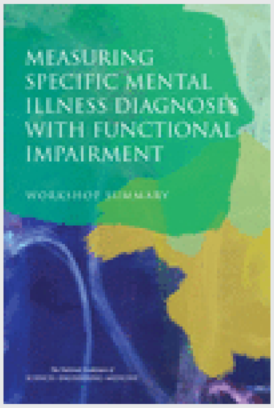
THE NATIONAL ACADEMIES PRESS


MEASURING SPECIFIC MENTAL ILLNESS DIAGNOSES WITH FUNCTIONAL IMPAIRMENT
WORKSHOP SUMMARY

Measuring Specific Mental Illness Diagnoses with Functional Impairment: Workshop Summary

DETAILS

106 pages | 6 x 9 | PAPERBACK
ISBN 978-0-309-39239-6 | DOI: 10.17226/21920

AUTHORS

Jeanne C. Rivard and Krisztina Marton, Rapporteurs; Committee on National Statistics; Board on Behavioral, Cognitive, and Sensory Sciences; Division of Behavioral and Social Sciences and Education; Board on Health Sciences Policy; Institute of Medicine; National Academies of Sciences, Engineering, and Medicine

MEASURING SPECIFIC MENTAL ILLNESS DIAGNOSES WITH FUNCTIONAL IMPAIRMENT

WORKSHOP SUMMARY

Jeanne C. Rivard and Krisztina Marton, *Rapporteurs*

Committee on National Statistics and
Board on Behavioral, Cognitive, and Sensory Sciences,
Division of Behavioral and Social Sciences and Education

and

Board on Health Sciences Policy, Institute of Medicine

The National Academies of
SCIENCES • ENGINEERING • MEDICINE

THE NATIONAL ACADEMIES PRESS
Washington, DC
www.nap.edu

THE NATIONAL ACADEMIES PRESS 500 Fifth Street, NW Washington, DC 20001

This activity was supported by Contract No. HHSP233201400020B/HHSP23337002 between the National Academy of Sciences and the U.S. Department of Health and Human Services. Support of the work of the Committee on National Statistics is provided by a consortium of federal agencies through a grant from the National Science Foundation (No. SES-1024012). Any opinions, findings, conclusions, or recommendations expressed in this publication do not necessarily reflect the views of any organization or agency that provided support for the project.

International Standard Book Number-13: 978-0-309-39239-6
International Standard Book Number-10: 0-309-39239-X
Digital Object Identifier: 10.17226/21920

Additional copies of this report are available for sale from the National Academies Press, 500 Fifth Street, NW, Keck 360, Washington, DC 20001; (800) 624-6242 or (202) 334-3313; http://www.nap.edu.

Printed in the United States of America

Suggested citation: National Academies of Sciences, Engineering, and Medicine. (2016). *Measuring Specific Mental Illness Diagnoses with Functional Impairment: Workshop Summary.* J.C. Rivard and K. Marton, Rapporteurs. Committee on National Statistics and Board on Behavioral, Cognitive, and Sensory Sciences, Division of Behavioral and Social Sciences and Education. Board on Health Sciences Policy, Institute of Medicine. Washington, DC: The National Academies Press. doi: 10.17226/21920.

The National Academies of
SCIENCES · ENGINEERING · MEDICINE

The **National Academy of Sciences** was established in 1863 by an Act of Congress, signed by President Lincoln, as a private, nongovernmental institution to advise the nation on issues related to science and technology. Members are elected by their peers for outstanding contributions to research. Dr. Ralph J. Cicerone is president.

The **National Academy of Engineering** was established in 1964 under the charter of the National Academy of Sciences to bring the practices of engineering to advising the nation. Members are elected by their peers for extraordinary contributions to engineering. Dr. C. D. Mote, Jr., is president.

The **National Academy of Medicine** (formerly the Institute of Medicine) was established in 1970 under the charter of the National Academy of Sciences to advise the nation on medical and health issues. Members are elected by their peers for distinguished contributions to medicine and health. Dr. Victor J. Dzau is president.

The three Academies work together as the **National Academies of Sciences, Engineering, and Medicine** to provide independent, objective analysis and advice to the nation and conduct other activities to solve complex problems and inform public policy decisions. The Academies also encourage education and research, recognize outstanding contributions to knowledge, and increase public understanding in matters of science, engineering, and medicine.

Learn more about the National Academies of Sciences, Engineering, and Medicine at **www.national-academies.org**.

STEERING COMMITTEE FOR THE WORKSHOP ON INTEGRATING NEW MEASURES OF SPECIFIC MENTAL ILLNESS DIAGNOSES WITH FUNCTIONAL IMPAIRMENT INTO THE SUBSTANCE ABUSE AND MENTAL HEALTH ADMINISTRATION'S DATA COLLECTION PROGRAMS

BENJAMIN G. DRUSS (*Chair*), Rollins School of Public Health, Emory University
FREDERICK G. CONRAD, Institute for Social Research, University of Michigan
ROBERT F. KRUEGER, Department of Psychology, University of Minnesota
RONALD MANDERSCHEID, National Association of County Behavioral Health & Developmental Disability Directors, Washington, DC, and Johns Hopkins University
NORA CATE SCHAEFFER, Department of Sociology, University of Wisconsin–Madison

KRISZTINA MARTON, *Study Director*
JEANNE C. RIVARD, *Senior Program Officer*
MICHAEL SIRI, *Program Associate*

COMMITTEE ON NATIONAL STATISTICS

LAWRENCE D. BROWN (*Chair*), Department of Statistics, The Wharton School, University of Pennsylvania
JOHN M. ABOWD, School of Industrial and Labor Relations, Cornell University
FRANCINE BLAU, Department of Economics, Cornell University
MARY ELLEN BOCK, Department of Statistics (emerita), Purdue University
MICHAEL CHERNEW, Department of Health Care Policy, Harvard Medical School
DONALD DILLMAN, Social and Economic Sciences Research Center, Washington State University
CONSTANTINE GATSONIS, Department of Biostatistics and Center for Statistical Sciences, Brown University
JAMES S. HOUSE, Survey Research Center, Institute for Social Research, University of Michigan
MICHAEL HOUT, Department of Sociology, New York University
THOMAS MESENBOURG, U.S. Census Bureau (retired)
SUSAN MURPHY, Department of Statistics and Institute for Social Research, University of Michigan
SARAH NUSSER, Office of the Vice President for Research, Iowa State University
COLM O'MUIRCHEARTAIGH, Harris School of Public Policy Studies, University of Chicago
RUTH PETERSON, Criminal Justice Research Center, Ohio State University
ROBERTO RIGOBON, Sloan School of Management, Massachusetts Institute of Technology
EDWARD SHORTLIFFE, Department of Biomedical Informatics, Columbia University and Arizona State University

CONSTANCE F. CITRO, *Director*
BRIAN HARRIS-KOJETIN, *Deputy Director*

BOARD ON BEHAVIORAL, COGNITIVE, AND SENSORY SCIENCES

BOARD ON HEALTH SCIENCES POLICY

Acknowledgment of Reviewers

This workshop summary has been reviewed in draft form by individuals chosen for their diverse perspectives and technical expertise, in accordance with procedures approved by the Report Review Committee of the National Academies of Sciences, Engineering, and Medicine. The purpose of this independent review is to provide candid and critical comments that will assist the institution in making its published report as sound as possible and to ensure that the report meets institutional standards for objectivity, evidence, and responsiveness to the charge. The review comments and draft manuscript remain confidential to protect the integrity of the process.

We thank the following individuals for their review of this report: Bruce Dohrenwend, Mailman School of Public Health, Columbia University; Robert Gibbons, Departments of Medicine and Public Health Sciences, University of Chicago; and James Wagner, Institute for Social Research, University of Michigan.

Although the reviewers listed above provided many constructive comments and suggestions, they were not asked to endorse the content of the report nor did they see the final draft of the report before its release. The review of this report was overseen by Susan A. Murphy, Department of Statistics and Institute for Social Research, University of Michigan. Appointed by the Academies, she was responsible for making certain that an independent examination of this report was carried out in accordance with institutional procedures and that all review comments were carefully considered. Responsibility for the final content of this report rests entirely with the rapporteurs and the institution.

Contents

APPENDIXES

1

Introduction

BACKGROUND

This report summarizes the presentations and discussions at the Workshop on Integrating New Measures of Specific Mental Illness Diagnoses with Functional Impairment into the Substance Abuse and Mental Health Services Administration's (SAMHSA) Data Collection Programs, which was held in Washington, D.C., in September 2015. The workshop was organized as part of an effort to assist SAMHSA and the Office of the Assistant Secretary for Planning and Evaluation (ASPE) of the U.S. Department of Health and Human Services (HHS) in their responsibilities to expand the collection of behavioral health data in several areas. The workshop was structured to bring together experts in the measurement of specific mental illness diagnoses in adults, related functional impairment, and health survey methods to facilitate discussion of measures and mechanisms most promising for expanding SAMHSA's data collections in this area.

The overall effort is being overseen by the Standing Committee on Integrating New Behavioral Health Measures into SAMHSA's Data Collection Programs.[1] In addition to the topics covered by this workshop, SAMHSA and ASPE are interested in expanding data collection on serious emotional disturbance in children, on trauma, and on recovery from

[1] For a description of the overall study, see http://sites.nationalacademies.org/DBASSE/CNSTAT/Behavioral_Health_Measures_Committee/index.htm [October 2015].

substance use or mental disorder. Workshops on all four topics are being convened as part of the overall effort.

WORKSHOP FOCUS

Neil Russell of SAMHSA described his agency's goals in exploring how to best measure and expand SAMHSA's data collection programs to include specific mental illness diagnoses with functional impairment and the inherent challenges in this effort. He first explained that the parameters of the expanded data collection are (1) to produce direct national estimates and state estimates for adult mental disorders with functional impairment, (2) to collect data on a wider variety of disorders than would be needed to estimate the prevalence of "serious mental illness,"[2] and (3) to collect these data at a frequency of not less than every 5 years.

Russell next summarized a previous SAMHSA effort to collect data on adult mental health disorders through adding a module to the National Survey of Drug Use and Health (NSDUH): that module was the Mental Health Surveillance Study (MHSS). The purpose of the MHSS was to produce national and state-level estimates of the prevalence of serious mental illness in accordance with SAMHSA's legislative mandate.

Russell further explained that the MHSS covered noninstitutionalized civilians aged 18 years and older who completed the NSDUH questionnaire in English. There was no Spanish version of the MHSS. NSDUH respondents were sampled and recruited for a one-time follow-up clinical interview that was administered following the NSDUH main study. The MHSS was fielded for 5 years, between 2008 and 2012. At its conclusion, a total of 5,653 respondents had participated, for an overall weighted response rate of 64.6 percent. Data from the MHSS were used to develop a model to estimate serious mental illness and apply it to the full sample of NSDUH participants.

Russell then described the interview process for the MHSS: on average, the clinical telephone interview was 72 minutes. The instrument used was the Structured Clinical Interview for DSM-IV-TR Axis I Disorders (SCID-I)–Non-patient Edition (SCID-I/NP), which was administered by interviewers who had undergone extensive training. The SCID includes standardized questions that are read verbatim and sequentially, followed

[2] On May 20, 1993, SAMHSA's Center for Mental Health Services (CMHS) published its definition of serious mental illness in the *Federal Register* (58 FR 29425): Persons aged 18 and over, who currently or at any time during the past year, have had diagnosable mental, behavioral, or emotional disorder of sufficient duration to meet diagnostic criteria specified within DSM-III-R [*Diagnostic and Statistical Manual of Mental Disorders, III, Revised*] that has resulted in functional impairment, which substantially interferes with or limits one or more major life activities.

by unstructured follow-up questions that the interviewers tailor to each respondent on the basis of clinical judgment and respondent answers. Clinical judgment was used to code each item in the SCID as “1” (absent or false), “2” (subthreshold), “3” (threshold or true), or “?” (inadequate information).

Data were collected on the following disorders: past year mood disorders (including major depressive disorder, bipolar disorder-manic episode, dysthymic disorder); past year anxiety disorders (including specific phobia, social phobia, generalized anxiety disorder, panic disorder with and without agoraphobia, agoraphobia without history of panic disorder, obsessive compulsive disorder, posttraumatic stress disorder); past year substance use disorders (including alcohol abuse or dependence, and drug abuse or dependence); past year eating disorders (including anorexia nervosa, bulimia nervosa); past year adjustment disorders; past year impulse control disorders (including intermittent explosive disorder); and past year psychotic symptoms (including delusions or hallucinations).

Russell pointed out that several mental disorders were excluded because some types of disorders are not amenable to the structure of the MHSS, including bipolar II disorder, personality disorders, other disorders typically identified in childhood, schizophrenia, and other psychotic disorders. However, a screener for two psychotic symptoms was included in the assessment. Developmental disorders were also excluded because they are excluded from the definition of serious mental illness.

The MHSS also included the Global Assessment of Functioning (GAF), a global measure of functional impairment. With scores ranging from 1 to 100, the GAF scale is a measure of global functional impairment rather than functional impairment specific to an individual mental disorder.[3]

Russell described challenges SAMHSA faces in considering how to collect data on a wider range of disorders while factoring in impairments related to those disorders. The first challenge is measuring disorder-specific functional impairment when there are multiple disorders and medical conditions. It is not clear whether respondents can accurately attribute functional impairment to a specific mental disorder in the presence of two or more mental disorders (including substance use disorders) and medical conditions, such as stroke or heart disease.

The second challenge is identifying measures of functional impairment. SAMHSA used a global measure of impairment in the MHSS,

[3] For a report that provides a global overview of the data, see Center for Behavioral Health Statistics and Quality. (2014). *2012 National Survey on Drug Use and Health: Methodological Resource Book* (Section 16a, 2012 Mental Health Surveillance Study: Design and Estimation Report). Rockville, MD: Substance Abuse and Mental Health Services Administration.

but the current question is whether data on impairment for a particular disorder can be measured and collected. The GAF was omitted from the *Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition* (DSM-5), because it lacked conceptual clarity and had questionable psychometric properties in routine practice. DSM-5 advises clinicians to use the World Health Organization Disability Assessment Scale (WHODAS, 2.0), either the 12-item or 36-item version; SAMHSA has been using a truncated, 8-item version in the NSDUH since 2008.

Third, Russell underscored that assistance is needed in identifying instruments to measure more DSM-5 disorders than are covered in the MHSS; considering the issues and instruments related to measuring functional impairment; and determining the most suitable approach for collecting the data. He listed SAMSHA's main options for data collection, noting that some of these options could require guidance on implementing model-based estimation procedures:

- Using the NSDUH, which would require a major redesign effort to incorporate the collection of data on specific mental health diagnoses with functional impairment.
- Reinstating the MHSS to accommodate more disorders and functional impairment, perhaps as was previously done by gathering data from small subsamples over time.
- Developing a new data collection program.
- Using existing data sources if they are representative at the national and state levels and the questions used have good psychometric properties.

D.E.B. Potter of the Office of the Assistant Secretary of Planning and Evaluation of HHS, cosponsor of the study, extended her appreciation for the expertise of the standing committee, workshop steering committee, and workshop presenters. She emphasized that the workshop would be informing not only SAMHSA, but also other surveys that HHS administers. She encouraged discusion that could inform multiple surveys; multiple purposes, including epidemiological and policy; and short-term as well as longer-term solutions.

WORKSHOP CHARGE

The specific statement of task for the workshop (shown in Box 1-1) was developed on the basis of the charge for the overall project, which was to expand data collections on several behavioral health topics. The main goals of the workshop were to discuss options for collecting data and producing estimates on specific mental illness diagnoses with func-

BOX 1-1
Statement of Task

A steering committee will organize a public workshop that will feature invited presentations and discussions on options for expanding SAMHSA's behavioral health data collections to include measures of specific mental illness diagnoses with functional impairment. The discussion will explore new measures and efficient mechanisms for collecting the data. Possibilities include adding new measures to existing surveys, initiating new data collections, or implementing model-based estimation procedures that take advantage of existing data sources, in the event that primary data collection methods are cost prohibitive or not necessary. Survey and questionnaire design tradeoffs, as well as the potential impact of any changes to existing surveys, will also be discussed. An individually authored summary of the presentations and discussions at the workshop will be prepared by a designated rapporteur in accordance with institutional guidelines.

tional impairment, including available measures and associated possible data collection mechanisms.

ORGANIZATION OF THE REPORT

This summary describes the workshop presentations and the discussions that followed each topic: see the workshop agenda in Appendix A. Biographical sketches of the presenters and of the steering committee members are in Appendix B.

Chapter 2 covers two topics that were used to set the stage for later presentations: the historical context of collecting data on mental illness diagnoses and functional impairment in the United States, which has been driven by evolving federal definitions of serious mental illness over the last 60 years, and a new study that is presently being planned by the National Institute of Mental Health (NIMH) to collect prevalence data on a range of mental illness diagnoses.

Chapter 3 describes studies that have been conducted to estimate specific mental illness diagnoses with functional impairment, and instruments that are available for this purpose.

Chapter 4 looks at existing data and data collection methods for measuring disorders, severity, and impairment, including the approaches used by the Global Burden of Disease study and the National Health Interview Survey, as well as the strengths and weaknesses of the administrative and other data sources available for potentially estimating the prevalence of mental disorders. Chapter 5 discusses innovative approaches to measurement: computerized adaptive testing and the Patient Reported Outcome Measurement System.

Chapter 6 covers the final workshop discussions, summarizing the major themes and implications for SAMHSA's planning efforts.

This report has been prepared by the workshop rapporteurs as a factual summary of what occurred at the workshop. The steering committee's role was limited to planning and convening the workshop. The views contained in the report are those of individual workshop participants and do not necessarily represent the views of all workshop participants, the steering committee, or the National Academies of Sciences, Engineering, and Medicine.

2

Data Needs and Studies Planned

HISTORICAL OVERVIEW OF THE DATA NEEDS

Ron Manderscheid (National Association of County Behavioral Health & Developmental Disability Directors and Johns Hopkins University) discussed the evolution of federal definitions for adults with serious mental illness over the past 60 years. He noted that the definitional work has kept abreast of the evolution of mental health services research. Definitions have changed as various versions of the DSM of the American Psychiatric Association (APA) were developed and as understanding of functional impairment has grown. He suggested that the definitions have had a major impact on the perception of services and services research. Right now, for example, the Murphy-Johnson bill in the U.S. House of Representatives and the Cassidy-Murphy bill in the U.S. Senate use the term "adults with serious mental illness," which is the definition developed by SAMHSA in the late 1990s. In addition to services and research applications, the definitions have policy and legal implications.

Manderscheid noted key concepts that are important in discussing estimation: *prevalence*, the total number of cases for a defined period of time; *incidence*, the number of new cases for a defined period of time; *treated prevalence*, the number of cases under care in specialty mental health settings; and *community prevalence*, the total number of cases in the community, including those under care.

From the 1950s to the 1970s, the field relied on treated prevalence and defined persons with mental illness by their diagnoses. There were no national epidemiological surveys, so prevalence rates of various diagnoses were based on the number of patients treated. The first national epidemiological survey, launched in the early 1980s, was the Epidemiological Catchment Area (ECA)[1] Project that Darrel Regier led and will be talking about later in the workshop. The ECA set new benchmarks that gave the field the ability to collect data in the community using lay interviewers. The ECA provided the capacity to produce national community prevalence estimates with adults.

During the 1980s, additional work was conducted to categorize adults with serious and persistent mental illness, which was defined by five different disorders (schizophrenia, bipolar disorder, depression, anxiety, and personality disorders), a Global Assessment of Functioning (GAF) score of 50 or less, and duration of 1 year. This was termed the "diagnosis-disability-duration" definition. Calculating prevalence in this way resulted in estimates of serious and persistent mental illness of about 2.8 percent of adults.

In the 1990s, the legislation that created SAMHSA in 1992, Public Law 102321, the Alcohol Drug Abuse Mental Health Administration Reorganization Act, required that SAMHSA develop a definition of adults with "serious mental illness" and that this definition be operationalized and applied to the Community Mental Health Services Block Grant Program. Manderscheid explained that the development of the new definition started with diagnosis and disability and eliminated duration. It included any disorder and a GAF score of 60 or less. Using this definition, he said, the prevalence of serious mental illness was about 5.8 percent of the adult population. The definition was tested with the ECA data that were collected in the early 1980s, and the definition was applied to the first iteration of the National Comorbidity Survey. The prevalence rates were consistent over time.

In the 2000s, SAMHSA used a proxy measure that was a variant of the Kessler Six Items Scale (K6), plus the World Health Organization Disability Assessment Schedule (WHODAS). When standardized on a global assessment scale score of 50 or less, this proxy measure produced a prevalence rate of about 4 percent of the adult population.

Manderscheid said that the definitions were mainly used for policy applications and then later for legal applications. He emphasized that the development and application of the definition of serious mental illness is not just for intellectual interest; it also has tremendous implications for the funding that Congress appropriates for this population.

[1] For details, see Regier, D.A., Myers, J.K., Kramer, M., Robins, L.N., Blazer, D.G., Hough, R.L., Eaton, W.W., and Locke, B.Z. (1984). The NIMH Epidemiologic Catchment Area Program: Historical context, major objectives, and study population characteristics. *Archives of General Psychiatry, 41,* 934-941.

Looking to the future, Manderscheid suggested that the era of electronic health records may change how prevalence data are collected, with increased reliance on samples constructed from these records. Having the right variables measured in the same way in electronic health records would be critical, and much developmental work would be needed to develop nationally comparable systems that could communicate with each other. He pointed out that there has been congressional interest in providing funds to the behavioral health world to adopt electronic health records.

Manderscheid also remarked that the field is in the midst of another transition from focusing on problems using deficit-based measures (i.e., diagnoses and functional impairments) to strength-based measures with concepts and measures being developed by consumers, peer programs, and researchers. One example of this is the development of measures of well-being and of dimensions of well-being—physical, mental, and social. Manderscheid concluded by noting three recent papers and reports that address these topics.[2]

Manderscheid's presentation was followed by a discussion that focused primarily on points he made related to electronic health records and the capacity of this data source to estimate prevalence. Hortensia Amaro (University of Southern California) asked for Manderscheid's thoughts on how well estimates based on electronic health records would reflect population-level estimates, considering that some populations are uninsured, underinsured, or do not use health services as frequently as others. Manderscheid replied that his assumption is that a universal system of electronic health records would be developed. A system that is not universal would indeed have inherent biases, underrepresenting immigrants who are not citizens, people who are in jail, those not currently enrolled in a health insurance program, and others. He pointed out that some of those same biases exist in the estimates based on current national surveys. For example, when people began to abandon landline telephones in favor of cell phones, the biases introduced into telephone surveys, which at the time were relying primarily on samples of landlines, had to be corrected. In his opinion, there might be a shift to new ways

[2] Institute of Medicine. (2015). *Vital Signs: Core Metrics for Health and Health Care Progress.* Washington, DC: The National Academies Press.

Manderscheid, R.W., Ryff, C.D., Freeman, E.J., McKnight-Eily, L.R., Dhingra, S., and Strine. T.W. (2010). Evolving definitions of mental illness and wellness. *Preventing Chronic Disease, 7*(1). Available: http://www.cdc.gov/pcd/issues/2010/jan/09_0124.htm [November 2015].

Schulte, P.A., Guerin, R.J., Schill, A.L., Bhattacharya, A., Cunningham, T.R., Pandalai, S.P., Eggerth, D., and Stephenson, C.M. (2015). Considerations for incorporating well-being in public policy for workers and workplaces. *American Journal of Public Health, 105*(8), e31-e44.

of collecting data, including more investment into the development of electronic health records.

Commenting further on the population considerations, Dean Kilpatrick (Medical University of South Carolina) added that, even if electronic health records were universal, the data would only reflect situations when a person decides to visit a physician because he or she has a problem. Furthermore, the nature of the health care system is such that health records reflect encounters to address problems and not a more general understanding of the patient. Kilpatrick also agreed with Amaro that people would be missing from the system if they do not have access to health care. Manderscheid noted that as part of the development of electronic health records there are parallel efforts to develop personal health records, which are electronic records that belong to individuals and centralize all of their health information. Electronic health records and personal health records could be brought together in a systematic way to develop a new system. He emphasized his agreement with Kilpatrick that electronic health records are not currently ready for this type of use but said that these ideas need to be put on the table for planning for the future.

Along the same lines, Theo Vos (University of Washington) added that another subset of the population that would be missing from electronic health records are people with unrecognized disease who have not sought care, which is particularly applicable for mental disorders. In addition, he said, in health records one would be relying on the different ways that clinicians determine diagnoses. Comparability would be lost in terms of being able to control the inclusion and exclusion criteria, as well as the case definitions. Manderscheid agreed that comparable definitions and structures are needed in electronic health records.

Nora Cate Schaeffer (University of Wisconsin) asked about informed consent issues for research using electronic health records and about how the need for covariates—which are available in population studies, but typically not in electronic health records—could be addressed. Manderscheid replied that there is a need for incorporating a systematic plan into the design of any system for electronic health records, not only obtaining permission for the possible use of the data in research, but also advance directives for the assignment of medical power of attorney and permission for sharing private information. In terms of covariates, Manderscheid said that it would also be important to decide early on the basic structure and scope of an electronic health records system, and whether it should include only service use data or additional covariates to enable research.

Robert Krueger (University of Minnesota) commented that the usability of electronic health records as a basis for prevalence data would be affected by how clinicians on the front lines actually use diagnostic systems. For example, in mental health settings, clinicians often use "not otherwise specified" diagnoses on encounter forms: that is, they often do not use the diagnostic system in the way it was intended. This is important to keep in mind when considering the use of electronic health records for prevalence data.

In a similar vein, Graham Kalton (Westat) commented that the survey research concept of reliability is important to consider when discussing the use of any administrative records. Administrative data are often collected by a variety of people who apply definitions in different ways, which would affect the quality of estimates derived from those data.

Robert Gibbons (University of Chicago) remarked that it is very easy to dismiss the usefulness of electronic health records, but this process can change from what is now a passive process to a more active process. However, it is important to think about two distinct issues. First, the measurement process could be greatly improved and made more comparable to the data collection involved in large-scale surveys. Second, the population coverage bias inherent in electronic health records is more difficult to address.

Mark Olfson (Columbia University) agreed with the concerns about reliability of the data because of the differences among the raters who would be entering information into electronic health records systems. He suggested that one way forward may be to begin integrating the routine collection of brief self-report measures into electronic health records.

Darrel Regier (Uniformed Services University) added that one of the things the APA revision team for the DSM-5 has been trying to do with the DSM-5 cross-cutting measures and the WHODAS disability measure is to eventually include self-report measures in electronic health records. However, this inclusion can happen only if there is an electronic platform for collecting these types of cross-cutting and disability measures. An electronic platform would inform clinicians and physicians and guide them through a more rigorous diagnostic process of examining symptom profiles of patients and following those symptom profiles over time for outcome measures. He said that some of the thinking that has gone into the understanding of diagnoses can apply to how researchers and others approach electronic health records and diagnosis in clinical settings in general.

David Cella (Northwestern University) said that the conclusion seems to be that there should be some investment in the capture of standardized information in electronic health records that could be used as the basis for policy decisions and funding allocation. Clinicians may continue to use "not otherwise specified" or not provide any documentation at all, or they may choose to assume a more active role in the development of a standardized diagnostic approach.

Vos pointed out that, as part of the Global Burden of Disease study, the researchers started analyzing large volumes of U.S. medical data from private health insurers, Medicare, and Medicaid. They found that the data on chronic, persistent illness from these records are very comparable to survey data. However, for chronic episodic or shorter duration conditions, it is more difficult to decipher whether a diagnosis, seen at one point in a record, was still present or not over the course of a year. The other issue they encountered was that medical records often do not provide information on severity, or severity may not be defined in a standard way. He said that, in his view, electronic health records are a wonderful source to work with, but they will never do away with the need to collect survey data as a complement.

A NEW NATIONAL INSTITUTE OF MENTAL HEALTH INITIATIVE

Lisa Colpe (National Institute of Mental Health) discussed a new initiative to field a nationally representative, in-person, household survey to assess mental disorders and their correlates among youth and adults: NIMH began conversations with SAMHSA last year about a possible follow-up study to the National Survey of Drug Use and Health (NSDUH). Colpe said that the two agencies have a history of collaboration and that NIMH also supported the expansion of SAMHSA's Mental Health Surveillance Study in order to produce disorder-based estimates.

NIMH's current plans are to collect data from an age-stratified sample of 13,500 people: one-third between 13 and 17 years old, one-third between 18 and 30 years old, and one-third over 30 years old. This strategy will allow for an oversample of people in the younger age group. Another goal is to follow people over time, so the sample will be designed to result in a sufficient number of people who are eligible for the types of follow-up studies that are planned.

Colpe said that NIMH plans to use a 5-minute household screener followed by a 65-minute personal interview that is administered by computer-assisted personal interview (CAPI) and audio computer-assisted self-interview (ACASI). This data collection method follows the procedure used in the NSDUH. For a subset of the sample, NIMH plans to administer a telephone follow-up clinical interview at 2-4 weeks post-interview for clinical validation and calibration of the disorder modules in the survey. The current plan is to use the SCID for the follow-up, along with a psychotic symptom scale.

The first step will be to conduct a field test with 1,500 respondents to test self-administered versions of scales that have generally been interviewer administered in the past. Although the study has cross-sectional aspects for producing prevalence rates, it will also be used as a platform for follow-up longitudinal studies with subsets of respondents, such as those who score above a threshold on the psychotic symptom scale. Other populations may be identified either for longitudinal follow-up to track whether a disorder develops over time or to participate in a more complete evaluation. All respondents will be asked for consent to be contacted again in the future, which will offer flexibility for future studies.

The planned survey module topics include comprehensive demographics; mental disorders (including substance use and personality measures); suicidality (including past and recent suicidal behavior, access to firearms); psychotic-like experiences (including diagnostic history, family history); traumatic experiences (including childhood adversities, exposures to violence, disasters, life-threatening events), research domain criteria (RDoC) dimensional measures,[3] NIMH common data elements,[4] and chronic health conditions (including head injury, health behaviors).

In addition to questions about disorders, the survey will include a relatively large module covering health and mental health service use, frequency, and how effective respondents find those services. Another module pertains to lifetime as well as past year homelessness in persons with mental illness, which will allow NIMH to examine the number of persons who are chronically homeless or who are rotating in and out of homelessness. NIMH also plans to collect data on people who have served in the military and the types of exposures during their service. Colpe added that some modules will be rotated or administered to a subset of the full sample.

With regard to specific disorders, Colpe provided a list that NIMH hopes to include in the study but noted that there may be others. For adults, these include depression, mania, posttraumatic stress disorder, panic disorder, social phobia, agoraphobia, generalized anxiety disorder, eating disorders, obsessive-compulsive disorder, attention deficit hyperactivity disorder, and substance use. For children, the same list of disorders is planned with the addition of specific phobia, oppositional defiant disorder, conduct disorder, and separation anxiety disorder.

Colpe explained that the project will become part of the existing contract SAMHSA has to conduct the NSDUH study. The deliverables will be reports based on the survey data and clinical calibration; datasets for public and restricted use, which will be added to the NIMH data repositories; and the survey instrument modules and documentation as a resource for researchers.

[3] See https://www.nimh.nih.gov/research-priorities/rdoc/index.shtml [December 2015].

[4] See https://www.phenxtoolkit.org/index.php [December 2015].

Colpe closed by laying out the project timeline:

- 2015 to 2016: consulting with the field, instrument programming, interviewer training, and field materials development;
- 2016 to 2017: carrying out the pilot test, integrating findings, and adjusting the main survey;
- 2017 to 2018: conducting the main cross-sectional survey, launching a planned clinical reappraisal study (first follow-up study by telephone), and launching the second follow-up study; and
- 2019 and subsequent years: producing the national findings report, manuscript drafts, and the public-use data file.

Olfson asked Colpe whether the NIMH team has given any thought to oversampling low income people, ethnic or racial minorities, or other groups of policy interest. Colpe replied that right now NIMH only plans to oversample by age groups. Kalton asked if Colpe could explain further the reason for the way the age stratification is designed and whether this will be the most efficient design. Colpe explained that budget limitations drove their decision to make sure they had an adequate sample for youth and "transition age" people. She said that the team acknowledges that it will take more screening of households to achieve these specifications.

Kalton also asked for clarification on whether the NIMH sample will be drawn from the NSDUH sample. Russell clarified that the NIMH study will be using retired segments from the NSDUH sample, but the study will not include households that participated in the NSDUH. He added that the sample is about 2-3 years old, and because of that some updating of the records will be necessary.

Kilpatrick asked Colpe about the purpose of the 5-minute household screener. He also asked whether they will interview all household members or a randomly selected individual within each household. Colpe replied that the purpose of the screener is to find out who is in the household, after which 0, 1, or 2 members will be selected for the interview.

Kilpatrick also asked whether the decisions to administer the full modules will be based only on the responses to the screening questions. Colpe responded that most of the disorder modules will be asked on the basis of the answers to stem symptom questions. Kilpatrick expressed concern that the choice of stem questions can have a large effect on the estimates if they are restricted to only certain symptoms and exclude people with the disorder who have other symptoms. He suggested considering a planned missing data design, which would involve administering full modules to predetermined subsets of the sample.

Regarding the screening questions for psychotic symptoms, Krueger asked whether there are plans to conduct a clinical reappraisal. Colpe replied that the team plans to use the PQ-16, which has been identified as an NIMH common data element (see above) and will use a psychotic symptom scale for a clinical reevaluation. The team plans on dividing the responses into score bins—low, medium, and high—and they will examine a proportion of the cases in each bin in order to get a sense of what these scores mean in the general population and in light of the clinical evaluations.

Regier asked Colpe what measures will be used for personality disorders and for RDoCs. Colpe answered that they will be using some of the RDoC recommended measures. For personality disorders, a screen will be used, one of the personality inventory scales in the DSM-5, and some items from the other DSM-5 cross-cutting measures. For clinical follow-up they will use the International Classification of Diseases (ICD-10) International Personality Disorders Examination.

Stephen Blumberg (National Center for Health Statistics) asked about plans for using proxy respondents for some individuals with disorders who may be either unwilling or unable to participate in the survey. Colpe said that they do not anticipate a routine use of proxy responses, but they will track why people could not respond for themselves. In the case of the youth interviews, there will be some modules that the researchers would prefer to have the parents answer, such as those on insurance and income.

Kalton asked whether there are any concerns related to the potentially wide variation in the amount of time it will take respondents to complete the interviews. Colpe replied that they do anticipate this to vary greatly by respondent, and they have a cap to make sure it is not too long. They also planned the instrument layout so that the most important questions are at the beginning.

Regier noted that some disorders of high interest in the context of disability issues, such as schizophrenia and autism spectrum disorder, are not included in the new NIMH survey. He said that he would like to see attention given to these severe mental disorders by either clinical follow-up, subsampling, or screening, to obtain clinically valid diagnoses for these disorder areas. Regier went on to say that he has been concerned for years about the autism surveys of the Centers for Disease Control and Prevention (CDC) because the prevalence rates produced based on those data collections do not match what clinicians report. He said that the new NIMH initiative is a fantastic opportunity to take advantage of some of the more current knowledge about psychopathology and about using a more dimensional approach to psychopathology. However, he emphasized that not including the two areas of schizophrenia and autism spectrum disorder would be a missed opportunity to address the problem of not having better data on those disorders.

Colpe explained that there are a couple of ways they are exploring psychotic symptoms. The survey will be asking whether participants have a diagnosis of autism or schizophrenia. In addition, a psychotic symptom scale will be used in the clinical evaluation to determine where a person might manifest symptoms in terms of the spectrum of psychosis. For autism, the best the survey may be able to do is to identify that group and include them in one of the follow-up clinical components. She said that the survey will also include questions about family history for those disorders. Regier pointed out that one of the problems with autism spectrum disorder, in particular, is the incentive for families to say that they have a child on the autism spectrum, because this enables them to receive a range of services that would otherwise not be available to them. He thinks that this incentive may explain in part the currently observed higher autism prevalence rates.

Olfson agreed with Regier's concerns about the lack of data on both autism and schizophrenia; he noted that the reasons for the deficit in that area are real and very difficult to get around. In the case of schizophrenia, estimates obtained from surveys will not reflect the overall population rate unless people in various institutional settings are included, in addition to the household population. In other words, beyond the measurement challenges and the need for clinical judgment, there are also difficult sampling and statistical issues to consider.

Olfson asked Vos how credible community-based estimates of those disorders are obtained in the Global Burden of Disease study. Vos said that relying on household surveys is likely to grossly underestimate a number of mental and substance use disorders, including psychotic disorders, drug use disorders, and some of the childhood disorders like autism and Asperger's. For those conditions Vos said that the Global Burden of Disease study tends to rely on different data collection methods, and the researchers pay special attention to selection bias issues. For the drug use disorders, they use indirect estimates that combine survey data, mental health records, needle exchange program data, and judiciary data. For psychotic disorders, they rely on studies that explicitly sample from mental health services records. Individuals with these disorders are often not included in surveys because they are either homeless or institutionalized, and they are also much more likely to be nonresponders.

Schaeffer asked whether estimates based on some other sampling frame, such as a frame based on clinic or community services records, have been considered. Colpe replied that NIMH has not considered this yet, and she said she would like to get the workshop participants' thoughts about it. Along the same lines, Manderscheid asked whether Colpe is considering integrating a sampling frame based on the public mental health system in order to include people served by these systems. Colpe said that the National Comorbidity Survey Replication used a school frame for adolescents. In other programs, NIMH is working with states with the 5 percent set-aside for coordinated specialty care for first-episode psychosis. If the states develop systems to allow them to store that type of data, NIMH might be able to use those programs as a frame.

Regier reminded the group that the ECA did sample the institutionalized population of nursing homes, prisons, and long-stay psychiatric hospitals. In terms of prevalence rates for schizophrenia, they found that far more people with schizophrenia are being served in primary care than are being served in the institutions, although the most severely ill may be in institutions.

Still on the subject of sampling frames, Kilpatrick raised a point about the low level of long-term inpatient mental health care that is available, which results in a disproportionate number of people with mental illness being in prisons and jails before they are incarcerated or adjudicated. He also asked whether as part of identifying primary sampling units it would be possible to identify institutions such as jails, prisons, and long-term care facilities. Colpe replied that they are able to identify homeless shelters and similar facilities and include people living in those in the sampling.

Vos remarked that integrating these approaches within a household survey may not work. If the sampling units are small, including people in prisons or long-term institutions that happen to fall into the sample will not help. He suggested that for these kinds of disorders, separate endeavors may be needed where all available sources of information are consulted in a geographic area, in order to develop a sampling frame. These sources of information could include school-based services, disability services for autism spectrum disorders, mental health services, primary care facilities, institutions for psychotic disorders, and various service providers for drug use disorders.

Kilpatrick commented that it is important to think about whether excluding a population that is not typically included in household surveys would have a meaningful impact on prevalence estimates. For example, even if the rate of a disorder is much higher among the chronically homeless than the general population, the proportion of homeless relative to the total population might be so small that not including them would not affect the prevalence estimates. Furthermore, it is also important to consider how the data will be used beyond the purposes of prevalence estimates. For example, if knowing the number of individuals with serious mental disorders is necessary to plan for services, then knowing the concentration of risk in specific places may be more important than the prevalence rate for the nation.

Regier reminded the group of some of the advantages of the ECA study, which used the five areas of Baltimore, Durham, New Haven, Los

Angeles, and St. Louis as the catchment areas. In each catchment area, the researchers identified all the prisons, nursing homes, and hospitals that served the population of interest, so they knew they were capturing the service delivery system for each of those five catchment areas. The researchers sampled the institutions with the goal of combining those data with the data from the general population survey: with this approach, the study was able to address such disorders as schizophrenia and bipolar disorder. For some surveys in the future, the catchment area approach, in addition to a national survey, might be the strategy that enables researchers to link prevalence estimates with information about service use. Regier added that a disadvantage of some national surveys is that they do not include data on incidence or short-term acute disorders. The ECA collected data on 1-month prevalence and 1-year incidents. On the basis of that information, the researchers were able to identify information about disorder onset, offset, and duration.

Connie Citro (Committee on National Statistics) noted from Colpe's presentation that the NIMH study is going to collect information about military service and homelessness. She asked if the researchers had also considered asking questions about involvement with the criminal justice system, such as being on parole or having been incarcerated. Colpe responded that the survey will include such items as ever having been in jail and having been in jail the past year.

James Jackson (University of Michigan) pointed out that in the National Survey of American Life the researchers devised a methodology of estimating whether there were people in the household who had some attachment to the household but had a different living arrangement, such as being institutionalized or homeless, at the time of the interview. He said that those survey data are available for analysis.

Vos commented that this does not address the issue of the differential response rate across groups of people with different disorders. Typically, people with drug use disorders, psychotic symptoms, or autism, which are often associated with intellectual disability, are much more likely to be nonresponders. Even if the overall response rate for the survey is high, the majority of the people in the sample with specific disorders could be missing.

Potter mentioned that there are a couple of new data sources that could be useful in producing estimates of the institutional population. First, all nursing homes are required to submit extensive assessment data to the Centers for Medicare & Medicaid Services for all of their patients, regardless of payers. Data are submitted when a person is admitted, at 90 days, and when there is a change. The second data source is a new Medicare payment mechanism for collecting claims data for people who use inpatient psychiatric facilities.

Manderscheid added that it does not appear that the NIMH survey is designed in a way that can address the legal requirements that SAMHSA has for producing estimates of adults with serious mental illness. He urged NIMH to give some consideration to strength-based measures, which is a growing area of interest. The Healthy People 2020 initiative has made major investments in this area and will continue to do that over time.[5] There are opportunities to build some synergies on these topics. Colpe responded that the new NIMH survey is not expected to replace SAMHSA's existing procedures for estimating serious mental illness. For example, the NIMH study will produce national estimates, not the state-level estimates that are required of SAMHSA for administering the block grant programs. With respect to strength-based measures, Colpe said that her presentation did not include all of the measures planned for the survey, such as the Strength and Difficulties Questionnaire, which is planned to be used for youth. She added that there are other measures in the study that are not necessarily disorder based.

[5] See http://www.cdc.gov/nchs/healthy_people/hp2000.htm [December 2015].

3

Instruments Available for Measuring Specific Mental Illness Diagnoses with Functional Impairment

OVERVIEW OF EPIDEMIOLOGICAL STUDIES

Darrel Regier (Uniformed Services University) began his presentation by saying that he is pleased to see the very impressive efforts of the SAMHSA team to research the history of surveys and measures and reevaluate how to move forward with their data collections. This work builds on a rich tradition of updating epidemiological studies, which started with the advent of the DSM-III. Regier also commented that it was good to see that the survey program, which started with NIMH's Epidemiological Catchment Area (ECA) study and continued through the National Comorbidity Study (NCS) and the National Comorbidity Study Replication (NCS-R), has continued to develop.

Regier provided an overview of prevalence rates of different disorders across some of the earlier epidemiological studies that were also mentioned by Ron Manderscheid (see Chapter 2). The ECA study found that the diagnostic criteria were not congruent with what would be identified as treatment need and treatment use. As defined by the DSM-III at that time, about 28 percent of people had a disorder, and about 15 percent received some mental or addiction services: approximately 6 percent received specialty mental health services, 5 percent received general medical services, and 4 percent received other services. However, about one-half of those who were receiving services were not identified as having a mental or addictive disorder. Among the 28 percent of the population that had mental or addictive disorders in the ECA, the rate of those with a mental disorder only was 19 percent; comorbid mental and addictive disorders was 3 percent; and addictive disorder only was 6 percent.

Regier also discussed the Marshfield Primary Care Study, which predates the ECA and used the Research Diagnostic Criteria (RDC) and the Global Assessment Scale (GAS), which later became the Global Assessment of Functioning (GAF). This study found that about 28 percent in the primary care population had mental or addictive disorders, and a little over half of those, 15 percent, had a score of 70 or less on the GAS, which indicates minimal impairment. Excluding people with minimal impairment and those with less than minimal impairment from the analysis reduced the number of those who met criteria for an RDC disorder by half. Lowering the threshold for the GAS score to less than 60 resulted in an estimate of 10 percent, and lowering it to less than 50 resulted in an estimate of about 2 percent.

The results from the ECA were not consistent with data from the NCS, which estimated the annual prevalence rate of any mental or addictive disorder to be around 38 percent. Regier said that the two surveys were reconciled through the addition in the DSM-IV of the clinical significance criteria that are required for any mental disorder.[1] Scoring individual symptom areas in terms of their clinically significant distress involved asking several questions, including: Did it interfere with your life a lot? Did you ever take any medication for it? Did you ever talk with anybody about these symptoms? With these criteria in the DSM-IV, the prevalence rates in the ECA and NCS dropped to 18.5 percent for any mental or substance use disorder and 14.9 percent for any mental disorder.

Regier recounted that Manderscheid's earlier reference to the definition of severe mental disorders was occasioned by the National Advisory Mental Health Council being asked by Senator Pete Domenici to develop a study of the cost of parity insurance coverage for the severely mentally ill. In response to this, the council looked at the ECA data and specifically at disorders with psychotic symptoms (i.e., schizophrenia, schizoaffective disorder, manic depressive disorder, and autism) and severe forms of other disorders, including major depression, panic disorder, and obsessive compulsive disorder. Personality disorder was not included because it was not part of the Senate definition of severe mental disorders at that time. The council found the prevalence of those disorders to be 2.8 percent. However, those receiving any services in the ECA were only 1.7 percent, and those with a disorder lasting 1 year (the duration criterion that was mentioned by Manderscheid) were only 0.8 percent.

[1] Narrow, W.E., Rae, D.S., Robins, L.N., and Regier, D.A. (2002). Revised prevalence estimates of mental disorders in the United States: Using a clinical significance criterion to reconcile two surveys' estimates. *Archives of General Psychiatry, 59*(2), 115-123.

Regier said that they also looked at other sources of information, such as Social Security Administration (SSA) data for the Supplemental Security Income (SSI) and the Social Security Disability Insurance (SSDI) programs. At that time, 0.5 percent of the population was receiving SSI or SSDI for severe mental disorders, based on the SSA definition. Using the Wisconsin and New Hampshire definitions of the severely mentally ill, the data showed 0.4 percent under treatment for intermittent care and 0.1 percent for continuous care. For nursing home long-term hospitalization, data from the Center for Mental Health Services Client/Patient Sample Survey showed 0.05 percent for mental illness. In other words, the different ways of defining severe mental disorders varied and resulted in different prevalence rates.

One of the conclusions drawn from the ECA, the Marshfield Primary Care Study, and earlier research was that the GAF was the best predictor of service use, when compared with any specific diagnosis, even a diagnosis of schizophrenia. Consequently, the GAF was adopted for DSM-III-R, and then it remained as Axis 5 for the DSM-IV. Regier said that at the time it was the best measure of functioning, although it somewhat conflated symptoms, suicide risk, and impairment. The GAF was also widely used by insurance companies as a criterion for hospitalization: a GAF score of less than 60 was needed for admission, and a score of more than 60 for discharge. However, there were reliability challenges with the GAF, and it required a lot of training for administration.

Regier also discussed the developments of the Kessler six-item, eight-item, and ten-item scales (K6, K8, and K10), which were used in the National Comorbidity Study and other studies as a screener for psychopathology. He noted that these measures are more accurately described as distress measures, because the items assess anxiety and depression. Since the DSM-IV introduced the idea of clinically significant distress in its additional criteria, these measures have become the most well-validated distress measures in the field. They were also used by SAMHSA to obtain an estimate of severe mental illness based on various cut points, rather than administering the Composite International Diagnostic Interview to the entire sample, and to determine the prevalence of severe mental illness on the basis of the Senate definition.

Regier then discussed more recent developments associated with the DSM-5 and the need to develop dimensional measures. The two-question Patient Health Questionnaire (PHQ-2) had been recommended by the federal task force on prevention as a screener for depression. Based on the PHQ-9 screener, the PHQ-2 includes items on mood and interest. But the goal was to assess a range of domains besides depression, such as mood, anxiety, sleep disturbance, substance use, and suicide. The researchers also wanted cross-cutting Level 1 and Level 2 measures and a dimensional severity rating to be freely available for download online.[2]

In describing the cross-cutting measures, Regier said that they call attention to symptoms that are relevant to most psychiatric disorders, such as mood, anxiety, sleep disturbance, substance abuse, and suicide. These measures are self-administered and include 13 symptom domains for adults and 12 for children. The measures are brief, with one to three questions per symptom domain: they screen for important symptoms but are not specific screens for individual disorders. The Level 2 items are completed when the corresponding Level 1 item is endorsed as mild or greater for most but not all items. The Level 2 measures provide more detailed assessment of symptom domains, and they are largely based on long-standing, well-validated measures including the revision of the Swanson, Nolan, and Pelham [SNAP] Questionnaire (SNAP-4) for inattention; the National Institute on Drug Abuse (NIDA)-modified Alcohol, Smoking and Substance Involvement Screening Test (ASSIST) for substance abuse, and the Patient Reported Outcomes Measurement Information System (PROMIS) forms for anger, sleep disturbance, and emotional distress.

For documenting the severity of a specific disorder, the frequency and intensity of its component symptoms are assessed for individuals with either a diagnosis, those meeting full criteria, or an "other" specified diagnosis, especially a clinically significant syndrome, that does not meet diagnostic threshold. Some of the severity measures are clinician rated and some are patient rated.

THE WORLD HEALTH ORGANIZATION DISABILITY ASSESSMENT SCALE

Based on the work that has already been completed on the World Health Organization Disability Assessment Schedule (WHODAS), Regier said that this measure became the recommended assessment for disability in the DSM-5. The WHODAS corresponds to the disability domains of the International Classification of Functioning, Disability and Health, is developed for use in all clinical and general population groups, and was tested worldwide and in the DSM-5 field trials. The team that worked on the development of the WHODAS was composed of researchers at different WHODAS centers around the world. Researchers from NIMH, the National Institute on Alcohol Abuse and Alcoholism, and NIDA collaborated closely with researchers at the World Health Organization.

[2] See http://www.psychiatry.org/psychiatrists/practice/dsm/dsm-5/online-assessment-measures [January 2016].

Regier said that there are compelling arguments for using a measure of this type. Studies have shown that diagnosis alone fails to predict service needs,[3] length of hospitalization,[4] outcome of hospitalization,[5] receipt of disability benefits or work performance,[6] and social integration.[7] In contrast, diagnosis combined with disability can predict health service utilization,[8] outcome after hospitalization,[9] and work performance,[10] among other positive outcomes. The WHODAS has the advantage of being an internationally recognized classification of functioning, disability, and health that can be used for physical and mental disorders. It also has cross-cultural comparability, good psychometric properties, ease of use, and availability.

The WHODAS captures functioning in the domains of cognition, mobility, self-care, getting along, life activities, and participation. There are six questions in each domain, which produce a score and a disability profile. The full version is 36 items and provides the most detail, but there is also a 12-item version, which is used for brief assessments. There is also a hybrid version with 12 items to screen for problematic domains of functioning: on the basis of positive responses to those 12 items, respondents may be asked up to 24 additional questions. The WHODAS can be administered by interview or computerized adaptive testing (discussed below).

Regier commented that computerized adaptive testing will eventually enable researchers to develop efficient surveys using these measures. Large pools of data can be used to standardize the approach for different population groups, which is likely to be the direction of research in the future. There is a lot more work to be done, however. For example, the

[3] National Advisory Mental Health Council. (1993). Healthcare reform for Americans with severe mental illness: Report of the National Advisory Mental Health Council. *American Journal of Psychiatry, 150*, 1447-1465.

[4] McCrone, P., and Phelan, M. (1994). Diagnosis and length of psychiatric inpatient stay. *Psychological Medicine, 24*, 1025-1030.

[5] Rabinowitz, J., Modai, I., and Inbar-Saban, N. (1994). Understanding who improves after psychiatric hospitalization. *Acta Psychiatrica Scandidiavica, 89*, 152-158.

[6] Massel, H.K., Liberman, R.P., Mintz, J., and Jacobs, H.E. (1990). Evaluating the capacity to work of the mentally ill. *Psychiatry: Journal for the Study of Interpersonal Processes, 53*, 31-43.

[7] Ormel, J., Oldehinkel, T., Brilman, E., and van den Brink, W. (1993). Outcome of depression and anxiety care: A three wave 3~HF year study of psychopathology and disability. *Archives of General Psychiatry, 50*, 759-766.

[8] Ormel, J., Oldehinkel, T., Brilman, E., and van den Brink, W. (1993). Outcome of depression and anxiety care: A three wave 3~HF year study of psychopathology and disability, *Archives of General Psychiatry*, *50*, 759-766.

[9] Rabinowitz, J., Modai, I., and Inbar-Saban, N. (1994). Understanding who improves after psychiatric hospitalization. *Acta Psychiatrica Scandidiavica, 89*, 152-158.

[10] Massel, H.K., Liberman, R.P., Mintz, J., Jacobs, H.E., Rush, T.V., Giannini, C.A., and Zarate, R. (1990). Evaluating the capacity to work of the mentally ill. *Psychiatry: Journal for the Study of Interpersonal Processes, 53*, 31-43.

hybrid version of the WHODAS is not yet in the DSM-5, and neither are the adaptive testing versions of the measures in PROMIS.

Using the WHODAS, meaningful distinctions have been found among subgroups of people with mental health problems, alcohol problems, drug problems, physical health problems, and the general population. For example, people with mental health problems have greater disabilities on the domain of "understanding and communicating" in comparison with people who have physical health problems. People with mental health problems also show high levels of disabilities in the domains of "getting along with people," "work," "household functioning," and "participation with society."

Regier commented that it will be valuable to start to disaggregate the WHODAS into the different subscales and start associating these with the specific disorders because of the different profiles for different disorders. Exactly how these are going to inform clinical judgments and how to go forward is something that no one has studied yet. But it is an important developmental area that needs attention.

Several papers have been published that examine the WHODAS as part of the DSM-5 field trials. The January 1, 2013, volume of the *American Journal of Psychiatry* contains several papers on methodology and design,[11] reliability of the findings,[12] and outcomes from dimensional measures.[13] There have also been results published from the routine clinical practice field trials with participation by more than 600 psychiatrists, psychologists, social workers, counselors, and psychiatric nurses.[14]

Findings from the field tests at one of the sites in Houston underscored the problem with relying only on diagnoses. There was a very high proportion of comorbidity in persons with major depressive disorder, posttraumatic stress disorder, alcohol use disorder, and generalized anxiety disorder, which accounted for almost 70 percent of the patient

[11] Clarke, D.E., Narrow, W.E., Regier, D.A., Kuramoto, S.J., Kupfer, D.J., Kuhl, E.A., Greiner, L., and Kraemer, H.C. (2013). DSM-5 field trials in the United States and Canada, Part I: Study design, sampling strategy, implementation, and analytic approaches. *American Journal of Psychiatry, 170*(1), 43-58.

[12] Regier, D.A., Narrow, W.E., Clarke, D.E., Kraemer, H.C., Kuramato, S.J., Kuhl, E.A., and Kupfer, D.J. (2013). DSM-5 field trials in the United States and Canada, Part II: Test-retest reliability of selected categorical diagnoses. *American Journal of Psychiatry, 170*(1), 59-70.

[13] Narrow, W.E., Clarke, D.E., Kuramoto, S.J., Kraemer, H.C., Kupfer, D.J., Greiner, L., and Regier, D.A. (2013). DSM-5 field trials in the United States and Canada, Part III: Development and reliability testing of a cross-cutting symptom assessment for DSM-5. *American Journal of Psychiatry*, *170*(1), 71-82.

[14] Moscicki, E.K., Clarke, D.E., Kuramoto, S.J., Kraemer, H.C., Narrow, W.E., Kupfer, D.J., and Regier, D.A. (2013). Testing DSM-5 in routine clinical practice settings: Feasibility and clinical utility. *Psychiatric Services*, *64*(10), 952-960.

population at this site. Diagnoses other than those four accounted for 27 percent of that patient population.

CHALLENGES AND OPTIONS FOR SAMHSA

One question of interest to SAMHSA is whether disability can be measured for specific disorders: Regier said that this cannot be done. If someone has an impairment, especially across several physical and mental disorders, trying to disentangle the attribution—whether using WHODAS or any other measure—is very problematic. Neither an individual nor a clinician would know which diagnosis is causing the problems. After years of testing with the DSM-III, DSM-IV, and more recently with DSM-5, it has been found that there are no firm boundaries between disorders.

Genetic studies have revealed that there are common genetic vulnerabilities across the first four major disorder groups of neurodevelopmental disorders, including autism spectrum and attention deficit hyperactivity disorder, schizophrenia, bipolar disorder, and depression. There are a large number of genes, perhaps 1,000, for a disorder like schizophrenia, but they contribute a very small vulnerability risk for an individual getting schizophrenia. Whatever the combination of the genetic risk is with environmental exposures, the epigenetic influences that turn these genes on and off can produce an almost infinite number of combinations that are not going to break cleanly across the diagnostic boundaries.

The introduction to the DSM-5 manual advises looking at diagnoses primarily as central tendencies. The challenge for population surveys is to figure out how best to use a combination of the dimensional profiles, categorical diagnostic criteria, disability impairment measures, and severity measures in order to get at information that is clinically relevant and that will provide predictability in terms of clinical course, response to treatment, need for disability insurance coverage, or disability payments for the individual.

In his conclusion, Regier noted that WHODAS has the potential to draw attention to disability concepts in clinical settings and to better integrate them into routine practice. He also pointed out that preliminary analysis of mean WHODAS scores indicates some interesting age, diagnosis, and informant effects. He said that the self-report WHODAS appears to be very reliable, but the clinician assessment that uses the six-item WHODAS is not reliable. More items are needed in order to get a reliable estimate of the six domains than just one question for each domain. He also said that there is evidence for the validity of the WHODAS based on disability scores that are higher for patients with two disorders than for patients with one disorder, and based on total disability scores that are higher for specific disorders than for "not otherwise specified" disorders.

Regier emphasized that the field needs more epidemiological studies. He suggested that if NIMH and SAMHSA could meld this into their ongoing research programs, it would be possible to produce population rates for some of the disability measures for the full range of disorders. Future research is also needed to understand what the potential of the WHODAS is to affect clinical care, assist clinical decision making, improve patient care outcomes, and enhance patients' involvement in their own care.

In relation to SAMHSA's challenge to link impairment with specific diagnoses, Mark Olfson (Columbia University) asked about whether statistical methods, such as factor analysis or path analysis, could be used in population studies to examine the extent to which the variation in impairment or disability is accounted for uniquely by disorders. Regier replied that there is value in understanding the statistical associations between the disorders and the level of disability. However, analysis is also needed for assessing the severity of those disorders. It would be useful to understand how much of the disability is associated with the severity of the individual disorders and how much of the disability is associated with the comorbidity with other disorders. He said that if all of the variables are available in a dataset, then statistically it would be possible to examine what happens to disability when severity or comorbidity are added to the analysis. Dean Kilpatrick (Medical University of South Carolina) agreed that asking someone to attribute their disability to a specific disorder may not be feasible, and that it is best to measure disability in functional areas. He also agreed that it may be possible to sort some of this out in the analysis stages.

4

Data Collection Approaches

MEASURING MENTAL AND SUBSTANCE USE DISORDERS IN THE GLOBAL BURDEN OF DISEASE STUDY

Theo Vos (University of Washington) discussed the Global Burden of Disease (GBD) study, focusing on estimates of mental and substance use disorders. He described the GBD as a systematic scientific effort to quantify the comparative magnitude of health loss due to diseases, injuries, and risk factors by age, sex, and geographic areas for specific points in time. The emphasis is on the concept of health loss, rather than a broader concept of general welfare loss.

Vos explained that the GBD is based on three key principles: (1) everyone deserves to live a long life in full health; (2) searching for answers to what is preventing people from achieving that goal; and (3) mapping out a comprehensive picture of what disables and kills people across countries, time, ages, and by gender. The GBD measures health loss for a population in comparison with a reference or a normative goal for a population that is living in full health with a life expectancy at birth of 86 years. As derived from a life table, at age 86 one still has remaining life expectancy. Even at age 105, according to the standard life table, one can expect 1½ additional years of life.

Vos explained further that the GBD values disabling consequences equally across countries and over time: this explicit egalitarian approach is used because the goal for everyone is the same, no matter where they live. The GBD's main measure is the disability-adjusted life year (DALY), which combines mortality and disability information in a time metric.

Mortality is translated into years of life lost due to premature mortality, calculated by number of deaths multiplied by the remaining life expectancy from the standard. For disability, the GBD estimates the years of life lost due to time lived in states of less than full health: it is calculated by the prevalence of diseases and all the major disabling consequences of diseases, which are termed sequelae. Each sequela is multiplied by a severity weight between 0 and 1. This indicates the relative severity of that particular sequela relative to all other sequelae, and is anchored by zero disability (full health) to 100 percent loss of health at death.

The GBD study, commissioned by the World Bank, started in the early 1990s. Since 1997, it has been funded largely by the Bill & Melinda Gates Foundation, along with other sources of funding. The study design underwent several ad hoc revisions and updates in the 1990s. These revisions led to reduced internal consistency between estimates, which caused the team to revisit the methods in the 2000s. A capstone paper based on GBD was published in 2015,[1] which represented a commitment for future annual updates of all of the data. The version dubbed GBD 2015 is under way and is expected to be published in the first half of 2016. Vos emphasized that the GBD is a "public good" that is the work of a network of over 1,400 recognized international collaborators with representation by 106 countries.

The GBD currently examines approximately 320 disease and injury categories. The number of sequelae is well over 2,000. A very large chunk of the latter are the sequelae related to the causes and nature of various injuries. For example, fractures, contusions, and head injury fall into 48 categories, but they are multiplied by 27 causes of injury. Currently, 78 individual risk factors or combinations of risk factors are taken into account.

Estimates are made for 188 countries in the world. Increasingly, subnational estimates are also produced based on the data to meet policy needs. For the GBD 2013, subnational estimates were provided for provinces in China, states in Mexico, and 11 delineations of the United Kingdom. For the current version they are making estimates by state for the United States, by prefecture for Japan, by province for South Africa, by district for Kenya, and by state for Brazil.

Vos provided an overview of the data collection and estimation steps, which start with a demographic component that estimates the total level

[1] Global Burden of Disease Study 2013 Collaborators. (2015). Global, regional, and national incidence, prevalence, and years lived with disability for 301 acute and chronic diseases and injuries in 188 countries, 1990-2013: A systematic analysis for the Global Burden of Disease study 2013. *The Lancet, 386*(9995), 743-800. Available: http://www.thelancet.com/pdfs/journals/lancet/PIIS0140-6736(15)60692-4.pdf [February 2016].

of mortality regardless of cause and yields the total number of deaths. Cause-of-death information is collected from vital registrations in an increasing number of countries, with increasing levels of completeness and better quality. These data are supplemented with verbal autopsy information from many countries that do not have functional vital registration systems. Police records and mortuary records are also used. Vos said that a significant effort is involved in cleaning the data from different sources and editing them for consistency in order to make them usable in statistical modeling. An additional step involves scaling all of the individual cases by demographics and time to provide estimates of years of life lost. Another major task involves attributing diseases to underlying risk factors or combinations of risk factors.

Producing the disability estimates involves deriving the disability weights and combining them with severity distributions, because many of the major disabling conditions have a very wide range of severity. Estimates of years lived with disability are produced after a comorbidity simulation is performed. The team also conducts systematic reviews of epidemiological parameters, largely incidence and prevalence, but also "risk of death," and remission, which is defined as a cure rate, meaning an individual is totally free of disease and severity distributions. A tool was specifically developed for these purposes called DisMod-MR. MR (meta-regression) uses Bayesian statistical techniques, in which *fixed effects* on the study level characteristics facilitate cross-walking between different recall periods, different instruments, or different case definitions used in the studies in various sites. Fixed effects are also put on any country-level covariates that help predict the estimates: for instance, the per capita alcohol consumption in a country for alcohol use disorders. A hierarchy of *random effects* is used in which countries are grouped into 21 world regions on the basis of their geography and epidemiological profiles. Those regions are then grouped into seven super-regions.

Disability weights are gathered from nine population surveys and an open-access Internet survey using pair-wise comparisons. Vos said that finding adequate data to use for severity distributions has been one of the bigger challenges for the GBD. For some conditions, comparable information can be obtained from systematic review and using meta-analysis techniques. Mental disorders and musculoskeletal disorders make up more than 50 percent of the estimates of disability, but finding enough comparable information for analysis is the most challenging for these disorders.

As an example, Vos pointed to the estimates produced for major depression among females in the United States, for six time periods. Figure 4-1 shows the prevalence estimates by age groups, with the uncertainty coming from the DisMod-MR tool. The vertical bars are the 95

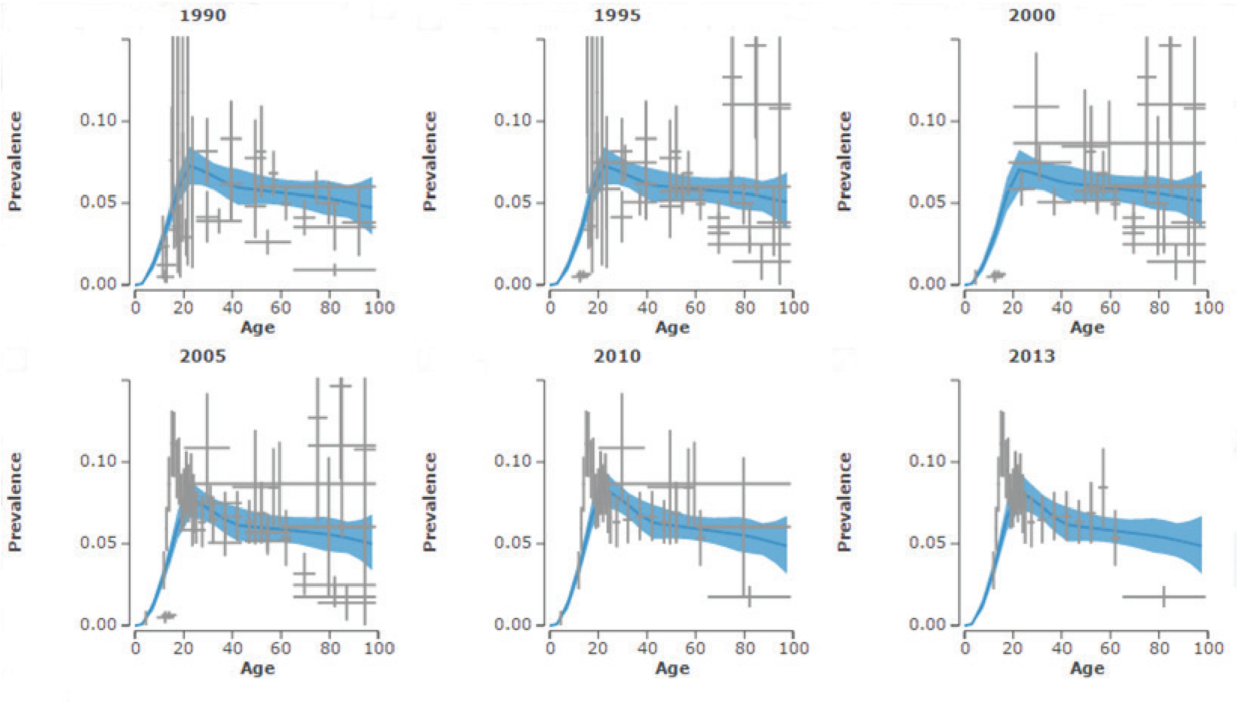


FIGURE 4-1 Estimates of major depression among females in the United States, 1990-2013.
SOURCES: Workshop presentation by Theo Vos, September 2015; data from the Global Burden of Disease study.
NOTE: Global Burden of Disease data and the analysis and visualization tools demonstrated during the workshop are available at http://www.healthdata.org/gbd/data [December 2015].

percent confidence intervals for the data points. The figure illustrates that even after adjusting for the differences in study-level covariates, there are still wide ranges of values, and there are some very high and very low estimates. The next step for the researchers is to try to statistically identify the best estimate, based on all of the available information.

Discussing disability weights in further detail, Vos explained that, in constructing the DALY, disability weights are the bridge between mortality and nonfatal outcomes. To measure health loss from nonfatal outcomes, weights were needed for all of the sequelae defined for the 300+ disorders with nonfatal outcomes. A parsimonious set of 235 health states were determined to cover all of the sequelae. Weights were then established that quantify the severity of outcomes as a percentage reduction from perfect health. For example, if the weight for blindness is 0.2 (i.e., a health loss of 20% resulting from blindness), then, according to the metric, five people living with blindness in a year are equivalent to 1 year of life lost due to disability.

Vos then described the data that provided the estimates for the disability weights. In the GBD 2010, information was collected from more than 30,000 respondents through face-to-face interview surveys in Bangladesh, Indonesia, Peru, and Tanzania; a telephone survey in the United States; and an open-access web survey.[2] The survey included 108 of the then 220 health states, but all of the health states were included in the web survey. For the GBD 2013, information came from four European surveys.

Since the GBD 2010, the primary mode in which the responses are elicited is by paired comparisons, in order to make fielding these sorts of questions as simple as possible. Respondents are presented with two descriptions of hypothetical people, each with a randomly selected health state, and then are asked which of the two hypothetical people is the healthier. The questions are chosen for relative ease of administration, comprehension, and analysis. The researchers found that high literacy and numeracy levels were not essential in order for respondents to be able to answer the questions. The health states are presented with a lay description, but without a label to avoid some of the stigma associated with some disorders, such as epilepsy or AIDS.[3]

The estimated disability weights in the GBD 2010 survey showed considerable consistency among the results in the different sites, with

[2] Salomon, J.A., Vos, T., Hogan, D.R., Gagnon, M., Naghavi, M., Mokdad, A., . . . Murray, C.J.L. (2012). Common values in assessing health outcomes from disease and injury: Disability weights measurement study for the Global Burden of Disease study 2010. *The Lancet, 380*(9859), 2129-2143.

[3] For further details on how the paired comparison questions are asked, see Salomon et al. (2012).

perhaps the exception of Bangladesh, for which the data were "noisier." Average educational attainment varied across the sites, and it was particularly high in the web survey, which involved a convenience sample that included many respondents with college degrees or higher, but there was still remarkable consistency in the estimated weight results. Further statistical analysis validated the consistency in how people respond to the framing of the questions and the methods that were used, across different geographies.

In examining disability weights across various disorders, which ranged from 0 (no disability) to 1.0 (maximum disability), the researchers found the highest weight for the psychotic phase of schizophrenia (0.77) and the lowest for mild anemia (0.004). The majority of health states were at the lower end of the scale: this finding is important because many of the health states are conditions with high prevalence. The severity of conditions showed internal consistency. For example, the values for depression were 0.16 for mild, 0.41 for moderate, and 0.67 for severe.

Vos noted that the disability weights for the different health states have to be combined with epidemiological data to yield distributions across the different sequelae. Table 4-1 shows the most recent top 20 ranking of disabling conditions globally, with varying levels of severity.

Vos showed the workshop participants one of the GBD's online data visualization tools with 2013 data years lived with disability, which is the disability component of the DALY. It showed that mental and substance use disorders represent a considerable proportion of the overall burden (21.2%) in the United States, with a very small annual rate of change between 1990 and 2013. He encouraged the workshop participants to explore the data for greater detail on specific disorders, through the GBD study website.[4]

Vos explained that, for some conditions, there are standard ways of describing severity. For example, vision impairment can be measured by defined thresholds of visual acuity, which is then mapped onto different levels of severity. For a number of diseases, people have converged on using similar sorts of methods over time, such as the Hoehn and Yahr[5] classification for Parkinson's. However, for many disorders, such as dementia, Vos and his colleagues had to search across several studies to get a consistent breakdown by severity for that disorder. Meta-analysis

[4] Global Burden of Disease data are available from http://www.healthdata.org/gbd/data. Also see http://vizhub.healthdata.org/gbd-compare/ and http://ihmeuw.org/3qad [December 2015].

[5] Hoehn, M.M., and Yahr, M.D. (1967). Parkinsonism: Onset, progression and mortality. *Neurology, 17*, 427-442.

TABLE 4-1 Top 20 Ranking, Globally, of Disabling Conditions with Levels of Severity, 2013

Disabling Conditions		Severity
1.	Low back pain	8
2.	Major depression	4
3.	Iron deficiency anemia	3
4.	Neck pain	4
5.	Other hearing loss	12
6.	Diabetes	By sequelae
7.	Migraine	2
8.	Chronic obstructive pulmonary disease	4
9.	Anxiety disorders	4
10.	Other musculoskeletal	7
11.	Schizophrenia	2
12.	Fall	By sequelae
13.	Osteoarthritis	3
14.	Refraction and accommodation	3
15.	Asthma	4
16.	Bipolar disorder	3
17.	Dysthymia	2
18.	Medication headache	2
19.	Dermatitis	8
20.	Other mental and substance	4

SOURCES: Workshop presentation by Theo Vos, September 2015; data from the Global Burden of Disease study.

techniques were then used to pool the data on the proportions with mild, moderate, and severe dementia.

For the severity of mental disorders, Vos said he and his colleagues had hoped to obtain comparable information from the World Mental Health Surveys (WMHS), but the data from the Sheehan Disability Scale, which is used by the WMHS, was skewed toward higher proportions of severe disease. There were similar problems in obtaining comparable data for severity in the case of several of the other disabling conditions, such as chronic obstructive pulmonary disease, asthma, osteoarthritis, back pain, and neck pain. Because of this, the GBD research team turned to three surveys with data available at the individual level, along with rich diagnostic information on mental disorders and physical disorders and a general health status measure. In this case, the common measure was the 12-Item Short Form Health Survey (SF-12), which has mental and physical health items. Vos noted that the drawback of the SF-12 is that it is not a freely available measure, but the cost for its use is relatively low.

The SF-12 has been embedded in the Medical Expenditure Panel Survey (MEPS) since 2000 and two waves of the National Epidemiological

Survey on Alcohol and Related Conditions (NESARC) from 2001-2002 and 2004-2005, as well as the Australian Mental Health Survey in 1997. A translation of SF-12 summary scores into the GBD disability weights was derived from small surveys asking individuals to fill in SF-12 based on GBD health as presented with their "lay description." Thus, for each individual in the survey a measure of the total amount of disability experienced could be derived.

To make estimates for specific disorders, the researchers statistically parsed out the aggregate disability in people who have more than one condition to determine what component of that disability was contributed by each disorder of interest. The GBD relies on the notion that disability is multiplicative rather than additive. If a person has a condition with a severe disability of 0.7 and another condition with a disability of 0.4, then the aggregate is 0.72, calculated as 1– [(1 – 0.7) × (1 – 0.4)].

For each individual, the GBD derives the contribution to the overall disability from each individual disorder. To be consistent with the mapping of all severity distributions and distributions by sequela, the researchers use a continuous measure artificially derived from the SF-12 information to determine the proportions of people who have severe, mild, and moderate disease. The researchers also estimate explicitly the proportion of people who have no disability that can be attributed to the underlying disorder of interest.

Vos stated that the team has learned that not only is it important to take comorbidity into account in the overall estimates of disability, but also that when severity is analyzed, it is necessary to tease out what is contributed to disability by comorbidity and what is contributed by an individual disorder. An advantage of this method is that it can be used consistently across a range of prevalent disabling outcomes, ranging from back pain to mental disorders. The limitation of this method was that the data used were from two high-income countries and that data from the NESARC and the Australian Mental Health Survey are relatively old. It is not clear whether this approach would work equally well in other countries.

Vos further noted that another big disadvantage of having very limited high-quality data on severity is that the GBD cannot capture variations in severity distribution over time that could be attributed to treatment effects. For instance, it cannot capture recent advances in medications for rheumatoid arthritis, which are making a big difference in the severity of symptoms for people with this condition.

Vos concluded that significant progress could be made if greater consistency can be achieved in the way studies measure different levels of severity for major disabling conditions. He encouraged the use of similar methods that can better capture all aspects of diseases of interest and the use of generic data collection instruments that can provide comparable information across a wide range of diseases. He acknowledged that the GBD approach on severity is limited in its ability to address the need for data on conditions that are not easily captured based on self-report, such as some of the childhood conditions or health states that limit people's capacity to respond. Possible ways around this include proxy reports and clinician-administered tools.

USING ADMINISTRATIVE DATA, GENERAL HEALTH SURVEYS, AND PRACTICE-BASED SURVEYS

Mark Olfson (Columbia University) focused on sources of data that differ from the household surveys that were discussed by other presenters. His goal was to discuss how these alternative data sources might be able to contribute to the understanding of mental disorders in the United States and to provide an overview of their limitations, as well as their strengths.

Olfson divided the datasets he covered into three groups: administrative datasets, population-based surveys, and provider-based surveys. He defined administrative datasets as databases that are generated as a by-product of medical billing. These datasets are not assembled with the idea that they are going to provide insights into the prevalence of mental disorders, but it is possible that researchers can obtain some information from them. Although others have discussed population surveys that collect data about mental disorders, Olfson concentrated on other surveys that contain some information about behavioral health, general health, and health care. The provider-based surveys discussed have information about the populations served by health care providers and the services that are delivered.

Administrative Data

As noted during the earlier session (see Chapter 2), electronic health records are limited to the population that uses the health care services. Olfson emphasized that this is true for all encounter data and claims data: it is an important limitation and one that is easy to lose sight of. However, in some cases this limitation may be less of a concern, if there is some evidence that a large proportion of a population of interest uses the services captured in the database, especially if the population is difficult to reach in other ways. For example, this situation might be the case for people with schizophrenia and other psychotic disorders.

It is also important to keep in mind that the diagnoses appearing in encounter databases have been generated by clinicians in the field. As noted by others during the workshop, there are concerns about the reliability of these diagnoses in comparison with diagnoses that result from applying the DSM criteria in a more systematic way, as would be the case in a structured interview as part of a research study.

One of the potential strengths of administrative data is that there is typically a well-defined sample. Researchers know who is enrolled and typically also have relevant dates associated with each record. In addition to diagnoses, administrative databases also include information about treatment, which allows for an examination of changes in rates of disorders or the natural course of service use over time. This information is more precise than is the information available from population surveys. Although the populations change over time, the databases can provide a steady stream of information.

As has been discussed earlier, the mental disorders of people who do not receive treatment for them are not included in these databases, and the coverage rates vary by disorder. For disorders like circumscribed phobia or alcohol abuse, only 10 or 20 percent of the affected population may be receiving treatment during the course of one year. As pointed out by Regier, only about one-half of the people who meet the criteria for mental disorder on the basis of a structured interview will receive some mental health services in the course of a year, either from the specialized mental health sector or from the general medical sector.

When considering the mental disorders of people who are missing from these claims databases, it is important to recognize that they are not missing at random. The mental disorders that are captured in the claims data tend to be more severe, and people who are included tend to have other characteristics that are associated with treatment seeking.

For example, one study found that among people with diabetes who meet depression threshold with a Patient Health Questionnaire (PHQ) score of at least 10, about one-half were clinically detected and would be represented in a claims database in the course of a year, and the other half were clinically undetected and thus would not be in a claims database.[6] The disorders that are present in a claims database tend to be associated with people who are somewhat younger, somewhat more severely depressed, and have more comorbidity, with a higher rate of panic attacks than those not in the claims database. In other words, those disorders included in the claims database are not representative of the disorders in the overall population.

Another limitation is illustrated by a study using data from local

[6] Katon, W.J., Simon, G., Russo, J., Von Korff, M., Lin, E.H., Ludman, E., Ciechanowski, P., and Bush, T. (2004). Quality of depression care in a population-based sample of patients with diabetes and major depression. *Medical Care, 42*(12), 1222-1229.

clinics in Pittsburgh.[7] This study involved administering the Structured Clinical Interview for DSM (SCID) to patients and comparing them with chart diagnoses. The agreement was low, with kappa values in the slight to fair range, even when using broad diagnosis criteria. For more refined diagnoses, such as type 2 bipolar disorder in remission, the kappa values would be especially low. Olfson reiterated that it is important to be cautious about assuming that diagnoses appearing in claims databases are the same as the diagnoses obtained from administering a structured psychiatric interview the way it is done in a specialized community epidemiological survey.

Olfson said that there are a wide variety of different types of administrative databases that include both public payers (e.g., Medicaid, Medicare, Veterans Health Administration/Tricare) and commercial insurance (e.g., MarketScan, Health Care Cost Institute, IMS Pharmetrics). The databases share some general characteristics, but they differ in terms of how they are generated, aspects of their basic structure, and therefore their fundamental strengths and weaknesses.

Medicare and Medicaid Data

One of the most important administrative databases in the United States, particularly for adults and young people with more serious psychiatric disorders, is generated by the Medicaid system. Medicaid is the largest public payer for mental health services in the United States, covering about 60 million people. Even though the Medicaid system is quite large, Medicaid beneficiaries are very different from the overall population in terms of the burden of psychiatric disorders. A large proportion of people with severe mental illness are in the Medicaid program. Their eligibility is often due to their disability, not because of poverty, although many of them also have low incomes.

Olfson described a study that used MEPS data to compare the sources of health care coverage for people with schizophrenia: it found that 67 percent reported that they had coverage from Medicaid, 46 percent from Medicare, and 15 percent from private health insurance for at least one day during a given year.[8] These rates are very different from the general population. As discussed throughout the workshop, people with

[7] Shear, M.K., Greeno, C., Kang, J., Ludewig, D., Frank, E., Swartz, H.A., and Hanekamp, M. (2000). Diagnosis of nonpsychotic patients in community clinics. *American Journal of Psychiatry, 157*(4), 581-587.

[8] Khaykin, E., Eaton, W., Ford, E., Anthony, C.B., and Daumit, G.L. (2010). Health insurance coverage among persons with schizophrenia in the United States. *Psychiatric Services, 61*(8), 830-834.

schizophrenia are a difficult population to reach with traditional surveys because of nonresponse and because the prevalence rate is relatively low. However, a great majority of adults with schizophrenia are included in the Medicaid and Medicare databases, which perhaps represents another approach counting persons with mental illness that are missing from community surveys.

Commercial Insurance Administrative Data

The second major source of administrative claims data is commercial insurance databases, which generally have similar structures. Olfson said that their weaknesses and limitations are similar to those of the Medicaid and Medicare databases, in terms of only having information about treated populations, that is, not representing the entire population. They also have further limitations. They tend not to collect data on race and ethnicity, and they are generally not as rich in other demographic and geographic information as are the Medicaid and Medicare databases. It may also be more difficult to link them to other data sources. On the positive side, data from commercial insurance databases tend to be available more rapidly than the national Medicare data, which have a lag of about 2 years, and national Medicaid data, which have a lag of 3-4 years.

Summarizing the strengths of administrative claims data in comparison with population surveys, Olfson pointed out that these sources have fewer problems related to nonresponse and the coverage of difficult-to-survey populations. The data are also less susceptible to response bias. They are not dependent on respondent recall, because they are essentially archival information about visits that occurred and the diagnoses that were assigned. Administrative records are also less susceptible to self-report bias due to stigma, although social processes govern who accesses care and sometimes what diagnoses are entered into claims databases for the purposes of reimbursement. Olfson also reiterated that administrative claims databases have the strengths of being able to provide information about the diagnoses and treatment patterns of some difficult-to-survey populations and, for some of them, of being quite large.

Data from Population Surveys

Olfson then turned to several ongoing, federally funded general health surveys that can be a source of mental health data. He said that, although these surveys cannot produce a precise estimate of the prevalence of individual disorders, they have some information about either treated disorders or distress.

National Health Interview Survey

He first briefly mentioned the National Health Interview Survey (NHIS), which was covered in further detail in the presentation of Stephen Blumberg (National Center for Health Statistics). Olfson noted that the survey, which has been administered continuously since 1957, has a respectable response rate and provides an opportunity for characterizing some aspects of distress through data collected with the K6 instrument (see above). In some years, the NHIS also includes other items that are of interest and relevance to mental health, such as whether a person has ever been told that he or she has bipolar disorder, schizophrenia, mania, or psychosis, as well as a history of mental health care or counseling.

The Behavioral Risk Factor Surveillance System

Box 4-1 provides an overview of another health population survey, the Behavioral Risk Factor Surveillance System (BRFSS), which Olfson pointed out, has the distinct benefit of providing state-level estimates. Very few surveys, other than the National Survey of Drug Use and Health (NSDUH), do so, and for some policy purposes it serves an important function. The BRFSS is a telephone survey, and like most telephone-based surveys, it has a low response rate. In addition to the standard set of questions, states can elect to have additional optional questions administered. The box lists some of these optional state modules that have been administered at different points in time, including measures of depression (PHQ-8), distress (K6), diagnoses of anxiety and depression, and treatment related to mental health condition. Olfson noted that the PHQ-8 is the same instrument as the PHQ-9, but without the ninth item that asks about self-harm and suicide.

The National Health and Nutrition Examination Survey

Olfson then briefly highlighted the National Health and Nutrition Examination Survey (NHANES), which involves mobile survey units sent to local communities. The survey produces national estimates based on interviews with approximately 5,000 adults over two administration periods. Box 4-2 provides an overview of the NHANES and shows the mental health information that is collected as part of the survey. This includes the PHQ-9, which gives a more robust measure of depression with the inclusion of the ninth item on suicidality. Olfson added that the NHANES has a richer array of general medical information and physical health information than many other federal surveys. He pointed out that the NHANES has a rather narrow range of person-level psychopathology

BOX 4-1
Behavioral Risk Factor Surveillance System (BRFSS)

Size: 450,000 adults/year

Informant: Individual self-report

Design: State-based, telephone survey, state-based weights forced to U.S. population

Mental health information (optional state modules):

- PHQ-8 [Patient Health Questionnaire] (45 states in 2006 and 2008, 12 in 2010)
- Lifetime diagnosis of anxiety, lifetime diagnosis of depression (2006, 2008, 2010)
- K6: past 30 days (2007, 2009) (37 states)
- Mentally unhealthy days in last 30 days: 1 item, 50 states (2007, 2009)
- Treatment related to mental health condition: 1 item, 50 states (2007, 2009)

Strengths:

- State-level estimates
- Large sample size

Limitations:

- Few years
- Did not cover cellular phones before 2011 or those without phones
- Recall bias and social desirability effects
- Low response rates

SOURCE: Workshop presentation by Mark Olfson, September 2015.

data, although different modules have been used in previous years. The structure of the survey allows researchers to combine several years of data to accrue a larger sample and derive more stable estimates.

The Medical Expenditure Panel Survey

Olfson next described the MEPS, noting that Vos had also discussed it as one of the datasets used in the Global Burden of Disease study (see above). Olfson pointed out that the MEPS has elements of interest to mental health services researchers. As shown in Box 4-3, the MEPS is an annual community-based survey, with a sample based on the NHIS. Household respondents report on family members' service use. The survey also includes the SF-12, PHQ-2, and K6. Outpatient service data from the MEPS have shown reasonable psychometric properties when compared with confirmed diagnoses in primary care samples.

Olfson said that estimates are derived every year and that the survey provides valuable trend data. One of the problems that the MEPS has struggled with, as other surveys, is decreasing response rates. Because of the gradual decline in the response rates, there are questions about whether results are as representative as they were in years past.

Olfson reiterated that he was only covering basic information about the population surveys he discussed, but as potential sources of existing data, they have several strengths: they are representative of the general population; unlike the administrative data sources, they yield information on untreated individuals; they are typically administered on an annual basis; and they can be analyzed cross-sectionally and over time. The drawbacks are that they do not have large sections dedicated to the assessment of mental health; they tend to cover only household populations, not people who are institutionalized in various settings; and their response rates have been declining.

BOX 4-2
National Health and Nutrition Examination Survey (NHANES)

Size: 5,000 adults/year

Informant: Individual self-report, physical health examination, lab testing

Design: Cross sectional, complex sampling, noninstitutionalized population

Mental health information:

- PHQ-9 [Patient Health Questionnaire], sleep disorders questionnaire, smoking status, mentally unhealthy days
- Prescribed medications past month
- Generalized anxiety disorder, panic disorder for young adults aged 20-39 (1999-2004)

Strengths:

- Nationally representative sample, acceptable response rate (69.5%, 2011-2012)
- Wealth of physical health data

Limitations:

- Small sample size
- Limited mental health information
- No expert validation of depression

SOURCE: Workshop presentation by Mark Olfson, September 2015.

BOX 4-3
Medical Expenditure Panel Survey (MEPS)

Size: Approximately 14,000 families, 35,000 persons (household component)

Design: Complex, household population, panels followed for up to 2 years

Mental-health related variables:

- Conditions
- Psychotropic medication purchases
- Psychotherapy/counseling visits
- Visits to mental health specialists
- Activity limitations
  –SF-12 Mental Component Summary (adult self-report)
  –Patient Health Questionnaire-2 (adult self-report)
  –K6 (adult self-report)

Strengths:

- Nationally representative, continuous sample
- Three interviews per year

Limitations:

- Modest response rate: 56.3% (2012)
- Household informant, except for SF-12, PHQ-2, and K6
- No systematic mental health status information

SOURCE: Workshop presentation by Mark Olfson, September 2015.

Practice-Based Surveys and Data

In the final part of his presentation, Olfson discussed practice-based surveys and data, which include information from health care providers. He said that the most well-known and well-trodden of these surveys is the National Ambulatory Medical Care Survey (NAMCS), which is focused primarily on office-based medical practice: see Box 4-4. Outpatient visits in hospitals and other emergency departments are captured in a companion survey, the National Hospital Ambulatory Medical Care Survey (NHAMCS). Personnel in the practices complete forms that describe the characteristics of the practice and provide data about visits in a particular sampling week. Information is provided on the reason for the visits, what the diagnoses were, and the treatment and services provided. Psychiatrists and other medical specialists are also included, in addition to general medicine and primary care providers. However, he noted, community mental health centers, substance abuse clinics, and other specialty outpatient clinics are not covered in these data, which means that persons with substance use disorders and severe mental disorders would only appear in small numbers. The unit of analysis in these datasets is the visit, and one metric that can be derived from the NAMCS is visits per population. The difference between visits and person prevalence is an important distinction because people consume multiple visits in the course of a year, and there can be some duplication of individuals. It is tempting to think about the data as treated prevalence rates, but it is important to remember that they are not.

Olfson said that these databases in some ways resemble the administrative data discussed previously in that they contain similar encounter data but are based on the abstracts and include all payers. They provide a more robust look at outpatient care and patterns of diagnoses over time nationally, rather than the commercial administrative databases that cover different groups of insurers.

The NAMCS has recently been redesigned and is now about twice the size it used to be in terms of the number of visits, and state-level esti-

BOX 4-4
National Ambulatory Medical Care Survey (NAMCS)

Size: 30,000 visits/year (1993-2010), 76,000 (2012)

Design: Office-based physician visits during sampling week, complex design

Mental health information:
- Mental health reasons for visit
- Clinical diagnoses
- Medications prescribed or monitored
- Psychotherapy/counseling
- Depression regardless of diagnosis (2005-2010, 2012)
- Includes visits to psychiatrists

Strengths:
- Covers all payers
- Measures mental illness burden in office-based practice
- Trend analyses possible

Limitations:
- Counts visits, not unduplicated patients
- Modest to low response rate (60.6%, 2005-2010; 38.4%, 2012)
- Does not capture outpatient care provided in community mental health centers, substance abuse clinics, and other specialty outpatient settings

SOURCE: Workshop presentation by Mark Olfson, September 2015.

mates can now be derived. However, the response rate for the survey has fallen by almost one-half. Olfson believes this drop is a significant concern because of the risks of selection bias that can occur with low response rates. The 2011 data are not yet available, but they will become available in the next few months.

The final database Olfson described was the Hospital Cost and Utilization Project National (Nationwide) Inpatient Sample (HCUP-NIS): see Box 4-5. The HCUP-NIS is not a survey, but a compilation of discharge summary abstracts. This data collection has also been recently redesigned, and it now represents a 20 percent national sample of all the discharges from hospitals. The unit of analysis is each hospital discharge, not each unique patient, so there is also a potential for duplication of individuals in the data. The database contains only sparse information on characteristics of the population, but it does include diagnoses and procedures

BOX 4-5
Hospital Cost and Utilization Project National (Nationwide) Inpatient Sample (HCUP-NIS)

Size: 8 million discharges from approximately 1,000 hospitals (annually)

Scope: Nonfederal, short-term general and other specialty hospitals

Design: 1988-2011 (participating states, weighted by hospital ownership, size, teaching status, location, region), 2012 (20% national sample of discharges, community hospitals)

Mental health information:
- Discharge diagnoses
- Procedures
- Disease severity measures (based on diagnoses, demographics, length of stay)

Strengths:
- Large sample size
- Covers all payers and includes uninsured patients
- National estimates
- Can be used to analyze trends

Limitations:
- Counts discharges, not unduplicated individuals
- Limited clinical information
- Does not include psychiatric hospitals, alcoholism, or chemical-dependency treatment facilities

SOURCE: Workshop presentation by Mark Olfson, September 2015.

that have been delivered in inpatient settings. The database also includes a disease severity measure that is derived from an algorithm using diagnoses, demographics, and length of stay. This four-point scale measure of severity allows comparisons across different diagnostic groups. Olfson noted that the HCUP-NIS would yield information on a very small slice of psychopathology in the community, but it is useful to consider it as part of the data sources available.

Olfson stressed that none of the databases he had described were designed for use in estimating prevalence of mental disorders in the population. Nonetheless, they can yield some insights on the topic. In particular, he said, they can be of value with regard to trends in treatment because their structures are largely conserved over time. He concluded by listing the elements that he would ideally want to see included in a mental health surveillance program:

- Major disorders (mood, anxiety, substance use, psychotic disorders)
- Impact on function (work, household, family, social)
- Quality of life
- Educational attainment
- Access to health care and mental health care
- General health outcomes

Discussion

D.E.B. Potter (Office of the Assistant Secretary of Planning and Evaluation) asked if Olfson could comment on the all-payer claims databases that some states are developing. Olfson replied that he is primarily familiar with the activities being undertaken in New York. He said that substantial funds are being set aside in some states to bring together the claims databases to try to obtain a complete picture of all reimbursed care that is provided. He said that he has yet to see much research come out of these efforts, but that he thinks they will fulfill an important gap at the state level, which is not filled by existing administrative databases that only reflect parts of the national picture, and only one payer. He said these all-payer claims databases may suffer from the same kinds of general issues that he described earlier with regard to being based on treated individuals and diagnoses that are assigned by clinicians.

Vos added that, from his experience working with commercial administrative data from Blue Cross/Blue Shield, he has seen surprisingly fleeting populations with few people steadily in the system, even over a period of 1 year. Over 3 years, it may be that only one-third of the people are continuously in the same system. The new state-based initiatives to develop all-payer claims databases will substantially increase the usefulness of these data sources if they can link individuals between separate payer systems.

Olfson underscored Vos' point by saying that the length of time people stay in a given job is declining in the United States, with the average around 6 years. In addition, people who do not change jobs can still change health plans annually, and employers may change health plans for an entire employee group. It is also important to note the turnover in the Medicaid and Medicare population. As a result of these fluctuations, if one wants to analyze rates in a year, a proportion of cases are lost because some people are not eligible for the full year. Vos added that there may be selection bias, and Olfson agreed that the changes may not be random. Robert Gibbons (University of Chicago) said that the average length of tenure in a system is about 2½ years.

Gibbons also noted that it would be possible to look for cross-validation for these data in other databases, such as those of the Department of Veterans Affairs, or the Karolinska Institute, where researchers have linked Scandinavian databases. He mentioned that other relevant existing data sources would be ones like DARTNet, which is an integrated medical practice database run by the University of Colorado. It has about 400 practices covering approximately 4 million people. Olfson remarked that in the area of mental health there have been efforts over the years to put together practice-based research networks, which build on that idea. The difficulty is in having to rely on the good will and volunteerism of the participants.

Dean Kilpatrick (Medical University of South Carolina) asked whether the databases Olfson covered in his presentation could be used to supplement data from population-based surveys and what the implications would be in terms of potential double counting. Olfson said that, rather than trying to overcome difficulties that would arise in trying to supplement a population survey, the administrative databases may be more useful in providing information about rare events that are psychiatric in nature and about populations with high use and their characteristics. The strength of these databases is that they can provide information that is not possible to capture in population-based surveys. Olfson added that he would have to give more thought to whether administrative databases could actually help in estimating the size of populations.

Gibbons suggested that one possibility would be to develop a practice-based network, with the aim of adding practices in a way that is nationally representative. A good example of this approach is that of IMS Health, a network of 30,000 nationally representative pharmacies, which is based on a cluster sampling approach. Darrell Regier (Uniformed Services University) added that the design of the practice research network of the American Psychiatric Association (APA) is similar. In that case, the researchers had access to the American Medical Association master file, and they were able to select a random sample of psychiatrists in the master file, regardless of their membership in the APA. This design was used for a number of major studies on the topics of parity insurance, Medicare Part D, and a number of other policy issues. He added that long-term support from associations is needed to sustain these types of efforts, and obtaining that support is challenging. Regier said that the Colorado DartNet network structure—in which a facilitator helps keep the providers engaged—is a better long-term strategy than the APA one in terms of maintaining stability of the groups.

THE NATIONAL HEALTH INTERVIEW SURVEY

Stephen Blumberg (National Center for Health Statistics) discussed the NHIS, one of the surveys covered briefly by Olfson. Blumberg said that it has never been a goal for the NHIS to measure serious mental illness, but some data on mental disorders have been collected since 1997, and efforts have been made over the years to enhance the data collected.

The NHIS is the primary source of information on the health of the U.S. civilian population living in households at the time of the interview. It is one of the major data collection programs of the National Center for Health Statistics (NCHS), which is part of the Centers for Disease Control and Prevention. The primary objective of the NHIS is to monitor the health and health care access of the population in the United States, through the collection and analysis of data on a broad range of health topics. The data are used widely throughout the Department of Health and Human Services to monitor trends in illness and disability and to track progress toward achieving the goals of the Healthy People initiative and other national health objectives. The NHIS has received much recent attention because of its ability to monitor the effects of the Affordable Care Act on health insurance, access to care, challenges that people encounter in paying medical bills, and other policy-relevant topics.

A multistaged clustered national sample of housing units from every state is used to represent the civilian noninstitutionalized U.S. population. The survey oversamples black, Hispanic, and Asian individuals, as well as adults aged 65 and older. The interviews are conducted in person, by Census Bureau interviewers. The NHIS has primarily been used for national and regional estimates. However, for the past few years, there were significant sample size increases in an effort to improve state-level estimates. In 2014, it became possible to produce estimates for all 50 states for certain measures that ask about every individual in the household (e.g., health insurance measures).

The base NHIS sample used to be 35,000 households and was later increased to almost 45,000. The survey domains are broad and cover health status; functional limitations; health conditions; health behaviors and risk factors; injuries; health insurance, access, utilization, and barriers; and a host of demographic and socioeconomic characteristics. The core questions are a short set that remain unchanged from year to year. They are supplemented annually to collect data on current issues of national importance in three sections: a family core, a sample adult core, and a sample child core.

In the family core, a knowledgeable respondent in the family is asked about all of the members of the family. For the other two sections, one adult and one child are randomly selected from each family. For the sample child, the interviews are conducted with the parent or guardian of that child. The health conditions are predominantly measured in the sample adult section. For most of the conditions, the survey items ask whether a doctor or other health care professional diagnosed a particular condition, such as diabetes or hypertension.

The NHIS generally does not rely on medical screening or tests, as does the NHANES; but for the mental health data, a decision was made in the 1990s to include a screener in order to distinguish cases based on the severity of symptoms rather than purely based on the receipt of a diagnosis. At that time, there was no short battery of questions that could identify clinically significant community cases using lists of symptoms. Therefore, when the NHIS was redesigned in 1997, Ronald Kessler from Harvard University was commissioned to develop a short questionnaire of about six to eight items that would assess the severity of symptoms. This work ultimately resulted in what has come to be known as the K6, a measure of nonspecific psychological distress.

The development of the K6 started with more than 600 questions on symptoms and then used item response theory (IRT) methods to reduce that to a much smaller set. A 10-item measure was first developed, which then was reduced to the current 6-item measure. In applying IRT, the goal was to maximize the precision of the scale around the 90th-95th percentile because that was the expected threshold for clinical significance.

The version of the scale used in the NHIS asks about feelings *during the month prior to the interview* (see Box 4-6). Blumberg pointed out that there is another version of the scale, which is also referred to as the K6, that asks questions about *the one month during the past year when the individual had the most severe and persistent emotional distress.* The latter terminology was used in an effort to better measure serious mental illness using the 12-month "look back" that would be required by the definition of serious mental illness. Because the NHIS is not measuring serious mental illness, the K6 used in the NHIS is measuring distress in the past month.

In order to get information for the 90th-95th percentile goal, each

BOX 4-6
NHIS K6 Instrument for Measuring Serious Psychological Distress

Now I am going to ask you some questions about feelings you may have experienced over the past 30 days. During the past 30 days, how often did you feel…

- So sad that nothing could cheer you up?
- Nervous?
- Restless or fidgety?
- Hopeless?
- That everything was an effort?
- Worthless?

ALL of the time, MOST of the time, SOME of the time, A LITTLE of the time, or NONE of the time

SOURCE: National Health Interview Survey Questionnaire. Available: ftp://ftp.cdc.gov/pub/Health_Statistics/NCHS/Survey_Questionnaires/NHIS/2015/english/qadult.pdf [January 2016].

of the symptoms is asked with response categories of all of the time (4 points), most of the time (3 points), some of the time (2 points), a little of the time (1 point), or none of the time (0 points). A cutoff score of 13 or higher means that the individual had to respond "most of the time" to at least one of these items and had elevated scores on everything else, and this yields prevalence estimates of serious psychological distress.

Figure 4-2 illustrates that, in the 2009-2013 NHIS, 3.4 percent of adults aged 18 and over reported serious psychological distress and that the rate was higher among women than men in every age group. Blumberg said that an analysis of the relationship between prevalence of serious psychological distress and income showed that lower income adults were more likely to experience serious psychological distress. The prevalence of serious psychological distress among non-Hispanic whites was lower than among Hispanics and non-Hispanic blacks.

Blumberg emphasized that the NCHS values the K6: it is considered to be one of the 15 key measures from the NHIS. Data from the K6 are therefore part of the NHIS early release program, which is a program that releases reports every 3 months prior to final processing and weighting of the annual data, in order to provide access to the most recent information. Blumberg pointed out that the percentage of adults with serious psycho-

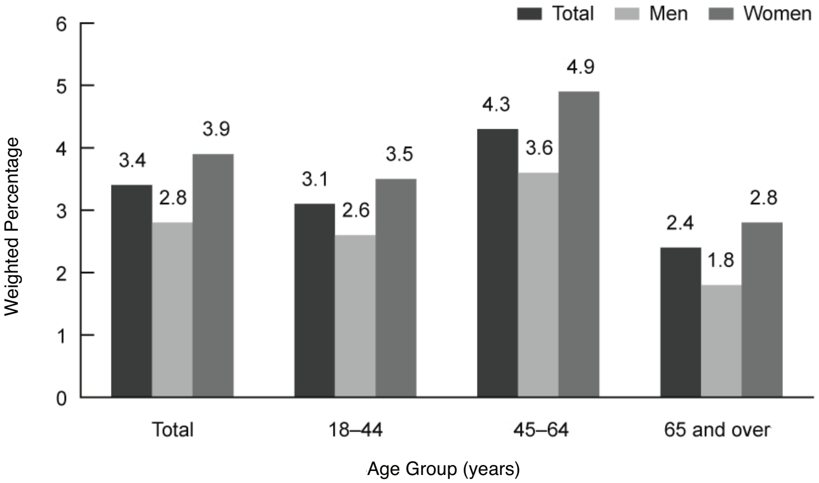


FIGURE 4-2 Adults with serious psychological distress, by sex and age: 2009-2013 National Health Interview Survey.
SOURCE: Weissman, J., Pratt, L.A., Miller, E.A., and Parker, J.D. (2015). *Serious Psychological Distress Among Adults: United States, 2009-2013.* NCHS data brief #203. Hyattsville, MD: National Center for Health Statistics.

logical distress between 1997 and 2015 has stayed around 3 percentage points, with a low of 2.4 percent in 1999 and a high of 3.8 in 2013. There was a significant increase of 1.1 percentage points between 2012 and 2013, which Blumberg said may have been related to a change in location of the psychological distress items on the survey. In 2014, the percentage came back down to 3.1 percent with the items in the same new place on the survey.

Blumberg noted that in the NHIS there is a follow-up question to the K6 that asks: How much did these feelings interfere with your life or activities? It is asked of all adults who reported at least one feeling on the K6 as experienced at least some of the time. However, NCHS has not used this item in any analyses. David Cella (Northwestern University) remarked that it is an item on functioning and could be analyzed to see which of the six symptoms affects functioning more or less than the others. Blumberg encouraged workshop participants to do additional analyses on the NHIS data, which are publicly available.

Blumberg also explained that the design of the NHIS incorporates annual supplements to periodically collect more information. In the 1997 redesign, a mental health supplement was added in an effort to obtain more detail about specific mental illness diagnoses. At the same time that

Ronald Kessler was working on the K6 as a core NHIS item, he was also under contract with NCHS to modify the Composite International Diagnostic Interview Short Form (CIDI-SF) to fit with DSM-IV criteria, with the idea that it would be used as a periodic supplement. The CIDI-SF was designed to estimate the prevalence of adults meeting the DSM-IV criteria for six different psychiatric outcomes and the DSM-III-R criteria for two addictive disorders.

The CIDI-SF is designed as a short series of symptom questions that follow the diagnostic stem questions. As an example, the series for major depression starts with two questions about whether the respondent felt sad, blue, or depressed for at least 2 weeks within the past 12 months and whether the respondent had lost interest in most things for at least 2 weeks within the past 12 months. A "yes" answer to one of these questions leads to a series of questions about how often this was experienced, and about depressive symptomology, feeling tired, having trouble concentrating, and feeling worthless. The aim is to determine whether the diagnostic criteria for major depression were met.

Despite developing the CIDI-SF for a number of different psychiatric outcomes in the 1999 periodic supplements, the only disorders that were included were major depression, generalized anxiety disorder, and panic attacks. Figure 4-3 presents the prevalence rates of these three disorders in 1999. As can be seen in the figure, nearly 9 percent of adults had any one of these selected mental disorders.

Blumberg said that the adult mental health supplement was never repeated. It is not clear why, but in 2001 and 2002 the NHIS was under some budget pressures, and there was more interest in supplements that other agencies were paying for than supplements that NCHS was paying for itself. Also, the validation studies on the CIDI-SF were never funded. Ultimately, it was only calibrated to the National Comorbidity Study data. Confirmatory clinical follow-up interviews were also not carried out for the CIDI-SF. Blumberg also presented information on a new initiative that is being undertaken with the Washington Group on Disability Statistics. The Washington Group was authorized by the U.N. Statistical Commission, following the U.N. International Seminar on the measurement of disability. Working from the structure of the International Classification of Functioning, Disability, and Health, this group was tasked with developing a small set of general disability measures to be used in censuses and other sample-based national surveys throughout the world.

The guiding principle for this work was that disability is the outcome of an interaction between a person and his or her environment, and it is therefore best measured as the ability of people to participate in their current environments. Some of the parameters for the new items were that the measures should be usable in surveys throughout the world, should

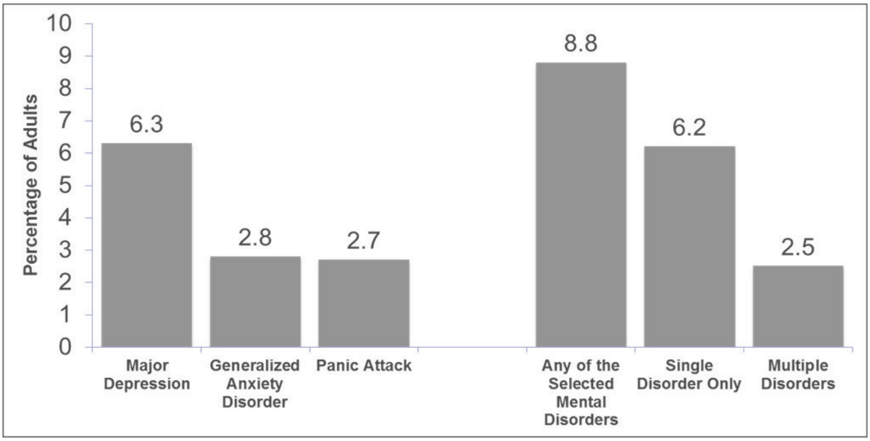


FIGURE 4-3 Prevalence rates for selected mental disorders among adults, past 12 months, 1999.
SOURCE: Dickey, W.C., and Blumberg, S.J. (2004) Prevalence of mental disorder and contacts with mental health professionals among adults in the United States: National Health Interview Survey, 1999. In R.W. Manderscheid and M.J. Henderson (Eds.), *Mental Health, United States, 2002* (Chapter 8). Rockville, MD: Substance Abuse and Mental Health Services Administration.

provide comparable data nationally, should have elements that crossed cultures and varying economic backgrounds, and should be short. For example, the World Health Organization Disability Assessment Scale was not a candidate because it is too long.

The group first developed a six-item measure of disability, based primarily on body systems: seeing, hearing, walking, remembering, self-care, and communicating. The items were intended to be somewhat dimensional, and four answers were available for each: no difficulty, some difficulty, a lot of difficulty, or cannot do at all. After these dimensional measures were developed and tested, the group developed an extended question set on affect, pain, fatigue, and the disability that results from these. For example, the questions on affect ask: How often do you feel worried, nervous, or anxious? Do you take medication for these feelings? Thinking about the last time you felt worried, nervous, or anxious, how would you describe the level of these feelings? The three questions are then repeated, substituting depressed for worried, nervous, or anxious. The items were cognitively tested in 15 countries and field tested in 9 countries. The researchers concluded that the questions are well understood across cultures and across economic situations, and they yield a continuum that correlates well with functioning difficulties.

Blumberg pointed out that there is no item on functioning difficulty that specifically asks, for example, if that feeling limits people's ability to carry out their daily activities. Functional items were part of the initial testing set, but they did not have the desired psychometric results. As discussed in earlier sessions, possible reasons may have been the difficulty in attributing one's inability to carry on their daily activities to certain feelings. But, he added, people may compensate for their limited ability to do something by changing their environment so they can function better or will not be required to function in that way. For example, someone who starts to have difficulty hearing may choose not to go to the movies or not to go to a crowded restaurant. A person may or may not even be aware that the change is due to a functional limitation related to hearing. Another issue the researchers faced in measuring functional limitation concerned the cross-cultural nature of the functional item—what daily activities would apply across the board from western Europe to Sri Lanka to the Maldives?

Without a functional item, the Washington Group now faces the challenge of determining a cut point for clinical significance with the frequency and intensity questions. Nevertheless, the questions have been included in the NHIS as a supplement since 2010 and were added to the 2014 European Health Interview Survey. The Washington Group questions have also been endorsed by the Budapest Initiative, which is another U.N. Statistical Commission group that was tasked with developing measures of health states for inclusion in the European Health Interview Survey.

In concluding the presentation, Blumberg discussed the questionnaire redesign that the NHIS is currently undergoing for the 2018 data collection. Although no decisions have been made, it is likely that the Washington Group questions will become part of the NHIS core. The goals of the redesign are to improve measurement and to incorporate recent advances in survey methodology. The researchers are dealing with the challenges of shortening a 90-minute survey in order to increase response rates. They also want to harmonize the NHIS content with that of other federal health surveys as appropriate. In addition, they want to establish a long-term structure of ongoing and periodic topics rather than supplements that are used only once and then discontinued due to changes in funding or priorities. A more stable structure would allow NCHS and its stakeholders to better predict the topics that will be included in certain survey years and the data that will be produced. He said that the mental health topic may be built into the new structure, but this has not yet been decided.

As part of the questionnaire redesign process, NCHS is currently involved in stakeholder engagement and outreach. The overall timeline calls for qualitative and quantitative assessments in 2015 to 2016, the required review by the Office of Management and Budget, and public comments from spring 2016 to spring 2017, questionnaire reprogramming in late spring 2017, and fielding of the new questionnaire in January 2018.

Cella remarked that the K6 cut point on the NHIS is sometimes a little lower than the 95th percentile and asked if there was a difference in data collection methods that Kessler used versus those used for the NHIS. Blumberg said that Kessler used data that were based on interview surveys in the United States and Australia, like the interview methods used in the NHIS. Lisa Colpe (National Institute of Mental Health) added that the cutoff of 13 was established through a pilot clinical calibration study and work done by SAMHSA as the researchers were refining it for their study. Ron Manderscheid (National Association of County Behavioral Health & Developmental Disability Directors and Johns Hopkins University) commented that the scale score of 13 was standardized in effect on a Global Assessment of Functioning (GAF) score of 50, but he noted that other work was done on a GAF of 60. That is why Kessler always talked about the cut point being between 5 and 6 percent.

Vos noted that Blumberg's discussion of the Washington Group's challenges in deriving functional limitation measures was very interesting in view of the GBD study's experience. The GBD found large differences in how people in different countries respond to functional items, which may reflect economic circumstances and culture. He said that one of the major reasons the GBD concentrates more on health loss and impairments, rather than general welfare and functioning in overall life, is to have valid comparisons.

Fred Conrad (University of Michigan) remarked that face-to-face household interviews that are conducted for the NHIS do not seem to be the ideal context in which to collect information about stigmatized behaviors. He asked if the NHIS data show any evidence of underreporting because of the nature of the questions. For example, are the higher levels of prevalence for females than for males perhaps related to males being less willing to report sensitive information about symptoms in an interview? Blumberg replied that this difference would probably also apply outside of the interview context: men just do not want to admit, even to themselves, these symptoms. Blumberg said he does not know whether the K6 is susceptible to mode effects, but it would not be surprising if it was. He said that it would be possible to look at this issue by examining differences between the sample adult interviews that are done in person versus the relatively few that are done by telephone.[9] Conrad added that an even better comparison would be with self-administration.

[9] The NHIS is primarily an in-person survey, but a small number of interviews are completed by telephone, after an initial in-person contact is made with the respondent: this is sometimes done to finish a partial interview.

Given that SAMHSA needs state-level estimates, Graham Kalton (Westat) asked Blumberg if NCHS has done any small-area estimation with the NHIS data. Blumberg said that they have, but he is not aware of small-area estimation using the K6. For example, he used it for estimates of cell-phone-only households. In addition, NCHS also used small-area estimation in connection with BRFSS data to obtain some estimates for the National Cancer Institute. Blumberg said that it would be interesting to look at the NHIS K6 state-level estimates from 2014 and see how they match up with the state-level estimates that SAMHSA has produced using the 30-day K6, if the samples are large enough for that comparison. Jonaki Bose (SAMHSA) said that SAMHSA has collected past year data at the state level, and it could also capture past 30 days data at the state level.

5

Innovative Approaches to Measurement

COMPUTERIZED ADAPTIVE TESTING

Robert Gibbons (University of Chicago) discussed how computerized adaptive testing can be applied to mental health measurement. He reminded workshop participants of the discussion about the challenge in creating the K6 from 600 items. As part of that effort, six items were derived to produce a score for psychological distress. An alternative to administering the same set of six items to each individual would be to keep the "K600" and use computerized adaptive testing (CAT) to produce a score using a subset of the items, averaging six items—plus or minus two items—that are best suited for each person.

In classic measurement theory, using the K6 as an example, there are six items, measured on an ordinal Likert-type scale. These items are like a series of hurdles in a race and they are added to produce a score. The score is then supposed to be a sufficient statistic to represent something in the universe. If an additional item (another hurdle) is added (to produce a "K7" for example), or the distances between the hurdles are changed, the scores between the two tests are no longer comparable, so everyone is administered the same set of items. By contrast, item response theory (IRT) is more similar to a high jump, with the height of the bars measured in inches. More skilled jumpers could start higher and end up jumping higher than less skilled jumpers, but everyone is still measured using the same metric. IRT is a model-based measurement and enables adaptive testing, where one person can be administered one set of items, and another person can be administered another set of items, while using the same metric.

Gibbons explained CAT with another metaphor. He asked the workshop participants to imagine a mathematics test that consists of 1,000 items, ranging in difficulty from simple arithmetic to advanced calculus, and two examinees, a fourth grader and a graduate student in statistics at the University of Chicago. Both could take a test consisting of all 1,000 items, and their scores would be very good estimates of their abilities, but this would not be an efficient use of their time. Alternatively, a test of only three items could be administered—one to measure arithmetic, one for algebra, and another one for calculus. This would be more efficient in terms of time, but we would learn very little in terms of their abilities. A better approach would be to start with an intermediate algebra item. If the fourth grader gets it wrong, he or she begins to receive easier items. If the graduate student gets it right, he or she moves to more difficult items. The process continues until the uncertainty in the estimated ability is smaller than a predefined threshold.

To use CAT, a bank of test items is first calibrated using an IRT model that relates properties of the test items (for example, their difficulty and discrimination) to the ability (or other trait) of the examinee. The paradigm shift is that, rather than administering a fixed set of items and allowing precision of measurement to vary between, or even within, individuals, CAT fixes measurement precision and allows the items to vary both in number and in content. The items are adaptively selected out of a much larger bank of items and the starting point of the adaptive testing process can also be informed by prior test results. The precision of the test can be adjusted depending on the application. For example, for an epidemiological study, less precision may be needed than in other situations, so that it would be sufficient to administer fewer items. More precision and more items may be desirable for screening in a primary care setting, while maximum precision and an even larger number of items may be needed in a randomized, controlled trial.

Gibbons said that historically CAT has been applied to unidimensional constructs in educational measurement. What is new about this work is the use of multidimensional IRT models as the foundation for CAT. This has particular advantages when measuring concepts such as depression, which are inherently multidimensional, with items drawn from cognitive or somatic domains, or domains related to mood, suicidality, or functional impairment. He said that the model used in CAT is complex, because different items may have different numbers of categories, different severity thresholds, and different abilities to discriminate high and low levels of the underlying latent variable of interest. The greatest complexity is introduced by the multidimensionality of the items. However, Gibbons noted, an important by-product of CAT is that the estimates of impairment are accompanied by estimates of uncertainty, which can be used to construct confidence intervals for the point estimates and characterize the resulting precision of the measurements. This is not possible to do in traditional mental health testing.

Gibbons pointed out that depression, and psychiatric rating scales in general (e.g., the K6, the [Patient Health Questionnaire] PHQ-9), work well at the extremes, that is, in differentiating the really depressed people from those who are not depressed. However, in the middle of the distribution, the traditional scales are less precise. With CAT, there is uniform precision because items can continue to be delivered until a desired level of precision is reached for everyone who is responding. This is possible because of the very large item banks.

Gibbons shared some results from his research on depression, anxiety, and bipolar disorder.[1] He pointed out that for depression—using a standard error of 0.3—the precision is about 5 points on a 100-point scale. Table 5-1 shows that with this standard error they were able to maintain a correlation of 0.95 with the 400-item depression bank, using an average of only 12 adaptively administered items. Relaxing the standard error to 0.4, which is about 7 on a 100-point scale, only an average of six items was needed to maintain a correlation of 0.92 with the 400-item bank. The results for anxiety are virtually identical using an average of about 12 items. The correlation for anxiety was 0.94 with a 430-item bank. Bipolar (mania) had a lower correlation of 0.91, using an average of 12 items: the reason may be that the mania items were dichotomous. Generally, polytomous items, ordinal items, or multicategorical items work best in multidimensional IRT-based CAT models.

Gibbons also described work in which he has participated to develop the first computerized adaptive diagnostic screener. He reminded workshop participants that for diagnostic screening the goal is to identify the tipping point between the probability of a positive and a negative diagnosis, while for measurement the goal is to differentiate severity levels. The researchers found that, with an average of four items and a maximum of six items, administered in an average of 36 seconds, they could maintain

[1] See Gibbons, R.D., Weiss, D.J., Pilkonis, P.A., Frank, E., Moore, T., Kim, J.B., and Kupfer, D.K. (2012). The CAT-DI: A computerized adaptive test for depression. *Archives of General Psychiatry, 69*, 1104-1112.

Gibbons, R.D., Weiss, D.J., Pilkonis, P.A., Frank, E., Moore, T., Kim, J.B., and Kupfer, D.J. (2014). Development of the CAT-ANX: A computerized adaptive test for anxiety. *American Journal of Psychiatry, 171*, 187-194.

Achtyes, E.D., Halstead, S., Smart, L., Moore, T., Frank, E., Kupfer, D., and Gibbons, R.D. (2015). Validation of computerized adaptive testing in an outpatient non-academic setting. *Psychiatric Services, 66*(10), 1091-1096.

TABLE 5-1 Results from Simulated Computerized Adaptive Testing

Test	Standard Error Term	Correlation	Mean Number of Items	Minimum Number of Items	Maximum Number of Items
Depression	0.30	0.95	12	7	22
	0.40	0.92	6	4	16
Anxiety	0.32	0.94	12	6	24
Bipolar Disorder	0.45	0.91	12	6	24

SOURCE: Workshop presentation by Robert Gibbons, September 2015.

the sensitivity of 0.95 and specificity of 0.87 of an hour-long face-to-face Structured Clinical Interview for DSM for major depressive disorder.

Gibbons also described an independent validation study that produced similar results.[2] He and his colleagues used a highly comorbid community mental health sample (N = 150) and found a sensitivity rate of 0.96 and a specificity rate of 1.0 for major depression disorder in comparison with the control population. Of the people who participated, 97 percent said that the test results accurately reflected their mood, and 86 percent preferred the computer interface to other testing modes. He noted that even older people who had less experience using computers were comfortable using it.

Gibbons also presented new data on detection rates in emergency rooms that involved screening approximately 1,000 people in the emergency department at the University of Chicago. Using a confidence level of over 50 percent, 26 percent of the participants screened positive for major depressive disorder. This proportion dropped to 22 percent with a confidence level of over 90 percent. When the CAT for depression was combined with the CAT diagnostic screener, 7 percent were found to be in the moderate to severe range. In addition, 3 percent had a positive suicide screen, which means ideation, in addition to intent, a plan, or recent suicidal behavior. Gibbons said that these are the people who need treatment, but, remarkably, these patients were not coming to the emergency department for a psychiatric indication. A health services implication of the findings is that the rate of emergency department visits in the past 2 years was three times higher for those who screened in the moderate to severe range than for those who screened in the none to mild range. The

[2] Achtyes, E.D., Halstead, S., Smart, L., Moore, T., Frank, E., Kupfer, D., and Gibbons, R.D. (2015). Validation of computerized adaptive testing in an outpatient non-academic setting. *Psychiatric Services, 66*(10), 1091-1096.

researchers also found the rate of hospitalizations in the past 2 years was four times higher among the moderate to severe positive screens in comparison with the mild to none negative screens. Gibbons suggested that there are enormous financial implications in terms of the service needs of depressed patients who show up in the emergency department for other than psychiatric indication.

Gibbons briefly described a study conducted in Spain and with Latino populations in the United States to examine whether or not the items mean the same thing in different cultures. With an IRT-based system of measurement, it is possible to look at the discrimination parameter and see whether there is differential item functioning. It is also possible to determine whether there are items that are excellent discriminators of high and low levels of depression in one culture but work less well in another culture (e.g., Latino population). Examining detection rates in primary care in Barcelona and Madrid, where depression screens were administered in Spanish, the researchers found similar high rates as in the emergency room study conducted at the University of Chicago.

Gibbons then described possible future directions for CAT in mental health assessments. He said that CAT has important applications for screening and monitoring in primary care; for conducting inexpensive phenotyping for large genome-wide association studies; for psychiatric epidemiology; and in comparative effectiveness and safety studies. Gibbons said he will also be using CAT as part of the Kiddie CAT study to assess the dimensions of depression, anxiety, mania, attention deficit hyperactivity disorder, oppositional defiant disorder, and conduct disorder. The study involves an item bank of 1,200 items for parents and 1,200 items for children, and the goal is to develop diagnostic screeners and measures for each of the dimensions. CAT can also be applied to autism, posttraumatic stress disorder, substance use, and research domain criteria dimensions (see Chapter 2). Gibbons said that CAT can also have a very useful application in military populations, for which the risk of suicide within the first 4 years after discharge is four times higher than the rate in the general population. One advantage of CAT is that the mental health applications can all be used in cloud computing environments, unless there is a reason not to do that, such as when screening for suicide.

Gibbons concluded his presentation with a demonstration of the CAT depression screen and a screen for suicide risk. After administering the test, the results show whether the depression screen was positive or negative and the severity level, along with the associated confidence level and precision. There is also a suicide warning displayed on the results screen. Gibbons noted that after the CAT is administered, a text message and email can be sent to several recipients, such as clinicians or a suicide hotline, as needed.

After the test is completed, it is possible to look up details about the interview, such as what questions were asked of the person, what the summary scores were, and how long it took to answer each question. If some items took longer than the rest, such as the suicide items, a clinician might want to follow up. These are additional unobtrusive measures that systems of measurement such as CAT can provide.

THE PATIENT REPORTED OUTCOME MEASUREMENT SYSTEM

David Cella (Northwestern University) discussed the Patient Reported Outcome Measurement Information System (PROMIS), a project under the Common Fund, which supports cross-cutting, trans-NIH programs. There was interest in standardizing a range of measures including those for pain, depression, physical function, social and cognitive function, dexterity, as well as other domains across different mental and physical diseases.

The PROMIS Cooperative Group, which operated from 2004 to 2015, was widely considered to be one of the success stories of the Common Fund. The project involved more than 250 investigators and more than 50 protocols, aligned around evolving PROMIS standards. More than 50 grants were funded not only by Common Fund grants, but also by different NIH institutes and other government and nongovernment entities, including the National Institute of Mental Health, the National Institute on Drug Abuse, the Centers for Disease Control and Prevention, the Patient Centered Outcomes Research Institute, the Department of Defense, the Department of Veterans Affairs, the Army, foundations, and industry.

Across the qualitative and quantitative databases, information was collected from more than 50,000 adults and children. All of the measures are available in English and Spanish, and many subsets of item banks are also available in Chinese and other languages. For adult health measures, there are about 1,500 items, which populate 71 distinct item banks and scales and are available in 20 languages. For pediatric health measures, there are about 280 items that make up about 40 distinct banks and scales in 10 languages.

From the beginning, PROMIS has been domain specific, not disease specific. By definition and by design, the work has focused on measuring traits, attributes, moods, and functional areas that cut across diseases. Cella said that item banks, as Gibbons' talk illustrated, are a great way to accomplish that. He defined item banks as large collections of items that measure a single domain, which is the specific feeling, function, or perception of interest. The domains cut across diseases.

As starting point for its domain framework, PROMIS uses the World

Health Organization (WHO) tripartite definition of health as a state of complete physical, mental, and social well-being. PROMIS has also been linked to the more recent WHO International Classification of Functioning, Disability and Health model. The domains of physical health are symptoms and function; the domains of mental health are affect, behavior, and cognition; the domains of social health are relationships and function. The item banks are spread across this broad framework, and they are unidimensional, although the PROMIS team has also experimented with multidimensional IRT and for some purposes uses a bifactor model developed by Robert Gibbons.

The goal for the PROMIS metrics is to capture the full spectrum of a concept or domain, such as physical functioning from 0 to 100 (e.g., getting out of bed, standing without losing balance, walking from one room to another, walking a block, jogging for 2 miles, running for 5 miles) and only ask those questions that are relevant. The approach is similar to that for a CAT environment. The metrics for PROMIS have a mean of 50 and a standard deviation of 10. The items in almost all cases are referenced to the U.S. general population.

One of the PROMIS tools is the Global Health Scale, which is a 10-item measure that can be thought of as a shorter version of the 12-Item Short Form Health Survey (SF-12). It is similar to the SF-12 in that it produces a global physical health score and a global mental health score. The index is conceptually comparable to the SF-12 and has the advantage of being free and publicly available. The Global Health Index is derived from item banks using CAT and averages about four or five items per domain. For some domains, for example, depression, very often there are just three items.

Also derived from item banks are fixed length forms of 4 to 10 items that are available "off the shelf," by individual domain. Short forms can also be customized for specific needs. If enough is known about a population, items can be selected that work better in a given range of the trait that is to be measured. PROMIS also has fixed length forms that cover seven domain health profiles: anxiety, depression, fatigue, pain, sleep, physical function, and role satisfaction. These profiles can also be used as short forms that are pulled from the calibrated item banks. Depending upon the desired sample size and level of precision needed, there are short forms that contain four, six, or eight items per domain.

As an example of how PROMIS is used, Cella said that the American Psychiatric Association is using the PROMIS depression and anxiety short forms in its DSM-5 field trials. In their approach, the PROMIS short forms are administered if screening items are answered with mild symptomatology evident. The PROMIS anxiety and depression short forms used are 6-8 items long, and each question uses a five-point frequency rating (never, rarely, sometimes, often, always). The t-score can then be identified from a patient's raw score, as long as all items have been completed. There are also cross-cutting Level 1 and Level 2 measures for child anxiety, depression, sleep, and anger. These cross-cutting measures are recommended to track severity of symptoms over time and as indicators of remission or of exacerbation of symptoms. They are completed at regular intervals, as clinically indicated, and consistently high scores identify an area that needs more detailed assessment, treatment, or follow-up.

Cella noted that a 2013 report[3] compared the DSM-5 approach for diagnosis (in other words, the use of information from cross-cutting measures, diagnostic criteria, and diagnostic-specific severity measures) to the DSM-IV approach for various disorders in pediatrics, and found that 80 percent of clinicians reported that, in their clinical experience, the DSM-5 approach was better or much better than the DSM-IV approach. Examining the same question by specific disciplines (i.e., psychiatrists, marriage and family therapists, clinical social workers, and counselors) again showed that about 70 percent of providers preferred the DSM-5 approach. Similar results were observed for adult patients who thought their clinicians would better understand their symptoms.

Cella also discussed how PROMIS measures are being used by Centers for Disease Control and Prevention and the Healthy People 2020 initiative. The measures have been approved for use in Healthy People 2020 and the National Health Interview Survey. The objectives were to increase the proportion of adults who report good or better physical health-related quality of life and to increase the proportion of adults who report good or better mental health-related quality of life. Four PROMIS global mental health items were approved as part of this effort, with excellent, very good, good, fair, and poor as response categories:

1. In general, would you say your quality of life is….
2. In general, how would you rate your mental health, including mood and ability to think?
3. In general, how would you rate your satisfaction with social activities/relationships?
4. How often have you been bothered by emotional problems?

Figure 5-1 shows 2010 NHIS data on the proportion of adults who reported good or better mental health among different demographic groups. The 2020 target that was proposed and approved by the Federal

[3] Moscicki, E.K., Clarke, D.E., Kuramoto, S.J., Kraemer, H.C., Narrow, W.E., Kupfer, D.J., and Regier, D.A. (2013). Testing DSM-5 in routine clinical practice settings: Feasibility and clinical utility. *Psychiatric Services, 64*(10), 952-960.

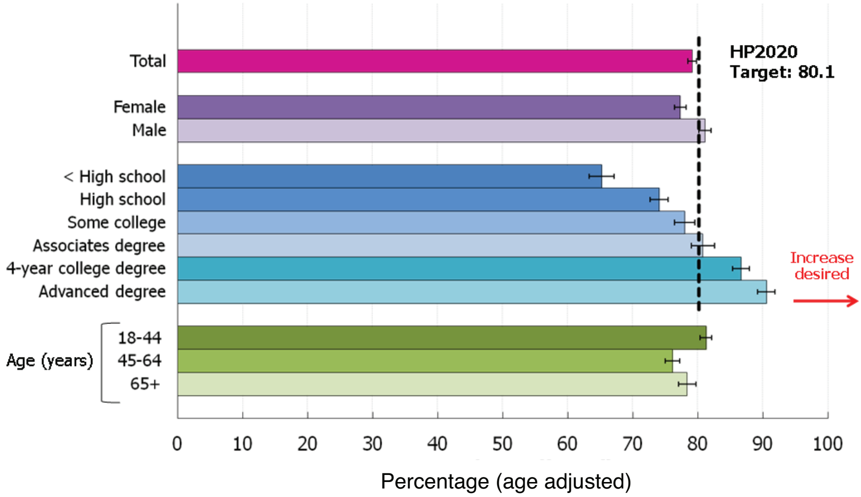


FIGURE 5-1 Adults who report good or better mental health, by demographic characteristics: 2010.
NOTE: Data (except data by age group) are age adjusted to the 2000 standard population.
SOURCES: Healthy People 2020 Spotlight on Health. Available: https://www.healthypeople.gov/sites/default/files/HP2020_SpotlightOnHealthHRQOL.pdf [January 2016]. Data from the National Health Interview Survey.

Interagency Working Group is the dotted line in the figure. The current 2010 status is the top magenta line. With regard to mental health, fewer women report good or better mental health than men. Education is a strong predictor of mental health, with a disparity in lower educational levels as shown by below high school, high school, and some college being below the line. People with advanced degrees are above the line. Cella also noted that there is less of an age disparity in mental health than in physical health (not shown in this figure).

Figure 5-2 also shows adults who reported good or better mental health but compares those with and without different physical disorders. The figure illustrates the mental health disparity among people with and without such conditions as diabetes, cancer, hypertension, heart disease, and, especially, disabilities.

Cella closed his presentation with a discussion of a project called PROsetta Stone. Though funded through the National Cancer Institute, its goal is to develop and apply methods to link the PROMIS measures with other related patient-reported outcome measures in order to have a common, standardized metric. Cella pointed out that the project website

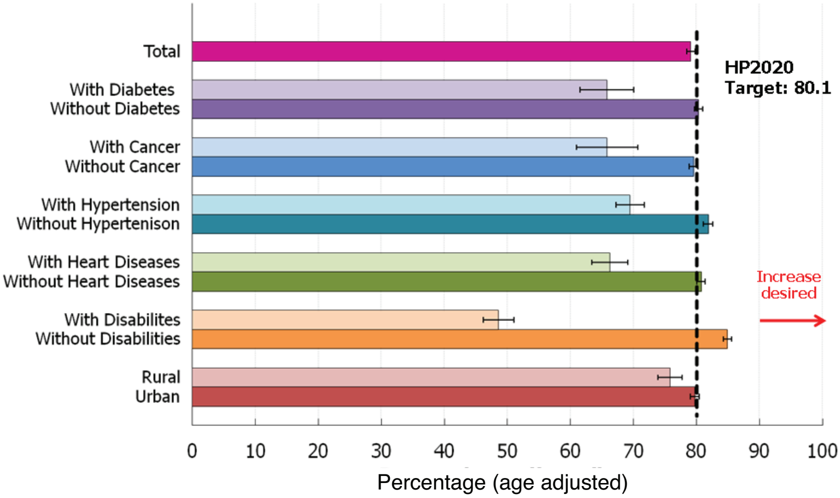


FIGURE 5-2 Adults who report good or better mental health, additional comparisons: 2010.
NOTE: Data are age adjusted to the 2000 standard population.
SOURCES: Healthy People 2020 Spotlight on Health. Available: https://www.healthypeople.gov/sites/default/files/HP2020_SpotlightOnHealthHRQOL.pdf [January 2016]. Data from the National Health Interview Survey.

has about four or five dozen tables that show different instruments linked and calibrated onto a common metric.[4]

Using the example of depression to link measures, the researchers first coadministered the PROMIS depression measure, the Center for Epidemiological Studies Depression (CES-D) measure, the [Patient Health Questionnaire] PHQ-9, and the Beck Depression Inventory-II, then calibrated all of the items. Figure 5-3 shows the cross-walk function between CES-D and PROMIS depression, with the scores mapping on top of one another. This is also true for other metrics.

Cella and his colleagues also produce a raw score to t-score conversion table that shows, for example, a PHQ-9 score and the PROMIS t-score equivalent. A PHQ-9 score of 10, which is moderate, is around 59 on a PROMIS t-score: 60 is a common t-score to use for mild to moderate symptomology and 70 for more severe symptomology, which would be a PHQ-9 score of around 19 or 20. He concluded by saying that the PROMIS team is working with organizations like the National Quality Forum and

[4] See http://www.prosettastone.org [December 2015].

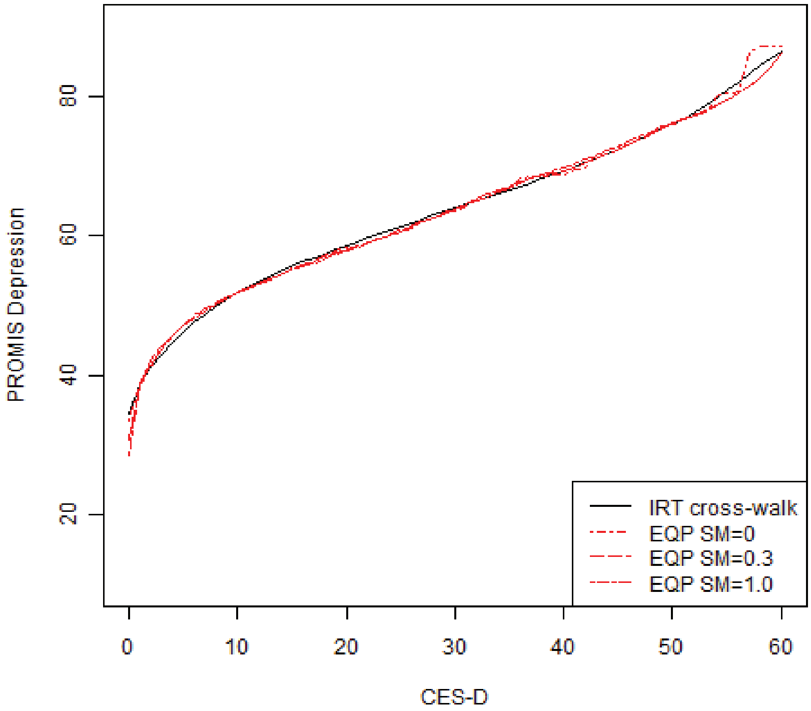


FIGURE 5-3 CES-D to PROMIS depression: IRT cross-walk function and equipercentile functions with different levels of smoothing.
NOTES: EQP, equipercentile; sm, post-smoothing. The IRT cross-walk function is based on fixed parameter calibration.
SOURCE: Choi, S.W., Schalet, B., Cook, K.F., and Cella, D. (2014). Establishing a common metric for depressive symptoms: Linking the BDI-II, CES-D, and PHQ-9 to PROMIS depression. *Psychological Assessment, 26*(2), 513-527. Published by the American Psychological Association, reprinted with permission.

National Committee on Quality Assurance to replace the use of the PHQ-9 with PROMIS metrics.

DISCUSSION

James Jackson (University of Michigan) asked Cella about the advantages of using PROMIS over the PHQ-9 depression measure. Cella replied that the PHQ-9 was driven by the DSM-IV and developed as a diagnostic tool, but it is now used as an outcome tool. For example, the PROMIS depression metric is on a near-interval level scale, so it is possible to begin to understand this underlying trait in a way that is not tied to DSM-IV clinical criteria. Regier pointed out that the PHQ-9 does try to integrate the multiple domains of major depression, as opposed to a univariate domain of depression alone. It includes mood, suicide risk, as well as various cognitive issues, which are not part of the depression univariate domain.

Regier then asked Gibbons about whether a multidimensional approach would enable researchers to capture more of the syndromal nature of mental disorders that contain more than just one domain. Gibbons replied that it could be done, but it would depend on what is being measured. Gibbons also commented on the PHQ-9, which he said in some sense is multidimensional, but it is scored with a single-value index. For example, one can have a PHQ-9 score of 13 for 4, 5, or 13 different reasons. In other words, one can have very different symptomology and have the same score. Using multidimensional IRT, it is possible to define the underlying domains from which the items were drawn, in the same way as the authors of the PHQ-9 did. It is also possible to preserve those underlying constructs and either map an unbiased estimate onto the construct of depression or score the individual subdomains and come up with something that is more multidimensional. If depression is composed of cognitive, mood, suicidality, sleep, and two or three other subdomains, it is possible to say that it is really depression, anxiety, and maybe some mania. It also becomes possible to obtain separate scores on each of the subdomains or obtain an overall composite score.

Gibbons also said that as part of one of his current projects he and his colleagues are working with 300 items on depression, mania, and psychosis, and are trying to produce an overall single-value index of severe mental illness that maintains the inherent multidimensionality of the item bank. One could then also score the individual subdomains, but that is not something his team has done yet.

Cella added that the issue of dimensionality is not all science or purely measurement, but also art to some extent. For example, the K6 includes four depression-like items and two anxiety-like items, and it fits an IRT model. However, it is not clear whether it is two-dimensional or one-dimensional. It is possible to make it one-dimensional. In fact, the bifactor model that Gibbons developed helps do that by removing some of the noise to purify the signal and allow content to stay in.

Cella went on to say that, for reasons related to conceptual elegance, PROMIS includes a separate depression/anxiety item bank, because these naturally work together. The assessment of both can be shortened, as Gibbons illustrated. If the depression test is administered, the anxiety test will be shorter because one knows where to start. It might only be shorter by one item because these polytomous items are efficient at determining where someone should start.

Robert Krueger (University of Minnesota) emarked that psychopathology in itself has a structure to it. In work with others, he and his colleagues looked at the structure of mental disorders using data from the National Comorbidity Survey Replication Adolescent Supplement.[5] They found a meaningful general factor that tends to bifurcate into internalizing and externalizing kinds of presentations, as well as further layers to the structure. Krueger noted that there is some recognition in DSM-5 that mental disorders have an underlying structure. If the goal is to measure overall psychopathology as a construct, he thought it would be possible to develop an efficient CAT method for doing so, and it would be akin to what has been called "distress" throughout the workshop.

Cella said that Gibbons' work is closer to doing what Krueger referred to than the PROMIS. Gibbons said that PROMIS has developed a very large series of measures to study a wide range of concepts using unidimensional IRT. The fact that it is unidimensional makes it more difficult to build very large item banks. The measures that Gibbons and his colleagues developed for depression, anxiety, and mania that were developed through CAT have been developed, with much larger item banks. The multidimensional IRT makes it possible to maintain that huge item bank, which allows for the adaptive selection of items that are tailored for each person. Because of the huge item bank, they would not be giving the same person the same items over and over again. As the item bank grows, CAT works better than a short-form test because it also provides uniformity of measurement throughout. Gibbons added, however, that it is very expensive to perform the original calibration in order to be able to maintain very large item banks.

Clinical Utility

Mark Olfson (Columbia University) commented that it is clear that the CAT is elegant and that getting to decisions more promptly, with fewer items, has advantages, if it can be integrated into large-scale surveys. However, he wondered whether the goal for both the CAT and the PROMIS is to be introduced into clinical practice. Gibbons replied that the CAT is suitable for the identification of people who need treatment. People who are not identified and not treated tend to consume health care services at high rates and are also at risk for sequelae, such as suicide. Once people are identified, CAT-based measurement is also ideal for

[5] Blanco, C., Wall, M.M., He, J-P., Krueger, R.F., Olfson, M., Jin, C.J., Burstein, M., and Merikangas, K.R. (2014). The space of common psychiatric disorders in adolescents: Comorbidity structure and individual latent liabilities. *Journal of the American Academy of Child & Adolescent Psychiatry, 54*(1), 45-52.

longitudinal assessment to determine whether people are responding to treatment and for making changes to the treatment plan.

Gibbons said that he is working on a project related to depression treatment, where frequent measurements are taken. The PHQ-9 or K6 are not suitable for administration every 30 minutes. When a CAT-based approach is used, two successive measurement occasions would not involve the same items, and thus the threat of response bias is lowered. He also noted that the test-retest reliability of the PHQ-9 is 0.80, and the test-retest reliability for the depression CAT test is 0.92. Despite asking different questions, the reliability of the estimates is improved, because CAT produces more precise estimates of depressive severity than traditional fixed-length tests.

From the perspective of a clinician, Olfson said the reason the PHQ-9 is so popular is its items are questions that clinicians want to ask because of their relevance to the person's mental health status. Clinicians are not just interested in the dichotomous decision of whether the person passes a threshold or not, which is possible to ascertain more rapidly and efficiently taking advantage of IRT. They want to know, within each of these domains, how well a person is sleeping, what his or her appetite is like, and whether they are having difficulty with decision making; clinicians will intend to follow up on any of the items that are positive. Ultimately, clinicians and patients may not be interested in the underlying construct but rather in probing more deeply about problem areas.

Gibbons said that his experience in working with clinicians has been just the opposite. The CAT helps them explore those areas where there is a density of symptomology. In fact, some clinicians use it as part of the therapeutic process. Patients go through the CAT, and then they review the responses and discuss what they were experiencing and why they answered in a certain way. Gibbons acknowledged that the CAT has some potential limitations. For example, the CAT can produce a valid measure of depression severity without having to ask questions from all possible domains, which means that if a clinician's goal is to learn about a particular symptom of interest that was not adaptively administered, he or she would have to supplement what was learned from the CAT.

Cella noted that on ClinicalTrials.gov, where people using PROMIS tools have the opportunity to use CAT, custom forms, or off-the-shelf short forms, the choice has been short forms, by a 5-to-1 margin over other options. He said that the reason might be that the CAT technology is not very accessible. There is also the desire of clinicians, clinical researchers, and regulators to want to see the answers to the same questions over time.

Krueger commented that one way to think about this is to consider the breadth of concepts that one would ideally like to cover and the amount of time that is available to cover them. There are many areas that are important to screen for, at least briefly, and if time is limited, CAT can be enormously useful.

Implications of the CAT's Precision in Identifying Disorders

Lisa Colpe (National Institute of Mental Health) commented that she was really impressed with the CAT and its precision, especially for measures that reference the past 2 weeks. She wondered whether SAMHSA would have a "duty to treat," if, based on a CAT method, one could identify people who are in need of treatment, given that the CAT approach can identify people who have the disorder now and not just "within the past year," as is the case with most national surveys. Typically, in the arena of public health, one does not conduct a screener unless it is possible to also take the next step and either do further assessment or offer treatment.

Gibbons replied that this comes up in his work of suicide screening. His team will only do suicide screening face to face so that if there is an issue, appropriate follow-up can be done. The question is whether something should be done even for untreated depression, and the right answer is, of course, yes, he said, because untreated depression leads to high health care costs and is also associated with a high suicide rate. Colpe said that procedures would have to be developed to address this issue, and perhaps participants could be given the instruction to go to a designated place for assessment or treatment, or could even be directed to an online cognitive behavioral therapy course, as needed.

Dean Kilpatrick (Medical University of South Carolina) said that research studies that involve sensitive topics typically have people available to help out, which is the standard for clinical epidemiological studies or epidemiological studies addressing mental health or substance use issues. Colpe replied that the standard at NIMH is to provide all participants with information about where they could go for treatment if they wanted to do so, after answering questions that are part of a study. However, collecting information about the past 2 weeks is different than asking about the past year, which is more typical for national surveys.

Calibrating for Language and Literacy

Thinking about SAMHSA's goals, Regier asked whether the PROMIS and the CAT have been calibrated well enough to be adapted to the entire U.S. population, including to specific demographic groups. Along those lines, Neil Russell (SAMHSA) also wanted to know whether the literacy level of the questions has been evaluated.

Gibbons replied that the literacy issue is one of the reasons that the questions are read out loud to the patients by a computer. One advantage of CAT is that it allows researchers to examine the issue of whether there are cultural differences. For example, if an item is a bad discriminator of high and low levels of depression among Latinos, then the item would not be included. Gibbons and his colleagues are continuing research on this topic to improve the CAT approach.

Use of Computerized Adaptive Testing by Federal Agencies

Stephen Blumberg (National Center for Health Statistics) commented that he considers computerized adaptive testing to be very useful in certain circumstances, but in the interest of transparency, government agencies need to be able to include in their datasets not just the scores, but also every item that the person responded to and exactly how the person responded. This is not impossible, but the documentation may be very extensive. He wondered whether government agencies would find the use of CAT more difficult to justify and whether there are any government agencies that are currently using CAT. Jonaki Bose (SAMHSA) replied that SAMHSA has used adaptive testing and that the National Center for Education Statistics also used it in an early childhood longitudinal study and in the National Assessment of Educational Progress.

Bose added that the transparency issue also raises the question of whether or not SAMHSA would be able to incorporate a proprietary set of questions in their surveys. Blumberg said that at one point NCHS was interested in such a possibility and this was discussed with the heads of all the federal statistical agencies: the conclusion was that a proprietary scale could only be used if it can be disclosed exactly what items are in the scale.

Technology for Administering CAT

Kilpatrick commented that CAT-type approaches have many benefits, but they can be confusing to implement. There would need to be a lot of education, so that people really understood how it works. Even if people understood it well, some might not be convinced about the advantages, given the transparency concerns. He asked whether CAT could be used on a free-standing laptop or tablet without Internet access, if one was interested in integrating it into a survey.

Gibbons replied that the system they have developed is designed for entire health care systems, so it primarily works through the Internet. He explained that their prototype versions are on dedicated computers, but University of Chicago undergraduates administered the tests in emergency departments, using tablets. Tablets were also used to collect data in primary care settings in Barcelona and Madrid, where Internet access was limited. He added that the team is in discussion with the Department of Veterans Affairs (VA) about the possible use of CAT, and they plan on giving the VA an executable version of the program that can be uploaded to their computers. Ultimately, from the perspective of Gibbons and his colleagues, the vision would be a cloud computing environment in which CAT could be administered on home computers, tablets, or cell phones.

6

Key Themes and Possible Next Steps

In considering the measures and data collection approaches discussed, Ron Manderscheid (National Association of County Behavioral Health & Developmental Disability Directors and Johns Hopkins University) remarked that, in the end, the purpose of the data collection will have to drive the methods selected, and SAMHSA would have to decide whether the primary purpose is epidemiological research, clinical assessment, policy development, or something else. When there is legislation with funding attached, congressional intent takes priority in terms of defining the parameters for how to carry out a data collection. If, as part of a congressional hearing, the question of how many adults are in the United States with serious mental illness is raised, there has to be an answer to that question, and there have to be data that can back up that answer.

Manderscheid pointed out that there are significant opportunities for synchrony in this work that could lead to substantial progress. Federal agencies tend to work in isolation as do many researchers. The challenge is to overcome that isolation and create synchrony. Kathleen Merikangas (National Institute of Mental Health) agreed that there are a large number of similar data collections in the United States, all using different methods, and she underscored the benefits of coordination.

Manderscheid noted that some of the relevant concepts that have not been discussed as part of the workshop include resiliency, recovery, wellness, and well-being. He reiterated that these are major themes in the world of mental health and that they are also becoming more and more aligned with funding and policy initiatives. He said that it would be useful if the workshop and the larger study contributed to the discussion in this area.

Darrell Regier (Uniformed Services University) said that some of the current legislation is critical of the emphasis that SAMHSA has placed on recovery, and that some argue that the agency has not paid enough attention to severe mental illness and to finding ways of helping individuals with severe mental illness into treatment. It is also important to note that the patient perspective is gathering greater influence in the United States. It is not clear that this is the case in government and areas of research such as the Global Burden of Disease, which has a focus on pathology and impairment, as opposed to on strengths. Nonetheless, it is important to begin to focus on strengths and resiliency, which will help modify the field's understanding of the treatment or disability implications of individuals with these illnesses. Regier noted that it will be very interesting to see if, in the future, some of the national surveys will start adding measures of the concepts of recovery and resilience.

Regier also commented that as part of the development of the DSM-5 there was much debate about whether to use a resilience, strength-based approach. However, the decision was that more research and evidence of the importance of these concepts is needed. At that point it did not appear that there were sufficient data to be able to develop a scale.

Returning to the possible goal of being able to produce an estimate of prevalence of severe mental disorders in the United States, Regier suggested that a computerized adaptive testing (CAT)-type approach that combined several measures might produce a better assessment of prevalence rates than any single measure. Such an approach might include (1) disorder measures; (2) specificity measured with a scale such as the Composite International Diagnostic Interview; (3) distress measured with the K6, K8, or K10; and (4) a measure of severity. He added that if the goal is to cover 13 domains, and not just depression and anxiety, the only way to do so is by adaptive testing or a sequence similar to the one used in the DSM-5 field trials, with cross-cutting measures at Level 1 and Level 2, followed by severity measures. Going forward, it will be important for the national surveys to find ways of using the new approaches and technologies being developed.

James Jackson (University of Michigan) added that CAT-type approaches that essentially adapt tests to a particular individual appear to fit the way in which serious mental illness manifests itself, which is highly individualized; they also fit with the ideas driving developments in the area of precision medicine and individualized approaches. Robert Gibbons (University of Chicago) commented that the Educational Testing Service (ETS) has been using adaptive testing for 30 years in one form or another. However, the ETS CAT approach is based on unidimensional item response theory, which might work well for the measurement of mathematical ability but is limited when applied to multidimensional constructs such as depression.

Fred Conrad (University of Michigan) commented, first, that he does not think that it would be wise for SAMHSA to substitute data from electronic health records for self-reported data. He said that it seemed clear based on the discussions that electronic health records are not designed for producing population estimates, and if they could be used for that purpose, it would be in the distant future. Second, he suggested, if SAMHSA fields its own survey, it would have to be a mixed mode survey, because it would be best to collect the data offering mode choices that are convenient for respondents and fit with their preferred modes of communication. This approach would help address the concern regarding the high nonresponse rates among people who suffer from mental illness. Third, Conrad noted that population coverage appears to also be a concern. There are people who are not going to be included in sampling frames using traditional methods because they are homeless or institutionalized. Although only a very small proportion of all the homeless are believed to be chronically homeless, if the chronically homeless are much more likely to suffer from mental illness than the rest of the population, then the risk of coverage error is quite high. Methods like respondent-driven sampling or other techniques focused on hard-to-reach populations may be more useful than traditional sampling. Neil Russell (SAMHSA) agreed with Conrad that electronic health records are not yet well developed enough to be considered for use of this type and that the coverage error question also deserves further attention.

Nora Cate Schaeffer (University of Wisconsin) noted that SAMHSA has the challenge of preparing for the next round of the National Survey on Drug Use and Health (NSDUH), as well as the opportunity to think about what might be possible after the next survey and further in the future. Even if electronic health records cannot be used now, such an approach might be something that is important to begin preparing for now. She said that this is a field that is in constant transition, and the technology is not very nimble, which means that introducing testing of new approaches early might be useful, where possible.

Stephen Blumberg (National Center for Health Statistics) added that although electronic health records may not work as a sampling frame or as primary data collection, they may work well as sources of secondary data. For example, it may be possible to collect the information that is needed about serious mental illness in a survey, and then, with permission, link to electronic health records to see if there is a related diagnosis in the medical record. The electronic health records could also be used to check whether people who did not meet clinical significance for a disorder in a survey might have a diagnosis in their medical records.

Schaeffer, moving to a different topic, remarked that there was little discussion of group comparisons for the different measures. One of the promises of IRT in the 1980s was the possibility of estimating different parameters for different socioeconomic groups. As the United States becomes more diverse, linguistically, culturally, and in other ways, it seems that computer-adapted testing could be really useful in making less biased comparisons across groups.

Jonaki Bose (SAMHSA) noted that one of the main questions for SAMHSA is whether it is possible to identify impairment associated with specific mental disorders. Based on the discussion, it appears that the answer is that this cannot be done through data collection. One may be able to do so in retrospective analyses of the Global Assessment of Functioning, World Health Organization Disability Assessment Scale, or data on days of disability, but it does not appear to be feasible to incorporate a more direct measurement approach into the data collection process. She said that this conclusion was useful to learn.

Bose added that both of the points that Manderscheid made early on and that Schaeffer made about looking forward to future applications of electronic health records were helpful. Administrative records inherently have a lot of problems for estimation purposes, and SAMHSA is well aware of these. But it will be important for the agency to have a voice in the potential development of a system of electronic health records for the purposes of estimating prevalence of mental illness disorders, and there is value in looking 15 years out into the future.

She also noted that the discussions have given SAMHSA cutting-edge ideas to consider that may be useful. For example, she had not thought of the possibility of applying computerized adaptive testing for specific disorder-level measurement, even though much of her work has been in the education field, where this technique is frequently used.

On the issue of coverage bias, Bose said that it is something that is discussed by the SAMHSA team all the time. The NSDUH cannot be everything for everyone, she noted. Another important take-home message from the discussions is to continue to pay close attention to what populations are not included and continue to investigate alternate data sources. She added that another reality is the need to balance priorities in funding for the present survey administration and development for the future.

Bringing the discussion back to some of SAMHSA's original questions, Benjamin Druss (Emory University) asked the workshop participants to comment on the goals of producing state-level estimates, and, in particular, whether there are meaningful state-level differences in rates of either symptomology or functioning. Bose said that even though

SAMHSA's estimates of serious mental illness are model based, they do see that some states have higher rates than others.

Bose also asked the participants to comment on the practice of using cutoff points for severity at the 95th percentile, and whether these are metrics that stand independent of the distribution. Manderscheid replied that, 25 years ago, they anchored the cutoff at 5.8 percent. They tested this between states using modeling and did not find significant statistical variations. States still use this today, and if there is a change in their population, it would appear in the estimate. However, it is not clear whether this is still valid today.

Druss encouraged the participants to also weigh in on the ideal frequency for a potential survey, keeping in mind that one of the parameters specified by SAMHSA was to collect data no less frequently than every 5 years. Given that SAMHAS' flagship survey is the NSDUH, Druss asked that comments address the frequency needed for both mental illness and substance use data.

Theo Vos (University of Washington) suggested that larger changes would probably be seen over time in substance use disorders than mental health, so more frequent survey administration would be needed on substance use. He thinks every 5 years on either topic would be acceptable. For the Global Burden of Disease study, some countries provide data every 10 years, and they are happy even with that interval. He also added that some data on substance use are available through monitoring of overdose deaths. In his view, collecting data on either substance use or mental health with surveys that are based on a traditional design is challenging, due to the factors that have been discussed throughout the workshop.

Appendix A

Workshop Agenda

WORKSHOP ON INTEGRATING NEW MEASURES OF SPECIFIC MENTAL ILLNESS DIAGNOSES WITH FUNCTIONAL IMPAIRMENT INTO SAMHSA'S DATA COLLECTION PROGRAMS

The National Academies of Sciences, Engineering, and Medicine
Keck Center, Room 208
500 Fifth Street, NW
Washington DC 20001
September 24, 2015

9:00-9:20 **Welcome and Introductions**

Benjamin Druss, *Workshop Chair, Emory University*

Connie Citro, *Director, Committee on National Statistics*

9:20-9:40 **SAMHSA's Goals and Challenges Related to Measuring Specific Mental Illness Diagnoses with Functional Impairment**

D.E.B. Potter, ASPE

Neil Russell, *Director, Division of Surveillance and Data Collection,* CBHSQ, SAMHSA

9:40-10:00 **Historical Overview of the Data Needs Related to Measuring Specific Mental Illness Diagnoses with Functional Impairment**

Ron Manderscheid, *National Association of County Behavioral Health & Developmental Disability Directors and Johns Hopkins University Bloomberg School of Public Health*

10:00-11:00 **Update on a New National Institute of Mental Health Initiative**

Lisa Colpe, *Office of Clinical and Population Epidemiology Research, National Institute of Mental Health*

11:00-11:10 ***Coffee Break***

11:10-11:50 **Advantages and Disadvantages of Instruments Available for Measuring Specific Mental Illness Diagnoses with Functional Impairment**

Darrel Regier, *Center for the Study of Traumatic Stress*

11:50-12:30 **The Global Burden of Disease Study**

Theo Vos, *Institute for Health Metrics and Evaluation*

12:30-1:30 **Working Lunch to Continue Discussion of Measures**

Third Floor Atrium

1:30-2:0 **Identifying Adult Mental Disorders with Existing Data Sources**

Mark Olfson, *Columbia University*

2:00-2:20 **The National Health Interview Survey**

Stephen Blumberg, *National Center for Health Statistics*

2:20-2:45 **Using Computerized Adaptive Testing for Mental Health Assessment**

Robert Gibbons, *University of Chicago*

2:45-3:10 **The Patient Reported Outcome Measurement System**

David Cella, *Northwestern University*

3:10-3:20 ***Coffee Break***

3:20-4:40 **Panel Discussion**

Benjamin Druss, *Emory University*

Ron Manderscheid, *National Association of County Behavioral Health and Developmental Disability Directors and Johns Hopkins University Bloomberg School of Public Health*

Robert Krueger, *University of Minnesota*

Nora Cate Schaeffer, *University of Wisconsin*

Frederick Conrad, *University of Michigan*

4:40-5:30 **Floor Discussion and Wrap-Up**

Benjamin Druss, *Workshop Chair, Emory University*

5:30 **Adjourn Public Session**

Appendix B

Biographical Sketches of Steering Committee Members and Speakers

STEPHEN BLUMBERG ***(Speaker)*** is associate director for science in the Division of Health Interview Statistics at the National Center for Health Statistics of the Centers for Disease Control and Prevention. Previously, he was the lead statistician for the State and Local Area Integrated Telephone Survey, which regularly fields some of the world's largest telephone surveys on children's health, health care, and well-being, including the National Survey of Children with Special Health Care Needs and the National Survey of Children's Health. His research interests focus on survey strategies to identify vulnerable populations, such as children with special health care needs and children with autism spectrum disorder, and on the prevalence of wireless-only households and the impact of cell phones on coverage bias for telephone surveys. His honors include the 2008 young professional achievement award from the Coalition for Excellence in Maternal and Child Health Epidemiology, and the Warren J. Mitofsky innovators award from the American Association for Public Opinion Research (AAPOR). He has served as president of AAPOR's Washington-Baltimore chapter. He has a Ph.D. in social psychology from the University of Texas at Austin.

DAVID CELLA ***(Speaker)*** is chair of the Department of Medical Social Sciences at Northwestern University Feinberg School of Medicine and also holds positions there as director of the Center for Patient-Centered Outcomes at the Institute for Public Health and Medicine, and professor in the Ken and Ruth Davee Department of Neurology, and the departments of Medical Social Sciences, Preventive Medicine-Health and Biomedical

Informatics, Psychiatry and Behavioral Sciences, and Weinberg College of Arts and Sciences. He is the developer of the Functional Assessment of Chronic Illness Therapy Measurement System for outcome evaluation in patients with chronic medical conditions. He is also the principal investigator of the statistical coordinating center for the NIH Roadmap Initiative to build a Patient Reported Outcome Measurement Information System and the principal investigator of a contract to develop item banks for the clinical trials supported by the National Institute of Neurological Disorders and Stroke. He has a Ph.D. in clinical psychology from Loyola University Stritch School of Medicine.

LISA J. COLPE ***(Speaker)*** is chief of the Office of Clinical and Population Epidemiology Research in the Division of Services and Intervention Research at the National Institute of Mental Health. (NIMH) A captain in the U.S. Public Health Service, she has previously served as senior program management officer at the Substance Abuse and Mental Health Services Administration; assistant director for Roadmap coordination at the National Institutes of Health, overseeing the agency's Roadmap activities, and chief of the Psychopathology Risk and Protective Factors Research Program at NIMH. She is a clinical psychologist with postdoctoral training in epidemiology and survey methodology.

FREDERICK G. CONRAD ***(Member, Steering Committee)*** is a research professor at the Survey Research Center and director of the Program in Survey Methodology at the University of Michigan. His recent work has focused on respondents' understanding of survey questions, biases in respondents' judgments about the frequency of their behaviors, the effect of automatic progress feedback on respondents' willingness to continue filling out a questionnaire, and the decision to participate in a survey among potential respondents. He has a Ph.D. in cognitive psychology from the University of Chicago.

BENJAMIN G. DRUSS ***(Chair, Steering Committee)*** is professor and Rosalynn Carter chair in mental health in the Department of Health Policy and Management and director of the Center for Behavioral Health Policy Studies at the Rollins School of Public Health at Emory University. He is working to build linkages between mental health, general medical health, and public health. He works closely with the Carter Center Mental Health Program, where he is a member of the Mental Health Task Force and Journalism Advisory Board. His research focuses on improving physical health and health care among persons with serious mental disorders. He has received a number of national awards for his work, including the health services research senior scholar award from the American

Psychiatric Association and the Armin Loeb award from the Psychiatric Rehabilitation Association. He has served as an expert consultant to the Substance Abuse and Mental Health Services Administration, the Centers for Disease Control and Prevention, and the Assistant Secretary for Planning and Evaluation. He has an M.P.H. from Yale University and an M.D. from New York University.

ROBERT D. GIBBONS ***(Speaker)*** is the Blum-Riese professor in the Departments of Medicine and Public Health Sciences, and director of the Center for Health Statistics, all at the University of Chicago. His research and policy interests involve the development and application of statistics to problems in the behavioral, biological, and environmental sciences, in particular, on the use of statistics in addressing questions in health care policy and the development of new statistical methods for the analysis of clustered or longitudinal data. He is a member of the National Academy of Medicine (formerly, the Institute of Medicine), a fellow of the American Statistical Association, and an elected member of the International Statistical Institute. He has received lifetime achievement awards from the American Statistical Association, the American Public Health Association, and Harvard University. He has a Ph.D. in statistics and psychometrics from the University of Chicago.

ROBERT F. KRUEGER ***(Member, Steering Committee)*** is a distinguished McKnight university professor and a Hathaway distinguished professor and serves as director of clinical training in the Department of Psychology at the University of Minnesota. His research focuses on the classification and etiology of psychopathology and personality, using psychometric, quantitative and molecular genetics, and neuroscience approaches. He is a fellow of the Association for Psychological Science and of the American Psychopathological Association and a member of the Society for Multivariate Experimental Psychology. He has received a number of national and international awards, including the American Psychological Association's award for Early Career Contributions, the early career contributions award from the International Society for the Study of Individual Differences, and an American Psychological Foundation mid-career award. He is the editor of the *Journal of Personality Disorders*. He has a Ph.D. in psychology from the University of Wisconsin–Madison.

RON MANDERSCHEID ***(Member, Steering Committee, and Speaker)*** is the executive director of the National Association of County Behavioral Health & Developmental Disability Directors and adjunct professor in the Department of Mental Health at the Bloomberg School of Public Health at Johns Hopkins University. Previously, he served as the director of Mental

Health and Substance Use Programs at the Global Health Sector of SRA International; as chief of the Survey and Analysis Branch of the Center for Mental Health Services at the Substance Abuse and Mental Health Services Administration; and as chief of the Statistical Research Branch of the National Institute of Mental Health. He serves on the boards of the Employee Assistance Research Foundation, the Danya Institute, the FrameWorks Institute, the Council on Quality and Leadership, the International Credentialing and Reciprocity Consortium, and the National Research Institute. He is a former member of the Advisory Committee on Healthy People 2020. He has received numerous federal and professional awards, including, most recently, the American Public Health Association Carl A. Taube lifetime achievement award in mental health. He is an elected fellow of the American Academy of Social Work and Social Welfare. He has an M.A. in sociology-anthropology from Marquette University and a Ph.D. in sociology from the University of Maryland.

MARK OLFSON ***(Speaker)*** is professor of psychiatry at the Columbia University Medical School. He also serves as codirector of the Agency for Healthcare Research and Quality Center for Education and Research on Mental Health Therapeutics. Previously, he served as the scientific director of the TeenScreen National Center for Mental Health Checkups at Columbia University. His research interests focus on national patterns and trends in the utilization of mental health services and quality of care. He currently directs several studies on the delivery of mental health services in community settings, with an emphasis on the pharmacoepidemiology of psychotic and mood disorders. He has served as a consultant to the World Health Organization, the National Institutes of Health, and the American Psychiatric Institute for Research and Education. He has an M.P.H. from the Columbia University School of Public Health and an M.D. from Northwestern University.

D.E.B. POTTER ***(Speaker)*** is program analyst with the U.S. Department of Health and Human Services Office of the Assistant Secretary for Planning and Evaluation (ASPE). Previously she was a senior survey statistician at the Agency for Healthcare Research and Quality (AHRQ). She leads an ASPE, AHRQ, and Centers for Medicare & Medicaid Services joint project to develop risk adjustment methods for quality measures for home- and community-based services populations. Other responsibilities include managing the development of behavioral health quality measures and advancing quality measurement for the population with dementia. She serves on numerous technical expert panels and cross-agency workgroups. She has an M.S. in biostatistics from Georgetown University.

DARREL REGIER ***(Speaker)*** is senior scientist at the Center for the Study of Traumatic Stress in the Henry M. Jackson Foundation for the Advancement of Military Medicine and the Department of Psychiatry at the Uniformed Services University. He also serves as an independent senior scientific consultant to the American Psychiatric Association (APA) on DSM-5 and research-related issues. Formerly, he was APA's research director and director of the American Psychiatric Institute for Research and Education. For a substantial part of his career, he directed three research divisions in the areas of epidemiology, prevention, clinical research, and health services research at the National Institute of Mental Health. He contributed to the planning of the DSM-5 and to the National Advisory Mental Health Council's reports to Congress on mental health insurance parity. He recently completed 20 years as the American editor of *Social Psychiatry and Psychiatric Epidemiology*. He has a medical degree from the Indiana University School of Medicine.

NEIL RUSSELL (***Speaker***) is director of the Division of Surveillance and Data Collection in the Center for Behavioral Health Statistics and Quality at the Substance Abuse and Mental Health Services Administration. His areas of expertise include behavioral health statistics and epidemiology; basic and applied research in behavioral health data systems and statistical methodology; as well as surveillance and data collection. He has a Ph.D. in sociology from Arizona State University with a focus in survey research.

NORA CATE SCHAEFFER ***(Member, Steering Committee)*** is Sewell Bascom professor of sociology in the Department of Sociology at the University of Wisconsin–Madison and faculty director of the university's Survey Center, where she teaches courses in survey research methods and conducts research on questionnaire design and interaction during survey interviews. She currently serves as a member of the advisory boards of *Public Opinion Quarterly,* the American Association for Public Opinion Research, and of the Board of Overseers of the General Social Survey. She recently completed terms as the Council on Sections Representatives for the Survey Research Methods Section of the American Statistical Association and as a member of the Census Advisory Committee of Professional Associations. Schaeffer is a fellow of the American Statistical Association. She has an M.A. degree in urban studies from the University of Chicago, and a Ph.D. in sociology from the University of Chicago.

THEO VOS ***(Speaker)*** is a professor of global health at the Institute for Health Metrics and Evaluation (IHME) at the University of Washington. He is a member of the research team for the Global Burden of Disease study, which is coordinated by IHME. Prior to joining IHME, he was director of the Centre for Burden of Disease and Cost-Effectiveness at the School of Population Health of the University of Queensland. While there, he led burden of disease studies in Australia and contributed to studies in Malaysia, Singapore, South Africa, Thailand, Vietnam, and Zimbabwe. Previously, he led two large economic evaluation projects: the Assessing Cost-Effectiveness in Prevention project in Australia and the Setting Priorities Using Information on Cost-Effectiveness project in Thailand. He has a an M.Sc. in public health in developing countries from the London School of Hygiene and Tropical Medicine, a medical degree from State University Groningen, and a Ph.D. in epidemiology and health economics from Erasmus University.

COMMITTEE ON NATIONAL STATISTICS

The Committee on National Statistics was established in 1972 at the National Academies of Sciences, Engineering, and Medicine to improve the statistical methods and information on which public policy decisions are based. The committee carries out studies, workshops, and other activities to foster better measures and fuller understanding of the economy, the environment, public health, crime, education, immigration, poverty, welfare, and other public policy issues. It also evaluates ongoing statistical programs and tracks the statistical policy and coordinating activities of the federal government, serving a unique role at the intersection of statistics and public policy. The committee's work is supported by a consortium of federal agencies through a National Science Foundation grant.