the proposed framework consists of the following sequential, and at times iterative, steps to carry out the analyses:

- data extraction from the database or data warehouse,
- data cleansing and preprocessing including feature extraction and dimensionality reduction,
- iterative analysis and prediction model development and verification using data mining tools.

In this section we first describe research issues related to feature extraction (dimensionality reduction), data analysis (including importance sampling), data mining (including execution of association rule mining and classification), and analysis and interpretation of results (including importance of domain experts in the loop). Finally, we describe how the lessons learnt from such an analysis can be made available at the point of care for improving healthcare delivery.

Feature Extraction—In our analyses, the main features (variables) we studied are the nursing diagnoses, interventions, and outcome goals and progress toward goals updated for each patient at the end of every shift. In addition, our analysis includes patient characteristics (e.g., age, gender, race, medical diagnoses, medical history), nurse characteristics (e.g., education, experience, work status, work load) and other derived variables such as nursing care characteristics (e.g., care continuity, worked hours per patient day), and episode characteristics (e.g., length of stay, time of admission, discharge status).

The first step of our analysis is to extract the relevant features to form the analytic dataset. With nursing diagnoses, outcomes, and interventions coded using the highly valid standardized nursing terminologies of NANDA-I, NOC, and NIC, respectively, the often difficult task of extracting nursing care plans is straightforward with HANDS data. On the other hand, medical diagnoses and history were entered as free text. To extract features from these data fields, an iterative process is needed. First, clinical experts need to generate a codebook for common medical terms including abbreviations. Natural language processing is then used to parse the free text entries for each patient episode. These two steps are

applied iteratively with both the codebook and the parsing algorithm being refined in each iteration until important features in these fields are extracted.

Some features need to be derived from the raw data in the database. For example, the length of stay for each patient episode can be obtained by computing the difference between the time of admission and time of discharge. The derivation of other features is much more involved. For example, a research study may want to examine the effect of part-time nurses, defined as nurses that have worked for less than 30% of the full time hours in the previous two weeks. To extract this feature, we need to compute for each nurse and for each day during the entire study period, the number of hours they worked in the two weeks prior. Another example is the care continuity, which needs to be aggregated for each patient taking into account multiple factors related to nurse staffing pattern, including the time a nurse last saw a patient, the amount of prior exposure to a patient, fragmented (4-hour) shifts, and so on (Stifter et al., 2014).

Feature Selection—Even with high quality structured data there are a number of factors that need reconciliation due to the complexity of the data. Specifically one that impedes the analysis of nursing care plan data, and for that matter any EHR related data, is the high dimensionality and sparseness of such data sets.

In our plan of care data for EOL patients, we identified 144 distinct NANDA-I nursing diagnoses labels, 268 distinct NOC outcome labels, and 412 distinct NIC intervention labels. Over 800 dummy variables are needed to represent all of these labels. In addition, patients typically had multiple medical diagnoses and multiple medical problems in their history, not to mention other patient and nurse characteristics. A model trying to utilize all of them as predictors will likely have both extremely high computational cost and abysmal prediction performance.

Firstly, the complexity of many traditional data mining algorithms and models grows exponentially with respect to the number of dimensions because of the exponential number of column combinations. Therefore, as the number of dimensions for a dataset increases, so does the search space as well as the complexity of the algorithm. Also, predictive power of prediction models reduces as the dimensionality increases, known as Hughes effect or Hughes phenomenon (Hughes, 1968).

A second problem is that the similarity between neighboring points in a high dimensional data-space reduces as the dimensionality increases. For any point in a high dimensional data space, the expected gap between Euclidean distance to the closest neighbor and to the farthest point in the space shrinks as the dimensionality grows (Beyer, Goldstein, Ramakrishnan, & Shaft, 1999). The notion of the nearest neighbor of a point is therefore meaningless unless the sample size also grows exponentially.

To handle the problem of dimensionality in data mining, different algorithms have been introduced, which instead of column-wise enumeration method, have used row-based enumeration method (Liu, Han, Xin, & Shao, 2006; Pan, Cong, Tung, Yang, & Zaki, 2003). These methods assume that datasets have thousands of columns (attributes or variables), but

a much lower number of rows (observations). Carpenter algorithm (Pan et al., 2003) uses a bottom-up approach, whereas (Liu et al., 2006) uses a top-down search method approach. However, both these methods work best for dense datasets. Data mining has been used to extract knowledge in some such data sets, albeit with limited success (Chen, Hu, & Zhang, 2006). Typical EHR data, on the other hand, are highly sparse. In the HANDS dataset, on average, a patient has four different diagnoses, five different outcomes and ten different interventions on his/her plan of care, resulting in a highly sparse dataset with 2% density.

Besides the high-dimensionality problem, sparsity poses significant challenges as well. Many data mining and machine learning methods work best for “denser” datasets (Zao, Albright, & Cox, 2014). To overcome the twin challenges of high dimensionality and sparsity, we therefore needed to perform feature selection.

Descriptive analysis is often a first step in feature selection. Descriptive statistics such as mean, standard deviation, frequency, and percentage are very useful in identifying predictors with low variability and thus most likely low impact that can be excluded. Examination of bivariate association can lead to identification of additional low impact features. When a considerable number of variables are present in a dataset, association rule mining algorithms can also be used to further reduce dimensionality. Association rules identify prevalent correlations and patterns. These patterns could be used as a basis for predictive modeling. Since the number of rules generated is often large, graphical approaches could be used to present them efficiently. For instance, an algorithm can generate a network of relations among the variables that are present in an association rule. In such a network, rules and associated variables that are connected in cliques, can help to determine the variables of interest and hence help in dimension reduction.

Besides excluding low impact features, we must also seek to consolidate and aggregate features. NANDA-I, NOC, NIC are standardized nursing terminologies with well-defined taxonomic structures (e.g., domains and classes). For a given research question, we can aggregate individual labels into domains and classes to greatly reduce the problem of dimension. Both expert input and model validation can be used to assess if these aggregations are appropriate. In addition to the taxonomic structures, implicit structures such as hierarchical relation among different labels can be exploited for data reduction.

For example, plan of care data in the our dataset were stored in a relational database, where each patient was assigned one or multiple nursing diagnoses (NANDA-Is) when admitted to a unit, each nursing diagnosis was linked with one or multiple nursing outcomes (NOCs), each of which in turn was linked with one or multiple nursing interventions (NICs). Besides this vertical structure among the three levels of NANDA-NOC-NIC hierarchy, there were also horizontal connections between the multiple nursing diagnoses and the outcomes and interventions that are linked to each. For example, when examining the EOL patients in a gerontology unit we found significant associations between multiple nursing diagnoses and the pain diagnosis (Almasalha et al., 2013). Finally, the plans of care were updated each shift, where the shift nurse would update a patient's condition by assigning numeric ratings to each of the nursing outcomes and add or remove NANDA-Is, NOCs, and NICs as needed, leading to a temporal structure in the data.

A thoughtful and methodical exploratory analysis of the plan of care structures could lead to significant insights on meaningful aggregation of features. In our analysis of plan-of-care data for patients who expired during the episode, we examined the connections between NANDA-I labels related to comfort-oriented care and those related to disease- or illness-oriented or health-restorative care. We were able to identify eleven NANDA-I labels including Death Anxiety and Anticipatory Grieving that could serve as infomarkers for palliative care (Yao et al., 2015). For a given research question, we should try to find similar meaningful infomarkers by examining the structure of plan of care data. Each of these infomarkers will likely include multiple NANDA-I, NOC, or NIC labels, meaning that when studying a related research question, it will make sense to aggregate these labels into a single predictor, leading to more substantial dimensionality reduction.

Concurrent approaches to the feature aggregation worth mentioning here are the principal component analysis (PCA) and factor analysis, employed to reduce dimensionality in datasets by identifying a small number of components or factors capable of reproducing with high fidelity the much larger set of original variables. PCA (Jolliffe, 2002) offers a sequence of best linear approximations to a given high-dimensional observation. It is one of the most popular techniques for dimensionality reduction. However, its effectiveness is limited by its reliance on the assumption of global linearity. Multidimensional scaling (MDS) (Cox & Cox, 2008) is similar to PCA, and suffers from the same drawback. Factor analysis (Frey, 1998) and independent component analysis (ICA) (Hyvärinen, Karhunen, & Oja, 2004) also assume that the underlying manifold is a linear subspace, but differ from PCA in how they identify and model the subspace. The main drawback with all these classical dimensionality reduction approaches is that they only characterize linear subspaces (manifolds) in the data. In addition to the linearity assumption, PCA and related methods may be suboptimal when data sets are very high-dimensional and sparse, and when variables are heterogeneous. Sparse, parallel and non-linear PCA have been developed in specific contexts, even as a generative probabilistic model, but an overall off-the-shelf approach for integrated data is still lacking. Another concurrent approach is cluster analysis, either model-based (e.g., latent class mixture modeling) or distance based (e.g., k-means, hierarchical clustering). The major drawback with these methods is the difficulty in defining the notion of cluster itself in determined contexts, as well as a proper probabilistic model or a multi-domain distance function.

Finally, we include penalty functions for model complexity in our model construction to further reduce dimensionality and to avoid overfitting. One such approach is the LASSO method, which can generate a series of models with an increasing number of features efficiently to allow us to evaluate them for model performance, complexity, and interpretability. For regression models, we can also utilize subset selection or shrinkage methods. Regularization or penalty functions can also be applied to non-parametric methods such as classification and regression trees, support vector machines, and various kernel based methods.

Data Mining Techniques

In the following section, we briefly describe the most common techniques for data mining purposes. Each of these techniques are highly relevant to data mining of nursing plan of care data. Table 1 also provides a list of data mining techniques along with different models that have been used in our framework.

Association Mining—Association mining is one of the most prominent and studied procedures in data mining. It aims to extract interesting correlations, frequent patterns, or associations between items in the transactional databases (Piatetsky-Shapiro, 1991) or other repositories for the purpose of making predictions. An association rule is an implication in the form of if A (antecedent) $\rightarrow$ then B (consequent), interpreted as A implies B. Association rule mining is to discover association rules that satisfy predefined minimum support and confidence from a given database. Support of an association rule is defined as the percentage of the records that contain both A (antecedent) and B (consequent) in the database. Confidence of an association rule is defined as the percentage of records containing A that also contains B and is a measure of the strength of the association rules (Agrawal, Imieliński, & Swami, 1993). Minimum values of confidence and support are usually predefined by the user below which rules are dropped, deemed as not so interesting and useful.

Clustering—Clustering is another technique of data mining. In clustering, objects are grouped together (or clustered) on a particular feature or set of features such that the items in one cluster are more similar to each other than to the items in another cluster (Bailey, 1994). Clustering is mainly used to discover groups or partition the data into distinct groups.

Classification—Classification is the problem of mapping (or classifying) a new data instance into one of several predefined classes or categories (Alpaydin, 2014). This is done on the basis of a training set of data consisting of observations whose category membership is given or known.

Regression: The regression functions are used to determine the relationship among the dependent or target variable and one or more independent variables. Regression analysis attempts to find a function between dependent and independent variables to model the data with the least error (Alpaydin, 2014).

Results

In this section, we demonstrate the broad applicability of our proposed knowledge discovery and analysis framework to extract clinically significant findings. Depending on the research question, different methods can be applied. In a healthcare setting, there are three different stakeholders who have vested interest in the clinical decision. These stakeholders are the patients, the healthcare providers (doctors and nurses) and the organization. Therefore, the results can be categorized into three classes based on the stakeholders.

Patient-oriented results

Patients might be more interested in finding out which hospital provides better care to the patients. This can be visualized using our database that consists of records from 4 different hospitals and 9 different hospital units. As an example, we found that patients suffering from Death Anxiety had a better chance of meeting the Comfortable Death goal when they were admitted to Gerontology units of the hospitals compared to other units of the hospitals (Figure 6).

Care-provider-oriented results

Data mining of the nursing database has the potential to identify the best practices or hidden associations that can help the caregivers to identify and provide better and more effective care to the patients. We used *association mining* to find correlations/anomalies within the data. Following are some examples. Among the patients that had Pain along with a diagnosis of Cardiopulmonary problems, it was observed that patients who had an outcome of client satisfaction: symptom control had a better chance of meeting their expected outcome where the patients did not suffer from coping/stress tolerance problems, as compared to those patients that had suffered from the coping/stress tolerance problems. The finding was statistically significant ($p = .039$). One possible explanation of the findings is that the diagnoses from the Coping/Stress Tolerance Domain include patient diagnoses like ineffective activity planning, ineffective coping, ineffective denial, fear, grieving, powerlessness, impaired resilience, chronic sorrow, and disorganized behavior. These diagnoses infer patients who may be grappling with dysfunctional behaviors at the EOL and thus are not as receptive, motivated, or ready to accept care provided by nurses (e.g., Client Satisfaction indicators include such things as assistance to achieve self-care, assistance to cope with emotional concerns, instructions on how to improve participation in care) who are trying to help them achieve a more functional state. Thus, this can lead to the NOC not being met as patients don't positively perceive these nursing care efforts due to the presence of one of these potentially dysfunctional diagnoses. This is compared to a patient who does not have any diagnosis from the Coping/Stress Tolerance Domain. These patients are most probably coping effectively with their illness and are trying to be involved in their plan of care, and working toward planned outcomes (effective pain management, addressing death anxiety and fear, comfortable death, etc.). These patients are typically much more receptive to nursing efforts and interventions that would be moving them along the care continuum and thus would perceive their level of involvement more positively.

In another example, it was observed that the patients having a diagnosis of risk for falls and suffering from pain, the outcome of Fall Prevention Behavior was met for patients when the nurses had used an intervention from the Family domain. On the other hand, when the nurses did not apply any intervention from the Family domain, those patients did not reach their expected outcome of Fall Prevention Behavior. Once again, this result was significant ($p = .019$). When investigated further, common NICs from Family domain included Family Involvement Promotion and Family Support. Family Involvement Promotion is "Facilitating family participation in the emotional and physical care of the patient" and Family Support is "Promotion of family values, interests and goals" (Bulechek et al., 2013; Dochterman & Bulechek, 2004). Inclusion of these NICs may indicate that the nurses realize that the

patients are now at EOL, and families may be needed to be involved as the patient enters their last days. This may lead to higher chances of NOC being met by the patients. This is compared to patients where family is not being involved, which may lead to depression in the patients as they enter their last days and hence result in NOC not being met.

In the last example, younger patients suffering from pain and cardiopulmonary problems had a higher chance of meeting their fall prevention behavior outcome than the older patients ($p = .034$), as shown in Figure 7. Elderly patients become much weaker and cannot tolerate well the typical forms of activity (walking, position changes from sitting to standing) that they may have historically been able to address. Confusion or disorientation can make it especially difficult for elderly patients to remember to ask for help before getting up or to use assistive devices like a cane or a walker. The confusion or disorientation can be caused by the hospitalization itself (unfamiliarity of room, change in routines, difficulty differentiating between day and night in a hospital), a disease process (possible dementia, Alzheimer's, sepsis, fluid and electrolyte imbalance), or even being sedated (pain, anti-anxiety or sleep medications). This phenomenon is noted more commonly in older patients as compared to the younger age groups which may contribute to the younger patients seeming to meet their outcomes more frequently than older patients.

Organization-oriented results

A key goal of any organization is to find cost-saving measures to reduce the ever increasing healthcare costs without compromising care quality. One of the reasons of the increasing costs of healthcare are readmissions. Reducing repeat hospitalizations can greatly reduce these costs. The readmissions often occur due to poor treatment provided to the patients (Billings et al., 1993; Smith, 2009), more specifically, they are often caused by premature discharges (Goodman, Fisher, Chang, Raymond, & Bronner, September 2011). These readmissions can cost up to \$45 billion annually to taxpayers (Yam et al., 2010). Hospitals are also being fined for early readmissions. Therefore, avoidable readmissions are a huge burden on hospital resources, including the workforce.

Our experiments on predictive modeling for re-admitted patients suffering from pain showed patterns that can be useful to reduce readmission to the hospitals. For example, it was observed that among young patients with a long length of stay, nutrition problems were associated with a significantly higher chance of re-admissions (Figure 8, Lodhi et al., 2016 in review).

Discussion

Nurse informatics experts advocate for a variety of approaches for advancing nursing care knowledge through use of big data and data science techniques. There is merit for the various approaches and in this article we showcase the challenges and present a knowledge discovery and analysis framework that is suitable for a variety of outcomes. We also demonstrate the merits of using a standardized nursing EHR system that is well tested and designed for generating reliable and valid standardized nursing care planning data and supported by substantial research evidence (Almasalha et al., 2013; Anderson, Keenan, & Jones, 2009; Boyd, Funk, Schwartz, Kaplan, & Keenan, 2009; Febretti et al., July 2013;

Johnson et al., 2014 in review; Keenan, Stocker, Barkauskas, Treder, & Heath, 2002; Keenan, Yakel, Lopez, Tschannen, & Ford, 2013; Keenan et al., 2005; Keenan, 2014; Keenan et al., 2002; Keenan, Tschannen, & Wesley, 2008; Keenan & Wilkie, 2014; Keenan et al., 2012; Lodhi et al., 2016 in review; Lodhi et al., 2014; Lodhi et al., 2015; Schneider, Barkauskas, & Keenan, 2008; Sousa et al., 2015; Stifter et al., 2014; Stifter et al., 2015; Tastan et al., 2014; Tschannen et al., 2011; Yao et al., 2015). In this work we have also illuminated important issues inherent with analysis of nursing healthcare data. These issues require thoughtful consideration of appropriate analytic strategies to make meaning of the high-dimensional and highly sparse data. For example, feature extraction and selection issues require careful attention and collaboration among data science experts and healthcare research experts—a team science approach.

Although our dataset and exemplars focused only on nursing related data, our work demonstrates the fundamental importance of this singular focus. The availability of standardized nursing data offers for the first time ever the capability to discover new knowledge about nursing but to do this it has been necessary to devise new methods for extracting this knowledge. This paper included a description of our systematic approach to knowledge discovery that involved tailoring existing methodologies to our standardized nursing dataset. As we have argued in this paper, before merging complex nursing data with other EHR and application data it is important to first understand the richness of this data alone. We have made major strides in understanding how to analyze standardized nursing data but plan additional work in this area. Simultaneously, however, there are also plans to merge the standardized nursing data with other data gathered in the EHRS to evaluate the impact of nursing within a multidisciplinary context. We strongly believe that the tipping point is near at hand and the work described in this manuscript is one of the catalysts to opening the floodgates to the rapid discovery of nursing knowledge. Such knowledge can be fed back to practice, education and research to continually improve nursing care that optimizes patient outcomes.

Acknowledgments

This publication was made possible by Grant Number 1R01 NR012949 from the National Institutes of Health, National Institute for Nursing Research. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the National Institute for Nursing Research. The final peer-reviewed manuscript is subject to the National Institutes of Health Public Access Policy.

References

- Agrawal R, Imieliński T, Swami A. Mining association rules between sets of items in large databases. *ACM SIGMOD Record*. 1993; 22(2):207–216.
- Almasalha F, Xu D, Keenan GM, Khokhar A, Yao Y, Chen YC, Wilkie DJ. Data Mining Nursing Care Plans of End-of-Life Patients: A Study to Improve Healthcare Decision Making. *International Journal of Nursing Knowledge*. 2013; 24(1):15–24. [PubMed: 23413930]
- Alpaydin E. *Introduction to machine learning*. Cambridge, MA: MIT press; 2014.
- Anderson CA, Keenan G, Jones J. Using bibliometrics to support your selection of a nursing terminology set. *Computers Informatics Nursing*. 2009; 27(2):82–90.
- Bailey K. *Numerical Taxonomy and Cluster Analysis, Typologies and Taxonomies*. Sage Atlanta, Ga. 1994:34–65.

- Beyer K, Goldstein J, Ramakrishnan R, Shaft U. When is “nearest neighbor” meaningful?; Paper presented at the International Conference on Database Theory; 1999.
- Billings J, Zeitel L, Lukomnik J, Carey TS, Blank AE, Newman L. Impact of socioeconomic status on hospital use in New York City. *Health Affairs*. 1993; 12(1):162–173.
- Boyd A, Funk E, Schwartz S, Kaplan B, Keenan G. Top EHR challenges in light of the stimulus. Enabling effective interdisciplinary, intradisciplinary and cross-setting communication. *Journal of Healthcare Information Management*. 2009; 24(1):18–24.
- Bulechek GM, Butcher HK, Dochterman JMM, Wagner C. *Nursing interventions classification (NIC)*. St Louis: MO: Elsevier Health Sciences; 2013.
- Chen Y, Hu D, Zhang G. *Knowledge Enterprise: Intelligent Strategies in Product Design, Manufacturing, and Management*. USA: Springer; 2006. Data Mining and Critical Success Factors in Data Mining Projects; 281–287.
- Cox TF, Cox MA. *Multidimensional scaling*. Berlin, Heidelberg: Springer Berlin Heidelberg; 2008.
- Dochterman J, Bulechek G. *Nursing interventions classification (NIC)*. St. Louis, MO: 2004.
- Febretti A, Lopez KD, Stifter J, Johnson AE, Keenan GM, Wilkie DJ. Design, user experience, and usability Design philosophy, methods, and tools. Berlin Heidelberg: Springer; 2013. A component-based evaluation protocol for clinical decision support interfaces; 232–241.
- Frey BJ. *Graphical models for machine learning and digital communication*. Cambridge, MA: MIT press; 1998.
- Goodman DC, Fisher ES, Chang CH, Raymond S, Bronner K. After hospitalization: a Dartmouth atlas report on post-acute care for Medicare beneficiaries. Vol. 28. Hanover, NH: The Dartmouth Institute for Health Policy and Clinical Practice; Sep, 2011
- Herdman TH, Kamitsuru S. *NANDA International: Definitions and classification*. 10. Chichester/Ames, IA: Wiley-Blackwell; 2014.
- Hilbert M, López P. The world's technological capacity to store, communicate, and compute information. *Science*. 2011; 332(6025):60–65. [PubMed: 21310967]
- Hughes G. On the mean accuracy of statistical pattern recognizers. *IEEE Transactions on Information Theory*. 1968; 14(1):55–63. DOI: 10.1109/TIT.1968.1054102
- Hyvärinen A, Karhunen J, Oja E. *Independent component analysis*. Vol. 46. New York: John Wiley & Sons; 2004.
- Johnson J, Lodhi MK, Cheema U, Stifter J, Dunn-Lopez K, Yao Y, Wilkie DJ. Outcomes for End-of-Life Patients with Anticipatory Grieving: Insights from Practice with Standardized Nursing Terminologies within an Interoperable Internet-based Electronic Health Record. 2014 in review.
- Jolliffe I. *Principal component analysis*. New York: Verlage; 2002.
- Keenan GM. Big Data in Health Care: An Urgent Mandate to CHANGE Nursing EHRs! On-line *Journal of Nursing Informatics*. 2014; 18(1)
- Keenan G, Stocker J, Barkauskas V, Treder M, Heath C. Toward collecting a standardized nursing data set across the continuum: case of adult care nurse practitioner setting. *Outcomes Management*. 2002; 7(3):113–120.
- Keenan GM, Stocker JR, Geo-Thomas AT, Soparkar NR, Barkauskas VH, Lee JL. The HANDS project: studying and refining the automated collection of a cross-setting clinical data set. *Computers Informatics Nursing*. 2002; 20(3):89–100.
- Keenan GM, Tschannen D, Wesley ML. Standardized nursing terminologies can transform practice. *Journal of Nursing Administration*. 2008; 38(3):103–106. [PubMed: 18327054]
- Keenan GM, Wilkie DJ. Integration of NNN into EHRs: How are we doing? *International Journal of Nursing Knowledge*. 2014; 25(2):68. [PubMed: 24828765]
- Keenan G, Yakel E. Promoting safe nursing care by bringing visibility to the disciplinary aspects of interdisciplinary care. *AMIA Annual Symposium Proceedings*. 2005:385–389. [PubMed: 16779067]
- Keenan G, Yakel E, Lopez KD, Tschannen D, Ford YB. Challenges to nurses' efforts of retrieving, documenting, and communicating patient care information. *Journal of the American Medical Informatics Association*. 2013; 20(2):245–251. [PubMed: 22822042]

- Keenan G, Yakel E, Marriott D. HANDS: A revitalized technology supported care planning method to improve nursing handoffs. *Studies in Health Technology and Informatics*. 2005; 122:580–584.
- Keenan GM, Yakel E, Yao Y, Xu D, Szalacha L, Tschannen D, Wilkie DJ. Maintaining a Consistent Big Picture: Meaningful Use of a Web-based POC EHR System. *International Journal of Nursing Knowledge*. 2012; 23(3):119–133. [PubMed: 23043651]
- Liu H, Han J, Xin D, Shao Z. Mining Frequent Patterns from Very High Dimensional Data: A Top-Down Row Enumeration Approach; Paper presented at the 2006 SIAM International Conference on Data Mining (SDM'06); Bethesda, MD. 2006.
- Lodhi K, Ansari R, Yao Y, Keenan GM, Wilkie DJ, Khokhar A. Hierarchical Framework to Predict 30-day Re-hospitalizations from Nursing Care Data of Hospitalized Patients; Paper presented at the IEEE International Conference on Data Mining; Barcelona, Spain. 2016. in review
- Lodhi MK, Cheema UI, Stifter J, Wilkie DJ, Keenan GM, Yao Y, Khokhar AA. Death anxiety in hospitalized end-of-life patients as captured from a structured electronic health record: Differences by patient and nurse characteristics. *Research in Gerontological Nursing*. 2014; 7(5):224–234. [PubMed: 25157534]
- Lodhi MK, Stifter J, Yao Y, Ansari R, Keenan GM, Wilkie DJ, Khokhar AA. Predictive Modeling for End-of-Life Pain Outcome Using Electronic Health Records. In: Perner P, editor *Advances in Data Mining: Applications and Theoretical Aspects*. Vol. 9165. Hamburg, Germany: Springer International Publishing; 2015. 56–68.
- Mearian L. Scientists calculate total data stored to date: 295 + exabytes. 2011. Retrieved Feb 14, 2011 from <http://www.computerworld.com/article/2513110/data-center/scientists-calculate-total-data-stored-to-date--295--exabytes.html>
- Moorhead S. *Nursing Outcomes Classification (NOC), Measurement of Health Outcomes*. 5. St. Louis, MO: Elsevier Health Sciences; 2013.
- Pan F, Cong G, Tung A, Yang J, Zaki MJ. CARPENTER: Finding closed patterns in long biological datasets; Paper presented at the International Conference on Knowledge Discovery and Data Mining; Washington, D.C. 2003.
- Piatetsky-Shapiro G. *Knowledge Discovery in Databases*. Cambridge, MA: MIT Press; 1991. Discovery, analysis, and presentation of strong rules; 229–238.
- Schneider JS, Barkauskas V, Keenan G. Evaluating home health care nursing outcomes with OASIS and NOC. *Journal of Nursing Scholarship*. 2008; 40(1):76–82. [PubMed: 18302595]
- Smith PC. *Performance measurement for health system improvement: experiences, challenges and prospects*. Cambridge: Cambridge University Press; 2009.
- Sousa VE, Lopez KD, Febretti A, Stifter J, Yao Y, Johnson A, Keenan GM. Use of Simulation to Study Nurses' Acceptance and Nonacceptance of Clinical Decision Support Suggestions. *Computers Informatics Nursing*. 2015; 33(10):465–472.
- Stifter J, Yao Y, Lodhi MK, Lopez KD, Khokhar A, Wilkie DJ, Keenan GM. Nurse Continuity and Hospital-Acquired Pressure Ulcers: A Comparative Analysis Using an Electronic Health Record" Big Data" Set. *Nursing Research*. 2014; 64(5):361–371.
- Stifter J, Yao Y, Lopez KD, Khokhar A, Wilkie DJ, Keenan GM. Proposing a New Conceptual Model and an Exemplar Measure Using Health Information: Technology to Examine the Impact of Relational Nurse Continuity on Hospital-Acquired Pressure Ulcers. *Advances in Nursing Science*. 2015; 38(3):241–251. [PubMed: 26244480]
- Tastan S, Linch GC, Keenan GM, Stifter J, McKinney D, Fahey L, Wilkie DJ. Evidence for the existing American Nurses Association-recognized standardized nursing terminologies: A systematic review. *International Journal of Nursing Studies*. 2014; 51(8):1160–1170. [PubMed: 24412062]
- Tschannen D, Keenan G, Aebbersold M, Kocan MJ, Lundy F, Averhart V. Implications of nurse-physician relations: Report of a successful intervention. *Nursing Economics*. 2011; 29(3):127. [PubMed: 21736176]
- Yam C, Wong E, Chan F, Wong F, Leung M, Yeoh E. Measuring and preventing potentially avoidable hospital readmissions: a review of the literature. *Hong Kong Medical Journal*. 2010; 16(5):383–389. [PubMed: 20890004]

- Yao Y, Stifter J, Ezenwa MO, Lodhi M, Khokhar A, Ansari R, Wilkie DJ. Infomarkers for transition to goals consistent with palliative care in dying patients. *Palliative and Supportive Care*. 2015; 13(05):1427–1434. [PubMed: 25711431]
- Zao Z, Albright R, Cox J. Processing and Storing Sparse Data in SAS® Using SAS® Text Miner Procedures. Paper presented at the SAS Institute Inc; Washington, D.C: 2014.

Figure 1. Depiction of Nursing Plan of Care Data Life Cycle (Ecosystem), including data extraction, preprocessing, mining and analysis phases. Copyright 2011 HANDS Research Team, reprinted with permission.

Figure 2. Knowledge Discovery Framework. This figure depicts the different steps involved in the knowledge discovery and analysis processes. Copyright 2015 HANDS Research Team, reprinted with permission.

Figure 3.
Data Structure of an Episode in HANDS.

Figure 4.
HANDS Data Structure View 2.

-
- ✓ **Impaired Gas Exchange**
 - ❖ **Mechanical Ventilation Response: Adult (Goal=5; Current=3)**
 - Sedation Management
 - Aspiration Precautions
 - Artificial Airway Management
 - Infection Control
 - ❖ **Respiratory Status: Gas Exchange (Goal=5; Current=2)**
 - Oxygen Therapy
 - Chest Physiotherapy
 - Acid-Base Management
 - ✓ **Impaired Physical Mobility**
 - ❖ **Immobility Consequences: Physiological (Goal=5; Current=3)**
 - Embolus Precautions
 - Bed Rest Care
 - ✓ **Risk For Injury**
 - ❖ **Personal Safety Behavior (Goal=5; Current=2)**
 - Physical Restraint
 - Surveillance: Safety

Figure 5.
Example of Part of Admission Plan of Care.

Figure 6.

Patient-oriented Results: Comfortable Death Outcome Being Met/Not Met by Hospital Unit Type.

Figure 7. Care Provider-oriented Results: Fall Prevention Behavior Outcome for Patients Suffering from Pain and Cardiopulmonary Problems

Figure 8.
Organization-oriented Results: Readmission Decision Tree for Young Patients

Table 1

Data Mining Techniques and Examples of Models.

Data Mining Techniques	Models
Association Mining	Association rules A priori algorithm
Classification	Decision Trees SVM Naïve-Bayes K-NN
Clustering	K-means K-median clustering
Regression	Linear Regression Polynomial Regression Logistic Regression