METHODOLOGY FOR DATA SET DEVELOPMENT

by

Laura K. Heermann Langford

A dissertation submitted to the faculty of
The University of Utah
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

College of Nursing
The University of Utah

May 2002
Copyright © Laura K. Heermann Langford 2002

All Rights Reserved
SUPERVISORY COMMITTEE APPROVAL

of a dissertation submitted by

Laura K. Heermann Langford

This dissertation has been read by each member of the following supervisory committee and by majority vote has been found to be satisfactory.

Chair: Eeryl B. Thompson

Todd Allen

Sue E. Huether

Stanley Huff

Rita Snyder-Halpern
I have read the dissertation of Laura K. Heermann in its final form and have found that (1) its format, citations, and bibliographic style are consistent and acceptable; (2) its illustrative materials including figures, tables, and charts are in place; and (3) the final manuscript is satisfactory to the supervisory committee and is ready for submission to The Graduate School.
ABSTRACT

The use of computers for multidisciplinary patient documentation has increased rapidly over the last 15 years. Each computerized patient care documentation system includes a predetermined set of data that is collected and stored based on the individual patient's clinical course and needs. To make the best use of the data collected, it is desirable for computerized documentation systems to use free text as minimally as possible. Coded data are easier to search and use for future analyses. The purpose of this study was to develop and formatively evaluate a methodological approach for use when developing data sets for computerized patient care documentation systems. The study identified knowledge engineering techniques that could be useful during data set development. Four knowledge acquisition techniques were identified as particularly applicable to data set development. These included literature review, expert interview, data analysis and focus groups. These identified techniques were formalized into a methodology for data set development. The methodology was presented in document form including text, diagrams and examples. Evaluation of the development methodology document was done by expert review and solicitation of expert opinion of the methodology via a questionnaire. The research questions addressed the clarity and usefulness of the newly proposed methodology. Eleven experts in data set development and/or data set use participated in the evaluation. The study sample represented a broad range of knowledge and experience in data set development and use.
The results of the questionnaire and the comments received from the reviewers indicated the Data Set Development Methodology (DSDM) was fairly well defined and is useful for data set development. The mean scores for all questions asking about the clarity and the usefulness of the document were above 5 (range=1-7, with 1 = strongly disagree, 7 = strongly agree). These reviewer scores support the formalization of this DSDM for data set development. The DSDM provides clinical data set developers of various skill levels and resources, a documented, consistent, structured and economical tool for the creation of their needed and/or desired data set.
This manuscript is dedicated to my husband, Patrick E. Langford and to my parents, Dale and Betty Heermann. Their constant love and support has been invaluable. I also truly appreciate and treasure the support of other family members, friends and teachers who have helped me throughout the years.

It is only with the support of others we truly obtain our goals.
# TABLE OF CONTENTS

ABSTRACT ........................................................................................................................................ iv

LIST OF TABLES .......................................................................................................................... x

LIST OF FIGURES ..................................................................................................................... xii

Chapter

I. THE PROBLEM .......................................................................................................................... 1

  Introduction ............................................................................................................................... 1
  Background ............................................................................................................................... 7
  Purpose of Patient Care Documentation ........................................................................ 7
  Problems With Paper Patient Care Documentation ....................................................... 7
  Benefits of Computerized Patient Care Documentation .............................................. 8
  Free Text versus Coded Data .............................................................................................. 8
  Challenges Faced With Paper to Computer Conversion .............................................. 10
  Previous Data Set Development Methods .................................................................... 13
  Knowledge Engineering Techniques ............................................................................ 13
  Statement of the Problem ............................................................................................. 15
  Purpose of the Study ....................................................................................................... 16
  Objectives ........................................................................................................................... 16
  Research Questions ......................................................................................................... 17
  Definition of Terms .......................................................................................................... 17
  Significance of Study ....................................................................................................... 17

II. REVIEW OF RELATED LITERATURE .............................................................................. 19

  Existing Data Sets ............................................................................................................... 20
  Purposes of Data Sets ....................................................................................................... 22
  Examples of Problems With Current Data Sets ............................................................ 24
  Past Data Set Development Techniques ...................................................................... 27
  Knowledge Acquisition Techniques for Data Set Development .................................. 34
  Evaluation ............................................................................................................................ 36
    Importance of Evaluation .............................................................................................. 36
    Methodology Evaluation ............................................................................................... 36
    Potential Methods for Evaluation of a Methodology ................................................... 39
  Summary .............................................................................................................................. 44
III. METHODOLOGY AND PROCEDURES

Formalization of Data Set Development Method ................................................. 45
Evaluation of the Methodology ............................................................................ 47
Sample .............................................................................................................. 47
Instrumentation ............................................................................................... 51
Instrumentation Evaluation .............................................................................. 53
Setting .............................................................................................................. 55
Data Collection ................................................................................................ 55
Data Analysis ................................................................................................... 56
Ethical Considerations ....................................................................................... 56

IV. RESULTS ................................................................................................................... 58

Research Question 1: Clarity of the Proposed Methodology ................................. 58
  The Methodology ............................................................................................. 59
  The “Data Set Development Methodology” Flow Diagram ............................ 61
  The “Application of Knowledge Acquisition Techniques for Formalization of Data Set: Flow Diagram” ....................................................................................... 63
  Literature Review Techniques ........................................................................ 64
  Data Analysis Techniques ............................................................................... 66
  Expert Interview Technique ............................................................................ 67
  Focus Group Technique .................................................................................. 68
  General Overview ........................................................................................... 70

Research Question 2: Usefulness of the New Methodology for Data Set Development ..................................................................................................................... 72
Suggestions for Validation and Distribution of the Data Set Development Methodology ..................................................................................................................... 74
Summary of Results ............................................................................................... 76

V. DISCUSSION AND RECOMMENDATIONS .............................................................................. .79

Discussion of Significant Findings ....................................................................... 79
Study Limitations ............................................................................................... 81
Recommendations for Future Research ................................................................. 82
  Further Development Needed ........................................................................ 82
  Application of the Methodology ................................................................... 83
  Further Evaluation ......................................................................................... 83
Implications for Clinical Informatics and Nursing .................................................. 86
Appendices

A. DATA SETS FOUND IN THE LITERATURE WITH PURPOSE, METHOD OF DEVELOPMENT, METHOD OF EVALUATION AND REFERENCE LISTED FOR EACH ................................................................. 89

B. CODING AND CLASSIFICATION SYSTEMS ................................................................. 100

C. GENERAL CLINICAL DATA SETS .............................................................................. 104

D. SPECIALIZED CLINICAL DATA SETS ...................................................................... 107

E. DATA SET DEVELOPMENT METHODOLOGY ......................................................... 110

F. LETTER OF INVITATION TO PARTICIPATE AS AN EXPERT REVIEWER OF THE PROPOSED DATA SET DEVELOPMENT METHODOLOGY ......................................................... 164

G. DATA SET DEVELOPMENT METHODOLOGY SURVEY ........................................ 166

H. MATRIX RELATING SURVEY QUESTIONS TO THE RESEARCH QUESTIONS ................................................................. 174

I. EVALUATION FORM FOR "DATA SET DEVELOPMENT METHODOLOGY SURVEY" .............................................................................. 178

J. DATA SET DEVELOPMENT METHODOLOGY SURVEY DRAFT .......................... 182

K. SUMMARY OF FINDINGS FROM "EVALUATION FORM FOR 'DATA SET DEVELOPMENT METHODOLOGY SURVEY'" ................................................................. 188

L. DATA SET DEVELOPMENT METHODOLOGY SURVEY NARRATIVE RESPONSES .............................................................................. 197

REFERENCES .................................................................................................................. 205
# LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Methods Used for Data Set Development</td>
<td>30</td>
</tr>
<tr>
<td>2. Basic Development Techniques With List of Data Sets Using Each Technique</td>
<td>31</td>
</tr>
<tr>
<td>3. Journal Databases Searched With Listing of Topics Covered by Each</td>
<td>38</td>
</tr>
<tr>
<td>4. Roles Held By Study Participants During Past Data Set Development Projects</td>
<td>51</td>
</tr>
<tr>
<td>5. Mode, Mean, Standard Deviation and Range of Scores Evaluating the Data Set Development Methodology in General</td>
<td>60</td>
</tr>
<tr>
<td>6. Mode, Mean, Standard Deviation and Range of Scores Evaluating the “Data Set Development Methodology” Flow Diagram</td>
<td>62</td>
</tr>
<tr>
<td>7. Mode, Mean, Standard Deviation and Range of Scores Evaluating the “Application of Knowledge Acquisition Techniques for Formalization of Data Set” Flow Diagram</td>
<td>63</td>
</tr>
<tr>
<td>8. Mode, Mean, Standard Deviation and Range of Scores Evaluating the Literature Review Technique</td>
<td>65</td>
</tr>
<tr>
<td>9. Mode, Mean, Standard Deviation and Range of Scores Evaluating the Data Analysis Technique</td>
<td>66</td>
</tr>
<tr>
<td>10. Mode, Mean, Standard Deviation and Range of Scores Evaluating the Expert Interview Technique</td>
<td>68</td>
</tr>
<tr>
<td>11. Mode, Mean, Standard Deviation and Range of Scores Evaluating the Focus Group Technique</td>
<td>69</td>
</tr>
<tr>
<td>12. Mode, Mean, Standard Deviation and Range of Scores Evaluating Redundancy and Organization of the DSDM Document</td>
<td>70</td>
</tr>
</tbody>
</table>
13. Mode, Mean, Standard Deviation and Range of Scores for Questions Evaluating the Usefulness of the Data Set Development Methodology ................................ 73

14. Frequency Potential Method of Distribution Chosen by Reviewers ...................... 76
## LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Coded documentation of wheezing heard in upper airways</td>
<td>11</td>
</tr>
<tr>
<td>2. Coded documentation of decreased breath sounds in bilateral bases</td>
<td>11</td>
</tr>
<tr>
<td>3. Process for formalization of data set development method</td>
<td>46</td>
</tr>
</tbody>
</table>
CHAPTER I

THE PROBLEM

Introduction

The use of computers for multidisciplinary patient documentation has increased rapidly over the last 15 years. Each computerized patient care documentation system includes a predetermined set of data that is collected and stored based on the individual patient's clinical course and needs. The goal of many of the computerized systems is to use free text as minimally as possible and to have the majority of data coded for ease of computer search and analysis (Hannah, Ball, & Edwards, 1994a).

Computerized patient documentation systems are often modeled after existing paper records. The content of these paper documents is largely dictated by tradition and history (Hannah et al., 1994a). Elements included in paper documentation systems have evolved over the years to include elements important to the care provided by several different disciplines including nursing, medicine, respiratory therapy, nutrition, physical therapy and many others (Hannah et al., 1994a; Shortliffe & Barnett, 1990). Often the elements included may not be specific to the individual patient's needs but are collected merely because "that is what has always been done" or "just in case we need them." When care practices change based on new research findings or clinical evidence, data elements are added, but rarely are any taken away from the patient care documentation.
As Andolina (1998) writes in *Nursing Informatics: Where Caring and Technology Meet*, (Ball, Hannah, Newbold, & Douglas, 1995), “part of the difficulty with documentation has been the historical perception that more was better” (p. 181).

Using the paper record as a model for computerized documentation systems may be a mistake. Hannah and colleagues (1994a) purport “nurses should only be gathering data that is essential for nursing decisions about patient care” (p. 84). This may be extreme, as nurses are the primary data collectors of patient data for quality assurance and legal purposes as well as research, but the point of not “over collecting” data is valid. Paper documentation systems are often lengthy, repetitive and full of information collected but never used. In addition, hand-written documentation often lacks structure (Lee, Martinez, Rutledge, & Maull, 1998). Because gleaning useful information from a narrative paper record is difficult, clinicians and others interested in patient data develop “flow sheets,” “spreadsheets” and “reports” specific to their interests. This results in multiple points of data entry and the repetition of common data points. Over time, with changes in policies, procedures and care provision, some of these multiple data entry points may lose their importance and/or usefulness and may no longer be needed at all. In paper documentation systems, however, no consistent ongoing system exists to review the use of data elements and see which ones are always left blank and never used. Without this type of evaluation it is unknown what data elements are necessary and used for decision making during patient care.

A big advantage of using computers for patient care documentation is their ability to manage and manipulate data and information. Computers allow data to be displayed in many formats and permit displaying a single piece of data in more than one place at one
time or on multiple reports. Computers can prompt for missing data and evaluate the data for accuracy at the time of data entry, as well as provide reports in a timely and readable manner (Andolina, 1998; Metzger, 1995). Many of these characteristics have been desired in the pen-and-paper documentation arena but have been difficult to accomplish. Recognizing the differences in the functionality of computer and paper tools and realizing that the model for one is not optimal for the other is important.

One method used to circumvent the inadequacies of the paper documentation system model when creating a computerized documentation system has been the development of “data sets” and “minimum data sets.” A data set is a related group of data elements that can be manipulated as a unit (Woodcock et al., 1994). Coenen, McNeil, Bakken, Bickford and Warren (2001) cite the International Organization for Standardization (ISO) definition of a data element as “a unit of data for which the definition, identification, representation and permissible values are specified by means of a set of attributes (ISO 1119-1)” (p. 243). Data sets facilitate the recording, storage and transmission of computerized data across organizational and geographic boundaries (Harris, Graves, Solberg, Elkin, & Chute, 2000). A minimum data set is the smallest number of data elements with uniform definitions and classifications that is to be collected for a specified purpose (Hannah et al., 1994a). A minimum data set is the smallest set of data that will fulfill the identified purpose. “They are called minimum because they represent the fewest points believed necessary to capture and convey the essence of required information” (Harris et al., 2000). Data sets are more global and generally do not have as narrowed of focus as minimum data sets. A minimum data set could actually be contained within a data set.
Appendix A contains 26 different data sets found and described in the literature. The name of each data set is listed along with an acronym for the data set if available and at least one reference is listed for each. Most of the data sets listed in Appendix A are considered “minimum data sets.” Other sets listed do not claim to be the “minimum” set required but rather aim to provide uniform specifications of data used by their intended field or audience (National Center for Injury Prevention and Control, 1997).

It is important to differentiate between data sets and classification or coding schemas. Classification is the “systematic arrangement in groups or categories according to established criteria” (Webster, 1985). Classification systems are used to organize similar health care terms and concepts into related categories (Gorden, 1998). Nursing classification systems are used to describe the nursing process and facilitate data aggregation for comparisons within and across sites (Henry, Warren, & Zielstorff, 1998). A coding system is a structured concept-based terminology used to support the capture, retrieval, analysis and transfer of electronic data (College of American Pathologists, 1999; Huff et al., 1998). Appendix B lists several coding and classification schemas found in the literature.

Several different groups developed the data sets in Appendix A. In addition, each data set was developed for different purposes. Some of the data sets were developed with a general-purpose use in mind (Appendix C), whereas others were developed for specific purposes (Appendix D). A stated purpose found in the literature associated with each data set is also listed for each data set in these appendices.

Many of these data sets have overlapping elements and inconsistent definitions or level of detail. For example, the Nursing Minimum Data Set (NMDS) (Werley, 1988) is
intended to collect the core minimum data elements needed to describe nursing practice. This data set is, however, a very high-level data set, containing only 16 elements organized in three categories. This example highlights one of the problems with minimum data sets. The descriptions of the purpose of the data sets are so weak; the user is left without direction on specifically how the minimum data set will be used.

As evidenced by Appendix C and Appendix D, data sets have been developed for a variety of reasons. Unfortunately none have been developed with the primary intention being to document patient care. Data sets used in computerized patient care documentation tend to come from the previously existing paper record—not from a thoughtful, methodological approach. Current data sets used in computerized patient care records are often designed to meet billing and quality assurance needs, but other purposes such as supporting clinical decision making and providing a complete, legal documentation of care are not a primary focus.

Additional drawbacks to many current data sets are the methods used to develop them. Development of these data sets occurred either via governmental organizations, national expert steering committees, national invited conferences or lengthy repeated surveys and/or Delphi studies. The development methods used are very time consuming, lengthy and expensive. The use of governmental organizations, national steering committees or invited conferences for data set development requires large amounts of funding to gather and host even a small number of people to work on developing a data set. The time between calling such a meeting, holding it, refining the work and disseminating the product is significant. Surveys and Delphi studies might be slightly cheaper but still consume a considerable amount of resources for development, mailing,
collection and analysis. Years of intermittent work and “review periods” can go into one data set! Many of the data sets listed in Appendix A continue to go through further development and refinement.

These methods of data set development are respectable and may produce a “reliable” and “valid” data set based on the criteria presented by the developing agency or special interest group. In the eyes of an individual interested in building a charting package that meets the needs of patient care, clinical communication and several vested interest groups, however, the data sets produced are often too limited in scope, incomplete, contradictory to each other and too confusing to be of use.

The problem addressed in this study is the lack of an efficient, cost-effective method for development of a data set for use in computerized patient care documentation systems. The premise is that data sets should be based on data elements needed to support patient care and/or required by regulatory agencies but not include unnecessary data elements. The purpose of this study was to develop and formatively evaluate a methodological approach for use when developing data sets for computerized patient care documentation systems. The methodological approach was to examine and include as appropriate methods that are efficient and cost-effective and that produce useful data sets. This chapter includes background information about challenges faced in the conversion from paper patient care documentation to computerized documentation, previous data set development methods and an introduction to knowledge engineering techniques. The problem and purpose are then discussed in detail prior to presentation of the research questions. The importance of the study is addressed at the end of the chapter.
Purpose of Patient Care Documentation

Patient care documentation has many purposes. McDonald and Barnett (1990) credit Lawrence Weed with the first analysis of the purposes of the medical record. These leaders determined the purposes of the medical record can be classified in three different functional areas: a) Facilitate patient care, b) serve as a financial and legal record and c) aid in clinical research (McDonald & Barnett, 1990). Although these functional areas are distinct, it can be argued the primary purpose of the medical record is for communication. Clinical communication may occur between caregivers in the same department providing immediate care to the patient as a team or to clinicians who care for the patient in the future. Patient care documentation is used for communicating information regarding a patient’s treatment to any interested party whether it is the patient, clinical researcher, risk manager, regulatory agency, insurance company or lawyer. The qualitative difference between the uses is timing of collection, the discrete or aggregate presentation of the data and the personal or anonymous nature of the data (Nelson, 1997).

Documented patient care data also assist and support decisions made by clinicians when providing patient care. A detailed yet concise, accurate presentation of relevant data adds value to the limited time a clinician has with a patient by assisting the clinician to derive a clinical impression or decision (Nelson, 1997).

Problems With Paper Patient Care Documentation

Paper patient care documentation systems are often inefficient and of inconsistent quality. A significant amount of health care resources are spent on the creation, storage
and retrieval of information in redundant, time-consuming, ineffective efforts to
document patient care (Metzger, 1995). Even with the considerable amount of time,
energy and resources, the patient care record is often little help to the clinician trying to
provide timely, appropriate and efficient health care (Metzger, 1995). With paper patient
care records, it is not uncommon for health care providers to spend more time finding and
aggregating the patient’s data than in providing direct patient care (Metzger, 1995).

Benefits of Computerized Patient Care Documentation

The use of computers enhances the possibility of improving the uses of patient
care documentation for more advanced decision support and for better, more easily
conducted research. Computerized patient care documentation systems can improve data
collection by automating data collection from patient monitoring devices, fluid pumps
and other devices used in health care. Computers can also improve communication
between clinicians by providing increased access, not only by more than one person at a
time but also at more than one location (from bedside to remote). Computers allow
flexible formatting depending on the user’s needs. Computers can improve accuracy,
legibility and readability of data. Computers can also improve timeliness of data
availability, improve organization and provide higher quality, more efficient access to the
data for individual or aggregate review (Metzger, 1995).

Free Text Versus Coded Data

Paper records often contain large amounts of unstructured, narrative writing
which is considered “free text.” “Free text” is similar to natural language and does not
have format or word choice restrictions (van Bemmel & Musen, 1997). When narrative free text is stored in computers, words and concepts are strung together in such a way the computer cannot differentiate one word or concept from another but sees these data as one large "blob" of characters. This form of data storage makes retrieval of specific pieces of information very difficult and in some cases impossible. Without structure the computer is not able to search, sort or index data. Therefore, all data stored in free text are retrieved in the same large piece as they are entered. In addition, free text data entry is mostly done via keyboard, which is not only time consuming but also allows more room for typing and grammatical errors. These types of errors decrease the accuracy and readability of computerized records.

A solution to the problems of free text data is coded data. "Coded data" refers to data elements that are represented by a preassigned string of alphanumeric characters that a computer can recognize, identify or recall at any given time (van Bemmel & Musen, 1997). Coded data allow computers to provide some of the most desired functionality such as automated data collection, efficient retrieval and aggregated data review. A common method of representing coded data is with check boxes or pick lists. These check boxes or pick lists contain possible values for common data elements used in the computerized system.

In health care, lists can be made of assessment findings, interventions, medications and a multitude of other values for common data points. An example of data points commonly used in documenting breath sounds after a patient assessment includes "normal, diminished, absent, wheezes, crackles/rales, stridor, rhonchi and E to A change/bronchial breath sounds." These data elements can be preprogrammed into the
system as choices to be used when documenting a patient's breath sound assessment. For example, when a nurse listens to a patient's breath sounds and hears wheezes in the upper airways and decreased breath sounds in the bases, his or her charting using a pick list with these items available would look like Figure 1 and Figure 2. The * indicates the items chosen by the nurse when documenting her findings.

**Challenges Faced with Paper to Computer Conversion**

Nurses face several challenges in the conversion from paper to computer charting. These challenges stem from the change in medium used for documentation, the difficulty of coding patient care data and the costs. Changing from paper to computer creates the challenge of using different configurations to collect, present and retrieve data, as well as high costs for developing, implementing and supporting hardware and software.

Computers provide much greater flexibility in collection, presentation and retrieval of data than paper records. However, the amount of data that can be displayed on one screen at a time is limited. The amount of screen space available is often referred to as "real estate." The real estate available from screen to screen is often influenced by mandatory elements that must occur on each screen. For example, an institution may require patient name, age, ID number and current date and time to be visible on all screens. Navigation tools are required to be displayed either as buttons or commands for keystrokes. The overall size of the display unit also impacts the display of data considerably. None of these considerations was a factor in paper documentation. With paper documentation, developers could use larger sheets of paper that could be unfolded
<table>
<thead>
<tr>
<th>Breath Sounds</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>R Apex</td>
</tr>
<tr>
<td>Diminished</td>
<td>RUL</td>
</tr>
<tr>
<td>Absent</td>
<td>RML</td>
</tr>
<tr>
<td>*Wheezees</td>
<td>RLL</td>
</tr>
<tr>
<td>Crackles/Rales</td>
<td>R Base</td>
</tr>
<tr>
<td>Stridor</td>
<td>L Apex</td>
</tr>
<tr>
<td>Rhonchi</td>
<td>LUL</td>
</tr>
<tr>
<td>E to A Change/</td>
<td>LLLL</td>
</tr>
<tr>
<td>Bronchial</td>
<td>*Bilateral Apex</td>
</tr>
<tr>
<td></td>
<td>*Bilateral Upper Lobes</td>
</tr>
<tr>
<td></td>
<td>Bilateral Lower Lobes</td>
</tr>
<tr>
<td></td>
<td>Bilateral Bases</td>
</tr>
<tr>
<td></td>
<td>Throughout Lung Fields</td>
</tr>
</tbody>
</table>

*indicates the items chosen by the nurse when documenting her findings.

Figure 1. Coded documentation of wheezing heard in upper airways

<table>
<thead>
<tr>
<th>Breath Sounds</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>R Apex</td>
</tr>
<tr>
<td>*Diminished</td>
<td>RUL</td>
</tr>
<tr>
<td>Absent</td>
<td>RML</td>
</tr>
<tr>
<td>Wheezes</td>
<td>RLL</td>
</tr>
<tr>
<td>Crackles/Rales</td>
<td>R Base</td>
</tr>
<tr>
<td>Stridor</td>
<td>L Apex</td>
</tr>
<tr>
<td>Rhonchi</td>
<td>LUL</td>
</tr>
<tr>
<td>E to A Change/</td>
<td>LLLL</td>
</tr>
<tr>
<td>Bronchial</td>
<td>Bilateral Apex</td>
</tr>
<tr>
<td></td>
<td>Bilateral Upper Lobes</td>
</tr>
<tr>
<td></td>
<td>Bilateral Lower Lobes</td>
</tr>
<tr>
<td></td>
<td>*Bilateral Bases</td>
</tr>
<tr>
<td></td>
<td>Throughout Lung Fields</td>
</tr>
</tbody>
</table>

*indicates the items chosen by the nurse when documenting her findings.

Figure 2. Coded documentation of decreased breath sounds in bilateral bases
for use and folded into the desired storage size or could add additional sheets of paper.

The addition of more screens in a computerized charting package is not the same as
adding more sheets of paper. With paper the user can hold them side by side to see
groups of data as they please; with computer screens the user does not have that
flexibility and is forced to view only the data available on one screen at a time. The
change in layout and organization of data for computer screens can be quite a challenge
for developers and users accustomed to paper patient documentation.

Using coded data to collect, present and retrieve data is desirable, but the
challenge in creating the coded data set for use in patient care documentation is
formidable. Coding health care data is difficult due to the large amount of data involved
in patient care documentation, problems with consensus on what should be included and
what should not be included and what words or definitions should be used to describe a
finding or action. For coded data to be most useful the words and terms used must be
clear, concise, relevant and never extraneous. This is difficult to do in health care where
concepts are not always well defined and consensus is not always present or easily
reached. In addition, health care is complex and it is often difficult to assign and
maintain codes that are meaningful in unique patient situations. Other challenges include
ensuring codes are represented, interpreted and used as they are intended.

The cost of computerized patient care documentation is sometimes the biggest
challenge of all. The reader must be remembered that the cost of collecting, managing
and storing patient care information is high regardless of the method, computer or paper,
used. Unfortunately, computerized documentation systems are not cheap. Costs for
computerized records include hardware, software, building space and human support.
Large costs are involved with designing and developing the user interface, the database for storage of patient information, backup mechanisms, "downtime" contingencies and seemingly continual updates of both hardware and software. Very few health care administrators would argue against the benefits of computerized patient care documentation. The difficulty is allocating the large number of dollars in one or two years' budgets for the computerized system. This huge upfront cost is sometimes seen as more formidable than a current paper system which, although expensive, drains the costs not entirely from the monetary budget but also takes some of its toll on intangible costs such as personnel frustration and the quality of patient care.

Previous Data Set Development Methods

The literature indicates several different methods are used for developing data sets. For the 26 data sets listed in Appendix A, the methods used for development are mentioned within the literature only for 15 of them. Six different methods were used either singularly or in some combination for these 14 data sets. These methods include conferences, committees, expert focus groups, literature review, surveys and case reviews. The information provided regarding data set development does not support one accepted way of developing a data set nor even suggest a preferred method.

Knowledge Engineering Techniques

The inconsistent nature of data set development could be improved by looking at techniques used in other areas of software development. Knowledge engineering is a development technique that has long been associated with the decision support specialty
in health care informatics. Examination of knowledge engineering components, however, indicates parts of the process may work well for clinical data set development.

Chignell and Peterson (1988) define knowledge engineering very broadly stating, “The process of building an expert system is generally referred to as knowledge engineering” (p. 383). Chignell describes the steps of building an expert system as the following (Parsaye & Chignell, 1988):

1. Feasibility Analysis
2. Conceptual Design
3. Knowledge Acquisition
4. Knowledge Representation
5. Validation
6. Technology Transfer and Maintenance

Other authors focus the definition of knowledge engineering further. Greer (1992) defines knowledge engineering as “converting the knowledge about facts and rules and their use into a computer program or encoding the context and process knowledge of human experts” (p. 249). Haley and Williams (1986) write, “The job of gathering and encoding the knowledge and experience of human experts is called ‘knowledge engineering’” (p. 83). Other authors (Buchanan & Shortliffe, 1985; Mulsant & Servan-Schreiber, 1984; Patton, 1985; Stefik et al., 1983) dance around the definition, but all seem to include in their description of knowledge engineering, if not their definition, steps 2, 3 and 4 of the Parsaye and Chignell (1988) list.

The use of knowledge engineering techniques, specifically those associated with knowledge acquisition, could be very useful during the development of data sets for use
in patient care documentation systems. Knowledge acquisition, the third step of knowledge engineering as described by Parsaye and Chignell (1988), is the process of acquiring required knowledge to perform a task from sources such as human experts, case histories and reference sources. Previous data set developments have included some of these practices, such as literature review and expert focus groups. Knowledge acquisition techniques are usually referred to as being useful to determine "knowledge" to be included in an expert system, but they could be used at a more granular level to assist with determining data elements to be included in a computerized patient care documentation system. A structured approach using literature review to formulate a rough structure and basic "starter" data set that could be refined by experts’ interviews and focus groups could streamline the data set development process while not compromising its integrity. The application of knowledge engineering techniques to clinical and administrative data sets also facilitates and promotes evidence-based practice and decision making (Rodrigues, 2000).

**Statement of the Problem**

The problem addressed in this study is the lack of an efficient, cost-effective method for development of a data set for use in computerized patient care documentation systems. The methodology of data set development has not been studied. The lack of concrete direction in data set development methodology has resulted in various techniques being applied as each developer has the resources for or sees fit. The assumption is that data sets should include only data points needed to support patient care and/or required by regulatory agencies, but not include unnecessary data points.
Data set development is not a well-defined process, but it is being addressed from multiple clinical perspectives to many different degrees. Data sets are needed for unit specific documentation systems (such as the emergency department), specialty specific documentation systems (such as critical care) and large multidisciplinary hospital-wide documentation systems. A cost-effective, well-defined method applicable to data set development for these varied instances is needed.

**Purpose of the Study**

The purpose of this study was to develop and formatively evaluate a methodological approach for developing data sets for computerized patient care documentation systems. The study identified knowledge engineering techniques that could be useful during data set development. These identified techniques were formalized into a methodology for data set development. Evaluation of the development approach was done by expert review and solicitation of expert opinion of the data set development approach via a questionnaire.

**Objectives**

The objectives for this study were as follows:

1. To develop a methodology for data set development that does not require an expensive national invitational conference of domain experts but results in a useful data set for documenting patient care in a local setting. The methodology should be generalizable to multiple settings for others to use when developing a data set.

2. Evaluate the data set development methodology via expert review.
Research Questions

The research questions addressed by this study were as follows:

Is the proposed methodology clearly defined?

Is the new methodology useful for data set development?

Definition of Terms

Terms needing definition and clarification in the context of this project include a) clearly defined and b) useful. For this project the following definitions were used:

1. **Clearly defined** means being plainly described, understandable and without doubt or obscurity.

2. **Useful** was defined as having practical purpose, advantage, helpfulness or good effect.

Significance of Study

Development of computerized patient care documentation systems is increasing rapidly while health care budgets are decreasing. With the expanding use of computers for patient care documentation, the demands for improving the use of the patient data collected are also increasing. In order to keep up with the demand for accurate, concise, data sets, developers need a more cost-effective, efficient data set development method. The traditional, popular data set development methods of invited workshop or government mandate used in the past are simply too slow and expensive for the current needs of developers. Knowledge engineering techniques commonly used in development of computerized decision support systems may provide a faster and less expensive alternative to data set development. This study will define a methodological approach to
data set development using knowledge engineering techniques and will evaluate the proposed method to determine its clarity and usefulness to clinical data set development. An alternative data set development methodology as proposed in this study would be very useful for clinicians of all disciplines who are needing a data set to meet their clinical documentation needs and yet do not have extensive time or resources.
CHAPTER II

REVIEW OF RELATED LITERATURE

The purpose of this study is to develop and evaluate a methodological approach for creating data sets for computerized patient care documentation systems. A broad approach was used to review the related literature to ensure past data set development and evaluation methodologies are known, critiqued and considered. First the investigator searched for all existing data sets published in health-care-related journals referenced in Medline®, CINAHL® and HealthStar®. Special attention was paid to the purposes, development methods and evaluation methods of each data set. The lack of clearly defined data set development methodologies and evaluation techniques led the investigator to research alternative methods for data set development and evaluation. The review of related literature presented in this chapter begins with an overview of existing data sets and their purposes and identified strengths and weaknesses. Next, a review of current data set development techniques and potential alternative development techniques is presented. Following the review on data set development techniques is a review of current and potential data set evaluation methods. The chapter concludes with the presentation of an opportunity available to develop and evaluate a methodological approach for developing data sets for computerized patient care documentation systems.
Existing Data Sets

A concerted effort was made to find all clinically related data sets published in the medical literature. Publication of descriptions of data sets promoted for use by various groups for the purposes of standardizing data began in the early 1970s, but it was not until the 1990s that the pace of data set development accelerated. Appendix A includes the approximate year each of the 25 data sets found in the literature were developed. No approximate development date was found for the Ambulatory Medical Care Minimum Data Set.

The earliest data set was the Uniform Hospital Discharge Data Set created in the 1970s by the U.S. Government for the purpose of documenting information about patients receiving Medicare reimbursement (McCormick, Renner, Mayes, Regan, & Greenberg, 1997). In 1972 and 1974 the Uniform Ambulatory Care Data Set and Emergency Medicine Services Minimum Data Set (respectively) were created to promote standardized record keeping (McCormick et al., 1997). The Nursing Minimum Data Set was created in 1985 to describe the practice of nursing (Werley, 1988). It was not until 1991 that a data set is reported as being created to “aid in the delivery of health care” by the Canadian Hospital Medical Records Institute (Hannah, Ball, & Edwards, 1994b p. 63). Creating data sets for the purpose of reimbursement and/or standardizing data is one of the most common impetuses for data set development as seen in Appendix C and Appendix D. As noted earlier, many of these data sets are considered “minimum” data sets or are strictly focused for a specific use such as regulating Medicare/Medicaid payees.
Not all data sets are published for review and use by others. However, *ALL* clinical information systems used for patient care documentation have a data set of some sort. The size and level of detail may vary from system to system depending on its uses, but any set of data elements combined in a database for the purpose of collecting data would be considered a data set. It is largely unknown how the data elements included in most of these data sets were chosen. Some may be combinations of published data sets, some may be developed by vendors interested in selling their product and some may be entirely “home grown” or developed by the site where the data set is used or any combination of these. These data sets are often a working and evolving part of the system, with items being added when identified as missing.

Data sets for use with computerized patient care documentation systems have been in existence since the beginning of such systems. Development of clinical information systems began in the mid-to late-1960s (Collen, 1995). The Health Evaluation and Logical Processing (HELP) System is one example of the very early computerized systems used in collecting and documenting data related to patient care (Gardner, 1990). The HELP system was initially installed at LDS Hospital in Salt Lake City, Utah, in the late 1960s. Although it is not described specifically in the literature or distributed, a very specific data set was determined for use in this system to enable the system to meet its purpose of information management. The data set is a large, all-encompassing set of data elements used to record patient care, to support the care provided, to meet regulatory requirements or to contribute to research. Additionally, the HELP system is well known for its provision of decision support to clinicians at the bedside (Wiederhold & Perreault, 1990). A clearly defined data set used within the
programming of this system enabled the system to analyze and evaluate specific data elements and generate patient (data) specific warnings, alerts, diagnostic suggestions and limited management advice (Wiederhold & Perreault, 1990). Without the specifically defined data set included in the HELP system, none of the other applications that make the HELP system so highly regarded would be possible.

**Purposes of Data Sets**

The intended purpose of a data set has a strong influence on its development, size and contents. Data sets can be created to standardize data, provide guidelines, measure performance, promote communication and support clinical care. Each of the data sets listed in Appendix A was developed for a specific purpose. Detailed descriptions of the data sets and their purposes can be seen in Appendix C and Appendix D. A concise purpose "label" was assigned to each of the data sets based on the detailed purpose statement. These concise purpose labels can be seen associated with each respective data set in Appendix A. Generally the purposes stated for these data sets are to standardize the data that are collected and to provide guidelines for submitting data to a government agency. Providing guidelines for submission of data is essentially the same as standardizing data, but in this instance they are separated to indicate the significant number of data sets developed specifically for submission of data to government agencies.

The purpose of the majority of these data sets is primarily to promote uniformity and standardization of data. Six of the data sets, HCFA 1500, UDS, HRMI, LTC MDS, McData and UHDDS, were developed by governmental agencies such as the U.S.
Department of Health and Human Services (DHHS), Health Care Financing Administration (HCFA) and National Center for Vital Health Statistics (NCVHS) for the direct purpose of reporting data to their agencies for either Medicaid/Medicare reimbursement or for regulation purposes. Only three of the data sets listed, HEDIS, NMMDS and OASIS, have the direct purpose of measuring performance and outcomes for use in quality assurance and improvement activities.

Standardized and uniform data are important for data transfer, sharing data and comparison of data across sites (Renner & Swart, 1997). These traits are important for governmental agencies if they are to use data from several health care providers for monitoring purposes. Data points included in data sets intended for uniformity and standardization of data are useful for tracking trends and resource utilization, but they are often a small snapshot specifically related to the needs of the monitoring agency and do not reflect the complete clinical picture of a patient encounter. “Standardized” data do not guarantee adequate data for documenting all information that is involved in patient care and its related decision making.

The purposes listed for the identified data sets are quite different from the purported goals and purposes of computerized patient care documentation systems. As stated in Chapter 1, the primary purpose for a patient care record is for communication between caregivers and others involved in the provision of the patient's care (Shortliffe & Barnett, 1990). Computer-enhanced documentation offers to promote better communication by increasing accessibility and improving accuracy, legibility and readability. Additionally, computers enhance the possibilities of improving patient care through decision support applications and more easily conducted research. Although the
standardization of data is important to meeting many of the goals and purposes of computerized patient care documentation, it is not enough to ensure the data set is adequate to support clinical decision making. The complete data set used in a computerized patient care documentation system directly influences how well the system will meet its goals. If the goal of the system is to document and/or support or perhaps even improve patient care, then a data set needs to be developed with that purpose in mind. The resulting data set may very well include one or more of the smaller data sets listed in Appendix A but can certainly not be limited to one of these data sets. A close examination of the data sets described in the literature indicated a variety of data sets have been developed, but none are intended for the documentation of care.

**Examples of Problems With Current Data Sets**

Major problems with current data sets are their limited descriptions of how they are to be used and their limited scope. The NMDS is a good example of a data set with little description of how it is to be used. Werley and Lang (1988) state that the NMDS is to be used to describe nursing practice. This is a very general, uninformative description. To what or who exactly is the data set to describe the practice of nursing? To the government? Clients? Other nurses? Other disciplines? And how exactly is the data set to be used to describe nursing practice?

The three categories in the NMDS include Nursing Elements, Patient Demographic Elements and Service Elements (Hassett & Farver, 1995). These three categories included 16 data elements. Some of these data elements are a single datum (such as gender or date of birth) where others may be complex, poorly defined data items.
(such as nursing intervention or outcome). It is difficult to appreciate how 16 data elements of differing complexity are supposed to describe the intricacies of nursing practice. The developers of the NMDS did not provide much direction for implementation or use of the data set, but did leave the interpretation up to the user. The developers of the NMDS had a very good idea to develop a data set to describe the practice of nursing. Unfortunately, the lack of direction on how to implement the data set may result in users being unable either to implement the data set or to implement it in such a variety of ways that its purpose of standardization of nursing data is defeated.

Additionally, the NMDS does not include clinical assessment factors. Data elements related to a nursing or clinical physical exam or assessment are not included in the NMDS. Without assessment data, there is no avenue to ascertain how or why other data elements such as the nursing diagnosis or the nursing outcome (two data elements that are included in the NMDS) are determined. Although assessment data were initially addressed by a specific task force of the original NMDS development conference, it was later eliminated. Westlake (1988) writes in the textbook entitled *Identification of the Nursing Minimum Data Set:*

> Given the diffusiveness of assessment data and the frequent overlap between assessment and the other activities of the nursing process, as well as the joint task force recommendation that the list of the diagnoses of the North American Nursing Diagnosis Association (NANDA) be used in the NMDS, the Nursing Assessment Task Force recommended that assessment data not be included in the NMDS. (p. 386)

It is amazing to this author to think the workgroup members considered that NANDA-described nursing diagnoses were sufficient to define nursing assessments. It is interesting to note, however, that, although not included in the final cut of the NMDS, the nursing assessment task force of the NMDS workgroup did develop a 16-category
classification scheme to organize the collection of assessment data across nurse practice settings (Westlake, 1988). Westlake (1988) reports the Nursing Assessment Task Force of the NMDS workgroup did recommend that a classification scheme should drive the collection and organization of all nursing data collected via nursing information systems and the NMDS, but that assessment data should not be included as such in the NMDS.

As the NMDS is an example of a data set with a weak description of how it is to be used, the DEEDS data set is an example of a data set with a limited scope. The data set entitled *Data Elements for Emergency Department Systems (DEEDS)* was developed by an interdisciplinary workgroup to promote uniform specifications for data elements that may be included in emergency department computerized documentation systems (National Center for Injury Prevention and Control, 1997). Representatives from six professional associations and three federal agencies convened in 1994 and 1995 to develop DEEDS Version 1.0. This version was then discussed in a public forum including 160 participants representing 35 professional associations and 12 federal agencies. Version 1.0 of DEEDS contains 156 data elements in eight sections. Of the 156 data elements, approximately 106 are demographic/administrative and 50 are directly related to clinical care. This lack of clinically related data elements makes the DEEDS data set inadequate to meet the needs of patient care documentation.

DEEDS attempts to identify and define specific data elements used in emergency care such as gender, heart rate and name of medications given. Each data element included in DEEDS has a definition, detailed specification (such as data type and field length), a description of possible repetition of the data element and field values designating valid data entries. The overall data set does not include, however, specific
items related to assessment data or test result data such as breath sounds clear or a chest x-ray finding of consolidation. The limited scope of DEEDS was intentional. The workgroup never intended for DEEDS to be a complete data set to record a patient’s visit to the emergency department (ED). The workgroup was more interested in providing a guideline to collecting uniform data elements that could be shared nationally and reused after they have been used for immediate patient care and administrative purposes (DEEDS Writing Committee et al., 1998). However, the demand for a more complete data set is pushing the workgroup for DEEDS Version 2.0 to add many more elements to the data set.

Although incomplete, DEEDS provides a much better map for the computerized patient care documentation system developer to use when determining what data elements to include in the database than the NMDS, but DEEDS is still far from including all data elements needed even for its defined ED focus. The NMDS and DEEDS data sets are helpful for standardizing data, but their design and purpose make them inadequate to meet the needs for defining a complete and useful data set for a comprehensive multidisciplinary computerized patient care documentation system.

**Past Data Set Development Techniques**

Descriptions of the techniques used for individual data set development are often sketchy or not available at all. The development techniques for data sets of legacy documentation packages such as the HELP system are not well documented in the literature. Traditionally, these data sets were developed based on paper documentation previously used in the developing institutions. Data sets developed by researchers
associated with a university or a private organization, such as NMDS (Ryan & Delaney, 1995; Werley & Lang, 1988), NMMDS (Huber, Schumacher, & Delaney, 1997) and UDSHCH (Pace, 1995) often have more published than those developed by governmental agencies such as UDS, UACDS and UHDDS (McCormick et al., 1997; Renner & Swart, 1997). This difference may be related to the drive for publications at universities that does not exist within government agencies. Since information was not available in the literature describing development techniques for all data sets included, a presumption was made about the techniques used based on the group organization or agency who developed the data set. For example, if the developing agency for a particular data set is the National Committee on Vital and Health Statistics (NCVHS) and no development method is described, the data set is presumed to have been developed by a committee.

Techniques used for previous data set development are varied. Appendix A lists each of the data sets found in the literature with a short description of the technique(s) used for their development. Evaluation of the development techniques listed in Appendix A indicates some commonalities in the methods used. Collapsing similar methods into a single approach leaves only seven basic techniques previously used in data set development. These basic techniques include: conference (consensus, invited and/or working, workshop), committee (public/private sponsorship, government sponsorship or steering), expert focus group (hearing), literature review, survey, case review and Delphi Study. A “committee” may be the same as a “conference” in the way that work is conducted. The true difference may lie in the number of people involved with each group (a “conference” being larger than a “committee”). However the number of people involved in either instance is not mentioned for any of the data sets, nor is a detailed
description provided as to how the work of developing a particular data set was conducted. For this review a committee and a conference will remain separated.

Table 1 presents these seven basic methods and the frequency with which the methods were used for the 25 data sets and a brief definition of each method. Table 2 presents a more detailed breakdown of the seven basic methods used for each data set. For several of the data sets more than one method was used for development. The methods used, the order of use or the combinations of methods as reported in the literature were not consistent across data sets. For those data sets with more than one technique, the techniques are listed in the order used. For example, for the Core Health Data Elements first a Committee was convened, then a survey was conducted and finally a Focus Group was used to validate the data set.

Understanding how each of these methods might be used to develop a data set is important. Because the literature has limited descriptions of the methods and processes used for existing data sets, the use of these methods must be surmised. Data sets could be developed by a conference or committee by having the group “brainstorm” as a group and then determine by consensus which items should be included. A focus group could act like a committee and use brainstorming and consensus techniques or, perhaps the facilitator of the group could present them with certain questions, activities or even a potential list of data elements for them to review and critique. The literature review technique could be used to determine a data set by looking in literature related to the topic of interest and finding the pertinent data elements to be included. The survey technique could be used to develop the initial set of data elements to be included by
### Table 1

**Methods Used for Data Set Development**

<table>
<thead>
<tr>
<th>Method of Development</th>
<th>Frequency</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conference</td>
<td>5</td>
<td>A meeting of two or more persons for discussing matters of a common concern or a representative assembly for the formal interchange of views*</td>
</tr>
<tr>
<td>Committee</td>
<td>5</td>
<td>A body of persons delegated to consider, investigate, take action or report on some matter*</td>
</tr>
<tr>
<td>Expert focus group</td>
<td>5</td>
<td>A group of individuals considered experts in the particular topic of interest assembled to discuss or answer questions on a given topic**</td>
</tr>
<tr>
<td>Literature Review</td>
<td>3</td>
<td>A critical summary of research on a topic of interest**</td>
</tr>
<tr>
<td>Survey</td>
<td>2</td>
<td>To query someone or a group of people, in order to collect data for the analysis of some aspect of a group or area of interest*</td>
</tr>
<tr>
<td>Case Review</td>
<td>3</td>
<td>A thorough, in-depth analysis of an individual, group, institution or other social unit; could also include the clinical record of a health care visit.*</td>
</tr>
<tr>
<td>Delphi Study</td>
<td>1</td>
<td>A method of obtaining judgments from a panel of experts. The experts are questioned individually; then a summary of these judgments is circulated to the entire panel. The experts are questioned again, with further iterations introduced as needed until there is consensus.**</td>
</tr>
<tr>
<td>Unknown</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>
| Unknown. Presumed to be (governmental) committee | 8 | *(Webster, 1985)*  
** (Polit & Hungler, 1999b)
Table 2

*Basic Development Techniques With List of Data Sets Using Each Technique*

<table>
<thead>
<tr>
<th>Conference</th>
<th>Committee</th>
<th>Expert Focus Group</th>
<th>Literature Review</th>
<th>Survey</th>
<th>Case Review</th>
<th>Delphi Study</th>
<th>Unknown (presumed to be committee)</th>
<th>Unknown</th>
</tr>
</thead>
<tbody>
<tr>
<td>AMC MDS</td>
<td>Core Health Data Elements</td>
<td>Core Health Data Elements</td>
<td>NM-MDS</td>
<td>Core Health Data Elements</td>
<td>EMS MDS (Ver. 2)</td>
<td>Rotor Wing Transport MDS</td>
<td>EMS MDS (Ver. 1)</td>
<td>MQIS</td>
</tr>
<tr>
<td>DEEDS</td>
<td>UB-92</td>
<td>NMMDS</td>
<td>PCDS</td>
<td>PNDSS</td>
<td>EMDS</td>
<td></td>
<td>HCFA 1500</td>
<td>RAI</td>
</tr>
<tr>
<td>EMDS</td>
<td>McData</td>
<td>NMDS</td>
<td>PNDSS</td>
<td></td>
<td>PCDS</td>
<td></td>
<td>HEDIS 3.0</td>
<td></td>
</tr>
<tr>
<td>NMDS UDSHC H</td>
<td>OASIS</td>
<td>PCDS</td>
<td>UHDDS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*See Appendix A for more information and references for each data set*
asking one or several people for their opinion on which data elements should be included or it could be used as a technique to refine a rough data set by asking qualified individuals their opinion on the data set and the items included in it. Case reviews could be used to determine the data elements by reviewing individual cases related to the area of interest and determining what data elements are needed for inclusion in a particular data set. The Delphi Study technique would require an initial data set to be determined using another technique but could be useful in refining a data set of interest. To use the Delphi Study technique the data set of interest would need to be reviewed and critiqued or judged by a qualified group of people, their feedback summarized then sent back to the group at least once and maybe even two or three times in an effort to gain consensus.

Each of these techniques has the potential of being useful in data set development, but there are pros and cons to each. Conferences, committees and focus groups have the advantage of having several people involved with the data set development process allowing for different knowledge bases, points of view, experiences and opinions to be considered. Working with large groups however can be difficult and progress may be hampered in an effort to gain consensus if willingness for compromise is not evident in the group. Literature reviews are useful for determining data elements for inclusion based on published standards of practice and other requirements but are often limited to the views of only a few people (the authors) or organizations who have taken the time or had the support to publish their findings. One of the best reasons to use literature when determining data elements for a data set would be the access to evidence-based literature of research reports regarding patient outcomes. Case reviews are useful for determining data elements by looking at what data elements are included in each case. Limitations of
case review are related to the number of cases reviewed. The fewer cases reviewed, the less likely a complete data set will be obtained. Each case is dependent on circumstances specific to the case and therefore findings from one case may not be generalizable to other cases. One solution is to review several cases, but then again the question of how many cases are sufficient is not easily answered. The Delphi Study technique is useful for the refinement of a data set. It works best if a “skeleton” data set is already determined via one of the other previously mentioned methods so the panel of experts included in the study have something to place judgment on to work towards consensus. Thompson and Schaffer (2000) have had good success using the Delphi technique when refining an air transport data set but have found some difficulty in collecting a sufficient number of responses from panel members in a timely manner when conducting the Delphi study via e-mail.

Unfortunately it is not known if the data sets developed by any of the methods listed in Table 1 result in the best data set possible or even the correct data set for the intended purpose. The majority of the data set development described in the literature relied heavily on conferences, committees and expert focus groups. The description of these specific development methods in the literature indicated each of these activities were done on a national level, including members from all over the United States and not just local to the development site. These methods of data set development are expensive and time consuming. Evidence is not available to support the conclusion that the data sets generated from these workshops or committees are the correct or “best” data sets for their intended purpose and worth their high cost. It is unfortunate more is not published regarding the development of many of these data sets. Without such information
confidence in the resulting data sets is compromised and the ability to replicate the
development process, thereby increasing the confidence in the data set, is all but
impossible. The actual use of these data sets is also unknown.

Knowledge Acquisition Techniques for Data Set Development

Knowledge acquisition pertains to acquiring the pertinent domain knowledge
from various sources for a specifically intended use (Mulsant & Servan-Schreiber, 1984;
Parsaye & Chignell, 1988; Patton, 1985). The term “knowledge acquisition” is often
confused with knowledge engineering. In the world of expert systems, knowledge
acquisition is considered part of the knowledge engineering process (Buchanan &
Shortliffe, 1985; Parsaye & Chignell, 1988; Stefik et al., 1983). Medical informatics
literature however will often substitute the term “knowledge engineering” when
describing knowledge acquisition (Middleton, Anderson, Fletcher, Masarie, & Leavitt,
medical knowledge, the medical literature (documented knowledge) and experts in a
specific domain (clinical experience). Techniques used for knowledge acquisition
include literature review, expert interview, focus groups and data extraction from
databases such as patient care data repositories and case studies (Mulsant & Servan-
Although it is not described as such in the data set literature, it is evident the methods
used for data set development are very much like those used for knowledge acquisition.
Literature review, focus groups and case review were methods used for the development
of several data sets listed in Appendix A, but there has been no discussion on how these methods are best applied for data set development.

The primary method of extracting knowledge from the medical literature is to conduct an extensive literature review (Guise et al., 1997). Literature reviews include a systematic search for articles, books or other printed or electronic materials related to the topic at hand. Each of the source materials is then read and studied for specific information relevant and presumed to be useful to the topic. A literature review should be conducted by someone with at least basic knowledge of the topic at hand so that relevancy can be determined. Problems encountered with literature reviews include the difficulty of extracting meaningful pieces from large, wordy narrative texts and being confident all related literature is included in the review (Shortliffe, 1990).

Data extraction from experts can be focused on a certain group of national or locally defined experts or clinical staff. The methods of invited workshop and special committees used in current data set development could also be considered methods of knowledge extraction from experts. The expert data extraction is usually seen as an approach for resolving discrepancies or gaps found during the literature review. One advantage of knowledge extraction from experts is it allows access to valuable experiential knowledge that may not be available in any other source (Warner, Sorenson, & Bouhaddou, 1997b). Review of databases is valuable for resolving discrepancies or gaps but may also be helpful in identifying items not previously "discovered" through the literature review or extraction from experts.

With knowledge acquisition for decision support systems, knowledge engineers are interested in extracting the "rules" of a certain discipline or care process. With
knowledge acquisition for data set development, data set developers (or data set engineers as they could be called) are interested in extracting data elements relevant to the desired goals and purposes of the new data set. These goals may include the standardization of data, to support communication between caregivers, to support caregivers decision making (and the care they provide), to measure quality and to improve care or support any other reporting function. Ultimately data set engineers strive for the appropriate data elements to support and guide the user’s need for information and generation of knowledge regarding the topic at hand, whether it be a specific patient’s care or the care provided across patient populations.

Evaluation

Importance of Evaluation

The importance of evaluation is well known to the science of nursing. Since the time of Florence Nightingale nurses have used evaluation techniques to determine how well a program, practice, procedure or policy is working. The need for evaluation in clinical informatics and data set development is also true. Without evaluation we cannot be sure the methods for data set development and the resulting data sets meet the goals and purposes for which they are developed.

Methodology Evaluation

A large part of this study is the formulation of a methodology for the development of clinical data sets. Because the evaluation of any new procedure or method is important to determine its effectiveness, it was necessary to evaluate the newly developed method.
A literature review of methodology evaluation was done to determine how best to conduct this evaluation.

Locating information in the literature regarding evaluation of methodologies was difficult. Table 3 lists 14 different databases searched regarding this topic. Keywords used for the search included “evaluation of methodology,” “methodology evaluation” and “evaluate method.” All searches were limited to English only. These search terms resulted in 0 to over 35,000 results. For those searches with 250 or less results, every citation was evaluated for potential applicability to this study. For those with greater than 250 resulting citations, the first 250 were evaluated and spot evaluations were done for the rest of the list.

Unfortunately, even with the large number of matches to the keywords, very few articles reporting an evaluation of a methodology were found. Most of the citations related to evaluation methodologies (i.e., the method of evaluation) or evaluation of study methodologies. Nine citations were found related to the evaluation of a methodology. Seven of these were found in the Dissertation Abstracts database. Of the nine, three are related to methods of chemical extraction (Attalla, 1986; Han, 1989; Kemple, Ray, & Lipkowitz, 1988), one to a method of crisis intervention in a child care facility (Nunno, 1996), one to a method of bronchoalveolar lavage (Gardiner, 1993), two to methods used in construction (Anderson, 1989; Goodwin, 1987), one to an agricultural method of rating corn (Figueroa-Ruiz, 1997) and one to a method of designing a computer interface (Pritchard, 1994). Methods to evaluate methodologies are not well documented, as evidenced by the extensive literature search. Thus, for this study it was decided by the
Table 3

Journal Databases Searched With Listing of Topics Covered by Each

<table>
<thead>
<tr>
<th>Database Name</th>
<th>Topics Covered</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABI -Inform</td>
<td>Business and Management</td>
</tr>
<tr>
<td>CINAHL</td>
<td>Nursing &amp; Allied Health</td>
</tr>
<tr>
<td>ClinPSYC</td>
<td>Psychology and Related Disciplines</td>
</tr>
<tr>
<td>Dissertation Abstracts</td>
<td>U. S., Canadian, British &amp; European Dissertations</td>
</tr>
<tr>
<td>ERIC</td>
<td>Education</td>
</tr>
<tr>
<td>HealthStar</td>
<td>Health Sciences, Technology, Administration &amp; Research</td>
</tr>
<tr>
<td>MEDLINE</td>
<td>Biomedical</td>
</tr>
<tr>
<td>PAIS</td>
<td>Political, Social &amp; Public Policy Issues</td>
</tr>
<tr>
<td>PsycINFO</td>
<td>Psychology &amp; Related Disciplines</td>
</tr>
<tr>
<td>Sociological Abstracts</td>
<td>Sociology &amp; Related Disciplines</td>
</tr>
<tr>
<td>Wilson Business Abstracts</td>
<td>Business &amp; Related Disciplines</td>
</tr>
<tr>
<td>Wilson General Science Abstracts</td>
<td>Physical, Life &amp; Health Sciences</td>
</tr>
</tbody>
</table>
author in conjunction with the doctoral supervisory committee to evaluate the proposed methodology for its clarity, applicability, appropriateness and usefulness in data set development.

Potential Methods for Evaluation of a Methodology

Issues of reliability and validity are important for scholars to consider when conducting an evaluation. Reliability is defined as “the extent to which an experiment, test or measuring procedure yields the same results on repeated trials” (Carmines & Zeller, 1979 p. 11; Webster, 1985 p. 11). Reliability is the measurement of an instrument’s consistency in measuring its intended attribute (Burns & Grove, 1993a; Polit & Hungler, 1999a). These definitions are not easily related to a methodology, because it is not considered to be “measuring” anything (as a questionnaire would). From these definitions it could be implied successful repetitive use of a methodology would indicate the methodology’s reliability.

Nursing research texts explain validity as the degree to which an instrument measures what it is intended to measure (Burns & Grove, 1993a; Polit & Hungler, 1999a). The dictionary (Webster, 1985) gives the definition of “well grounded or justifiable, being at once relevant and meaningful.” Determining the definition of validity in relation to methodologies is difficult. Again, methodologies are not measuring anything, but to be valid they must be relevant, meaningful and useful.

Brinberg and McGrath (1985) purport “validity has to do with truth, strength and value... validity is like integrity, character or quality to be assessed relative to purposes and circumstances [sic]” (p. 13). Their Validity Network Schema (VNS) relates validity
to different phases or stages within the research process. Brinberg and McGrath identify different meanings for validity in each of the different phases of the research process. These dimensions are “validity as value or worth”, “validity as correspondence or fit” and “validity as robustness or generalizability” (Brinberg & McGrath, 1985).

The VNS is based on the interaction of three domains (substantive, conceptual and methodological) and three stages of research (preparatory, central and generalization) (Brinberg & McGrath, 1985). The substantive domain includes the area of research that focuses on real-world phenomena. The conceptual domain includes abstract representations of the substantive phenomena. The methodological domain refers to techniques used to gather, process and interpret information regarding substantive phenomena (Brinberg & McGrath, 1985).

In the VNS, Stage 1) the preparatory stage, is in research when data are generated, developed, clarified or evaluated in each of the three domains. This is the “validity as value or worth” dimension. In Stage 1, elements and relations are evaluated based on their value, worth, importance or usefulness. Stage 2, the central stage, is the point when elements are combined and relationships are determined to form intermediate or instrumental structures for further investigation or clarification. This is the ‘validity as correspondence or fit’ dimension. In Stage 2, “the degree to which the features of the relations a researcher is examining match across domains” is investigated (Brinberg & McGrath, 1985 p. 27). The final and third stage is the point in research when findings from Stage 2 are investigated for generalizability and credibility. This is usually accomplished by verifying, replicating, extending or delimiting the second stage findings (Brinberg & McGrath, 1985; Grobe & Hughes, 1993). This is the “validity as robustness
or generalizability” dimension. In this stage, the external validity or the range in which
the findings of Stages 1 and 2 hold true are evaluated.

Content validity, a type of validity, is the extent to which an instrument represents
the content of the universe of content intended to be included in the instrument (Polit &
Hungler, 1999a). The “universe” of content is of course influenced by the scope of the
instrument. Lynn (1986) purports the measurement of content validity using a two-stage
process is fundamental to the validation of almost all instrumentation. The two stages
described by Lynn (1986) include the Development Stage and the Judgment-
Quantification Stage. Stage I, the Development Stage includes the following:

1. Identification of full content domain
2. Sampling and item generation
3. Assimilation of items into useable form.

Stage II, the Judgment-Quantification Stage includes the following:

4. Judgment/quantification of content validity of items
5. Judgment/quantification of content validity of instrument.

These stages of content validation could fit the development cycle of a
methodology as well. As described in Stage I, methodologies begin with an identified
domain for which they are intended for use. The sampling and item/concept generation
of a methodology is done through literature review, application of experience by the
methodology developer and expert review and critique. The methodology itself is the
result of assimilating the items/concepts into a useful form presented in an organized
manner.
The Judgment-Quantification Stage, or Stage II as described by Lynn (1986), is also applicable to methodology development. The two steps in Stage II include the evaluation of the content validity of specific items as well as the content validity of the entire instrument. Evaluating a methodology's content validity on these levels is important when determining the methodology's clarity and usefulness.

To test the content validity of the instrument items and the overall instrument itself, Lynn (1986) suggests using a content validity index. The index described by Lynn is a 4-point ordinal rating scale (1 = irrelevant, 4 = extremely relevant). Using this scale, experts rate each item included in the instrument. The content validity of each item is determined by the proportion of experts who rate them as valid (a rating of 3 or 4 on the scale). The instrument content validity is determined by the proportion of total items judged by the experts as being valid. If the instrument is assessed as having content validity based on the scoring but items are determined to have been omitted, then further clarification may be needed (Lynn, 1986).

Lynn describes a 4-point scale as being preferable because it eliminates the ambivalent middle rating common in odd number rating scales (Lynn, 1986). Park and colleagues (1986) however describe a similar process using a 9-point rating scale with success. The 9-point scale described by Park and colleagues is often referred to as the RAND Appropriateness Methodology (Piccirillo et al., 1998) or RAND Formal Consensus Technique (Shiffman, Leape, & Greenes, 1994). The RAND technique requires experts to rate each item included in an instrument on a scale of 1 (extremely inappropriate) to 9 (extremely appropriate). This is an odd-numbered scale and would allow an ambivalent
middle rating, but the larger scale allows for a little more flexibility in scoring and analysis.

The scoring systems described by Lynn (1986) and Park et al. (1986) are both variations of a Likert scale. A Likert scale consists of several declarative statements expressing a point of view. Respondents are asked to rate their agreement or disagreement to the statement on a numbered scale (Burns & Grove, 1993b; Polit & Hungler, 1999b). Likert scales typically offer a 5- or 7-point numbering schema for use by respondents when rating each item. One end of the scale, typically the low end, is labeled "disagree," whereas the other end of the scale, the high end, is labeled "agree."

There are arguments for the simplicity of the smaller scale and for the increased diversity of the larger scale, but no concrete evidence says one is better than another (Burns & Grove, 1993b; Polit & Hungler, 1999b). There are also debates regarding the use of an odd vs. even number of choices. Even numbers do not allow for a middle "undecided" or noncommittal category. The size and design of the scale is left to the discretion of the researcher depending on the needs of the study (Burns & Grove, 1993b; Polit & Hungler, 1999b).

These types of content validation rely upon experts in the domain of the methodology to rate the methodology using a consistent set of questions. Traditionally the number of experts to include in such a study has been somewhat arbitrary, depending on the number of experts available and willing to complete the rating. Lynn (1986) recommends at least a minimum of five experts to control for chance agreement. The maximum number of experts has not been determined but is thought by Lynn (1986) to be unlikely to exceed 10. The number of experts can be determined by applying the
standard error of the proportion (Lynn, 1986). This is done by “calculating the proportion of the number of experts who might agree out of the total number planned for use and then setting the standard error of the proportion to identify the cut off for chance versus real agreement.” (Lynn, 1986 p. 383). Using this formula and desiring a level of significance beyond .05, it is beneficial to include at least six or more experts (Lynn, 1986). The more experts used, the more disagreement between experts is allowed while still maintaining the desired level of significance.

Summary

The literature contains several reports of existing data sets. Each of these data sets were created for distinct purposes and are not necessarily transferable or usable in all domains. The development techniques used for these data sets vary widely. There is no evidence that any one development technique is superior to the others or is worthy of promotion throughout the field. Other methods for data set development such as formalized knowledge acquisition are potentially useful. Investigating the use of these other methods for data set development and the evaluation of these methods would be useful in the effort to define a more efficient, cost-effective method for the development of data sets for use in computerized patient care documentation systems.
CHAPTER III

METHODOLOGY AND PROCEDURES

A significant portion of this work included the formalization of a methodology for data set development. The resulting Data Set Development Methodology (Methodology) is found in Appendix E. The remainder of the project focused on the evaluation of the methodology designed. This chapter begins with an explanation on how the data set development methodology was developed and then proceeds to the method and procedures used for its formative evaluation. The participant sample, instrumentation, procedures, data collection and data analysis are described.

Formalization of Data Set Development Method

The process used to formalize a data set development method is depicted in Figure 3. The initial steps to formalizing a method for data set development included describing the knowledge acquisition process and the stages of data set development. From this information, the different knowledge acquisition techniques were identified and evaluated for their usefulness in data set development. Each technique was then carefully studied, reviewed and described in writing. Any differences in how the technique would be applied to data set development as opposed to the more familiar use of knowledge acquisition processes for expert system development were determined.
A logical process for the application of knowledge acquisition techniques to data set development was developed and carefully arranged in flow diagram form. The flow diagram, along with the definitions, descriptions and application of each technique was compiled into one document, thereby creating a “handbook” for data set creation (Appendix E). This “handbook” is referred to as the Data Set Development Methodology (DSDM) Document. Background information regarding the knowledge engineering process, stages of data set development and the fit of the DSDM into the Systems Development Life Cycle of software development was also included in the DSDM document.
Evaluation of the Methodology

To know if a new methodology is valuable the methodology must be evaluated. The validity of a new methodology must be established before beginning tests of its reliability. The first two dimensions of validity testing as stated by the Brinberg and McGrath Validity Network Schema (Brinberg & McGrath, 1985) fit the initial formative evaluation of a newly defined methodology. The first two dimensions are “validity as value or worth” and “validity as correspondence or fit” (Brinberg & McGrath, 1985). To determine if the methodology is valuable and worthwhile, it is important to know if it is clear, understandable and accurate. This dimension of validity is represented in the first research question of the study, “Is the proposed methodology clearly defined?” The applicability of the methodology to the problem it addresses determines its validity as correspondence and fit. This dimension of validity is represented in the second research question of the study, “Is the new methodology useful for data set development?” To determine the validity of the methodology, experts in data set use and data set development from across the United States were asked to review the DSDM document and provide their feedback via a questionnaire-type survey.

Sample

The experts invited to participate as reviewers of the DSDM methodology were chosen using purposive convenience sampling. Purposive sampling is often not considered a strong sampling method, but in this case where experience and knowledge of the subjects needed to be ensured, it was the best available method. Subjects in this part of the study were not required to be clinicians or working in healthcare, but they must have knowledge of clinical informatics, data sets, data set development and data set
application. Subjects were required to speak and write English and be able to provide knowledgeable responses to questions asked on the survey. Potential subjects included authors of previously published data sets and nationally and/or locally recognized experts in informatics and/or data set development. No awards, rewards or incentives were given for participation.

No gold standard exists for the number of subjects to include in an opinion survey such as this. The number of subjects included is often influenced by the number of candidates available and willing to participate (Lynn, 1986). Lynn (1986) purports a minimum number of five experts are needed to provide a sufficient level of control for chance agreement. A maximum number of judges has not been established, but Lynn states the maximum number needed to determine content validity in similar instances to this study is unlikely to exceed 10. Since there is no gold standard and available potential candidates were limited to the strict qualifications, willingness and schedules of the subjects, a goal of 5-10 participants for this study was set following Lynn’s (1986) guidelines.

Telephone calls, personal contact, or personal e-mail were made or sent to individuals identified by the researcher or doctoral supervisory committee members as meeting the study criteria. Potential participants were identified as experts in data set development and/or data set use by noting their documented involvement (publications, job titles and duties) in data set development and or use. With each contact, the study was explained and the individual was asked if he or she would be willing to participate. If the individual was willing to participate, a cover letter (Appendix F) explaining a) the study, b) the purpose of their invitation to participate in the study, c) the consent process
and d) instructions on how to complete and return the DSDM Survey was sent to him or her along with the DSDM document (Appendix E) and the DSDM Survey to be completed (Appendix G). Participants could choose to have the documents delivered to them via e-mail or U. S. Postal Service. All documents were also posted and available on a web server that was accessible to all participants for download of the documents.

Participants were given a 2-week time period to read the DSDM document and return the completed survey. All individuals agreeing to participate in the study were explicitly informed of this 2-week deadline for return of completed surveys. Completed surveys could be returned to the researcher via e-mail, web form completion or the U. S. Postal Service. Addressed, stamped envelopes to cover costs for the use of the U.S. Postal Service were offered and available to all participants.

At the end of 2 weeks, a reminder notice was sent by e-mail to all individuals agreeing to participate in the study. The message thanked those who had completed a survey and submitted it for their participation and reminded those who had not yet completed and submitted a survey that the deadline had passed. The e-mail message asked for all participants still wishing to participate to submit a completed survey and/or contact the researcher to make arrangements for the timing of the return. As indicated in the cover letter and according to the University of Utah Institutional Review Board (IRB) procedures, the return of the survey served as consent to participate in the study.

Thirty personal contacts were initiated to invite reviewers to participate in the study. Of these 30, 18 individuals agreed to participate in the study, 2 declined and 10 were unable to be reached over a 3-day period. Three individuals requested all documents be sent to them via the U.S. Postal service with an enclosed, stamped
envelope for return of the completed survey. The other 15 requested for all documents to be e-mailed to them. It is unknown if the "electronic" participants downloaded the necessary documents from the web site or via their e-mail.

At the end of the 2-week deadline, eight completed surveys had been received. Three were received via the U.S. Post office, four via the web site and one via facsimile submission. One week after the 2-week deadline reminder notice, three more surveys had been returned. No other arrangements for submitting surveys had been arranged. Data analysis was done on the 11 completed surveys received. This sample represented 37% of the original 30 potential participants identified, and 61% of the 18 participants who had personally agreed to participate in the study.

All participants completing a survey and participating in the study were at least 19 years old. Six participants listed their highest level of education as an MS/MA. Four listed their highest level of education as PhD and one listed his or hers as EdD. Occupational fields of the participants included nursing (9), biomedical engineering/informatics (1) and information systems/informatics (1). The average number of years participants had used clinical data sets was 13.2 \((SD = 8.4)\). The participant group had an average of 9.9 \((SD = 9.1)\) years of experience in developing data sets. The roles held by each participant for the development of data sets are listed in Table 4. Participants could choose more than one past role. The most commonly held role during data set development by this study group was "participant." Other frequently held roles included "project leader" and "coordinator."
Table 4

Table of Roles Held By Study Participants During Past Data Set Development Projects

<table>
<thead>
<tr>
<th>Role</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Participant</td>
<td>9</td>
</tr>
<tr>
<td>Project Leader</td>
<td>6</td>
</tr>
<tr>
<td>Coordinator</td>
<td>6</td>
</tr>
<tr>
<td>Other</td>
<td></td>
</tr>
<tr>
<td>Analyst/Designer</td>
<td>1</td>
</tr>
<tr>
<td>Consultant</td>
<td>1</td>
</tr>
<tr>
<td>Educational Requirement</td>
<td>1</td>
</tr>
</tbody>
</table>

Instrumentation

A questionnaire regarding clarity, understandability, applicability and respondent’s perception of the proposed data set development methodology was used to collect responses from the subjects. The questionnaire was developed by the investigator to query participants about the clarity, understandability and applicability of the DSDM document. The questionnaire developed for data collection for this study was titled “The Data Set Development Methodology Survey” (Appendix G).

Data Set Development Methodology Survey

The Data Set Development Methodology Survey (Appendix G) was a seven-page questionnaire containing 49 questions. Of these 49 questions, 29 were answered via a Likert scale, 14 were open-ended questions allowing for narrative feedback from the respondents, 5 were multiple choice and 1 was dichotomous.

The first page of the survey contained six brief demographic items such as level of education, level of involvement in data set development and level of involvement in data
set use for sample description purposes. No identifying information was requested in the demographic portion of the survey. Verification of each subject’s age being ≥ 19 years old was done to meet IRB requirements. Gender and geographic location of participants were not collected from the respondents of this survey because this information could compromise subject confidentiality and/or anonymity due to the small sample size. It should be noted, there were no obvious variations in responses of the reviewers in relation to their demographic characteristics.

The questions were organized according to subject matter in the DSDM document. The first six questions inquired about the methodology in general. The next 5 questions examined the “Data Set Development Methodology” flow diagram. Five more questions examined the “Application of Knowledge Acquisition Techniques for Formalization of Data Set” flow diagram. The topics of “Literature Review Techniques,” “Data Analysis Techniques,” “Expert Interview Techniques” and “Focus Group Techniques” were each examined by 4 questions. The final 12 questions of the survey referred to a general overview of the Data Set Development Methodology’s generalizability, organization, needed improvements, potential validation and dissemination.

Thirty-nine of the 49 questions on the survey addressed the research questions posed by the study. Thirty-one of the questions on the survey addressed this study’s first research question, “Is the proposed methodology clearly defined?” Eight of the questions on the survey addressed the second research question for this study, “Is the new methodology useful for data set development?” Appendix H contains a matrix matching questions on the survey to the specific research question they address. Six of the
questions were aimed at obtaining demographics to describe the sample. Three questions were directed at gathering opinions from the reviewers regarding potential methods for validation of the methodology, current availability of a similar methodology, as well as the dissemination of this methodology. The final question was an open-ended question allowing the reviewer to provide any additional comments or suggestions not solicited by the other questions.

**Instrumentation Evaluation**

The survey used by study participants to provide feedback on the DSDM document was a new tool developed solely for the purpose of evaluating the DSDM document. Before the use of any new data collection tool the reliability and validity of the measurement tool must be established.

Examining the internal consistency and/or using test-retest procedures tests the reliability of a questionnaire. Both of these testing measures require a representative group to complete the questionnaire. The alpha coefficient (internal consistency) or Cohen's Kappa (test-retest) or Percent agreement (test-retest) is then calculated based on the responses of the representative group. Due to the specialty knowledge required by participants, the length of the DSDM document and the time required to read it and complete the survey, the field of qualified subjects was not large enough to conduct reliability testing of the survey used AND collect data for the evaluation of the DSDM document. Therefore reliability testing of the data collection tool was not done.

Face validity is useful in determining if an instrument measures what it is intended to measure (Nunnally & Bernstein, 1994). Face validity is a cursory inspection of an instrument and, if present, may increase response rates (Waltz, Strickland, & Lenz,
Face validity of the survey was determined by the doctoral committee members of this study and by selected faculty members at the University of Utah College of Nursing.

Content validity is useful in determining if the items included on a measurement tool represent the domain of content addressed by the instrument (Waltz et al., 1991). Content validity can be established by having a small group of area experts (at least two) evaluate the items included on the measurement tool in relation to the objectives of the tool, the relevancy of the items and the adequacy of the items to represent the content of interest (Waltz et al., 1991). Six local clinical informatics experts with experience in data set development and/or use established content validity of the survey. These nurse informaticists evaluated the DSDM survey by completing a second questionnaire (Appendix I). The original survey evaluated by this group is included in Appendix J.

The evaluation questionnaire included in Appendix I consisted of 19 questions. Each of the questions were dichotomous “Yes or No”-type questions with an accompanying option to provide narrative feedback if desired. An additional area for general impressions and/or comments was included at the very end. A summary of responses gathered from the content validity questionnaire in Appendix I is found in Appendix K. The DSDM survey was revised based on the comments and suggestions made by the pretesting panel. Changes made or not made to each question based on the responses to the content validity questionnaire (Appendix I) and the reasoning for each are also found in Appendix K. The final, resulting survey used for the formative evaluation of the DSDM document is found in Appendix G.
The responses from the nurse informaticists group were very helpful in refining the DSDM survey. Based on their suggestions several changes were made to the DSDM survey. These included reorganizing questions, rewriting some questions for clarity, rewriting some questions to break out potentially "double barreled" and difficult to answer items, adding examples to items and verifying information needed to answer questions was available in the DSDM document itself.

**Setting**

The setting for reading the DSDM document and completing the survey was not dictated by the researcher. Each subject determined the setting for the reading of the DSDM document and completion of the survey. The exact time and or place of participation in study activities was not requested or recorded.

**Data Collection**

All completed surveys were handled in a confidential manner. No identifiers were collected on the survey. Although all surveys submitted could be identified via e-mail, IP address, or postmark, the confidentiality of respondents was maintained. Once data were removed from the e-mail, web site or postal envelope no effort was made to maintain any identifiable information with the completed surveys. Raw data were only accessible to the primary researcher and the doctoral committee members involved with this study as needed for analysis. All hard copies of completed surveys were kept in a locked filing cabinet to which the researcher had the only key. The locked filing cabinet was located in the office of the researcher. All electronic data were kept on a computer that was password protected.
Data Analysis

Data from the DSDM Survey (Appendix G) were analyzed through content analysis, generation of scores for specific survey items as well as for topical areas. Demographic data collected were tabulated to describe the overall sample. Data collected via the Likert scales were considered interval data. This approach to handling Likert scale data was based on Burns and Grove (1993b). These authors state values of individual Likert responses are technically ordinal level data (Burns & Grove, 1993b). However, individual responses are often summed to obtain a single score for each item in question or for each subject and are often treated as interval-level data allowing for more sophisticated parametric statistics (Burns & Grove, 1993b). Polit and Hungler support the use of adding Likert scores together, calling it a “summated rating scale” (Polit & Hungler, 1999b, p. 341). Likert data collected were entered into the computerized statistical program Statistical Package for the Social Sciences (SPSS, Version 11.0) for analysis. Modes, means, standard deviations and ranges were calculated for each Likert-scored question. Scores greater than 4 were considered to be in agreement with the question stated. Scores less then 4 were considered to be disagreement with the question stated. Scores of 4 were considered to be neutral. Narrative responses were read, collated for similarities and reported.

Ethical Considerations

Each participant was fully informed the study was related to doctoral student work regarding the development of a consistent, efficient, cost-effective method to develop data sets for use in clinical care. They were informed there were no risks to them as a participant except for the expenditure of their time. The benefit of their participation
is possibly an improved methodology for data set development for clinical data sets. All IRB procedures for the University of Utah were followed.
CHAPTER IV

RESULTS

The results of the survey are most meaningful when organized in relation to the research questions addressed. The research questions addressed by this study were as follows:

1. Is the proposed methodology clearly defined?

2. Is the new methodology useful for data set development?

As mentioned in Chapter 3, the alignment of questions on the survey and the research question they address are contained in Appendix H.

Research Question 1: Clarity of the Proposed Methodology

The first research question asked, "Is the proposed methodology clearly defined?"

For the analysis of the survey results aimed at this research question, it is helpful to group and evaluate the results according to the different topics of the DSDM document and the headings of the DSDM survey. These subject headings are as follows:

1. The Methodology

2. The "Data Set Development Methodology" Flow Diagram

3. The "Application of Knowledge Acquisition Techniques for Formalization of Data Set" Flow Diagram.

4. Literature Review Technique
5. Data Analysis Technique
6. Expert Interview Technique
7. Focus Group Technique
8. General Overview

Most of the topics are listed as distinct bolded headings in the survey. However, distinct headings for questions related to topics 2 and 3 were overlooked and not included on the survey. These items are found under the heading “The Methodology” on the survey (questions 7-11 and 12-16 respectively). For each subject heading listed, the results from the Likert-type questions are reported first, followed by the narrative responses collected by the open-ended questions. The consequence of this organization of the results presentation is questions within one subject heading may not be covered in numerical order.

The Methodology

Questions 1, 2 and 4 asked participants (also referred to as subjects or reviewers) to rate several distinct aspects of the Methodology. These included the presentation of the Methodology, the reviewer’s perceived ability to apply the Methodology after reading the document and the Methodology’s fit with the Systems Development Life Cycle (SDLC). Not all reviewers answered all questions on the survey. This left some results with an N < 11. The mode, mean, standard deviation and range results of the Likert scores from these questions can be found in Table 5.

The Likert score results indicated the study group agreed the presentation of the DSDM was clear and understandable. The scores also indicate the reviewers felt they
would be able to apply the DSDM after having read the document and they felt the fit of the DSDM with the SDLC was clear. The range of scores was larger for the scores evaluating the presentation of the DSDM and the fit of the DSDM with the SDLC than the scores for agreement on being able to apply the DSDM after having read the document. Scores regarding the presentation of the DSDM had the highest variability of the group. One reviewer disagreed the presentation of the DSDM was clear and understandable, rating the item a “3,” as did one reviewer regarding the fit of the DSDM with the SDLC. The modes for these questions indicate the majority of reviewers rated the methodology very highly for these questions.

Reviewers were also asked for narrative feedback on what may be missing or unclear about the document and how the proposed process could be improved. The exact responses to the narrative questions can be found in Appendix L (questions 3 and 6) with a listing of responses to all narrative questions. Not all reviewers gave narrative feedback to all open-ended questions. Ten comments total were made to both questions regarding the Methodology. Responses indicated the document was in need of more examples, a
more specific "prescription" for the application of the methodology, further delineation of the steps in development, a stated definition of the term "data set," a method for finding existing data sets not readily available in the literature and a method for evaluating data sets resulting from the methodology.

The reviewers suggested the Data Set Development Methodology could be improved by expanding its application to the analysis phase of the Systems Development Life Cycle and not limiting it to the design phase of the SDLC. Other comments towards the improvement of the proposed process flow included considering the addition of other expertise on the data set development team such as individuals with expertise in data set work or informatics.

The "Data Set Development Methodology" Flow Diagram

An important part in describing the Data Set Development Methodology is presenting the methodology in an easy to understand, quickly comprehended diagram. This was done via the "Data Set Development Methodology" flow diagram of the DSDM document (Appendix E). Four of the questions on the survey were related to this diagram. The questions elicited feedback from the reviewers on the diagram's representation of the proposed process flow, the logic of the DSDM process, the diagram's clarity and its application to describe the methodology process. The results of the Likert scores from the questions related to the "Data Set Development Methodology" flow diagram are found in Table 6.

The Likert score results indicated the majority of reviewers found that the diagram was logical, followed the proposed process flow and was clear and helpful in
Table 6

Mode, Mean, Standard Deviation and Range of Scores Evaluating the “Data Set Development Methodology” Flow Diagram

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>Mode</th>
<th>Mean</th>
<th>SD</th>
<th>Min-Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>7. Flow diagram is logical</td>
<td>11</td>
<td>7</td>
<td>6.09</td>
<td>1.22</td>
<td>3-7</td>
</tr>
<tr>
<td>8. Flow diagram follows the proposed process flow</td>
<td>11</td>
<td>7</td>
<td>6.09</td>
<td>1.22</td>
<td>3-7</td>
</tr>
<tr>
<td>9. Flow diagram is clear</td>
<td>11</td>
<td>7</td>
<td>6.00</td>
<td>1.27</td>
<td>3-7</td>
</tr>
<tr>
<td>10. Flow diagram helps describe the methodology</td>
<td>10</td>
<td>6*</td>
<td>5.91</td>
<td>1.22</td>
<td>3-7</td>
</tr>
</tbody>
</table>

* Multiple modes exist. The smallest value is shown.

describing the Methodology. The range of scores for each of these questions was the same and there was not much difference in the variability of these scores. The mode of each question suggested the reviewers rated the diagram describing the methodology very highly. One reviewer disagreed (rating of 3) with all questions related to this diagram.

Nine comments were submitted regarding the “Data Set Development Methodology” flow diagram. The narrative feedback from the reviewers suggested this flow diagram could be improved by clarifying the “Quality Analysis” and “Freeze” points of the methodology in the document. Reviewers also suggested providing more information on the criteria used during the quality analysis points of the methodology. Formatting issues such as being consistent in the level of detail included in the diagram, including a key to explain abbreviations and arranging the phases left to right to show “true chronology” were also addressed by reviewers. One reviewer found the diagram to be “quite clear.” (For a complete listing of specific comments made by reviewers, please refer to Appendix L, question 11.)
The "Application of Knowledge Acquisition Techniques for Formalization of Data Set: Flow Diagram"

Another diagram was developed to assist with explaining the process of the Data Set Development Methodology. This diagram, titled "Application of Knowledge Acquisition Techniques for Formalization of Data Set" is found in Appendix E. The survey contained similar questions about this diagram as the earlier diagram. The Likert score results from the questions related to the "Application of Knowledge Acquisition Techniques for Formalization of Data Set" flow diagram are found in Table 7.

The Likert scores indicate the reviewers found the diagram logical; it followed the proposed process flow and was clear and helpful in describing the methodology. The modes and means of these scores are very similar, but the ranges and standard deviation of these scores are quite varied. The agreement of scores for question 12, regarding the diagram being logical, had the largest range and the most variability among scores in this table.

Table 7

<table>
<thead>
<tr>
<th>Question</th>
<th>Mode</th>
<th>Mean</th>
<th>SD</th>
<th>Min-Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>12. Flow diagram is logical</td>
<td>11</td>
<td>7</td>
<td>6.18</td>
<td>3-7</td>
</tr>
<tr>
<td>13. Flow diagram follows the proposed process flow</td>
<td>11</td>
<td>7</td>
<td>6.27</td>
<td>5-7</td>
</tr>
<tr>
<td>14. Flow diagram is clear</td>
<td>11</td>
<td>7</td>
<td>6.18</td>
<td>4-7</td>
</tr>
<tr>
<td>15. Flow diagram helps describe the methodology</td>
<td>10</td>
<td>7</td>
<td>6.36</td>
<td>5-7</td>
</tr>
</tbody>
</table>
section. One reviewer disagreed with the diagram being logical, rating the item a “3.” This reviewer did add the comment “Seems like a need LR is needed initially before Lit Review [sic].” Unfortunately, the comment is limited in its meaning and not very useful to the improvement of the methodology or diagram.

Six reviewers gave narrative comments regarding the “Application of Knowledge Acquisition Techniques for Formalization of Data Set” diagram. Suggestions made by reviewers for the improvement of the diagram included defining more clearly how one might know the data set is “complete” and what type of literature impacts a data set. Again reviewers’ commented on formatting issues such as alignment of figures and the use of abbreviations. One comment indicated the diagram was “simple and clean.” The specific comments made for this diagram are also found in Appendix L, question 16.

**Literature Review Techniques**

The Data Set Development Methodology document contains specific sections explaining each of the techniques included in the DSDM process. Each of the sections included a definition and description and information for the application of the specific technique. The Likert questions regarding the clarity and understandability of the definition and description of the techniques were used in the evaluation of the results in regards to Research Question 1, the clarity of the methodology. The Likert question regarding the application of the technique was used in the evaluation of the results in regards to Research Question 2, the usefulness of the methodology. The survey queried reviewers for feedback on the reader’s understanding and the clarity of each topic and asked for narrative feedback on how the presentation of each technique could be
improved. The responses to the open-ended questions regarding suggested improvements will be presented after the clarity Likert scores, while discussing each subject. The specific Likert scores for each topic addressing the usefulness of the methodology will be presented later, when discussing Research Question 2.

The first technique presented in the Data Set Development Methodology Document is the Literature Review technique. The results of the Likert scores for the Literature Review section are found in Table 8.

The Likert results in Table 8 show the reviewers found the definition and description of the Literature Review technique to be clear and understandable. Five specific narrative responses from reviewers regarding the Literature Review technique are found in Appendix L, question 20. The comments suggested including more information on using MESH/search terms and a description of all types of literature useful to data set development, including fugitive literature.

Reviewers also suggested the methodology include a recommendation for search strategies used during the development of a specific data set to be documented and saved.

Table 8

*Mode, Mean, Standard Deviation and Range of Scores Evaluating the Literature Review Technique*

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>Mode</th>
<th>Mean</th>
<th>SD</th>
<th>Min-Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>17.</td>
<td>11</td>
<td>7</td>
<td>6.27</td>
<td>1.10</td>
<td>4-7</td>
</tr>
<tr>
<td>18.</td>
<td>11</td>
<td>6</td>
<td>6.99</td>
<td>1.04</td>
<td>4-7</td>
</tr>
</tbody>
</table>
for future reference. The formatting of the section was a concern to one reviewer who suggested the document be written in a more “handbook/recipe” style in an effort to assist readers in the readability of the definitions etc. One reviewer had no suggestions, liked the hint section and felt the information was helpful for a beginner or an expert.

**Data Analysis Techniques**

A technique using analysis of data previously collected in similar instances is one of the techniques included in the Data Set Development Methodology. It was the least understood or accepted techniques of the Methodology as indicated by the results presented below. The mode, mean, standard deviation and range of the Likert Scores for clarity and understandability of the definition and description of the Data Analysis technique are found in Table 9.

These Likert scorings indicate the reviewers found the definition and description of the Data Analysis technique to be clear and understandable, although not quite as

<table>
<thead>
<tr>
<th>21. Definition is clear and understandable</th>
<th>N</th>
<th>Mode</th>
<th>Mean</th>
<th>SD</th>
<th>Min-Max</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>11</td>
<td>7</td>
<td>5.91</td>
<td>1.22</td>
<td>4-7</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>22. Description is clear and understandable</th>
<th>N</th>
<th>Mode</th>
<th>Mean</th>
<th>SD</th>
<th>Min-Max</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>11</td>
<td>7</td>
<td>5.64</td>
<td>1.29</td>
<td>4-7</td>
</tr>
</tbody>
</table>
much as those of the Literature Review technique. The mean scores for the Data Analysis technique are lower than those for the Literature Review technique, but the modes are equal to or higher than those for the Literature Review. The variability of scores for the Data Analysis technique is higher than those for the Literature Review technique.

Comments collected from reviewers regarding the Data Analysis Technique suggested clarification is needed in this area of the Data Set Development Methodology document. One reviewer asked for more definition of the data set engineer position and the skills needed for this position. Three reviewers implied the section was not clear to them and suggested more clarification and simplified formatting of the section. One reviewer specifically wrote “no suggestions” and I wrote the example given in the section was helpful in clarifying the intent of the technique. Specific responses regarding the Data Analysis Technique are included in Appendix L (question 24).

**Expert Interview Technique**

A third technique included in the Data Set Development Methodology is the use of Expert Interviews. Similar questions were asked of the reviewers regarding the Expert Interview technique as were asked about the Literature Review technique and Data Analysis technique. The mode, mean, standard deviation and range of the Likert score results regarding the clarity and understandability of the Expert Interview technique definition and description are in Table 10.

The Likert results in Table 10 indicate the majority of reviewers found both the definition and description of the Expert Interview technique clear and understandable.
Table 10

Mode, Mean, Standard Deviation and Range of Scores Evaluating the Expert Interview Technique

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>Mode</th>
<th>Mean</th>
<th>SD</th>
<th>Min-Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>25. Definition is clear and understandable</td>
<td>11</td>
<td>7</td>
<td>6.09</td>
<td>1.22</td>
<td>3-7</td>
</tr>
<tr>
<td>26. Description is clear and understandable</td>
<td>11</td>
<td>6</td>
<td>5.73</td>
<td>1.19</td>
<td>3-7</td>
</tr>
</tbody>
</table>

The definition scored higher than the description but had more variability. One reviewer again disagreed and rated the definition and description of the Expert Interview Technique a "3" on the rating scale, indicating the definition and description of this technique were not clear and understandable.

The five narrative responses given by reviewers regarding the Expert Interview technique are found in Appendix L, question 28. The remarks were varied. One reviewer wrote "no suggestions—quite clean, thoughtful." Others made suggestions for improvement such as the inclusion of more examples and specific word tenses to use. One reviewer questioned the overall value of using experts and this technique in the Methodology at all.

Focus Group Technique

The final technique included in the Methodology is the use of Focus Groups. The questions regarding the clarity and understandability of the definition and description of
this technique were posed to reviewers as well. The mode, mean, standard deviation and range of Likert results for the Focus Group Technique are found in Table 11. The mode, mean, standard deviation and range of scores regarding the clarity/understandability scores of the definition and description of the Focus Group Technique were very close. The mode and mean of the Likert scores indicate the majority of reviewers did find the Focus Group Technique clearly defined and described. One reviewer however did not feel the definition and description were clear and understandable and gave the items a rating of “3.”

Seven narrative responses were collected from reviewers regarding the Focus Group Techniques. In response to the Focus Group Technique 1 reviewer suggested the section be less wordy, more direct and user friendly. The use of a scribe during the focus group session was suggested to allow the moderator to focus on managing the group process. One reviewer suggested including more samples and generic questions directed at data set development. One reviewer stated the section was clear, but again questioned

Table 11

| Mode, Mean, Standard Deviation and Range of Scores Evaluating the Focus Group Technique |
|----------------------------------|----------|--------|--------|---------|         |
| 29. Definition is clear and understandable | 11       | 7      | 6.18   | 1.25    | 3-7     |
| 30. Description is clear and understandable | 11       | 7      | 6.00   | 1.27    | 3-7     |
the value of focus groups with the development of data sets. For a listing of all specific
responses see Appendix L, question 32.

General Overview

Questions 36 and 37 were related to the general overview of the DSDM document
and asked reviewers to score the redundancy and organization of the document. The
mode, mean, standard deviation and range of the Likert results for the redundancy and
organization of the document are listed in Table 12.

The mean score for redundancy in the document is the lowest of all scores from
the survey. This finding indicates reviewers find some redundancy in the document. The
ranges and variability for these two items are some of the largest seen from all the items
indicating the most dissension amongst reviewers. The results to these questions were
bimodal but continued to indicate generally high scores from the reviewers with the
modes of 6 and 7. Two reviewers thought the document was somewhat redundant giving
the item a rating of “3.” One reviewer did not think the document well organized and
gave the organization item a rating of “2.”

Table 12

| Mode, Mean, Standard Deviation and Range of Scores Evaluating Redundancy and Organization of the DSDM Document |
|---------------------------------|---|---|---|---|
| 36. No redundancy in the methodology content | N | Mode | Mean | SD | Min-Max |
| 11 | 6a | 5.36 | 1.50 | 3-7 |
| 37. Document is well organized | N | Mode | Mean | SD | Min-Max |
| 10 | 6a | 5.80 | 1.62 | 2-7 |

*aMultiple modes exist. The smallest value is shown.*
Two questions asked reviewers to provide suggestions for the improvement of the Data Set Development Methodology Document. Six responses were given for general improvements and seven regarding the readability of the document. These specific responses are listed in Appendix L, questions 38 and 39.

Several reviewers suggested the document could be improved with the addition of more examples, sample forms, bullets and tables. One reviewer made grammatical suggestions for font use and capitalization. Suggestions were made by other reviewers to include a better description for the process of finding existing vocabularies and to include "user groups" as part of the process. It was also suggested to include evidence of the testing of the methodology. Interestingly, one reviewer suggested the document could be improved by shortening it, while also suggesting the addition of examples, bullets and tables.

Comments regarding the readability of the document were split. Two reviewers found it easy to read. One liked the headings, bulleted lists and table of contents. One reviewer suggested each section begin with a "taste" of the entire process to be discussed followed by the subject matter. It was suggested the introduction of cited material be more varied and to include a description of each myth in the Focus Group section. Other reviewers suggested the document be simplified by using more examples, specifically ones that are tracked throughout the document and include an overview similar to an executive summary.
Research Question 2: Usefulness of the New Methodology for Data Set Development

The second research question asked, "Is the new methodology useful for data set development?" The usefulness of the Data Set Development Methodology was evaluated by asking reviewers to rate the reasonableness of the proposed process flow, the applicability of each of the techniques used in the methodology for data set development and the generalizability of the methodology to all healthcare disciplines as well as to data sets used on paper or in computerized systems. The results from the eight questions of the survey aimed at evaluating the usefulness of the methodology are found in Table 13. The results of the Likert scoring indicate the reviewers felt the Data Set Development Methodology was useful. The mean scores for the Expert Interview and Focus Group Techniques (questions 27 and 31 respectively) were a little higher in applicability to data set development then the Literature Review (question 19) and Data Analysis Techniques (question 23). The higher scores for the Expert Interview and Focus Group Techniques are interesting since they both had comments questioning the usefulness and/or helpfulness of the technique for data set development. Not much difference was seen in the modes for these questions. The highest mean score given by reviewers regarding the usefulness of the methodology was for the generalizability of the methodology to all disciplines within healthcare. Scores regarding the use of the methodology for data sets used in paper systems or for data sets used in computerized systems were equal. The scores related to the use of the methodology for paper systems or computerized systems also had the least variability among scores. The most variability among scores was found in the responses to the applicability of the Data Analysis
Table 13

*Mode, Mean, Standard Deviation and Range of Scores for Questions Evaluating the Usefulness of the Data Set Development Methodology*

<table>
<thead>
<tr>
<th>Question</th>
<th>N</th>
<th>Mode</th>
<th>Mean</th>
<th>SD</th>
<th>Min-Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>5. Process flow is reasonable</td>
<td>11</td>
<td>6</td>
<td>5.91</td>
<td>1.14</td>
<td>3-7</td>
</tr>
<tr>
<td>19. Literature Review Technique is applicable to data set development</td>
<td>10</td>
<td>7</td>
<td>5.80</td>
<td>1.23</td>
<td>4-7</td>
</tr>
<tr>
<td>23. Data Analysis Technique is applicable to data set development</td>
<td>11</td>
<td>7</td>
<td>5.82</td>
<td>1.33</td>
<td>4-7</td>
</tr>
<tr>
<td>27. Expert Interview Technique is applicable to data set development</td>
<td>11</td>
<td>6</td>
<td>6.09</td>
<td>0.94</td>
<td>4-7</td>
</tr>
<tr>
<td>31. Focus Group Technique is applicable to data set development</td>
<td>10</td>
<td>7</td>
<td>6.10</td>
<td>1.10</td>
<td>4-7</td>
</tr>
<tr>
<td>33. Methodology is generalizable to other disciplines in health care</td>
<td>11</td>
<td>7</td>
<td>6.36</td>
<td>0.81</td>
<td>5-7</td>
</tr>
<tr>
<td>34. Methodology is applicable to data sets used on paper</td>
<td>11</td>
<td>7</td>
<td>6.27</td>
<td>0.79</td>
<td>5-7</td>
</tr>
<tr>
<td>35. Methodology is applicable to data sets used in computerized systems</td>
<td>11</td>
<td>7</td>
<td>6.27</td>
<td>0.79</td>
<td>5-7</td>
</tr>
</tbody>
</table>
Technique. One reviewer disagreed with the process flow of the methodology being reasonable, rating the item a “3.”

At the end of the survey, reviewers were asked to provide general comments or suggestions. Eight reviewers provided general comments. Several of the comments were supportive and complimentary of the overall project and the methodology. More suggestions for improvement were made such as clarifying the difference between a data set, a database and a system of data elements. One reviewer suggested the inclusion of a brief overview or diagram of the Systems Development Life cycle as a review or for readers unfamiliar with the process. It was suggested in these comments for the methodology to include a recommended method for picking sources when multiple similar sources exist. Comments related to the need to present the methodology more clearly with the use of more examples and tools before it can be really useable were reiterated. Two reviewers confirmed the methodology in stating the techniques presented were quite valid and are used by their own data set development groups.

Suggestions for Validation and Distribution of the Data Set Development Methodology

The survey presented to the national reviewers included three questions related to the future validation and distribution of the Data Set Development Methodology. The first question in this series asked the reviewers “In your opinion, what kind of validation study of this methodology would be useful?” This question was presented to reviewers to gather a broad set of ideas for what procedures may be useful for the validation of the methodology. The nine responses received for this question are found in Appendix L,
question 40. Several of the responses suggested the application of the methodology in practice would provide validation. Recommendations were made for comparison studies between data set development groups and for the development of a data set either in a complex yet focused domain or for a specialty unit with a specific need.

The remaining two questions of the survey targeted the current availability of the proposed methodology and potential methods of distribution for this methodology. The reviewers were asked if the methodology was already available in another format. Six reviewers stated “No” it was not available in other formats, 1 indicated “yes” and 1 wrote in “unknown” on the survey. For the responses of “Yes” to this question, reviewers were asked to provide information on where the methodology could be found. The 1 reviewer marking “Yes” noted similar information could be found in a textbook with the accompanying note, “check Saba and McCormick book—developing a system.” Other nonsolicited narrative responses written in response to this question included the following:

- Not a complete methodology, only parts in textbooks or journal article
- I don’t remember ever seeing this published, however I believe that vendors/software developers use these techniques/methodology with the addition of “user groups”
- Not explicitly, but the methods are certainly around in other formats and for other purposes
- I don’t know of any

To ascertain potential methods of distribution for the methodology, reviewers were asked to provide feedback on how they thought it should be distributed. The survey included a list of possible methods for distribution. These methods and the responses received are listed in Table 14. Reviewers were allowed to choose more than one potential
Table 14  

*Frequency Potential Method of Distribution Chosen by Reviewers*

<table>
<thead>
<tr>
<th>Method</th>
<th>f</th>
<th>% of Reviewers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Journal Article</td>
<td>7</td>
<td>64</td>
</tr>
<tr>
<td>Published Manuscript</td>
<td>4</td>
<td>36</td>
</tr>
<tr>
<td>Textbook</td>
<td>6</td>
<td>55</td>
</tr>
<tr>
<td>Published Flow Diagram</td>
<td>3</td>
<td>27</td>
</tr>
<tr>
<td>World Wide Web</td>
<td>4</td>
<td>36</td>
</tr>
<tr>
<td>Other</td>
<td>2</td>
<td>18</td>
</tr>
</tbody>
</table>

distribution method. The most popular method for distribution among reviewers was “Journal Article,” followed by “Textbook.” The popularity of using a “Published Manuscript” or the “World Wide Web” was tied. Using a “Published Flow Diagram” was the least popular of the suggested methods. Two other methods suggested by 2 reviewers for the distribution of the methodology included a chapter in an informatics textbook and as a presentation at a conference.

**Summary of Results**

The results of the survey provided evidence to answer the research questions. The 11 study participants represented a wide range of knowledge and experience in data set development, but all had experience in the field and were qualified to participate as a reviewer of the Data Set Development Methodology Document.

The results of the survey and the comments received from the reviewers indicated the Data Set Development Methodology was fairly well defined and, in the opinion of the expert reviewers, is useful for data set development. The mean scores for all questions asking about the clarity of the document were above 5, indicating the reviewers agreed with the document being clearly defined. Certain areas of the Methodology had higher
mean clarity scores than others. The ratings indicated the DSDM document was somewhat redundant and the fit of the Data Set Development Methodology with the Systems Development Life Cycle was not as clearly explained as other aspects of the methodology. The generalizability of the methodology to other healthcare disciplines received the highest mean rating from all reviewers.

The four core techniques included in the Data Set Development Methodology all received ratings indicating they were clearly defined. The Literature Review technique received the highest ratings for clarity. Although the Data Analysis technique was the least understood and the most unclear method to the readers, it still received marks indicating the reviewers found the technique clearly defined. The clarity scores for each of the techniques hover around the mean of 6.0 indicating they are clearly defined but could use some improvement.

The two flow diagrams included in the DSDM document were also rated by reviewers as being clearly defined and helpful in describing the Methodology. The flow diagram titled “Application of Knowledge Acquisition Techniques for Formalization of Data Set” received the highest mean rating of all ratings for its helpfulness in describing the methodology. The ratings for the “Data Set Development Methodology” flow diagram were also on the higher end of the scale indicating agreement in its clarity and helpfulness in describing the Methodology.

All scores related to the usefulness of the Data Set Development Methodology were on the upper end of the agreement scale indicating the reviewers thought each part of the Data Set Development Methodology is useful to the development of data sets. Interestingly the Literature Review technique, having some of the highest clarity scores,
was rated the lowest of all the usefulness scores. The second lowest score for usefulness in the opinion of the expert reviewers was received by the Data Analysis technique.

Comments received from reviewers were very constructive. The themes of the comments received centered around four areas. These areas are as follows:

1. Suggestions for improvement to the Methodology overall
2. Suggestions for improvement to specific areas of the Methodology
3. Positive statements supporting the Methodology
4. Positive statements regarding the presentation of the Methodology.
CHAPTER V

DISCUSSION AND RECOMMENDATIONS

The discussion of findings is organized into four major sections. The first section contains a discussion of the findings of the study, the second section examines the limitations of the study and the third section lists recommendations for further research. The fourth and final section of the chapter will address the implications of this study on the fields of clinical informatics and nursing.

Discussion of Significant Findings

The purpose of this study was to develop and formatively evaluate a methodological approach for developing data sets for computerized patient care documentation systems. The objectives were as follows:

1. Develop a methodology for data set development that does not require an expensive national invitational conference of domain experts but results in a useful data set for documenting patient care in a local setting. The methodology should be generalizable to multiple settings for others to use when developing a data set.

2. Evaluate the data set development methodology via expert review.

A methodology for data set development that does not require the gathering of domain experts at a national conference was developed. The resulting methodology may require the gathering of domain experts (if the focus group technique is used) but
considers and uses individuals with expertise related to a data set that can be found locally, closer to the site of the data set development.

The second objective of this study was to conduct a formative evaluation of the Data Set Development Methodology via an expert review. This expert review was conducted with the assistance of 11 individuals across the United States with experience in the use and development of data sets. Nineteen individuals were personally approached with the invitation to participate. The study required 3-4 hours of “work” by the subjects to read, review and critique the 47-page document. Therefore, the author is pleased that a majority did participate and returned completed surveys.

The comments and responses gathered from the reviewers show the group was receptive to the idea of the Data Set Development Methodology. Most of the comments were supportive of the project and praised the majority of the Methodology. Multiple suggestions for improvements questioned the value of the Expert Interview and Focus Group techniques, but no comments questioned the value of the Methodology itself. The sample group was diverse in its data set development experience and its geographical location. The acceptance of the Methodology by this diverse sample indicates the Methodology may be generalizable to a variety of settings.

The study group of 11 meets the recommendation made by Lynn (1986) to include at least 5 experts to control for chance agreement and exceeds the maximum of 10 thought necessary by Lynn (1986). One reviewer frequently gave the Methodology ratings of 3 on the Likert scale indicating disagreement with the Methodology. According to Lynn (1986), the larger study group allows reviewers to be in disagreement while still allowing the methodology itself to be considered valid.
The expert review concentrated on two specific research questions, one being the clarity of the newly proposed Data Set Development Methodology and the second being the usefulness of it. The results of the study indicate the DSDM is clearly presented, although suggestions were made for improvement. The usefulness of the overall Methodology was supported and not questioned by the review group.

**Study Limitations**

Limitations of this study include the use of purposive sampling and the instrument used to collect responses from the experts. Purposive sampling cannot guarantee a representative sample of the target population for a study. Efforts were made, however, to invite the widest variety of qualified subjects as possible. The researcher could not control who would and would not decide to participate. The resulting study sample did, however, contain individuals with a wide range of experience both in years and in positions held for data set development providing a fairly representative sample of the target population.

The instrument used for data collection was developed specifically for this study. Efforts were made to validate the survey but the reliability of the tool was not tested. Although considered, the repetitive use of the survey needed to determine its reliability was impractical. The pool of qualified subjects was limited in size. Finding two or more groups or asking one group to complete the survey twice was not feasible. The effort required by reviewers to participate in this study was not trivial. The evaluation conducted was a formative evaluation. The responses gathered from the survey were very constructive and helpful to the further development of the methodology regardless of whether the same results would have been obtained again at a later date.
Recommendations for Future Research

Three main areas of focus remain for future development and research regarding the Data Set Development Methodology. These areas include further development of the Methodology based on comments and suggestions received from reviewers, the application of the Methodology to the development of a data set and further evaluation and validation of the Methodology.

Further Development Needed

The Methodology itself could be improved with further refinements and developments as suggested by the reviewers. Several areas such as the “Quality Assurance” and “Freeze” points of the Methodology, the fit of the Methodology with the Systems Development Life Cycle and the Data Analysis technique need clarification.

The addition of a stronger example carried as a thread throughout the DSDM document, as well as more example forms and documents, might make the Methodology clearer to the reader.

Several reviewers made suggestions for additional thoughts, tasks or comments to be added to the Methodology. Including reviewer ideas such as more information on MESH/search terms and the use of a scribe during Focus Group sessions are valuable and would strengthen the overall DSDM document.

Several other reviewer comments touched on formatting issues that would enhance the readability of the DSDM document. The use of abbreviations in the diagrams and the use of varied fonts to highlight different areas of text were confusing to some readers. Reviewers also suggested using more bullet points and tables throughout the document and perhaps moving the definitions to a glossary. Changing and clarifying
the document from these and other suggestions made by the reviewers would improve the readability of a long and somewhat tedious document.

Application of the Methodology

The truest evaluation of the Data Set Development Methodology's usefulness will be the application of the Methodology to the development of a data set with a defined scope. The application of the Methodology to develop a data set may find errors or holes in the Methodology or the included techniques. Applying the Methodology will allow for the discovery of overlooked steps or recommendations that should be included.

When reviewers were asked what type of validation study would be useful for this methodology, several of them responded with a statement suggesting the development of at least one data set. Suggestions were made to compare a data set developed using this Data Set Development Methodology with another similar data set developed with an alternative approach or "previous" methodology. Although it may be difficult to identify another specified "structured" approach to data set development, it would be useful to compare data sets developed by the DSDM and other ad hoc development methods. A comparison of a resulting data set created using the DSDM with other data sets, regardless of their development method, would be helpful in determining the usefulness of the DSDM and its contribution towards the improvement of data set development.

Further Evaluation

Another method being pursued by the researcher to evaluate the Methodology through its application is the evaluation of the resulting data set. The premise is, if the resulting data set is complete and useful to its specified purpose then the Methodology for
obtaining that data set is valid. Two approaches are being investigated for the evaluation of a data set developed using the Data Set Development Methodology. The first includes the individual items of the data set being evaluated by potential users of the data set (i.e., experts in the use of a data set for the specified domain) for the appropriateness of inclusion in the data set. The second potential evaluation for a data set resulting from the applied Methodology is the actual use of the data set in the setting for which it is designed.

The first approach requires the expert users to evaluate each item included in a data set and to give it a score regarding the appropriateness of its inclusion in the data set. Appropriateness scoring described by Park and colleagues (1986) could be applied to the evaluation of the appropriateness of the inclusion of data elements in a data set. This technique, referred to as the RAND Appropriateness Methodology (Piccirillo et al., 1998) or RAND Formal Consensus Technique (Shiffman et al., 1994), requires experts to rate each item included in an instrument on a scale of 1 (extremely inappropriate) to 9 (extremely appropriate). The ratings are then analyzed according to a method prescribed by Park and colleagues (1986) to determine the appropriateness of the data element being included in the resulting data set. The outcome of these scorings would be a good indication of how well the Data Set Development Methodology worked to produce an acceptable data set.

The actual use of a data set resulting from the application of the Data Set Development Methodology is one of the most definitive tests that could be done to evaluate the usefulness of the Data Set Development Methodology. All data sets are
produced for a specific purpose. Evaluating the resulting data set against the needs of the users and the goals of the data set is essential.

To conduct this evaluation the data set needs to be put to its intended use and then evaluated for the use and/or "nonuse" of data elements within the data set. The data elements contained in the data set need to be put in some usable form for the users whether it is on paper or in a computer and then monitored over a set period of time, patients or instances to see what data elements are used or not used. At the end of the designated study period the use of each data element can be determined by simply counting the number of times it was used. This total can be compared to the potential number of uses and the data element determined to be "useful and needed" or "extraneous." If the data element is never used to record data, then it should be evaluated for possibly being irrelevant to the purpose of the data set. The acceptable or desired percentage of specific data element use targeted for each data set is dependent on the overall goals of the data set and should be set by each developer based on the consequences of missing or extraneous data elements.

Determining how often an element is viewed or "read" in the life of a data set is more difficult. Evaluating the number of times a data element is "used" by simply being read or viewed is doable but unrealistic. It is reasonable to expect, however, that if a data element is needed for aggregate reporting or "read only"-type functions that its absence will be noted and requested by users (i.e., viewers) to be added to the overall data set. It is important during this evaluation period not only to note the data elements actually being used but to allow users to have the ability to note any additional data elements that are needed but not present in the current data set.
Implications for Clinical Informatics and Nursing

The developed methodology and subsequent evaluation are potentially useful to the field of data set development and to clinical informatics in general. Presently, many data set development activities are occurring throughout the country. Past data set development projects reported in the literature do not include much information regarding the specific methods used for their development. The Data Set Development Methodology provides a structure to use for data set development. No other current reports were found in the literature regarding the development and/or use of a structured methodology for data set development.

The development of an efficient, cost-effective system for the development of data sets for use in patient care documentation systems is useful to nursing and nursing informatics in general. A growing number of nurses are interested in and involved with healthcare informatics. These nurses may or may not have a formal education in informatics principles. A formal education in informatics principles provides individuals with tools and skills needed to make the determination of what data elements to include in a data set. This education, however, is not readily available or of interest to all persons charged with the development of data sets or employed as a “Nurse Informaticist.” Even those with a formal education or extensive “on the job” training in informatics do not have a consistent method to apply when developing a data set. The preparation and publication of documents containing principles that can be accessed and used as resources are useful to both of these audiences. These resources provide the opportunity for some to learn new knowledge and to provide structure and consistency for others.
Nurse informaticists are often involved in the determination of what data elements should be included in a computerized documentation system. Even when a computerized system is purchased from a vendor, users and/or the purchasing site are often faced with the vendor charge of “we’ll put whatever YOU want in there.” This leaves the users and site with the responsibility of determining what data elements need to be included in the system for their site.

This methodology allows nurses at all locations to develop data sets for their own use in a timely manner without depending on “national experts” to create and publish the data set for them. Data set development work at a local level and the DSDM may also have applicability at a national level. Locally developed data sets may be very useful during the literature review activities of the formalization stage for a national data set. Each of the local data sets could be examined for overlapping data elements and unique data elements. The overlapping data elements might automatically be included in the national data set, while unique items would indicate areas for consideration of possible inclusion or exclusion based on the defined purpose and goals of the national data set.

The developed methodology also has implications beyond the field of informatics. Bedside nurses, who may be primary users of a data set developed using this Methodology, benefit by having a data set well grounded in the domain. Using a data set tailored to their needs without missing elements and not containing numerous extraneous data elements they must wade through to complete their documentation tasks may decrease frustration and documentation time.

In summary, this work has implications across the discipline of nursing. The American Nurses Association supports the creation of standards and advancements
promoting the development of data sets (Coenen et al., 2001). The Data Set Development Methodology is one such tool providing a consistent structure and set of techniques for use in data set development. The formalization of this Data Set Development Methodology provides clinical data set developers of various skill levels and resources a documented, consistent, structured and economical tool for the creation of their needed and/or desired data set.
APPENDIX A

DATA SETS FOUND IN THE LITERATURE WITH PURPOSE, METHOD OF DEVELOPMENT, METHOD OF EVALUATION AND REFERENCE LISTED FOR EACH
<table>
<thead>
<tr>
<th>Name of Data Set</th>
<th>Acronym</th>
<th>Approx. Dev. Date</th>
<th>Concise purpose</th>
<th>Method of Development</th>
<th>Method of Evaluation</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Air Transport Minimum Data Set</td>
<td></td>
<td>1999</td>
<td>Standardize data</td>
<td>5 Round Delphi Study</td>
<td>Agreement rating on 3-point Likert scale (survey)</td>
<td>(Thompson &amp; Schaffer, in press)</td>
</tr>
<tr>
<td>*Ambulatory Medical Care Minimum Data Set</td>
<td>AMC</td>
<td></td>
<td>Standardize data</td>
<td>Working conference</td>
<td>Field test for availability and feasibility</td>
<td>(McCormick et al., 1997)</td>
</tr>
<tr>
<td>*Core Health Data Elements</td>
<td></td>
<td>1996</td>
<td>Standardize data</td>
<td>Government committee, mail survey, &quot;hearings&quot; (expert focus group)</td>
<td>Unknown. Mailings and hearings were repetitive – indicating evaluation and refinement</td>
<td>(McCormick et al., 1997)</td>
</tr>
<tr>
<td>*Data Elements for Emergency Department Systems</td>
<td>DEEDS</td>
<td>1997</td>
<td>Standardize data</td>
<td>Public/private sponsored National workshop</td>
<td>Open national review</td>
<td>(McCormick et al., 1997; National Center for Injury Prevention and Control, 1997; NII-HIN Consortium, 1997; Renner &amp; Swart, 1997)</td>
</tr>
<tr>
<td>Name of Data Set</td>
<td>Acronym</td>
<td>Approx. Dev. Date</td>
<td>Concise purpose</td>
<td>Method of Development</td>
<td>Method of Evaluation</td>
<td>Reference</td>
</tr>
<tr>
<td>----------------------------------</td>
<td>---------</td>
<td>------------------</td>
<td>-----------------</td>
<td>-----------------------</td>
<td>---------------------------------</td>
<td>-----------------------------------------</td>
</tr>
<tr>
<td></td>
<td>MDS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(McCormick et al., 1997), (Renner &amp; Swart, 1997)</td>
</tr>
<tr>
<td>Essential Medical Data Set (The)</td>
<td>EMDS</td>
<td>Version 1.0 1996</td>
<td>Standardize data</td>
<td>Public/private sponsored workgroup, Case review (&gt;1000 cases)</td>
<td>Open national review</td>
<td>(NII-HIN Consortium, 1997)</td>
</tr>
<tr>
<td>Name of Data Set</td>
<td>Acronym</td>
<td>Approx. Dev. Date</td>
<td>Concise Purpose</td>
<td>Method of Development</td>
<td>Method of Evaluation</td>
<td>Reference</td>
</tr>
<tr>
<td>------------------</td>
<td>---------</td>
<td>-------------------</td>
<td>-----------------</td>
<td>-----------------------</td>
<td>----------------------</td>
<td>-----------</td>
</tr>
<tr>
<td>*Federal Recommended Data Elements Medicare Quality Indicator System (Formerly, Uniform Clinical Data Set)</td>
<td>MQIS (UCDS)</td>
<td>1992</td>
<td>Standardize data</td>
<td>Unknown</td>
<td>Unknown. McCormick reports MQIS data elements have been used to collect data from &gt;400,000 records and that a draft was being tested for usefulness</td>
<td>(McCormick et al., 1997; Renner &amp; Swart, 1997)</td>
</tr>
<tr>
<td>Financial Uniform Minimum Data Set</td>
<td>UB-92</td>
<td>1992</td>
<td>Standardize data</td>
<td>Public/private sponsored committee</td>
<td>Unknown. UB92 is used nationwide to meet reimbursement requirements for Medicare and Medicaid</td>
<td>(McCormick et al., 1997; Renner &amp; Swart, 1997)</td>
</tr>
<tr>
<td>Name of Data Set</td>
<td>Acronym</td>
<td>Approx. Dev. Date</td>
<td>Concise Purpose</td>
<td>Method of Development</td>
<td>Method of Evaluation</td>
<td>Reference</td>
</tr>
<tr>
<td>----------------------------------</td>
<td>---------</td>
<td>-------------------</td>
<td>--------------------------------------------------------------------------------</td>
<td>-----------------------------------------------</td>
<td>--------------------------------------------------------------------------------------</td>
<td>---------------------------</td>
</tr>
<tr>
<td>*Health Insurance Claim Form</td>
<td>HCFA 1500</td>
<td>1990</td>
<td>Provide submission guidelines (to government agency)</td>
<td>Unknown. Presumably by government committee</td>
<td>Unknown. Is used to submit MD charges to Medicare/Medicaid (and possibly other insurances)</td>
<td>(McCormick et al., 1997)</td>
</tr>
<tr>
<td>Health Plan Employer Data and Information Set</td>
<td>HEDIS 3.0</td>
<td>Version 3.0 1996</td>
<td>Measure performance</td>
<td>Unknown. Presumed to be by a private (organizational) committee</td>
<td>Unknown. NCQA website reports HEDIS is used by 90% of U.S. managed care organizations to measure performance</td>
<td>(McCormick et al., 1997)</td>
</tr>
<tr>
<td>*Health Resources Service Administration Uniform Data System</td>
<td>UDS</td>
<td>1995</td>
<td>Provide submission guidelines (to government agency)</td>
<td>Unknown. Presumed to be by a public (government) committee</td>
<td>Unknown. Used by recipients of several gov.’t grants for annual reporting</td>
<td>(McCormick et al., 1997)</td>
</tr>
<tr>
<td>Name of Data Set</td>
<td>Acronym</td>
<td>Approx. Dev. Date</td>
<td>Concise Purpose</td>
<td>Method of Development</td>
<td>Method of Evaluation</td>
<td>Reference</td>
</tr>
<tr>
<td>----------------------------------</td>
<td>---------</td>
<td>------------------</td>
<td>------------------------------------------------------</td>
<td>----------------------------------------------------------------------------------------</td>
<td>---------------------------------------------------------------------------------------</td>
<td>-----------------------------------------------</td>
</tr>
<tr>
<td>*Hospital Medical Records Institute</td>
<td>HMRI</td>
<td>1991</td>
<td>Provide submission guidelines (to government agency)</td>
<td>Unknown. Presumed to be done by public (Canadian government) committee</td>
<td>Unknown</td>
<td>(Hannah et al., 1994b)</td>
</tr>
<tr>
<td>Körner Minimum Data Set</td>
<td>Körner MDS</td>
<td>1982</td>
<td>Standardize data</td>
<td>Unknown. Presumed to be by a public (British government committee)</td>
<td>Unknown</td>
<td>(Hannah et al., 1994b)</td>
</tr>
<tr>
<td>*Long Term Health Care Minimum Data Set</td>
<td>LTC MDS</td>
<td>1992</td>
<td>Provide submission guidelines (to government agency)</td>
<td>Unknown. Presumed to be by a public/private committee</td>
<td>Extensive pre/post implementation study was done-collecting 6 mos. of data (pre and post) of medical record review, assessments, interviews and observations</td>
<td>(McCormick et al., 1997; Renner &amp; Swart, 1997)</td>
</tr>
<tr>
<td>Name of Data Set</td>
<td>Acronym</td>
<td>Approx. Dev. Date</td>
<td>Concise Purpose</td>
<td>Method of Development</td>
<td>Method of Evaluation</td>
<td>Reference</td>
</tr>
<tr>
<td>----------------------------------</td>
<td>---------</td>
<td>-------------------</td>
<td>--------------------------</td>
<td>----------------------------------------</td>
<td>---------------------------------------</td>
<td>----------------------------------</td>
</tr>
<tr>
<td>Nursing Management Minimum Data Set</td>
<td>NMMDS</td>
<td>1997</td>
<td>Measure performance</td>
<td>Literature review</td>
<td>Consensus surveys</td>
<td>(Huber et al., 1997)</td>
</tr>
<tr>
<td>Nursing Minimum data set (The)</td>
<td>NMDS</td>
<td>1985</td>
<td>Standardize data</td>
<td>Invited conference</td>
<td>Open national review</td>
<td>(Werley, 1988)</td>
</tr>
<tr>
<td>*Office of Managed Care McData</td>
<td>McData</td>
<td>1994</td>
<td>Provide submission guidelines (to government agency)</td>
<td>Governmental Steering committee</td>
<td>Pilot feasibility study done in Washington State (details no found in literature search)</td>
<td>(McCormick et al., 1997)</td>
</tr>
<tr>
<td>Name of Data Set</td>
<td>Acronym</td>
<td>Approx. Dev. Date</td>
<td>Concise Purpose</td>
<td>Method of Development</td>
<td>Method of Evaluation</td>
<td>Reference</td>
</tr>
<tr>
<td>---------------------</td>
<td>---------</td>
<td>-------------------</td>
<td>---------------------</td>
<td>--------------------------------------------------------------------------------------</td>
<td>--------------------------------------------------------------------------------------</td>
<td>------------------------------------</td>
</tr>
</tbody>
</table>
| Patient Core Data Set | PCDS    | 1997              | Standardize data    | Literature review  
Frequency of occurrence study  
Expert panel/focus group                                                                               | Data availability study  
CVI survey  
Invitational conference of experts  
Plan for ease of collection and use of data study —details not found in literature | (McCormick et al., 1997) |
| Perioperative Minimum data set | PNDS  | 1996              | Standardize data    | Literature review  
Mail study with RNs matching interventions to outcomes                                                                                                              | 2nd round of mail intervention/outcome matching study  
Committee investigating relevancy (details or results not found in literature) | (Kleinbeck, 1996) |
<table>
<thead>
<tr>
<th>Name of Data Set</th>
<th>Acronym</th>
<th>Approx. Dev. Date</th>
<th>Concise purpose</th>
<th>Method of Development</th>
<th>Method of Evaluation</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recommendations for Uniform Reporting of Data</td>
<td></td>
<td>1998</td>
<td>Standardize data</td>
<td>Consensus (working group from International Trauma Anaesthesia and Critical Care Society (ITACCS))</td>
<td>Worldwide review and critique</td>
<td>(Dick &amp; Baskett, 1999)</td>
</tr>
<tr>
<td>*Standardized Outcome and Assessment Information Set for Home Health Care</td>
<td>OASIS</td>
<td>1996-1997</td>
<td>Measure performance</td>
<td>Public/private committee</td>
<td>Public/Private committee to evaluate usefulness (methods/details not found in literature) Field testing</td>
<td>(McCormick et al., 1997)</td>
</tr>
<tr>
<td>Name of Data Set</td>
<td>Acronym</td>
<td>Approx. Dev. Date</td>
<td>Concise Purpose</td>
<td>Method of Development</td>
<td>Method of Evaluation</td>
<td>Reference</td>
</tr>
<tr>
<td>------------------</td>
<td>---------</td>
<td>-------------------</td>
<td>----------------</td>
<td>-----------------------</td>
<td>---------------------</td>
<td>-----------</td>
</tr>
<tr>
<td>*Uniform Ambulatory Care Data Set</td>
<td>UACDS</td>
<td>1972, 1981, Rev. 1989</td>
<td>Standardize data</td>
<td>Unknown. Presumed to be by Public/private committee</td>
<td>Unknown. This data set is closely related to the HCFA 1500 which is regularly used for government purposes</td>
<td>(McCormick et al., 1997)</td>
</tr>
<tr>
<td>Uniform Data Set for Home Care and Hospice</td>
<td></td>
<td>1993</td>
<td>Standardize data</td>
<td>Consensus conference</td>
<td>Open national review Volunteer use (in &gt;60 agencies)</td>
<td>(McCormick et al., 1997; Renner &amp; Swart, 1997)</td>
</tr>
<tr>
<td>*Uniform Hospital Discharge Data Set</td>
<td>UHDDS</td>
<td>1970s, Rev. 1985</td>
<td>Provide submission guidelines (to government agency)</td>
<td>Expert focus group Public/private Committee</td>
<td>Unknown. Currently in use (closely related to the UB92)</td>
<td>(McCormick et al., 1997; Renner &amp; Swart, 1997)</td>
</tr>
<tr>
<td>Name of Data Set</td>
<td>Acronym</td>
<td>Approx. Dev. Date</td>
<td>Concise purpose</td>
<td>Method of Development</td>
<td>Method of Evaluation</td>
<td>Reference</td>
</tr>
<tr>
<td>--------------------------------------</td>
<td>---------</td>
<td>------------------</td>
<td>----------------</td>
<td>-----------------------</td>
<td>----------------------</td>
<td>----------------------</td>
</tr>
<tr>
<td>Uniform Pre-Hospital Emergency Medical Services (EMS) Data Set</td>
<td></td>
<td>1994</td>
<td>Standardize data</td>
<td>Expert focus group Working conference (Cosponsored by 9 Federal agencies)</td>
<td>Unknown</td>
<td>(Dean, 1994; Spaite et al., 1995)</td>
</tr>
</tbody>
</table>
APPENDIX B

CODING AND CLASSIFICATION SYSTEMS
<table>
<thead>
<tr>
<th>Name</th>
<th>Acronym</th>
<th>Approx. Development Date</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diagnostic and Statistical Manual of Mental Disorders</td>
<td>DSM-IV</td>
<td>4&lt;sup&gt;th&lt;/sup&gt; edition published in 1994</td>
<td><a href="http://www.psych.org">www.psych.org</a></td>
</tr>
<tr>
<td>Diagnostic-Related Groups</td>
<td>DRG</td>
<td>1983</td>
<td>(MEDI CODE, 1997)</td>
</tr>
<tr>
<td>International Classification of Diseases for Oncology</td>
<td>ICD-O</td>
<td>2&lt;sup&gt;nd&lt;/sup&gt; edition released in 1990</td>
<td><a href="http://www.who.int/dsa/cat98/disease8.htm">www.who.int/dsa/cat98/disease8.htm</a></td>
</tr>
<tr>
<td>International Classification of Diseases- ‘Revision Number’</td>
<td>ICD-‘X’ -</td>
<td>First release in 1900, Revision 9 released in 1979, Rev. 10 released in 1999</td>
<td><a href="http://www.cdc.gov/nchswww/about/major/dvs/icd9des.htm">www.cdc.gov/nchswww/about/major/dvs/icd9des.htm</a></td>
</tr>
<tr>
<td>International Classification of Primary Care</td>
<td>ICPC</td>
<td>1984</td>
<td>(Britt, 1997)</td>
</tr>
</tbody>
</table>

www.raecgp.aone.net.au/papers/integrity/sld039.htm www.wonca.org
<table>
<thead>
<tr>
<th>Name</th>
<th>Acronym</th>
<th>Approx. Development Date</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logical Observation Identifier Names and Codes</td>
<td>LOINC</td>
<td></td>
<td><a href="http://www.regenstrief.org/loinc/">http://www.regenstrief.org/loinc/</a> (Huff et al., 1998)</td>
</tr>
<tr>
<td>Medical Subject Headings</td>
<td>MeSH</td>
<td></td>
<td><a href="http://www.nlm.nih.gov/mesh/meshhome.html">www.nlm.nih.gov/mesh/meshhome.html</a></td>
</tr>
<tr>
<td>Medical Subject Identifier Names and loinc Codes (Huff et al., 1998)</td>
<td></td>
<td></td>
<td><a href="http://www.health.library.mcgill.ca/eguides/mesh.htm">www.health.library.mcgill.ca/eguides/mesh.htm</a></td>
</tr>
<tr>
<td>North American Nursing Diagnosis Association Taxonomy of Nursing Diagnoses</td>
<td>NANDA</td>
<td>Early 1980s</td>
<td>nelle.mc.duke.edu/standard s/termcode/nanda.htm</td>
</tr>
<tr>
<td>Nursing Interventions Classifications</td>
<td>NIC</td>
<td>1992 1st Ed.</td>
<td>(McCloskey et al., 1996)</td>
</tr>
<tr>
<td>Nursing Outcomes Classifications</td>
<td>NOC</td>
<td>1997 1st ed.</td>
<td>(Johnson et al., 1997)</td>
</tr>
<tr>
<td>Omaha Classification System</td>
<td>OCS</td>
<td>Development began in 1970</td>
<td>(Bowles &amp; Naylor, 1996)</td>
</tr>
<tr>
<td>Read Clinical Classification (AKA Clinical Terms Version 3 (CTV3) and Read Thesaurus)</td>
<td>RCC</td>
<td>Introduced in early 1980s</td>
<td><a href="http://www.coding.nhsia.nhs.uk">www.coding.nhsia.nhs.uk</a></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(Robinson, Schulz, Brown, &amp; Price, 1997)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td><a href="http://www.snomed.org">www.snomed.org</a></td>
</tr>
<tr>
<td>Name</td>
<td>Acronym</td>
<td>Approx. Development Date</td>
<td>Reference</td>
</tr>
<tr>
<td>-------------------------</td>
<td>---------</td>
<td>--------------------------</td>
<td>----------------------------</td>
</tr>
</tbody>
</table>
APPENDIX C

GENERAL CLINICAL DATA SETS
<table>
<thead>
<tr>
<th>Name of Data Set</th>
<th>Acronym</th>
<th>Purpose of Data Set</th>
</tr>
</thead>
<tbody>
<tr>
<td>Core Health Data Elements</td>
<td></td>
<td>To serve as a consensus point of data elements needed on persons and encounters or events that can serve multipurposes and benefit from standardization</td>
</tr>
<tr>
<td>Federal Recommended Data Elements</td>
<td>MQIS</td>
<td>Developed in response to the Effectiveness Initiative issued in April 1989 by the HCFA. Intended to obtain standardized clinical data for use by both Peer Review Organizations (PROs) as well as the general clinical research community</td>
</tr>
<tr>
<td>Medicare Quality Indicator System</td>
<td></td>
<td>A uniform billing system intended for use by third party payors, hospitals, skilled nursing and long term health care facilities and home health agencies. Required for Medicare/Medicaid reimbursement therefore implemented nationwide in US.</td>
</tr>
<tr>
<td>Financial Uniform Minimum Data Set</td>
<td>UB-92</td>
<td>The basic form prescribed by HCFA for the Medicare program for claims from physicians and suppliers, except for ambulance services. Also adopted by the Office of Civilian Health and Medical Program of the Uniformed Services (OCHAMPUS) Approved by the American Medical Association (AMA) Council on Medical Services.</td>
</tr>
<tr>
<td>Health Insurance Claim Form</td>
<td>HCFA 1500</td>
<td>Required of all grantees of a few agencies for their annual reports of aggregate data to the federal government</td>
</tr>
<tr>
<td>Health Resources Service Administration Uniform Data System</td>
<td>UDS</td>
<td>Health care data management to aid in the delivery of health services</td>
</tr>
<tr>
<td>Hospital Medical Records Institute</td>
<td>HMRI</td>
<td>Provides relevant, timely accurate information to assist health service managers</td>
</tr>
<tr>
<td>Körner Minimum Data Set</td>
<td>Körner MDS</td>
<td>Provides reporting requirements for Medicare-Medicaid recipients to support access and utilization monitoring, quality assurance and improvement review and capitation rate-setting</td>
</tr>
<tr>
<td>Name of Data Set</td>
<td>Acronym</td>
<td>Purpose of Data Set</td>
</tr>
<tr>
<td>---------------------------</td>
<td>---------</td>
<td>-----------------------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Patient Core Data Set</td>
<td>PCDS</td>
<td>Provides a basis for a longitudinal record of a patient’s health and medical history. Meant to allow transmission of patient data across all care settings for use by clinicians, administrators, researchers and health policy makers</td>
</tr>
</tbody>
</table>

**See Appendix A for references to individual data sets**
APPENDIX D

SPECIALIZED CLINICAL DATA SETS
<table>
<thead>
<tr>
<th>Name of Data Set</th>
<th>Acronym</th>
<th>Purpose of Data Set</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ambulatory Medical Care Minimum Data Set</td>
<td>AMC-MDS</td>
<td>Similar to the UHDDS but is based on an encounter in which the patient is neither hospitalized or institutionalized</td>
</tr>
<tr>
<td>Data Elements for Emergency Department Systems</td>
<td>DEEDS</td>
<td>Provide uniform specifications for data elements used in 24-hour, hospital based ED’s and provide a foundation for sharing data and communicating at the local, state and federal levels. Primarily developed for the reporting of conditions to the CDC.</td>
</tr>
<tr>
<td>Emergency Medicine Minimum Data Set</td>
<td>EMS-MDS</td>
<td>To standardize record keeping for each of the 300 EMS regions nationwide.</td>
</tr>
<tr>
<td>Essential Medical Data Set (The)</td>
<td>EMDS</td>
<td>Intended to supply a concise medical history data set. Complements DEEDS, not intended to be a complete longitudinal EMR</td>
</tr>
<tr>
<td>Health Plan Employer Data and Information Set</td>
<td>HEDIS 3.0</td>
<td>Measures the performance of managed health care plans</td>
</tr>
<tr>
<td>Long Term Health Care Minimum Data Set</td>
<td>LTC-MDS</td>
<td>All long-term health centers are required to submit patient assessments using the MDS for regulation purposes.</td>
</tr>
<tr>
<td>Nursing Management Minimum Data Set</td>
<td>NMMDS</td>
<td>Intended to empower nurses and nurse administrators with data to answer questions related to costs, quality and nursing resource allocation.</td>
</tr>
<tr>
<td>Nursing Minimum Data Set</td>
<td>NMDS</td>
<td>Core minimum data elements needed to describe nursing practice</td>
</tr>
<tr>
<td>Perioperative Minimum Data Set</td>
<td>PNDS</td>
<td>The purpose of the PNDS is to provide perioperative nurses with a standardized language for documenting and thus evaluating the care they provide.</td>
</tr>
<tr>
<td>Resident Assessment Instrument</td>
<td>RAI</td>
<td>To enhance clinical care and clinical understanding via standardized assessment.</td>
</tr>
<tr>
<td>Rotor Wing Transport Minimum Data Set</td>
<td></td>
<td>To standardize critical care transport data.</td>
</tr>
<tr>
<td>Name of Data Set</td>
<td>Acronym</td>
<td>Purpose of Data Set</td>
</tr>
<tr>
<td>-------------------------------------------------------</td>
<td>---------</td>
<td>---------------------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Standardized Outcome and Assessment Information Set for Home Health Care</td>
<td>OASIS</td>
<td>Proposed by HCFA for the purpose of advancing Medicare's outcome based quality improvement</td>
</tr>
<tr>
<td>Uniform Ambulatory Care Data Set</td>
<td>UACDS</td>
<td>Recommended as a uniform data set for ambulatory encounters. The HCFA 1500 Form is seen as a primary method of collecting the UACDS.</td>
</tr>
<tr>
<td>Uniform Data Set for Home Care and Hospice</td>
<td></td>
<td>Standardized data for home care and hospice.</td>
</tr>
<tr>
<td>Uniform Hospital Discharge Data Set</td>
<td>UHDDS</td>
<td>Prescribed by NCHS for Medicare reimbursement. Used to document different populations hospitalized in US, their reasons for using hosp. Resources and place of residence</td>
</tr>
<tr>
<td>Uniform Needs Assessment Uniform Pre-Hospital Emergency Medical Services (EMS) Data Set</td>
<td>UNAI</td>
<td>To evaluate needs for extended care and/or home health services for Medicare patients. To provide a uniform data set for EMS prehospital data collection for reporting, management and evaluation purposes.</td>
</tr>
</tbody>
</table>

**Please See Appendix A for references to individual data sets**
APPENDIX E

A DATA SET DEVELOPMENT METHODOLOGY
A Data Set Development Methodology

Developed by
Laura K. Heermann Langford

© 2001

“A Data Set Development Methodology” © L. Heermann Langford 2001
# TABLE OF CONTENTS

## INTRODUCTION
- Purpose of Patient Care Documentation ........................................... 114
- Knowledge Engineering and Knowledge Acquisition ......................... 115
  - Knowledge (Data Set) Engineers ...................................................... 116
- Requirements for Data Set Development ............................................. 116

## DATA SET DEVELOPMENT METHODOLOGY
- Stages of Data Set Development .......................................................... 118
  - Identification ...................................................................................... 118
  - Conceptualization ............................................................................. 121
  - Formalization .................................................................................. 122
  - Implementation ............................................................................... 127
  - Testing Stage .................................................................................. 128
- Documentation of Process ................................................................... 130
- Difficulties in Data Set Development .................................................. 130

## TECHNIQUES USED DURING FORMALIZATION
- Literature Review .............................................................................. 132
  - Definition ......................................................................................... 132
  - Description ....................................................................................... 132
  - Application of Technique ................................................................. 133
- Analysis of Previously Collected Data ................................................. 136
  - Definition ......................................................................................... 136
  - Description ....................................................................................... 136
  - Application of Technique ................................................................. 137
- Expert Interview/Knowledge Acquisition Interviews ........................ 138
  - Definition ......................................................................................... 138
  - Description ....................................................................................... 138
  - Application of Technique ................................................................. 139
- Focus Group Interviews ...................................................................... 149
  - Definition ......................................................................................... 149
  - Description ....................................................................................... 149
  - Application of Technique ................................................................. 150
- Summary ............................................................................................ 159

## REFERENCES ....................................................................................... 160

*Page numbers accurate for this "Appendix" Version, not the stand alone version.

"A Data Set Development Methodology" © L. Heermann Langford 2001
INTRODUCTION

Historically the development of data sets has occurred either via governmental organizations, national expert steering committees, national invited conferences, lengthy repeated surveys and/or Delphi studies. The development methods used are very time consuming, lengthy and expensive. The use of governmental organizations, national steering committees or invited conferences for data set development requires large amounts of funding to gather and host even a small number of people to work on developing a data set. The time between calling such a meeting, holding it, refining the work and disseminating the product is significant. Surveys and Delphi studies might be slightly cheaper but still consume a considerable amount of resources for development, mailing, collection and analysis. Years of intermittent work and “review periods” can go into one data set!

These methods of data set development are respectable and may produce a “reliable” and “valid” data set based on the criteria presented by the developing agency or special interest group. But, in the eyes of a developer interested in building a charting package that meets the needs of patient care, clinical communication and several vested interest groups, the data sets produced are often too limited in scope, incomplete, contradictory to each other and confusing.

The data set development methodology presented in this handbook addresses the need for an efficient, cost-effective method for developing data sets for use in computerized patient care systems. The premise of the methodology is that data sets should be based on data elements needed to support patient care and/or meet regulation requirements, but not include unnecessary data elements.

The standardized data set development methodology fits nicely within the larger Systems Development Life Cycle (SDLC) framework. The SDLC framework guides design, development, implementation and evaluation of computerized patient care systems (Whitten & Bentley, 1997). The data set development methodology is specifically useful during the design and development phases of the SDLC when requirements for a computerized patient care documentation system are determined. Specifically, the data set methodology supports identification of data elements for inclusion in the system.

“A Data Set Development Methodology” © L. Heermann Langford 2001
The methodology presented is intended to assist those who may have little or no formal training in clinical informatics principles but are charged with the development of a data set for use in a health care setting. It has been found that knowledge acquisition techniques as used in knowledge engineering for decision support systems are very applicable to data set development. Knowledge acquisition techniques are familiar to most clinicians and healthcare related professionals and do not require much training or instruction prior to application. In many instances, knowledge acquisition techniques are already being used by developers when determining data elements to include in a data set, but a formalized method to follow and apply the techniques has not been presented until now.

The benefit of this methodology is its capitalization of local resources and expertise. The use of these techniques results in a complete data set for the documentation of patient care without the time and expense involved in planning, organizing and hosting an invitational workgroup meeting of various interested parties and experts. The methodology recognizes that most often the "experts" in documenting care for any particular clinical area are the clinicians themselves and the users of the data set (Whitten & Bentley, 1997).

Purpose of Patient Care Documentation

Patient care documentation has many purposes. McDonald and Barnett (McDonald & Barnett, 1990) credit Lawrence Weed with the first analysis of the purposes of the medical record. These leaders determined the purposes of the medical record can be classified in three different functional areas, 1) facilitate patient care, 2) serve as a financial and legal record and 3) aid in clinical research (McDonald & Barnett, 1990). Although these functional areas are distinct, the primary purpose of the medical record for all these areas is communication (Meiner, 1999). Clinical communication may occur between caregivers in the same department providing immediate care to the patient as a team or to clinicians who care for the patient in the future. Patient care documentation is used for communicating information regarding a patient's treatment to any interested party whether it is the patient, a clinical researcher, risk manager, regulatory agency, insurance company or lawyer. Documented patient care data also assists and supports decisions made by clinicians when providing patient care.

While a data set itself would never qualify to be considered a decision support system, Hebda, Czar and Mascara (1998) use the definition of decision support systems by Semples (1996) that could apply to data sets used in patient care documentation. This definition states that decision support systems "aid in and strengthen the selection of viable options using the information of an organization or field to facilitate decision making and overall efficiency" (Hebda et al., 1998 p. 252). Data sets, when used inpatient care documentation systems, are often in the position of presenting "viable options" to clinicians. This is especially true when patient care documentation is done using predefined checklists, menus and other "data sets."

"A Data Set Development Methodology" © L. Heermann Langford 2001
Documentation consists of data and information in a variety of formats. Data are simply numbers, names, facts or other designators that are discrete entities not yet interpreted. Data elements that are given a form and structured are considered information (Saba & McCormick, 1996). Data sets are a method for providing some structure to a group of data elements. A data set such as the Nursing Minimum Data Set (NMDS) (Werley & Lang, 1988) is designed and structured for the specific purpose of “describing” nursing. The sixteen data elements included in the NMDS and only those sixteen, were chosen by the NMDS developers as having meaning to the science and service of nursing. The latest price of tomatoes is not included in the NMDS because it has no relevance or meaning in the context of nursing. It does however have meaning when included in the inventory data set of a grocery store and provides information regarding the revenue of the grocery store. So although individually, the data elements included in a data set are discrete items without meaning, once they are grouped into a data set for a specific purpose they do provide information. Information that is synthesized to identify and formalize interdependencies and relationships is knowledge (Graves & Corcoran, 1989). Data is not knowledge. But, knowledge has data as its foundation. The ability to create and use knowledge and the quality of knowledge is entirely dependant on the presence and quality of data and information.

Haug and colleagues (1994) present four categories of decision support in medicine. These four categories are alerting, critiquing, suggesting and retrospective (Haug et al., 1994). Of these categories “suggestion systems” could be applied to data sets. Data sets do not “suggest” therapies to be done or actions to be taken during patient care, but they do suggest what data items need to be recorded and possibly collected. Suggestions for documentation of data elements can be very useful to busy clinicians when juggling a busy patient assignment and all of the regulatory, legal and reimbursement needs of patient care documentation. The organization and arrangement of data elements can also be very useful to clinicians when providing care to patients and making decisions as to what actions should be taken. Careful planning of data elements included and their arrangement, whether it is on paper or in a computer, can also decrease confusion and promote efficiency (Cooper, 1995).

Knowledge Engineering and Knowledge Acquisition

With the influence of patient care documentation on the decision making done by clinicians, great care should be used in the development of data sets used for clinical care and the documentation of patient care findings and activities. The inconsistent nature of data set development could be improved by looking at techniques used in other areas of software development. van Bemmel and Musen (1997) define knowledge as facts and relationships used or needed to obtain insight or to solve problems. Acquisition is defined as the act of acquiring, i.e. to gain into possession (Gramercy, 1996). Knowledge acquisition is the process of extracting required knowledge to perform a task from sources such as reference sources, case histories and human experts (Hayes-Roth,
Waterman, & Lenat, 1983). Previous data set developments have included some of these practices, such as literature review and expert focus groups. Knowledge acquisition techniques are usually referred to as being useful to determine “knowledge” to be included in an expert system, but they can be used at a more granular level to assist with determining data elements to be included in a computerized patient care documentation system. A structured approach using the knowledge acquisition techniques of literature review to formulate a rough structure and basic “starter” data set that is refined by expert interviews and focus groups can streamline the data set development process while not compromising its integrity.

Knowledge (Data Set) Engineers

Knowledge acquisition is usually done by a knowledge engineer. A knowledge engineer is someone who engineers the knowledge, that is, converts knowledge from its existing form into a form applicable to the problem addressed and the tools (i.e., computer or otherwise) used (Hayes-Roth, 1984). With knowledge acquisition for decision support systems, knowledge engineers are interested in extracting the “rules” of a certain discipline or care process. With knowledge acquisition for data set development, data set developers (or data set engineers as they could be called) are interested in extracting data elements relevant to the desired goals and purposes of the new data set. These goals may include the standardization of data, to support communication between caregivers, support caregivers decision making (and the subsequent care they provide), measure quality and improve care or support any other reporting function. Ultimately the data set engineers strive for the appropriate data elements to support and guide the user’s need for information and knowledge generation regarding the topic at hand, be it a specific patient’s care or the care provided across several patients

Data set engineers, as is true of knowledge engineers, often have far less knowledge of the domain in question than an expert in the domain. Data set engineers act as an intermediary between domain experts and the tools and methodologies used for data set development and implementation. The job of the data set engineer is to structure the domain data and to identify and formalize domain concepts. This can often be a formidable task for a data set engineer who is not an expert in the domain. Even with the involvement of a supportive and interested domain expert, it is essential for the data set engineer to become somewhat familiar with the terminology, tasks and processes of the domain for the data elements and structure of the data set to be useful.

Requirements for Data Set Development

Certain requirements must be met before knowledge acquisition techniques can be successfully used for data set development. These requirements include:

“A Data Set Development Methodology” © L. Heermann Langford 2001
1. The purpose of the data set is explicitly stated
2. The scope of the data set is carefully defined
3. Reliable sources of the data and information used in the domain are available (including literature, domain experts and potential data set users)
4. Methods for resolving conflicts must be defined and applied when the information from separate sources differs
5. Procedures for reviewing, testing and updating data set content are clearly delineated. (Plan should include testing over time and revising based on the performance of the data set with actual patient cases.)
DATA SET DEVELOPMENT METHODOLOGY

The data set development methodology presented in this handbook is based on the knowledge acquisition process presented by Buchanan and colleagues (1983) in Hayes-Roth, Waterman and Lenat (1983). There are three phases to data set development. The first phase includes the identification and conceptualization stages, the second phase is the actual formalization stage of the data set and the third phase includes implementation and evaluation stages. Figure 1 demonstrates the three different phases of data set development and the activities (or stages) of each phase. An evaluation period is planned systematically after each stage of the development process to assure all objectives of the stage are met and required activities are completed before moving onto the next stage. On the diagram these evaluation periods are denoted as “QA.”

The methodology is broken into phases to indicate points where the product of the development process should be “frozen.” Freezing is used to control the development process thereby keeping structure and orderliness to the methodology. Once a phase has been completed and the product frozen, a developer cannot decide to “just go back and pick up or do one more thing.” The entire process is followed through to the present point before resuming the development process. Freezing the product at certain phases and not allowing haphazard “jumping” from one activity to another avoids the risk of incorrectly or inadequately implementing one piece (McConnell, 1996).

Stages of Data Set Development

Identification

The first stage to data set development is the identification stage. Activities included in the identification stage of data set development include:

1. Identifying current needs, problems or deficiencies
2. Identifying potential development team members and resources
3. Determining the roles of team members
4. Identifying the domain of the new data set
5. Determining the purpose of the new data set
6. Determining the scope of the new data set

“A Data Set Development Methodology” © L. Heermann Langford 2001
Figure 1. Data Set Development Methodology

Identification
1. Current Needs, Problems or Deficiencies
2. Team Members, Resources
3. Roles of team members
4. Domain of new data set
5. Purpose of new data set
6. Scope of new data set
7. Goals of new data set

Conceptualization
Key Concepts and relations clearly and explicitly defined

Criteria Not Met
QA
Criteria Met

Formalization
Data elements determined and organized using all available resources

Criteria Not Met
QA
Criteria Met

Implementation
On paper or computerized system

Criteria Not Met
QA
Criteria Met

Testing/Evaluation
Evaluate data set against predefined purpose and goals

Criteria Not Met
QA
Criteria Met

Data Set Completed
7. Determining the goals of the new data set

The data set development process begins with the identification of any problems, or deficiencies in the current system or processes for which a data set could improve. The initial identification of current needs, problems or deficiencies may be made by one of several activities. Instances that would instigate the development of a data set include the desire to install a computerized documentation system, the desire or need to redesign current documentation tools or the identification of unmet needs of a special interest group or research team. The needs, problems and/or deficiencies noted should be clearly written down.

Potential resources are identified before embarking on any development project. Most importantly the team members, specifically the data set engineer and a supportive domain expert, are needed for a data set development project to be successful. The number of data set engineers and domain experts collaborating on the data set development process depends on the domain and the intended scope of the data set. For simple, well confined data sets one in each role is probably sufficient. For larger, more complicated data sets, especially those with more than one domain (for example rehabilitation nursing and physical therapy) more data set engineers or domain experts may be required. The data set engineer is an essential position. The data set engineer is often the “glue” that holds the project together, keeping the project focused and on track. A supportive domain expert is also essential to the success of the development process. Frequent access to a willing expert is very helpful when deciding on scope restrictions and when determining basic concepts, primitive relations and definitions of data elements needed for the domain of interest.

All resources including knowledge sources, human and otherwise, time, money and computing facilities (if needed) are critical to the overall success of the project. Depending on the overall goal of the data set, the level of computing resources may differ. If the goal is to have a computerized application to collect, store and retrieve the data the database development personnel and computer resources required would be much greater than if the goal of the data set use is to be on paper forms. There must be adequate administrative support for all personnel, especially the data set engineer and collaborating expert, to spend the time required on the project. In addition, other sources such as books, journals, regulation manuals data sources, potential users must be available and accessible if needed.

The roles of the team members are dependent on the number of team members involved. Usually the domain expert(s) act as informants and monitors of the data elements evaluated. The data set engineer learns about the domain, formulates questions, keeps records of data elements considered, included and excluded from the data set and is the “project manager” for the data set development process. The data set engineer needs to discover enough about the domain of interest to feel comfortable conversing with experts and other users of the resulting data set.

"A Data Set Development Methodology" © L. Heermann Langford 2001
Once the data set development team is identified, the domain, purpose, scope and goals of the data set to be developed must be determined and clearly explicited. If these areas are not well defined, the project may tend to “creep” in its development, resources may be consumed and the endpoint becomes either difficult or impossible to obtain. During the formalization of the data set’s domain, purpose, scope and goals other problems or deficiencies in the current system or process are often identified. These should each be carefully examined, recorded and referred to while determining the scope, purpose and goals of the data set as they are often one of the primary reasons the data set is being developed or they may put additional constraints on the project.

Questions to ask that may be helpful during the identification stage of the development process include:

- Why do we need this data set?
- How do we see this data set fulfilling the need, solving the problem or correcting the deficiency?
- What resources are available to help with or use when developing this data set?
- What domain, or aspects of a domain is this data set to encompass?
- What is the purpose of this data set?
- What are the goals of this data set?
- How is this data set intended to be used?
- By whom is this data set to be used?
- Who is the intended audience of this data set?
- What is the nature and extent of the processes, activities and knowledge and/or decisions this data set will address?

The identification stage is an extremely important step and should not be taken lightly. Addressing each of the issues mentioned is essential to the success of the data set development process and the resulting product. It should be noted that the development of the domain, purpose and scope may require several rewrites before a workable and achievable scope for the project is determined (Buchanan et al., 1983).

Conceptualization

During the conceptualization stage the key concepts, relations and basic structure of the data set that characterize the domain of interest are clearly and explicitly defined. For example, a data set being addressed could have the domain of emergency nursing, the purpose to assist with and improve the documentation of nurses and a scope including only patients seen in the emergency department with a chief complaint of chest pain. Key concepts associated with this data set would include those important to the

“A Data Set Development Methodology” © L. Heermann Langford 2001
documentation of nursing care and findings such as past medical history, present illness history, physical assessment findings, interventions, patient responses and discharge/destination information. Each of these concepts may or may not have a relation to each other, such as the relationship of patient responses and certain interventions. For example, the intervention of pain medication and relief of pain are related, but the intervention of an intravenous line start and an increase in a patient’s oxygenation are not related. The basic structure of the data set would indicate relationships such as discharge/destination information being most likely the last data to be documented.

Questions that should be asked at this stage of the data set development include:

- What tasks are associated with the key processes, activities and decisions being supported?
- What are the data and information flow of these tasks, key processes, activities and decisions?
- Is there a hierarchy to these concepts or relationships? Can they be diagramed?

Several iterations of the data set may also evolve through this stage. The data set engineer will most likely have several interactions with the domain expert and other knowledge sources. Diagramming or outlining the concepts may be useful. This stage should not be too terribly time consuming. Only the KEY concepts are identified and related. More definition of the concepts will occur in the next, formalization stage.

Formalization

The formalization stage is where the key concepts are expanded and the data elements related to them are determined. The collaborating domain expert is very helpful and involved during this stage, but it is during the formalization stage when most development responsibility falls on the data set engineer. During formalization, the data set engineer makes the most use of knowledge acquisition techniques and sources of knowledge such as the literature, existing databases, other domain experts and potential users of the data set. It is during the formalization stage that resources such as literature sources, databases and domain experts are actually obtained, read, examined and interviewed. The formalization stage is when the data set actually begins to take shape and be developed.

A variety of knowledge sources can be used for the data set development process. van Bemmel and Musen (1997) divide knowledge sources into two categories: 1)
knowledge obtained from published sources and 2) knowledge obtained from domain experts. Published sources of knowledge can include textbooks, journals, policy and procedure manuals, regulatory agency publications, guidelines, protocols, standards, curriculums, training manuals, and user guides. In this electronic age, published sources of knowledge may even be found on the Internet on web pages and electronic journals. Data is usually obtained from the literature via structured literature searches. As with any source the quality of the knowledge source must be carefully scrutinized. Research and evidence-based sources are more desirable than unknown incidental types of sources. The data set engineer must have at least a basic understanding of the domain being researched or will need to work very closely with a domain expert when reviewing the literature.

Databases are also valuable sources of knowledge that should be considered for use during the knowledge acquisition process during data set development. Looking at how frequently data elements have been used in similar circumstances can be very useful when deciding which elements should be included in the new data set. Examining items stored as “free text” are also helpful in finding new data elements used that may not be evident in the literature or older databases.

Human experts are of course an invaluable resource when determining what data elements should be included in a data set. Although the published sources are extremely useful in identifying potential data elements, a human expert will have more information regarding the “idiosyncrasies” of data element use in the “real” world. For example, even if seven texts say a certain data element is essential to making a certain diagnosis, a human expert can explain how the test to obtain the data element is either too expensive and therefore not used, or not even available at the clinical site. Data is usually obtained from human experts either via a structured interview, unstructured interview or possibly even a focus group. It is not uncommon for the data set engineer to be an expert in the domain of study, but it is always prudent to involve another expert so as to not incorporate bias into the data set (Buchanan et al., 1983).

Figure 2 demonstrates a logical method to apply knowledge acquisition techniques for data set development during the formalization stage. The data engineer is able to obtain many of the data elements from nonhuman resources then confer with the collaborating domain expert(s) for review of the data elements for their accuracy and appropriateness for inclusion. Using sources other than human experts not only allows the identification of data elements to proceed in a methodical and consistent manner, but it also decreases the amount of time required of the domain expert. Presenting a data set obtained from textual and other nonhuman based sources to the domain experts will reduce the amount of time the experts need to “think of everything” for inclusion and will also aid the experts in remembering items they may otherwise forget.

The collaborating domain expert may be the only expert needed to provide input and feedback on the developing data set. Some projects may need other expert input on the
Figure 2. Application of Knowledge Acquisition Techniques for Formalization of Data Set

"A Data Set Development Methodology" © L. Heermana Langford 2001
data set and this should be done using structured or unstructured one on one interviews. In addition, it may be helpful to form a focus group of experts or other end users to provide feedback, input and direction on the data set as it is being developed. Involving a focus group of end users may not only help in the development of the data set, but may also improve acceptance of the data set by the users and its successful implementation (Whitten & Bentley, 1997).

Factors that are addressed during the formalization stage include the granularity of the data set and the characteristics of the data elements included. The granularity of the data refers to “how deep” the data set will define the task, key process, activity or decision supported by the data set. An example might be given using the key concept of interventions in the emergency nursing documentation data set. Figure 3 indicates how significantly decisions regarding the granularity of a data set affect its size and end user interpretation. To determine the granularity of the data set, ask, “what specific data are needed to support the key processes, activities and decisions being addressed?” The level of granularity needed is also affected by the purpose and intended uses of the data set. If the data set is being developed for a fully coded, pick list type computerized documentation system, the granularity detail needed will be much higher than if the data set will be implemented only on paper.

Data element characteristics addressed during the formalization stage includes determining how data are collected and or elicited, how frequently the data are collected (one time, a specified number of times or an unknown number of times), are the data readily available or difficult to obtain and once obtained are the data reliable, accurate and precise? Other characteristics of the data that should be considered are data collection costs and their order of occurrence over time.

It is important during the formalization stage of data set development to determine which standardized vocabulary will be used to represent the data elements included in the data set. Even if the data set being developed has a small scope and is intended for “just this one little stand alone clinic’s paper documentation” the use of a standardized vocabulary for the data elements included in the data set should be considered. Patient care documentation systems often “live” longer than ever expected during its developmental stages and the uses of the data collected change and evolve over time. To accommodate potential growth, data sharing and changes to the purpose and use of the data set it is important to have standardized terms and definitions for each of the data elements included. If the formalized vocabulary seems too “stiff” or unnatural for the intended use and purposes of the data set, at the very least a mapping of the data elements in the data set to related terms in the chosen standardized vocabulary should be made. Do not dismiss a specific standardized vocabulary if it does not contain specific terms needed by your data set. Submit any missing terms to the specific vocabulary’s parent group and contribute to the growth of the vocabulary in a meaningful way!
A. Intervention(s)
   a. Airway/Oxygen/Respiratory Management
   b. Intravenous Therapy
   c. Medications
      i. Currently taken on daily or PRN basis
         1. Name of drug
         2. Dosage
         3. Frequency taken
         4. Route taken
      ii. Given to Patient Prehospital
         1. Name of drug
         2. Dosage
         3. Time given
         4. Frequency given
         5. Route administered
         6. Patient response
      iii. Given to Patient in Emergency Department
         1. Name of drug
         2. Dosage
            a. A number range (0.00 - 000,000.00)
            b. Labeled as
               i. amp (ampule)
               ii. cap (capsule)
               iii. cc (cubic centimeter/milliliter)
               iv. g (gram)
               v. gr (grain)
               vi. in (inch)
               vii. kg (kilogram)
               viii. L (liter)
               ix. mcg (microgram)
               x. mEq (milliequivalent)
               xi. mg (milligram)
               xii. mg/ml (milligram/milliliter)
               xiii. ml (milliliter)
               xiv. oz (ounce)
               xv. supp (suppository)
               xvi. tab (tablet)
               xvii. tbsp (tablespoon)
               xviii. tsp (teaspoon)
               xix. u (unit)
         3. Time given
         4. Frequency given
         5. Route administered
            a. IV (intravenous)
            b. IM (intramuscular)
            c. PO (by mouth)
            d. SL (sublingual)

Figure 3: Data Set of Emergency Department Interventions with Granularity Indicated
Depending on the scope of the data set development process and the project's scope, the development of a database using the data set may need to be considered. If the data set is intended to be implemented in a computerized patient care documentation system, the data set engineer may consider employing some of the database development techniques such as developing an entity relationship diagram (ERD) with the accompanying data element attributes. Although these tools are usually not helpful for the domain expert(s) in assisting with the determination of data elements to include in the data set, they may be useful to the data set engineer to identify areas that need clarification from the domain expert(s) and formulate questions to be asked of the expert(s). Regardless of the actual implementation environment, it is useful for the data set engineer to be thinking of what methods would be most useful for collecting and presenting the data elements. Some data are likely to be conducive to the use of pick lists and or check boxes (such as past medical history) or graphical indication (laceration location and size/shape), while others need a repetitive spreadsheet type format for specific value entry (vital signs) or any combination of data entry possibilities.

Questions that should be asked during the formalization stage of the data set development include:

- What data elements characterize the data set?
- What data are required to support the key processes, tasks, activities and decisions included in the domain and scope of this data set?
- What data are available?
- What data are not available?
- If data are not available how could they become available?
- What is the granularity needed for this data set?

A significant amount of time can be spent in phase two during the formalization stage as the data set evolves. Accurately scoping the data set and carefully attending to the data and information needs of the users during phase one of the development process can dramatically affect the outcome. It is important to remember the process is developing a DATA SET not a DATA BASE. Once the inclusion or exclusion of data elements in the data set is decided then issues related to database development and the taxonomies used can be addressed.

Implementation

A data set development project may or may not continue through the implementation stage. If the purpose of the data set development is for a data set to be published or distributed to other sites and or users for implementation, it is very helpful for the data set to be clearly described and defined in the formalization stage. Any interpretation for implementation of the data set left to the end user allows for differences

“A Data Set Development Methodology” © L. Heermann Langford 2001
in actual implementation to occur. This ultimately defeats any purpose of standardization and possible sharing of data. Therefore, even for a "model" data set, the data set should be implemented as a prototype not only to demonstrate how the data set is to be implemented, but also to allow for needed refinements to the conceptualization and formalization of the data set.

Implementation of a data set can occur on several levels. Most easily the data set can be implemented in paper form or on computerized "screen" designs or for the more enthusiastic, a computerized database using the data set can be developed. The relations of the key concepts and data elements determined in the conceptualization and formalization stage should be closely adhered to and followed. The data types and repetitiveness of each data element are also considered and also represented in whatever medium is chosen for the prototype implementation. During the implementation of a data set any ambiguities not fully addressed in the formalization stage will be found.

Database development and taxonomy use are addressed during the implementation stage of data set development. However, this methodology does not address these issues as plenty of excellent resources are available for these activities if needed.

Questions that should be asked during the implementation stage of the data set development include:

- How is the data set to be used?
- Can the data set be installed on paper (lower cost and investment) instead of computerized to initially test it?

Testing Stage

The final stage of the knowledge acquisition process for data set development is the testing stage. This stage involves evaluating the data set against its purpose and goals. No preestablished gold standard for data set evaluation exists, but it is important to determine if the resulting data set meets the needs of its users as intended.

Very little detailed information is found in the literature regarding the evaluation of clinical data sets. Of data set descriptions that did have evaluation information, the criteria for evaluation included anything from determining the availability of data and feasibility of collection to extensive pre/post implementation and content validity studies (Harris, Graves, Solberg, Elkin, & Chute, 2000; Huber, Schumacher, & Delaney, 1997; Kleinbeck, 1996; McCormick, Remer, Mayes, Regan, & Greenberg, 1997; National Center for Injury Prevention and Control, 1997; NII-HIN Consortium, 1997; Renner & Swart, 1997; Thompson & Schaffer, 2000a, 2000b; Werley, 1988). Several studies regarding the validity and reliability of the Long Term Health Care Minimum Data Set

"A Data Set Development Methodology" © L. Heermann Langford 2001
(LTC MDS, or often just MDS) have been published. Correlational studies, hypothesis testing and factor analysis have been used to test the validity and/or reliability of parts of this data set by several authors (Casten, Lawton, Parmelee, & Kleban, 1998; Crooks, Schnelle, Ouslander, & McNees, 1995; Frederiksen, Tariot, & De Jonghe, 1996; Lawton et al., 1998; Snowden et al., 1999; Zulkowski, 1998). None of the studies evaluated the entire minimal data set.

Certainly data will be standardized if the data sets are implemented exactly as they are intended to be used. However, studies for data availability and collection feasibility as well as open national reviews may not measure how well data are “standardized” once the data set is implemented. Researchers would be most supportive of rigorous in depth study of all data sets, but unfortunately such studies are often time consuming, expensive and difficult to do.

One of the most straightforward methods of evaluating a data set would be to use the data set in the environment or domain for which it is intended. Use the data set on a variety of cases and note any strengths, weaknesses or failures. Examine the data collected and determine if the data set is fulfilling its purpose and meeting the goals of the project. Note any weaknesses or failures in the data set for use in future revisions of the data set.

A certain amount of evaluation has occurred throughout the data set development process as indicated by the “QA” (Quality Assurance) processes in Figure 1. The evaluation periods are helpful, but they do not guarantee all considerations necessary for each stage have been completed. Therefore it is important for a global evaluation of the data set to occur after the formalization stage. It is almost impossible to adequately conduct a global evaluation without implementing the data set in some form or another. Ending the development process after the formalization stage is very tempting. Implementing a data set can be very tedious, time consuming and difficult. It is necessary however to complete all stages including the implementation and testing stages to verify the data set meets the predetermined purpose and goals. Theoretically if the data set does not meet its purpose and goals, the development process is not over, but should continue in a recursive fashion until the evaluation of the data set is satisfactory.

Questions that should be asked during the testing stage of the data set development include:

- Does the data set meet the previously defined purpose?
- Does the data set meet the previously defined goals?

Creating data sets using the method described in this handbook creates data sets that are operational representations of knowledge, as it exists at the time the data set was developed. It is important to periodically review the data set, reevaluate it based on current purposes and goals and update the data set for current processes and knowledge used as needed.
Documentation of Process

Throughout the knowledge acquisition process for data set development it is very important to keep a list of references and experts used. In addition, a diary noting the justifications made for either including or excluding data elements should be kept. This documentation not only supports the data set once it is developed, but will also provide an invaluable reference tool during future revisions and/or explanations. The documentation can be kept in a tabular form with the suggested headings in Table 1.

Difficulties in Data Set Development

A common problem encountered with the knowledge acquisition process is collecting knowledge from only the first available or most convenient source. In most domains, knowledge is spread across several sources and it may take considerable effort to find and research them (Parsaye & Chignell, 1988). Knowledge acquisition is more of an art than a science that “requires introspection on the part of the expert, interpretation on the part of the knowledge engineer and considerable communication between the two” (Grosso, 1998 p. 165). It is often very difficult for the data set engineer to recognize subtle differences in data elements and just as difficult for a domain expert to express the knowledge he possesses in precise, complete and consistent terms (Buchanan et al., 1983). Fortunately, the difficulties encountered during the knowledge acquisition process are often overcome with patience and tenacity.

Table 1

Example table for documenting data elements considered for data set

<table>
<thead>
<tr>
<th>Data Element</th>
<th>Source</th>
<th>Citation</th>
<th>Included</th>
<th>Why or Why Not</th>
<th>Inclusion Confirmed By</th>
<th>Data Element Definition Source*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patient Last Name</td>
<td>Institution Policy</td>
<td>Ocean Side Hospital documentation policy</td>
<td>Y</td>
<td>Required for institutional, regulatory and clinical care purposes</td>
<td>Expert: Dr. Jones Focus Group: 11/17/00</td>
<td>DEEDS</td>
</tr>
</tbody>
</table>

* Indicates vocabulary, coding, classification or standard used for data definition

“A Data Set Development Methodology” © L. Heermann Langford 2001
Another difficulty encountered with the data set development process is the honoring of the freezing that occurs between phases. It may be very tempting to widen the scope of the project when deep into the formalization stage. Perhaps the literature review or expert interview reveals areas, topics or concepts not previously included in the scope of the project. Including these findings in the data set may seem helpful and natural but should not be done without careful consideration. In smaller projects it may be feasible to return to Phase 1 in the middle of the formalization stage and broaden the scope or add key concepts to the data set. The ramifications of doing so in a large project however may be too great to "jump" between Phases. In these instances, making a list for future revisions or changes to be addressed after the initial data set development process is completed may be more prudent. The decision to honor any freezing between phases is in reality a judgment call of the data set engineer and fellow developers. A key point to keep in mind however is that any jump back in the process requires strict progression through the methodology as described back to the present "jumping off" position. Quick jumps between stages should not be done, as the ramifications of any changes will not be fully known.
TECHNIQUES USED DURING FORMALIZATION

The formalization stage of data set development extensively uses the knowledge acquisition techniques of literature review, data analysis, expert interview and focus groups. Following is a more thorough description and explanation of each of these processes.

**Literature Review**

*Definition*

A literature review is the examination of the body of writings on a particular subject. In the context of knowledge acquisition for data set development, the literature review involves examining the writings on a particular subject to find data elements that may be useful in the data set that is desired.

*Description*

A literature review involves the identification of all sources of written and/or published knowledge regarding a certain topic. Theses sources are then found either in the local institution, through a local library or on the Internet. The items are screened for relevancy. The irrelevant items are discarded. All relevant items are read with the intention of gleaning data elements that may be useful for inclusion in the data set. All potentially relevant data elements are recorded and then organized according to the key concepts identified in the conceptualization stage. Duplicate data elements are deleted, retaining only one of each item. Items that are similar but not exactly duplicative are retained for clarification from other knowledge sources (such as the human expert) consulted later in the process.
Application of Technique

Why conduct a Literature Review for Data Set Development?
The healthcare literature can be very useful in determining what data elements should be included in a data set. The literature regarding any specific topic will often indicate what data elements are needed for communication, regulatory issues and clinical decision making. Using the literature to build a foundation for a data set not only gives the data set a solid starting point, but also provides a skeleton from which further investigations, discussions and development can evolve.

Who conducts the Literature Review?
The data set engineer, or someone closely involved in the data set development knowledge acquisition process should conduct the Literature Review. The person conducting the literature review must have at least a cursory familiarity with the topic being explored.

Guideline to conducting a Literature Review:

- **Begin with a Strategy.**
  In the case of data set development, the strategy for the literature review should be based on the key concepts identified in the conceptualization stage. Using the key concepts, determine keywords that can be used when searching computerized bibliographic databases and indexes.

- **Conduct a search of Bibliographic Databases**
  Computerized bibliographic databases are available at most libraries and over the Internet. Choosing the correct database to search for the specific data set desired is very important. More than one computerized bibliographic database may need to be consulted. Unless very familiar with the topic and the databases available, ask for assistance from the librarian when choosing a computerized database to search. Knowing how to use a computerized bibliographic database is also very important. It is essential to be familiar with the Boolean logic ("AND," "OR" and "NOT") used in most computerized bibliographic database search engines (Saba & McCormick, 1996). Again, if not familiar with these search techniques, ask a librarian for assistance. Be sure to access databases for both the library card catalogs and journals. Fortunately, most bibliographic databases include an annotation or abstract for each reference so that high level screening of relevancy to the topic can be done. Not all relevance screening can be done from an abstract or annotation, so be sure to add any potentially useful references to the list of "saved" references.
• Locate items listed on bibliographic database search results
   The listings of book and journal references generated by the bibliographic database searches are used to locate written materials in the library. Some resources may not be available in the local library but would be available via interlibrary loan or in the case of journal articles, by ordering copies from other libraries. Consult a librarian for questions regarding interlibrary loans and article ordering. In some libraries with large book collections just "browsing" the books on the shelf near those that appeared on your computerized bibliographic search may be useful. Even if they were not found during the keyword search, some neighboring books have a wealth of information on the topic of interest.

• Locate items not listed in bibliographic databases
   Not all written material useful to the data set development process will be indexed in a computerized bibliographic database. Some items such as policy and procedure manuals and protocols may only be accessible at the institution or unit for which the data set is intended. Be sure to look for materials that are specific to the unit if appropriate for the data set being developed.

   Other written materials not always found through a search of bibliographic databases include standards of practice, state or federal government regulations, specialty interest group guidelines and consensus statements and training manuals. Be sure to ask colleagues for all possible resources. Ask the collaborating domain expert or any other domain educator or expert for access to their personal libraries and recommendations.

• Screen items for relevancy to the data set
   Screening the written materials for relevancy to the project is usually done as the items are collected from the variety of locations. If this is not done during the location process then be sure to take time to do so before proceeding. Unfortunately in some cases it is not until an indepth study of the source is done that relevancy or irrelevancy can be determined. If in doubt, keep the reference for further study.

• Review literature sources for data elements to be included in the data set
   Examine the written material related to the subject that has been located for data elements relevant to the topic of the data set being developed. This can be done via a structured or unstructured review.

   • Structured Literature Review

"A Data Set Development Methodology" © L. Heermann Langford 2001
A structured literature review is a review of the literature that follows a specific predefined pattern of inquiry through each source. The pattern of inquiry may be based on the key concepts identified in the conceptualization stage or by following the task, process, activity or decision process that will be conducted while using the data set being developed. An example would be, when developing the data set for use in emergency nursing documentation, search the literature for all data elements related to present illness history first, then go on to look at those associated with the past medical history, prehospital care, physical exam, interventions etc.

- Unstructured Literature Review
  - An unstructured literature review is a review of the literature that does not have a specific pattern of inquiry but rather has somewhat of a “life of its own.” With the unstructured literature review the data set engineer examines all relevant literature looking for any data element that may be relevant to the data set being developed. The unstructured review may be more time consuming than a structured review, but it may also find relevant data elements not found during a structured review.

For both structured and unstructured literature reviews be sure to keep in mind the scope of the data set. During literature reviews it is easy to forget the purpose of the search and begin recording all data elements found or get distracted and pursue avenues of inquiry that are not related to the project at hand. The literature can be interesting and distracting! Write a set of question to ask yourself frequently while conducting the literature review. Answering the questions may help in maintaining the proper focus.

Example questions to ask while conducting a Literature Review for data set development include:

- What do we need to know to care for this patient?
- What do we need to know to accomplish this task?
- What do we need to record to complete this process?
- What do we need to know to make this decision?

- Organize the data elements
  The organization of the data elements collected via the literature review is very important. The organization of the data elements
should follow some logical order evident from the key concepts, processes or activities associated with the data set. The organization of the data elements may start early as potential data elements are recorded from the first literature source or at any time during or after the literature review process. The logical organization of the relevant data elements found in the literature will allow for finding and deleting duplicates, finding similarities needing clarification, finding holes or gaps of knowledge and/or finding data elements also needing clarification. During the first pass through the literature the data set engineer is able to develop the “starter” data set. From this starter set all other work will revolve, through the addition, deletion and clarification of data elements. Therefore doing a thorough review of the literature when beginning the data set development process is important. Don’t be surprised however if the literature brings up more questions than it answers. Be careful not to agonize over the literature review. Time should be spent doing a thorough job, but don’t get stymied in the literature review. You can always come back if questions come up in the future!

- Represent the data set on paper
  Representing the starter data set on paper can be a challenge. Forms such as outlines, spreadsheets and tables may be helpful. Stay away from screen designs and flow diagrams until later in the process. The early starter data set is a “brain dump.” Space and other limitations set by screen designs and flow diagrams may restrict the growth of the data set too severely.

Analysis of Previously Collected Data

Definition

Data analysis is defined as the systematic organization and synthesis of discrete entities, such as numbers, names or facts in order to interpret and give meaning to them.

Description

Discrete data elements are organized in a logical and/or meaningful manner either manually or with the assistance of computerized tools. The organization allows for mathematical, statistical or visual interpretation of the data into meaningful information.
**Application of Technique**

Why conduct Data Analysis for Data Set Development?

Data previously collected during the tasks, key processes or activities intended to be documented or supported with the data set under development can be very useful in determining what data elements should be included in the new data set. Unfortunately reviewing and analyzing the data previously collected can be a formidable task. Even when available, data from computerized databases are often unstructured and incomplete. If the data stored in computerized databases is captured and stored in text blocks or other unstructured formats searching for the existence or frequency of use of individual data elements may require either sophisticated text parsing programs or manual review. “Manual chart reviews are an option but the data are inevitably incomplete or phrased in a wide variety of ways” (Warner, Sorenson, & Bouhaddou, 1997 p. 47). In addition, the retrieval of such data can be very time consuming and expensive for little or unpredictable reward. If resources allow however, performing a simple data analysis of existing databases may be useful when developing a new data set.

Who conducts the Analysis of Previously Collected Data?

The data set engineer, or someone closely involved in the data set development knowledge acquisition process should conduct the data analysis. Others outside of the data set development process (such as a database administrator) can run queries against the database if the queries needed are explicitly stated. The data set engineer or other involved data set development person should do the final analysis.

Guideline to conducting an Analysis of Previously Collected Data:

- Determine the frequency data element(s) have been used in the past.

  The analysis can be done using database and statistical methods that determine the frequency a data element has or has not been used. Or, looking at “free text” areas in computerized databases may indicate data elements that are not found in the existing dataset but are frequently used. Other valuable information found in free text areas of computerized databases are data elements that are included in an existing data set but are presented or defined in a way not useful to the user.

**EXAMPLE!** An example of data analysis for knowledge acquisition in the data set development process is examining previously documented heart rhythms. A literature review or older data set may indicate 30 different heart rhythms could be documented. An electronic or manual review of actual heart rhythms documented may show only 10 rhythms.
are routinely documented for 99.9% of patients seen. These findings may indicate including only those 10 specific data elements and an area for free text for the remaining 0.1% would be adequate in the new data set. Using data analysis can help simplify a potentially long and complex data set by not including "every" possible data element. Decisions for inclusion and exclusion of data elements based on data analysis should be made carefully and in accordance with the purpose and goals of the data set.

- Add or clarify data elements as indicated by the data analysis
  The data analysis may clearly indicate data elements that should be included in the new data set being developed. It may also assist in clarifying any data elements identified during the Literature Review process that were ambiguous or confusing. Rarely should a data element be deleted from the developing data set based on the findings of a data analysis. At most the data elements in question for deletion should be marked for further review and discussion with the experts and focus groups.

- Update the data set on paper as needed

**Expert Interview/Knowledge Acquisition Interviews**

**Definition**

"An expert is a person who does not have to think, he knows."

*Frank Lloyd Wright, from Frenzel, 1987 p. 129*

An expert interview is a face-to-face verbal interchange at which information is obtained from an expert, or someone who has special skills or knowledge representing mastery of a particular subject. In the case of data set development, the data set engineer conducts the interview with an expert in the domain or topic of the data set being developed. Information is usually obtained via direct questioning of the expert by the data set engineer (Waltz, Strickland, & Lenz, 1991).

**Description**

In the data set development process expert interviews consist of one person, the data set engineer, asking another person, the domain expert, specific questions regarding the data set being developed. These questions range from the scope of the data set (used
early in the data set development process) to which specific data elements should be included or clarification of data elements already chosen for inclusion. During the data set development process, several interviews are likely with the collaborating domain expert. Other domain experts with specific domain knowledge not possessed by the collaborating domain expert are also interviewed at least once during the data set development process. The number of domain experts will depend on the scope of the data set and its purpose(s). For data set development the expert interview uses the data set developed up to the point of the interview as a working document from which the interview questions are based. The output of the expert interview includes additions, deletions and clarifications of data elements in the data set. Careful documentation of interviews is kept for future reference during the development process and as supporting evidence for the resulting data set as well.

Application of Technique

Why conduct an Expert Interview(s) for Data Set Development?

An expert interview is used in data set development to elicit individual expertise, knowledge, standards, attitudes or opinions not accessible in the written literature sources. Gaps in the data set may exist following literature reviews and data analyses that are only filled by incorporating the expert interview knowledge acquisition technique. In addition, the knowledge contained in books may not be a part of practical reasoning or may have exceptions that the expert has learned to work around (Parsaye & Chignell, 1988). In many domains, certain knowledge may never have been written down and is only available to experts (Parsaye & Chignell, 1988). Caution must be used however as all knowledge obtained from an expert cannot be assumed to be correct.

Who conducts the Expert Interview?

The data set engineer, or someone closely involved in the data set development knowledge acquisition process should conduct the Expert Interview(s).

Obtaining the Expert(s)

- Find the Expert(s)

The best way to find local experts is to ask people within the domain area of the developing data set. Others will quickly tell you who the experts are in certain topic areas. Frequently the expert is the busiest and most highly regarded individual in the organization.

Remember to look for experts with all types of expertise that may be supported by the data set under development. This may include not only clinical experts of all disciplines, but also legal
experts, billing and reimbursement experts and regulatory experts such as those knowledgeable about HCFMA (Health Care Financing Administration), HIPAA (Health Information Portability and Accountability Act) and JCAHO (Joint Commission on Accreditation of Healthcare Organizations) requirements.

- **Contact the Expert**

  Experts may initially be contacted and invited to participate in the data set development process in person or by a personal telephone call. Initial contact should be made verbally as it enables the data set engineer to immediately answer any questions the expert may have and it is a “warmer” more inviting technique when recruiting an expert. A person to person verbal contact also allows the data set engineer to know the expert has been contacted, made aware of the invitation to participate and whether the expert is willing to participate. Once initial verbal contact has been made it is prudent and respectful to confirm via a letter or phone call (Scott, Clayton, & Gibson, 1991).

- **Qualify the Expert**

  It is important to qualify the expert with an initial interview to verify the individual has the knowledge needed and is willing to work with you.

- **Challenges of working with Experts**

  The greatest challenge working with experts is often scheduling the expert’s time to conduct the interview and work on the data set (Frenzel, 1987). Experts are often involved in many activities and have many demands on their time. Flexibility on the part of the data set engineer is essential to contacting and scheduling experts.

  When utilizing experts for the development of decision support or expert systems, the experts are frequently reluctant to readily share their years of education and experience. They often realize they have valuable knowledge and want to protect it for personal or organizational interests or security. There is often a “fear” of being replaced by a computer (Frenzel, 1987). Fortunately the threat or fear of “being replaced” by a data set is not as great as with a decision support or expert system. Experts may actually be excited to participate in the development of a data set if they see it as a tool that will assist them with their job.

“A Data Set Development Methodology” © L. Heermann Langford 2001
Goals of Expert Interviews

The goal of an expert interview for data set development is to provide a context within which the expert can explore and explain why certain data elements should or should not be included in a data set and how they should be meaningfully organized. This goal is accompanied by the goal of obtaining a full record of the expert’s descriptions and explanations (Parsaye & Chignell, 1988) for use in the development of the data set and its documentation. The expert interview should be considered a “team building” experience, an effort to establish a good working relationship and an opportunity to determine a shared vocabulary for communication between the data set engineer and domain expert (Scott et al., 1991).

Guidelines for Expert Interviews

Several guidelines are recommended when conducting an expert interview for the acquisition of knowledge for data set development. The interviewer (data set engineer) should have a good understanding of the knowledge domain being investigated via the interviewee (domain expert). The expert should also be treated with respect, patience and encouraged to foster comfort and a feeling of inclusion in the entire data set development process. Other important guidelines to follow during expert interviews for data set development include (Cordingley, 1989; Parsaye & Chignell, 1988):

1. Be specific about what is expected from the expert and the outcome desired
2. Always remain open to learning or discovering something unexpected
3. Learning will occur in both directions, but the data set engineer will always learn more from the expert than vice versa
4. Speak the language of the expert so that it is possible to concentrate on the task
5. Allow the expert to provide material in a way most natural to them. Note the organization of the presentation of information. It may be useful to the organization of the data elements
6. Listen to how the expert partitions data elements and knowledge
7. Listen. Do not interrupt. Allow the expert to complete thoughts without interruption.
8. Avoid attacks or excessive use of jargon
9. Monitor nonverbal behaviors of self and expert
10. Ensure the environment remains comfortable and conducive to the exchange of information
11. Be conscious and respectful of time
12. Record the interview in as complete a fashion as possible for thorough analysis at a later time
13. Plan adequately for the time and energy involved in eliciting the information as well as reviewing and supplementing notes taken during the interview and after the interview
Types of Expert Interviews:

- **Formal, planned event (May, 1991)**
  The date, time and place of interview is scheduled in advance of its occurrence

- **Informal, unplanned event (May, 1991)**
  The date, time and place of the interview are not scheduled, but occur “impromptu” as the need or opportunity arrives.

**HINT!**
In data set development informal, unplanned interviews are most useful when only a small amount of clarification is needed on one to two data elements. If more information is needed it is best to conduct a planned formal interview.

- **Structured interview**
  During structured interviews the interviewer exercises maximum control by using only predetermined wording and sequence of questions (Waltz et al., 1991). In strictly structured interviews no deviation or clarification from the planned questioning is allowed (Waltz et al., 1991).

- **Unstructured interview**
  Unstructured interviews have no prescribed wording or sequencing to the questions asked. These decisions are left up to the individual interviewer. The extreme unstructured interview consists of no structure at all. During these interviews the interviewee, or respondent, is asked to talk about whatever they wish regarding the topic. A list of questions or agenda can be followed, but the interviewee is not required to cover all of it. This type of interview is sometimes called an informal conversational interview (Waltz et al., 1991).

- **Focused interview**
  A focused interview begins with a loosely stated agenda such as the topics to be covered, but the flow between topics is flexible and the depth or amount of time spent on each topic is variable. During the focused interview, the interviewer keeps the agenda in mind and covers all topic areas, but will also allow the interviewee to discuss other relevant topics to various degrees. The only requirement of a focused interview is for all topic areas included on the agenda to be
covered by the end of the interview. In a focused interview, the order and approach to covering the topic is not dictated (Waltz et al., 1991).

**NOTE!**

Expert Interviews are often a combination of these interview types. Planned focused Expert Interviews are best early in the data set development process while unplanned, unstructured interviews may be needed later in the development process.

How to Conduct an Expert Interview

The key to a successful interview is careful planning and preparation. The expert has been notified by phone and a follow up letter sent. To adequately prepare for the interview the data set engineer plans the content of the interview, arranges the time and location of the interview, identifies necessary materials, prepares the expert and prepares him or herself (Scott et al., 1991).

- **Plan Interview Content**
  
  The content of the interview is centered on what specifically is needed to clarify or fill any gaps in the data set previously developed from the literature review and data analysis. The goals and purpose of the data set are reviewed, highlighting specific ones needing addressed during the interview with the expert (Scott et al., 1991; Waltz et al., 1991). Activities helpful to the data set development process are identified, such as presenting an intermediate representation of the dataset (i.e. an outline), discussing a particular case or situation or asking the expert to expound on a specific topic area. An agenda is developed listing each activity and questions or items needing to be addressed. The agenda should include a proposed schedule for the interview with an estimated amount of time for each agenda item. Do not under estimate the amount of time needed for each activity (Hart, 1992; Scott et al., 1991; Tuthill, 1990).

  When planning the content of the interview it is also important to identify the data set engineer’s own assumptions and prejudices. It is possible for the interviewer to influence the interviewee by wording of questions, or by nonverbal communications used. By identifying where the assumptions and prejudices lie, the data set engineer is better able to guard against biasing the responses obtained from the domain expert (Scott et al., 1991).

- **Arrange Time and Location**
  
  The time chosen for the interview should be convenient to all participants to maximize the effectiveness of the expert interview. Be
sure the time allotted for the interview with the expert is adequate. Scheduling more than one interview with the expert may be wise in order to adequately accomplish all that is needed (Hart, 1992; Scott et al., 1991; Tuthill, 1990). The frequency and duration of the expert interview(s) will vary according to the needs of the project and the schedules of participants (Scott et al., 1991). One long meeting may work better than several shorter meetings or vice versa. Meetings should be scheduled far enough in advance that a convenient time can be found for both the data set engineer and the domain expert and to allow each to adequately prepare (Scott et al., 1991). If more than one interview is necessary, subsequent interviews can be scheduled at the end of each preceding meeting or by setting a standing appointment time and place.

The location of the interview is also critical to the overall success of knowledge acquisition during an expert interview. A room large enough to comfortably accommodate all participants and accompanying materials is required. Each participant needs a writing surface, access to shared materials and adequate lighting of their workspace. Seating must allow for face-to-face communications without discomfort or straining. The room should be quiet and have a door that can be closed to decrease distractions. Consider meeting in a place where the expert cannot be found and interrupted. Remember the expert is often one of the busiest persons around and in the greatest demand! If possible, arrange with the expert for all business to be done before and after the interview so as not to be interrupted (Tuthill, 1990). Ask if pagers and cell phones can be left with another person for the retrieval of messages to be answered after the interview.

- Identify Necessary Materials

Using the planned agenda, identify what materials are needed for the interview activities. Materials that may be needed include those used for display of the intermediate data set, references, case studies etc. These display materials could be books, references, whiteboards, overheads, laptops with projectors, or copies of previously developed materials regarding the data set being developed (Scott et al., 1991). For each material item identified as being needed at the interview, arrangements for its procurement, transportation and setup need to be made. Most material items are usually the responsibility of the data set engineer, except in the case of specialty items or case studies agreed to be provided by the domain expert. Other materials standard to all interviews are those needed to record
the interview such as paper and pencil or a tape recorder (Scott et al., 1991 p. 364).

- Prepare the Expert

  Consideration for the expert includes adequately preparing the expert for the upcoming interview. Be sure the expert knows the date, time and location of the interview well in advance so that he or she may adequately plan and prepare. Provide the domain expert with a copy of the agenda in advance of the meeting and again at the time of the meeting. Remind the expert close to the time of the interview what items he or she is to bring if necessary. Offer to assist with the preparation, collection and transport of the materials provided by the expert if appropriate.

- Prepare the Knowledge/Data Set Engineer

  It is very important for the data set engineer to be familiar with the topic and the terminology of the data set domain. This preparation is usually done throughout the literature review process. Shortly before the interview, review the topic, terminology, agenda, goals and notes from earlier meetings to have all topics fresh in your mind. Have photocopies, overheads and other materials prepared in advance and set up before the meeting time begins. Comments and/or questions for discussion should be prepared and organized for easy referral. Allow plenty of time to travel to, arrive and prepare at the meeting site before the interview begins (Scott et al., 1991). Do not be late or make yourself feel rushed as these feelings could spill over into the interview and decrease its effectiveness.

Planning for an expert interview is imperative to its success. The interview will not be successful in obtaining the objectives and goals as desired without careful planning and preparation of the interview, the room, expert and data set engineer (Tuthill, 1990).

Structure of an Expert Interview

Each interview should consist of three segments. These segments are the introduction, body and closing (Scott et al., 1991). Following this structure not only allows for a smooth interview process but also keeps all participants oriented to the process and the task at hand.

- Introduction

  Each interview session should begin with the brief exchange of pleasantries and the introduction of any parties new to, participating in or observing the interview (if a third party is involved). Allow
everyone to get settled, materials prepared and take comfort measures, such as obtaining a glass of water, before starting. Optimally these activities would start before the actual meeting time so the interview can start on time. Start the meeting by reviewing the agenda with the domain expert and explaining the goals and objectives of the interview. If this is not the first interview with the expert, summarize the discussion of the previous meeting and accomplishments on the data set development since then. If this is the first meeting, be sure to explain and get permission from the expert for the use of any recording equipment, such as a tape recorder if needed.

➢ Set the tone

The tone of the meeting is set during the introduction presented by the data set knowledge engineer. “The tone should be professional, low risk and comfortable” (Tuthill, 1990 p. 199). The organization, statement of objectives, dress, manners and demeanor of the interviewer set the tone for a professional interview. A smile and a handshake are always a good start! The tone of low risk is communicated to the domain expert by reassuring the expert’s contribution is invaluable, by making status or progress reports on the project and by small talk (Tuthill, 1990). A comfortable tone is ensured with a well planned session, no schedule or time conflicts and flexibility of the interviewer (Tuthill, 1990).

• Body of an Interview

The body of an interview follows the prepared agenda using the questions and or discussion topics as planned. Adjustments to the scheduled time allowed for each topic area should be made as needed, but the discussion should not be allowed to get too far off track and consume too much time without a clear focus. Use the prepared questions and discussion topics to assist with maintaining a smooth flow throughout the interview. Use this time in the interview to ask questions, clarify meanings, receive and/or make suggestions and challenge inconsistencies (Tuthill, 1990).

➢ Types of Interview Questions

• Closed ended

Closed ended questions supply the respondent with two or more answers for the respondent to choose from. Examples of closed ended question are those that are answered by yes or no (dichotomous) and by choosing one of a list of choices (multiple choice).

• Open ended.
Open-ended questions are without a prespecified list of responses. Experts are asked to answer open-ended questions in their own words (Waltz et al., 1991).

Questions used during an expert interview for data set development are likely to be a combination of these depending on the objectives, goals and needs of the interview. Questions may or may not be written prior to the actual interview. The data set engineer may find that writing the topic areas and clarifications required are all that is needed and each question will be formulated during the actual discussion with the expert.

**HINTS:**

**Knowledge Acquisition Form:**

During the interview consider using a knowledge acquisition form for the data set being developed (Tuthill, 1990). A knowledge acquisition form is created for the specific task at hand and purpose of the expert interview. A knowledge acquisition form for an expert interview in the data set development process contains the outlined data set created from the literature review and data analysis with questions, gaps and topics for discussion noted and/or highlighted in some way. The knowledge acquisition form can be organized according to the key concepts of the data set being developed or in alignment with the agenda for the expert interview of the day.

**Verbal and Nonverbal Cues:**

It is important for the data set engineer to attend to the verbal and nonverbal cues exhibited by both himself/herself and the domain expert throughout the expert interview (Tuthill, 1990). Facial expressions, vocal inflections, speaking rate, pauses, hand gestures, body movements and posture can indicate how well an interview is going and the importance or insignificance of data elements for the data set. Also be cognizant of signals that indicate a tired, uneasy, frustrated or worried domain expert. Each of these emotions could be cues the discussion is not really being productive or effective for the data set development process.

- **Closing the Interview**

  The interview should end at or before the designated ending time of the interview. Allow time for all participants to make any final
remarks. To close the interview, a summary of what was discussed and accomplished during the interview is made and areas needing more discussion or on the agenda but not yet addressed are identified. New topics needing discussion or further attention that were identified during the interview should be noted. Plans for accomplishing completion of the outstanding topics are made and a tentative agenda or list of topics for the next meeting is noted. If more meetings will occur, interim tasks for the data set engineer and domain expert should be reviewed and the next meeting scheduled. Even if an actual meeting time and subsequent interview are not scheduled at the close of the first interview, build in provisions for later interactions or interviewing to allow for clarification on points that may become unclear or ambiguous to the data set engineer later in the development process (May, 1991).

Capture of Information During an Expert Interview

One of the most important tasks of the data set engineer during the expert interview process is the recording of the discussion and decisions made regarding the data set's data elements and their organization. Note taking is the most common and usually the most helpful, method used to record discussion and events during the interview. Note taking allows expert input to be highlighted, record explanations and ideas on structure and relationships of data elements, indicate particularly important information and note topics needing further discussion or research (Scott et al., 1991). Notes taken during an expert interview should be concise yet detailed enough to accurately record complex discussion after the interview has ended. It is very important to keep up with the discussion and not get too bogged down in the note taking. Use or create a “shorthand” note taking style if needed (Hart, 1992; Scott et al., 1991). Other methods of taking notes that could be considered are the use of flip charts, whiteboards, audio recording or even video recording. The use of electronic recording equipment can be invaluable for complex data set domains. An expert’s time is often very valuable and difficult to obtain. It would be unfortunate to have to repeat an interview or do extensive follow up due to the lack of adequate note taking (Frenzel, 1987). If using audio and/or video recording equipment be sure to know how to use the equipment well before relying on it to capture all information of the expert interview.

REMINDER! It is a good idea to get the expert’s consent for the use of electronic recording devices before doing so.
Potential Problems with Expert Interviews

The biggest problem encountered with expert interviews is the lack of adequate planning and/or preparation. Without some structure and domain familiarity on the part of the data set engineer, the interview encounter can be filled with inefficiencies, frustrations and the loss of time (Tuthill, 1990). In addition, a lack of preparation can result in insufficient exploration of the expertise the domain expert has to offer (Parsaye & Chignell, 1988). It is essential to cover the range of the domain expert’s contributions while at the same time not leading the expert into preconceived descriptions or behaviors (Parsaye & Chignell, 1988).

Focus Group Interviews

Definition

A focus group is an interview where the respondents are a group of individuals assembled to answer questions on a given topic (Polit & Hungler, 1999a). Focus groups are moderated by someone who presents questions for discussion, keeps the discussion on track as needed and records all relevant points or decisions made by the group (Morgan, 1988a, 1988b; Polit & Hungler, 1999a). In data set development, focus groups are used for discussions to obtain consensus on data elements questionable for inclusion or exclusion and at times will result in the generation of new data elements not yet considered for inclusion in the data set.

Description

Focus groups consist of 7 to 10 participants (Krueger, 1994). In data set development, the participants may or may not be intended users of the data set once it is developed, but they must have knowledge about the domain of the data set. The group is convened by the moderator or data set engineer and presented with a draft of the data set determined from the previous knowledge acquisition activities. Questions are posed to the group regarding the inclusion or exclusion of certain data elements, groups of data elements or their organization. Discussion is held for each question with the goal being to obtain a consensus from the group for each data element or data group discussed. If consensus is not obtained by the group decisions regarding the inclusion, exclusion or organization of data elements are left to the data set engineer and any domain experts acting as advisors for the project. As throughout the data set development process, careful documentation should be kept of all discussions and decisions made.
Application of Technique

Why Use Focus Groups for Data Set Development?

Morgan (Morgan, 1988a) contends a strength of using focus groups includes the ability to explore poorly understood topics and discover new insights. Focus groups allow considerable flexibility in the questions that are asked and the responses are inherently entirely up to the respondents, but they are not meant for all circumstances (Morgan, 1988a). Focus groups should be used for a well defined purpose (Morgan, 1988a). Focus groups are best when 1) there is a gap of knowledge between sources that has not been filled or resolved, 2) when investigating complex behavior and motivations, 3) when there is a desire to understand diversity and 4) when there is a need for a friendly respectful research method (Morgan, 1988a).

The greatest reason for using focus groups during the data set development process is its efficiency in obtaining the viewpoint of many busy individuals in a short amount of time (Polit & Hungler, 1999a). Focus groups should not be considered a substitute for individual expert interviews (Morgan, 1988b). But they do allow for discussion regarding data points between data set development participants of different backgrounds and or levels of experience.

When developing a data set for the documentation of clinical data it is often important to query the needs of clinicians at all experience levels since the data set may be used by them to support the care each is providing. Novices and experts make decisions regarding patient care and their documentation differently (Benner, 1984). However, clinical documentation tools and the data sets contained in them, whether computerized or on paper, must meet the needs of all clinicians regardless of their clinical expertise. It could be said a novice clinician is an expert in the needs of the novice level of documentation just as an expert clinician is an expert in the needs for their level of documentation. The consensus discussions of several individuals with varied knowledge and experience in the domain of the data set being developed can be extremely important in the decisions regarding the inclusion, exclusion or organization of data elements.

Morgan also states how useful focus groups are as a follow up technique to clarify findings of other research (Morgan, 1988b). Conducting the focus groups is best after a “skeleton” data set has been developed using the literature review, data analysis and expert interview techniques. This approach gives the group something to work with and a focus for their discussion. Without a baseline data set with which to work the discussion is likely to wander and have difficulty staying within the predefined scope of the project. Focus groups are useful as follow up activities to explore issues that arose during the analysis of information previously collected and are most helpful in clarifying areas in which there...
seemed to be several different viewpoints exhibited and/or elicited (Morgan, 1988b).

Goals for Data Set Development Focus Groups

The goal for using focus groups during the data set development process not only includes obtaining individual opinions for what data items should be included or excluded (which will happen) but more importantly to provide an environment in which a consensus for what should and should not be included in the data set. The goals for the focus group are closely related to the goals for the overall data set development project, which should be closely adhered to when developing specific goals for the focus group.

Focus Group Stages

There are five stages to conducting a focus group. These are planning, recruiting, moderating, analysis and reporting (Morgan, 1988a). The five stages and their application to data set development are discussed below.

- **Planning**
  - **General Planning**
    As with expert interviews, the key to success with focus groups is careful planning and preparation. The planning of a successful focus group includes considering all stages of the process. Explicit initial planning includes identifying issues that must be addressed throughout the project (Morgan, 1988c). These include defining the purpose and goals of the focus group, which are closely related to exactly, the purpose and goals of the data set development project. Resources for personnel and other materials are identified. Resources are usually a balance of time, talent and money. If any one of these is missing or short, more of the others will be required (Morgan, 1988a). During general planning the size, composition and structure of the group is determined. Questions and topics to be discussed by the group are identified, written and organized. The anticipated frequency of the group meeting is estimated and the dates, times and locations of the sessions are chosen and scheduled (Morgan, 1988c).

  - **Planning for Recruitment**
    Recruitment is the process of securing participants for the focus group (Greenbaum, 1993). The criteria and protocol for recruiting participants are determined during the planning stage of the focus group. A detailed plan for how the participants will be contacted, invited and reminded to come to the focus group is very helpful and should be documented (Morgan, 1988c).
The composition of the focus group is very important to the outcome of the focus group. The composition should be chosen based on the purpose and goals of the data set. If the data set is intended for multidisciplinary use then members of the group should be from all disciplines included in the scope of the data set. If the data set is very large having focus groups of each separate discipline as well as a mixed discipline group may be useful.

Eligibility and exclusion criteria are written for each focus group. These criteria are based on the purpose and goals of the data set. The criteria should reflect the qualifications and experience participants must have to contribute to the data set development process.

The recruitment strategy must be practical, economical and meet the screening criteria (Krueger, 1994). Although random sampling is the preferred method of recruitment for most marketing focus groups, Morgan (Morgan, 1988a) supports the use of purposive sampling of participants meeting the needs of a particular project if needed. For data set development, data set engineers may find purposive, or even a combination of purposive and convenience sampling the most effective method to find participants for the focus group. Purposive sampling is useful for targeting specific clinical or professional groups such as RNs of the oncology department, or all physicians of a specific clinic. Convenience sampling is the use of any volunteers from within the purposive sampled group(s) who are willing to participate in the data set development focus group. These methods are acceptable for data set development as long as the previously identified composition needs of the group are met. For example, if the data set is to be for multidisciplinary use and the criteria for the focus group states multiple disciplines are to be represented in the focus group, then the data set engineer must ensure all needed disciplines are represented. A multidisciplinary data set cannot be developed using a focus group comprised of only one discipline! Once the composition of the focus group is determined, sources for recruitment of contacts are identified.

Planning for the Meeting

Planning for the focus group meeting is very similar to the planning for an individual interview. The content of the discussion and materials or props for the discussion are identified, developed and scheduled or reserved. Additionally the role of the moderator needs to be determined. The moderator must be familiar with the topics and terminology of the data set domain (Morgan, 1988c). Luckily, in data set development it is unlikely the moderator will be someone hired

“A Data Set Development Methodology” © L. Heermann Langford 2001
from outside of the project to conduct the interview as is done in marketing and some research projects. Therefore the moderator should already be familiar with the data set and its domain.

➤ Planning for the Analysis

When planning the analysis of data from a focus group conducted for data set development, define what type of data will be generated from the discussion. Lists of data made during the discussion? Notes, tapes and transcripts of the entire discussion (Morgan, 1988c)? These decisions are again based on the goals and purposes of conducting the focus group in relation to the data set development process.

• Recruiting Participants for the Focus Group

Two methods of recruiting focus group participants are described by Krueger (Krueger, 1994). These methods are telephone or specific contacts and volunteers.

• Specific Contacts

Specific contacts are made from lists of available participants such as staff rosters or from personal recommendations. The success of recruiting with this method depends on the size of the list and the willingness of those contacted to participate (Krueger, 1994).

➤ Volunteer Participants

Recruiting focus group participants from volunteers can be very successful. Volunteers, especially those with a great interest in the project, are known to make extraordinary efforts to participate and contribute. Volunteers can also be helpful by recruiting other qualified participants. Again, it is important for the criteria for participation to be well defined for this “snowball” recruitment (Polit & Hungler, 1999b).

• Conducting (Moderating) the Focus Group Interview

Depending on the questions and topics discussed and the participants of the focus group itself the moderator may have a very active role in conducting the group through the questions or may be more passive and primarily listen. The most important factor is for the moderator to have a strong sense of what the group needs to accomplish and not allow the discussion to wander into unproductive areas.

The structure of a focus group interview is similar to that of the individual interview, consisting of an introduction, body and closing, but with a few unique considerations.
Introduction

The introduction of a focus group includes a welcome, an overview of the topic to be discussed, the ground rules of the group and an initial nonthreatening “ice breaker” type question. Ground rules include information on how participants will be recognized to speak and that only one person should speak at a time. Assurances that all will have a chance to speak should be made. The icebreaker question should be answerable in less than 20 seconds and not require deep reflection to answer. It works well to have the question answered by each participant as they introduce themselves to the group. If any recording equipment is to be used during the focus group it should be acknowledged during the introduction.

EXAMPLE: An example of an ice breaking question to use in the introductory stage of a focus group is:

“Let’s start with an introduction so that we are sure we all know each other. Please tell us your name, where you work and where you took your last vacation.”

Body of Interview

The body of a focus group interview is centered on the goals and purpose for convening the group. The moderator presents the questions or topic areas previously defined and the group is encouraged to make comments on and discuss them with each other. With data set development, since most discussion will relate to specific data elements in regard to the overall data set, it is best to provide each participant with a copy of the data set previously developed. Projecting the data set using an overhead or LCD projector or even writing it on a white board also works well to facilitate discussion.

Closing

The closing of a focus group should occur, as with the individual interview, before or at the designated ending time. This may be difficult if there is a significant amount of discussion happening as the closing time approaches. In this instance the moderator should break into the conversation and politely ask if the participants would like to continue the discussion now or at a future meeting time. Once the discussion is concluded the data set engineer/moderator should summarize the discussion of the day, review any decisions regarding data elements made and review any “action” items resulting from the
discussion. If appropriate, the plan for resolving the “action” items can also be reviewed. If this is the first of several focus groups, it may be during the closing that future meetings of this focus group are planned. Future meetings can be scheduled individually or as a standing appointment as the group and the data set engineer see fit. Finally, the data set engineer/moderator should thank every participant for his or her presence and contributions to the focus group (Krueger, 1994).

- **Data Analysis**

  Focus groups can produce large amounts of concentrated data in a short amount of time (Morgan, 1988a). The analysis of this data requires a subjective process of listening to and making sense of what is said during the focus group discussion (Morgan, 1988a). The data set engineer, who is also most likely the focus group moderator, is responsible for analyzing data collected from the focus group discussion and its impact on the development of the data set. This includes determining from the discussion what data elements should be included or excluded from the data set. It also includes understanding from the focus group discussion what the participants (potential users of the data set), feel is the best organization of the data elements. The organization of the data elements is key to the future implementation of the data set on paper and in computerized systems so that the data set is logical and intuitive to use. The composition and organization of the final data set absolutely affect overall success of the data set.

  During data set development focus groups, discussions regarding the inclusion, exclusion or addition of data elements to a data set may be quite explicit. The discussions could also be rather ambiguous. It is very important for the data set engineer/moderator to analyze the data (i.e. the discussions of the focus group) throughout the group discussion for clear indications as to the outcome of each data element. If the data set engineer is not getting a clear indication from the group as to the inclusion, exclusion, addition or organization of a data element in question then the data set engineer must probe deeper and ask the group more questions to clarify their answers. The goal of the group is to come to consensus on the data elements. If they are unable to make a consensus regarding an item the data set engineer will need to make a note of that item and return to other sources such as the literature or expert interviews to make a decision regarding the outcome of that specific element in the data set.
• Reporting

Reporting of focus group discussions in data set development is reflected in the data set itself. Data elements either retained or excluded and the organization of the data elements indicate the outcome of a data set development focus group discussion. Other forms of reporting from data set development focus groups include the documentation of discussion surrounding the reasons “why” and “why not” decisions regarding the data elements were made.

Capture of Information during Focus Group

Capturing the discussion and decisions made during a focus group discussion is just as important as during the expert interviews. Similar techniques can be used in either case. Note taking is almost always useful during focus group discussion. Note taking can be as simple as marking specific data elements with “include” or striking out data elements to exclude (with notation to the side stating why). In complex discussion, the data set engineer may find it useful to use a tape recorder or other device to completely record the entire discussion of the group. White boards may be extremely useful for discussing and recording discussions regarding the organization of data elements. Just be sure to recopy any diagrams on the white board onto paper or some other longer lasting device before erasing it!

Myths of Focus Groups

Focus groups are used extensively in marketing and qualitative research. From these arenas certain myths have surfaced, (Morgan, 1988a) not all of which are applicable to the use of focus groups for data set development. They are however worth mentioning and discussing briefly.

• Myth 1

The first myth is that focus groups are low cost and quick (Morgan, 1988a). In data set development, this may or may not be a myth. The domain, scope and chosen participants of a focus group influence the time and costs of focus groups considerably. If a data set had a well defined scope and the chosen participants are available locally, then the focus group can be low in cost and conducted efficiently. If however, the data set has a large, ill defined scope or the developers wish to include participants that are remote and will require reimbursement of travel expenses, then the focus group could in fact be very expensive and time consuming to conduct.

• Myth 2
The second myth is that focus groups require professional moderators (Morgan, 1988a). In the case of data set development, this is absolutely not true. In fact, it is best for the data set engineer or someone closely involved with the data set being developed to conduct the focus group. Since data analysis does occur during the focus group discussion as described above, it is imperative for someone very close to the project to conduct the focus group. It may be advantageous however, for the data set development moderator to receive some training as a facilitator for group discussions.

Myth 3

The third myth is that focus groups require special facilities (Morgan, 1988a). Data set development focus groups are not observed through a special “mirror” or use special recording equipment. Data set development focus groups are working meetings that function quite well in conference rooms and classrooms. The only requirements are an area to project any material as needed and well lighted writing surfaces for each participant if they should choose to use them.

Myth 4

The fourth myth of focus groups is that they must consist of strangers (Morgan, 1988a). This is not really necessary for data set development focus groups. Not all members may know each other before the group convenes, but there is no need to restrict the group due to members who do not know each other. In fact in many clinical arenas, it may be close to impossible to find enough members to participate who are strangers to each other! The only exception to this rule is the inclusion of members who do not get along personally or professionally for any reason. Bad feelings between members may inhibit contributions from either member. This however may not be foreseeable and may need to be dealt with on a case-by-case basis if difficulties arise.

Myth 5

The fifth myth is that focus groups will not work for sensitive topics (Morgan, 1988a). This myth does not usually apply to data set development. Rarely are data sets and the elements covered in them personally sensitive to individuals. Even if a topic does touch on a sensitive area for one or more participants, they should be able to distance themselves from the data element and approach the discussion objectively. If for some reason, a dataset is too sensitive for a person to objectively approach then allowing the participant to remove themselves from the group would be prudent. It is unlikely the entire group would be affected in the same way.

Myth 6
The sixth myth of focus groups is that they produce conformity (Morgan, 1988a). In the case of data set development this is one ‘myth’ that may actually be true! The goal of a focus group for data set development is to come to consensus on the status of the data elements.

- Myth 7
  The seventh myth is also one that in the case of data set development does not need to be dispelled. This myth states that focus groups must be validated by other methods (Morgan, 1988a). This is not a myth to be challenged by the use of focus groups in data set development because this methodology already proposes using more than one method of knowledge acquisition.

- Myth 8
  The last myth is one that should be closely noted with data set development. This myth is focus groups will tell how people behave (Morgan, 1988a). Focus groups will only tell how people ‘say’ they behave not their actual behaviors. This is a limiting factor to data set development due to the goal of acquiring information from the participants in what data elements should be included in the data set. The focus group will only indicate what data elements the participants ‘think’ should be included. Not the ones they always actually use.

Focus Group Beliefs to Encourage

In contrast to the myths about focus groups, Morgan (Morgan, 1988a) purports some beliefs about focus groups that should be encouraged. These beliefs are also applicable to the use of focus groups for data set development. These beliefs include the following: (1) No method of research (or knowledge acquisition) is ever perfect and should always be approached skeptically. (2) High quality moderating is crucial to focus groups (Morgan, 1988a), the moderator has a significant impact on the outcome of the focus group. The talents, preparation and attentiveness of the moderator directly affect the quality of the results. Moderation however, is only one aspect of the success. Adequate recruiting and good analysis are just as important (Morgan, 1988a). (3) Teamwork produces the best focus groups (Morgan, 1988a). A concentrated effort is required from all participants and the moderator to ensure success of the group. (4) The data set development team can always learn from the participants. The focus group is not just an “exercise” to conduct to imply user input to the data set development process. The participants are likely all experts in their own use of a data set. Listen to them! (5) There are as many ways to conduct a focus group, as there are purposes to have one (Morgan, 1988a). This is the glory of focus groups and their application to knowledge acquisition! The flexibility of focus groups and the information they provide can be invaluable to the data set development process.
Summary

The documentation of health care provided to patients and their families is critical in the ever changing health care arena of today (Meiner, 1999). Data sets are becoming a popular method for documenting patient care either on paper via checklists or in computers. It is imperative for clinical data sets to contain all data elements required for communication and decision making during the provision of care, as well as meeting regulatory and legal requirements. Including unnecessary data elements or missing vital elements are very distracting in the communication and decision making process. Therefore great care should be applied to the development of data sets for clinical use. Great efforts have ensued over the last decade on a national level to determine data sets and minimal data sets for clinical use (McCormick et al., 1997; Renner & Swart, 1997). The amount of time required however for the development of these data sets has been considerable. Frustration occurs when opportunities arise to implement a new documentation system but the data set needed for the intended environment does not exist. Knowledge acquisition techniques are effective methods to use for determining a complete and accurate data set in a shorter period of time. Literature Review, expert interview and focus groups are three very useful knowledge acquisition techniques applicable to data set development. The techniques as described in this methodology can be applied to a variety of circumstances in a variety of settings. Data sets resulting from the methodology presented are not only useful for the environments for which they are intended but are often readily accepted by the users who have often had an opportunity to participate in the development of the data set throughout stages in the data set development methodology.

“A Data Set Development Methodology” © L. Heermann Langford 2001
REFERENCES


“A Data Set Development Methodology” © L. Heermann Langford 2001


“A Data Set Development Methodology” © L. Heermann Langford 2001


"A Data Set Development Methodology" © L. Heermann Langford 2001
APPENDIX F

LETTER OF INVITATION TO PARTICIPATE AS AN EXPERT REVIEWER OF THE PROPOSED DATA SET DEVELOPMENT METHODOLOGY
Letter of Invitation to Participate
as an Expert Reviewer of the
Proposed Data Set Development Methodology

Laura K. Heermann Langford RN, MS
University of Utah, Doctoral Student
1070 E. Quail Park Dr. Apt. A
Salt Lake City, UT 84117
(801) 694-2343 (Telephone)
(801) 249-5010 (Digital Pager, available 24 hours/day, 7 days a week)
ldlheerm@ihc.com

Dear Expert Reviewer,

Greetings! I am a doctoral student at the University of Utah specializing in Nursing Informatics. For part of my dissertation I have proposed a methodology for data set development. The proposed method is an attempt to provide a consistent, efficient, cost-effective alternative for persons and or teams interested in determining a data set for their local needs. The purpose of this letter is to invite you to be a reviewer of this proposed methodology. It is imperative that the proposed methodology be reviewed by experts in the field of clinical data set development to determine if it is clear, accurate, understandable and applicable.

Many of the data sets available in the literature today have been developed with considerable costs in time and money. These sets, while very useful to the field of informatics, often are minimum data sets or are incomplete and do not fill the needs of people and or sites trying to implement coded computerized patient care documentation. A consistent, efficient, cost-effective methodology to determine what data elements to include is needed by those who are developing a computerized documentation application from scratch or just trying to determine what data elements should be included at their site when customizing and implementing a system purchased from a vendor. The techniques used in the proposed methodology are not new to the field of nursing informatics but are familiar knowledge acquisition techniques most often used in knowledge engineering.

You have been chosen as a potential reviewer for this proposed methodology because you have expertise, knowledge and/or experience in data set development for clinical use. I am inviting you to review and critique the proposed methodology for its clarity, accuracy, understandability and applicability. I realize you are a very busy person and your time is valuable. The methodology document titled “A Data Set Development Methodology” will take from 1 to 3 hours to read. The survey will take about 30 minutes to complete.

Enclosed you will find The Data Set Development Survey. The purpose of this survey is to evaluate the Data Set Development Methodology created by Laura Heermann Langford. If you are willing to participate in this study, please complete the survey and return the survey by November 16, 2001. Electronic copies may be e-mailed to LDLHEERM@ihc.com.Hardcopies should be sent by the U.S. Postal Service using the enclosed stamped envelope to:

Laura Heermann Langford
1070 E. Quail Park Dr. Apt A
Salt Lake City, UT 84117

You may omit any questions you do not want to answer. The return of the survey will constitute your consent to participate in this study. PLEASE do not write your name on the survey. All responses will be kept confidential. Any questions regarding the completion of this survey should be directed to Laura Heermann Langford at LDLHEERM@ihc.com or 801-694-2343.

Thank you for your time and I do hope you will choose to participate in this study.

Sincerely,
Laura K. Heermann Langford
APPENDIX G

DATA SET DEVELOPMENT

METHODOLOGY SURVEY
Data Set Development Methodology Survey

Respondent profile:
(This data will be separated from the questionnaire and used only to describe the study sample)

Please circle or fill in the blank as appropriate:

1. I am 19 years old or older: Yes  No

2. My highest level of education is:
   A. AD
   B. BS/BA
   C. MS/MA
   D. PhD
   E. Other __________________________
      (specify)

3. My education/occupational field includes:
   A. RN
   B. MD
   C. Other __________________________
      (specify)

4. Years of experience in using clinical data sets: _______________

5. Years of experience in developing clinical data sets _______________

6. My role in the development of data sets was as (may choose more than one):
   A. Participant
   B. Project leader
   C. Coordinator
   D. Other __________________________
      (specify)
Data Set Development Methodology Survey

Directions: Please read the document entitled “A Data Set Development Methodology”* (henceforth referenced as Document) before completing the questionnaire. It may be helpful to review the questions contained in this questionnaire prior to reading the longer document.

*The document “A Data Set Development Methodology” contains the description of a method to develop clinical data sets. Although this study is related to a doctoral dissertation, the methodology described in the document IS a data set development methodology, not the methodology of the doctoral dissertation.

Please indicate how much you agree or disagree with each item using the scale to the right of each question where 1 = Strongly Disagree and 7 = Strongly Agree

The Methodology:

1 = Strongly Disagree  7 = Strongly Agree

1. The presentation of the new Data Set Development Methodology is clear and understandable. 1 2 3 4 5 6 7

2. After reading this document, “Data Set Development Methodology,” I could begin to apply the methodology to develop a data set. 1 2 3 4 5 6 7

3. What is missing or unclear in the document “Data Set Development Methodology?”

4. The fit between the data set development methodology and the Systems Development Life Cycle is clear. (Document p.3) 1 2 3 4 5 6 7

5. The proposed process flow of the methodology is reasonable. (Document pp. 7-18) 1 2 3 4 5 6 7

6. How can the proposed process flow be improved?”
<table>
<thead>
<tr>
<th></th>
<th>1 = Strongly Disagree</th>
<th>7 = Strongly Agree</th>
</tr>
</thead>
<tbody>
<tr>
<td>7.</td>
<td>The “Data Set Development Methodology” flow diagram is logical. (Document p.8, Figure 1)</td>
<td>1 2 3 4 5 6 7</td>
</tr>
<tr>
<td>8.</td>
<td>The “Data Set Development Methodology” flow diagram follows the proposed process flow. (Document pp. 7-18)</td>
<td>1 2 3 4 5 6 7</td>
</tr>
<tr>
<td>9.</td>
<td>The “Data Set Development Methodology” diagram is clear. (Document p.8, Figure 1)</td>
<td>1 2 3 4 5 6 7</td>
</tr>
<tr>
<td>10.</td>
<td>The “Data Set Development Methodology” diagram helps describe the methodology. (Document p.8, Figure 1)</td>
<td>1 2 3 4 5 6 7</td>
</tr>
<tr>
<td>11.</td>
<td>How might the “Data Set Development Methodology” diagram be improved?</td>
<td></td>
</tr>
<tr>
<td>12.</td>
<td>The “Application of Knowledge Techniques…” flow diagram is logical. (Document p.12, Figure 2)</td>
<td>1 2 3 4 5 6 7</td>
</tr>
<tr>
<td>13.</td>
<td>The “Application of Knowledge Techniques…” flow diagram follows the proposed process flow. (Document p.12, Figure 2)</td>
<td>1 2 3 4 5 6 7</td>
</tr>
<tr>
<td>14.</td>
<td>The “Application of Knowledge Techniques…” diagram is clear. (Document p.12, Figure 2)</td>
<td>1 2 3 4 5 6 7</td>
</tr>
<tr>
<td>15.</td>
<td>The “Application of Knowledge Techniques…” diagram helps describe the methodology. (Document p.12, Figure 2)</td>
<td>1 2 3 4 5 6 7</td>
</tr>
<tr>
<td>16.</td>
<td>How might the “Application of Knowledge Techniques…” diagram be improved?</td>
<td></td>
</tr>
</tbody>
</table>
Literature Review Technique:

1 = Strongly Disagree  7 = Strongly Agree

17. The definition for Literature Review is clear and understandable. *(Document p. 19)*

18. The Literature Review description is clear and understandable. *(Document p. 19)*

19. The technique for Literature Review, as described, is applicable to data set development. *(Document pp. 19-22)*

20. How might the definition, description and/or application of the Literature Review Technique be improved?

Data Analysis Technique:

1 = Strongly Disagree  7 = Strongly Agree

21. The definition for Data Analysis is clear and understandable. *(Document p. 23)*

22. The Data Analysis description is clear and understandable. *(Document p. 23)*

23. The technique for Data Analysis, as described, is applicable to data set development. *(Document p. 23-24)*

24. How might the definition, description and/or application of the Data Analysis Technique be improved?
Expert Interview Technique:

1 = Strongly Disagree 7 = Strongly Agree

25. The definition for Expert Interview is clear and understandable. (Document p. 25) 1 2 3 4 5 6 7

26. The Expert Interview description is clear and understandable. (Document p. 25) 1 2 3 4 5 6 7

27. The technique for Expert Interview, as described, is applicable to data set development. (Document pp. 25-34) 1 2 3 4 5 6 7

28. How might the definition, description and/or application of the Expert Interview Technique be improved?

Focus Group Technique:

1 = Strongly Disagree 7 = Strongly Agree

29. The definition for Focus Group is clear and understandable. (Document p. 35) 1 2 3 4 5 6 7

30. The Focus Group description is clear and understandable. (Document p. 35) 1 2 3 4 5 6 7

31. The technique for Focus Group, as described, is applicable to data set development. (Document p. 35-43) 1 2 3 4 5 6 7

32. How might the definition, description and/or application of the Focus Group Technique be improved?
General Overview:

1 = Strongly Disagree  7 = Strongly Agree

33. The methodology is generalizable to other disciplines within healthcare.

34. The methodology is applicable to data sets used in paper systems.

35. The methodology is applicable to data sets used in computerized systems.

36. There is no redundancy in the content of the methodology.

37. The document is well organized.

38. What improvements to the document would you make?

39. How would you improve the “readability” of the document?

40. In your opinion, what kind of validation study of this methodology would be useful?

41. Is this methodology available in other formats? Yes No

If so what?

_____ Journal Article
_____ Published Manuscript
_____ Textbook
_____ Published Flow Diagram
_____ World Wide Web
42. How should this methodology be distributed?
   - Journal Article
   - Published Manuscript
   - Textbook
   - Published Flow Diagram
   - World Wide Web
   - Other (describe) __________________________

43. Please add any comments/suggestions you might have here:

   __________________________________________
   __________________________________________
   __________________________________________
   __________________________________________
   __________________________________________
   __________________________________________

Thank you for completing this survey!!

Please complete and return the questionnaire by November 16, 2001.

PLEASE do NOT write your name on the questionnaire. All responses will be kept confidential.
APPENDIX H

MATRIX RELATING SURVEY QUESTIONS
TO THE RESEARCH QUESTIONS
<table>
<thead>
<tr>
<th>Research Question</th>
<th>Survey: Item Number and Question</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Is the methodology clearly defined?</td>
<td>1. The presentation of the new Data Set Development Methodology is clear and understandable.</td>
</tr>
<tr>
<td></td>
<td>2. After reading this document, “Data Set Development Methodology,” I could begin to apply the methodology to develop a data set.</td>
</tr>
<tr>
<td></td>
<td>3. What is missing or unclear in the document “Data Set Development Methodology?”</td>
</tr>
<tr>
<td></td>
<td>4. The fit between the data set development methodology and the Systems Development Life Cycle is clear. (Document p.3)</td>
</tr>
<tr>
<td></td>
<td>6. How can the proposed process flow be improved?</td>
</tr>
<tr>
<td></td>
<td>7. The “Data Set Development Methodology” flow diagram is logical. (Document p.8, Figure 1)</td>
</tr>
<tr>
<td></td>
<td>8. The “Data Set Development Methodology” flow diagram follows the proposed process flow. (Document pp. 7-18)</td>
</tr>
<tr>
<td></td>
<td>9. The “Data Set Development Methodology” diagram is clear. (Document p.8, Figure 1)</td>
</tr>
<tr>
<td></td>
<td>10. The “Data Set Development Methodology” diagram helps describe the methodology. (Document p.8, Figure 1)</td>
</tr>
<tr>
<td></td>
<td>11. How might the “Data Set Development Methodology” diagram be improved?</td>
</tr>
<tr>
<td></td>
<td>12. The “Application of Knowledge Techniques…” flow diagram is logical. (Document p.12, Figure 2)</td>
</tr>
<tr>
<td></td>
<td>13. The “Application of Knowledge Techniques…” flow diagram follows the proposed process flow. (Document p.12, Figure 2)</td>
</tr>
<tr>
<td>Research Question</td>
<td>Survey: Item Number and Question</td>
</tr>
<tr>
<td>---------------------------------------------------------------------------------</td>
<td>---------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>1. Is the methodology clearly defined? (cont.)</td>
<td>14. The “Application of Knowledge Techniques…” diagram is clear. <em>(Document p.12, Figure 2)</em></td>
</tr>
<tr>
<td></td>
<td>15. The “Application of Knowledge Techniques…” diagram helps describe the methodology. <em>(Document p.12, Figure 2)</em></td>
</tr>
<tr>
<td></td>
<td>16. How might the “Application of Knowledge Techniques…” diagram be improved?</td>
</tr>
<tr>
<td></td>
<td>17. The <em>definition</em> for Literature Review is clear and understandable. <em>(Document p. 19)</em></td>
</tr>
<tr>
<td></td>
<td>18. The Literature Review <em>description</em> is clear and understandable. <em>(Document p. 19)</em></td>
</tr>
<tr>
<td></td>
<td>20. How might the definition, description and/or application of the Literature Review Technique be improved?</td>
</tr>
<tr>
<td></td>
<td>21. The <em>definition</em> for Data Analysis is clear and understandable. <em>(Document p. 23)</em></td>
</tr>
<tr>
<td></td>
<td>22. The Data Analysis <em>description</em> is clear and understandable. <em>(Document p. 23)</em></td>
</tr>
<tr>
<td></td>
<td>24. How might the definition, description and/or application of the Data Analysis Technique be improved?</td>
</tr>
<tr>
<td></td>
<td>25. The <em>definition</em> for Expert Interview is clear and understandable. <em>(Document p. 25)</em></td>
</tr>
<tr>
<td></td>
<td>26. The Expert Interview <em>description</em> is clear and understandable. <em>(Document p. 25)</em></td>
</tr>
<tr>
<td></td>
<td>28. How might the definition, description and/or application of the Expert Interview Technique be improved?</td>
</tr>
<tr>
<td></td>
<td>29. The <em>definition</em> for Focus Group is clear and understandable. <em>(Document p. 35)</em></td>
</tr>
<tr>
<td></td>
<td>30. The Focus Group <em>description</em> is clear and understandable. <em>(Document p. 35)</em></td>
</tr>
<tr>
<td>Research Question</td>
<td>Survey: Item Number and Question</td>
</tr>
<tr>
<td>------------------------------------------------------</td>
<td>-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>2. Is the new methodology useful for data set development?</td>
<td>36. There is <em>no</em> redundancy in the content of the methodology.</td>
</tr>
<tr>
<td></td>
<td>37. The document is well organized.</td>
</tr>
<tr>
<td></td>
<td>38. What improvements to the document would you make?</td>
</tr>
<tr>
<td></td>
<td>39. How would you improve the “readability” of the document?</td>
</tr>
<tr>
<td></td>
<td>5. The proposed process flow of the methodology is reasonable. <em>(Document pp. 7-18)</em></td>
</tr>
<tr>
<td></td>
<td>19. The <em>technique</em> for Literature Review, as described, is applicable to data set development. <em>(Document pp. 19-22)</em></td>
</tr>
<tr>
<td></td>
<td>23. The <em>technique</em> for Data Analysis, as described, is applicable to data set development. <em>(Document p. 23-24)</em></td>
</tr>
<tr>
<td></td>
<td>27. The <em>technique</em> for Expert Interview, as described, is applicable to data set development. <em>(Document pp. 25-34)</em></td>
</tr>
<tr>
<td></td>
<td>31. The <em>technique</em> for Focus Group, as described, is applicable to data set development. <em>(Document p. 35-43)</em></td>
</tr>
<tr>
<td></td>
<td>33. The methodology is generalizable to other disciplines within healthcare.</td>
</tr>
<tr>
<td></td>
<td>34. The methodology is applicable to data sets used in paper systems.</td>
</tr>
<tr>
<td></td>
<td>35. The methodology is applicable to data sets used in computerized systems.</td>
</tr>
</tbody>
</table>
APPENDIX I

EVALUATION FORM FOR “DATA SET DEVELOPMENT METHODOLOGY SURVEY”
Evaluation Form for
“Data Set Development Methodology Survey”

Name: ___________________________ Date: ______________

Phone: __________________________

Instructions: Using this evaluation form, please evaluate the attached questionnaire. Should you have comments regarding particular questions on the questionnaire, please note the question number next to your comment on the evaluation form. Please return this evaluation form and any other comments you have to LDLHEERM@ihc.com by September 21, 2001. Your assistance is greatly appreciated. Thank you!

Section 1. Content Evaluation

The objectives of this questionnaire are:

• Is the new methodology clearly defined?
• Is the presentation of the new methodology clear and understandable?
• Is the new methodology applicable to data set development?

1. Are the objectives of the evaluation tool represented?  
   Yes__ No__
   If not, what should be added?

2. Are there additional questions that should be added?  
   Yes__ No__
   If yes, what are they?

3. Are there questions that should be deleted?  
   Yes__ No__
   If yes, what are they?

4. Are the questions “emotionally loaded” or obviously biased? If yes, what would you change?  
   Yes__ No__
5. Would you feel comfortable answering all of the questions on this questionnaire? If no, please explain

Yes__ No__

Section II: Form Evaluation

6. Are the instructions for the questionnaire concise and clear? If no, please explain.

Yes__ No__

7. Do the questions, individually, ask for only one piece of information? If no, please explain.

Yes__ No__

8. Do the questions use unfamiliar abbreviations or technical terms? If yes, please explain.

Yes__ No__

9. Are any of the questions vaguely defined or overly general? If yes, please explain.

Yes__ No__

10. Are any of the questions ambiguous or do they convey double meanings which may cause misunderstandings? If yes, please explain.

Yes__ No__

11. Does the wording of any of the questions imply a desired answer? If yes, please explain.

Yes__ No__
12. Are the types of question structures used (e.g., 7 point Likert scales and open ended) appropriate for each question? Yes__ No__

13. Do the response options cover all reasonable alternatives? If no, please explain Yes__ No__

14. Are response options in the correct sequence/order for each question? If no, please explain. Yes__ No__

15. Are response options of sufficient length to answer questions effectively? If no, please explain. Yes__ No__

16. Are response options mutually exclusive for each question? If no, please explain. Yes__ No__

17. Is the sequence and grouping of questions clear and logical? If no, please explain. Yes__ No__

18. Is the questionnaire too short? If yes, what would you add? Yes__ No__

19. Is the questionnaire too long? If yes, what would you delete? Yes__ No__

Please provide any additional comments and/or suggestions have regarding the questionnaire here. Thank you!
APPENDIX J

DATA SET DEVELOPMENT

METHODOLOGY SURVEY

DRAFT
Data Set Development Methodology Survey

The purpose of this survey is to evaluate the Data Set Development Methodology created by Laura Heermann Langford. Please read the document entitled “A Data Set Development Methodology” before completing the questionnaire. It may be helpful to review the questions contained in this questionnaire prior to reading the longer document. Any questions regarding the completion of this survey should be directed to Laura Heermann Langford at LDLHEERM@ihc.com or 801-314-4329.

Please complete and return the questionnaire by Month, Date, 2001.

Please indicate how much you agree or disagree with each item using the scale to the right of each question where 1 = Strongly Agree and 7 = Strongly Disagree

The Methodology:

1 = Strongly Disagree  7 = Strongly Agree

1. After reading this document I could begin to apply the methodology to develop a data set. 1 2 3 4 5 6 7

2. If you are unable to apply the methodology to the development of a data set, what do you think is missing or unclear?

1 = Strongly Disagree  7 = Strongly Agree

3. The fit between the data set development methodology and the Systems Development Life Cycle is clear. 1 2 3 4 5 6 7

4. The proposed process flow is reasonable. 1 2 3 4 5 6 7

5. If the proposed process flow is not reasonable, how might it be improved?
1 = Strongly Disagree    7 = Strongly Agree

6. The flow diagram is logical and follows the proposed process flow.  
   1 2 3 4 5 6 7

7. The diagrams are clear.  
   1 2 3 4 5 6 7

8. The diagrams help describe the methodology.  
   1 2 3 4 5 6 7

9. If the diagrams are confusing, how might they be improved?

Literature Review Technique:

1 = Strongly Disagree    7 = Strongly Agree

10. The definition for Literature Review is clear and understandable.  
    1 2 3 4 5 6 7

11. The Literature Review description is clear and understandable.  
    1 2 3 4 5 6 7

12. The technique for Literature Review, as described, is applicable to data set development.  
    1 2 3 4 5 6 7

13. How might the definition, description and/or application of the Literature Review Technique be improved?

Data Analysis Technique:

1 = Strongly Disagree    7 = Strongly Agree

14. The definition for Data Analysis is clear and understandable.  
    1 2 3 4 5 6 7

15. The Data Analysis description is clear and understandable.  
    1 2 3 4 5 6 7

16. The technique for Data Analysis, as described, is applicable to data set development.  
    1 2 3 4 5 6 7

17. How might the definition, description and/or application of the Data Analysis Technique be improved?
**Expert Interview Technique:**

1 = Strongly Disagree  7 = Strongly Agree

18. The *definition* for Expert Interview is clear and understandable.

19. The Expert Interview *description* is clear and understandable.

20. The *technique* for Expert Interview, as described, is applicable to data set development.

21. How might the definition, description and/or application of the Expert Interview Technique be improved?

---

**Focus Group Technique:**

1 = Strongly Disagree  7 = Strongly Agree

22. The *definition* for Focus Group is clear and understandable.

23. The Focus Group *description* is clear and understandable.

24. The *technique* for Focus Group, as described, is applicable to data set development.

25. How might the definition, description and/or application of the Focus Group Technique be improved?
General Overview:

1 = Strongly Disagree  7 = Strongly Agree

26. The methodology is generalizable to other disciplines of healthcare.  
   1 2 3 4 5 6 7

27. The methodology is applicable to general data set uses (those for use on paper as well as in computerized systems)  
   1 2 3 4 5 6 7

28. The content of the methodology is *not* redundant.  
   1 2 3 4 5 6 7

29. The document is well organized.  
   ? 2 3 4 5 6 7

30. If the document is not well organized, what improvements would you make?

31. How would you improve the “readability” of the document?

32. In your opinion, what kind of validation study of this methodology would be useful?

33. Is this methodology available in other formats?

34. How should this methodology be distributed? (i.e., published book, manuscript, web site....)

35. Please add any comments/suggestions you might have here:
Basic Demographic Information:
Please circle or fill in the blank as appropriate:

36. Age: _____ years

37. Gender:  
   A. Female  
   B. Male

38. My highest level of education is:  
   F. AD  
   G. BS  
   H. MS  
   I. PhD  
   J. Other (specify)

39. Years of experience in using clinical data sets: _______________

40. Years of experience in developing data sets for clinical use _______________

41. My role in the development of data sets was as (may choose more than one):  
   E. Participant  
   F. Project leader  
   G. Coordinator  
   H. Other (specify)

Thank you for completing this survey!!

Return of this questionnaire constitutes your willingness to participate in the study. Please return the completed survey to Laura Heermann Langford at LDLHEERM@ihc.com or by the U.S. Postal Service to:

Laura Heermann Langford  
1070 E. Quail Park Dr. Apt A  
Salt Lake City, UT 84117

All postage costs via the U.S. Postal Service will be reimbursed upon receipt of the completed survey, or if you prefer, send e-mail to LDLHEERM@ihc.com asking for a stamped and addressed envelope to be sent to you.

Please complete and return the questionnaire by Month, Date, 2001.

PLEASE do NOT write your name on the questionnaire. All responses will be kept confidential.
<table>
<thead>
<tr>
<th>Appendix J Question #</th>
<th>Responses Validated Question “as is”</th>
<th>Responses Suggesting Change</th>
<th>No Answer</th>
<th>Reviewer Comment(s)</th>
<th>Changes Made/Not Made</th>
<th>Reason Behind Changes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td>0</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>2</td>
<td>1</td>
<td>Q6 should be in 2 Q’s</td>
<td>Questions regarding the 2 flow diagrams were separated out per diagram and the specific double barreled question was split into two questions.</td>
<td>Good suggestions made by reviewers</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Don’t know without reading the document</td>
<td>A question was added directly asking if the presentation was clear and understandable.</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>I would more directly query the “presentation of the new methodology is clear and understandable”</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>3</td>
<td>1</td>
<td>#7 seems no different than #6</td>
<td>Above mentioned changes on questions regarding flow diagrams made to clarify questions.</td>
<td>Good suggestions made by reviewers</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Don’t know without reading the document</td>
<td>Examples of other formats given for original Q33</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>I’m not sure what Q33 means</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Appendix J Question #</td>
<td>Responses Validated Question “as is”</td>
<td>Responses Suggesting Change</td>
<td>No Answer</td>
<td>Reviewer Comment(s)</td>
<td>Changes Made/Not Made</td>
<td>Reason Behind Changes</td>
</tr>
<tr>
<td>-----------------------</td>
<td>--------------------------------------</td>
<td>-----------------------------</td>
<td>-----------</td>
<td>----------------------</td>
<td>-----------------------</td>
<td>----------------------</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>2</td>
<td>0</td>
<td>#2. I would say “unable or would find it difficult” rather than just saying “unable” I think some of the open ended ‘?’s are slanted (e.g., 2 &amp; 5) see tool for suggestions</td>
<td>Q#2 rewritten for clarification (Q3# on final survey) Q#2 and Q#5 rewritten for clarification (Final Survey Q#3 and Q#6)</td>
<td>Good suggestions made by reviewers</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>0</td>
<td>1</td>
<td>I am not sure what you mean by “other formats” on #33</td>
<td>Examples of other formats given for original Q33</td>
<td>Consistent problem prompted change</td>
</tr>
<tr>
<td>6</td>
<td>5</td>
<td>0</td>
<td>1</td>
<td>On #9 I would just say “How might the diagrams be improved”</td>
<td>Simplification of Q’s made with above changes to all questions regarding diagrams</td>
<td>Consistent problem prompted change</td>
</tr>
<tr>
<td>Appendix J Question #</td>
<td>Responses Validated Question &quot;as is&quot;</td>
<td>Responses Suggesting Change</td>
<td>No Answer</td>
<td>Reviewer Comment(s)</td>
<td>Changes Made/Not Made</td>
<td>Reason Behind Changes</td>
</tr>
<tr>
<td>------------------------</td>
<td>-------------------------------------</td>
<td>-----------------------------</td>
<td>-----------</td>
<td>---------------------</td>
<td>-----------------------</td>
<td>----------------------</td>
</tr>
<tr>
<td>7</td>
<td>5</td>
<td>1</td>
<td>0</td>
<td>I think Q6 is double barreled and suggest splitting into 2 #6 Double barreled? (some conceivably could think it followed the process but wasn’t logical or vice versa)</td>
<td>Simplification of Q’s made with above changes to all questions regarding diagrams</td>
<td>Consistent problem prompted change</td>
</tr>
<tr>
<td>8</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>(I assume everyone reading it will be familiar with SDLC?) As long as you define and explain the SDLC in the document</td>
<td>Verification of SDLC explanation included in DSDM document done and verified by colleague</td>
<td>Good suggestions made by reviewers</td>
</tr>
<tr>
<td>9</td>
<td>4</td>
<td>2</td>
<td>0</td>
<td>#28 Need to define—redundant in what way? I’m not sure what Q33 means</td>
<td>No change made to redundancy question Q33 clarified with examples included</td>
<td>No other reviewers questioned the redundancy question.</td>
</tr>
<tr>
<td>Appendix J Question #</td>
<td>Responses Validated Question “as is”</td>
<td>Responses Suggesting Change</td>
<td>No Answer</td>
<td>Reviewer Comment(s)</td>
<td>Changes Made/Not Made</td>
<td>Reason Behind Changes</td>
</tr>
<tr>
<td>------------------------</td>
<td>----------------------------------------</td>
<td>-----------------------------</td>
<td>-----------</td>
<td>----------------------</td>
<td>------------------------</td>
<td>-----------------------</td>
</tr>
<tr>
<td>10</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>May want to define “definition” vs. “description” to make it easier to understand</td>
<td>No changes made to “definition,” “description” etc.</td>
<td>Q#6 clarified with changes made to questions regarding diagrams.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>#6 Double barreled?</td>
<td>Q#27 rewritten into Q#34 and Q#35 for clarification.</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Although the 1st time through #27 I read as “general data set USER” – not sure how to get around this</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>All are oriented the same direction though as “positive” (to the study) questions. I don’t know if this is a “bad” thing but many studies try to have at least a few Q’s oriented in the reverse direction</td>
<td>No changes made</td>
<td>Discussed with committee members and decision made to leave questions as is.</td>
</tr>
<tr>
<td>Question #</td>
<td>Responses Validated Question “as is”</td>
<td>Responses Suggesting Change</td>
<td>No Answer</td>
<td>Reviewer Comment(s)</td>
<td>Changes Made/Not Made</td>
<td>Reason Behind Changes</td>
</tr>
<tr>
<td>------------</td>
<td>--------------------------------------</td>
<td>-----------------------------</td>
<td>-----------</td>
<td>---------------------</td>
<td>-----------------------</td>
<td>----------------------</td>
</tr>
<tr>
<td>12</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>I’m not sure I like having the open ended ones mixed in with the Likert scale questions, but not sure how you’d put them all at the end either. I guess they’re OK</td>
<td>No changes made.</td>
<td>The questions were grouped according to content being evaluated and kept the survey more organized.</td>
</tr>
<tr>
<td>13</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>Not sure without reading document</td>
<td>No changes made</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>5</td>
<td>1</td>
<td>0</td>
<td>I liked that you had the questions in consistent order. Instructions in heading are reverse of the heading for the Likert scale.</td>
<td>No change in consistent ordering of questions. Instructions corrected</td>
<td>Necessary changes needed due to oversight.</td>
</tr>
<tr>
<td>15</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Appendix J Question #</td>
<td>Responses Validated Question “as is”</td>
<td>Responses Suggesting Change</td>
<td>No Answer</td>
<td>Reviewer Comment(s)</td>
<td>Changes Made/Not Made</td>
<td>Reason Behind Changes</td>
</tr>
<tr>
<td>-----------------------</td>
<td>-------------------------------------</td>
<td>-----------------------------</td>
<td>-----------</td>
<td>----------------------</td>
<td>-----------------------</td>
<td>----------------------</td>
</tr>
<tr>
<td>18</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>5</td>
<td>1</td>
<td>0</td>
<td>&quot;I think it is a little long but I don’t know what I’d cut. And I don’t know how long the document is that they’re reading&quot;</td>
<td>No changes made</td>
<td>All questions necessary to adequately evaluate the long DSDM document.</td>
</tr>
<tr>
<td>Appendix J</td>
<td>Responses Validated Question “as is”</td>
<td>Responses Suggesting Change</td>
<td>No Answer</td>
<td>Reviewer Comment(s)</td>
<td>Changes Made/Not Made</td>
<td>Reason Behind Changes</td>
</tr>
<tr>
<td>------------</td>
<td>-------------------------------------</td>
<td>-----------------------------</td>
<td>-----------</td>
<td>---------------------</td>
<td>-----------------------</td>
<td>----------------------</td>
</tr>
<tr>
<td>20</td>
<td>On highest level of education, I’d say BS/BA &amp; MS/MA (maybe just because I have an MA &amp; it’s not a choice)</td>
<td>On #41, change “then” to “than”</td>
<td>Overall, very nicely done. I think it is readable and well formatted</td>
<td>It’s a little difficult to evaluate the survey without reading the document. For example, I think the flow diagram questions are getting at whether the methodology is clearly defined, but I don’t know what the diagrams are depicting.</td>
<td>BA and MA added to education choices. Necessary grammatical changes made</td>
<td>Good suggestions made by reviewers</td>
</tr>
<tr>
<td>Appendix J Question #</td>
<td>Responses Validated Question “as is”</td>
<td>Responses Suggesting Change</td>
<td>No Answer</td>
<td>Reviewer Comment(s)</td>
<td>Changes Made/Not Made</td>
<td>Reason Behind Changes</td>
</tr>
<tr>
<td>-----------------------</td>
<td>--------------------------------------</td>
<td>-----------------------------</td>
<td>-----------</td>
<td>---------------------</td>
<td>-----------------------</td>
<td>-----------------------</td>
</tr>
<tr>
<td>General Layout</td>
<td></td>
<td></td>
<td></td>
<td>Respondent profile moved to front of Survey</td>
<td>Based on discussion with committee member</td>
<td></td>
</tr>
</tbody>
</table>
Each Question from the DSDM Survey allowing narrative response is listed below with its accompanying responses.

The Methodology:

3. What is missing or unclear in the document “Data Set Development Methodology?”
   - From an applied sense, a clearer "prescription." Frequently too general to be useful for the practitioner.
   - Mapping a local data set onto a canonical data set (or a group of sets that comprise a canonical set).
   - Include examples of forms you describe whenever possible. Such as example of "knowledge acquisition form" as described on page 32. The one on page 18 is very helpful. This is for the benefit of "visual learners."
   - I don't know if data sets are ever complete! As data sets emerge and used, the new discoveries are found. With new discoveries data sets need revised.
   - Perhaps how to evaluate/test actual use of data set. In the case of clinical data sets (example: Long Term Care MDS). How is the data set used and by whom?
   - A definition of a data set. Descriptive end product
   - A literature search may not reveal pertinent data sets/data dictionaries, other vocabularies. For example, IHC has an extensive data dictionary, so does Siemens. How do people find this information. Also HL7 is registering vocabs (info source)
   - Further delineation of the steps in development

6. How can the proposed process flow be improved?"
   - My comment is not on the proposed process per se but rather to question re: fit of your methodology with SDLC. It seems the methodology is also beneficial during analysis phase not just design when considering activities listed in first stage of identification. Seems this is clearly the analysis phase. Otherwise, the proposed process flow seems appropriate.
   - I wonder if additional expertise on the team. For example, individual with expertise in data set work or informatics. In prior work, individuals who have worked with online documentation/development software have enriched vocabulary efforts
11. **How might the “Data Set Development Methodology” diagram be improved?**

- I found it to be very clear.
- At this point, I don't know what QA or Freeze means. Need more clarification to the figure. (It happens later, but too much later to be useful initially)
- Minor formatting issues. Briefly describe what occurs in the "QA" process boxes in the diagram.
- Am I correct in assuming that the criteria being used for QA is that written in the boxes for each of the processes within the phases? That is a little unclear. For example, on page 6 requirements are listed which must be met before knowledge acquisition techniques can be used--yet they are not completely consistent with 1-7 of Identification box.
- Show a true chronology between the phases --so that each phase is followed left to right by the subsequent phase
- Same-either be more general or more detailed you have both in your scheme
- I don't think of a data set as ever complete. There most likely will be revisions. I would put a key on the chart to explain abbreviations.
- Either be more general or more detailed you have both in your chart

16. **How might the “Application of Knowledge Techniques…” diagram be improved?**

- Seems like a need LR? is needed initially before Lit Review
- Minor formatting issues -- line up "Need *" nodes with associated actions so the "yes" lines do not need to bend.
- Question I have, is how do you know whether the data set is complete? As I read the description, I am led to believe that the expert interview and focus groups particularly are used to determine if the data set is complete? Or did I misunderstand your intent? It appears from the diagram, for example that only literature review might be used if I decide at that point the data set is complete.
- Great diagram, simple, clean
- Must correlate what literature impacts a data set. Generally literature does not provide data dictionaries which are needed for your system design
- See earlier comment about abbreviations
Literature Review Technique:

20. **How might the definition, description and/or application of the Literature Review Technique be improved?**
   - Could be written more like a handbook--recipe style. I'm not sure that a user will sit still to read definitions.
   - I would highly recommend that one document in some way the search strategy that was used for future reference. Whether it be to clarify the terms you used, which were exploded, etc. Also would be helpful in future should one wish to reexamine the literature or update the data set. I would also recommend that the quality or strength of evidence be documented in some cases for determining at some points which terms might be best utilized in the data set. In health care today, this seems important as we strive for evidence based practice.
   - No suggestions--great hint section. Helpful info for a beginner or expert
   - You omit fugitive literature. You do not describe types of literature--each provides different types of info--books, from articles, from proceedings etc.
   - I would include more on MESH/search terms. Also a lit review will not yield all data sets. This approach cannot be the only approach to identify pertinent data sets. This section needs work to help others find pertinent data sets

Data Analysis Technique:

24. **How might the definition, description and/or application of the Data Analysis Technique be improved?**
   - We don't really know who the data set engineer is or what are his/her skills. I'd number the steps and be very direct, e.g., "use the concepts identified in the conceptualization phase as key words when searching...."
   - The example was very helpful in clarifying what is meant by this technique.
   - No suggestions
   - Not clear-need to be more specific. One analyzes data. I think you mean "data needs" for system
   - This section is not clear to me (it could be me). One problem in looking at "past data collection" is it might not have been the right data to collect. I simply am unsure about the purpose of this step in the process.
Expert Interview Technique:

28. How might the definition, description and/or application of the Expert Interview Technique be improved?

- Goal - last statement makes it seem the relationship is primary. Avoid passive tense ("are recommended," "should be treated,"). Instead, "Treat the expert with..." How experts partition data and knowledge is crucial, but I don't have a clue from this how to determine this. Would help to include examples of knowledge acquisition forms.

- Under guidelines, you might include in #12 the possibility of tape recording the interview (and obtaining permission of course as you mention later p.24.) Also note that on page 26 you refer to HCFA--that agency has been renamed so you might just want to update that.

- No suggestions--quite clean, thoughtful

- Step is clear but not sure how experts can provide assistance in the design--Just test it is better use of time.

- I would give some examples of specific and open ended questions to ask the interviewee

Focus Group Technique:

32. How might the definition, description and/or application of the Focus Group Technique be improved?

- Too wordy; be more direct and user friendly.

- I would suggest using a scribe if possible for the focus groups. It can be taxing to say the least to be both the moderator and the scribe. I would attempt to have a scribe who is also quite familiar with the domain in an attempt to assure accuracy of note taking and this person can also be helpful in summarizing agreements of the session. Just takes a little pressure off the moderator who then can focus more on managing the group process.

- Application-Issues to be excluded during discussion. Or another way to keep focus group focused.

- No suggestions
• I'm surprised there are not more recent articles about this, especially since "usability" is getting to be a hot topic.
• Step is clear but not sure this helps
• Again give sample/generic questions specific to data sets

General Overview:

38. What improvements to the document would you make?
   • Shorten; use examples, bullets, tables
   • Only use 1 font family check capitalization rules for "van Bemmel" at start of sentences.
   • As mentioned previously, sample forms you describe would be useful. Overall, very nicely done!
   • Give more clinical examples---procedures
   • I would have like to have seen the author's ideas "tested" ---at least retrospectively with a variety of vendors and others who have developed successful data sets. I think that "user groups" were left out as another mechanism.
   • One of the greatest weaknesses relates to finding already existing vocabularies. That needs to be better described

39. How would you improve the "readability" of the document?
   • Same as above; definitions may not be needed--or perhaps move to a glossary
   • Try to vary introduction of cited material (e.g. Don't always start with "x and y say...") In Myths of Focus Groups change the heading from "Myth x" to "Myth x: <description>" and then start out the paragraph discussing the myth. This will reduce the redundancy of the heading "Myth x" and the first sentence "The x myth is that...."
   • Actually, I find it quite easy to read. The headings, bulleted lists and the like are very helpful. Your table of contents also really helps orient the reader to your approach.
   • With in each section, give the reader a "taste" of the entire process you will be discussing, then follow with the subject matter.
   • reads well
• Simplify it - use examples. Use one example which you track all throughout the entire document (EMR)

• I think a simplified overview of the entire process like an executive summary would be helpful

40. **In your opinion, what kind of validation study of this methodology would be useful?**

• Compare datasets completed for the same content by groups trained with the methodology and those who were not.

• Try in practice; compare with other methods.

• Development of a data set for a focused, but complex domain.

• Use methodology as a cook book approach in developing a real data set while Simultaneously evaluate the methodology. I know this would be time and resource intensive, but would validate what you're trying to prove.

• Perhaps a unit specific need

• Two group data set experts using this methodology vs. previous methodology

• Minimally-interviews with those who have done successful data sets

• test the method

• I think working on specialty vocabs/data sets would be helpful

43. **Please add any comments/suggestions you might have here:**

• The Database comment on page 11, paragraph 3 is quite valid. We frequently query our free-text assessment findings and user-created numeric measurement fields for elements to be added to the data set. Great methodology document, I learned a lot!

• It's an interesting project and a good idea; before it becomes a really usable method, I think it will need some really good examples, tools and clearer presentation. Right now (and probably rightly so) it reads more like a thesis. Probably the handbook itself needs to be included within that thesis, but to focus on the user. Some of the background could then either be moved to an appendix or as an intro. Good luck with the project!

• Include a brief overview (diagram would be good) of the SDLC framework for readers who may not be familiar with SDLC. Might want to discuss or recommend methods for picking a source for data elements when multiple sources may have similar concept definitions. This might
also lead to a discussion of making sure that the new data set can be mapped back out to a canonical data set.

- Shows much effort on your part. Good luck in your future endeavors.

- Laura, You've done an excellent job! I don't know if you are validating our methodology used in developing AADE's National Diabetes Education Outcomes System (NDEOS) or if I'm validating your methodology research. This is congruent to our methodology used for the NDEOS, including the Focus Group technique. You might want to look at some of our published articles available on line at aadenet.org web site. Thanks for the opportunity to be a reviewer! May I share this with my project team members when you are done? Thanks,

- I think the document is pragmatic and useful; a good collection of best practices and advice for building any kind of data set. The document might be a little "soft" for a PhD thesis, but only if one evaluates its merits from a traditional academic perspective.

- Clarify the difference between a data set, a database and a system data elements

- Overall nice summary of an appropriate methods
REFERENCES


information systems: Successful design and implementation (pp. 1-30). New York: Springer-Verlag.


