A GUIDE TO THE DEVELOPMENT OF CERTIFIED SHORT FORM SURVEY INTERPRETATION AND REPORTING CAPABILITIES

Mark E. Maruish, PhD
Mark Kosinski, MA
# Table of Contents

Copyright .......................................................................................................................................................... ii  
Table of Contents ............................................................................................................................................ iii  
Preface ............................................................................................................................................................. vii  
How to Use This Guide ................................................................................................................................. xi  

## Chapter 1  Introduction
Reporting Short Form Survey Results: General Considerations ................................................................. 1  
Interpreting Short Form Survey Results: General Considerations .............................................................. 2  
QualityMetric’s Vendor Certification Program ............................................................................................ 3  
Purpose of this Guide ...................................................................................................................................... 3  

## Chapter 2  General Approaches to Reporting and Interpreting Short Form Survey Results
Reporting Short Form Survey Results ............................................................................................................... 5  
General Considerations .................................................................................................................................. 5  
The Score Report ........................................................................................................................................... 7  
The Profile Report ........................................................................................................................................ 7  
The Interpretive Report ................................................................................................................................ 7  
The “Four-Box” Report .................................................................................................................................. 7  
Interpreting Short Form Survey Results .......................................................................................................... 8  
Assumptions in Interpreting Short Form Survey Results ............................................................................... 8  
General Considerations .................................................................................................................................. 8  
Norm-Based Interpretation ................................................................................................................................. 9  
Content-Based Interpretation ............................................................................................................................. 10  
Item-Level Interpretation .................................................................................................................................. 10  
Criterion-Based Interpretation ........................................................................................................................... 11  
Guides to Short Form Survey Interpretation ..................................................................................................... 11  
Recommendations for Developing Interpretive Algorithms and Text ............................................................ 11  
Using Short Form Results to Trigger Immediate Intervention ....................................................................... 13  

## Chapter 3  Reporting and Interpreting Respondent-Level Data
General Considerations ................................................................................................................................. 15  
Disclaimer/Cautionary Statement .................................................................................................................... 15  
Identifying Information .................................................................................................................................. 17  
Data Quality Evaluation ................................................................................................................................. 17  
Missing Responses .......................................................................................................................................... 17  
Response Consistency ..................................................................................................................................... 17  
Response Pattern ........................................................................................................................................... 18  
Score Summary ............................................................................................................................................... 18
Table of Contents

Profile of Scores .......................................................... 19
Summary of Item Responses ............................................. 20
Norm-Based Interpretation .............................................. 20
  Using the Standard Error of Measurement (SEM) and
  Confidence Intervals (CIs) ........................................... 21
  Comparison to Age-by-Gender Norms ........................... 22
  Comparison to Disease-Specific Norms ................. 22
Content-Based Interpretation ......................................... 23
Criterion-Based Interpretation ........................................ 23
Depression Screening .................................................. 23
  MCS Score .................................................................. 23
  MH Score .................................................................... 24
SF-6D Health Utility Index ........................................... 24
Monitoring Change ..................................................... 24
Combining Results from Short Form Surveys and Other Instruments ........................................... 25

Chapter 4  Reporting and Interpreting Group-Level Data
General Considerations ................................................ 27
QualityMetric’s Outcomes Insight Consulting™ Services .......................................................... 28
Overview of Available Group-Level Interpretation and Reporting Services ......................... 29
  Data Quality Evaluation ................................................ 29
  Response Consistency .................................................. 30
  Missing Data Analysis .................................................. 30
  Tests of Scaling Assumptions ........................................ 30
Group-Level Scoring Summaries and Profiles .................................................. 31
Development of Group-Specific Interpretive Guidelines .................................................. 32
  Norm-Based Interpretation ........................................... 32
  Content-Based Interpretation ......................................... 32
  Criterion-Based Interpretation ......................................... 32
Burden of Disease Analyses ........................................... 32
Treatment Efficacy Analyses .......................................... 32
  Analyses of Change Within and Between Groups ....... 32
  Repeated-Measures Analyses ....................................... 33
  Analyses of Categorical Change Scores ..................... 33
Health Economics Estimations ....................................... 33
Mortality Risk Analyses ................................................. 33
Cross-Calibration of PRO Tools ........................................ 33

References ........................................................................ 35

Appendix  QualityMetric Health Outcomes™ Scoring Software 2.0 Sample Reports
Appendix A  Sample SF-36v2 Patient Report .......................................................... 38
Appendix B  Sample DYNHA General Health Survey Provider Report ....................... 39
Appendix C  Sample SF-12v2 Provider Report with Medical Expenditure .................... 40
Appendix D  Sample SF-12v2 Total Sample Professional Tabled Data Aggregate Report
  with Medical Expenditure ........................................... 41
Appendix E  Sample SF-36v2 Total Sample Aggregate Report for
  Management and Providers ........................................ 42
Some of the “Short Form family” of health status instruments developed by QualityMetric Incorporated—the SF-36v2® Health Survey (SF-36v2), DYNHA® SF-36® Health Survey (DYNHA SF-36), SF-36® Health Survey (SF-36), SF-12v2® Health Survey (SF-12v2), SF-12® Health Survey (SF-12), SF-8™ Health Survey (SF-8), and SF-10™ Health Survey for Children (SF-10)—are the most widely used health status assessment tools in clinical trials (Garratt, Schmidt, Mackintosh, & Fitzpatrick, 2002). The availability of more than 110 language translations and English-language adaptations of these instruments facilitates the adoption of the Short Form surveys in randomized clinical trials (RCTs), while the more than 1,800 published articles documenting RCTs utilizing the Short Form instruments attest to their utility and acceptance in the scientific community.

The increasing acceptance and use of the Short Form instruments for research and clinical purposes has been accompanied by the growing need for varied modes for administering, scoring, and reporting the results of administration, modes employing the latest technology. QualityMetric Incorporated has responded to this demand by making available a wide range of assessment capabilities for the Short Form surveys. In addition to the standard self-report paper Form format, the Short Form surveys can now be administered, scored, and/or have their results reported using interview scripts, approved translations of the surveys, and software developed for desktop computers, the Internet, handheld devices, interactive voice response systems (IVRS), and faxback systems.

Because of the rapidly growing worldwide demand for its commercial products and services, QualityMetric™ is expanding its licensing program for vendors and those involved in academic and commercially sponsored research who wish to develop means for administering, scoring, and/or reporting the results of the administration of Short Form instruments for their specific purposes using QualityMetric’s trademarks, copyrighted forms, scoring algorithms, norms, and interpretive guidelines. As a condition for licensing, QualityMetric now requires that any Short Form survey products developed (under license) by outside parties be certified as meeting standards of high quality, in terms of maintaining standardization and accuracy in survey administration, scoring, and reporting and interpretation of survey results. Certification can serve to ensure end-users (e.g., pharmaceutical companies) that Short Form survey products developed by a vendor other than QualityMetric (e.g., a contract research organization that develops and uses its own materials to collect, score, and report Short Form data for a clinical trial) meet QualityMetric quality standards. To assist licensees, QualityMetric has developed a series of three guides that describe important aspects of Short Form survey administration, scoring and reporting, and the guidelines and criteria that must be met in order for Short Form assessment products developed by licensed vendors to be certified by QualityMetric. The underlying theme of each of these guides is the importance of maintaining standardization in any assessment process utilizing any of the Short Form surveys.

The first guide in the three-part series, A Guide to the Development of Certified Modes of Short Form Survey Administration (Maruish & Turner-Bowker, 2009), provides
those developing Short Form survey forms and administration software with the guidelines and criteria that QualityMetric uses to license and certify modes of administration developed by parties other than itself, using QualityMetric’s Application Programming Interface (API). It also provides the background information necessary for understanding the requirements of QualityMetric’s certification program.

The primary purpose of the second guide, A Guide to the Integration of Certified Short Form Survey Scoring and Data Quality Evaluation Capabilities (Maruish & DeRosa, 2009), is to provide those licensed to develop Short Form scoring software with an overview of the scoring procedures that are contained in the API. This guide also describes QualityMetric’s methods and associated logic for its Data Quality Evaluation (DQE) and Missing Score Estimation (MSE) procedures.

This third guide in the series focuses on the standards for the development of certified interpretation and reporting capabilities for the Short Form surveys. The purpose of this guide is to discuss various ways in which Short Form survey data can be analyzed, interpreted, and reported to maximize its utility to the Short Form survey user. In addition, published resources for interpretation of data are identified, and report content and format considerations are addressed.

It is important for the reader to note that with the publication of this certification guide series, QualityMetric introduces three changes in terminology that it has been using in its commercial and peer-reviewed publications for over a decade. First, what was previously called “norm-based scores” is now referred to as “T scores.” Also, the set of procedures used to maximize the amount of useable Short Form data, previously referred to as “Missing Data Estimation (MDE),” is now called “Missing Score Estimation (MSE).” Finally, the “Reported Health Transition (HT)” item that is part of the SF-36 and SF-36v2 surveys is now referred to as the “Self-Evaluated Transition (SET)” item. The reason for these changes is to more precisely describe what each term represents and thus minimize misconceptions about what the term means among users of the Short Form surveys and other consumers of information derived from the Short Form surveys. Although the terminology has changed, what each term represents and how it’s used remains unchanged.

Information about each Short Form survey is available in its user’s manual, referenced in Chapter 2. In particular, the User’s Manual for the SF-36v2® Health Survey, Second Edition (Ware et al., 2007) provides a comprehensive history of the Short Form surveys, including a synopsis of the investigations leading to the development of the original SF-36 and a discussion of the improvements to the original version of the survey that have been incorporated into the SF-36v2 and SF-12v2. The SF-36v2 manual also provides form-to-form comparisons that can be useful for selecting the most appropriate Short Form survey for a given application.

Additional information about the history, development, psychometric properties, and use of the Short Form surveys is available from several sources. A comprehensive overview of the SF-36 survey is found in “SF-36 Health Survey Update,” a chapter by J.E. Ware, Jr., in The Use of Psychological Testing for Treatment Planning and Outcomes Assessment, Volume 3, Third Edition (Maruish [Ed.], 2004). More recently, Ware (2008) published an invited commentary in the Journal of Clinical Epidemiology summarizing the improvements to the SF-36 that led to the development of the SF-36v2. This commentary was the first article in a special series, called “Improvements in Short Form (SF) Measures of Health Status,” appearing in this journal. Other sources of information about the development and empirical testing of the Short Form surveys include several articles published in the November 1998 issue of the Journal of Clinical Epidemiology (Gandek & Ware [Eds.], 1998) documenting domestic and international studies of improvements in item wording and response categories that were evaluated pursuant to the International Quality of Life Assessment (IQOLA) Project, as well as more current studies on translations (e.g., Bjorner & Turner-Bowker, in press; Han, Lee, Iwaya, Kataoka, & Kohzuki, 2004; Turner-Bowker, DeRosa, & Ware, 2007).
Numerous other independent articles have been published and can be found at http://www.qualitymetric.com. Information about the Short Form surveys and the opportunity to discuss related issues with its developers are also available at http://www.sf-36.org/.
How to Use This Guide

Suggestions on how to use this guide and quickly find particular information are offered below. The reader is also referred to the Table of Contents to locate more specific information.

Introduction

Chapter 1 details the purpose of this guide and includes a discussion of general considerations for the interpretation and reporting of individual patient and group-level Short Form survey results.

Report content and format

Chapter 2 discusses four commonly used formats for reporting the results of health surveys and other psychometric measures, as well as considerations for determining which format and content are appropriate for the intended end users of the report.

Approaches to interpretation

Chapter 2 also discusses various approaches to the interpretation of Short Form results, including norm-based, content-based, criterion-based, and item-level interpretation.

Content of individual patient reports

Chapter 3 addresses important considerations pertaining to the development of software that presents interpretations of individual patient results obtained from the administration of a Short Form survey.

Content of group-level reports

Chapter 4 addresses important considerations pertaining to the development of software that presents interpretations of aggregated results obtained from the administration of a Short Form survey to groups of clinical patients or research subjects.
1

Introduction

The first two guides in this certification series (Maruish & DeRosa, 2009; Maruish & Turner-Bowker, 2009) address important issues related to the proper administration and scoring of the Short Form family of health status surveys. In these guides, the importance of maintaining standardization during the administration and scoring steps in the assessment process is emphasized. Standardization helps to ensure that results obtained from a survey administration are valid and can be compared with other Short Form survey results obtained using the same procedures. Being able to meaningfully compare findings is the basis for interpreting what they mean.

The purpose of this chapter is to provide QualityMetric’s certified software vendors with some general yet important considerations related to interpreting Short Form survey results and reporting these findings in a manner that will be useful to a clinician, researcher, or other interested party (e.g., patient, healthcare organization manager). More detailed discussions of interpreting results and reporting findings are presented in the remaining chapters of this guide. Additionally, this chapter presents a description of QualityMetric’s vendor certification program, as well as a brief overview of the purpose of this guide.

Reporting Short Form Survey Results: General Considerations

At first glance, the design of a software-driven system for reporting Short Form survey results and an automated (computer-generated) interpretation of those results seems to be a simple and straightforward task. However, QualityMetric’s experience in developing computer-generated reports and the feedback the company has received from Short Form survey users make it clear that this is not the case. Developing a report generated from algorithms, which are used to determine what data are presented and how those data are interpreted for the end user, requires careful consideration of several factors.

The first consideration is the intended audience for the report. Who is the target market, end user, or beneficiary of what is contained in the report? The wide range of individuals or organizations that may benefit from knowing Short Form results and/or their implications includes the clinical researcher conducting a drug trial, the individual himself (patient), the clinician who administered the survey, the managed care organization that requires its provider panel to administer the survey to their patients, and the organization that accredits managed care organizations. Moreover, it is possible that there will be more than one end user of the report. Discussed in more detail below, the target market/end user/beneficiary will significantly influence how the report is developed.

Another consideration is the purpose of the report. Here, the certified vendor must determine what need the report is meant to meet for the intended audience. The need of many clinical researchers is to have a straightforward report of scored data obtained from the administration of one of the Short Form surveys. Clinicians, however, may be less interested in the specific scores their patients obtained on the survey scales than in what those scores mean...
for their patients, in terms of self-perceived health status, and for the planning, monitoring, and determining of outcomes of treatment for the patient.

The established purpose(s) of the report will have implications for the next consideration: content. Content may include the raw and/or scored results of a Short Form administration, statistical analyses of those results, and/or interpretation of some or all of the available findings.

The final consideration is the design of the report. To best address the identified purpose(s) of the intended audience, it is important to determine the most effective way to present the desired content derived from the results of a Short Form survey administration. Here, several considerations must be taken into account, such as: Is the report primarily intended to convey the scored results from a Short Form survey administration, with or without some form of analysis of those results? If so, should the presentation be limited to just scores for the health domain scales and component summary measures, or should other scores or results (e.g., SF-6D, medical expenditure prediction) also be presented? Should the quantitative results be presented in tabular form, a graphic profile of scores, or both? Is the report also intended to offer the clinician or user an interpretation of the results and/or analyses thereof?

A more detailed discussion of important considerations in the development of Short Form survey reports is presented in Chapter 2 of this guide.

Interpreting Short Form Survey Results: General Considerations

Providing interpretations of Short Form survey scores and other quantitative results can add an enhancing dimension to a computer-generated report. Like other psychometric instruments, accurate interpretation of data from any of the Short Form surveys requires careful consideration. This is particularly important when the interpretation of the data is automated through the application of interpretive algorithms. In these instances, the interpretation is considered "blind;" that is, undertaken without knowledge or consideration of any aspect of the person(s) other than that which may be made known through any data accompanying the survey data. Blind interpretations of Short Form data—without or with any accompanying information—must therefore be limited to the available data and any or all aspects of functioning which research have empirically shown to be associated with that particular dataset for the person(s) in question.

The information provided by the analyses of the normative and supplemental data gathered during the development of the Short Form surveys can serve as a primary source of interpretive information. This information is reported in the manuals developed by QualityMetric for users of each of these surveys. These manuals provide guidelines or recommendations for interpreting Short Form surveys using a norm-based approach. With few exceptions, each manual also presents data enabling content-based and criterion-based interpretations of the data. Interpretation of the data at the item level provides yet another means of maximizing the amount of useful information that can be derived from Short Form survey data. These four approaches to interpretation are discussed in detail in Chapter 2.

Another source of interpretive information is the published literature. At the time of this guide’s publication, QualityMetric’s Short Form survey bibliographic database contained more than 12,000 entries. This catalog of publications attests not only to the widespread use and acceptance of the Short Form surveys for both research and clinical purposes, but also to the large body of empirically based information that extends the body of knowledge of Short Form health domain and component summary measure scores well beyond that provided by the surveys’ manuals alone. Identifying published literature on the Short Form surveys that is relevant to specific populations, treatment interventions, or assessment purposes is easily accomplished through literature searches of databases such as PubMed.
In summary, programmed algorithms used to generate interpretive text in automated Short Form survey reports should be based on empirical findings published in each survey’s technical manual and/or the published scientific literature. Unsupported interpretations should be avoided or, at least, labeled as such. An extensive discussion on issues pertaining to the development of computer-generated interpretive algorithms, narrative text, and reports is presented in Chapter 2 of this guide.

QualityMetric’s Vendor Certification Program

QualityMetric’s vendor certification program provides licensed vendors of QualityMetric survey products a means of demonstrating or verifying for their customers and other parties that the licensed products they develop and offer have met the high quality standards set by QualityMetric. These standards ensure that standardization of vendor-developed survey administration, scoring, and interpretation and reporting products have been maintained and that the results yielded by their use are consistent with those from products developed by QualityMetric. QualityMetric has adopted general certification standards for licensed vendors developing interpretation and reporting software. Vendors seeking certification for this type of software product must submit a software functional specification document or other form of documentation demonstrating that the product meets these standards. At the minimum, this documentation must address the following:

1. Because accurate scoring of data is a prerequisite of interpreting and reporting said data, certified vendors of Short Form interpretation and reporting software must also meet QualityMetric’s certification requirements for scoring, as presented in the second guide in this series (see Maruish & DeRosa, 2009).

2. The computer-generated report must include a unique identifier (e.g., name, ID number), date of survey administration, appropriate QualityMetric copyright and trademark information, and a cautionary statement (see Chapter 3).

3. Interpretive text and associated algorithms must be based on empirical findings from published research. Documentation must include references for the sources upon which the interpretive algorithms and text are based.

4. The computer-generated text must be clearly written at a level that is appropriate for the intended audience.

Purpose of This Guide

This guide was developed to meet the needs of vendors and other organizations licensed by QualityMetric Incorporated to develop software for the automated interpretation and reporting of individual and group-level (aggregate) data obtained through the administration of any of the Short Form surveys. The guide begins with an overview of common approaches to interpreting Short Form survey data and the assumptions that underlie them. Next, a detailed discussion of the possible content and structure of individual patient reports is presented, followed by a similar discussion pertaining to the development of reports for group-level, or aggregate, Short Form survey data. Because of their similarity in content and presentation, this guide focuses on the development of interpretive text for and the reporting of the SF-36v2, SF-12v2, and SF-8 data—the most frequently licensed Short Form surveys. The general guidelines for certified interpretation and reporting of data from these three instruments also apply to the SF-36, SF-12, and SF-10.

This guide is the third and final component in a series of guides developed to assist licensed vendors of Short Form survey materials and software in the development of products that meet QualityMetric’s certification standards. Those seeking certification from QualityMetric for their products should also review the guides on developing Short Form modes of administration (Maruish & Turner-Bowker, 2009) and integrating data quality evaluation and scoring capabilities (Maruish & DeRosa, 2009) into their products. Those who
are interested in learning more about the history, administration, scoring, interpretation, and psychometric properties of the Short Form instruments are encouraged to obtain copies of the manuals and user’s guides listed in Chapter 2 and available from QualityMetric Incorporated.
General Approaches to Reporting and Interpreting Short Form Survey Results

The use of the Short Form surveys for a clinical or research purpose is supported by years of empirical study and thousands of scientific publications. Contained within these sources is a large body of information that can provide a strong foundation for interpreting data obtained from the administration of the Short Form surveys. The amount of empirical support is much greater for some of the surveys (e.g., SF-36) than it is for others (e.g., SF-10), due in part to the length of time each survey has been available for scientific scrutiny. However, well-designed, scientifically sound investigations conducted by licensed vendors or their agents can provide almost any type of information needed to develop appropriate interpretive algorithms for Short Form survey data that will be useful for their particular purposes.

The purpose of this chapter is to present an overview of several approaches to reporting and interpreting Short Form survey data. First, matters to consider when developing a report for its intended end users are described, along with some common means of reporting findings from Short Form and other surveys. Next, assumptions underlying the interpretation of the Short Form surveys and a general strategy for understanding Short Form survey findings are presented. This is followed by a discussion of four interpretive approaches. These approaches are based on (a) a comparison of obtained scores with a survey’s normative data (norm-based interpretation); (b) a comparison of the individual’s or group’s specific item responses to the percentage of the comparison group(s) endorsing those responses across score levels of the health domain scales and component summary measures (content-based interpretation); (c) a straightforward examination of the responses to each of the individual items that the health domain scales and component summary measures comprise (item-level interpretation); and (d) an examination of the Short Form survey scores to established score relationships with external criteria (criterion-based interpretation). A referenced list of manuals and other publications useful in interpreting Short Form survey results accompanies this overview. Finally, general recommendations for developing interpretive algorithms and narrative text are presented.

Reporting Short Form Survey Results

There are many options for how certified vendors can present the results from the administration of a Short Form survey to the end user. Some of these options are discussed in the following sections. However, before the report content and presentation format are decided, several matters should be addressed.

General Considerations

Certified vendors developing Short Form reporting capabilities should carefully consider a number of factors when deciding what their automated report should include or look like. At a minimum, vendors should ask themselves the following basic questions:

1. **Who is the intended audience or end user of the report?** Computer-generated reports can be developed for many types of users, such as patients, providers, healthcare executives, researchers, and accreditation agencies. The potential
audiences for Short Form reports can run the gamut, from a very small, specific group (e.g., senior management staff at a large healthcare organization) to an assortment of users with wide-ranging information needs. For this reason, it is important to establish at the onset of development for whom the report is intended. This single piece of information is key to arriving at meaningful answers to the remaining questions in this exercise and, ultimately, to producing a report that is useful to the intended audience.

2. What types of information can the Short Form survey provide? The answer to this question will depend on the survey being considered. In terms of scales and scale content, available norms, derived measures (e.g., SF-6D, medical expenditure prediction), and the availability of established responder criteria and minimally important difference values, the SF-36v2 can provide the most information about adult respondents, whereas the SF-8 offers the least. Among all Short Form surveys, the SF-10 offers the least amount information; however, it is the only Short Form survey developed specifically for use with children and adolescents. In all cases, one must recognize the limitations in the types and amount of information that the Short Form survey results can provide in order to avoid offering the user more than can be empirically supported.

3. What types of available Short Form information best meets the end user’s needs, or what does the end user want to see in the report? In general, these questions speak to the need for clearly establishing the purpose of the report. This need should guide what is included in the report (i.e., content) and how it is presented. For example, if the primary audience for an SF-36v2 report is physicians who want to understand their patient’s level of health status burden or to monitor their patients’ progress over time, then it would be important to present results from both current and previous administrations of the survey in each report. Also, it would be useful to indicate whether changes in health domain and component summary measure scores are clinically meaningful, using the responder criteria reported in the SF-36v2 User’s Manual (Ware et al., 2007). On the other hand, the upper-level management at a large healthcare organization may only want a quarterly report of the mean of the aggregated SF-36v2 component summary scores and medical expense predictions for all patients seen by its provider panel. In this situation, narrative text and profiles of scores are neither necessary nor appropriate. Chapters 3 and 4 discuss the types of information that can be included in the Short Form survey reports for individual respondents and groups of respondents, respectively.

4. What is the end user’s level of understanding of the types of information that could be included in the report? The content of a Short Form survey report and the manner in which it is presented should be tailored to each user’s assumed ability to understand and assimilate the information. A report intended for use by the survey respondent will likely use language and a report structure that is quite different from a report intended for the respondent’s care provider.

5. What technical resources are available for developing the report software? In addition to software development expertise, certified vendors will need access to resources with a clear understanding of how the Short Form surveys are scored. In the case of reports that provide an automated interpretation, knowledge of the relevant published literature and the psychometric properties of the Short Form survey for which the report is being developed is also required. After answering these questions, the certified vendor will be better prepared to determine what type of report will meet the needs
Chapter 2: General Approaches to Reporting and Interpreting Short Form Survey Results

of different groups of end users. Described below are four common means or formats for presenting survey results through an automated reporting system that should be considered. In most cases, one of these formats can serve as a good blueprint for report design; however, vendors should not limit themselves to these options, as others may be found to be better suited to their particular needs.

**The Score Report**

A score report is the simplest of the automated reports. As the name implies, only scores for variables of interest are presented. These variables may include the health domain scales, component summary measures, SF-6D, medical expenditure predictions, relevant benchmark data, and/or scores from other surveys or measures that were administered at the same time as the Short Form survey. Raw data may be included along with the scored data (e.g., T scores). Identifying and/or demographic information regarding the respondent or groups of respondents whose results are being reported is also included. Typically, this type of report is no more than one page in length.

**The Profile Report**

The profile report plots Short Form survey scores on the standard Short Form T-score profile. The results from multiple administrations of the survey can be plotted on a single profile to show changes in scores (health status) over time. Moreover, relevant benchmark data available for the survey (e.g., age-by-gender norms, norms for diabetic patients) can be plotted along with the respondent’s scores, thus facilitating a greater understanding of the meaning of those scores. Commonly, profile reports also include the same content as the score report. Sample profiles displaying results from single and multiple Short Form survey administrations are presented and discussed further in Chapter 3 of this guide.

**The Interpretive Report**

As the name implies, interpretive reports provide the user with an interpretation of the results of an administration of a Short Form survey. In these reports, narratives provide an interpretation of a respondent’s survey results. The respondent-specific interpretation is based on the obtained Short Form scores, with or without consideration of other available data (e.g., demographics, medical information, data from one or more other measures), and empirically based algorithms. Datasets meeting specific algorithms trigger the inclusion of the interpretive text that is associated with those algorithms. Both the algorithms and the text can be as simple or complicated as the report developer wishes to make them; however, the narrative text (i.e., the interpretation) should be supported by empirical evidence. The development of interpretive text and associated algorithms is addressed in detail in later in this chapter.

In addition to narrative text, interpretive reports commonly contain the same information that is found in a profile report (e.g., score tables, profiles). Also, although interpretive reports are usually designed to address the findings for a single survey respondent, the same type of report can be developed to address group-level findings.

**The “Four-Box” Report**

The standard format used by QualityMetric in developing its own computer-generated reports for the Short Form surveys is a “four-box” model. A sample report using this format, designed to be given to the respondent, is present in Appendix A of this guide. Divided into four boxed areas, this one-page report is designed to provide the user with (a) a quick, easily understood visual, numerical, and brief narrative summary of the respondent’s results on the component summary measures; (b) a listing of the component summary measure scores obtained during the current and previous administrations of the survey; (c) brief, bullet-point interpretations of health domain and component summary measure scores relative to general population and appropriate age-by-gender norms; and (d) recommendations for follow-up. Similar reports can be made available to both the respondent and his/her healthcare provider. The primary difference between the “patient report” and the “provider report” is that the later substitutes the follow-up recommendations with medical expenditure, a color-coded
summary of patient change over time, and/or other information that is more useful to the provider.

Samples of some of the Short Form survey reports generated by QualityMetric are presented in the appendixes of this guide. Appendix A contains a sample of a four-box report intended for patient use. Appendixes B through D contain sample provider reports: one that includes a color-coded “dashboard” summary of the patient’s progress (Appendix B); one with the predicted medical expenditure for the patient over the next 6 months (Appendix C); and one displaying only a tabular summary of individual patient Short Form component summary measure, health domain, and predicted medical expenditure results from several survey administrations, for several survey respondents (Appendix D). The last three appendixes present sample reports that summarize aggregated Short Form group data. Appendixes E, F, and G present graphical summaries of average component summary measure, health domain, and MCS-based depression screening results by total sample, age group, and gender, respectively.

Interpreting Short Form Survey Results

Any of several approaches, briefly discussed below, can be taken when developing automated interpretive reports for the Short Form surveys. Regardless of the approach(es) employed, the validity and utility of the report’s interpretive algorithms and text will be based on several assumptions about them and the data to which they will be applied.

Assumptions in Interpreting Short Form Survey Results

When developing or using software for a computer-generated interpretation of findings from the administration of a Short Form survey or any other psychometric instrument, the developer of the interpretations and the user of those interpretations must make several assumptions. The certified developer of the report must assume that:

1. Standardized procedures were used in the administration of the survey.
2. Standardized procedures were used in the scoring of survey raw data (if scored by software/products other than the developer’s own).
3. The survey data being analyzed are determined to be valid when evaluated against current recommended data quality evaluation criteria; if not, this should be addressed in the report.
4. The intended audience for the report is capable of understanding the computer-generated interpretations. This has particular implications when the end user is someone other than a healthcare professional (e.g., a patient).

The user of the interpretive report must assume that:

1. Standardized procedures were used in the scoring of survey raw data.
2. The quality of the data was evaluated against current recommended criteria, and the results of this evaluation are accounted for in the interpretations of the data.
3. Interpretations are based on empirical research and/or the recommendations of the survey author(s); otherwise, appropriate caveats are provided in the body of the report.

It is important for the interpretive report developer to always be mindful of the assumptions that he/she makes, as well as those made by the end user of the report. This awareness will help to achieve the goal of producing the most useful report for the intended audience.

General Considerations

The interpretation of Short Form survey results, or the results of other health status surveys, can be accomplished through any of various approaches, employed either alone or in combination with other approaches. Each approach assumes that patient characteristics impact or are otherwise reflected in the patient’s responses to the survey’s questions. The empirically demonstrated association of these responses, either alone or in aggregate, with what the respondent does or feels outside of
the assessment situation provides the basis for their meaningfulness in describing how he or she currently functions or is likely to function in the future.

One approach, or strategy, to interpreting the results of the Short Form surveys is demonstrated in Figure 2.1. This figure represents Keller and Ware’s (1995) conceptualization of how well-validated health status surveys are sensitive to many circumstances (Causes) that can impact an individual’s functioning and well-being. Based on empirical research, the results yielded by the administration of a Short Form survey can be translated or interpreted on the basis their relationship to: (a) the causes themselves; (b) specific aspects of functioning and well-being that are directly assessed by the instrument (Content of Measure); (c) findings related to normative data and their relationships with other measures of health status (Concurrent Measure); and (d) health-related aspects of the patient’s functioning in the future (Consequences).

The interpretive approach illustrated in Figure 2.1 requires consideration of several variables and empirical data that may not be available for consideration in a computer-generated interpretation. Certified vendors may not want to incorporate an interpretive model of this complexity. Regardless, it is important for vendors to point out ways in which their interpretive software may be limited.

**Norm-Based Interpretation**

The interpretation of Short Form results has been simplified with the use of $T$ scores with the health domain scales and component summary measures, based on U.S. general population sample normative data. Specifically, $T$ scores have proven to be very useful when interpreting differences across the eight health domain scales and for purposes of comparing those domains with the two component summary measures. With $T$ scores, each scale is scored using the same mean (50) and the same standard deviation (10 points) found in the U.S. general population. Thus, each $T$-score point represents one-tenth of a standard deviation. With this method, one can determine the status of the health dimension (physical or mental) or domain represented by the measure or scale, relative to the average, without referring to tables of norms. Also, results for the PCS and MCS measures, which have always been scored using $T$ scores can be compared directly with results for the eight health

---

**Figure 2.1** A Strategy for Interpreting the Results of Health Status Surveys

---

**INTERPRETATION STRATEGIES**

**CAUSES**
- Diagnosis
- Severity of Disease
- Life Stress
- Age
- Other Factors

**MEASURE IN QUESTION**

**CONTENT OF MEASURE**

**CONCURRENT MEASURE**

**CONSEQUENCES**
- Utilization of Services
- Job Loss and Productivity
- Future Health
- Mortality
domain scales because all are standardized in relation to norms from the same population. SF-36v2, SF-12v2, and SF-8 age, gender, gender-by-age, and combined U.S. general population norms for both the standard and acute forms are presented in their respective user’s manuals (see Table 2.1).

Note that the recommendations for interpreting differences in individual respondent scores differ slightly from the recommendations given for interpreting group-level mean scores. These differences reflect the fact that group-level mean scores contain less measurement error than individual, respondent-level scores. One can therefore have greater confidence in the interpretation of group mean scores than in the interpretation of individual respondent scores. Consequently, smaller differences in group mean scores can be meaningfully interpreted. Because individual respondent scores likely contain more measurement error, there is less confidence that the obtained score represents the individual’s true score. Thus, more stringent or conservative guidelines for interpreting individual patient scores are required. This issue is discussed further in Chapter 3 of this guide.

Content-Based Interpretation

Content-based interpretation has proven to be a useful approach to interpreting and explaining Short Form survey results. The Short Form survey user’s manuals provide general strategies for interpreting scores of health domain scale and component summary measure scores based on normative data and the general content of the scale or measure. An interpretive approach based on analyses of the content of individual items is another strategy that can be used to understand the meaning and interpretation of differences in health domain scale and component summary measure scores that fall between the extreme scores (highest and lowest scores). This is accomplished by plotting specific responses to Short Form survey items across score levels of the health domain scales and component summary measures.

Content-based interpretation guidelines were prepared for each SF-36 and SF-36v2 component summary measure and health domain scale, across all score ranges. Development of these guidelines was completed in several steps. First, items that provided useful interpretations across the entire continuum or at particular levels of health domain scale scores were selected. Items selected for content-based interpretations of PCS and MCS scores came from the scales most highly correlated with PCS and MCS, respectively. Second, responses to each of the selected items were dichotomized in a way that is meaningful and that reveals differences across levels of the scale or measure in the score ranges of interests. Third, the percentage of responses to each dichotomous item at each health domain scale and component summary measure score level being interpreted was estimated and plotted.

Item-Level Interpretation

Item-related information can be gleaned from Short Form survey data using the content-based interpretation guidelines presented in the user’s manuals for the SF-36 and SF-36v2. These guidelines, developed primarily for use with group data, are also useful for interpreting data at the individual respondent level and can provide another way to understand the meaning of differences in health domain scale and component summary measure scores at various T-score levels. Unfortunately, this approach provides little specific information about variations in functioning within each domain.

More specific implications of health domain scale scores can be discerned through a direct examination of the response to each of the individual items that these scales comprise. Knowing a respondent’s specific response to a health domain item provides the examiner with the opportunity to understand what areas of functioning are contributing to or account for the observed scale score. This step is particularly useful in determining the types of functional limitations that are present in cases in which a scale score falls within the mid-range of potential health domain scale scores (e.g., T score = 30–45 for the SF-36v2). For example, it may be important to know that a patient with a T score of 38 on the PF scale cannot walk more than a mile or engage in vigorous activities but is able to carry groceries and climb several
flights of stairs. Analysis of item responses is particularly important when a score falls into the “gray” area between impaired and unimpaired functioning (i.e., T score = 40–44 for the SF-36v2). In these cases, examination of responses to individual items within a health domain scale may enable the examiner to determine if the scale score is more indicative of impaired or unimpaired functioning.

**Criterion-Based Interpretation**

Criterion-based interpretation guidelines are based on analyses of relationships between the measures in question and other variables, referred to as criteria, measured either concurrently or after a period of time. The empirical strategy for evaluating the meaningfulness of Short Form survey scores has been to link health domain scale and component summary measure scores to important clinical benchmarks. By this logic, information about importance is gathered by linking differences in scores to important clinical conditions that have a well-understood effect on the domains and components of health measured by the Short Form scales and measures, and by showing how differences in scores of a certain magnitude predict important clinical and social events.

The criteria relied upon for the analyses reported in the Short Form surveys’ user’s manuals were chosen because they are conceptually related to the health domain scales and component summary measures and, in the absence of a “gold standard,” can provide the most useful interpretation guidelines. Specifically, these criteria include variables that (a) are important clinically or socially (e.g., clinical diagnosis, job loss); (b) represent plausible outcomes of the variations in physical, social, and role functioning and in pain, vitality, and mental health; and (c) are measured independently of the health domain scales and component summary measures. Examples of criterion-based interpretation of the SF-36 can be found in Ware, Kosinski, and Gandek (2000), as well as in studies by Kravitz et al. (1992) and Wells et al. (1989). Also, Ware et al. (2007) present data for criterion-based interpretation of the SF-36v2, including data related to mortality, interference with work and social activities, and treatment and/or diagnosis of mental and physical problems.

**Guides to Short Form Survey Interpretation**

In addition to the extensive body of peer-reviewed publications, certified vendors have access to several QualityMetric publications that can assist them in developing sound, automated approaches to interpreting Short Form survey data. The most prominent of these are listed in Table 2.1. These documents provide the information necessary to offer a basic interpretation (e.g., norm-based interpretation) of their respective survey’s results. Many also provide the information that allows one to go beyond high-level or broad interpretations of the data, permitting the vendor to develop the means of offering a more descriptive and detailed picture (e.g., criterion-based interpretation) of the individual(s) whose data are being considered.

**Recommendations for Developing Interpretive Algorithms and Text**

Developing the algorithms and text for an automated interpretation of Short Form survey results is a relatively straightforward task. At the same time, it requires more than a modicum of forethought and attention to detail to ensure that the user of the report receives accurate, clear, and useful information. Developing interpretive statements that are not supported by the literature, do not take into consideration important factors that might mitigate the interpretation for certain patients, are the same for two different health states, are contradictory with other interpretive statements in the report, or incorporate other common mistakes can detract from the actual or perceived utility of the report and hinder its use.

Recommendations that certified vendors may find helpful when planning and developing interpretive reports for the Short Form surveys are listed below. These recommendations are broad and not intended to provide a complete guide to developing these types of reports. However, following them will assist certified vendors avoid costly mistakes that render the final product useless.
<table>
<thead>
<tr>
<th>Survey</th>
<th>Publication Title</th>
<th>Date</th>
<th>Authors</th>
</tr>
</thead>
<tbody>
<tr>
<td>SF-36v2</td>
<td>SF-36v2® Health Survey: A Primer for Healthcare Providers</td>
<td>2008</td>
<td>Ware et al.</td>
</tr>
<tr>
<td>SF-36</td>
<td>SF-36 Health Survey Update</td>
<td>2004</td>
<td>Ware</td>
</tr>
<tr>
<td>SF-36</td>
<td>SF-36 Health Survey Manual and Interpretation Guide</td>
<td>2000</td>
<td>Ware, Kosinski, &amp; Gandek</td>
</tr>
<tr>
<td>SF-36</td>
<td>Measuring and Improving Health Outcomes: An SF-36 Primer for the Medicare Health Outcomes Survey</td>
<td>2004</td>
<td>Ware, Gandek, Sinclair, &amp; Kosinski,</td>
</tr>
<tr>
<td>SF-36</td>
<td>Using the SF-36 Health Survey in Primary Care</td>
<td>2000</td>
<td>Wetzler, Lum, &amp; Bush</td>
</tr>
<tr>
<td>SF-12v2</td>
<td>User’s Manual for the SF-12v2® Health Survey (with a Supplement Documenting the SF-12® Health Survey)</td>
<td>2002</td>
<td>Ware, Kosinski, Turner-Bowker, &amp; Gandek</td>
</tr>
<tr>
<td>SF-12</td>
<td>SF-12: How to Score the SF-12 Physical and Mental Health Summary Scales, Second Edition</td>
<td>1995</td>
<td>Ware, Kosinski, &amp; Keller</td>
</tr>
<tr>
<td>SF-8</td>
<td>How to Score and Interpret Single-Item Health Status Measures: A Manual of Users of the SF-8™ Health Survey</td>
<td>2001</td>
<td>Ware, Kosinski, Dewey, &amp; Gandek</td>
</tr>
</tbody>
</table>
1. Keep it as simple as possible. Establish who the intended audience is, what the purpose of the report will be, what content will help meet that purpose, and what the design of the report will be (see Chapter 1). Do not try to develop a report that will meet all of the needs of all Short Form survey users. The “intended audience” consideration is especially important when a report is being developed for the respondent (e.g., a patient) to review versus for the person who requested that the respondent complete the survey (e.g., a physician).

2. Determine which variables will be used to generate report sections involving interpretation of data. In addition to Short Form survey scores, one might include associated demographic variables, self-reports of diseases, scores from other measures, and/or health data that accompanies the Short Form data.

3. Define and label all variables that will be considered in the report, including variables that must be computed (e.g., differences between scores).

4. Define and determine how to treat missing data (see Maruish & DeRosa, 2009).

5. In developing interpretive algorithms and text, account for the entire range of scores and/or other variables that will be considered for the generation of text for each interpretive section of the report. For example, if the interpretation of the SF-36v2 VT score is to take age into consideration, be sure that all possible combinations of VT \( T \) scores between 20.87–70.82 (i.e., the possible score range for the VT scale; see Ware et al., 2007) and ages 18 years and older (i.e., the appropriate age range for the SF-36v2) have associated algorithms and text.

6. Avoid any overlap of variable scores or score ranges in developing algorithms for interpretations. For example, if different interpretive text is developed for interpretive score ranges for SF-36v2 VT \( T \) scores above and below 45, algorithms written for VT \( \leq 45 \) and VT \( \geq 45 \) would overlap, such that a VT \( T \) score of 45 would meet both the conditions. In this case, one must choose either “VT \( < 45 \) and VT \( \geq 45 \)” or “VT \( \leq 45 \) and VT \( > 45 \).” Overlap considerations are particularly important when two or more variables are considered in the algorithm.

7. Do not provide interpretations that are not supported by data nor recommended by the survey’s authors.

8. Have the final set of interpretive text edited by a copyeditor, or at least reviewed by someone other than the person(s) involved in developing the algorithms and/or text.

9. Create a functional specification document that contains all the information necessary to develop the software to generate the report. This document will not only guide the programming and quality assurance of the report, it will also provide a means of easily reviewing the program if problems arise or changes to the report are being considered. This specification document should be updated whenever changes are made to content, programming logic, interpretive algorithms, or text, so that the document remains an effective tool for documenting and communicating information about the software.

**Using Short Form Results to Trigger Immediate Intervention**

Some organizations find it useful, if not imperative, to be instantly alerted about any individual who responds to a Short Form survey and/or other administered measure in a manner that suggests the presence of a significant problem requiring immediate attention. A hypothetical example is a health plan that requires new members to complete the online SF-36v2 at the time of enrollment. The purpose of completing the survey would be not only to establish a baseline measure of health status that can be compared to results from subsequent yearly assessments for population monitoring purposes, but also to identify individuals who may be at risk for depression. The plan’s claims data have shown that plan members
with depression increase the plan’s costs, so initiating treatment as soon as possible can help reduce these costs for members identified as being depressed. To this end, this hypothetical health plan’s IT department has programmed its online assessment system to automatically notify the plan’s care management call center whenever a member receives an SF-36v2 Mental Health Component (MCS) score that is predictive of the presence or risk of significant depression. At that point, the health plan’s staff would contact the member to further assess the risk and, as appropriate, make a recommendation and arrange for the member to begin treatment.

The preceding example demonstrates how data obtained from an online assessment can be used by organizations to have an immediate impact on a survey respondent and, in some cases, on its bottom line. More specifically, this example demonstrates using Short Form survey results to screen individuals for a health-related condition that an organization wants or needs to know about immediately. Be aware that care must be taken when using any of the Short Form surveys for this purpose. As with other tests and measures, many factors should be taken into consideration when developing algorithms to screen for the presence of any condition, such as:

- How common the condition is in the population being assessed (i.e., base rate)
- The appropriateness of a particular health domain scale or component summary measure for determining the presence of a specific condition
- Empirical evidence of the validity of the selected scale or measure for the intended screening purpose
- How much the organization is willing to follow-up with all individuals who screen positive for the condition but do not actually have the it (i.e., false positives), or to ignore individuals who screen negative for the condition but actually do have it (i.e., false negatives). Addressing this will help determine what score cutoff(s) should be used with the variable(s) that will be used to classify the condition as being present or not.

Other considerations may need to be taken into account, depending on the organization and its goals. Moreover, similar factors should be considered when what is being screened for is the likelihood of some adverse future event (e.g., hospitalization or inability to work occurring within the next 6 months).

Because of the potential complexity and difficulty in developing and incorporating a valid and useful screening mechanism into a health assessment system that meets specific organizational requirements, it is recommended that certified vendors and organizations contact QualityMetric’s Outcomes Insight Consulting™ Division to assist in this task. The Outcomes Insight Consulting staff’s knowledge of the Short Form instruments and experience in designing assessment systems for organizations of all sizes can help ensure that the end product is the most useful Short Form-based means of assessment for the organization’s needs and is successfully integrated into the organization’s work flow.
Short Form survey results can serve as a rich source of information for understanding the health status of individual survey respondents when different approaches to examining the data are taken. This includes information based on published empirical research on the survey, as well as information from a more individualized, qualitative examination of the survey data. How much and to what depth a computer-generated report of the findings goes into will depend on decisions made by the developer with regard to the intended audience, purpose, content, and design of the report. Regardless, a report developed to present the results of an administration of a Short Form survey to a single respondent or group of respondents can be as simple or complex, and as brief or lengthy, as one would like.

The purpose of this chapter is to present considerations for the development of an automated report for results obtained from the administration of a Short Form survey to a single respondent. Following some general considerations, approaches to data quality evaluation and data reporting for individual respondent results are discussed. This is followed by a discussion of each of the basic approaches to interpretation—norm-based, content-based, item-level, and criterion-based.

Next, interpretation of Short Form results in light of results from other psychometric perspectives is discussed. For example, the application of measure- or scale-specific standard errors of measurement (SEMs) allows the user to determine, within specified levels of confidence, intervals in which the respondent’s true score falls on each measure and scale. Moreover, the availability of gender- and/or age-based norms and norms for 18 disease groups provide additional means for better understanding the meaning of the obtained scores. Other useful measures of health status that can be derived from the Short Form survey scores—the SF-6D, medical expenditure prediction, and level of depression—are also described. Finally, means of monitoring and evaluating treatment outcomes is discussed. Chapter 4 addresses similar considerations for group-level data.

**General Considerations**

Before developing a report, certified vendors should carefully review their decisions about the intended audience, purpose, content, and design of the report to ensure that the type of product they have decided to develop is:

1. **Realistic**, from the standpoint of what can actually be included in the report or, more generally, what the report can provide to those who will be using it.
2. **Doable**, from the standpoint of the expertise, time, and money that are available to the vendor for report development.
3. ** Marketable**, from the standpoint of meeting the needs of potential customers at a price that they are willing to pay and that will result in the desired return on investment for the development and marketing of the report.

A major consideration in this regard is the planned content for the report. One cannot derive the same type of information from all of the Short Form surveys; taking the time to match the reporting and interpretation needs of...
the user to the appropriate Short Form survey is important. Table 3.1 presents a summary of the types report sections that are appropriate for each of the surveys. Generally, the ability to provide useful information in each of these sections reflects the content of the survey, the scales and other measures that can be scored from the survey data, and the research that is available to support a valid interpretation of results as they relate to the purpose of the section. The SF-36/ SF-36v2 reflects one end of the continuum while the SF-10 reflects the other end.

### Disclaimer/Cautionary Statement

By itself, a computer-generated interpretation of the results obtained from the administration of a psychometric instrument such as a Short Form survey cannot and should not be viewed as providing definitive, final-word answers to questions about the respondent. At best, it can provide a probable yet fairly accurate picture of a respondent based on empirical findings. The information that can be obtained from a computer-generated report will necessarily be limited by the type, amount, and quality of data the instrument elicits; the quality and amount of empirical data that supports the programmed algorithms and interpretations; and the report developer’s knowledge and expertise in the use of the instrument. Other factors (e.g., the appropriateness of the instrument for a particular respondent) may also come into play.

Like computer-generated reports for other instruments, Short Form survey reports should be used as just one of several sources of information about the respondent. In clinical settings, other types of information, such as information obtained from a clinical interview, medical history, lab tests, other psychometric instruments, and interviews with collaterals, need to be considered along with the survey findings. In short, Short Form survey reports should be viewed by the user as a means of generating hypotheses about the respondent (or, in the case of the SF-10, about the child) or as a means of supporting hypotheses generated on the basis of other information known

### Table 3.1

<table>
<thead>
<tr>
<th>Report Section</th>
<th>SF-36v2</th>
<th>SF-36</th>
<th>SF-12v2</th>
<th>SF-12</th>
<th>SF-8</th>
<th>SF-10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Identifying information</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>DQE – Missing responses</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>DQE – Response consistency</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DQE – Response pattern</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Profile</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Score summary</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Item response summary</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Norm-based interpretation</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Content-based interpretation</td>
<td>X</td>
<td>X</td>
<td>X*</td>
<td>X**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Item-level interpretation</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Criterion-based interpretation</td>
<td>X</td>
<td>X</td>
<td>X*</td>
<td>X**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Depression screening</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Medical expenditure prediction</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SF-6D</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age-by-gender norm comparison</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Disease-specific norm comparison</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X**</td>
<td></td>
</tr>
<tr>
<td>Monitoring change</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
</tbody>
</table>

* Available in late 2009.
** Available in early 2010.
about the respondent. For these reasons, it is important for each report to include a statement that ensures that the report user is informed of the limitations of the report and how it should be used.

Cautionary statements that appear on computer-generated reports vary according to several factors, including the specific type of instrument that was administered, how the developer intended the report to be used, the level of empirical support behind the interpretive statements, and the setting(s) in which it is likely to be used. The following example of a boilerplate cautionary statement appears on one of QualityMetric’s SF-12v2 reports developed for the respondent’s use:

Note: This survey is not a diagnostic tool. It is intended to supplement, but not replace or contradict, the advice of your personal physician. If you have any questions or concerns about your health, it is always good to seek one-to-one professional medical consultation.

The cautionary statement for the SF-12v2 report developed for the respondent’s care provider reads as follows:

Note: This report reflects information provided through patient self-report. It is not intended to replace clinical judgment or treatment.

Although cautionary statements may appear as a footer at the bottom of QualityMetric’s reports, they can also be conveyed in a separate section within the body of the report (usually at the beginning or end). Regardless, certified vendors should seek legal review of any proposed disclaimer or cautionary statement to ensure that it accurately conveys how the report should be used and protects the vendor from legal liability resulting from misuse of the report by the end user.

**Identifying Information**

Each report should begin with a section identifying the respondent (e.g., name, date of birth, patient/subject/ID number) and any important demographic (e.g., age, sex), health (e.g., disease indicated on a disease checklist), occupational (e.g., employment status, professional status), and/or other type of data (e.g., education level) that may be considered in any of the report’s interpretive algorithms. Presentation of this data can either be in tabular form or in narrative text. Date of survey completion and date of report generation should also be included here or elsewhere in the report (e.g., header/footer).

**Data Quality Evaluation**

Unlike the case with group-level data, there are only a few ways in which to evaluate the quality of Short Form survey data from a single respondent, most of which are qualitative in nature. However, running these data quality checks as part of data processing can result in information that lends support to the validity of the findings or their interpretations. For this reason, data quality information, if included, should be presented before other sections that report and/or interpret the respondent’s Short Form survey data.

**Missing Responses**

With the exception of the SF-10 and under certain other circumstances, maximum data recovery methods can be applied to datasets with missing data to yield scores on the health domain scales and component summary measures (see Saris-Baglama, Dewey, et al., 2007). In addition, the half-scale missing score estimation method can be applied to SF-36v2 and SF-36 datasets. Regardless of the ability to apply missing score estimation methods to survey datasets, a determination of the total number of missing or otherwise invalid responses can provide useful information about the results. Saris-Baglama, Dewey, et al. indicate that at least 90% of the responses in a group-level dataset should be valid. This criterion might also serve as an appropriate criterion for individual respondent datasets.

**Response Consistency**

Although developed for use with SF-36v2 and SF-36 group-level data, the Response Consistency Index (RCI; Ware et al., 2007) can also serve as a measure of response consistency for
an individual completing either version of the 36-item Short Form survey. Ware et al. suggest that a score of two or greater on the RCI be considered problematic for individual respondents.

**Response Pattern**

Another means of determining the quality of an individual’s responses to a Short Form survey is through an examination of the respondent’s response set. Generally, one should be suspicious of results generated from a response set in which:

1. The first response, last response, or another response (e.g., the third of five choices) was selected for all items;
2. The response indicating the worst level or the best level of functioning was always indicated; or
3. The response pattern is sequential from one item to the next (e.g., for the SF-36v2 PF scale items, the respondent answered the 10 items 1,2,3,1,2,3,1,2,3,1; or, throughout the entire survey, the sequence of responses is 1,2,1,2,1,2,1,2,1,2...).

Although possible, it is highly unlikely that these and other types of patterned responding reflect honest responses to survey items. When such patterns appear, the overall validity of the Short Form survey results should be questioned. At a minimum, the report user should be encouraged to ask the respondent to explain his or her item responses, thereby verifying or discrediting their accuracy.

When the user is confident that the Short Form survey results satisfy data quality standards, interpretation of those results can begin.

**Score Summary**

Presenting a brief summary of the Short Form survey results provides the report user with a means of quickly determining the important aspects of the respondent’s health status. This can be accomplished by using a simple tabular summary. Table 3.2 indicates the types of scores and other data that can be presented for each of the Short Form surveys in a tabular format. Note that interpretive or other narrative report sections do not have to accompany each of the scores or data presented in this summary.

As part of a tabular summary, one may also include additional psychometric data that can enhance the user’s understanding of the data. For example, presenting the relevant age-by-gender and disease-specific norms (i.e., means and standard deviations) for the Short Form survey’s scales and measures in addition to the obtained scores would enable the user of the report to better understand the meaning of these scores without having to refer to a survey’s user’s manual. Similarly, printing the SEM values for

<table>
<thead>
<tr>
<th>Quantitative Variable</th>
<th>SF-36v2</th>
<th>SF-36</th>
<th>SF-12v2</th>
<th>SF-12</th>
<th>SF-8</th>
<th>SF-10</th>
</tr>
</thead>
<tbody>
<tr>
<td>DQE – Missing responses</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>DQE – RCI score</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Component summary measure scores</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Health domain scale scores</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Depression equivalent score</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Medical expenditure prediction</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SF-6D scores</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age-by-gender norms</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Disease-specific norms</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td>X*</td>
</tr>
<tr>
<td>Change scores</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td>X</td>
</tr>
</tbody>
</table>

*Available in 2010.
PCS, MCS, and the health domain scales would enable the user to determine a confidence interval for each of scores (discussed in more detail later in this chapter).

**Profile of Scores**

Another means of providing a summary of component summary measure and health domain scale scores is through a graphic profile. The standard Short Form profile of observed scores for an individual respondent provides a visual overview of health status that is quickly and easily understood. The score profile first presents the PCS and MCS scores; placement of these measures at the beginning (left side) of the profile emphasizes the importance of first considering results with regard to overall functioning in the physical and mental health dimensions (see Figure 3.1). As a result, upon visual examination of the profile, one can quickly determine whether differences in health (if any) are more or less apparent in general physical or mental health status. PCS and MCS are followed on the profile by the health domain scales, arranged in order of those that are correlated highest with the PCS measure (PF, RP, BP, and GH) to those that are correlated highest with the MCS measure (VT, SF, RE, and MH), thus allowing for easy discernment of which domain(s) contributed to high or low scores on the component summary measure.

Data from each of several administrations of the same Short Form survey can be profiled simultaneously using a single, standard Short Form profile format. As shown in Figure 3.2, the component summary measure scores and health domain scale scores from each administration can be plotted using printed bars of differing patterns and/or shades of color. Note that Figure 3.2 also includes the scores represented by the bars, thus facilitating the user’s understanding of what is being represented in the profile. This same feature (i.e., including scores represented by the bars) can be included with profiles reporting the results of only single administrations of a Short Form survey.

Note that the standard Short Form profile is not appropriate for presenting SF-10 findings because the survey does not include health domain scales and its component summary measures are different from those of the other five Short Form surveys. One may, however,

---

**Figure 3.1**  Sample Short Form Survey Profile of Scores
Summary of Item Responses

As discussed in the previous section, a profile of Short Form survey scores provides a good overview of an individual’s health status on the assessed domains and components, relative to the survey’s normative sample, and the scores plotted on this profile will serve as the basis of a norm-based interpretation of the results (see below). A more detailed picture of the respondent can come from understanding what contributes to these scores, which requires an examination of his or her responses to the survey’s individual items. Providing a summary of each item’s scale placement, content, and entered response can greatly facilitate this task for the user. An example of how this can be done in a tabular format is presented in Figure 3.3. Information contained in this type of table can serve as the basis for a content-based interpretation of an individual respondent’s survey results (see below).

Norm-Based Interpretation

The general approach for norm-based interpretation employs a systematic examination of the Short Form profile, first from a broad perspective and then by conducting a more detailed analysis of the data. A simple yet informative approach involves determining if the T scores for the PCS and MCS measures deviate from what is considered the average range for the U.S. general population. If desired, this section could include interpretation of both measures, regardless of their scores.

To obtain a clearer picture of the respondent’s health status, a similar examination of the T scores for the health domain scales is also recommended. Note that the ordering of the health domain scales facilitates interpretation of the profile, with differences in the four scales on the left side of health domain profile
(PF, RP, BP, and GH) generally reflecting physical health status and differences in the four scales on the right side of the profile (VT, SF, RE, and MH) generally reflecting mental health status. In most instances, the four health domain scales at the left of the profile will correspond to what is observed on the PCS measure, whereas the four scales at the right of the profile will correspond to what is observed with the MCS measure.

Note that in the user’s manual for the SF-36v2 (Ware et al., 2007), the recommendations for interpreting differences in individual patient scores differ slightly from the recommendations given for interpreting group-level mean scores. These differences reflect the fact that group-level mean scores contain less measurement error than individual respondent-level scores. One can therefore have greater confidence in the interpretation of group mean scores than in the interpretation of individual patient scores. Consequently, smaller differences in group mean scores can be meaningfully interpreted. Because individual respondent scores likely contain more measurement error, there is less confidence that the obtained score represents the individual’s true score. Thus, more stringent or conservative guidelines for interpreting individual patient scores for the SF-36v2 are required. In general, these same recommendations apply to the interpretation of SF-12v2 and SF-8 results. At the same time, because of these surveys’ limited precision at the individual patient level, interpretation of the SF-12v2 and SF-8 health domain scale results should be focused on the extent to which they contribute to the obtained component summary measure scores.

Overall, examination of component summary measure and health domain scale scores can be used to determine if the respondent deviates from the norm in a manner that is indicative of increased health risk. This examination also provides the context in which the content-based and criterion-based approaches to interpretation of results (described below) should take place.

**Using the Standard Error of Measurement (SEM) and Confidence Intervals (CIs)**

When evaluating Short Form component summary measure and health domain scale scores, it is important to be mindful of the error

<table>
<thead>
<tr>
<th>Item No.</th>
<th>Scale</th>
<th>Abbreviated Item Content</th>
<th>Response</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>GH</td>
<td>Self-rating of health</td>
<td>Excellent</td>
</tr>
<tr>
<td>2</td>
<td>SET</td>
<td>Health compared to 1 year ago</td>
<td>Much better</td>
</tr>
<tr>
<td>3a</td>
<td>PF</td>
<td>Limited in vigorous activities</td>
<td>Missing</td>
</tr>
<tr>
<td>3b</td>
<td>PF</td>
<td>Limited in moderate activities</td>
<td>Limited a lot</td>
</tr>
<tr>
<td>3c</td>
<td>PF</td>
<td>Limited in carrying groceries</td>
<td>Limited a little</td>
</tr>
<tr>
<td>3d</td>
<td>PF</td>
<td>Limited in climbing several flights of stairs</td>
<td>Limited a little</td>
</tr>
<tr>
<td>3e</td>
<td>PF</td>
<td>Limited in climbing one flight of stairs</td>
<td>Not limited</td>
</tr>
<tr>
<td>3f</td>
<td>PF</td>
<td>Limited in bending/kneeling/stooping</td>
<td>Limited a little</td>
</tr>
<tr>
<td>3g</td>
<td>PF</td>
<td>Limited in walking more than a mile</td>
<td>Limited a lot</td>
</tr>
<tr>
<td>3h</td>
<td>PF</td>
<td>Limited in walking several hundred yards</td>
<td>Limited a lot</td>
</tr>
<tr>
<td>3i</td>
<td>PF</td>
<td>Limited in walking one hundred yards</td>
<td>Limited a little</td>
</tr>
<tr>
<td>3j</td>
<td>PF</td>
<td>Limited in bathing/dressing self</td>
<td>Not limited</td>
</tr>
<tr>
<td>4a</td>
<td>RP</td>
<td>Cut down time spent on work/activities due to physical health</td>
<td>Some of the time</td>
</tr>
<tr>
<td>4b</td>
<td>RP</td>
<td>Accomplished less due to physical health</td>
<td>Some of the time</td>
</tr>
<tr>
<td>4c</td>
<td>RP</td>
<td>Limited in kind of work/activities due to physical health</td>
<td>All of the time</td>
</tr>
<tr>
<td>4d</td>
<td>RP</td>
<td>Difficulty performing work/activities due to physical health</td>
<td>Most of the time</td>
</tr>
<tr>
<td>5a</td>
<td>RE</td>
<td>Cut down time spent on work/activities due to emotional problems</td>
<td>A little of the time</td>
</tr>
<tr>
<td>5b</td>
<td>RE</td>
<td>Accomplished less due to emotional problems</td>
<td>None of the time</td>
</tr>
<tr>
<td>5c</td>
<td>RE</td>
<td>Did work/activities less carefully due to emotional problems</td>
<td>None of the time</td>
</tr>
</tbody>
</table>
that is contained in each and, as necessary, use said errors to temper the interpretation of the obtained results. Confidence intervals (CIs) provide valuable information about the amount of fluctuation that can be expected in a single score due to measurement error. A CI around a measure or scale obtained score is a function of the standard error of measurement (SEM) for that measure or scale. A scale or measure with a relatively small standard deviation (SD) and high reliability has a small SEM and, consequently, small CIs around observed scores on that scale or measure. With smaller CIs, fluctuations in an individual patient’s scores due to chance are less likely, facilitating their use in monitoring individual patients in clinical practice.

Different CIs can be constructed for obtained Short Form scores, depending on the desired level of confidence one wishes to have about where the respondent falls on a scale or measure. CIs are established by determining the appropriate score value for the desired confidence level, and then by first adding and then subtracting said score value to and from the obtained score to establish the CI’s upper and lower limits, respectively. The appropriate score value for a determining a CI is a multiple of that scale or measure’s SEM, which will vary according to the desired CI level. For example, 1 times a scale’s SEM produces the value corresponding to a 68% CI, whereas 1.96 times a scale’s SEM produces the value to corresponding to a 95% CI (see Ware et al., 2007). When making decisions about reporting CIs and/or considering them in the interpretation of Short Form scores, vendors must also decide whether to compute CIs at only one level of confidence or to allow the user of the report to select the level of confidence (e.g., 68%, 80%, 95%) on a respondent-by-respondent basis. The latter option may make development of interpretive algorithms more complex but can provide an attractive, useful feature for the report user.

If CIs are used, certified vendors are encouraged to use the SEMs printed in the appropriate Short Form survey’s user’s manual to compute the CIs. The CIs found in the manual are based on the general population and reflect a more normal variation in health status scores than specific samples that utilize inclusion and exclusion criteria to define the sample under investigation. For example, inclusion and exclusion criteria in clinical trials produce homogeneous samples whose SF-36v2 scores vary less than general population scores. Consequently, SEM estimates will tend to be much smaller for these samples than the estimates observed in the general population. If the SEM can be proven to be substantially different in particular samples, some (e.g., Nunnally & Bernstein, 1994) would recommend that the CIs should be re-estimated using published formulas.

Comparison to Age-by-Gender Norms

Age and/or gender normative data are available for several of the Short Form surveys and can be found in each survey’s user’s manual. The SF-36v2 age-by-gender norms are useful in determining whether the respondent’s T scores for the component summary measures and health domain scales are above or below the average for same-sex individuals in his/her age range. Comparison of the respondent’s scores to a sample of individuals of the same age and gender can provide a clearer picture of the individual’s health status and can improve the accuracy and usefulness of any norm-based interpretation. For example, although a T score of 39 on the SF-36v2 PF scale is indicative of significant impairment in physical functioning in the U.S. general population, it is the mean PF score for men aged 75 years or older. Thus, consideration of these data might temper the more general interpretation of impairment for this score for men aged 75 years and older.

Comparison to Disease-Specific Norms

Like the age and gender norms, the availability of norms for 18 disease groups for both versions of the SF-36 and SF-12 surveys provides additional means for better understanding the meaning of the observed scores. Assuming the certified vendor has included in the software the ability to identify respondents with one or more of these specific diseases, comparison of the T scores of a respondent suffering from a specific disease or condition
to the normative data for a group of individuals suffering from the same disease or condition can provide an indication of how similar he or she is to the prototypical or average member of that disease group, in terms of health status and extent of limitations. For example, although a $T$ score of 41 on the SF-36v2 GH scale is indicative of borderline to impaired functioning in the U.S. general population, it is the mean GH score for individuals diagnosed with depression. Thus, the interpretation of a GH score of 41 may be different for those identified as having depression than for those not so identified.

**Content-Based Interpretation**

Content-based interpretation, as described in Chapter 2, is particularly useful when examining group-level data. However, it is limited by the items examined, the way in which responses were dichotomized, and the score ranges that were used in the analyses. When interpreting individual respondent data, direct examination of each of the respondent’s individual item responses (see Figure 3.3) is helpful in clarifying the specific aspects of his or her overall or domain-specific physical and/or mental health functioning. For certain SF-36v2 or SF-36 items, this approach can be supplemented by a comparison of a respondent’s scores with the data from the content-based interpretation tables published in the each survey’s user’s manual in order to determine the percentage of the survey’s normative sample that responded the same way as the respondent. For example, for a respondent with an SF-36v2 PF $T$ score of 48 who indicated that he or she is limited a little in walking 100 yards, it is useful to know that only 5.5% of those in the U.S. general population with a $T$ score in the 45–49.9 range on the SF-36v2 PF scale reported being limited in walking 100 yards.

**Criterion-Based Interpretation**

Another approach to understanding health status is the interpretation of a respondent’s scores based on demonstrated relationships between the Short Form scores and external criteria. Tables in the user’s manuals for the SF-36v2 and SF-36 provide information that can be used to generate criterion-based interpretative statements based on the results from these two Short Form surveys. For example, in the case of a respondent with an SF-36v2 PCS score of 38, report users may find it useful to know that 6 months after completing the Short Form survey as part of the Medical Outcomes Study (MOS; see Stewart & Ware, 1992), 41.3% of respondents who obtained a PCS score in the 35–39 $T$-score range reported that they could not work because of their health. In addition, during this same 6-month period, 16% of those with PCS scores in this same range were hospitalized overnight and 19.4% lost their job because of their health. Similar types of criterion-based interpretations can be made for the SF-36 and SF-36v2 MCS score and the health domain scale scores.

**Depression Screening**

Data from the SF-36v2, SF-36, SF-12v2, SF-12, and the new SF-12v2®–MH Enhanced™ (the standard SF-12v2 plus the additional 3 items from the SF-36v2 MH scale) can be used to screen survey respondents for depression. Given that depression screening is currently employed in many settings offering treatment to patients with either physical or emotional problems, providing a section that directly addresses the possible presence of depression and the need for further inquiry adds a useful feature to the report.

**MCS Score**

The MCS measure that can be scored from all versions of both the SF-36 and SF-12 has proven to be valid and useful for screening patients for mental health problems. Ware and Kosinski (2001) recommend a $T$ score of 42 or below on the MCS measure for identifying patients who are clinically depressed. This score was found to be the optimal cutoff for identifying clinical depression among the 3,445 MOS patients, 508 (14.6%) of whom
were diagnosed with dysthymia and/or major depression (see Burnam, Wells, Leake, & Landsverk, 1988; Wells, et al., 1989). This cut-off resulted in a sensitivity of 73.7% and a specificity of 80.6%.

MH Score

The 5-item MH scale from the SF-36v2, SF-36, and SF-12v2–MH Enhanced forms has been shown to be an effective screener for major affective disorders, such as depression (Berwick et al., 1991; Ware, Snow, Kosinski, & Gandek, 1993; Ware & Kosinski, 2001; Weinstein, Berwick, Goldman, Murphy, & Barsky, 1989). The recommended cutoff for detecting depression is a T score of 37 or below. In addition, the MH scale has been cross-calibrated with the Beck Depression Inventory® (BDI®, Beck & Steer, 1987; Beck, Ward, Mendelson, Mock, & Erbaugh, 1961) to enhance the understanding and interpretation of MH scores from the SF-36v2 and SF-12v2. As a result, the MH score can be linked to BDI units and placed into one of four mental health severity stages.

SF-6D Health Utility Index

The SF-6D is an index, which can be calculated using either version of the SF-36 or SF-12, designed for use in economic evaluation studies. Using data from the administration of either version of the SF-36, individual respondents can be classified on any of four to six levels of functioning or limitations and on each of six domains (with RP and RE considered a single dimension), thus allowing a respondent to be classified into any of 7,500 possible unique health states (Brazier & Roberts, 2004). While Brazier et al. (2002) were able to obtain utility values on 249 of the possible health states of 836 respondents for the SF-36, simplification of the PF and BP domains reduced the number of possible health states to 241 for the SF-12. Normative data for the SF-6D are included in the SF-36v2 and SF-12v2 general population, gender-, sex-, age-, and diseasespecific norms tables.

Note that utility weights for the SF-6D index have been developed for Great Britain. For a discussion of the advantages and disadvantages of developing country-specific weights, see Brazier et al. (2002), Brazier and Roberts (2004), and Walters and Brazier (2003). Also, see Brazier, Rowen, and Hanmer (2008) for a summary of recent improvement to the scoring of the SF-6D.

Monitoring Change

Interpretation of scores and score differences has become a focus area in the field of health status assessment. Researchers and clinicians have used the concept of minimally important difference, or MID, to differentiate between an important score difference and a trivial, unimportant difference. A significant motivation behind the MID concept is the proper design of clinical trials. Given a sufficiently large sample size, even trivial differences can show statistical significance. Therefore, the central question is whether a particular difference is clinically significant. From this question came the concept of minimal clinically important difference, shortened to minimally important difference in the context of patient-reported outcomes (PRO) to emphasize the perspective of the patient (rather than the clinician) and the importance of evidence of many types, including but not limited to clinical evidence. Thus, from the perspective of clinical trials, MID refers to mean group differences, as explicitly pointed out in the Food and Drug Administration
(FDA) draft guidance for the industry regarding PRO measures (U.S. Department of Health and Human Services, 2006).

It is well recognized for other endpoints that the MID for a comparison of group means is different from the MID for individual patient scores (Guyatt, Osoba, Wu, Wyrwich, & Norman, 2002). However, in much of the literature on PRO, this distinction is not made explicit, leading to some confusion regarding recommendations. Following suggestions from the FDA, this guide uses the concept of MID in relation to mean group differences and the concept of a responder in relation to an individual whose scores have shown important change over time. An MID is generally smaller than a responder criterion for the same scale, partly because the magnitude of group mean differences are not affected by measurement precision and partly because it is not reasonable to assume that everyone in a clinical trial will benefit from treatment.

Ware et al. (2007) recommend a different responder criterion for each of the SF-36v2 health domain scales and component summary measures based on an 80% confidence interval (i.e., obtained score ± 1.28 SEM for each scale or measure) and assuming a correlation of 0.4 from baseline to follow-up. For those wanting to take a more conservative approach to the evaluation of change, values for determining 90% and 95% confidence intervals could be computed using different multiples of the SEMs. The approach to establishing the responder criteria for the SF-36v2 can also be taken with other Short Form instruments. However, because most of these surveys have significantly less items, it is best to use only data from the component summary measures when making statements about change from one administration to another.

**Combining Results from Short Form Surveys and Other Instruments**

As with other psychometric instruments, Short Form survey results can be more clearly understood when interpreted within the context of other information known about an individual or a group of individuals. Common sources of additional information include lab tests, face-to-face interviews, chart reviews, and other self-report instruments, such as disease-specific measures of HRQOL. Data from all of these sources may provide insight into the nature and extent of any health status problems revealed by the Short Form component summary measures and health domain scales. When used primarily for research, data from other sources can serve several purposes, such as cross-validating the Short Form survey findings, identifying important co-variates, and determining the generalizability of the results. For individual patients in clinical settings, use of multiple assessment instruments can assist in arriving at more accurate and comprehensive diagnoses and lead to the development of more effective treatment plans. In most settings, data from many of these sources can be obtained at the time the individual completes the Short Form survey.

When evaluating an individual with the Short Form survey and another psychometric instrument that purportedly measures one or more of the same constructs, one could obtain results that appear contradictory. This can occur for any of several reasons, including (a) one instrument was developed to broadly sample a given domain, whereas the other was developed to provide a comprehensive or focused assessment of a specific aspect of that same domain; (b) the normative sample for one instrument differs from that of the other instrument; (c) one instrument is more appropriate than the other for assessing the person or population in question; (d) one instrument is more valid than the other for the purpose for which it is being used; or (e) errors occurred in scoring one or more of the instruments that were administered. Assuming that both instruments are valid and were administered at the same time and scored correctly, the occurrence of findings that appear contradictory should lead the examiner to determine if the difference in findings can be attributed to the content, norms, or intended purpose of the two instruments. Resolution of any contradictory findings obtained from the Short Form survey
and another instrument (or from any two instruments) often leads to a better understanding of the person being assessed.

The development of algorithms and narrative text to address the possible nature of contradictory findings can be a complex task requiring the expertise of an individual who is knowledgeable about both instruments. Thus, it is generally better to avoid trying to explain contradictions; instead, focus on identifying apparent contradictions and recommending their resolution by the report user.
Developing certified software for reporting and interpreting group-level, or aggregate, data for a Short Form survey presents a different set of challenges and considerations than those that occur when developing individual respondent reports. Like the individual respondent report, a report developed to present the results of an administration of a Short Form survey to a group of respondents can be as simple or complex, or as brief or lengthy, as one would like. However, the requirements for group-level reports often include going beyond the simple reporting of the scored results to also including advanced statistical analyses of those same results. At the minimum, the certified vendor must enlist the services of a statistical expert who is knowledgeable about the Short Form survey in question and who can correctly determine the most appropriate analyses that should be performed. Moreover, organizations needing group-level reports tend to have different data analytic and reporting needs at different points in time, requiring the certified vendor to develop a computer-generated report so specific that it is useful to only a relatively small number of organizations or so general that it is of limited utility to any organization. Thus, a computer-generated report may not be the best solution to an organization’s needs in this situation, especially if an updated report is not needed on a frequent enough basis to justify development costs. Instead, use of the services of a consulting group with expertise in the Short Form surveys and analyses of group-level results would be the more effective and efficient option.

The purpose of this chapter is to briefly discuss conditions that make it advantageous for certified vendors of Short Form products to develop software intended for the reporting and interpretation of group-level data. A tabular summary of the types of content that can be included in reports for each of the Short Form surveys is presented in this chapter, followed by a discussion of the analyses and report content that QualityMetric’s Outcomes Insight Consulting Division can provide to organizations whose needs and required level of specific analytic expertise differ from those which are common across various organizations and/or are required on a more frequent basis.

General Considerations

As when developing reports for individual respondents, certified vendors and other licensees should carefully review their decisions about the intended audience, purpose, content, and design of the report to ensure that the type of group-level Short Form survey report they intend to develop is realistic, doable, and marketable (see Chapter 3). When the need for a fairly standard (i.e., unchanging in content or required statistical analyses) Short Form survey group-level report occurs on a frequent basis (e.g., monthly), the investment of time and money for the development of software for group-level analyses and reporting may be well worthwhile. However, there are likely many instances in which the analyses of group data is required on a less frequent basis (e.g., annually) or is subject to frequent modification, making it an unwise decision to develop software for standard analyses and reporting.

Using information from documentation available from QualityMetric (e.g., manuals,
user’s guides) and other sources (e.g., peer-reviewed publications), certified vendors of Short Form reporting software can develop a standard group-level interpretive report that can provide a substantial amount of information about the group of individuals being evaluated. Table 4.1 presents a summary of the types of report sections that would be appropriate to include in group-level reports for each of the Short Form surveys. However, it is important to note that such reports are likely to be limited in the amount and type of information that can be derived from the group’s dataset. With its knowledge of the Short Form surveys, proprietary Short Form databases, and scientific and statistical expertise, QualityMetric’s Outcomes Insight Consulting Division can develop group-level reports that maximize the use of the data to meet the organization’s specific needs for point-in-time information.

QualityMetric’s Outcomes Insight Consulting™ Services

QualityMetric’s Outcomes Insight Consulting Division collaborates with clinical, research, product development, and marketing professionals to provide survey development, outcomes, Table 4.1

<table>
<thead>
<tr>
<th>Report Section</th>
<th>SF-36v2</th>
<th>SF-36</th>
<th>SF-12v2</th>
<th>SF-12</th>
<th>SF-8</th>
<th>SF-10</th>
</tr>
</thead>
<tbody>
<tr>
<td>DQE – Completeness of data</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>DQE – Responses within range</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>DQE – Estimable component scores</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>X</td>
<td>a</td>
<td></td>
</tr>
<tr>
<td>DQE – Convergent validity</td>
<td>b</td>
<td>b</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DQE – Discriminant validity</td>
<td>c</td>
<td>c</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>DQE – Consistent responses</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>DQE – Estimable scale scores</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DQE – Item internal consistency</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>DQE – Item discriminant validity</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>DQE – Reliable scales</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Profile – Total sample</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Profile – Total sample by age</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Profile – Total sample by gender</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Score Summary – Total sample</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Score Summary – Total sample by age</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Score Summary – Total sample by gender</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Norm-based interpretation</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Content-based interpretation</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Criterion-based interpretation</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Medical expenditure prediction</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Health economics estimations</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age-by-gender norm comparison</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Disease-specific norm comparison</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Treatment efficacy analyses</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mortality risk analysis</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Depression screening</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Physical functioning screening</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* Assessed as part of estimable scale scores.
* Assessed as part of item internal consistency.
* Assessed as part of item discriminant validity.
* Available in late 2009.
** Available in 2010.
data analytic, and other consulting services to the healthcare and medical products (e.g., pharmaceutical, medical device, biologics) industries. In general, the Outcomes Insight Consulting Division’s main goal is to help healthcare organizations integrate PRO assessment into their clinical service and quality improvement systems; for medical products organizations, the goal is to incorporate PRO measurement into all phases of development, testing, and marketing. The Outcomes Insight Consulting team’s resources and focus include:

- A team of outcomes scientists with expertise in PRO instrument development to help design outcomes assessment systems, develop surveys and studies, score and analyze data, interpret findings, and present results
- Unique experience with analysis and interpretation of Short Form surveys and other PRO surveys
- Pharmacoeconomic analyses—including cost-effectiveness and cost-utility analyses—of PRO data obtained from measures such as the SF-6D and EQ-5D
- Proprietary data assets for benchmarking and interpreting results across multiple diseases or conditions
- Distinctive methodologies for analyzing disease burden and predicting future medical expenditures
- Bibliographic assets that include a proprietary database containing more than 12,000 publications referencing the Short Form surveys

Overview of Available Group-Level Interpretation and Reporting Services

Outcomes Insight Consulting has the resources to analyze group-level data and develop a report that can be customized to meet the needs of an organization. In addition to presenting summaries of a standard data quality evaluation and the basic $T$-score results, its reports can also address organizational needs through the presentation of findings from other kinds of analyses. These include norm-based, content-based, and criterion-based interpretation of the data based on group-specific guidelines; analyses of burden of disease, treatment efficacy, and mortality risk; health economics estimations; and cross-calibration with other PRO instruments. Each of these is briefly discussed in the following sections. Although software could be developed by certified vendors to automate analytic and reporting tasks, this may not be possible for all tasks. Moreover, organizational needs are likely to be unique and to change frequently over time, thus other types of data analyses and reporting capabilities may be required beyond what vendor-developed software can provide.

Data Quality Evaluation

Data problems can have a major impact on any study, including on the ability of that study to reach its intended goals. In the past, data problems with studies utilizing the SF-36 and SF-36v2 have impacted regulatory decisions, delayed the availability of new treatments in the marketplace, and even eliminated a researcher’s ability to compare study results with those from other participating sites. Most importantly, erroneously scored data impacts one’s ability to appropriately interpret scores. Some of these errors include:

- Administration of a survey that has inaccurately formatted questions (e.g., questions are omitted or presented in the wrong order)
- Pages not included or missing from the survey during administration
- Incorrectly keying response values into a database during the data entry phase

To identify these types of errors, QualityMetric developed Data Quality Evaluation (DQE) algorithms that are part of the Outcomes Insight Consulting scoring and analytic services. In terms of data quality, after being provided with a file containing raw item-level response data, the Outcomes Insight Consulting team performs a thorough review and inspection of the quality of data using DQE algorithms and then implements QualityMetric’s proprietary scoring algorithms to score the data. Once complete, the Outcomes Insight Consulting team
The data quality indicators that are examined by the Outcomes Insight Consulting team are described in detail in the second guide in this vendor certification series, *A Guide to the Integration of Certified Short Form Scoring and Data Quality Evaluation Capabilities* (Maruish & DeRosa, 2009). Thus, these indicators will be only briefly reviewed here. Interested readers are referred to the Maruish and DeRosa guide for a more detailed description of these indicators.

**Response consistency.** One of many SF-36v2 and SF-36 data quality indicators is the Response Consistency Index (RCI). Scoring the RCI is an easy way of evaluating the consistency of responses to individual survey items. The RCI consists of checks of consistency between 15 pairs of items. If a pair of responses is consistent, the RCI score for that pair equals 0, whereas a pair of inconsistent responses earns a score of 1. The final RCI score for a respondent is the sum of the scores for the 15 consistency checks. Thus, the best (i.e., most consistent) RCI score is 0 while the worst score (i.e., least consistent) is 15. This data quality indicator is considered satisfactory when the percentage of the total number of individual respondents with an RCI of zero is at least 90%.

**Missing data analysis.** Each Short Form survey dataset is reviewed for missing and out of range responses. An extraordinarily high rate of missing data for any one item is an indication that respondents may not have understood the question being asked or (in instances of paper-form survey administration) that there may have been a formatting problem with the survey items (if the standard Short Form survey format was not used). High missing data rates for a block of questions could also be an indication that a page of the survey may have been missing during the administration of the survey. Out of range responses typically indicate a problem during data entry; for example, errors in keying item response values into the database. Missing data rates for any one item that exceeds 5% is considered problematic. Fortunately, the Outcomes Insight Consulting group can apply proprietary missing score estimation algorithms for recovering partially completed surveys.

**Tests of scaling assumptions.** The scales of the Short Form surveys are scored using the method of summated rating scales. To compute a scale score using the method of summated rating scales, scores assigned to responses options of each survey item are simply summed (assuming item response values are all in the same direction, either positive or negative). The simplicity of this method is based on a number of assumptions that must be tested. These tests determine the appropriateness of including an item in a particular scale and whether it is appropriate to simply sum item scores to estimate a scale score. Further, these tests lend support to the validity of the measurement model developed for a survey instrument. QualityMetric tests the assumptions underlying the scoring of SF-36 scales and summary measures using the Multitrait Analysis Program (MAP; Ware, Harris, Gandek, Rogers, & Reese, 1997). For the SF-36v2 and SF-36, the MAP program tests the following assumptions necessary to support the scoring and interpretation of Short Form health domain scales and component summary measures:

1. Items measuring the same concept (scale) have approximately equal variances; otherwise item standardization may be required.
2. Each item is substantially, linearly related (r > 0.40 corrected for overlap) to the underlying concept being measured (item-internal consistency).
3. Each item correlates significantly (higher) with its hypothesized scale than with scales measuring other concepts (item-discriminant validity).
4. Items in the same scale contribute roughly equally to the same scale as defined by the item-scale correlation; if this assumption is not met, a strong basis should exist for including the item in the scale without weighting.
5. Internal-consistency reliability estimates for multi-item scales satisfy minimum standards for group-level comparisons (α > 0.70).
6. The Physical Functioning (PF), Role-Physical (RP), Bodily Pain (BP), and General Health (GH) scales correlate higher with the Physical Component Summary (PCS) measure than with the Mental Component Summary (MCS) measure.

7. The Mental Health (MH), Role-Emotional (RE), Social Functioning (SF), and Vitality (VT) scales correlate higher with MCS than with PCS.

**Group-Level Scoring Summaries and Profiles**

As with individual respondent reports, a tabular summary of aggregate scores and other findings allows the report user to obtain a quick, overall picture of the Short Form survey findings. In the case of group data, findings are commonly represented by the mean of the aggregate scores on each of the Short Form component summary measures, health domain scales, and SF-6D. Results for other variables may be reported as frequency counts (e.g., number of respondents with computable scales), percentages (e.g., percentage of respondents with complete sets of Short Form responses), or other values (e.g., mean change score from first to last administration of the Short Form survey). Depending on the homogeneity of the group, inclusion of relevant disease-specific normative data may be appropriate for the sake of comparison. Table 4.2 indicates the types of scores and other data that could be presented in a summary table for each of the Short Form surveys.

Generally, a single table summarizing Short Form results for all the members of the group will meet the needs of the organization, particularly if the group is relatively homogeneous (e.g., male diabetic patients aged 65 years or older). However, with a group of respondents who are more diverse on one or more important variables (e.g., gender, age, diagnosis, education, health insurance benefits), it is often useful to develop separate tables for presenting the results, broken down by meaningful categories within that variable (e.g., male vs. female; respondents 65 years and older vs. respondents 18–64 years; cancer patients vs. non-cancer patients).

Furthermore, scoring summaries can be supplemented by other tables that an organization would find useful. For example, a table presenting the mean raw score for each SF-36v2 item, along with correlations of the score of each item with the score of each of the 8 health domain scales, can be presented in support of the survey’s item convergent validity and discriminant validity. Or, a table showing the correlation of each SF-36v2 scale with each of the other scales and with the PCS and MCS measures provides support for the expected underlying two-dimensional structure of the survey, whereas a table presenting the computed Cronbach’s alpha coefficients for the health domain scales can serve as evidence of the SF-36v2’s internal consistency.

### Table 4.2

**Potential Content of Short Form Survey Score Summaries for Group-Level Data**

<table>
<thead>
<tr>
<th>Quantitative Variable</th>
<th>SF-36v2</th>
<th>SF-36</th>
<th>SF-12v2</th>
<th>SF-12</th>
<th>SF-8</th>
<th>SF-10</th>
</tr>
</thead>
<tbody>
<tr>
<td>DQE indicators</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Component summary measure scores</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Health domain scale scores</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Depression equivalent score</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Medical expenditure prediction</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SF-6D scores</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Disease-specific norms</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td>X*</td>
<td>X</td>
</tr>
<tr>
<td>Age-specific norms</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Change scores</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td>X</td>
<td>X</td>
</tr>
</tbody>
</table>

* Available in early 2010.
SF-36v2, SF-12v2, and SF-8 group mean health domain and component summary measure scores, for single or multiple administrations, can be profiled on the same standard Short Form score profile that is used for profiling individual respondent scores (see Figures 3.1 and 3.2). Also, as in the case of individual respondent scores, one may use the component summary measure portion of the standard profile as a template for profiling mean PCS and MCS scores for the SF-36 and SF-12 or the mean PHS-10 and PSS-10 scores for the SF-10.

Development of Group-Specific Interpretive Guidelines

Analyses can be conducted to develop Short Form interpretation guidelines for the PRO burden and treatment outcomes observed for the clinical or research group. This can be done using the same three general strategies used in interpreting individual respondent data.

Norm-based interpretation. Using a norm-based interpretation approach, the obtained \( T \)-score results from a single administration (i.e., mean \( T \) scores) are compared to the distribution of \( T \) scores in the U.S. general population to determine general levels of health status for the group on each of the Short Form health domains and component summary measures. Similarly, interpretation of PRO benefits of treatment can be determined from a normative perspective by examining change in Short Form \( T \) scores from one administration of the survey to another.

Content-based interpretation. As in individual patient-level interpretation, another group-level strategy utilizes content-based interpretations. In this approach, the units of analyses are the items of each scale on the Short Form survey. Meaningful cutoff points for each item in a scale are determined \textit{a priori}. For example, a meaningful, response-based cutoff point for SF-36v2 PF Item 3i (Walking one hundred yards) would be those indicating no limitations (\textit{not limited at all}) versus those having any level of limitation (\textit{limited a little} or \textit{limited a lot}). Knowing the percentages of patients above and below this cutoff point at baseline and at a subsequent visit can provide a salient interpretation of what the change in the PF scale score means.

Criterion-based interpretation. Also, as with patient-level interpretation, criterion-based interpretation for group-level data can be developed through analyses that link differences or changes in Short Form survey scores to differences or changes in conceptually related criterion variables. Here, the criterion variables that are selected can be specific to the desired outcomes of the study, such as number of hospitalizations or outpatient visits, employment status or missed days of work, or mortality rates.

Burden of Disease Analyses

The functional burden associated with various conditions can be assessed by comparing baseline Short Form survey scores of study participants with U.S. general population norms and, for the SF-36v2 and SF-12v2 administrations, disease-specific benchmarks. Based on QualityMetric’s available datasets, the Outcomes Insight Consulting team can also compare these scores against other disease-related benchmarks for disease burden comparison and interpretation. In addition, regression methods can be applied to adjust the U.S. general population norms and disease-specific benchmarks to the age and gender distribution of the user’s sample. The consulting team can also control for the presence of comorbidities, if applicable. Moreover, item-level analyses can be conducted in order to provide additional interpretation of the functional burden observed for the study group.

Treatment Efficacy Analyses

Clinical or research data can be analyzed to determine the statistical significance and meaningfulness of changes in Short Form scores using pre- to posttreatment datasets.

Analyses of change within and between groups. Statistical tests can be conducted to determine whether the changes in Short Form survey scores within study samples differ from zero. Statistical tests can also be conducted on clinical trials data to compare changes in Short Form scores between treated and placebo groups, using appropriate methods for group comparisons. These analyses can employ intent-to-treat designs, and effect sizes can be used to determine the magnitude of change scores within groups and between groups.
Repeated-measures analyses. For Short Form surveys administered at multiple follow-up visits, a repeated-measures analysis of scores can be conducted to evaluate the trends in scores over time. Furthermore, differences in trends in Short Form survey scores between treated and placebo groups in clinical trials can be evaluated with appropriate statistical tests.

Analyses of categorical change scores. Categories of change can be derived for each Short Form survey scale and summary measure. Each patient’s final score can be categorized as better, same, or worse according to the magnitude and direction of the difference in scores from baseline to final visit assessments. The analysis of categorical change scores provides a means of assessing the underlying variability in outcomes that are masked by the analysis of mean change scores. Results can be presented in terms of the percentages of patients who improve (better), stay the same (same), and decline (worse). Moreover, the amount of change in each Short Form survey score from baseline to final visit that meets the threshold for improvement (better) or decline (worse) can be determined from various approaches (e.g., SEM, 0.5 SD, MID differences).

Health Economics Estimations

Three sets of health economics-related estimates can be derived from SF-36v2 and SF-12v2 results. One is a medical expenditure prediction: Using algorithms developed by Fleishman, Cohen, Manning, and Kosinski (2006), PCS and MCS scores can be used to predict monthly medical expenditures during the subsequent 6 months. Another is concurrent work status prediction and risk of work loss due to health. The third health economics-related estimate is the relative benefits gained from treatment in terms of QALYs using SF-6D scores and cost-utility and cost-benefit analyses.

Mortality Risk Analyses

SF-36v2 scores, particularly PCS scores, have a strong association with 2-year and 5-year mortality (Ware et al., 2007). However, this association has been established primarily in general patient populations and the elderly (ages 65 and older). Using survival analysis methods and Cox proportional hazard models, a mortality risk model can be developed specifically from an organization’s Short Form data sets.

Cross-Calibration of PRO Tools

Using item response theory (IRT) models, PRO instruments measuring the same health concepts can be calibrated and scored on the same scale, permitting the equating of PRO scale scores and reporting the scores on the metric of the source instrument. For example, QualityMetric has developed cross-calibration tables to translate SF-36v2 MH scale scores into BDI score units. These score units can serve as the metric for evaluating mental health burden at baseline, as well as for analyzing the benefits of treatment. In addition, the efficacy of treatment for those screening positive for depression (based on standard BDI score cutoffs) can be compared to treatment outcomes for those who screened negative at baseline.

In another example of cross-calibration, clinical trial data can be used to explore approaches for developing a composite score algorithm that combines the SF-36v2 PF scale with the Health Assessment Questionnaire (HAQ; Fries, Spitz, Kraines, & Holman, 1980; Fries, Spitz, & Young, 1982). This can improve the assessment of treatment efficacy, thereby enabling the measurement of the physical functioning health domain beyond the ranges measured by each scale separately.


Walters, S. J., & Brazier, J. E. (2003). What is the relationship between the minimally important difference and health state utility values? The case of the SF-6D. *Health and Quality of Life Outcomes*, 1, 4.


Ware, J. E., Jr., Harris, W. J., Gandek, B., Rogers, B. W., & Reese, P. R. (1997). *MAP-R for Windows: Multitrait/multi-item analysis program. Revised user’s guide*. Boston, MA: Health Assessment Lab.
Appendix

QualityMetric Health Outcomes™
Scoring Software 2.0 Sample Reports
Appendix A  Sample SF-36v2 Patient Report

SF-36v2® Health Survey

YOUR SCORES

Survey Date: April 22, 2008
User ID: 34
Name: Daisy Smith
Age: 42
Gender: Female

PHYSICAL HEALTH SUMMARY  51

Based on your answers about health in the past 4 weeks, our research shows that:

Compared to the general population...
Physically, your...
- functioning is worse
- pain is much less
- performance of work, home or school activities is the same or better

Emotionally, your...
- bothered more than most
- participation in social activities is less limited
- performance of work, home and school activities is limited less

Overall, your...
- rating of your health is worse
- energy level is lower

WHAT YOUR SCORES MEAN

Mental Health Summary  47

Compared to other women of similar age...
- your physical health appears to be about the same
- your emotional health appears to be about the same

WHAT YOU SHOULD DO

YOUR PROGRESS

Date  Physical Health Summary  Mental Health Summary
Current: 04/22/08  51  47

- This is the first time you have completed the survey. We will report changes in scores the next time you complete this survey.

- In three months, take this survey again to continue monitoring your progress.
- By showing this report to your doctor or other health care provider, you can work together to make sure that your physical and mental health are as good as they can be.

Note: This survey is not a diagnostic tool. It is intended to supplement, but not replace or contradict the advice of your personal physician. If you have any questions or concerns about your health, it is always a good idea to seek one-on-one professional medical consultation.

Copyright © 1999-2008 QualityMetric Incorporated
Appendix B  Sample DYNHA General Health Survey Provider Report

DYNHA® General Health Survey  Provider Report for Jane Doe

January 10, 2007

YOUR SCORES

Report Date:  January 10, 2007
Survey Mode:  Web
Age:  38
Gender:  Female
Conditions:  OA, Problem with arm/leg

PHYSICAL HEALTH SUMMARY  38

38

WORST  30  40  50  60  70  US AVG.  BEST

Your physical health summary score is very much below average, even taking into account the margin of error.

MENTAL HEALTH SUMMARY  36

36

WORST  30  40  50  60  70  US AVG.  BEST

Your mental health summary score is very much below average, even taking into account the margin of error.

WHAT YOUR SCORES MEAN

<table>
<thead>
<tr>
<th>Scale</th>
<th>Current Score</th>
<th>Status Flags</th>
</tr>
</thead>
<tbody>
<tr>
<td>GH</td>
<td>29.0</td>
<td>⬤</td>
</tr>
<tr>
<td>PF</td>
<td>40.3</td>
<td>⬤</td>
</tr>
<tr>
<td>RP</td>
<td>36.1</td>
<td>⬤</td>
</tr>
<tr>
<td>BP</td>
<td>44.5</td>
<td>⬤</td>
</tr>
<tr>
<td>VT</td>
<td>39.1</td>
<td>⬤</td>
</tr>
<tr>
<td>SF</td>
<td>35.1</td>
<td>⬤</td>
</tr>
<tr>
<td>NH</td>
<td>44.2</td>
<td>⬤</td>
</tr>
<tr>
<td>RE</td>
<td>19.9</td>
<td>⬤</td>
</tr>
</tbody>
</table>

What the status flags mean...

- Current Score shows fast deterioration from last survey administration
- Current Score shows deterioration in relation to respondent's average score
- Respondent shows continuous deterioration over time
- Current Score shows improvement compared to last survey administration
- Respondent's score shows no significant change or this is the respondent's first survey administration

WHAT YOU SHOULD DO

- Schedule today, an appointment with your doctor to discuss the impact that your personal and emotional problems are having on your life.
- In one month, take this survey again to continue monitoring your progress.
- By showing this report to your doctor or other health care provider, you can work together to make sure that your physical & mental health are as good as they can be.

YOUR PROGRESS

Date  Physical Health Summary  Mental Health Summary
Current:  01/10/07  38  36
Previous:  01/10/07  44  27
Change:  worse  better

- Your physical and mental health summary scores changed significantly compared to the last time the survey was taken. Be sure to mention this to your doctor.
- The margin of error can cause small changes in scores. This survey focuses on changes of 5 points or more. However, you should always report to your doctor changes that are important to you.

Note: This survey is not a diagnostic tool. It is intended to supplement, but not replace or contradict the advice of your personal physician. If you have any questions or concerns about your health, it is always a good idea to seek one-on-one professional medical consultation.

Powered by DYNHA®

Copyright © 1995-2007 QualityMetric Incorporated
Appendix C  Sample SF-12v2 Provider Report with Medical Expenditure

SF-12v2™ Health Survey
(Advanced Reporting)

Patient Results
Report Date: July 21, 2005
Survey Mode: Web
Age: 39
Gender: Male
Conditions: Back problems

Physical Health Summary

45

Worst 30 40 50 60 70 US AVG. Best
This individual’s physical health summary score is about average, taking into account the margin of error.

Mental Health Summary

31

Worst 30 40 50 60 70 US AVG. Best
This individual’s mental health summary score is very much below average, even taking into account the margin of error.

Patient Progress

<table>
<thead>
<tr>
<th>Date</th>
<th>Physical Health Summary</th>
<th>Mental Health Summary</th>
<th>Medical Expenditure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Current: 07/21/05</td>
<td>45</td>
<td>31</td>
<td>$139</td>
</tr>
<tr>
<td>Previous: 07/20/05</td>
<td>35</td>
<td>44</td>
<td>$174</td>
</tr>
<tr>
<td>Change:</td>
<td>better</td>
<td>worse</td>
<td>-25.18%</td>
</tr>
</tbody>
</table>

- This individual’s physical and mental health summary scores changed significantly compared to the last time the survey was taken.
- The margin of error can cause small changes in scores. This survey focuses on changes of 5 points or more.

Interpretation
Based on this individual’s answers about health in the past 4 weeks, our research shows that:

Compared to the general population...
Physically, this individual’s...
- functioning is much worse
- pain is much worse
- performance of work, home or school activities is worse
Emotionally, ...
- bothered more than most
- participation in social activities is more limited
- performance of work, home and school activities is limited much more
Overall, this individual’s...
- rating of health is much better
- energy level is higher
Compared to other men of similar age...
- this individual’s physical health appears to be worse
- this individual’s emotional health appears to be much worse

Predicted Medical Expenditure Guideline

|                  | Predicted per month expenditure | This individual’s predicted expenditure is ...
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>This Individual</td>
<td>$139</td>
<td>at the 32 percentile of predicted expenditures.</td>
</tr>
<tr>
<td>General U.S.</td>
<td>$81</td>
<td>is 72% above the U.S. General Population Average.</td>
</tr>
<tr>
<td>population,</td>
<td></td>
<td></td>
</tr>
<tr>
<td>same age</td>
<td></td>
<td></td>
</tr>
<tr>
<td>and gender</td>
<td></td>
<td></td>
</tr>
<tr>
<td>General U.S.</td>
<td>$137</td>
<td>1% higher than the norm.</td>
</tr>
<tr>
<td>population,</td>
<td></td>
<td></td>
</tr>
<tr>
<td>healthy person</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Condition Benchmarks:
(Note: These are not additive for the individual)

<table>
<thead>
<tr>
<th>Condition</th>
<th>Cost</th>
<th>Equal to or below the norm</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anemia</td>
<td>$349</td>
<td></td>
</tr>
</tbody>
</table>

1 References: Fleishman JA, Cohen JW, Manning WG, Kosinski M. Using the SF-12 Health Status Measure to Predict Medical Expenditures: Medical Care (forthcoming); Fleishman JA, Cohen JW, Manning WG, Kosinski M. Using the SF-12 Health Status Measure to Predict Medical Expenditures. Presented at Academy Health Annual Meeting, 2006.

Note: This survey is not a diagnostic tool; it is intended to supplement clinical decision making. Visit us at: www.anhealthy.com for more information.
### Tabled Reporting

**Export to Excel**  
**Export to Text**

### Tabled Report Query Parameters:

All Survey Administrations Between 7/17/2005 12:00:00 AM AND 12/6/2005 12:00:00 AM; Including the Following Sites: '18'; Including Surveys: SF-12v2™ Health Survey (Advanced Reporting) (52430).

<table>
<thead>
<tr>
<th>Member</th>
<th>ID</th>
<th>DOB</th>
<th>Gender</th>
<th>Date</th>
<th>GH</th>
<th>PF</th>
<th>RP</th>
<th>BP</th>
<th>VT</th>
<th>SF</th>
<th>MH</th>
<th>RE</th>
<th>MCS</th>
<th>PCS</th>
<th>MedicalExpenditure</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>01/01/1936</td>
<td>M</td>
<td>11/30/2005 11:17:14</td>
<td>44.74</td>
<td>30.70</td>
<td>29.54</td>
<td>26.87</td>
<td>37.69</td>
<td>36.37</td>
<td>40.16</td>
<td>22.53</td>
<td>36.03</td>
<td>33.17</td>
<td></td>
<td>1135.8</td>
</tr>
<tr>
<td>Doe, John</td>
<td>02/07/1940</td>
<td>M</td>
<td>07/27/2005 14:36:04</td>
<td>29.65</td>
<td>30.70</td>
<td>29.54</td>
<td>37.06</td>
<td>47.75</td>
<td>26.27</td>
<td>40.16</td>
<td>22.53</td>
<td>34.93</td>
<td>33.00</td>
<td></td>
<td>1159.5</td>
</tr>
<tr>
<td>Smith, John</td>
<td>02/09/1940</td>
<td>M</td>
<td>07/19/2005 09:23:03</td>
<td>44.74</td>
<td>39.29</td>
<td>38.75</td>
<td>26.87</td>
<td>37.69</td>
<td>36.37</td>
<td>52.35</td>
<td>33.71</td>
<td>43.70</td>
<td>35.21</td>
<td></td>
<td>569.3</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>07/19/2005 10:58:53</td>
<td>44.74</td>
<td>47.88</td>
<td>57.18</td>
<td>47.25</td>
<td>67.88</td>
<td>56.57</td>
<td>64.54</td>
<td>56.08</td>
<td>65.63</td>
<td>45.52</td>
<td></td>
<td>456.7</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>07/25/2005 11:07:29</td>
<td>29.65</td>
<td>39.29</td>
<td>47.96</td>
<td>57.44</td>
<td>67.88</td>
<td>56.57</td>
<td>58.45</td>
<td>33.71</td>
<td>55.32</td>
<td>43.76</td>
<td></td>
<td>559.6</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>07/25/2005 14:26:00</td>
<td>29.65</td>
<td>30.70</td>
<td>38.75</td>
<td>26.87</td>
<td>47.75</td>
<td>36.37</td>
<td>40.16</td>
<td>11.35</td>
<td>32.65</td>
<td>35.07</td>
<td></td>
<td>1085.6</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>07/27/2005 15:12:23</td>
<td>44.74</td>
<td>39.29</td>
<td>34.14</td>
<td>37.06</td>
<td>47.75</td>
<td>36.37</td>
<td>52.35</td>
<td>22.53</td>
<td>40.79</td>
<td>39.26</td>
<td></td>
<td>815.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>09/26/2005 16:43:32</td>
<td>55.52</td>
<td>30.70</td>
<td>38.75</td>
<td>37.06</td>
<td>47.75</td>
<td>36.37</td>
<td>46.25</td>
<td>33.71</td>
<td>43.92</td>
<td>39.13</td>
<td></td>
<td>790.2</td>
</tr>
</tbody>
</table>
Appendix E  Sample SF-36v2 Total Sample Aggregate Report for Management and Providers


Abbreviation
PCS = Physical Component Summary
MCS = Mental Component Summary
GH = General Health
PF = Physical Functioning
RP = Role Physical
BP = Bodily Pain
VT = Vitality
SF = Social Functioning
RE = Role Emotional
MH = Mental Health

Note: The Depression Screening is not a diagnostic tool.
Appendix F  Sample SF-36v2 Total Sample by Age Aggregate Report for Management and Providers

Scores By Age Group

<table>
<thead>
<tr>
<th>Demographic Profile</th>
<th>Survey</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample Size 86</td>
<td>SF-36v2</td>
<td>Sample 2B</td>
</tr>
<tr>
<td>Male (*) 84%</td>
<td></td>
<td>Site: Dallas</td>
</tr>
<tr>
<td>Female (*) 16%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean Age 48</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age Range 20-73</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* Of those reporting gender (n=32)

** PCS Scores by Age Group **

<table>
<thead>
<tr>
<th>Sample</th>
<th>Norm</th>
<th># In Sample</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

** MCS Scores by Age Group **

<table>
<thead>
<tr>
<th>Sample</th>
<th>Norm</th>
<th># In Sample</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Appendix G  Sample SF-36v2 Total Sample by Gender Aggregate Report for Management and Providers