The University of Iowa, College of Nursing, Center for Nursing Classification & Clinical Effectiveness

NURSING

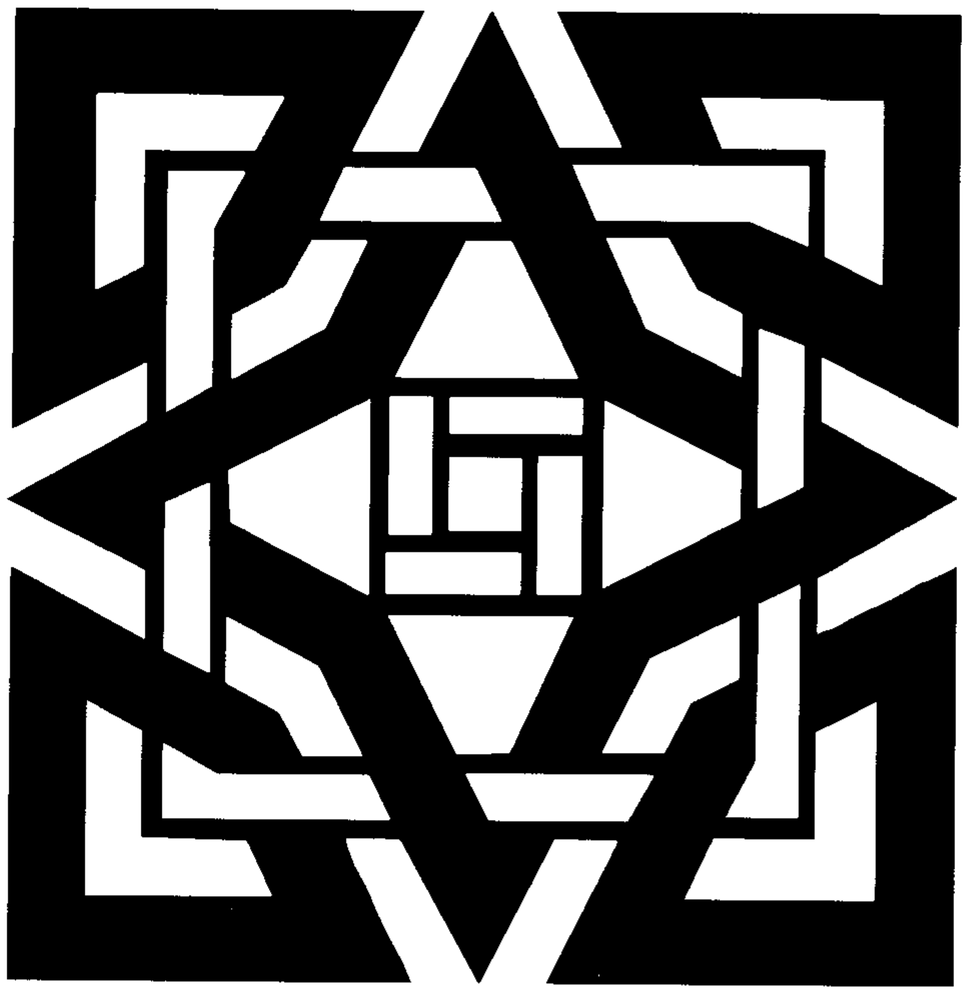
GUIDELINE FOR CONDUCTING EFFECTIVENESS RESEARCH IN NURSING & OTHER HEALTHCARE SERVICES

Authors

Marita Titler, PhD, RN

Joanne Dochterman, PhD, RN

David Reed, PhD



EFFECTIVENESS RESEARCH IN NURSING

Guideline for Conducting Effectiveness Research in Nursing and Other Healthcare Services

August, 2004

Authors

This guideline was developed as part of a funded NINR/AHRQ research study ("Nursing Interventions and Outcomes in 3 Older Populations." RO1NR05331, M. Titler, PI, 2001-2005) that required the design and use of a large electronic data repository that included standardized nursing data.

Authors of the guideline are Marita Titler, PhD, RN, Joanne Dochterman, PhD, RN, and David Reed, PhD. Other members of the research team at the time that this was written included: Paul Abramowitz, Pharm D, Ginette Budreau, MA, MBA, RN, Gloria Bulechek, PhD, RN, Connie Delaney, PhD, RN, Linda Everett, PhD, RN, Mary Kanak, MA, RN, Vicki Kraus, PhD, RN, Mary Mathew Wilson, MA, Sue Moorhead, PhD, RN, Debra Pettit PhD, RN, and Jenny Wang, PhD.

Guideline for Conducting Effectiveness Research in Nursing and Other Healthcare Services

Abstract

The purpose of this guideline is to assist nurse researchers and other investigators to construct and use clinical data for effectiveness research. While nursing care and other health care services are more frequently documented using a computer, electronic documentation of clinical care is generally not used for research to improve care. This guideline identifies the issues and processes of using electronic data repositories of clinical and administrative data to answer important research questions.

The guideline is directed towards health care providers, administrators, and researchers who are

1) implementing standardized nursing and other provider languages in a documentation system;
2) evaluating clinical information systems with an eye toward purchase; or
3) involved in constructing or using a clinical documentation system for effectiveness research.

While this guideline may be of most interest to nurses, the general principles apply to those in any health discipline, and all researchers and administrators desiring to analyze clinical data for best care practices. It is best to think about the use of data for effectiveness research while planning for implementation of a clinical information system to facilitate getting the data output in useful form. However, if this has not been done, the guideline will still be helpful to assist in transferring the clinical data into a useful research format.

Although the guideline is primarily directed toward the use of clinical data in acute care agencies as that is where computerized databases most often exist and where we have the most experience, it is also applicable to other settings such as home care and long-term care. For use in other than acute care settings, the reader should apply the information here as it fits the particular setting and the available electronic repositories.

Table of Contents

Guideline for Conducting Effectiveness Research in Nursing and Other Healthcare Services

Purpose of the Guideline

While there is a great desire among nurses and other health care providers to base care on sound research findings (often referred to as evidence-based practice), as well as a growing trend to document care delivered with computer information systems, there has been little use of the resulting electronic data to improve nursing care. There are many reasons for this, including: a focus on putting the data in rather than getting the data out, lack of agreement on how to document specific variables, lack of resources, and lack of knowledge about how to construct electronic clinical data repositories that can be used for both a clinical record and a research database. **The purpose of this guideline is to assist others in using electronic data repositories for outcomes research**.

The guideline will help those who are using the repositories to understand some of the issues they will encounter and strategies that might be effective. It will also benefit those who are building clinical databases to construct repositories that will be useful for research. The guideline is based on our past experience in classification and evidence-based practice research. It results most directly from our experiences during a funded NINR/AHRQ research study (Titler, 2001) that required the use of several large electronic data repositories that included standardized nursing data.

Effectiveness Research

The term effectiveness research is used to indicate the study of the effect of provider interventions on patient outcomes. Effectiveness indicates the benefits of health care that are actually achieved under ordinary circumstances for typical patients (Lohr, 1988; Muenning, 2002). Health services/interventions are considered effective to the extent they achieve health improvements in real practice settings (Mandelblatt, Fryback, Weinstein, Russell, & Gold, 1997). Patient outcomes are the results of health care interventions as experienced by the recipient of the intervention (e.g. better functional ability or a nosocomial infection) or broader measures related to the impact of the intervention (e.g. length of hospital stay or cost of care). Outcomes describe behaviors, responses, or feelings of the patient in response to care provided. Many variables influence outcomes including the health care providers, the treatments prescribed, the environment, the patient, and the patient's significant others.

Effectiveness research in nursing can facilitate better clinical decision-making and better use of scarce resources. Systematic documentation of interventions used by nurses allows us to study and compare the use of particular interventions by type of unit and facility. Determining the interventions used most frequently in a specific unit or agency will help determine the content of the unit's nursing information system; assist in formulation of the unit's skill mix; and provide direction for staff continuing education programming. Knowing which interventions work best for specific diagnoses and result in certain outcomes can be used to assist nurses to make better clinical decisions. When information is systematically collected

about the treatments nurses perform, clusters of interventions that typically occur together for certain types of patients can be identified. There is a need to identify interventions that are frequently used together for certain types of patients so that they can be studied to determine their interactive effects.

Documentation of nursing practice through the use of standardized language (see section on standardized nursing language) creates many exciting possibilities for nursing in effectiveness research. The identification of research questions to be addressed is referred to by two health policy authors at Harvard Medical School as the "effectiveness space." (Guadagnoli & McNeil, 1994). That is, nurses and other providers must identify the variables (e.g., interventions, outcomes, specific patient characteristics, specific provider characteristics, specific treatment setting characteristics) and their measures necessary to evaluate the effectiveness of their care. This guideline will assist in the identification of these variables.

Identification of Research Questions

In the early stages of planning for implementation of a clinical information system, one should identify key research questions that can be addressed with the data collected through documentation. After the research questions are identified, the variables needed to address the questions can be determined. Then, for each variable, one needs to ascertain if the data are currently collected (say, in other places in the institution's databases) or should be collected in the new system. These data must be linked with each other at the individual patient and specific encounter level. Addressing these concerns when designing a nursing information system will enhance the ability to use these data in effectiveness research.

As implementation of a clinical documentation system may involve several phases and could be a lengthy process, it is advisable to prioritize the research questions; that is, to identify those questions that can be answered early on, say at the end of 3 years time, as well as those that need more time, say, 5 or 7 years of implementation.

Examples of research questions related to nursing care are the following:

1. What interventions are used most frequently for a given patient population? On a certain type of unit?
2. What types of nursing personnel typically use which interventions?
3. Does the type and amount of staff in a facility relate to the achievement of specific patient outcomes?
4. What are the related diagnoses and outcomes for particular interventions?
5. Do nursing interventions reduce the patient's length of stay? Cost of care?
6. Do nursing interventions reduce patient complications, such as urinary tract infection, wound infection, falls?
7. What interventions typically occur together?
8. Does the effectiveness of specific nursing interventions depend on the type of medical procedure or the medication administered?

In following sections we discuss database variables that will assist researchers in determining answers to these and other questions. Each variable that is to be included should have both a definition and an operational measure. For example, in question 1 in the above list, one needs to define and measure interventions: are these nursing, medical, pharmacy, other? What measures will be used for each? In the same question one needs a definition and measurement of the patient population and unit. Consistent definition and measurement are necessary to aggregate and compare data from different units in different settings.

Electronic Data Repositories

For a long time, health care agencies did not collect or store patient information electronically. This is changing and most institutions have some form of clinical electronic data repositories. The information on a specific individual is identified using key indicators such as a unique number, admit/discharge dates and department(s), service(s) and unit(s) associated with the patient's stay. Because electronic data are used for internal decision-making (e.g. pay; staffing, unit census, patient volume) and audited, motivation for accuracy of the data by the users is high. Furthermore, accuracy of electronic data is monitored by personnel in the agencies that collect the data. Users of the electronic systems receive training on the purpose and use. Legal and professional accountability for documenting care delivered is emphasized. Electronic data entry systems are often programmed to alert users that an error has been made (e.g. out of range value), requiring reentry of a correct value. Checks are done to validate key data with the primary data source, for example, 10% of the records in the incident report database are compared with the manual copy of the incident reports. It is important to be aware of the checks that ensure accuracy in the identified system.

Several data repositories may exist in one agency; these may or may not be linked with each other. It is important to identify the various repositories that contain the variables of interest, the time frame that data have been collected for each variable, and the number and type of units/clinics/other places where the data have been collected. A description of the repository should be available; two examples follow:

Pharmacy Repository

Pharmaceutical data for inpatients is captured via PharmNet and stored in a non-relational database that has been functioning since June 23, 1997 and resides as a mainframe data file. All data related to medication orders can be retrieved following patient discharge. Data elements in the database are type of medication, start date, stop date, dose, route of administration, interval of administration, duration of therapy, and total doses dispensed. Patient demographic data including height, weight, lab data, allergies, and adverse reactions are also retrievable.

Medical Record Abstract Repository

The Medical Record Abstract, available on every hospitalized patient, is overseen by the Health Information Management (HIM) Department. Coding is performed concurrently, using ICD-9-CM coding manuals, by trained medical record technicians who review the medical records on the inpatient

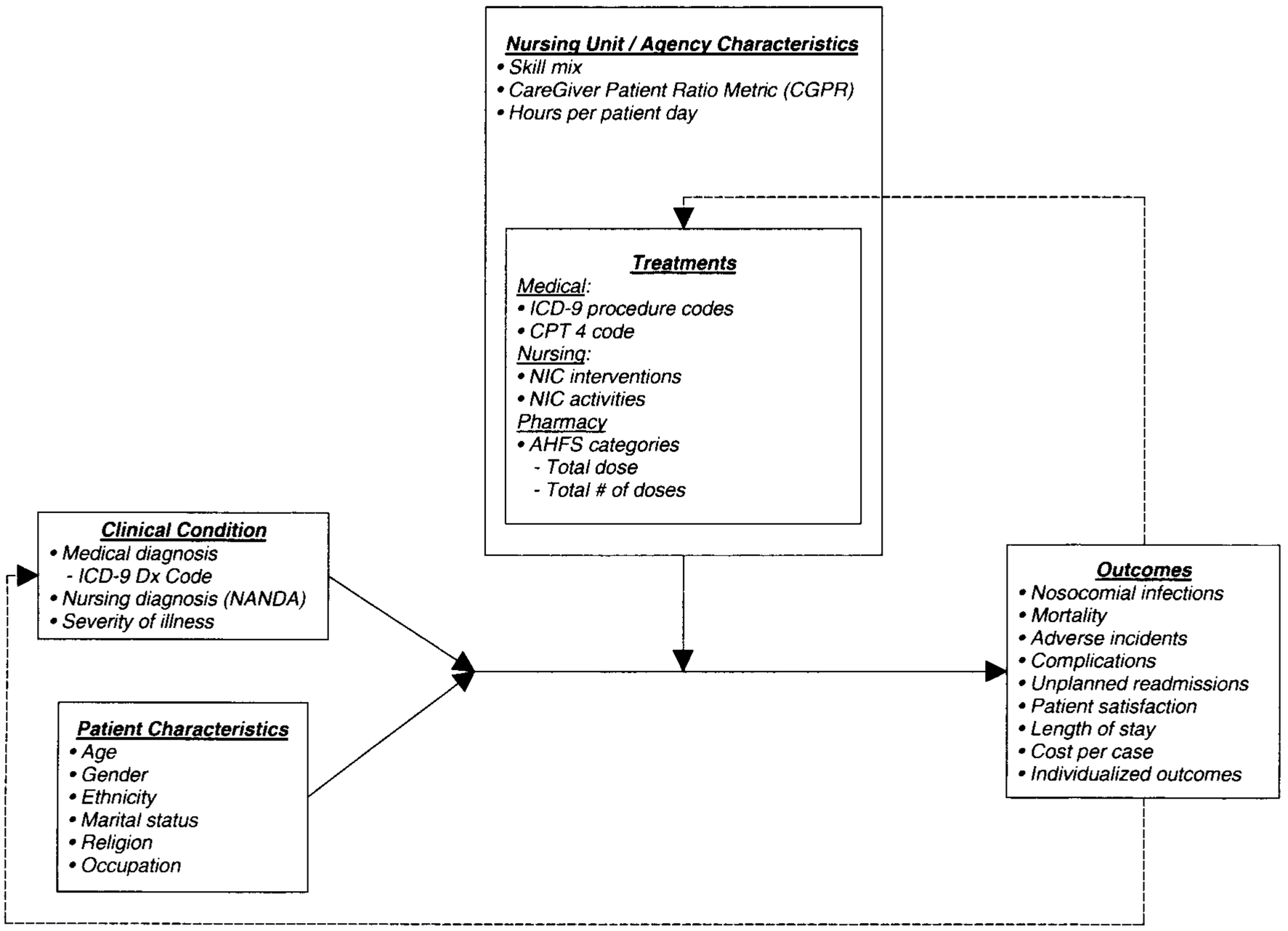
units. A medical record abstract is completed for each inpatient encounter and data are reported as required by state and federal guidelines. Data elements in the medical record abstract include patient hospital number, inpatient admit and discharge dates; demographic information (e.g. date of birth, residency, ethnicity); admission data (e.g. source such as emergency room); discharge data; mortality/expiration and type (e.g. within 48 hours of surgery, within 48 hours of admission); admission/discharge physician; type of visit (e.g. acute, inpatient, emergency); admission source (e.g. patient home, other acute care hospital); discharge disposition; length of stay (hospital; each patient care unit) principle diagnosis type by ICD-9 codes, and date of onset; procedures (type and date) by ICD-9 codes, physician code and department; and DRG assignment.

Factors to Consider in Selecting Database Variables for Research Analysis

After reviewing the research questions, the investigators determine the central variables of interest. In the case of nursing, these are ones that describe clinical nursing care (i.e. nursing interventions), as well as the dependent variables (i.e. patient outcomes) in which one hopes to affect changes, as well as any control variables (i.e. data about the patient, other providers and their actions, or the setting) that might interact with this process. Variable definitions and measurement indicators ideally should be consistent with those of the standards of the field; for example, the Uniform Hospital Discharge Data Set (UHDDS) (Pearce, 1988), the Nursing Minimum Data Set (NMDS) (Werley & Lang, 1988), and the Nursing Management Minimum Data Set (Delaney & Huber, 1996).

A conceptual model is useful to guide variable selection for statistical analysis. Figure 1 depicts a general model that can guide effectiveness research. Patient outcomes of acute care services are influenced by characteristics of the patient (demographics), their clinical conditions (nursing diagnoses, medical diagnoses, severity of illness), the treatments they receive (medical, nursing, and pharmacological) and the context in which care is delivered (unit or agency characteristics) (Kane, 1997; Titler, 2001). The characteristics of the patient and nursing environment (units) influence the delivery of medical, nursing, and pharmacological treatments thereby impacting patient outcomes and thus, can be used as control variables in nursing studies examining the impact of nursing interventions. Patient outcomes, in turn, affect the characteristics of the patient, and subsequent treatments as depicted by the feedback loops. The variables for each of the categories described in the model are described in the section entitled Data Definitions.

Figure 1: Model for Effectiveness Research



Standardized Nursing Language

The core variables for nursing effectiveness research are the collection and coding of nursing data in standardized format. To this end, we briefly overview three standardized languages that can be used to document nursing diagnoses, interventions, and outcomes: NANDA, NIC and NOC. Multiple facilities are using these languages to plan and document nursing care. This guideline will assist providers and researchers in these and other institutions to provide meaningful data that can assist in improving the quality of nursing care. Other standardized nursing languages can also be used but these three are comprehensive and have ongoing processes to keep them current.

NANDA: The use of standardized language in nursing began with the development of the NANDA classification in the 1970s. NANDA has been translated into multiple languages and is used in more than 20 countries throughout the world. The 2003 edition of the classification (NANDA International, 2003) includes 167 diagnoses (e.g. Activity Intolerance, Impaired Verbal Communication, Wandering) (See Appendix One for an example of a complete NANDA diagnosis). A nursing diagnosis is defined as a "*clinical judgment about individual, family, or community responses to actual or potential health problems/life processes. A nursing diagnosis provides the basis for selection of nursing interventions to achieve outcomes for which the nurse is accountable*" (NANDA International, 2003, p. 263). The NANDA classification is maintained by NANDA International, a membership organization with members in multiple countries (www.nanda.org).

NIC: The Nursing Interventions Classification (NIC) (Dochterman & Bulechek, 2004) names and describes interventions that nurses perform. An intervention is defined as *"any treatment, based upon clinical judgment and knowledge that a nurse performs to enhance patient/client outcomes"* (p. xxiii). Each of the 514 interventions in the fourth edition of the classification is composed of a naming label, a definition, and a list of activities that describe what a nurse does to carry out the intervention. (See Appendix One for an example of a complete NIC intervention.) The classification includes all treatments that nurses perform, from the most basic (e.g., Body Mechanics Promotion) to those that are highly complex and specialized (e.g., Anesthesia Administration). There are interventions for illness treatment, injury prevention, and health promotion. Interventions are included for individuals, families, and communities; indirect care interventions are also included in the classification. While the entire classification describes the domain of nursing, some of the interventions in the classification are also done by other providers. Work between editions and other relevant publications that enhance the use of the classification are available from the Center for Nursing Classification and Clinical Effectiveness at the University of Iowa, Iowa City, http://www.nursing.uiowa.edu/cnc).

NOC: The Nursing Outcomes Classification (NOC) (Moorhead, Johnson, & Maas, 2004) is a companion language to the NIC interventions and the NANDA diagnoses and allows for measurement of the effectiveness of nursing interventions. An outcome is defined as: "*An individual, family, or community state, behavior or perception that is measured along a continuum in response to a nursing intervention(s). Each outcome has an associated group of indicators that are used to determine patient status in relation to the outcome. In order to be measured, the outcome requires identification of a series of more specific indicators.*" (Moorhead, Johnson, & Maas, 2004, p. xix) The third edition of the Nursing Outcomes Classification contains 330 outcomes (Moorhead, Johnson, & Maas, 2004). (See Appendix One for an example of a complete NOC outcome.) The outcomes assist nurses and other health care providers to evaluate and quantify the status of the patient, caregiver, family or community outcome. Degrees of variation on achievement of the outcome are measured on a five point Likert scale where 1=the least desirable outcome and 5= the most desirable outcome. Change in rating scores can be determined for outcomes over time. NOC, like NIC is housed in the Center for Nursing Classification and Clinical Effectiveness at the University of Iowa College of Nursing. (For more information see http://www.uiowa.nursing.edu/cnc.)

Data Definitions

Each of the sections in Figure 1 is defined here (also refer to variable definitions in Appendix Two). In your agency, you will need to identify who gathers these data and where the data are stored. (In the table in Appendix Two there is a column to put this information in.)

Clinical Condition of Patients. A clinical condition is defined by the medical and nursing diagnoses and severity of illness. The medical diagnoses are typically coded in the medical record abstract, using the ICD-9 diagnostic codes to identify the principle and secondary medical diagnoses of each patient. NANDA is used in many institutions to document the nursing diagnoses. Severity of illness is defined as the extent of physiological decompensation or organ system loss of function. For example, as assigned by the APR-DRG

system 1=minor, 2=moderate, 3=major, 4=extreme, or as measured in a particular facility. These data are usually available from the Medical Record Abstract database; severity of illness can be used as a control variable in statistical analyses.

Patient Characteristics. Patient characteristics (also called demographics) are usually available from a census system and include gender, date of birth, admission/discharge date, marital status, religion, and occupation. These data are usually collected in a standardized format, but the specific format may vary by agency and may need to be recoded for research purposes (see next section).

Unit/Agency Characteristics. Unit or agency characteristics include information about the unit the patient is on, the nursing staff on that unit, and the agency. Figure 1 gives three examples of this data: skill mix of the staff (meaning the percentage of RNs to assistants on the unit), the caregiver-patient ratio (the average number of patients to a caregiver), and the nursing hours per patient day. If the information system does not collect standard data on some variables (such as caregiver-patient ratio), a decision to collect data for the variable needs to be addressed. Would this information be helpful to have in order to address the impact of other variables on specific outcomes? If so, should the agency begin to collect the information? When and in what form? Characteristics of patients and nursing units are typically used as control variables in statistical analyses. (In a non-hospital setting the unit of care may be the whole agency or setting.) See Appendix Three for one example on how to group data about types of nursing units.

The coding of demographic variables by hospitals and other health care agencies tends to be idiosyncratic, resulting in a several problematic aspects for research purposes. Codes may reflect categories unique to the locality, codes may vary greatly in the level of abstraction they represent, and codes may not have any relevance to health status. Furthermore, in a particular locality, a large proportion of persons may fall into a single category (e.g. higher than expected number of older patients that come for treatment or a higher percentage of nursing assistants due to agency location). Therefore, judgment must be used in recoding variables to produce categories that have substantial numbers in them and that, so far as possible, are categories that are meaningful in relation to the research purpose.

Treatments. The treatments that patients receive during the acute episode of care are categorized in Figure 1 as medical treatments, nursing interventions, and pharmacological treatments.

Medical treatments are defined as any medical procedure patients received during the episode of care, coded as ICD-9 (Ingenix, 2004)) and CPT-4 procedure codes (AMA, 1999) available from the medical record abstract and/or financial management databases respectively. The date and time that the treatment was delivered also needs to be captured.

Nursing interventions can be documented by using NIC; in addition, the nursing activities for each nursing intervention can also be collected. It should be noted that the collection of both NIC labels and activities results in a very large database. Initially a facility may want to only capture and study the effect of the overall intervention (at the NIC label level). The date and time that the intervention was delivered also needs to be captured in order to determine the frequency of use for each intervention.

Pharmacological treatments may be documented in a separate database (e.g. PharmNet medication management system from Cerner Corporation) and include medications ordered (start/stop dates, name, route, and dose), medications administered (name, route, dose, frequency), and medication allergies for each patient. The American Hospital Formulary Service Categories (AHFSC) (American Hospital Formulary Service, 2000) can be used to categorize the medications administered.

Patient Outcomes. Outcomes that are frequently collected in health care institutions are grouped for definitional purposes in Figure 1 as nosocomial infections, mortality, adverse incidents, complications, unplanned readmissions, patient satisfaction, length of stay, and cost per case. (An individual agency may group or define these somewhat differently.) Within each group, outcomes can be designed generic and/or condition specific. Generic outcomes assess the overall effects of services on general health and across patient conditions (e.g. nosocomial infections) (Maciejewski, 1997). Condition specific outcomes assess specific diagnostic groups or populations of patients with the same condition, and tap the domains of greatest interest for a particular condition (e.g. MI for patients with heart failure) (Atherly, 1997). Use of both generic and condition specific outcomes are recommended for effectiveness research (Atherly, 1997).

Nosocomial infections are infections acquired during a hospitalization that were not present at the time of admission, in our institution defined as 48 hours or more after admission. The Center for Disease Control (CDC) criteria are used to classify an infection as nosocomial and to classify each type of nosocomial infection (Garner, Jarvis, Emori, Horan, & Hughes, 1988; University of Iowa Hospitals & Clinics, 2003). The total nosocomial infection rate and nosocomial infection rates for urinary catheter associated urinary tract infections (UTIs), pneumonia, surgical wounds, and intravascular site infections can be used as generic outcome variables across some populations.

Mortality, a traditionally used generic outcome, (Atherly, 1997) is defined as death during an acute episode of care.

Adverse incidents are undesired outcomes such as falls and medication errors. Falls and medication error data are found typically in the incident report system. Complexity of medication regimens increases the risk for medication errors in some patients.

Complications are defined as onset of additional diseases or conditions associated with treatment of a condition during hospitalization, e.g. cerebral vascular accident (CVA), deep venous thrombosis, myocardial infarction (MI), pneumothorax, pulmonary embolus (PE), and tissue/organ injury. Complications are, by definition, condition specific as they are untoward outcomes associated with the treatment of a condition.

Unplanned readmissions are defined as unplanned admissions within a specific number of days (e.g. 10 days) after discharge.

Satisfaction with care can be used as a generic outcome for most patient populations as measured by willingness to recommend the hospital to others, and overall satisfaction with services. These two areas have been shown to be robust measures of patient satisfaction with healthcare (ANA, 1995; Davies & Ware, 1991)).

Length of stay and cost per episode of care are outcomes that are available in all institutions and can usually be acquired from the medical record abstract database and financial database respectively.

Figure 1 also lists individualized outcomes or outcomes selected to determine the effect of particular interventions for each patient. These may be the same or different from outcomes in a previous category, but are measured at the individual level. Ideally, the Nursing Outcomes Classification (NOC) is used to measure the level of specific individual outcomes achieved at specified times. Please see the earlier section on standardized nursing language. The data should include the name of the outcome, the times rated and the ratings themselves. When NOC is used to document patient outcomes one has a rich condition and intervention specific database.

Potential Data Sources

Once the research questions and desired variables have been identified the next step is to evaluate the availability and usefulness of data sources. If the study is being conducted within a single health care organization or a small group of related organizations, the logical starting point for data sources is within the structure of the organization(s). The ever-increasing integration of computer technology into both the operational and clinical sides of health care organizations has resulted in an abundant supply of electronic data within such organizations. Usually multiple electronic databases exist and pertain to very specific purposes (e.g. census, finance/billing) or the work of specific disciplines or specialties (e.g. nursing, physical therapy, surgery department, risk management, epidemiology) within the organization. These individual databases have often been developed and maintained with the assistance of the organization's Information Systems department. Consequently, in the search for potential data sources, a good starting point is to speak with the Chief Information Officer (CIO) of the organization to determine what electronic databases exist within the organization. (Another good source is the Chief Quality Officer if the organization has this person.) The CIO can then direct the investigator to the various database administrators for discussion of availability of the database for research purposes as well as the match between existing variables within the database and the research questions/variables.

The following are examples of departments or organizational units within a facility that possess databases that might be tapped for effectiveness research:

- The Health Information Management or Systems Department often maintains the Medical Record Abstract (MRA) database and/or census database. As discussed previously, these databases provide information about patient characteristics (demographic information), primary and secondary medical diagnoses (DRGs, ICD 9-CM diagnostic codes), severity of illness (APR-DRG), medical treatments (ICD 9-CM procedure codes, CPT codes), and details about each patient stay/visit (name of nursing unit where care was provided, length of stay, discharge disposition, etc.).

- The Department of Nursing usually maintains the Nursing Information System (NIS) database as well as information about nursing staffing. The NIS provides information about the nursing diagnoses, interventions, and outcomes that pertained to the care of the patient sample. The staffing database provides information on the overall hours of nursing care provided as well as a breakdown of hours of care by type of nursing staff (RN, LPN, unlicensed assistive nursing personnel).

- The Pharmacy Department maintains databases that provide information on the medications, dosages and times of administration for the patient sample being studied.

- Some facilities have an Office of Quality Management that may collect and maintain an incident report database as well as a patient satisfaction database. These two databases provide valuable information about outcomes such as adverse incidents and the perceptions of the patient and his/her significant others about the quality of health care provided during a hospital episode of care.

- The Program of Hospital Epidemiology maintains an infection control database that provides data about nosocomial infections.

- The Financial Management Department keeps a financial database that provides information about various kinds of charges related to supplies, treatments and physician time, and room cost (nursing care is typically included in this category).

It is possible that a variable desired by the investigator may not be present in the available databases or present in the form desired by the investigator. In some cases it may be possible to use existing variables in the databases to construct a derived variable. For instance, length of stay may not be a variable in the databases but can easily be calculated from the admitting and discharge dates.

Methods of Requesting Data

Defining the sample. The repository of data in a clinical information system is very large, in fact overwhelming. In order to reduce the volume to a manageable size and to make sense of the data one needs to define a particular sample on which to collect the identified variables. There are various ways to identify a sample: unit or location specific, patient population specific, provider specific, time specific, or a combination. Each approach has its advantages and disadvantages and some may not be possible in a given facility due to limitations of the clinical information system. For example, it is not possible to compare data between two units if one unit is not computerized.

Usually, the investigator is interested in a specific patient population during an identified period of time. How the population is identified determines how the data will be accessed. One common way is to use medical Diagnosis Related Groups (DRGs) categories. Accessing patient data by DRG group is common in outcomes effectiveness research as these groupings of diseases are designed to be similar in cost (Muenning, 2002). The use of DRGs in all US hospitals facilitates the comparison of nursing care within each DRG across institutions. This is not as easy as it sounds, however. For example, in a recent study we wanted to access

patients who had been admitted for a hip fracture; we thought we could pull records by DRG 209 defined as major joint/limb reattachment procedure of the lower extremity. We found however, that in addition to DRG 209 we also needed DRG 210 (hip and femur procedure excluding a major joint>17 with complications), 211 (hip and femur procedure excluding a major joint> 17 without complications), 236 (fracture of the hip and pelvis), or 471 (bilateral or multiple major joint procedures) and/or one of 7 specified ICD-9-CM primary or secondary diagnosis codes (e.g. 820 fracture of neck of femur) or one of 6 specified procedure codes (e.g. 8151 total hip replacement). In selecting the sample, the ICD-9 procedure codes were reviewed and those with in-hospital fractures, chronic osteomyletis of the pelvis, pyogenetic arthritis, and malignant neoplasm of the pelvis were first removed. As you can see from the example, selecting the sample by DRG category involves careful review of multiple related codes.

But other ways also are possible and have advantages, for example accessing data by patient age or by specific nursing diagnoses or interventions. Often a combination of these is used. For example, in a recent study, data were requested on adults greater than 60 years of age who had been admitted to the facility from January 1, 1998 through December 31, 2001, who received care on an inpatient non-critical care unit, and who were in one of two DRG groups. In the same study we requested data on all patients who were at risk for falls (determined by a specific score on a risk assessment scale required on all those admitted) or on those who had received the NIC intervention of Fall Prevention for the same time period and on the same units. This is more of a nursing approach and results in a more heterogenous medical diagnosis group of patients than using only medical diagnoses or DRG categories. Defining a specific sample takes some thought and often requires choice or modification. The user will need to work with the information system's representatives to determine what is available, taking into consideration the desired sample as well as the desired variables in the defined sample.

Methods of acquiring data. Once appropriate data repositories have been selected and the data elements needed to construct individual research variables identified, and the desired sample identified, a formal written request for acquiring electronic data should be routed to the individual in charge of each repository. As noted in the next section on Helpful Hints (see Hint Eleven), obtaining test files first is a useful strategy for developing and refining database structures and testing relationships prior to receiving the "real" data. When acquiring electronic data files—whether test data or the entire set of "real" data—the request should be made in terms of the specific data elements that are to be retrieved from the repository and that correspond to particular patients (stated as the patient identification number) during particular visits (stated as visit number, admit date and discharge date).

It is a good idea, to assure consistency from the very start, to specify in writing how data from each original source should be transmitted to the research team. This includes specification of a preferred spreadsheet program for transmission of the data and detailing how various types of data fields should be formatted (such as dates, patient ID numbers, address information, rounding of numerical data, financial data, and so forth). Consistency in format of data received from various sources will be of obvious benefit when programming begins. It should also be noted, however, that the investigator(s) might have to be content to receive the data in the form that it is maintained and do the transformation of format at the researcher's end (See

Appendix Four: Example of data format request). Any passing back and forth of confidential patient information linked to patient ID numbers, either in electronic file form or on paper, should involve hand carrying of the data. Confidential information in electronic form should never be relayed via email attachments or via disks put in interdepartmental or U.S. mail.

Depending on the number of data repositories to be accessed, a tracking mechanism can be an essential piece of the project management strategy. This can be as simple as a multi-column chart showing the type of data requested, date requested, date received, and whether, when checked, the data received matches the data request. Other notes can also be added. The tracking chart should be updated as requests are made and data is received and checked. (See Appendix Five: Example of data tracking form).

Helpful Hints to Data Access

Access to clinical data for research purposes is not a given, especially with increased concern about privacy and confidentiality. In this section we include "hints" to make the process of data access easier for all involved. Depending on your system or agency, not all of these hints may be relevant but for those that are, access to the data may depend on your careful attention.

Hint One: *Involve an influential insider as part of the research team*
Although effectiveness studies may be conducted by persons employed by the agency where the data reside, some projects/studies will be initiated by those who are not employed at the agency. It is much easier to gain access to data if a key person on the team is someone who has some administrative clout in the agency and who knows the persons to contact and the system. If this person also is familiar with the data elements in the data repository, all the better.

Hint Two: *Gather the needed background information*
The first step for using clinical data is to know something about it. At an early stage of the process you should be able to answer the following questions:

- Where are the data located in the institution?
- What variables are collected?
- Who has charge of the data? Who has access?
- Is there one data set or many?
- How long are the data saved/ can be retrieved?

Hint Three: *Gain top administrative support*
In order to access data, you will need to show that you have support from key administrative people. The more controversial the research, the more important it is to have support from top administrators (e.g. the chief nursing officer). This will likely mean talking with the appropriate individuals and convincing them that they should lend support.

Hint Four: *Have someone in charge for a coordinated effort*
If the project/research study has a Principal Investigator/designated leader this may not be an issue. This person may or may not be the influential insider. The point is that the agency and people who are being asked to participate need to know the contact person to go to if there is

an issue. This could be and often is someone other than the person to whom the data are being sent. The rest of the team need to keep this person informed and the person must be one whom others trust and can easily communicate.

Hint Five: *Don't waste people's time-- Be able to clearly articulate your project and its importance*
The more important the research questions, the more likely that you will get access to the data. Importance should address both how the research can promote the knowledge base of the field but also address how it can assist the institution in its decision-making and resource allocation. For example, knowing the most frequently used nursing interventions for specific patient populations can help the institution plan for the level of nurse staffing and competency evaluation. Importance is necessary but not sufficient; persons requesting data also have to demonstrate that they are competent to conduct the research, will treat the data with the utmost care and confidentiality, and produce some results in a reasonable time.

Hint Six: *Figure out where you will store the data, what equipment/software you need to access it and what programming resources you will need*
In most clinical facilities the clinical data that is being requested will need to be copied and "moved" somewhere so the researchers can manipulate and transform it without changing the original patient data. As the number of data records of most variables is very large (because nursing assessments and care delivered are often documented on a frequent basis) the storage area needs to be able to accommodate large files and manipulation of a large data set. For example, access to a SQL server (explained in next section), analysis software, and programming time must all be considered. Each researcher will need to know what the requirements and expectations are for accessing the data.

Hint Seven: *Address confidentiality issues*
Confidentiality is always a concern, but especially with clinical patient information. Be prepared to address who has access to the data, on what machines, in what offices? How will the patient's identity be protected? How will reports of analyzed data be written? Who will get the reports? If you are using multiple databases that need to be connected, you will need to do this through the patient and visit information but after this has been done you will want to scramble the patient identifiers so that they are no longer meaningful.

Hint Eight: *Determine who will have access to the data in the future*
It is important to determine early on who has access to the data, now and in the future. If this is unknown then an alternate approach is to develop a system to request and review this. This is more than a confidentiality issue; it is a quality control issue. It is one of those items that does not seem all that pressing at the time, but in the future, when it arises as an issue, one will be glad this was discussed ahead of time.

Hint Nine: *Meet with key people who control each electronic repository*
Expect to have several meetings. At the first one, establish a positive upbeat relationship and be clear about the purpose and objective of the research and what kind of help you need from them. Even if you think you know the data, get a list of data variables/element, definitions, and codes in each repository. Definitions may limit your inclusion of data, for example if you have a study that involves only the first two days of stay and nosocomial infections are only

recorded 72 hours after admission, these will not be able to be included. You may be surprised as to what is included in the repository that you did not know about. Often these individuals have a great amount of information but they are reluctant to share information unless they are convinced this is: important, confidential, and that you know what you are doing. Ask about years/dates of availability and any limitations. For example, if you want to include cost data, meet with the financial personnel and learn about the system that collects and reports cost and charge data. Find out if data are available at the patient level and determine what is included in the direct and indirect care categories.

Hint Ten: *Keep written notes of dates and people you have met with and the decisions/action steps that were agreed upon*
When people get busy, things sometimes get forgotten and it is helpful to have these notes to refer to. The notes should include the names and contact information for all key people. Later on, when you want to acknowledge the persons who have helped, these notes will be valuable.

Hint Eleven: *Obtain "test files" to develop database structure and test out relationships*
After the meeting, design a request of what is desired; include variable name, dates collected, and the necessary information about how to pull the data (e.g. by certain patient identifiers). We found it very helpful to get only a sample of cases early on and then to build test files to test relationships. When we were sure we had it right, we then requested all the data and had a ready-made structure in which to house the data.

Hint Twelve: *Have or hire a programmer that is paid from or is assigned to the project*
In most facilities, programmers are in short supply and often working on many projects at once. It is best to have a programming person hired and paid to work directly on your project, one that reports to the investigator(s) in charge and who participates in your team meetings. This person will have to set up the files for the relational database and communicate with the technical persons who are in charge of the clinical databases. If you anticipate having a grant-funded project or otherwise are requesting resource help, be sure to include such a position.

Hint Thirteen: *Keep a balance between comprehensiveness and feeling overwhelmed*
The detail and amount of clinical data are overwhelming. Often data are collected hourly or more often and each clinical recording is detailed in the clinical electronic database. For example, a sample of 12,592 patient hospitalizations over 4 years in one hospital resulted in 603,449 records of pharmacy data, 919,756 records of caregiver-patient ratio data, and 6,104,761 records of nursing interventions (label level only, not activities)! One is tempted to "get it all" but this is not a great strategy unless one is sure it will be used in the future. For some variables, more detail is desirable; for others, something simple is satisfactory. For some needs, getting it all is desirable and the right thing to do, but for others it is simply a waste of time and resources. In addition, at some point an overwhelming amount of data leads to the feeling of being overwhelmed and may eventually lead to discouragement and abandonment of the project. Of course, if it is known that the data or resources will not be available in the future one may decide to obtain it now and find ways (e.g. set it aside in the research database) to try to minimize the size. One guideline that may help is to stay focused on the study's aims and not get sidetracked by multiple vague possibilities.

Developing a Relational Database

Commonly, the data elements that are needed to answer the research questions of an effectiveness study are spread across several different databases maintained separately within an organization. There may be separate databases for nursing documentation, medical information, incident reporting, staffing levels, and so on. For example, in our study, variables from nine different databases had to be merged to bring together all the variables needed for the study. The different databases may even be maintained on different computer systems and represent the same data elements in different ways. The process of bringing all the data together into a single database for the purpose of analysis begins with developing a relational database structure. The general method for doing this is described on the following pages. Some of the details in this section are rather technical and are best performed by a database manager with specific training in data management.

Step 1. *Model the Data*

A small amount of data that fits into a single table can be saved in a spreadsheet. When a large amount of data is collected, a more complex database management system is needed. There are several types of database management systems, such as hierarchical, relational, and object-oriented. Relational data base management systems (RDBMS) are very popular and efficient systems that store raw data into tables and define relationships between those tables. RDBMS provide efficient storage of data while maintaining flexibility in manipulation of the data. Users can extract selected variables from many different tables and merge them in a single query to answer specific questions.

To build a relational database, one first should model the data. This should be done before any raw data is transferred to a relational database server. The purpose of data modeling is to best represent the actual data relationships that exist in the real world. In the process of data modeling the following database components are defined: data entity, attributes, relationship between entities, and an identifier for each unique instance within an entity. An example of a data entity is a table called Patient Visits (see Appendix Six), which has general information from the medical record abstract for an acute care patient stay. Each record within this table represents one episode of a patient's acute hospital say. The attributes that describe each visit are Patient ID, Visit Number, Admission Date, Discharge Date, Length of Stay, DRG, etc. That is, the variables that are characteristics of a patient visit are included in this table.

Note: To protect confidentiality, the Patient ID used in the analysis database should not be the actual identification number used by the hospital but an arbitrary number generated as a substitute. In the event that it would be necessary to retrieve additional data for the patient visits, a table can be created that shows the correspondence between the actual patient ID and the arbitrary patient ID generated for the analysis database. It is recommended that such a table be created and kept in a separate database accessible only to the database administrator and project director.

Besides defining database components, data modeling involves the process of normalization. This process usually is completed in five stages, with each stage dedicated to achieving one of five normal forms (Connolly & Begg, 2002). Eliminating repeating fields for a single variable

imposes the *first* normal form. For example, in a raw data file, up to twenty transfer units may be represented for a visit and are stored in variables called unit 1, unit 2, unit 3 ... unit 20. Obviously, transfer unit is a repeating field. To normalize it, a new table called Transfer Unit is defined. It has five attributes: Patient ID, Visit Number (identifying the episode of care), Transfer Unit, Transfer Date, and Transfer Time. If a patient has 12 transfer units during a hospital visit, this patient visit will have 12 records in the table Transfer Unit. With the data structured in this way, there is no limit on how many transfer units can be stored for an episode of care and there are no empty fields if one episode of care has fewer transfer units than others. *Second* and *third* normal forms deal with the relationship between key (common to multiple entities) and non-key fields. *Fourth* and *fifth* normal forms deal with multi-valued facts (subjects who have multiple values for one variable). An example is a person who speaks more than one language; the variable of language would need its own table. Discussing each of these normal forms (stages) is outside the scope of this guideline. Readers can find additional information on the process of normalization in Connolly and Begg's widely used text (2002) or one of the many other texts on database design. The person who is in charge of establishing a relational database should fully understand how to implement five-stage database normalization. Data modeling is completed when all the tables (entities) have been normalized.

Modeling of the data also requires the specification of relationships among entities. A unique identifier common to all entities must be specified in order to link the tables. This link is called a *key*. For example, the identifier for Patient Visits is a *primary key* made up of two attributes—Patient ID and Visit Number. It is a primary key because it can uniquely identify a record in that table: each hospital visit is associated with only one combination of a Patient ID and a Visit Number. To link related entities together, the primary key for Patient Visits may also become a foreign key in other entities (not a part of that Table but essential for linking data). For example the Patient Visits key in the Nursing Interventions table is a foreign key, essential for matching nursing intervention records in the intervention table to a specific visit in the Patient Visits table.

Note: The use of the tables may be more efficient if the patient ID and the visit number are combined into a single variable. The Patient ID and Visit Number variables can be retained as separate variables in the Patient Visits table to permit linkage to a patient table with information about a patient that is constant across visits, while other tables need only contain the single variable created by combining patient ID and visit number.

The second aspect of the relationship between tables that must be specified is whether the relationship is one-to-one, one-to-many, or many-to-many. Relationships between one entity and another are usually one-to-one or one-to-many. For example, because a patient may receive many nursing interventions during a hospital visit, the relationship between the Patient Visits table and the Nursing Interventions table is one to many, with the Nursing Interventions table having one record for every intervention. Appendix Six, the example of a simple database, graphically displays the relationships among tables which is useful for clarifying the relationship structure. An example of such a chart is presented in Appendix Six. Software exists to accomplish this, but the chart may also be sketched out with paper and pencil.

Step 2. ***Document Data Definitions in Codebook***

To use the data, a data codebook is needed as a guide to the database. For a relational database, a section of the codebook can be created for each table. The fundamental elements of a codebook are data field (variable) name, data type (text, numeric, date/time), field length, description of the data field, its possible values, and the descriptive labels for those values (if there are a limited number of possible values). The database administrator creates a codebook for each table or several conceptually related tables together, such as Patient Visits, Medical Diagnoses, Medical Procedures, and Nursing Diagnoses. An example of a simple data codebook is shown in Appendix Seven; an example of a small portion of a data dictionary is in Appendix Eight. The data dictionary reflects the refinements that were made to create and modify variables for the analyses. The data dictionary may include two types of definitions: conceptual, more abstract definitions; and operational definitions, or the way in which the variable is specifically measured for this study. It might also contain supporting materials, such as a lookup-table for a particular data field that has a very large number of possible values. .

Step 3. ***Establish a File Share System (optional) and a Storage Area***

In some cases, a raw data file will simply be too large to be saved efficiently on disks, even high capacity ones. It may be more convenient to use a file server with a large storage capacity to store raw data before doing data transformation. In our case, those who provided raw data files were able to write files directly to a designated file space on a server using a local network. Then the database administrator could access the data on a pc computer by remotely accessing the file server. The file server may also provide data backup during the time before the data are transferred to a SQL server, if the file server is part of a routine data backup system functioning on the network. The term SQL server may refer to a set of software tools used to manage data, to the hardware on which those software tools run, or to both of these together. SQL stands for Structured Query Language (Knight, 2003). SQL server software tools provide a scalable data platform and the means to manage information, build a data warehouse, or generate a backend to support other software applications. SQL server software packages are offered by a number of vendors. Microsoft SQL Server is widely used. Other options available are Oracle DB 9i and DB2 from IBM. The hardware on which the SQL software runs should have a large storage capacity that exceeds the size of the relational database. **The available storage capacity should be at least twice the size of the database.** The extra storage allows for manipulation of data. If the same hardware will be used to conduct analyses with statistical software and store output from these analyses, then more capacity will be needed. It is possible to work around limited storage capacity to some extent, but given the low cost of storage media at the present, it is generally not worth the effort.

Step 4. ***Test the Database Transformation Using a Small Test Database***

Given that data files may be quite large and thus can take a long time to process, it is advisable to test out all data transfer and merging processes on a small test database. If the process fails at any point, it will have taken less time to get to that point and one can more quickly have the problem corrected and be moving forward again. To create the test data files, a small number of hospitalizations can be selected, assuming the episode of care is the

unit of analysis, and the data on those hospitalizations in all the different databases can be requested. If the programming to accomplish this with each database is retained, then it will take little additional effort to modify the programming to retrieve the complete data when all problems that have come to light in processing the test data have been corrected. This small test database can also be used for some sample data runs to determine if the obtained data can in fact address the research questions. Elaborated below are more details on the parts of Step 4: transforming the data, checking the data transformation, and doing sample data runs.

Transforming the database. A temporary place on a SQL server (or alternative storage space) should be established for testing purposes. The testing is conducted in two steps. In the first step, the database administrator creates small programs to transform the test data and transfer it to the test files on the SQL Server. In our study these small programs were created using Data Transformation Services (DTS), an element of the Microsoft SQL Server software package. The programs may alter variable formats as well as transform the data into SQL tables. Changing formats may be needed to assure the data are consistent across tables in the analysis database. If the system generates data transformation errors, for example, notice of an unrecognized data type, the database administrator deletes any tables created on the test server, debugs the errors, and reruns the program. Even if no errors are generated, the logic of the programming should be reviewed to ensure that the results created are the results intended. Once debugging is completed, all the valid DTS packages are saved on the file server or disks to be used later for transferring the real data.

Checking the data transformation. In the second step, the database administrator tests for the validity of database structure and also the data quality. Usually, a data quality check is completed in four steps:

1. Determine if the principal variables characterizing a patient episode of care occur in each table, e.g. hospital number, episode of care number, admission date, and discharge date -- match these with the corresponding variables in the raw data table.

2. Count the total number of records of each table and determine that the count matches the count in the raw data table. When the exact number of records expected is known, for example, it is known that there are 1,000 patient visits, determine that the count is 1,000 in any tables with one record per patient visit. When an expected number is not known, determine that the number of records is reasonable. For example, if there were 1,000 visits and a table with one record for every fall that occurred during the selected patient visits had 1,100 records, the count would be suspect.

3. Check whether the data in the electronic database are defined exactly as the codebook specifies. For example, check that a variable that should have time values in a certain format does have time values in the specified format. If the codebook says that a variable should have values ranging from 1 to 7, then it should be verified that no values fall outside this range. Sometimes certain codes were discontinued and this fact was not noted in the documentation or a special code was used for rare instances. When the original data structure contained a limited number of fields to represent a variable that could repeat an indefinite number of times, such as the number of transfer units, then the last field might contain a special code indicating that there were more

repetitions than the data structure could accommodate. Descriptive statistics can be generated using queries or statistical software, although, with a small number of records, visual inspection may reveal some problems.

4. Determine if redundant information, information available from more than one of the original databases, is consistent. For example, in one study, we had a total length of stay variable, an admission date/time variable, and discharge date/time variable as well as information on the amount of time spent in each unit during the hospital stay. This provided three alternative measures of length of stay. Comparison of all three measures verified that they produced nearly identical values.

Conduct some sample data runs. To test the validity of a database structure, the database administrator usually checks the links between tables, identifiers (primary keys, foreign keys) and constraints by querying the database. One way to do this is to identify a few simple research questions, each of which requires using a set of variables that cross two or more of the original databases, such as the financial database and the pharmacy database. The database administrator would try to answer these questions by analyzing the test data on the server. The results with this small sample are not statistically meaningful, but permit testing whether tables can be linked and variables in the tables manipulated to create new variables needed for analyses.

Step 5. ***Load Entire Database to a Permanent Storage Space***

The database administrator can create a relational database on the permanent SQL server (or other storage space) by simply copying the database structure that has already been established with the test data. The administrator then can load the entire data from the file server to the SQL server using the Data Transformation Services (DTS) packages developed and debugged with the small test files.

In this step, the principal concern may be the time needed to transfer the data. Some of the files may have several million records and may take several hours for the records to be transformed and transferred to the SQL server from the file server. Others may be able to achieve transfers of very large files in shorter times. The time needed is, in part, due to whether data are stored locally. Also, as computers continue to gain in power and speed, the time needed to process large files will decrease. Nevertheless, the time needed to process very large datasets will continue to be non-trivial for some time. Therefore, attention should be given to the timing of the processing of very large files. It may be advisable to start the process at the end of the day and check back on the transfer later in the evening or the next morning so that the computer is available for use during normal working hours. This problem could be avoided by having a second computer dedicated exclusively to the transformation process, but, if the processing is done over a network, carrying it out after normal business hours also means that the network will be less congested.

The database administrator will need to perform checks of the data quality and the validity of the data transformation again, but in this, too, the fact that procedures and programming exist should permit this to be done quickly, with no, or at worst very few, problems being detected.

The data are now ready to be accessed by those who will use it to conduct effectiveness research. The database administrator will need to create database users and grant appropriate privileges to them. It is recommended that people who can access the SQL server should have different levels of read and write privileges. In our case, the database administrator has full privileges to access the data, and all other users only have the privilege to read the data (SELECT statement). The database administrator also facilitates statisticians linking to the SQL server to access data using statistical software, such as SAS or SPSS. In our case, Open Database Connectivity (ODBC) was used to employ SPSS and SAS almost seamlessly in the analysis of data.

Detailed Data Quality Checking

The fact that the data are ready to be accessed does not mean that all data checking is complete. Once it has been determined that the data transformation was successful, the data are ready for more detailed checking. This is the most challenging part of verifying the data. It involves finding those things in the data that are not what you expect them to be. These may be actual errors in the data that the researchers are not yet aware of, or they may be differences between what the data actually are and what they were assumed to be. Outsiders (those who did not create the original database) using a database will always make assumptions about what the data represent. Problems arise when the outsiders' assumptions are both unexamined and incorrect.

Discussions with the creators/managers of each of the original databases (the word liaison will be used here) will aid the researchers greatly in understanding the character of those databases. Nevertheless, even the most cooperative liaisons will not tell you everything you need to know. There are a number of reasons for this. Understanding what those reasons are and what misunderstandings they lead to can help you avoid serious misinterpretations of data. Listed below are five reasons why important characteristics of the data may remain unknown even after discussions with the liaisons of the databases and must be discovered by the investigators. Each is accompanied by at least one example.

1) *You and the liaison may have different ideas about what a record represents because you have different ideas about what a term means.* For example, to you a transfer record represents a between unit transfer, which, in fact, it usually does, but to the liaison, a transfer record represents any change in bed number, including a within unit transfer from one bed to another. Because neither you nor the liaison make explicit what you think a transfer record represents, you may talk about "unit transfer records" extensively without ever discovering that you don't share an understanding of what a record represents and you realize this only when you notice that some patients appear to have more "unit transfers" than they have units on which they resided.

This is such a common problem, that we will provide a second example. Records of charges commonly include records of charges that are "backed out." That is, when a charge in cancelled for whatever reason, the original record of the charge is left in the database and another record is added with the same fields, except the dollar amount has a negative sign, resulting in a sum of zero dollars. If one naively counts the records of a certain category of transactions, the two records will both be added to the count for that category, when these two records actually should result in nothing being added to the count.

2) *You may not be fully aware of the process by which the data were generated, but the liaison is so accustomed to the process that she assumes "everyone" knows how the data were generated.* An example is, in the case of NIC, allowing nurses to sometimes document at only the activity level without an attached intervention label (not a process that we recommend but one that may happen in some agencies). In our data we had a large number of so-called "null interventions" which, in fact, were nursing activities (in NIC, more concrete descriptions of nursing actions) with no label attached. We had assumed that each nursing activity was linked to a nursing intervention label, but, actually, the documentation system gave nurses the flexibility to select nursing activities without linking them to a nursing intervention label. We handled this particular situation by recoding the unlinked activities to the appropriate interventions, using guidelines available in the literature (Coenan, Ryan, & Sutton, 1997; Delaney & Moorhead, 1997) as well as the clinical expertise of the research team.

3) Liaisons may also know things about a database that they do not recall until something prompts them. As an example, the liaison might provide a list of codes for a repeating data field, leaving out a rarely used code that indicates there were more repetitions than the database could accommodate. Once the researcher notices that this code appears and recognizes that it is not simply an error, the liaison can tell the researcher what the code means and may even be able to say that it indicates additional data can be retrieved from the paper documentation.

4) There may be things about the database that no one knows because no one has ever attempted to retrieve that particular data from the database before. An example is discovering that data, believed to have been archived, do not exist. The general process of archiving almost certainly will have been tested, but it probably has not been determined that every piece of data was archived successfully and that nothing catastrophic subsequently happened to it, because in most situations no attempts or very few attempts have been made to retrieve and analyze data for any purpose.

5) The liaison did not tell you what you needed to know because you did not know what you wanted. An example of this is failing to adequately think through the selection criteria for the data. Exclusion criteria often must be quite specific to ensure that the data are appropriate for testing the research question and it is not easy to think of every limitation that is important. If you forget an important limitation, you will not get what you want unless the liaison understands what you want better than you do. In some cases the liaison will know better than you do what you want, but you should not count on this.

Because the liaison does not make explicit all he or she knows about the database and the research team does not know what assumptions about the database they should question, a protocol must be implemented to discover discrepancies between what the data are and what they are expected to be. The time invested in this data checking will prevent the waste of much time and effort later performing analyses that give incorrect results.

The *first rule* to follow is to examine the results of every analysis to see if they are consistent with assumptions about the data. This should start with the descriptive statistics generated to check that only valid values appear. Even if there are only valid values, there still may be problems with the data. Perhaps nearly all responses fall into one category when it would be expected that responses would be distributed across categories. An inquiry could reveal that for many cases it was unclear which category should be used and a coding rule directed choice of a certain code in those instances. Everyone on the research team should be alert to apparent anomalies in the data. Yet some people will be better at this than others. If there is someone on the research team who is disposed to question why the data look the way they do, encourage this tendency.

The *second rule* is to examine important variables broken down by time period and setting, if the data cover an extended period of time and different settings (such as different units within a hospital). There may be other important subgroupings of data to examine, but time period and setting will be significant ones with almost any data used in effectiveness research. Time period is important because data collection rules often change over time and data collection may start and stop at different times or never get started in certain settings. In hospitals, a new computer or data collection system will gradually be brought up unit by unit over time. Data archiving may have failed for a period of time and this was not known because the data had not been accessed.

Sometimes data collection begins at the same time but is implemented differently in different parts of a setting/agency. This can occur because personnel on a unit see their unit as having unique circumstances that require a unique approach to data collection. It may be that the research team will be aware of some of the deviations from standard practice, but not all of them, because there are so many idiosyncratic adjustments that have been made. Consistency of data collection across settings cannot be assumed.

Having several people review the data can help find problems. The person running the analyses may be so caught up in the details of data manipulation, which can take hours, that by the time the results are obtained, he or she is just glad to have them and may not be attentive to minor anomalies. Someone else looking at the results may see something immediately that raises questions about the assumptions that have been made regarding the data.

A *third rule* is to continue to question results throughout the course of the research. If you are continuing to ask new questions, you are also continuing to see the data in a way that you haven't seen it before. This means that it is possible that you will see problems with the data that you have not seen before. It is not pleasant to discover problems after months or years of work, but it is even more undesirable to accept incorrect conclusions.

Document, Document, Document

It is so important to continually document what you have done, why it was done, and what the results were. Summarize the most important information about the data in one or two sheets that are readily available for reference. Keep all programming and document the programming so that anyone who looks at it can understand the purpose and the logic of it. This makes it possible to quickly modify and rerun programs if any errors are discovered later.

The summary of important things about the data should include basic information about the description of the sample, including the total sample size and the overall time frame for the data. It should also include a listing of all the qualifiers that apply to the data. For example, if information from a particular database is only available for a limited period of time, this should be noted. It is easy to forget specific limitations when focusing on other aspects of the data. When you first hear about some limitation in the data, the implications of that information may not be immediately apparent. This increases the chance that you will not remember this information when it does become important later. Therefore, you should document the limitations you have learned from the liaison and from your own investigation and *review* that documentation frequently to ensure that you are not forgetting something of substantial consequence for the investigation.

Another aspect of the data that needs good documentation is when you decide to measure one variable in several different ways. This can get complicated and without good documentation one forgets why decisions were made and exactly what the different measures are. We strongly suggest that the names and operational definitions of variables be kept up-to-date and organized by topic as different measures evolve during the course of data analysis.

Data Analysis

It is not our intent to describe all possible statistical techniques that might be appropriate or required to analyze the data from a relational dataset, but large data sets do generate unique challenges and sometimes require techniques that might not be used elsewhere. It is best to have a statistician as part of the research group and have this person provide leadership for the techniques that are used. In addition to those rather routine descriptive and inferential statistical tests (frequencies, chi square tests, correlations, t-test, analysis of variance, ordinary-least-squares regression) that one often uses with quantitative data, analysis of large databases may require the use of more sophisticated techniques, such as generalized estimating equations (GEE) analysis, logistic regression, poisson regression, or propensity scores (Hosmer & Lemeshow, 2000; Hox, 2002; Stevens, 1996). In addition, the regression analyses will often contain many variables within a general category (e.g. types of drugs within the pharmacy group) that require procedures and decisions about how to reduce the number of independent variables to a meaningful size. We discuss some of these issues and related methods below.

Statistical modeling. For the analysis, some form of multivariable analysis should be carried out, that is, an analysis with multiple independent variables. The analysis must have multiple independent variables because, as discussed, many variables will likely have an effect on the outcome and those getting the intervention will likely differ on these variables from those not getting the intervention. The analysis generally may be constructed as some form of regression because developments in statistics have led to various methods of analyses being subsumed under the regression framework. Thus, analysis of variance and analysis of covariance, as well as crosstabulations of categorical variables, may all now be treated as a form of regression.

The regression framework has also been extended with the development of new statistical methods to cope with the lack of independence among cases that exists when cases are clustered in some way, for example, when the cases are patients grouped within hospital units. These developments of the regression framework in statistics have made regression a more versatile and powerful tool. They also require that those planning to carry out statistical analyses stretch themselves to understand and employ more sophisticated approaches to analyses than have been used in the past. Discussions of basic multivariable regression analysis can be found in a number of textbooks, such as Agresti and Finlay's (1997) well-written text. Those needing details on how to carry out a statistical analysis should, at a minimum, study one of the basic texts thoroughly, and preferably obtain the assistance of a statistician.

A simple model. The following equation displays a simple statistical model with the dependent variable being a measure of satisfaction with care. In this simple model there are only two variables other than the nursing intervention included as independent variables. The b's represent the effect of each independent variable on the satisfaction score, with the effect of the other independent variables controlled, so that b_3 represents the actual effect of the nursing intervention, assuming there are no other variables related to satisfaction and the intervention.

$$\text{Satisfaction score} = a + b_1\text{Patient\_Age} + b_2\text{Patient\_Gender} + b_3\,\text{Nursing Intervention} + e$$

This statistical model could be run in SPSS with the following command. The satisfaction score is represented by *satsfctn*, while *pt_age*, *pt_gendr*, and *intvn* represent the patient age, patient gender, and nursing intervention variables, respectively. *Pt_gendr* and *intvn* are categorical variables with two categories. *Pt_gendr* is scored 1 if the patient is female, 0 if the patient is not. *Intvn* is scored 1 if the patient received an intervention and 0 if the patient did not. This formulation extends the basic regression model to subsume analysis of covariance.

```
REGRESSION
  /STATISTICS COEFF R ANOVA
 /DEPENDENT satsfctn
 /METHOD=ENTER pt_age pt_gendr intvn.
```

The same statistical model could be run in SAS using PROC GENMOD with the following command. The CLASS statement identifies *pt_gendr* and intvn explicitly as categorical variables. The LINK statement indicates that the dependent variable *satsfctn* is to be treated as an interval-level, normally distributed variable (See following).

```
PROC GENMOD data=testdata;
CLASS pt_gendr intvn;
model satsfctn = pt_age pt_gendr intvn
                    / LINK= identity type3;
```

The statistical model actually represents an endpoint. Before commands to complete an analysis are actually set up, a number of questions must be answered. These follow.

Questions to answer. Eight questions that relate to the selection of the appropriate statistical model for data analysis are addressed here.

1. What software should be used? SPSS and SAS statistical software are the most widely used general-purpose statistical software, with SPSS probably having the edge among researchers in nursing. Both have substantial data management capabilities (merging and restructuring data files, computing and transforming variables) and include procedures for a wide variety of statistical techniques. SAS tends to be more on the cutting edge in terms of the procedures available, but it also tends to be less accessible to the ordinary user. The providers of these two packages, SPSS, Inc. and SAS Institute, Inc., both have websites (www.spss.com and www.sas.com) and both have special offers for academic institutions. You may check with them regarding special offers for governmental entities and non-profit institutions, too. There are other general purpose software packages from smaller vendors. Some such as S-Plus (www.insightful.com) offer sophisticated packages. There are also some widely used specialized programs for multilevel modeling (See the question "Are the cases grouped in clusters?" section below). These are HLM (www.ssicentral.com) and MLwiN (multilevel.ioe.ac.uk/index.html). Because they are designed specifically for analysis of multilevel models, also known as hierarchical linear models, these programs cam make it easier to set up and run this type of analysis.

2. How many cases are needed? Power analysis (Cohen, 1988) is widely recommended and applied to ensure that number of cases used is sufficient to detect an effect when there actually is an effect. With effectiveness research, there may well be a very large number of cases available, possibly thousands. This makes power analysis less of a concern. No hard and fast rule can be given, but, in general, if one has a large number of cases, perhaps greater than 300-400, it is likely that there will be power to detect any effect that is clinically significant enough to want to detect it. If there any doubt, a power analysis should be performed. This is best done by working together with a statistician, because a well-done power analysis requires both statistical expertise and substantive knowledge of what constitutes a clinically important effect.

Another consideration is the number of cases used in relation to the number of independent variables in the analysis. The primary concern from this point of view is whether reliable (reproducible) results will be obtained. Stevens (1996) argues that practical experience, as well as the limited research done relevant to the issue, indicates there be at least 15 cases per independent variable, even though a commonly heard rule of thumb is at least ten cases per independent variable. Pedhazur and Schmelkin (1991) recommend 30 cases per independent variable. Even when there are thousands of cases, it is generally desirable to have in the final model for analysis, a relatively small number of independent variables, perhaps no more than ten to fifteen, because the results are more likely to be reproducible (Iezzoni, 2003).

3. Does a subset of cases need to be selected for the main analysis or some part of the analysis? It is not uncommon for data to be unavailable for certain time periods or for definitions of variables to have changed at some point in time. The most practical solution for this may be to restrict the analysis to data collected within a limited time frame. There may be other circumstances for which the most practical solution is to restrict the cases included in

the database in some way. Thus, the first task in analysis is to make sure that you have defined the dataset appropriately for the question(s) that you want answered. If you have not, you might end up having to redo analyses.

4. What are the characteristics of the cases selected for the analysis? Before analyzing the effect of the intervention of interest, descriptive measures should be generated for all the variables in the dataset. For categorical variables the descriptive measures would be frequency counts and percentages for the categories. If some categories are found to have few cases, one can consider whether it makes sense to combine categories with small frequencies. For interval-level data, measures of skewness (lopsidedness of the distribution) should be generated as well as means and standard deviations. A graphical representation of the distribution can also be helpful in seeing how well or poorly the distribution matches a normal distribution. Generating measures of bivariate associations among the variables is also helpful in providing an understanding of the character of the dataset. All of the above these things should have been done when checking the data and you should have a good idea of the description of the data already, but all this descriptive information should be generated and saved so that it can be presented to others along with the findings regarding the intervention(s) of interest.

5. What is the level of measurement for the outcome(s) of interest? The answer to this question is important for determining the specific statistical technique that will be used, as well as for clarifying what question you are asking. Much of the statistical training that non-statisticians receive focuses on techniques for analyzing outcome variables that are interval-level, that is, have equally-spaced, ordered values (e.g., age), and are normally-distributed, that is, have the familiar bell-shaped distribution. However, many of the common outcome variables are not interval-level, normally-distributed variables. Some have only two values. Mortality is an example; the patient did or did not die within a specified period. Other outcomes, such as falling, may be thought of as having two values (dichotomous variables). A patient had a fall (at least one) or did not have a fall. Such outcomes may also be thought of as counts. The patient had X number of falls during a specified time period. Counts of adverse events, such as falling, will generally not have a normal distribution. The largest number of patients will have no falls, the next largest number will have one fall, and so on. The maximum number of falls will be limited to a fairly low count.

The characteristics of the outcome measure will determine the specific type of regression model that is most appropriate to use. Ordinary-least-squares regression is robust to some violation of the assumptions about the level of measurement and the normality of distribution of the outcome variable, meaning that one can get valid results even when the assumptions are not met (Hosmer & Lemeshow, 2000). However, when the assumptions are violated substantially, such as they are when the outcome variable is dichotomous, the use of ordinary-least-squares regression is questionable. With a dichotomous outcome variable, logistic regression (Hosmer & Lemeshow, 2000) will be the most appropriate statistical technique. Generally, the distribution of responses for a dichotomous health-related outcome will be very unequal. For example only a small percentage of patients will have mortality as an outcome, while the great majority of patients will survive. The more unequal the distribution of patients between a good outcome and a bad outcome, the more desirable it is to use logistic regression over ordinary regression. Both SPSS and SAS statistical software packages have

procedures that will carry out logistic regression. Logistic regression yields odds ratios as a measure of the effect of an independent variable on an outcome. The correct interpretation of odds ratios is, however, sometimes difficult (Scott, Mason, & Chapman, 1999; Zhang, 1998).

When the outcome is a count, poisson regression may be the most appropriate statistical technique. The procedure PROC GENMOD in SAS will carry out poisson regression. Poisson regression is a step up in difficulty from logistic regression and you will almost certainly want a statistician familiar with it, if you want to examine outcome variables as counts.

Other outcome variables commonly examined can be thought of as interval-level, but may deviate substantially from being normally distributed. Hospital costs (or charges) and length of stay are two commonly examined outcome measures that are typically not normally distributed, because the distributions are skewed to the right. That is, there are a small number of cases with large values that stretch out the right tail of the distribution. A fairly simple way to handle these variables is to transform them by taking the natural log of each of the actual values. The log value is then analyzed as the dependent variable using ordinary regression. More sophisticated approaches are possible, but the log transformation is, in our opinion, an acceptable, relatively simple solution. The parameters obtained in the ordinary regression can be translated back into actual values by exponentiating. Further information on using log transformations can be found in Chapter 10 of Iezzoni (2003).

6. What is the form of the relationship between independent variables and outcomes? When thinking about what variables to include in the dataset, some thought should have been given to how these variables relate to one another. Not all independent variables will have a simple linear relationship with outcomes, such that a higher level is always better (or always worse) for an outcome than a lower level no matter what the actual level of that variable. For example, a systolic blood pressure of 120 is better than a pressure of 50, but a pressure of 170 is not better than a pressure of 120. Depending upon the nature of the outcome and the independent variable, it may be desirable to group values of such a variable into three categories, low, normal, and high, to see if outcomes are significantly better when the value is in the normal category than when it is in either the low or high categories. More detailed categorization might be done if that seemed to make clinical sense. There are a number of possibilities for deviation from a linear relationship. We cannot detail and discuss all of these, but review of the literature and consideration of the way a specific variable is expected to work clinically in relation to a specific outcome is an important part of setting up the analysis in an appropriate way (Iezzoni, 2003).

There is a method that eliminates the need to know the form of the relationship between the outcome and all the risk-adjustment variables. This method, which has only recently come into wide use, is the use of propensity scores (Rosenbaum & Rubin, 1983; 1985). It has similarities to case-control methodology in that a procedure is followed resulting in the selection from all the cases available a set of cases with the intervention and a set of cases without the intervention that are comparable to one another. The two sets of cases are comparable in the sense that they do not differ on all those variables known to affect the choice of assignment to getting or not getting the intervention. They differ (as far as can be determined) only on getting the intervention. If the nature of the relationship of the risk-

adjustment variables is not of interest, this is a possible approach focusing only on the effect of the intervention among a group of similar cases, but if the effect of the risk adjustment variables themselves are of interest, then it would not be a good choice.

7. Are the cases grouped in clusters? With health care data, the cases one wants to examine are almost always clustered within some larger unit. For example, patients are clustered within clinics or within hospital units or within geographical areas. Also, individual patients may have several episodes of care, each of which appears as a case in a dataset. These episodes can be viewed as clustered within individuals. The basic reason for thinking about whether the cases are grouped in clusters is that cases within a cluster cannot be assumed to be statistically independent (Fitzmaurice, 2001). They can be expected to be more like one another than they are like cases in a different cluster. To get the least biased tests of intervention effects, the statistical technique used for analysis should take into account the intraclass correlation, sometimes called the intracluster correlation, created by the lack of independence of cases.

There are basically two statistical approaches to dealing with the problem of clustered data. One is the generalized estimating equations (GEE) approach (Liang & Zeger, 1986). A GEE analysis can be carried out with PROC GENMOD in SAS statistical software. The following displays the SAS syntax for a GEE analysis. The Data statement, Class statement, Link statement, and the outcome variable are all the same as in the PROC GENMOD statement given above. The REPEATED statement generates the GEE analysis. The *subject =pt_id* part of the REPEATED statement designates the variable that identifies clusters. All the cases with the same value for this variable are in the same cluster. Here the variable identifies individual patients, who have repeated visits that appear as cases in the dataset.

```
PROC GENMOD data=testdata;
CLASS pt_gendr intvn;
model satsfctn = pt_age pt_gendr intvn
                    / LINK= identity type3
  repeated  subject=pt_id / type=cs ;
    run;
```

If the concern is simply to obtain correct tests of the effects of the independent variables, a GEE analysis would be the approach of choice. In a GEE analysis, the intraclass correlation of cases within a cluster is regarded as a "nuisance parameter" that needs to be taken into account, but the effect of the clusters themselves is not of interest. If there is concern for the effects of the clusters themselves, then the second approach, multilevel modeling, also called hierarchical linear modeling, would be the approach to choose.

In a multilevel analysis, the cases are viewed as level 1 and the clusters are viewed as level 2. Analyses are carried out simultaneously for levels 1 and 2. For example, the level 1 cases could consist of data on individual patients with length of stay as an outcome. Then the level 2 clusters could be different hospitals in which the patients received care, and there would be data to describe characteristics of the hospitals, such as size. The Level 1 analysis could examine what characteristics of the patients are significantly related to the length of stay for a visit. The level 2 analyses could examine what characteristics of the hospital predicted the

mean length of stay for patients within a hospital. If the multilevel model is properly specified, the estimate of effects at the individual patient level will be adjusted for the effect of the larger units in which they are clustered, while simultaneously the estimate of effects at the hospital level will be adjusted for the effect of the characteristics of the individual cases. Detailed discussion of multilevel models (hierarchical linear models) can be found in works by Hox (2002), Goldstein, Browne, and Rabash (2002), and Bryk and Raudenbush (1992). Multilevel modeling may be applied to the analysis of repeated measures. This type of repeated measures analysis has an advantage over traditional repeated measures analysis of variance methods when repeated measures are missing for some cases at some time periods and when the time period between measurements varies among cases. Multilevel modeling can be implemented with both SPSS and SAS statistical software, as well as specialized programs such as HLM, and MLwiN, mentioned above.

8. Is there missing data and, if so, why is it missing? The best solution to missing data is not to have any, but unfortunately this is rarely possible. How much of a problem missing data creates depends upon the amount of missing data and the mechanism that creates the missing data. If the amount of missing data is small and the mechanism creating missingness is completely random, perhaps clerical error unrelated to the nature of the variable, then missing data might have no substantive effect on research findings, even if nothing is done to correct for missing data. The only effect then is to reduce statistical power, the ability to detect a relationship when there truly is one. If the number of cases lost is small and the number of cases remaining is large, the effect on statistical power may well not be a problem. Knowing what is "small enough" and "large enough" would require a power analysis, of course, and cannot be precisely specified otherwise.

Missing data for a variable is most problematic when it is not truly random (Little & Rubin, 1987), but rather depends upon the value of other variables, either ones included in the analysis or ones not included in the analysis. The latter, missingness due to variables not in the analysis, presents the greatest threat to the validity of results.

There are many approaches to the handling of missing data that are preferable to the traditional approaches of listwise deletion or mean substitution. These newer approaches include, incidentally, the use of GEE analysis and multilevel modeling. Multiple imputation (Graham & Hofer, 2000) is an approach that permits assessment of the extent of the effect of the missing data on the results. This approach can be handled now with SAS or SPSS statistical software. Discussion of this topic is beyond the scope of this monograph. Researchers should read current discussions of missing data (e.g. Graham & Hofer, 2000; Little & Rubin, 1987) and preferably have a statistician with knowledge of contemporary methods.

Statistical model selection. When developing a statistical model, it may become apparent that a large number of variables are potentially related to the outcome. Sometimes there are a large number of variables just within a particular category of variable, such as pharmacy treatments or medical treatments. This makes it desirable, even essential, in the course of the analysis to reduce the number of independent variables to a manageable size, and not simply throw in all the potential independent variables at once.

A common approach to reduction of the number of variables is to first screen the variables by testing each variable individually for its relationship with the outcome of interest. Only those variables that meet a predefined criterion for the significance level of its relationship with the outcome are retained for further analyses. The p value may be set above the conventional .05 level when exploratory analyses are being conducted (Iezzoni, 2003); for example, the p value might be set at .10 or .15. This protects against the exclusion of variables that might prove to have a significant relationship when they are included with other independent variables in a multiple variable analysis.

When there are a large number of potential predictors within a category of variables, another step might be added, as suggested by Hosmer and Lemeshow (2000). For example, in one study we had a fairly large number of variables to consider with each of several categories of variables (demographic characteristic, clinical conditions, etc.). We followed the procedure of testing all the individual variables within a category separately for their relationship to the outcome. Then all the variables within a category that were found to have a statistically significant zero-order relationship with the outcome (using p<= .15 as the criterion) were considered together in a multiple variable analysis. Variables within a category that continued to have a statistically significant relationship, using the same criterion, were retained for the next step in development of the analytical model.

Those variables that pass the significance level test are then added together in a multiple variable analysis to determine which still have a significant effect when other variables are controlled for in the analysis. We recommend against the common practice of using an automatic variable selection process, such as forward or backward stepwise regression, to trim the number of variables. We believe that the researcher should be more active in the process of variable selection and that the selection be guided by a literature-based theoretical model and by the best clinical judgment available. For example, if two independent variables are related to one another and each has a significant zero-order correlation with the outcome, when they are entered together into an analysis, neither may have a significant relationship with the outcome. However, the researcher may judge that clinically one of these variables could be seen as more directly related to the outcome and this one would be selected for retention in the analytical model. In the study just mentioned, after testing all the variables within each category separately, variables retained within each category were added in a predetermined order, applying a research-based theoretical model and the investigators' clinical judgments. Demographic characteristics were viewed as acting through their effect on other variables. For example, age was perceived as having its effect, at least primarily, through its impact on clinical conditions (primary diagnosis, co-morbidities, and severity of illness). Based on this assumption, clinical conditions were added to the model after the demographic characteristics to see if demographics continued to have a significant effect after clinical conditions were considered. They did not.

Statistical model selection is as much an art as a science. No specific algorithm for the selection of variables for inclusion in the final model will guarantee that the final model represents the true model of relationships among the independent variables and the outcome. Given that the correct specification of the relationship among variables is unknown, any strategy used to reduce the number of independent variables used possibly could eliminate variables that are related to the outcome and retain variables that are not. Nevertheless, every

completed analysis can help build an understanding of what nursing interventions contribute to outcomes in real-world settings. That is why it is important to pursue analyses despite the imperfections in the process.

Interpretation of the parameters in regression output. The b's in the regression equation, the parameters obtained from a regression analysis, can be interpreted to provide an estimated effect of the independent variables, including the intervention of interest. An overview of the interpretation of output from both ordinary-least-squares regression and logistic regression can be given here, but a detailed explanation of the interpretation of output is outside the scope of this guideline. A basic text in regression analysis such as the text by Agresti and Finlay (1997) should be consulted for further information.

Ordinary-least-squares regression. For the output from an ordinary-least-squares regression, assuming that the effects are statistically significant and that the parameters are unstandardized regression coefficients, then the interpretation can be made in terms of the raw units of the outcome measure and the independent variable. For example, if the outcome measure is cost of care (costs) and the independent variable is patient age, then the b associated with the patient age variable could be interpreted as the number of dollars by which costs increase (or decrease, if the sign is negative) for one year increase in patient age, given that patient age is entered as years in the regression.

An example of output from an analysis done in SPSS using the menus to choose Analyze, GLM, Univariate, then clicking on Options and checking the Parameter Estimates box. The results are simulated. For the purposes of this example, it is assumed that the model is correctly specified.

Parameter Estimates

Dependent Variable: Total Costs for Visit

Parameter	B	Std. Error	t	Sig.
Intercept	45607.453	1535.951	29.693	.000
[PT_GENDR=F]	-772.214	392.574	-1.967	.049
[PTGENDER=M]	0(a)	.	.	.
[SEVR=1]	-32302.166	770.611	-41.918	.000
[SEVR=2]	-31219.223	568.433	-54.922	.000
[SEVR=3]	-26109.556	587.424	-44.448	.000
[SEVR=4]	0(a)	.	.	.
PT_AGE	-158.349	21.122	-7.497	.000

a This parameter is set to zero because it is redundant.

In the table above, the results indicate that costs decrease by approximately $158 dollars for every additional year of patient age, because the b parameter for the PT_AGE is equal to a negative (-) 158.349. Furthermore, costs are approximately $772 lower for women than for men. In this model, patient gender (PT_GENDR) is a categorical variable and the male category is a referent category, that is, the exponent for the PT_GENDER = F entry represents costs for women relative to costs for men. SEVR is a measure of severity of illness that has four categories. The referent category is category 4, the most severe level. The parameters for levels 1,2, and 3 of the SEVR variable indicate the difference in cost for the respective level in comparison to level 4, the referent category. As would be expected, the visits categorized in a lower level of severity, have lower costs, with costs for visits in level 3 being approximately $26,110 lower than costs for visits in level 4.

Logistic regression. Interpretation of parameters in a logistic regression are more complicated because the parameters are actually natural logarithms. An odds ratio is calculated parameter by raising the base *e* to the power b. The base *e* equals roughly 2.718. Negative parameters result in odds ratios less than 1, while positive parameters result in odds ratios greater than 1. A zero value for the parameter results in an odds ratio of 1, which indicates the odds are equal, i.e., there is no effect. This means that negative, positive and zero parameters can be interpreted in the same way as with an ordinary-least-squares regression as indicating that the effect of the independent variable on the outcome is negative, positive, or zero.

The following table shows the output from running a logistic regression in SPSS. Again the results are simulated and it is assumed for the purposes of the example that the model is correctly specified. Some formatting of the original output has been done to improve readability.

Parameter Estimates

Dependent Variable: Mortality

	B	S.E.	Wald	df	Sig.	Exp(B)
PT_GENDR(F)	-.283	.088	10.248	1	.001	.753
SEVR			663.566	3	.000	
SEVR(1)	.899	.432	4.330	1	.037	2.458
SEVR(2)	2.081	.418	24.734	1	.000	8.012
SEVR(3)	3.916	.413	89.824	1	.000	50.191
PT_AGE	.025	.005	23.484	1	.000	1.025
Constant	-6.921	.560	152.638	1	.000	.001

The Exp(B) column displays the exponentiated value of the b. That is, it displays the odds ratio corresponding to that parameter. With odds ratios as the measure of effect, the effect of a categorical variable is generally somewhat easier to express then the effect of continuous variables, like age, so interpretation of the above table will begin with gender. The parameter for the PT_GENDR(F) entry is a negative (-) .283. This converts to an odds ratio of 0.753, indicating that for women the odds of mortality being the outcome of the visit are about three-fourths of the odds of mortality for men.

The b parameter associated with PT_AGE is .025. This converts to an odds ratio of 1.025, indicating that, for a one-year increase in patient age, the odds of mortality being the outcome of a visit are 1.025 times higher. With many continuous variables, one unit increase may not be a very meaningful measure. If the ranges is large, perhaps in the thousands, then the odds ratio associated with the variable could be very low, something like 1.002, yet still be statistically significant. It would make the odds ratio a more meaningful value to transform the scale of the variable, perhaps by dividing all values by 1,000. With patient age as used in this example a transformation of scale by dividing by 10 results in an odds ratio of 1.28, indicating that for a 10-year increase in patient age, the odds of mortality are 28% higher. Conversely, if meaningful differences in the independent variable are expressed in decimals, then very large, and not meaningful, odds ratios will result unless a transformation is made to adjust the scale.

The most confusing aspect of interpreting logistic regression results is that odds ratios are not the same as risk ratios. A risk ratio indicates relative probability of an outcome occurring. Typically, a risk ratio interpretation is the interpretation made of logistic regression output. If the odds ratio for an independent variable representing a medical diagnosis is 1.5, then the common interpretation is that the probability of the outcome for those with this medical diagnosis is 1.5 times the probability of the outcome for those without the diagnosis. When

the outcome is fairly rare, which outcomes such as adverse events often are, then this interpretation is roughly correct. However, if the outcome is fairly common, then the odds ratio and the risk ratio can diverge greatly. There is an extensive literature on this topic. Scott, Mason, and Chapman (1999) provide one of the clearest discussions on the difference between odds ratios and risk ratios.

Presentation of Findings

The way in which findings are presented will depend to a large extent upon the audience. If publication of results in an academic journal is anticipated, then one will want to prepare tables similar to the tables shown in the preceding section on interpretation of output. For each independent variable list there should be a parameter value, a standard error value, a test statistic (such as a t value), and the significance level (p value). The exact format in which the results are presented will depend upon the format specified by the journal. For those journals using the American Psychological Association format, the Publication Manual of the American Psychological Association, 5th edition (2001) provides extensive information on how to present results.

If the presentation of findings is aimed at persons in clinical practice, then the principal concern should by how understandable the presentation is and how important the results appear to be. It is not enough to have obtained the correct answers. You have to convince clinicians that these are the correct answers and that the answers matter enough for them to make changes in their practice. Understandable means most of all that clinicians find the results to be credible. If clinicians see the results as contradictory to their clinical experience, they will not accept them. Ensuring that the results are credible to clinicians begins with developing a theoretical model that reflects reported research and clinical expertise. That is the most important element in credibility. However, connecting the dots, making a clear, clinically-based argument about the mechanisms that produce the results, is also important. Understandable also means expressing the effect of the intervention in clinically meaningful terms, such as "If you do intervention Z two more times per day, outcome Y for a patient of type M will will be 25% better." Furthermore, if doing intervention A two more times a day is a practical goal from the clinicians viewpoint and a 25% improvement in outcome Y appears to be clinically meaningful, you will also convince clinicians that the findings matter.

Conclusion

An increasing amount of nursing and other clinical data is being collected through computer documentation of care delivered in practice. The data are now routinely saved and can be used to address questions related to improvement in quality. Standardized nursing languages are available and now more frequently used to document the provision of nursing care. With the use of standardized language in clinical information systems, the availability of large data storage systems, and the relative ease of using personal computers for data analysis, the use of clinical data to answer compelling research questions is now timely. Few people, in nursing or other health care disciplines, have experience using actual clinical data from a clinical data repository for research. The use of national databases containing data that have been aggregated across many facilities and already cleaned and arranged in a relational structure requires large data base analysis skills. However, none of these national or regional databases

include nursing data. In order to address nursing questions, one needs to go through the process as outlined in this guideline. In the future we hope that many of these steps will become incorporated into vendor systems or be a part of the information technology departments in the clinical settings. To assist that to become a reality and to help all those who are currently building or want to use clinical documentation systems, we have written this guideline. Transforming data into meaningful and useful information is a goal worth achieving.

Feedback on this Guideline

If you have found this publication helpful, or have suggestions for improvement, please send your comments to Center for Nursing Classification and Clinical Effectiveness, College of Nursing, University of Iowa, Iowa City, Iowa 52242. (email: classification-center@uiowa.edu) Your comments will be given to the authors and suggestions for improvement will be addressed in future revisions of this Guideline.

References

Agresti, A., & Finlay, B. (1997). *Statistical methods for the social sciences* (3rd ed.). Upper Saddle River, NJ: Prentice Hall.

Atherly, A. (1997). Condition-specific measures. In R.L. Kane (Ed), *Understanding health care outcomes research* (pp. 53-66). Gaithersburg, MD: Aspen Publishers.

American Hospital Formulary Service. (2000). *AHFS drug information 2000*. Bethesda, MD: American Society of Health System Pharmacists.

American Medical Association. (1999). *Current procedural terminology* (4th ed.). Chicago: Author.

American Nurses Association. (1995). *Nursing care report card for acute care*. Washington, DC: American Nurses Publishing.

American Psychological Association. (2001). *Publication manual of the American Psychological Association* (5th ed.). Washington, DC: Author.

Bryk, A.S., & Raudenbush, S.W. (1992). *Hierarchical linear models*. Newbury Park, CA: Sage.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Lawrence Erlbaum.

Coenen, A., Ryan, P., & Sutton, J. (1997). Mapping nursing interventions from a hospital information system to the Nursing Interventions Classification (NIC). *Nursing Diagnosis, 8*(4), 145-151.

Connolly, T., & Begg, C. (2002). *Database systems: A practical approach to design, implementation, and management*. New York: Addison-Wesley.

Davies, AR, & Ware Jr., JE. (1991). *GHAA's consumer satisfaction survey and user's manual*. Washington, DC: Group Health Association of America.

Delaney, C. & Huber, D. (1996). *A Nursing Management Minimum Data Set (NMMDS): Report of an invitational conference*. Chicago: American Organization of Nurse Executives.

Delaney, C., & Moorhead, S. (1997). Synthesis of methods, rules and issues of standardizing nursing intervention language mapping. *Nursing Diagnosis, 8*(4), 152-156.

Dochterman, J.M. & Bulechek, G.M. (Eds.). (2004). Nursing *Interventions Classification (NIC)* (4th ed). St. Louis: Mosby.

Fitzmaurice, G. (2001). Clustered data. *Nutrition, 17*, 487-488.

Garner, J.S., Jarvis, W.R., Emori, T.G., Horan, T.C., & Hughes, J.M. (1988). CDC definitions for nosocomial infections. *American Journal of Infection Control, 16*(3), 128-140.

Goldstein, H., Browne, W., & Rabash, J. (2002). Multilevel modeling of medical data. *Statistics in Medicine, 21*, 3291-3315.

Graham, J.W., & Hofer, S.M. (2000). Multiple imputation in multivariate research. In T.D. Little, K.U. Schnabel & J. Baumert (Eds.) *Modeling longitudinal and multilevel data: Practical issues, applied approaches, and specific examples*. Mahwah, NJ: Lawrence Erlbaum Associates.

Guadagnoli, E., & McNeil, B.J. (1994). Outcomes research: Hope for the future or the latest rage? *Inquiry, 31*(1), 14-24.

Hosmer, D., & Lemeshow, S. (2000). *Applied logistic regression*. New York: Wiley.

Hox, J. (2002). *Multilevel analysis: Techniques and applications*. Mahwah, NJ: Lawrence Erlbaum Associates.

Iezzoni, L. (2003). *Risk adjustment for measuring health care outcomes*. Chicago: Health Administration Press.

Ingenix. (2004). *International Classification of Diseases, 9th revision (ICD-9 CM)*. Salt Lake City, Utah: Author.

Kane, R.L. (Ed.). (1997). *Understanding health care outcomes research*. Gaithersburg, MD: Aspen Publishers.

Knight, B. (2003). *SQL Server for experienced DBA's*. Berkeley, CA: Osborne/McGraw Hill.

Liang, K., & Zeger, S. (1986). Longitudinal data analysis using generalized linear models. *Biometrika, 73*(1), 13-22.

Little, R.J.A., & Rubin, D.B. (1987). *Statistical analysis with missing data*. New York: Wiley.

Lohr, K.N. (1988). Outcome measurement: Concepts and questions. *Inquiry, 25*(1), 37-50.

Mandelblatt, J.S., Fryback, D.G., Weinstein, M.C., Russell, L.B., & Gold, M.R. (1997). Assessing the effectiveness of health interventions for cost-effectiveness analysis. *Journal of General Internal Medicine, 12*(9), 551-558.

Maciejewski, M. (1997). Generic measures. In R.L. Kane (Ed), *Understanding health care outcomes research* (pp. 19-52). Gaithersburg, MD: Aspen Publishers.

Moorhead, S, Johnson, M., & Maas, M. (Eds.). (2004). *Nursing Outcomes Classification (NOC)*, (3rd ed.). St. Louis: Mosby.

Muennig, P. (2002). *Designing and conducting cost-effectiveness analyses in medicine and health care*. San Francisco: Jossey Bass.

NANDA International. (2003). *Nursing diagnoses: Definitions & classification, 2003-2004*. Philadelphia: Author.

Pearce, N.D. (1988). Uniform minimum health data sets: Concept, development, testing, recognition for federal health use, and current status. In H.H. Werley & N.M. Lang (Eds.), *Identification of the nursing minimum data set*. New York: Springer.

Pedhazur, E.J., & Schmelkin, L. P. (1991). *Measurement, design, and analysis: An integrated approach*. Hillsdale, NJ: Earlbaum.

Practice Management Information Corporation. (1989). ICD-9-CM International classification of diseases (9th rev., ed 3) clinical modification. Los Angeles: Author.

Rosenbaum, P.R., & Rubin, D.B. (1983). The central role of the propensity score in observational studies of causal effect. *Biometrika, 76*, 41-55.

Rosenbaum, P.R., & Rubin, D.B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *American Statistician, 39*, 33-38.

Scott, K.G., Mason, C.A., & Chapman, D.A. (1999). The use of epidemiological methodology as a means of influencing public policy. *Child Development, 70*(5), 1263-1272.

Stevens, J. (1996). *Applied multivariate statistics for the social sciences*. Mahwah, NJ: Earlbaum.

Titler, M.T. (PI). (2001). *"Nursing Interventions and Outcomes in 3 Older Populations."* (NINR and AHRQ, RO1NR05331-01A1), funded 2001-2005.

University of Iowa Hospitals & Clinics. (2003). *Infection control manual*. Retrieved January 12, 2004, from http://policies.uihc.uiowa.edu/Infection%20Control/epidem.htm

Werley, H.H. & Lang, N.M. (Eds.). (1988). *The identification of the Nursing Minimum Data Set*. New York: Springer Publishing.

World Health Organization. (1992). International statistical classification of diseases and related health problems, rev 10. Geneva, Switzerland: Author.

Zhang, J. (1998). What's the relative risk? A method of correcting the odds ratio in cohort studies of common outcomes. *JAMA: The Journal of the American Medical Association, 280*, 1690-1691.

Appendix One:
One example each of a NANDA diagnosis, NIC intervention and NOC outcome

NANDA Diagnosis

ACTIVITY INTOLERANCE

Definition Insufficient physiological or psychological energy to endure or complete required or desired daily activities

Defining Characteristics

- Verbal report of fatigue or weakness
- Abnormal heart rate or blood pressure response to activity
- Electrocardiographic changes reflecting arrhythmias or ischemia
- Extertional discomfort or dyspnea

Related Factors

Bed rest or immobility
Generalized weakness
Imbalance between oxygen supply / demand
Sedentary lifestyle

Source: NANDA International (2003). Nursing diagnoses: Definitions & classification, 2003-2004. Philadelphia: NANDA International.

NIC intervention

Activity Therapy - 4310

DEFINITION: Prescription of and assistance with specific physical, cognitive, social, and spiritual activities to increase the range, frequency, or duration of an individual's (or group's) activity

ACTIVITIES:

Collaborate with occupational, physical, and/or recreational therapists in planning and monitoring an activity program, as appropriate

Determine patient's commitment to increasing frequency and/or range of activity

Assist to explore the personal meaning of usual activity (e.g., work) and/or favorite leisure activities

Assist to choose activities consistent with physical, psychological, and social capabilities

Assist to focus on what patient can do, rather than on deficits

Assist to identify and obtain resources required for the desired activity

Assist to obtain transportation to activities, as appropriate

Assist patient to identify preferences for activity

Assist patient to identify meaningful activities

Assist patient to schedule specific periods for diversional activity into daily routine

Assist patient/family to identify deficits in activity level

Instruct patient/family regarding the role of physical, social, spiritual, and cognitive activity in maintaining function and health

Instruct patient/family how to perform desired or prescribed activity

Assist patient/family to adapt environment to accommodate desired activity

Provide activities to increase attention span in consultation with OT

Facilitate activity substitution when patient has limitations in time, energy, or movement

Refer to community centers or activity programs

Assist with regular physical activities (e.g., ambulation, transfers, turning, and personal care), as needed

Provide gross motor activities for hyperactive patient

Make environment safe for continuous large muscle movement, as indicated

Provide motor activity to relieve muscle tension

Provide noncompetitive, structured, and active group games

Promote engagement in recreational and diversional activities aimed at reducing anxiety: group singing; volleyball; table tennis; walking; swimming; simple, concrete tasks; simple games; routine tasks; housekeeping chores; grooming; puzzles and cards

Provide positive reinforcement for participation in activities

Assist patient to develop self-motivation and reinforcement

Monitor emotional, physical, social, and spiritual response to activity

Assist patient/family to monitor own progress toward goal achievement

BACKGROUND READINGS:

Glick, O.J. (1992). Interventions related to activity and movement. In G.M. Bulechek & J.C. McCloskey (Eds.), Symposium on nursing interventions. *Nursing Clinics of North America, 27*(2), 541-568.

MacNeil, R., & Teague, M. (1987). *Aging and leisure: Vitality in later life*. Englewood Cliffs, NJ: Prentice-Hall.

McFarland, G.K., & McFarlane, E.A. (1997). *Nursing diagnosis and intervention*. (3rd ed.) St. Louis, MO: Mosby.

Warnick, M.A. (1985). Acute care patients can stay active. *Journal of Gerontological Nursing, 11*(3), 31-35.

McFarland, G.K., & McFarlane, E.A. (1997). *Nursing diagnosis and intervention*. (3rd ed.) St. Louis, MO: Mosby.

Warnick, M.A. (1985). Acute care patients can stay active. *Journal of Gerontological Nursing, 11*(3), 31-35.

Source: Dochterman, J.M. & Bulechek, GM (Eds.) (2004). Nursing Interventions Classification (NIC), 4th ed. St. Louis: Mosby.

NOC outcome

Endurance--0001

Domain-Functional Health (I)
Class-Energy Maintenance (A)
Scale(s)-Severely compromised to Not compromised (a) and Severe to None (n)

Care Recipient:
Data Source:

DEFINITION: Capacity to sustain activity

OUTCOME TARGET RATING: Maintain at_____ Increase to_____

		Severely compromised	Substantially compromised	Moderately compromised	Mildly compromised	Not compromised	
ENDURANCE OVERALL RATING		**1**	**2**	**3**	**4**	**5**	
INDICATORS:							
000101	Performance of usual routine	1	2	3	4	5	NA
000102	Activity	1	2	3	4	5	NA
000103	Rested appearance	1	2	3	4	5	NA
000104	Concentration	1	2	3	4	5	NA
000105	Interest in surroundings	1	2	3	4	5	NA
000106	Muscle endurance	1	2	3	4	5	NA
000107	Eating pattern	1	2	3	4	5	NA
000108	Libido	1	2	3	4	5	NA
000109	Energy restored after rest	1	2	3	4	5	NA
000112	Blood oxygen level	1	2	3	4	5	NA
000113	Hemoglobin	1	2	3	4	5	NA
000114	Hematocrit	1	2	3	4	5	NA
000115	Blood glucose	1	2	3	4	5	NA
000116	Serum electrolytes	1	2	3	4	5	NA
		Severe	**Substantial**	**Moderate**	**Mild**	**None**	
000110	Exhaustion	1	2	3	4	5	NA
000111	Lethargy	1	2	3	4	5	NA
000118	Fatigue	1	2	3	4	5	NA

1st edition 1997; Revised 3rd edition

OUTCOME CONTENT REFERENCES:

Ades, P.A., Ballor, D.L., Ashikaga, T., Utton, J.L., & Streekumaran Nair, K. (1996). Weight training improves walking endurance in healthy elderly persons, *Annals of Internal Medicine, 124*(6), 568-572.

+Dartmouth Primary Care Cooperative Information Project. (1987). *COOP Charts*. Hanover, NH: Department of Community and Family Medicine, Dartmouth Medical School.

Ellis, J.R., & Nowlis, E.A. (1994). *Providing nursing care within the nursing process* (5th ed.). Philadelphia: J.B. Lippincott.

Johns, M.E. (1991). Activity and exercise. In S. Wingate (Ed.), *Cardiac nursing: A clinical management and patient care resource* (pp. 141-145). Gaithersburg, MD: Aspen.

Lubkin, I.M. (2002). *Chronic illness: Impact and interventions* (5th ed.). Boston: Jones & Bartlett.

Potter, P.A., & Perry, A.G. (2001). *Fundamentals of nursing* (5th ed.). St. Louis: Mosby.

Pugh, L.C., & Milligan, R. (1993). A framework for the study of childbearing fatigue. *Advances in Nursing Science, 15*(4), 60-70.

Tiesinga, L.J., Dassen, T.W.N., & Halfens, R.J.G. (1996). Fatigue: A summary of the definitions, dimensions, and indicators. *Nursing Diagnosis, 7*(2), 51-62.

Titler, M.G. (2001). Activity intolerance. In M. Maas, K. Buckwalter, M. Hardy, T. Tripp-Reimer, M. Titler & J. Specht (Eds.), *Nursing care of older adults: Diagnoses, outcomes & interventions* (pp. 324-336). St. Louis: Mosby.

Topf, M. (1992). Effects of personal control over hospital noise on sleep. *Research in Nursing and Health, 15*(1), 19-28.

Source: Moorhead, S, Johnson, M. & Maas, M. (Eds.) (2004) Nursing Outcomes Classification (NOC), 3rd ed. St. Louis: Mosby.

Appendix Two:
Variable definitions: Conceptual and operational

VARIABLE	CONCEPTUAL *DEFINITIONS*	OPERATIONAL *DEFINITIONS*	DATA SOURCE (TO BE FILLED IN BY USER)
PATIENT CHARACTERISTICS			
Demographic Characteristics of Patients	Age, gender, ethnicity, marital status, religion, occupation, etc.	Age = number of years when admitted Gender = male, female, not determined Ethnicity= white, black, Hispanic, American Indian or Alaskan Native, Asian or Pacific Islander, other Marital Status= married, single, separated, widowed, and divorced Religion=Protestant, Catholic, Other faiths, and None/no preference Occupation=retired, working, homemaker, not retired/not working	
CLINICAL CONDITION:			
Medical Diagnoses	Principal and secondary illnesses of patients treated by physicians	Principal = ICD-9 diagnosis code Secondary = ICD-9 diagnoses codes	
Nursing Diagnoses	Human response to illness	NANDA diagnoses	
Severity of illness	Extent of physiological decompensation or organ system loss of function.	APR – DRG Score of 1 (minor) to 4 (extreme)	
TREATMENTS:			
Nursing Interventions	Any treatment, based on clinical judgment and knowledge that a nurse performs to enhance patient/client outcomes	NIC Intervention	
Nursing Activities	Discrete nursing actions for each NIC intervention	NIC activities	
Pharmacological Treatments	Medications used in care of patients during an acute episode of care	Total dose and total number of doses for each *American Hospital Formulary Service Category.*	
PATIENT OUTCOMES:			
Nosocomial Infections	Infections (all sites) acquired during hospitalization for each patient (presence or absence of)	Presence/absence of. Total Nosocomial Infection Rate (calculated) = (# of total nosocomial infections ÷ # of patients) x total patient days	
• Urinary Tract Infections	Presence/absence of UTIs in patients with urinary catheters for each patient	Presence/absence of. Nosocomial UTI Rate = (# of UTIs ÷ # of patients with foley catheters) x total days of catheterization	
• Nosocomial Pneumonia	Development of inflammation of the lungs, with exudation & consolidation, during hospitalization	Presence/absence of. Nosocomial Pneumonia Rate = (# of pneumonias ÷ number of patients) x total patient days	
• Nosocomial Surgical Wound Infection	Infections in surgical wounds within 30 days after the operative procedure	Presence/absence of. Nosocomial Surgical Wound Infection Rate = (# of surgical wound infections ÷ # of surgical wounds) x total patient days	
• Nosocomial Intravascular Site Infections	Infections at site of an intravascular device (arterial line, central venous line, peripheral IV, peripherally inserted central catheter)	Presence/absence of. Rate of Nosocomial Intravenous Site Infection = (# of IV site infections ÷ # of IV devices) x total patient days	
Mortality	Number of patients who die following admission to the hospital	Presence/absence of. Mortality Rate = (# of deaths ÷ total # of patients admitted) x total patient days	

VARIABLE	CONCEPTUAL *DEFINITIONS*	OPERATIONAL *DEFINITIONS*	DATA SOURCE
PATIENT OUTCOMES (CONT'D):			
Adverse Incidents	Number of adverse incidents experienced by patients during hospitalization	Rate of Adverse Incidents = (# of adverse incidents ÷ total # of patient admission) x total patient days	
• Medication Errors	Errors in administration of medications	*Number and type of;* Rate of Medication Errors = (# of medication errors ÷ total # of medications administered) x total patient days	
• Falls – *Fracture* – *Head injury*	Assisted and unassisted falls	Presence/absence of. Rate of Falls = (# of falls divided by number of patients) x total patient days	
Complications	Onset of additional diseases or conditions during hospitalization	Presence/absence of each complication. Total Complication Rate (calculated) = (# of cardiac arrests, CVA, DVT, MI, pneumothorax, PE, tissue/organ injury) ÷ total # of patient admission) x total patient days (NOTE: rate for each complication will also be calculated)	
Unplanned Readmission Rate	Unplanned admission within 10 days after discharge from a UIHC acute admission and related to the same diagnosis addressed during the prior admission	Presence/absence of. Unplanned readmission rate = # of unplanned readmissions ÷ total # of admissions	
Satisfaction			
• Willingness to recommend hospital to others	Level of agreement that patient/family would recommend hospital to others	Score of 1 to 5 on Satisfaction Questionnaire.	
• Overall satisfaction with care	Level of satisfaction with healthcare received.	Score of 1 to 5 on Satisfaction Questionnaire	
Total Length of Stay	Duration of inpatient hospitalization (admission date to discharge date)	Number of acute care days	
Cost per Case	Hospital costs, procedure costs, and physician costs per acute episode of care will be added to arrive at a total cost adjusted for the most current fiscal year	*Cost in dollars* per episode of acute care.	
Individual Outcomes	A behavior or perception that is measured along a continuum in response to a nursing intervention	NOC outcome label and scale ratings	
UNIT CHARACTERISTICS			
Nursing Unit of Service	Type of unit(s) patient receives care in during hospitalization	Unit code(s). See attached for our suggestion grouping.	
Supply and Demand for Nursing Care	CareGiver Patient Ratio (CGPR)	*Hours of Nursing Care Available, Designated by Skill Mix* --- *Hourly Patient Census and Number of Patient Admissions, Discharges, and Transfers*	
Derived Nursing Hours Per Patient Day	Number of hours of care needed per patient day for each unit (Demand for nursing care)	Unit's average daily CGPR x 24 = derived HPPDs	
Skill Mix of Caregivers	Type and Number of caregivers each day (Supply of caregivers)	Proportion of RNs to other (LPNs, NAs) personnel delivering patient care	

Appendix Three:
Types of nursing units

1. General Medicine Unit: provides general medical care of the internal organs (cardiology, endocrine/metabolic, pulmonary, gastrointestinal, urology, renal).

2. General Surgery Unit: provides care to patients who receive surgical treatment for diseases of the bowel, gallbladder, stomach, and other digestive organs.

3. Specialty Surgery Unit: provides care to non-general surgery specialties such as Ear/Eye/Nose & Throat, opthamology, oral surgery, neurosurgery, urology, orthopaedics, organ transplant, oncology, plastic reconstructive, dermatology, vascular, cardiovascular/cardiothoracic.

4. Specialty Medicine Unit: provides specialty care such as bone marrow transplant, dermatology, opthamology, neurology, hematology, oncology, orthopaedics, otolaryngology, renal dialysis, and medical psychiatry. Also includes non-ICU intermediate care units that provide cardiac monitoring and mechanical ventilatory support to patients with medical problems.

5. Specialty Medicine/Specialty Surgical Inpatient Unit: cares for a combination of specialty medicine/specialty surgery patients.

6. Emergency Unit : emergency room or trauma center .

7. Peri-Operative Unit: includes the operating room, post operative recovery room(s), and second stage recovery facility.

8. Adult Psychiatry Unit: provides treatment to adult patients with psychiatric disorders that include mood disorders, psychotic disorders, eating disorders, dementia, personality disorders, substance abuse disorders.

9. Intensive Care Unit: includes Medical ICU, Surgical ICU, Cardiovascular ICU

10. Other: units that do not logically fit into the other nine categories.

Appendix Four:
Example of data format request

Suggestions for how data comes from original data source to investigative team

1. Data files in Excel format are preferred.

2. All the DATE type fields, such as date the patient was admitted, are in the length of 8 characters, and in the order of century (2 characters), year (2 characters), month (2 characters) and day (2 characters)—CCYYMMDD.
   e.g.
   DATE ADMITTED 19990712 -- patient was admitted on July 12, 1999,

3. No hyphen in patient ID numbers.
   e.g.
   PATIENT ID# 12345678

4. No space, hyphen (-), slash (/ and \ ), quotes (‘ and “) and sign (*, $, %, & and #) in any character fields. Basically, values in character fields have no sign or space in between.

5. Break any array data fields into the smallest units, and make those smallest units into separate data fields.
   e.g.
   Break one single Address array data field into five data fields—Street, Apt, City, State, and Zip.

   Address (Street, Apt, City, State, Zip)
   100 Main St, 2, Iowa City, IA, 52242
   →
   Street 100 Main St
   Apt 2
   City Iowa City
   State Iowa
   Zip 52242

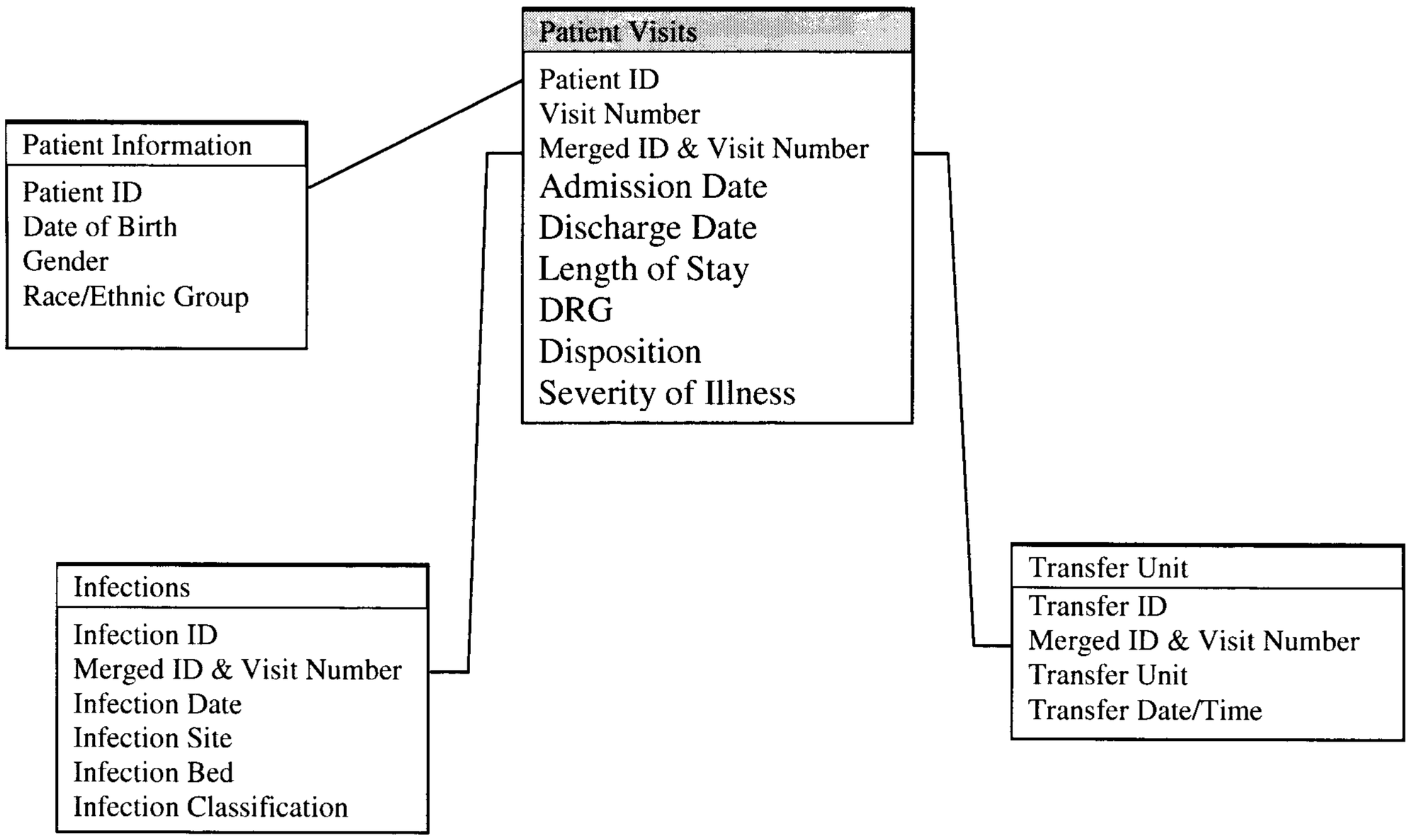
6. Numerical data is rounded to no more than the fourth digit after the decimal point.
   e.g.
   PATIENT WEIGHT 150.4325

7. Financial data and any data of dollars and cents are in numeric data type with 9 digits for dollars and 2 digits for cents.
   e.g.
   TOTAL ACTUAL CHARGES TO DATE ($) 1599.32

Appendix Five:
Example of data tracking form

Data Base	Date of Data Requested	Date of Data Received	Data Req Matches Data Rec'd?	Data Field Report / Dictionary	Reviewed By Team	Overall Status	Use for Analysis
Pt. Satisfaction							
Sat00							
Sat98							
MRA							
98-01 data							
CHF							
HIP FX							
01-02 data							
CHF							
HIP FX							
Fall (PID, Visit, Admit, Discharge)							
98-01 data							
01-02 data							
Finance (TSI)							
Caregiver/Pt. Ratio							
98-01 data							
01-02 data							
NIS							
Incident Report							
Infection Control							
Pharmacy							
PharmAdm							
PharmNet							
Lab							
Allergy							
Census							

Appendix Six:
Example of simple database



Notes:
A single-variable key for each visit was created by merging the Patient ID & Visit Number, which together uniquely identify the visit. This was done to simplify the relationships between tables.

Only a small number of potential tables and variables are included in this example.

Appendix Seven:
Example of a data codebook

Table: Infection

SQL Server				Original Data		
Key	**Field name**	**Data type**	**Length**	**Original Name**	**Description**	**Option value**
PK	PID	Char	8	InfPatientID	Patient ID	8-digit number
PK	VisitNum	Char	5	InfVisit	Visit Number	5-digit number
PK	InfInfectDate	smalldatetime	10	InfInfectDate	Infection Date	MM/DD/YYYY
PK	InfSiteCode	Int	4	InfSiteCode	Site Code	See section Site Code - "Lookup_Site_Code_Tables"
	InfKeySequence	Int	4	InfKeySequence	Key Sequence - Indicates the sequence of infection for multiple infections in the same patient during the same visit.	1 - First Infection; 2 - Second Infection, and so forth.
	DOB	smalldatetime	10	PatBirthDate	Date of Birth	MM/DD/YYYY
	PtGender	Char	1	PatSex	Gender	'F' - Female 'M' - Male
	AdmitDate	smalldatetime	10	VisAdmitDate	Date when admitted to hospital	MM/DD/YYYY
	DischDate	smalldatetime	10	VisDischargeDate	Date when discharged	MM/DD/YYYY
	PrincipalICDDx	Varchar	10	VisPrincipalICDDx	PrincipalICDDx - ICD-9 Code	ICD-9 codes

Appendix Eight:
Example from a data dictionary

Average care-giver patient ratio: RN (Average CGPR RN) for an entire visit

Conceptual Definition: For an entire visit, the average number of all hourly CGPR RN values for the visit. The hourly CGPR RN values are calculated by dividing the total RN hours for a one-hour period by the total patient hours for that same one hour time period. Patient acuity is not accounted for.

Operational Definition: For each hour of the visit, calculate:

$$\frac{\text{Total \# of RN hours for that one hour time period}}{\text{Total \# of patient hours for that same hour}}$$

And then calculate:

$$\frac{\text{Sum of hourly CGPR RN values for the entire visit}}{\text{Total visit hours for which hourly CGPR RN values are available}}$$

The resultant values for all visits within the sample are ranked and divided into quartiles (1^{st}, 2^{nd}, 3^{rd}, or 4^{th} quartile). These categories were coded and used in the analyses as follows:

Quartile	Percentile	Code Used in the Regression
Fourth	76-100%ile	4
Third	51-75%ile	3
Second	26-50%ile	2
First	1-25%ile	1

The ranking/assignment to quartiles allows the investigator to concurrently use both the Average CGPR RN variable as well as the CGPR RN dip variable in the analysis in a meaningful way. It also allows for non-linearity in the relationship between the independent variable, Average CGPR RN, and the dependent variable, thereby allowing a more detailed look at the effect of the independent variable on the dependent variable.


THE UNIVERSITY
OF IOWA