Finding that there was little advice in the quality improvement literature on specific methods for conducting chart review, the authors set about formulating their own recommendations. They describe how to integrate data collection with project design, choose the appropriate data collection format, design the chart abstraction tool, precisely define each variable, best use chart review personnel, and monitor data quality.

TOOLS, METHODS, AND STRATEGIES

The Art and Science of Chart Review

JEROAN J. ALLISON, MD, MS
TERRY C. WALL, MD, MPH
CLAIRE M. SPEETTLE, PHD
JAIMEE CALHOUN, M AEd
CRAYTON A. FARGASON, JR, MD
RICHARD W. KOBYLINSKI
ROBERT FARMER
CATARINA KIEFE, PHD, MD

In 1997, in the course of a newly funded study, “Using Achievable Benchmarks of Care to Improve Quality of Care for Outpatients with Depression,” we turned to the quality improvement (QI) literature for specific methodologic assistance with chart review and found little.1–4 Chart review was an integral part of this ongoing national cooperative project conducted by a large managed care organization (MCO; U.S. Quality Algorithms, Inc [USQA; Blue Bell, Penn]) and an academic medical center (University of Alabama at Birmingham [UAB] School of Medicine). Our multidisciplinary team consisted of researchers from the MCO and UAB, including a doctoral-level psychologist; a computer programmer; a research assistant; five physicians, including a board-certified psychiatrist; and three chart abstractors.

The overall aim of the project was to evaluate a QI plan for Aetna U.S. HealthCare® beneficiaries with depression. The project consists of two phases:

Jeroan J. Allison, MD, MS, is Assistant Professor, Department of Medicine, Division of General Internal Medicine, and Center for Outcomes Effectiveness Research and Education (COERE), University of Alabama at Birmingham, and University Hospital, Birmingham, Alabama. Terry C. Wall, MD, MPH, is Assistant Professor, Department of Pediatrics, University of Alabama at Birmingham, and The Children’s Hospital, Birmingham. Claire M. Spettell, PhD, is Research Director, U.S. Quality Algorithms® (an affiliate of Aetna U.S. HealthCare®), Blue Bell, Pennsylvania. Jaimee Calhoun, MAEd, is Program Coordinator, COERE, University of Alabama at Birmingham. Crayton A. Fargason, Jr, MD, is Associate Professor, Department of Pediatrics, University of Alabama at Birmingham, and The Children’s Hospital. Richard W. Kobylinski is Senior Healthcare Analyst, U.S. Quality Algorithms®. Robert Farmer is Data Engineer, Alabama Quality Assurance Foundation, Birmingham.
1. Using medical record review to validate the USQA algorithm for identifying depressed patients based on claims data. This ascertainment of depression based on chart review will be applied symmetrically across two separate populations of algorithm-positive and algorithm-negative patients.

2. Ascertaining the effect of physician performance feedback on quality of care.

In the first phase of the project, we determined the operating characteristics (sensitivity, specificity, predictive value) of a previously developed algorithm designed to identify patients with depression from encounter, claims, and pharmacy data. Because the gold standard for the diagnosis of depression (the structured psychiatric interview based on criteria from the Diagnostic and Statistical Manual of Mental Disorders, fourth revision [DSM-IV]) is resource intensive and involves direct patient contact, we used surrogate measures from chart review.

We have now begun the second project phase, which contains no chart review, and after the subsequent brief summary, it is not discussed further in this article. In this phase, offices of 400 primary care physicians (family practitioners, pediatricians, general internists, and general practitioners) will be randomly contacted.

Catarina Kiefe, PhD, MD, is Professor, Department of Medicine, Division of Preventive Medicine and COERE, University of Alabama at Birmingham, and University Hospital.

The work reported in this article was supported by a grant from the Academic Medicine and Managed Care Forum (to which Aetna U.S. HealthCare is a contributor) and by grant no. S09446 from the Agency for Health Care Policy and Research. Please address requests for reprints to Jeroan J. Allison, MD, MS, The University of Alabama at Birmingham, 621 Medical Education Building, 1813 6th Avenue South, Birmingham, AL 35294; phone 205/934-3007; fax 205/975-7797; e-mail jallison@gim.dom.uab.edu.
assigned to experimental and control groups of 200 each. Physician offices in the experimental group will receive baseline feedback on their management of depression relative to several quality of care indicators derived from clinical practice guidelines issued by the Agency for Health Care Policy and Research and the American Psychiatric Association.7,8 (For our purposes, quality indicators capture appropriate aspects of medical care to ascertain the frequency with which a given intervention has been appropriately used or withheld.9) In addition to receiving the performance of their individual group practice and the mean performance of all physicians in the study, physicians from experimental offices will also receive an Achievable Benchmark of Care, a data-driven measure of excellence in peer performance.10 Physicians in the control group will receive no feedback. After one year, performance will again be measured and compared with baseline performance.

Although the national cooperative project and this resulting article focus on using chart review to provide the data necessary to derive quality of care indicators, we recognize that there are many other purposes and types of chart review.5,9,11–17 We concentrate on explicit chart review, in which trained personnel ascertain (abstract) specified data elements from a chart according to a standardized template (data collection tool) that may be in either electronic or paper format. In contrast, in implicit chart review, the abstractor evaluates the quality of care received by the patients based on “clinical judgment.” That is, the abstractor provides an analytic function at the time of review. Implicit chart review, formerly used by peer review organizations (PROs; now called Quality Improvement Organizations), produces unacceptably high disagreement rates.18,19

In the remainder of the article, we review those concepts learned from the literature, as well as those learned from our experience with this project. We explain by example the process of chart review from beginning to end with frequent illustrations based on our real-world experience. We demonstrate that these concepts are theoretically straightforward but operationally difficult. Because the specific application of these concepts is intimately linked to each specific project, we do not lay out a “cookbook” approach to chart review, nor do we provide a complete chart review manual. Likewise, this article is not designed to teach the process of chart review de novo but to assist those currently engaged in the process.

**Literature Review**

Chart review provides a convenient source of data not available elsewhere, but the complexities involved in obtaining high-quality data are sometimes overlooked. Some investigators consider chart review to be a complete and reliable source of clinical information, neglecting the fact that most chart review data are usually two steps removed from the patient. (Clinicians examine the patient and record this information in the chart; often there is an intermediate step of transcription, and this is followed by chart abstraction.) Banks elucidates the difficulty of chart review and provides an overview of the design of data collection forms.20 However, other articles that provide step-by-step instructions on the development, implementation, and evaluation of QI projects do not describe the underlying complexity inherent in chart review.21–24

In our review of the QI literature, we seldom found significant methodologic detail about chart abstraction. Instead, we found such statements as, “The team developed a new data collection tool,”25 “Charts of all patients who arrive at the emergency department with cardiac-related symptoms are reviewed on an ongoing basis to evaluate the timeliness and appropriateness of assessment and treatment,”26 and “Residents were trained to perform the chart reviews and faculty physicians checked their work.”27 A notable exception lies in Marciniak’s description of the Cooperative Cardiovascular Project.28 Shortell surveyed the QI literature and noted many opportunities for improvement, including the need to obtain high-quality data.29 In their review of the emergency medicine literature, Gilbert and colleagues30 noted that a large number of studies examining emergency medicine care relied on chart review but provided little methodologic detail. One explanation for this lack of description of exactly how charts were reviewed in QI projects lies in limited journal space and editorial restraint.

Garnick et al provided a detailed description of a computerized chart review system for use with ambulatory medical records31 although they did not discuss at length the process by which the system was developed. This system was the centerpiece of
DEMPAQ (Development and Evaluation of Methods to Promote Ambulatory Care Quality; sponsored by the Health Care Financing Administration [HCFA]), which is no longer in widespread use. Yet several advances in chart review systems, many originating with the DEMP AQ developers themselves, have since been made. For example, researchers have learned not to abstract indicators but instead to abstract the individual components of an indicator. When data collection is linked to a specific indicator, as with DEMP AQ, each section of the chart must be referenced more than once (as opposed to collecting all necessary data from one section at one time). In addition, DEMP AQ requires subjective judgments by the abstractor at the computer, which, as is subsequently discussed, may introduce significant bias. An advancement in computer technology allows data to be collected in a window-based format, avoiding a series of pull-down menus. A standard roster of diagnostic codes, medications, and variable descriptions can thereby be incorporated into the program itself. (Additional details are provided in the discussion of the MedQuest chart abstraction software, pp 119-120.)

Two clinical data abstraction centers (CDACs) collect chart review data under contract from HCFA and work closely with PROs across the United States. Together, the CDACs and PROs have accumulated substantial experience and expertise in chart review, but have published no summary of their work in the peer reviewed literature. (Many of the insights presented in this article stemmed from our work with Alabama Quality Assurance Foundation and DynKe-PRO [York, Penn], the central zone CDAC.)

**Methods**

**Integration of Data Collection and Project Design**

For our project, the design of the study (and the subsequent definition of the precise data elements needed from chart abstraction) dictated the design of the chart review tool. Lack of availability of certain items from chart abstraction may necessitate important changes in study protocol. For example, chart abstraction does not always capture all mammograms received by female Medicare beneficiaries, who may frequently receive services from multiple providers, diminishing the probability that any one chart contains complete information. In addition, we did not relegate the design of the chart review tool to “ancillary personnel,” nor did we assign this task to any one individual working in isolation. We recognized that development of an optimal abstraction tool requires a diverse team. Clinical expertise was needed if the questions were to be constructed for capture of meaningful information. Because we chose an electronic format, a programmer provided valuable assistance with software development and remained in contact with other members of the team throughout the entire process of tool development and refinement. The personnel actually performing the chart review often suggested revisions to the tool which helped them access information quicker and more easily.

**Format: Electronic Versus Paper**

When selecting a data collection format, there are several important issues to consider: the complexity of the review; the cost and importance of coupling data collection and data entry; data security; and where the review will be conducted. Data collection may occur on site or centrally. The broader the geographic area covered, in general, the more costly on-site abstractions become. With central data collection, records must be photocopied and then forwarded, which may create unexpected problems. Relevant clinical details may be lacking, either because they were not photocopied or because of copy quality. When requesting photocopies of clinical records, it is therefore important to be specific about which chart elements should be copied.

Data may be abstracted from the chart onto a paper abstraction tool or directly into a computerized abstraction system. A paper abstraction system makes the instrument easier to copy and use, meaning a lower initial setup cost. However, there may be a higher cost per variable relative to computer-based abstractions.

With complex projects, significant additional personnel costs are associated with duplicate data recording, as are additional opportunities for errors. Even with a paper system, the usual practice is for data to be ultimately entered into a computer. Cost-benefit ratios should be compared, but, in general, as the complexity and length of the survey tool increases, so do the benefits of using a computerized abstracting tool. (Although no precise definition exists to differentiate between simple and complex abstraction projects, we consider a project to be complex if the abstraction...
module contains more than 20 to 30 variables or there are more than 50 charts to abstract.)

Especially when the data are extensive and complex, a properly designed abstraction software package greatly facilitates the maintenance of data quality, the performance of routine analyses, and the preparation of reports. For example, the software package we chose, MedQuest (Fu and Associates, Arlington, Va), was developed for HCFA and is publicly available on the Internet (www.hcfa.gov). MedQuest, which consists of a suite of tools designed for customized data collection, is an interpreted data entry system that stores abstraction results in a single Access database. This format allows chart abstraction data to be easily exported to many common data analysis tools or programming languages. The MedQuest software package meets all the specifications in Table 1 (right). Other benefits are as follows:

- The extensive resources expended in MedQuest's development and refinement saved us unnecessary duplication of programmer time.
- MedQuest also provides flexibility in variable presentation, which, as is discussed later in this article, plays an important role in abstraction.
- MedQuest allows sophisticated error checks while entering data. For example, MedQuest can prevent out-of-range entries, perform logical consistency checks (for example, flagging prostate cancer in a woman), and remind the abstractor when critical questions are unanswered upon exiting the module.
- MedQuest contains a standardized medication database from Multum Medisource Lexicon and the International Classification of Diseases, Ninth Revision, Clinical Modification database. The Multum Lexicon is updated weekly and is available from the Multum Internet site (www.multum.com).
- Other standardized systems of nomenclature may be easily added. Many MedQuest chart review modules developed for specific Medicare QI projects are available on the HCFA MedQuest site. Use of these modules can save substantial time in the development of a chart review tool.

Our choice of an electronic format for data collection reflected our plans to abstract more than 400 charts and to include at least 100 variables. We also planned to export the chart review data to a mainframe computer for linkage with administrative data.

<table>
<thead>
<tr>
<th>Table 1. Requirements of an Ideal Data Abstraction Software Package</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Modular organization, based on individual diseases or projects</td>
</tr>
<tr>
<td>- Presentation of related questions in window-based screens, with tabs for navigation</td>
</tr>
<tr>
<td>- Design of custom questions using decision rules (eg, error checks, parent-child relationships)</td>
</tr>
<tr>
<td>- Notification of uncompleted questions upon exit from the module</td>
</tr>
<tr>
<td>- Inclusion of built-in synonyms, medication databases, and diagnostic codes</td>
</tr>
<tr>
<td>- Real-time monitoring of data quality</td>
</tr>
<tr>
<td>- Derivation of adherence rates for quality indicators</td>
</tr>
<tr>
<td>- Exportation and importation of data in standardized formats</td>
</tr>
<tr>
<td>- Generation of printed versions of the tool, including data entry forms and variable dictionaries</td>
</tr>
<tr>
<td>- Ability to function as a stand-alone or network unit</td>
</tr>
</tbody>
</table>

Because charts would be obtained from a large geographic area, we chose to have physicians' offices provide photocopied charts with specified components.

Abstraction Tool Design

After we had designated MedQuest as the software package, we began development of the abstraction tool. (A sample screen is shown in Figure 1, p 120.) Without the aid of actual charts, progress was slow. However, once we had a preliminary version of the module, we moved rapidly to testing with actual patient charts. We abstracted the same three to five charts that were chosen by convenience but thought to be representative of the charts that would be encountered in the actual project. In the prepilot phase, we used basic observations of interrater agreement to improve the tool. (A formal pilot phase was conducted as a “final” test after the tool was fully developed but before actual chart review.)

The prepilot phase revealed confusing and misleading questions in the abstraction tool. We compared our individual responses, determined where discrepancies lay, and discussed our reasons for choosing specific options. Each revision after prepilot testing made the content of the tool more succinct and clear. For example, we discovered inconsistencies within the tool in the definitions of “yes,” “no,” and “not mentioned,” which led us to.
reconsider the collection of negative information. We also found differential interpretation of some questions which reflected the level of patient complexity; for example, one chart documented several changes in the patient’s marital status. The addition of “other” as an option offered an acceptable solution to this problem. However, the definition of “other” depends on the role of the variable in the research questions and analytic plans of the specific project requiring the chart abstraction.

Perhaps the greatest problem evident in prepilot testing was abstractor assumption and interpretation. For example, we found that documentation of “depressive symptoms” in some charts was marked as positive for “depressed mood.” Theoretically, depressive symptoms could refer to any of the DSM-IV criteria, not specifically depressed mood, as had been assumed. We concluded that the term “depressive symptoms” alone does not imply depressed mood. Therefore, this issue was addressed both in the help screen for “depressed mood” and in the training manual introduction to the project.

The prepilot also highlighted the need for a clear distinction between symptoms of depression or anxiety and physician diagnosis of major depression or anxiety disorder. In addition, every DSM-IV symptom required extensive documentation of synonyms. For example, cognitive dulling could be documented in a medical chart as impaired ability to focus, indecisiveness, decreased memory or retention, or impaired thinking. Anhedonia could be noted as loss of interest, loss of pleasure, feeling “blah,” or not caring about anything. Listing synonyms for variable options made it less likely that the abstractors would make assumptions about the meaning of a particular variable, thereby increasing interrater reliability.

Figure 1. This sample screen from MedQuest shows the user-friendly, window-based format for data entry.
Design of Chart Abstraction Variables

In designing the chart review module, we considered four types of abstraction variables, pertaining to classification, format, definition, and presentation. In general, chart abstraction variables are classified as patient identifier, demographic, or study-specific variables. The format of a variable may be free text, numeric, or categoric. The complete definition of a variable includes source designation (location in chart where information is obtained), synonyms, time frame, dependence on other variables ("parent"--"child" relationships, that is, in which variables are based on a branching logic format), precision of numeric variables, and specifications for negative data capture.

When possible, we followed the specifications of The National Committee on Vital and Health Statistics (NCVHS), which is working on the identification, definition, and implementation of standardized data in the health care and health care information fields. To date, this committee has defined several key core health data elements, but these variables constitute only a small fraction of those usually needed in a chart review project. The variable list is available on the Internet at www.ncvhs.gov.

Before proceeding with the creation of new variables for abstraction, we considered the use of administrative data. In our case, use of preexisting electronic databases minimized the complexity of the abstraction process and avoided costly duplication. Patient identifiers, consisting of lengthy text or numeric strings, provoked abstraction errors and increased abstraction time. Therefore, we preloaded selected fields that were not expected to change (for example, patient name, identification codes, birthdate) from administrative data into the chart abstraction database. The chart abstraction tool allowed the abstracter to type the first several digits of a patient numeric identifier and then select a preloaded patient record.

The decision to integrate administrative data into the chart abstraction module must be based on the quality and type of the administrative data. In general, administrative data lack the clinical detail found in charts, and certain items, such as diagnostic codes assigned for billing purposes, sometimes mislead the clinical investigator. However, MCOs often collect rich administrative databases that may be linked to chart review data. A full discussion of the risks and benefits of administrative versus chart review data is beyond the scope of this article and necessitates several steps, each with hidden subtleties.

Classification of Chart Abstraction Variables

Patient identifiers. Patient identifiers may include names, Social Security numbers, medical record numbers, and unique identifiers. If identifiers are not imported from administrative data, certain principles guide their abstraction. If the patient’s name is to be used as an identifier, then the exact format (that is, first name then last or last then first, whether to use middle initial or entire middle name) must be specified. Whether to use commas and periods may seem to be a trivial issue, but variation in name entry may have significant repercussions when the need to merge databases arises. More than one identifier is often necessary for adequate matching with other data sources.

For privacy concerns, encryption of the patient identifier in the abstracted data should be considered. In cooperative research efforts, it is generally necessary to encrypt patient identifiers when passing data from one research phase to another. Patient identity can often be imputed from incomplete identifiers; therefore, confidentiality issues must be given adequate consideration. Additional concerns about security and confidentiality are addressed in the “Chart Procurement” section later in this article.

Demographics. The precise specification of age, sex, race, occupation, and marital status varies, depending on the purpose of the project. In addition, collection of each of these variables contains subtleties. Although demographic variables cannot be defined without reference to the specific project, some general principles warrant discussion. For example, one might assume that all patients would be either male or female. However, the chart may not state the sex, and the name of the patient may be neutral. Therefore, for sex designation, one might use four categories: male, female, transgendered, and unknown/not stated/other. Race and ethnicity are far more complicated. However, these designations are often based on a description from the health care provider or are missing from the chart. Some MCOs intentionally exclude race information from their administrative databases.
Marital status and living arrangements are often intertwined. For example, even though one might assume that married persons would be living together unless stated otherwise, this may not be the case. Descriptions of marital status and/or living arrangements vary widely, from not being mentioned at all to being described in detail or described in vague terms, for example, "estranged." In addition, marital status and living arrangements often change over the time frame of interest. A comprehensive list such as that used for the United States Census may provide a template for the collection of occupational data.

**Study-specific variables.** Study-specific variables may measure a wide array of content. Examples include symptoms, diagnoses, physical findings, objective data (weight or vital signs), diagnostic tests, medications, surgical procedures, preventive counseling, hospitalization, referrals, emergency room visits, phone calls, severity of illness, complications of care, other outcomes, and utilization. For a given symptom, additional variables may describe the nature, severity, time course, associated symptoms, ameliorating factors, and exacerbating factors.

**Format of Chart Abstraction Variables**

**Free-text and numeric variables.** Free-text variables, such as verbal descriptions of patient symptoms, cannot be quantitatively analyzed with ease; therefore, their use should be limited for most QI projects. If certain verbal descriptions appear with regularity, these descriptions may be incorporated as variable options in future revisions of the tool. In content analysis—a quasiquantitative approach to analyzing text—verbal descriptions are more appropriate.

For numeric variables, several techniques, such as creating the exact number of slots in each response field for the number of digits to be collected and using leading zeros when a one-digit number is entered into a two-digit space, reduce data entry errors. Proper visual alignment means that all numbers are recorded in parallel format, with an exact correspondence of decimal places. Built-in error checks can be used to prevent the abstractor from recording a value that is outside a reasonable range.

The precision of data must be determined by the referent question. Some situations may require that time be recorded in minutes or seconds, whereas for other situations days may be sufficient. For example, timing of electrocardiogram after admission to the ER with chest pain may be measured in minutes, and duration of time for which a patient was held NPO (nothing by mouth) may be recorded to the nearest hour. Use of decimal places for numeric values also needs to be considered during the design phase.

**Categorical variables.** Categorical variables present the abstractor with a pick list containing one or more possible responses (Figure 1). When constructing a pick list, the options may be mutually exclusive, or it may be possible to choose multiple options. Investigators may wish to convert numeric data in the chart to categorical variables in the abstraction tool. For example, the abstractor may count the number of blood transfusions for a given hospitalization. Duration of depression might be categorized into "≤ 2 weeks," "> 2 weeks," and "unable to determine." Laboratory results and vital signs are often best recorded with such categorical ranges. Often the conversion rule may demand collection of only the first, last, best, worst, or average values. For example, in a project designed to increase readiness for discharge among diabetic patients, abstractors may record the highest and lowest glucose values within the 24 hours before discharge. Although some information is lost, these approaches simplify data recording and increase accuracy. (Additional details can be found in the discussion on data quality, p 125.) Such transformations may or may not be appropriate on the basis of the question being asked.

**Medication variables.** Medication variables can be especially challenging for abstraction projects. Complete specification of a medication variable requires abstraction of (1) generic/trade name, (2) strength, (3) dosing, (4) administration route, (5) start time, and (6) stop time. Medication names may be selected from a master list, eliminating errors from textual entry of drug names. However, new medications are constantly introduced to the market, making it essential to keep medication files updated. The abstractor may use free text to enter medications not found in the master list. Dosages of medications taken at home may not be recorded in the chart. Finally, as with other abstracted elements, the degree of specification depends on the question being asked.

**Variable Definition**

**Source designation.** Source designation gives the abstractor explicit directions to specific chart
locations that are acceptable sources of information for a given variable. Often source designation lists for- 
bidden sources of information. This is more compli- 
cated than it appears on the surface. For example, 
physicians may use the same term for a diagnosis, a 
symptom, or even an observation, for example, 
depression/depressed or obesity/obese. If depression as 
a diagnosis is the question, then it may be found in the 
assessment section of the progress notes or in the past 
medical history. Depression as a symptom may be 
recorded in the history of present illness, in a review of 
systems, or on a new patient enrollment form. A 
physician may record in the physical examination that 
a patient appears depressed. The investigative team 
must clearly define for the abstractors where to find 
the information and what constitutes a positive response. 

Synonyms. Synonym lists are obviously impor- 
tant for accurate abstraction because a multitude of 
terms may refer to the same concept. We found 
numerous examples in our project in which the syn- 
onym list was key to accurate data abstraction. We 
also used extensive synonym lists in ascertaining 
comorbidity. While a synonym list applies to a par- 
ticular option of a categoric variable, the different 
options themselves form the pick list for that variable. 
Variables usually have one pick list but many options, 
each with a corresponding synonym list. An example 
of a synonym list developed by DynKePRO can be 
found under “Definition of Variables” in Appendix 1 
(p 133).

Time frame. Time frame issues complicate chart 
abstraction if the period of interest is more than one 
encounter. Time frame for the chart abstraction must 
be clearly stated. Many variables, including demo- 
graphics, occupation, marital status, comorbidities, 
and symptoms change over time. Abstracting dates 
and times allows a more accurate calculation of time 
intervals than having the abstractor compute the time 
during the abstraction process.

Dependence. Parent–child variables are useful 
for situations in which the need for information is 
dependent on the response to another question. The 
abstractor answers the child question only after 
recording a specified response for the parent variable. 
If a computerized abstraction tool is being used, the 
instrument may be designed so that the child variables 
appear only if there is a positive response to the parent variable. We found this design useful for quantifica-
tion questions (for example, dosage, duration, fre-
quency, intensity). For example, in recording 
depressive symptoms, the MedQuest module kept all 
questions regarding duration inactive until the abstracter marked a symptom as present.

Negative information. Recording negative infor-
mation requires special attention. Useful classifications for negative data include “not applicable,” “no 
documentation in chart,” “documentation of negative 
response in chart,” and “documentation in chart is not 
clear.” When there is no documentation in the chart 
regarding a particular variable, the team must decide if 
this will be interpreted as a positive or negative 
response. For example, lack of documentation for a 
rare disease such as pheochromocytoma probably 
indicates absence of the disease. In contrast, lack of 
documentation of a common symptom, such as sleep 
disturbance, cannot be automatically interpreted as a 
negative response.

Presentation of Chart Review Variables

Careful consideration of variable presentation 
conserves resources and reduces abstractor fatigue. In 
the abstraction module, we grouped variables accord- 
ing to their source in the chart. For example, we pre- 
sented the screens containing clinical history variables 
before the screens related to laboratory variables. Vari- 
able presentation in the chart review module may be 
complemented by the preprocessing of charts. In pre- 
processing, someone with less training than the 
abstractors reviews the chart and flags the major sec-
tions such as the history, physical exam, or laboratory 
data. Preprocessing may save time by rapidly orienting 
the abstractor to the chart.

Abstractor Selection, Training, and 
Management

Available positions were advertised on a bulletin 
board within the MCO, and several employees 
expressed interest. Two aspects of these part-time posi-
tions were key for potential candidates: (1) previous 
experience with abstracting medical chart data, and 
(2) ability to work during evenings and weekends. To 
avoid highly complex interrater reliability checks, a 
maximum of four abstractors was considered ideal.

Three abstractors were hired—two master’s level 
nurses and one health care analyst with a master’s in
public health. The abstractors participated in a five-hour training session led by the members of the research team who developed the abstraction tool. Three charts were abstracted during the training session—one done jointly, and then two done independently by all three abstractors. Responses were then reviewed together. The abstractors recommended certain modifications to the abstraction tool, and these modifications were implemented during the training session.

Variable-specific help, to which the abstractors have easy access, proves invaluable in the long term. In the short term, compiling detailed help demands a significant time investment. Our experience reveals that although individuals often interpret an item differently, the item itself does not necessarily need to be changed. Clarification of such disagreements belongs in the help section. The help section should be generous and document the following for each variable: (1) intention, (2) source, (3) time frame, (4) synonyms, (5) format for recording the variable, and (6) explanations of each response item contained in the pick list from which abstractors choose alternatives.

Because the purpose of the training manual was to aid the abstractors in understanding the general and specific goals of our project, the following items were included: (1) table of contents, (2) an explanation of the purpose of our project, (3) a listing of the DSM-IV criteria for the diagnosis of depression, (4) definition of quality indicators specific to the project, (5) a table of antidepressant medications and therapeutic dosages, (6) a description of the MedQuest tool, (7) helpful hints that explained how to access and use MedQuest, (8) abstractor tips for previously encountered difficulties, and (9) printouts of the help screens for all the variables. The help screens in the manual were organized with tabs that resembled the layout of the MedQuest tool. This layout offered an alternative for abstractors who found it easier to locate information from a printed desktop manual than to navigate to a different computer screen for the same information. Examples of variable-specific help and excerpts from our training manual can be found in Appendix 1.

Until interacting with the abstractors, we did not appreciate the magnitude of the bias that could be introduced by abstractor interpretation. For example, consider the comorbid variable obesity. An abstractor may interpret a weight of 200 pounds recorded in the chart as being obese. However, depending on sex and height, the patient may not be obese. Differential knowledge bases of the abstractors may cause problems with interrater reliability. On occasion, prior medical training interferes with abstraction quality by leading to overinterpretation.

The obesity variable in our study was more complex because obesity may be interpreted in two contexts. First, we were mostly concerned with physician-recognized obesity. We carefully instructed the abstractors to enter a positive response only if the physician actually recorded the diagnosis in the problem list, history and physical, or assessment and plan. Second, if we were also interested in the absolute presence of obesity regardless of physician recognition, the abstractor should record the actual weight and height of the patient.

Proper construction of quality indicators and research questions also minimizes inappropriate interpretation. As stated in the literature review section, each variable should contain only one unit of information, and care should be taken to avoid presenting the abstractors with multivariable quality indicators or algorithms. For example, defining a variable as positive when (1) the patient is female, (2) the patient is between the ages of 50 and 70 years, and (3) the patient received a mammogram leads to poor reliability and validity. Such a composite approach makes it difficult to discern the exact reason for poor reliability and validity and impedes refinement of the indicator. In addition, the components of a composite indicator cannot be analyzed individually, and this makes running supplementary analyses more difficult.

As in the example above, most quality indicators should be derived in the analysis phase rather than in the chart abstraction phase. Abstracting each element of the indicator as a separate variable allows for postabstraction modification of the indicator. For example, collecting each component variable separately allows the team to modify the cutoff ages for receipt of a mammogram. If the indicator were abstracted in one step as one variable at the time of abstraction, postabstraction modification may become almost impossible.

Chart Procurement

We initially planned to abstract 440 medical charts from primary care providers distributed over a large geographic area. Therefore, central abstraction with photocopies was necessary. Members were selected as
previously described, and letters were sent by the MCO to the appropriate primary care providers. In addition to explaining the purpose of the study and describing the investigators, the letter asked the providers to copy all relevant parts of the chart. The letter specifically requested that all forms containing medical information (progress notes, history and physical examination notes, problem lists, medication lists, consultation reports, and medical information forms completed by the patient) be copied for a defined 18-month period. We noted that regardless of this cooperative nature of the project, the chart review would be done by the MCO and no member-identifying information would be released to the academic institution. All team members were trained regarding confidentiality concerns and are bound by the MCO’s confidentiality safeguard policies. All charts were tracked through a manual logging system. When the charts were not being used by the abstractors, they were kept in a secure location.

We expected that a fair number of providers would not comply with a request by mail and that some charts would be illegible. Therefore, on the basis of previous experience, we oversampled by 100%, that is, we requested charts for 880 randomly selected members. First, we divided all MCO members into two groups: those who were algorithm positive for depression and those who were algorithm negative. Next, we randomly sampled from each group. This sample included 420 members with depression, as defined by the algorithm under study, and 460 controls. Because we wanted the cases and controls to be matched on certain elements such as age, sex, and number of comorbidities, we used a stratified randomization scheme.

Measurement, Monitoring, and Maintenance of Data Quality

We view the quantitative measures of reliability and validity as indispensable tools for improving and maintaining the quality of chart review data. Desirable levels of validity and reliability cannot be set outside the context of a specific research project. There is a clear trade-off between achieving high data quality by maximizing validity and reliability and increasing chart review cost. Also, certain variables central to the main question of the project may warrant higher levels of validity and reliability than ancillary variables. Before proceeding with specific methodologic and implementation issues and with the reporting of the results of our quality measurement, a clarification of these concepts is in order.

Reliability

In the strictest sense, reliability is a measure of both reproducibility and discrimination. A chart review tool that provides consistent results (when different operators abstract the same charts) possesses high reproducability. An abstracted variable that correctly divides patients into different groups (for example, those with disease versus those without) provides a high level of discrimination.

Frequently, interrater reliability can be increased at the expense of discrimination by making a continuous variable discrete or by combining categories of a discrete variable. This principle is illustrated in our choice of variable format for duration of depressive symptoms. In ascertaining duration of depressive symptoms, we gave the abstractors the choice of three options, as already stated, instead of requesting that the exact time be recorded. We did not need more detail for the project, since DSM-IV uses only symptoms present for two or more weeks in diagnosing depression. However, for other projects, the need for detail depends on the research question.

Interrater reliability may be measured as the percentage of agreement when multiple operators abstract the same charts. Agreement on a continuous variable may be ascertained in a dichotomous fashion by giving credit when the two values are within a certain margin of error, such as ±10%. Intrarater reliability, in which one abstractor abstracts the same chart twice, may substitute for interrater reliability when there is only one abstractor. In addition, measurement of intrarater reliability provides the abstractor useful guidance during the training phase.

We recommend a goal of 95% reliability for important variables. At first glance, this standard appears excessively stringent, but it can be achieved, as illustrated by the Cooperative Cardiovascular Project. However, certain circumstances often encountered in chart review artificially inflate interrater reliability, for example, when the response to a categoric variable is heavily weighted, as in ascertaining the presence or absence of a rare condition in a population. Because the absence of this condition is not documented in the charts in a majority of the cases, abstractors would tend to agree with each other by chance alone. The kappa
Validity

Validity, which reflects the degree to which the measured variable captures “the truth,” defies a simple yet precise definition. Traditionally, psychometric writers divide validity into content, construct, and criterion validity. Content validity reflects the adequacy with which the variables relate to the topic of interest and provide comprehensive coverage of all elements in the topic. As a chart review tool becomes broader and more detailed, it usually becomes more representative, with a corresponding increase in content validity. Results of experiments based on the underlying theory or construct of the chart review project demonstrate construct validity. For example, a chart review tool designed to measure past utilization that also predicts future utilization and cost possesses high construct validity. Measurement of content and construct validity lies outside the scope of chart review and depends on the current study design as well as future studies.

Criterion validity, the form of validity most useful in improving the quality of medical record abstraction data, compares a given measure with other known measures, one of which is often the gold standard. For the purposes of chart review, we focus on criterion validity, using expert clinical judgment as the gold standard. Practically, we measure criterion validity through a process described as chart adjudication, through which a gold standard for each element in each chart is constructed, and agreement of each abstractor with this standard is calculated. Although using adjudication to measure criterion validity involves assessing interrater reliability, adjudication depends on the subjective judgment of a clinical expert to resolve discrepancies between abstractors. This differs from the measurement of reliability described previously, in which the agreement of independent abstractions on supposedly objective variables from the same chart are assessed.

In developing the gold standard, multiple experts perform independent abstractions of the same data elements for a series of charts. During adjudication, agreement by the experts is examined for each variable in each particular chart. When there is disagreement on a particular variable, clinical consensus determines which response, if any, is correct. The consensus process may be as simple as a discussion between two clinicians. For practical purposes, it is sometimes assumed that when two abstractors agree on a given variable, both abstractors chose the correct response. However, this approach leads to an overestimate of validity because experts may obviously agree on the incorrect answer. During the process of adjudication, the developers of MedQuest suggest classifying reasons for disagreement as resulting from one of the following: data entry error, missed information, computer mismatch, poor record copy, unclear element definition, unclear location, not following rules, or conflicting information.

Reasons for Measuring Reliability and Validity

Measuring reliability and validity offers the following advantages to chart review projects: (1) assistance with development of the chart review tool, (2) continuous maintenance of data quality throughout the production phase of chart review, and (3) final documentation of data quality. For projects that require ongoing abstraction of large numbers of clinical records, the team may monitor data quality by using control charts based on the principles of statistical process control. Pursuing quality control consumes additional resources and, as usual, the team must find the balance between cost and quality.

The practice of statistical process control originated in manufacturing industries, where it is necessary to continuously monitor a stream of products made on the assembly line. The philosophic underpinnings of statistical process control state that before any process can be improved, it must be understood because acting on random variation pushes the system toward chaos.

Control charts are useful in documenting data quality and in determining when a fall in performance indicates systematic problems that must be addressed versus random variation requiring no immediate action. (For smaller projects involving a one-time abstraction, control charts may not be necessary.) Control charts, based on rational sampling, provide a longitudinal view of system performance and provide limits within which random variation acts. In general, control limits are set at three sigma units above and below an average quality measure. Sigma units, which often differ from standard deviations, reflect an average of several different measures of dispersion. With the exception of binomial
distributions, control limits computed based on a single measure of dispersion are overly sensitive to random variation, and their use poses practical barriers. Additional discussion is provided in Appendix 3 (pp 135–136).

Results

Chart Procurement

Because 385 copies of charts were obtained within four weeks of the first request, a second request was deemed not necessary. By ten weeks after the request, 451 charts had been obtained, and, ultimately, we received 465 unique charts. An additional 36 providers responded that they had not seen the patient during the specified time. The overall response rate to the mailing was 57%. Additional details are provided in the time table shown in Table 2, above.

Quality of Chart Abstraction

We divided pilot testing into three phases: prepilot, formal pilot, and production. In the prepilot phase, we conducted three sets of dual chart abstraction, which resulted in interrater agreement rates of 80%, 90%, and 91%. At this point, we relied on preliminary data to locate charts thought to be representative of our study population. (We did not count the charts used in prepilot testing in the total number of dually abstracted charts.) After training, the abstractors themselves performed the formal pilot phase to document the final quality of their ability to use the chart abstraction tool. For the formal pilot abstraction test, 5 charts were abstracted by all three abstractors, with an overall agreement rate of 97%. During the production phase, a total of 82 charts were abstracted by two different abstractors. In contrast to our approach in the prepilot and formal pilot phases, we randomly selected the charts for dual abstraction in the production phase. The charts were approximately evenly distributed between a reabstraction set at the beginning of production, a quality check when about one-half of the total charts had been abstracted, and a final quality check after abstraction was complete. All quality checks resulted in a reliability of at least 96%.

We noted that agreement rates obtained from pilot testing were better than those from prepilot testing. The prepilot testing, which was conducted when the chart abstraction tool was at an early point in development, was intended for internal use only.

Also, the abstractors who conducted the pilot testing, as compared with the investigators performing the prepilot testing, were probably more meticulous and less hurried.

Resource Use

Because of the multiple variables involved, a detailed economic analysis is beyond the scope of this article. Here we provide only estimates of the number of hours we required to complete the major tasks in the chart review phase of the project. Because salaries may vary over time and by region and institution, dollar amounts are not given. In presenting resource use, we take a marginal approach, excluding the cost of existing infrastructure (for example, computer and network hardware and software, office space, routing office functions).

Time required for chart abstraction. The abstractors used different methods for abstracting charts, according to their own preference. For
example, one abstractor read the entire chart first and highlighted key pieces of information before entering any data, and another entered data as it was encountered in the chart, skipping from screen to screen. The interrater reliability estimates suggest that both methods for chart abstraction may be equally effective. The average abstraction time per chart for all abstractors was approximately 10 minutes (Table 3, above), with most charts taking 30 minutes or less to abstract (Figure 2, p 129). Because we measured abstraction time by a computer clock, time spent reviewing the chart before accessing the MedQuest chart abstraction program leads to underestimation of abstractor time. This is most relevant for the abstractor who read the entire chart first. Abstractors tended to improve in efficiency over time.

**Effort of project team.** The time required to complete each major portion of the project relevant to chart review is provided in Table 4 (p 130). In interpreting Table 4, one must consider two important aspects of this project. First, this project is an academic endeavor designed to scientifically test specific hypotheses. Therefore, building and implementing a similar chart review project as part of a more routine QI project may require less time. Also, we encountered a significant number of unexpected obstacles, and we anticipate that on future projects our efficiency will improve. Finally and most important, the amount of time required to develop the chart abstraction module, train the abstractors, and perform quality monitoring was appreciable, approaching the amount of time actually spent in data entry.

**Project time table.** As indicated in the project time table (Table 2), reaching the point of data analysis took more than one year. However, this interval overestimates the actual time required because the investigators were funded at a level of 0.2 full-time equivalent (FTE). We estimate that it would have required our team working full time two months to reach the point of data analysis. Based on the lessons learned from this project, we anticipate much shorter abstraction module completion times in the future.

Because of delays encountered in the time table, we now know to include more flexible time for chart abstraction when planning future studies. For convenience, we have categorized the reasons for time delay as project related, technical, and personal. The most important project-related delay resulted from the necessity of multiple revisions of the abstraction module during the prepilot phase of module design. Other delays resulted from the part-time status of the abstractors and the time lag in receiving photocopies of the charts from the physician offices. Technical delays included network and computer problems. We encountered difficulty in installing MedQuest on a local area network. Finally, personal factors such as incompatible schedules, vacation time, and holidays also introduced delay.

**Conclusion**

In completing the chart review phase of our QI research project, we learned four important lessons, which will enable us to proceed with the project more efficiently. First, we had difficulty finding answers to our methodologic questions. Many investigative teams had developed excellent chart review protocols, but it was difficult to tap into this pool of expertise. In searching the medical literature we did not uncover many useful articles. We recommend that anyone planning a chart review project become familiar with MedQuest and its documentation. Even if a paper-based collection format will be used, the MedQuest documentation offers many good suggestions. The HCFA Web site (www.hcfa.gov) has examples of well-developed MedQuest modules for several diseases. These modules contain extensive synonym lists for variables. Also, seeking expertise in the PRO community or a CDAC should be considered.
Second, the precise specification of abstraction variables is more difficult than we anticipated. We now understand the difficulties in attempting to formulate an acceptable variable definition without examining representative charts and without testing all proposed definitions in a variety of settings. The abstractors suggested modifications to the variable specifications not originally considered by the investigators. In fact, much of the increment in interrater reliability that we achieved reflects the abstractors’ suggestions.

Third, in planning the project, we did not allow adequate time for developing the chart review module. Also, unanticipated problems caused significant delays after the chart module was completed. Only a portion of the delay we experienced resulted from easily remedied problems. Even with the insights we gained, we will build more time for chart review into our next project.

Fourth, it is important to obtain an in-depth understanding of available administrative data. Using existing administrative data saves times and prevents frustration. However, the properties of all administrative data sets differ, and improper use may produce misleading results. We used administrative data for medication variables. At first, we planned to collect medication variables through chart review. After we experienced the magnitude and difficulty of this task, we realized that, in our particular case, the administrative files contained better medication data than we could collect with chart review.

In general, chart review is more difficult than it appears on the surface. Chart review is also project specific, making a cookbook approach difficult. Many factors, such as imprecisely worded research questions, vague specification of variables, poorly designed abstraction tools, inappropriate interpretation by abstractors, and poor or missing recording of data in the chart, may compromise data quality. Table 5 (p 131) summarizes important points that provide a good starting place for planning future projects. Table 6 (p 131) provides a succinct checklist for the evaluation of methodologic descriptions of chart review found in the literature.
Table 4. Estimated Personnel Time (Hours)

<table>
<thead>
<tr>
<th>Task</th>
<th>Doctoral-level investigators</th>
<th>Research assistant</th>
<th>Computer programmer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Preparation of chart review protocol</td>
<td>30</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>Team meetings</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MedQuest module construction</td>
<td>12</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Pre-plot testing</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chart acquisition</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Algorithm for chart identification</td>
<td>10</td>
<td></td>
<td>24</td>
</tr>
<tr>
<td>from administrative data</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Preparation of letters to providers’ offices</td>
<td>2</td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>Receiving/tracking of charts</td>
<td></td>
<td></td>
<td>20</td>
</tr>
<tr>
<td>Telephone contact with providers’ offices</td>
<td>6</td>
<td></td>
<td>5</td>
</tr>
<tr>
<td>Abstractor training</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Development of training manual</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Didactic session/pilot testing</td>
<td>10</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>Production</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data entry by abstractors*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Network support</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Quality monitoring</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data entry by abstractors*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Analysis of duplicate abstractions</td>
<td></td>
<td></td>
<td>20</td>
</tr>
<tr>
<td>Adjudication</td>
<td>4</td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>Unanticipated</td>
<td>5</td>
<td>5</td>
<td>10</td>
</tr>
<tr>
<td>Totals</td>
<td>79</td>
<td>60</td>
<td>88</td>
</tr>
</tbody>
</table>

* The average time per chart for data entry by abstractors was 9.76 ± standard deviation of 14.36 minutes.

References

Table 5. Important Points for Planning a Chart Review Project

<table>
<thead>
<tr>
<th>Abstraction module development</th>
</tr>
</thead>
<tbody>
<tr>
<td>Consider module development an integral component of study design</td>
</tr>
<tr>
<td>Build a multidisciplinary team</td>
</tr>
<tr>
<td>Move quickly from the cerebration phase to pilot testing with actual charts</td>
</tr>
<tr>
<td>Use prepiilot testing to</td>
</tr>
<tr>
<td>- identify confusing or misleading questions</td>
</tr>
<tr>
<td>- force clarification of questions</td>
</tr>
<tr>
<td>- revise the tool from the abstractors’ point of view</td>
</tr>
<tr>
<td>- perform pilot testing with charts similar to those that will be encountered in production</td>
</tr>
<tr>
<td>Provide help screens describing option choices, synonyms, and instructions</td>
</tr>
<tr>
<td>Invest up front in building a quality abstraction tool for significant subsequent dividends</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Chart procurement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Generate a “pull list” from administrative data</td>
</tr>
<tr>
<td>Send a letter briefly describing the project along with the request for charts</td>
</tr>
<tr>
<td>For large projects, copy charts on site, but abstract them centrally</td>
</tr>
<tr>
<td>Anticipate questions from the physicians’ offices about the purpose of the project, financial reimbursement, and logistical considerations</td>
</tr>
<tr>
<td>Consider preloading patient identifiers and demographics from administrative data</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Abstractor training</th>
</tr>
</thead>
<tbody>
<tr>
<td>For projects not requiring specialized expertise, value prior abstraction experience over medical background</td>
</tr>
<tr>
<td>Provide detailed educational presentation and abstractor training manual</td>
</tr>
<tr>
<td>Treat abstractors as integral to the project and make sure that they understand the importance and significance of what they are to do</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Production</th>
</tr>
</thead>
<tbody>
<tr>
<td>Provide abstractors with flexible working hours</td>
</tr>
<tr>
<td>Discourage abstractors from going too fast or working too long without a break, as the process is tedious</td>
</tr>
</tbody>
</table>

Table 6. Checklist for Description of Chart Review Methodology

- Was a chart review protocol documented?
- Was the chart review protocol refined through pilot testing?
- Were standard lists used for variable synonyms, medications, diagnoses, and procedures?
- Was abstractor training described?
- Was data quality monitored during production?
- Were reliability and validity reported?


Continued
References (continued)


Appendix 1. Excerpts from the Abstractor Training Manual

The MedQuest Tool

MedQuest is a data abstraction tool developed by the Health Care Financing Administration (HCFA) through Fu and Associates. It is a window-based computer program that allows for tool designs specific to a given project. You will find this program user-friendly, including the availability of help screens with a click of the right mouse button.

Subject matter is separated by a series of screens that appear as file folder tabs in the tool. One click on a screen tab enables access to the variables relevant to that screen. Furthermore, as variable questions are answered, the variable changes color (to blue) to save time looking over each variable to ensure you have answered it. In addition, once all of the variables on a given screen have been answered, the screen tab will appear green in color, also saving valuable abstraction time.

We have endeavored through our own prepilot studies here at UAB to design the variable questions as clearly and concisely as possible, for reliability and abstraction ease. However, please feel free to offer suggestions and comments during this pilot phase as to how we may improve upon the MedQuest tool.

The MedQuest Screens

The MedQuest tool designed for this project has a total of six screens, which include:

- demographics, including patient’s name, date of birth, Social Security number, Aetna identification number, gender, occupation, and race. The asterisked items denote Aetna identifier variables. This information will not be forwarded to UAB to protect the confidentiality of the patients.
- DSM-IV Criteria 1 and 2.
- ancillary, including such variables as diagnosis of depression, referral of depression, family history of depression, change in libido, progress notes, and patient phone call frequency.
- comorbid Conditions 1 and 2, describing the illnesses suffered by the patient in addition to depression, which may also influence the patient’s diagnosis and/or response to treatment.

Helpful Hints

1. Getting into the system:
   - From the computer’s Program Manager screen, click on the MedQuest icon.
   - A screen will appear with the title, “MedQuest - Clinical Data Collection Design System.” In the box labeled “List of Modules” you will see “AED Aetna Depression.” Click here once.
   - To the right you will see two empty boxes entitled, “User Name” and “Password.” In each of these boxes type in “id1.”
   - Next, click on the “Load Module” button.
   - On this new screen, you will see an “Options” box on the left. Click on the “Data Entry” option.

2. Starting a new case:
   - On this same screen, find the long, thin Case ID box and type in the case identification of the patient, and click “OK.”
   - A message box will appear informing you that the case does not exist, would you like to create one? Click “Yes.”
   - You should now see the six file tabs that represent your subject screens. Single click on any one of the screen tabs to view the variable questions.

3. The questions will appear in red until answered.
   - When you move on to the next question, the answered question will turn blue. This will make life easier when switching from screen to screen to answer different questions.

4. When to use the right mouse button. Place the mouse pointer on a variable and right click:
   - to view the help screen of the variable
   - to clear a variable of a selected answer(s)
   - For the “Choose One Or More” variables, the “Clear Variable” option will not appear with a right click. To clear an answer from one of these variables, simply left
Appendix 1. Excerpts from the Abstractor Training Manual (continued)

click on the answer a second time. The “x” will disappear.
5. To change to another screen, click on the tab of the screen you wish to view.
6. When all of the variable questions on a screen have been answered, the screen tab will turn green, letting you know that you have completed the screen.
7. When all of the screen tabs have turned green for a given case and you wish to exit the case (or you have not completed the case, but need to shut it down for a while), click on “File” in the upper left-hand corner of the screen, and then click “Exit.”
8. A screen will appear, called the “Case Status Screen,” which will ask why you wish to exit. Please only mark the “Completed” option if all of the screen tabs are green. Either of the other two options will allow you to re-enter the case with no difficulty. After choosing your option, click on the “Exit Case” button on the bottom right-hand side of the screen. From here, you return to the case identification screen, where you may begin a new case or exit MedQuest by clicking on the “Exit” button beside the “OK” button.

Abstractor Tips
1. The difference between “yes,” “no,” and “not mentioned,” using obesity as an example variable.
   - YES indicates that the chart documents that the patient is obese, or one of the synonyms noted on the help screen for obese, such as heavy, stout, or fat.
   - NO indicates that the chart documents that the patient is NOT obese, heavy, stout, or fat.
   - NOT MENTIONED indicates that the chart does not address obesity, or at least not in a negative fashion with regard to the patient.
2. Note the two “warning” labels on the DSM4 screens.
   - Label one: DSM4 Criteria Concern Patient SYMPTOMS, NOT physician DIAGNOSIS.
   - Label two: If any DSM4 symptom is stated as a SIDE EFFECT of medication, DO NOT MARK.
3. Do not make judgments or “read in” information from the chart. For example, hypersomnia or insomnia. Absence of one does not imply presence of the other.
4. The term “depressive symptoms” by itself does NOT imply depressed mood.
5. Note the variables with “Choose One Or More” after the variable question, such as the occupation variable and most of the variables pertaining to the Comorbid Conditions screens.

Definition of Variables
The example provided below was developed by DynKePRO and used in our chart review module.

Cardiovascular
Question:
Record on this screen conditions the patient may have that involve the heart or the arterial circulation.

Sources:
All information available in chart, including patient information sheet.

Options:
a. Mark “Coronary Artery Disease (CAD)” if the chart documents a diagnosis of coronary artery disease. Synonyms include:
   - Acute coronary insufficiency, Angina, Angina any type, history or chronic, Angina decubitus (Angina at rest), Angina pectoris, Angioplasty in past, Atherosclerotic cardiovascular disease, Arteriosclerotic heart disease, ASCVD (atherosclerotic/arteriosclerotic cardiovascular disease), ASHD (Atherosclerotic/arteriosclerotic heart disease), Atheroma, Atherosclerosis, Atherosclerotic heart disease, Bypass surgery/CABG, CABG (Coronary artery bypass graft), CAD, Cardiac bypass surgery, Cardiac insufficiency, Cardiomyopathy, CHF, Coronary angioplasty, Coronary angioplasty with or without stent placement (palmaz stent), Coronary artery angioplasty, Coronary artery bypass surgery, Coronary artery surgery, Coronary atherectomy, Coronary balloon angioplasty, Coronary disease/condition, Coronary insufficiency, Coronary occlusion, Coronary revascularization, Coronary thrombosis, Crescendo angina, Heart bypass surgery, History of chest tightness, or pain attributed to coronary artery disease, Intermediate coronary syndrome, Ischemic cardiomyopathy, Ischemic chest pain (history of), Ischemic disease or condition, Ischemic heart disease (IHD), Isosorbide use, Myocardial disease/condition, Myocardial revascularization, Nitrobid use, Nitrobid/nitroglycerin use, Nitroglycerin (NTG) use, NTG patch use, Percutaneous coronary angioplasty (PTCA) in the past, Percutaneous transluminal angioplasty (PTA) of the coronary artery, Percutaneous transluminal coronary angiography (PTCA), Percutaneous transluminal coronary excimer, Percutaneous
transluminal coronary rotorblade, Post myocardial infarction syndrome, Prinzmetal angina, Stable angina, Subendocardial injury, Unstable angina, Variant angina.

b. Mark “Arrhythmia” if the chart documents a diagnosis of arrhythmia.

c. Mark “Congenital Heart Disease (CHD)” if the chart documents a diagnosis of congenital heart disease.

d. Mark “Congestive Heart Failure (CHF)” if the chart documents a diagnosis of congestive heart failure.

e. Mark “Hypertension (HTN)” if the chart documents a diagnosis of hypertension/high blood pressure.

f. Mark “Mitral Valve Prolapse” if the chart documents a diagnosis of mitral valve prolapse.

g. Mark “Myocardial Infarction (MI)” if the chart documents a diagnosis of myocardial infarction/heart attack.

i. Mark “Other Cardiovascular Conditions” if the chart documents a diagnosis of a cardiovascular condition not mentioned above.

j. Mark “Not Mentioned” if the chart does not address the patient’s cardiovascular condition in a negative fashion.

Appendix 1. Excerpts from the Abstractor Training Manual (continued)

The following example provides a more detailed consideration of the kappa statistic (*k*). Suppose that we need to record the presence or absence of a diagnosis of hypertension for 2,000 charts from a general medicine practice. The charts are completed, and dual chart abstraction is performed for 100 charts (5% of the total) for quality control.

Array the data in a $2 \times 2$ table, where the rows contain the responses of the first abstractor, the columns contain the responses of the second abstractor, and the margins contain row and column totals (Table 1, right). The table reveals that both abstractors agreed that 20 charts documented the presence of hypertension and that 40 charts did not. The abstractors disagreed on the remaining 40 charts. The simple agreement rate is 0.60, but this rate must be judged against the agreement that would be expected by chance.

In Table 1 (right), the expected agreement counts are derived from the products of the corresponding row and column sums. For example, take the first cell where both abstractors agreed that 20 charts recorded the diagnosis. The marginal row sum is 30 and the marginal column sum is 50. The expected count for this cell is $(30 \times 50)/100 = 15$. Continuing this process yields the following table of expected values. Therefore, the total agreement expected by chance alone is $(15 + 35)/100 = 0.50$. Therefore, for this example, $k = 0.2$, which is obviously much lower than the simple rate of agreement. However, because the kappa statistic imposes a more stringent standard than does the simple agreement rate, interpretation must be different. Rosner suggests the following: (1) $k$ greater than 0.75 denotes excellent reproducibility, (2) $k$ between 0.40 and 0.75 denotes good reproducibility, and (3) $k$ less than 0.40 denotes marginal reproducibility.

The kappa statistic is now defined as follows:

$$k = \frac{(P_o - P_e)}{(1 - P_e)}$$

where $P_o$ is the observed agreement and $P_e$ is the expected agreement.

**Table 1. Observed and Expected Agreement Counts**

<table>
<thead>
<tr>
<th></th>
<th>Yes</th>
<th>No</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>20</td>
<td>10</td>
<td>30</td>
</tr>
<tr>
<td>No</td>
<td>30</td>
<td>40</td>
<td>70</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>50</td>
<td>100</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Yes</th>
<th>No</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>15</td>
<td>15</td>
<td>30</td>
</tr>
<tr>
<td>No</td>
<td>35</td>
<td>35</td>
<td>70</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>50</td>
<td>100</td>
</tr>
</tbody>
</table>

One may choose to monitor reliability or validity with agreement counts for individual variables, meaningful groups of variables, and/or by aggregating all chart variables into one group for an overall rate. Because agreement rates should approximately follow a binomial distribution, \( np \)-charts provide an appropriate mechanism of ongoing quality assessment.* Using an \( np \)-chart requires the following assumptions: (1) the number of dually abstracted charts is the same in each sample; (2) each variable is classified as reflecting or not reflecting agreement; (3) the probability of agreement may change from sample to sample, but must remain constant within a given sample; and (4) the likelihood of agreement of any variable is not influenced by the finding of agreement on any other variable in the sample. Given these assumptions, then

\[
UCL = np + 3 \sqrt{np(1-p)}, \\
CL = np, \\
LCL = np - 3 \sqrt{np(1-p)},
\]

where \( UCL \) is the upper control limit, \( CL \) is the control limit, \( LCL \) is the lower control limit, \( n \) is the number of charts per sample, and \( p \) is the pooled agreement rate of all samples.

Now consider a hypothetical project, where we want to abstract 2,000 charts and monitor the overall agreement rate (Figure 1, below). A 5% reabstraction rate means we must dually abstract 100 charts. During the initial phases of production, we reabstract a higher percentage of charts and use these to establish the control parameters to guide the remainder of the abstraction. This unequal sampling scheme with an emphasis on sampling more charts early in production lets the team understand how the system is performing and, if performance is satisfactory, provides the parameters by which the quality of the remaining abstractions will be judged.

Please refer to Figure 1 for a schematic example of a sampling scheme. Take the first 200 of the 2,000 charts as pilot charts to determine the control limits. Divide the production of these charts into four segments, each containing 50 charts. Reabstract 10 charts from each segment and use the resulting data from to provide the control limits. Divide the remaining 1,800 charts of the production phase into six segments of 300 charts, and for each segment abstract 10 charts for ongoing quality measurement. This procedure will result in a 20% sampling rate for the first 200 charts, a 3.33% sampling rate for the remaining 1,800 charts, and a 5% sampling rate overall.

Make the additional assumption that the interobserver agreement counts for the first four samples from the 200 charts were as follows: 9, 10, 8, 9. Then \( n = 10, p = 36/40 = 0.90, CL = 9, UCL = 10, \) and \( LCL = 6. \) (Note that the true \( UCL \) was reduced to the maximum agreement count.) Also assume that the following counts were obtained on each of the subsequent reabstraction samples: 10, 9, 7, 10, 9, 9. Then the overall agreement for the 100 dually abstracted charts is 0.90, and the low count of 7 reflected random variation, which required no immediate corrective action (Figure 2, p 136).

Appendix 3. Adaptation of Control Charts for Medical Record Review (continued)

Figure 2. This figure shows a control chart based on data from Figure 1 (p 135). The third sample had an interrater reliability of 70%, but this decrement in performance reflects only random variation.