Research White Paper

Outcome Measure Harmonization and Data Infrastructure for Patient-Centered Outcomes Research in Depression: Final Report

Agency for Healthcare Research and Quality

Research White Paper

Outcome Measure Harmonization and Data Infrastructure for Patient-Centered Outcomes Research in Depression: Final Report

Prepared for:
Agency for Healthcare Research and Quality
U.S. Department of Health and Human Services
5600 Fishers Lane
Rockville, MD 20857
www.ahrq.gov

Contract No. 75Q80119C00005

Prepared by:
OM1, Inc.,[1] with subcontractors:
American Board of Family Medicine[2]
American Psychiatric Association[3]
Baystate Health[4]
Elimu Informatics[5]

Investigators:
Michelle B. Leavy, M.P.H.[1]
Costas Boussios, Ph.D.[1]
Robert L. Phillips, Jr., M.D., M.S.P.H.[2]
Diana Clarke, Ph.D.[3]
Barry Sarvet, M.D.[4]
Aziz Boxwala, M.D., PhD.[5]
Richard Gliklich, M.D.[1]

AHRQ Publication No. 21(22)-EHC032
June 2022

This report is based on research conducted by OM1, Inc. with subcontractors the American Board of Family Medicine, American Psychiatric Association, Baystate Health, and Elimu Informatics under contract to the Agency for Healthcare Research and Quality (AHRQ), Rockville, MD (Contract No. 75Q80119C00005). This work is supported by the Office of the Secretary Patient-Centered Outcomes Research Trust Fund under Interagency Agreement 18-596R-18. The findings and conclusions in this document are those of the authors, who are responsible for its contents; the findings and conclusions do not necessarily represent the views of AHRQ. Therefore, no statement in this report should be construed as an official position of AHRQ or of the United States (U.S.) Department of Health and Human Services.

None of the investigators have any affiliations or financial involvement that conflicts with the material presented in this report.

The information in this report is intended to help healthcare decision makers—patients and clinicians, health system leaders, and policymakers, among others—make well-informed decisions and thereby improve the quality of healthcare services. This report is not intended to be a substitute for the application of clinical judgment. Anyone who makes decisions concerning the provision of clinical care should consider this report in the same way as any medical reference and in conjunction with all other pertinent information, i.e., in the context of available resources and circumstances presented by individual patients.

AHRQ appreciates appropriate acknowledgment and citation of its work. Suggested language for acknowledgment: This work was based on a Final Report, Outcome Measure Harmonization and Data Infrastructure for Patient-Centered Outcomes Research in Depression, by OM1, Inc., with subcontractors the American Board of Family Medicine, American Psychiatric Association, Baystate Health, and Elimu Informatics through the Agency for Healthcare Research and Quality (AHRQ).

Suggested citation: Leavy MB, Boussios C, Phillips RL, Clarke D, Sarvet B, Boxwala A, Gliklich R. Outcome Measure Harmonization and Data Infrastructure for Patient-Centered Outcomes Research in Depression: Final Report. (Prepared by OM1, Inc. and its subcontractors the American Board of Family Medicine, American Psychiatric Association, Baystate Health, and Elimu Informatics under Contract No. 75Q80119C00005.) AHRQ Publication No. 21(22)-EHC032. Rockville, MD: Agency for Healthcare Research and Quality; June 2022. DOI: 10.23970/AHRQEPCWHITEPAPERDEPRESSIONFINAL. Posted final reports are located on the Effective Health Care Program search page.

Acknowledgments

The authors gratefully acknowledge the following individuals for their contributions to this project: Eric Bickelman, Mingliang Dai, Baily Egan, Sarah Hajjar, Zachary Morgan, and Lars Peterson from the American Board of Family Medicine; Debbie Gibson and Barbara Casanova from the American Psychiatric Association; Jennifer McCaffrey and Jessica Wozniak from Baystate Health; Bob Dolin, James Shalaby, Hank Wallace, and Richard Wang from Elimu Informatics; and Pedro Alves, Daniel Brenner, Matthew Budday, Danielle Cooke, Gary Curhan, Andy DeMayo, Ashish Deshpande, Haley Friedler, Cameron Jones, Kathleen Laliberte, Adrian Law, Dan Levy, Kathleen Mortimer, Noa Palmon, Robin Pepper, Safiyy Momen, Zhaohui Su, Anna Swenson, Ia Topuria, and Ariella Yosafat from OM1.

Outcome Measure Harmonization and Data Infrastructure for Patient-Centered Outcomes Research in Depression: Final Report

Structured Abstract

Objective. The objective of this project was to demonstrate the feasibility and value of collecting harmonized depression outcome measures in the patient registry and health system settings, displaying the outcome measures to clinicians to support individual patient care and population health management, and using the resulting measures data to support patient-centered outcomes research (PCOR).

Methods. The harmonized depression outcome measures selected for this project were response, remission, recurrence, suicide ideation and behavior, adverse effects of treatment, and death from suicide. The measures were calculated in the PRIME Registry, sponsored by the American Board of Family Medicine, and PsychPRO, sponsored by the American Psychiatric Association, and displayed on the registry dashboards for the participating pilot sites. At the conclusion of the data collection period (March 2020-March 2021), registry data were analyzed to describe implementation of measurement-based care and outcomes in the primary care and behavioral health care settings. To calculate and display the measures in the health system setting, a Substitutable Medical Apps, Reusable Technology (SMART) on Fast Healthcare Interoperability Resource (FHIR) application was developed and deployed at Baystate Health. Finally a stakeholder panel was convened to develop a prioritized research agenda for PCOR in depression and to provide feedback on the development of a data use and governance toolkit.

Results. Calculation of the harmonized outcome measures within the PRIME Registry and PsychPRO was feasible, but technical and operational barriers needed to be overcome to ensure that relevant data were available and that the measures were meaningful to clinicians. Analysis of the registry data demonstrated that the harmonized outcome measures can be used to support PCOR across care settings and data sources. In the health system setting, this project demonstrated that it is technically and operationally feasible to use an open-source app to calculate and display the outcome measures in the clinician's workflow. Finally, this project produced tools and resources to support future implementations of harmonized measures and use of the resulting data for research, including a prioritized research agenda and data use and governance toolkit.

Conclusion. Standardization of outcome measures across patient registries and routine clinical care is an important step toward creating robust, national-level data infrastructure that could serve as the foundation for learning health systems, quality improvement initiatives, and research. This project demonstrated that it is feasible to calculate the harmonized outcome measures for depression in two patient registries and a health system setting, display the results to clinicians to support individual patient management and population health, and use the outcome measures data to support research. This project also assessed the value and burden of capturing the measures in different care settings and created standards-based tools and other resources to support future implementations of harmonized outcome measures in depression and other clinical areas. The findings and lessons learned from this project should serve as a roadmap to guide future implementations of harmonized outcome measures in depression and other clinical areas.

Contents

Tables

Figures

Appendixes

1. Introduction

Background

Patient registries provide valuable information to describe the course of a disease, understand treatment patterns and outcomes, examine the effectiveness, safety, and value of products and interventions, and measure and improve quality of care. A patient registry is defined as "an organized system that uses observational study methods to collect uniform data (clinical and other) to evaluate specified outcomes for a population defined by a particular disease, condition, or exposure and that serves one or more pre-determined scientific, clinical, or policy purposes."[1]

Patient registries fulfill different purposes for a wide range of stakeholders, as documented in the publication, *Registries for Evaluating Patient Outcomes: A User's Guide*.[1] For clinicians, registries that collect data on disease course and outcomes in large patient populations can provide information on current treatment practices and outcomes to inform decision-making. Clinicians may participate in registries to engage in research, complete maintenance of certification requirements, and/or collect and report quality measures, such as those required under the Merit-based Incentive Payment System (MIPS). Recently, many efforts in the United States have focused on the potential role of registries as the foundation of research data infrastructure and learning health systems. Registries can be a central component of these systems by providing data and tools to support population health management, clinical decision-making, quality improvement, research, and collection of patient-reported outcomes (PROs). Many professional associations have developed registries to support the needs of their clinicians. For example, the PRIME Registry[2] and PsychPRO[3] are designed to help clinicians meet quality reporting requirements under MIPS and track and manage patient outcomes through progress reports and clinical decision support tools. In some cases, registries such as these have provided data for research purposes as well.

While many registries serve multiple purposes, some are designed specifically for clinical research and safety surveillance. Drug and device manufacturers use registries to inform the development of new products by gathering information on treatment patterns and patient populations. For marketed products, registries can provide real-world evidence (RWE) of product performance to support reimbursement decisions and help manufacturers meet post-marketing commitments. The passage of the 21st Century Cures Act[4] in 2016 generated new interest in registries as a source of real-world data and RWE to inform regulatory decision-making. The U.S. Food and Drug Administration (FDA) identified registries as a source of real-world data in its framework for RWE; some registries have already been used as a source of real-world data to support regulatory decision-making. The FDA also uses data from registries under the National Evaluation System for Health Technology (NEST) project.[5] While typically not registry sponsors, public and private payers use registry data to track how devices, procedures, or pharmaceutical products are used in practice and to monitor effectiveness in different populations; of note, the Centers for Medicare and Medicaid Services (CMS) uses registry data for decisions under the Coverage with Evidence Development program.[6]

Other sponsors and users of registry data include patient advocacy groups, academic researchers, and public health professionals. Patient advocacy groups may sponsor or participate in registry development to increase understanding of the natural history of a disease and to support efforts to develop new treatments; this is particularly common for rare diseases. Academic researchers use registries for a wide range of purposes, such as tracking long-term patient outcomes, examining the effectiveness or comparative effectiveness of procedures or therapies, investigating genetic or environmental factors related to specific diseases, or examining the role of new technologies. For public health professionals, registries provide an important tool for monitoring prevalence and incidence of diseases and tracking the impact of public health interventions.

Given their myriad purposes, it is unsurprising that a large number of registries exist – over 7,300 according to ClinicalTrials.gov. Together, registries represent an enormous investment in research infrastructure and a tremendous data resource that could be used to address new research questions in a timely and efficient manner. Registries also occupy a unique role in the health data landscape, in that they capture data across all components of a learning health system. Registries provide a bridge connecting research and clinical practice, and they can offer tools to support clinical decision-making at the individual patient level and data to support population health management and quality improvement initiatives. Yet, the value of registries as a foundation for research data infrastructure and learning health systems is currently limited by the variation in the data collected in different registries, even within the same clinical areas. This variation makes it more challenging to reuse registry data for other purposes, and at the same time increases the burden of data collection at the clinician and registry level.

The development and implementation of core sets of standardized outcome measures in patient registries and clinical care would address these challenges and enable registries to realize their potential as the foundation for learning health systems and research data infrastructure. For example, use of standardized outcome measures in registries would create opportunities to compare, link, and aggregate registry data to address new research questions, while use of standardized outcome measures in clinical practice would create opportunities to compare outcomes across care settings and better compare the outcomes achieved in real-world settings with those reported in research studies. To realize this vision, the Agency for Healthcare Research and Quality (AHRQ) has supported the creation of the Outcome Measures Framework (OMF), a conceptual model for classifying outcomes that are relevant to patients and clinicians across most conditions,[7] and the use of the OMF to develop standardized outcome measures in five clinical areas.[8-13]

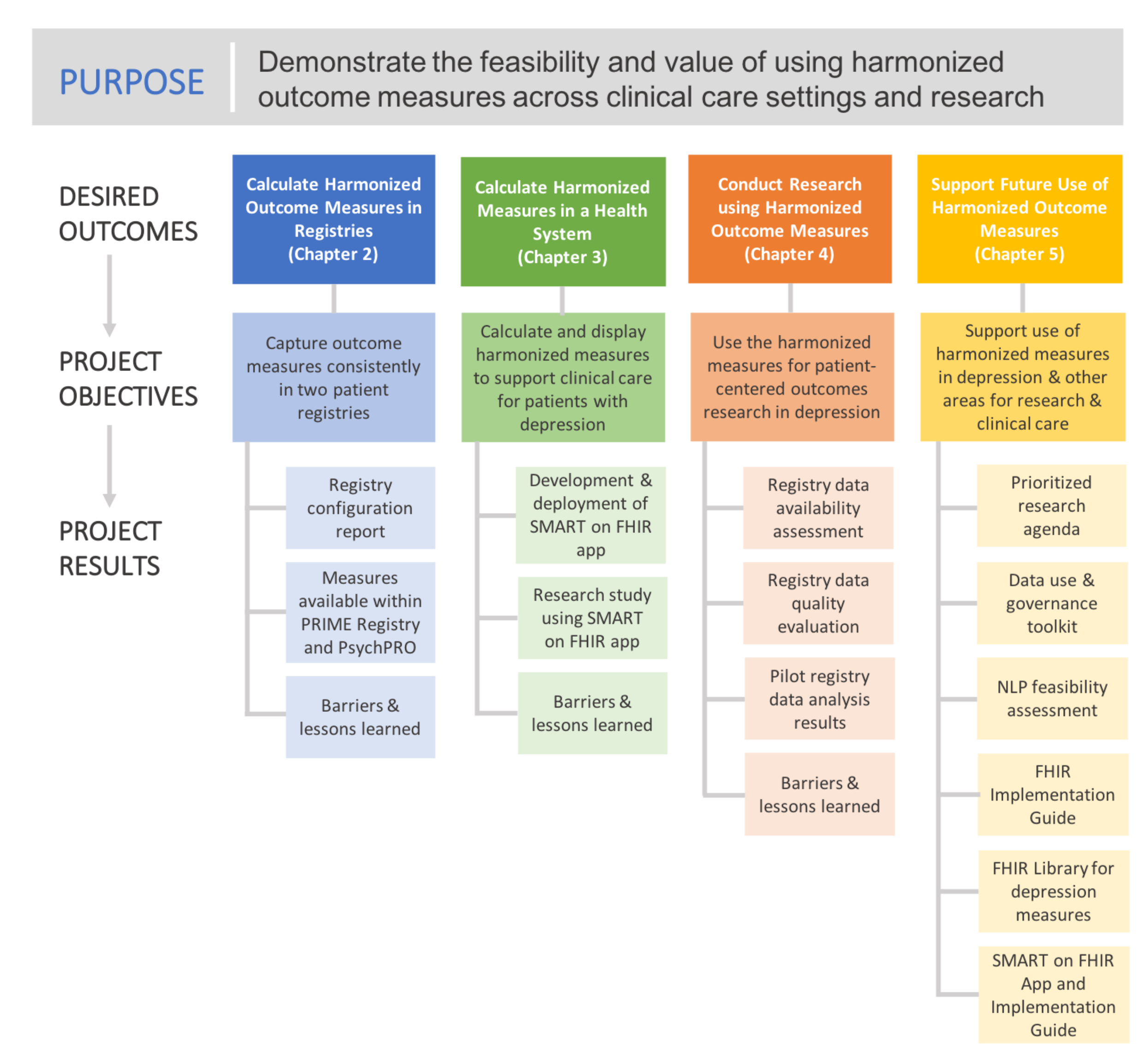
The purpose of this project was to assess the feasibility and value of implementing standardized outcome measures in multiple care settings, using the harmonized depression outcome measures as a test case.[11] Major depressive disorder (MDD) is a common mental disorder that affects an estimated 16.2 million adults and 3.1 million adolescents in the United States.[14] Characterized by changes in mood, cognitive function, and/or physical function that persist for two or more weeks, MDD can reduce quality of life substantially, impair function at home, work, school, and in social settings, and result in increased mortality.[15] Research on depression diagnosis, treatments, and outcomes is complicated by heterogeneity in care settings and treatment approaches. Clinicians use different instruments, such as the Patient Health Questionnaire-9 (PHQ-9)[16] and the Hamilton Depression Rating Scale (HAM-D),[17] to assess symptom severity and different definitions and different timeframes to measure concepts such as remission, response to treatment, and recurrence.

The harmonized depression outcome measures developed under the prior project (Appendix A) provide a core set of harmonized definitions intended for use in routine clinical care across care settings and to support patient-centered outcomes research.

Objectives

The purpose of this project was to demonstrate the feasibility and value of collecting the harmonized depression outcome measures in the patient registry and health system settings, displaying the outcome measures to clinicians to support individual patient care and population health management, and using the resulting measures data to support patient-centered outcomes research. The project purpose, primary and secondary objectives, and resulting products are presented in Figure 1.

Figure 1. Project purpose, objectives, and results


PURPOSE
Demonstrate the feasibility and value of using harmonized outcome measures across clinical care settings and research
DESIRED OUTCOMES
PROJECT OBJECTIVES
PROJECT RESULTS
Calculate Harmonized Outcome Measures in Registries (Chapter 2)
Calculate Harmonized Measures in a Health System (Chapter 3)
Conduct Research using Harmonized Outcome Measures (Chapter 4)
Support Future Use of Harmonized Outcome Measures (Chapter 5)
Capture outcome measures consistently in two patient registries
Calculate and display harmonized measures to support clinical care for patients with depression
Use the harmonized measures for patient-centered outcomes research in depression
Support use of harmonized measures in depression & other areas for research & clinical care
Registry configuration report
Measures available within PRIME Registry and PsychPRO
Barriers & lessons learned
Development & deployment of SMART on FHIR app
Research study using SMART on FHIR app
Barriers & lessons learned
Registry data availability assessment
Registry data quality evaluation
Pilot registry data analysis results
Barriers & lessons learned
Prioritized research agenda
Data use & governance toolkit
NLP feasibility assessment
FHIR Implementation Guide
FHIR Library for depression measures
SMART on FHIR App and Implementation Guide

Organization of This Report

This final report describes the technical approaches used in this project, the project findings, and the barriers encountered and lessons learned. This report may serve as a roadmap for future implementations of the harmonized depression outcome measures and harmonized outcome measures in other clinical areas. The report is organized into chapters that align with the project objectives, as presented in Figure 1. Specifically:

- Chapter 2 discusses the calculation of the measures within the patient registry setting.
- Chapter 3 discusses the calculation of the measures within the health system setting using a Substitutable Medical Apps, Reusable Technology (SMART) on Fast Healthcare Interoperability Resource (FHIR) application.
- Chapter 4 describes the use of the measures as implemented in the patient registries for conducting patient-centered outcomes research.
- Chapter 5 describes tools to support future implementations of the measures.
- Chapter 6 reviews lessons learned and discusses implications for future implementations of the depression measures and measures in other clinical areas.
- Chapter 7 describes the project conclusions.

In addition to the content included in this report, artifacts and tools that were produced under this project may be useful for future measure outcome implementations. These are listed in Appendix B.

2. Calculation of Harmonized Depression Outcome Measures in Registries

Introduction

Calculation of the harmonized outcome measures in patient registries is an important step toward creating a robust, national-level data infrastructure that could be used to support patient-centered outcomes research in depression. This task tested the feasibility of calculating six of the harmonized outcome measures developed under the prior project[11] in two patient registries. The primary objective was to demonstrate the feasibility and value of capturing the harmonized depression outcome measures in the clinical workflow and submitting these data to two patient registries. Each registry enrolled a diverse set of pilot sites to gain a broad perspective on feasibility and value. The project team assessed barriers to measure calculation across care settings and identified lessons learned that could inform future implementations of the harmonized depression outcome measures in patient registries and clinical practice settings. The methods, results, barriers, and lessons learned are discussed in the following sections.

Methods for Calculating the Measures in Patient Registries

Two patient registries participated in this project. The PRIME Registry, sponsored by the American Board of Family Medicine (ABFM), was established to provide family physicians and primary care clinicians with tools to evaluate practice performance, support population health and risk stratification, improve primary care practice and patient outcomes, and reduce the burden of reporting for the Centers for Medicare and Medicaid Services (CMS) payment programs. The PRIME Registry has over 2,500 active clinicians participating from 47 States and data on 5.4 million patients. PsychPRO, sponsored by the American Psychiatric Association (APA), was established to help psychiatrists and mental health professionals achieve optimal patient outcomes using tools to measure, chart, and benchmark clinical care, validate quality patient care through measurement and analysis, and avoid payment penalties and instead achieve bonuses for meeting quality reporting requirements for the CMS payment programs. PsychPRO has over 600 active clinicians participating from 46 states and data on over 180,000 patients. Each registry recruited pilot sites to participate in this project, for a total of 21 pilot sites (10 from PsychPRO and 11 from the PRIME Registry).

Both registries captured individual-level clinical data that was generated and documented during the course of patient treatment and care. Data were electronically extracted directly from electronic health records (EHRs) and from online portals. Data fields and elements varied with respect to standardization and included structured and unstructured data. Data included patient demographics; diagnosis(es) and intervention(s) (e.g., medications, therapy); encounter data; PROs; and limited clinician details. Data were collected during routine assessment and clinical care of patients and used primarily to support a practice's quality improvement activities and quality reporting to the CMS. Only de-identified data were secondarily used for research, including this project. As such, patient informed consent and institutional review board (IRB) approval were not required. However, participating sites were not precluded from seeking IRB approval.

The six measures selected for this project were response, remission, recurrence, suicide ideation and behavior, adverse effects of treatment, and death from suicide. The rationale for the selection of these measures and the measure definitions are described in a separate publication[11] and summarized in Appendix A of this report. To calculate the measures within these two patient registries, we used the standardized definitions developed for each measure.[18] The standardized definitions described the initial population for measurement (e.g., all depression patients), the outcome focused population (patients who experienced the outcome of interest), and the data criteria and value sets for each measure. The purpose of the standardized definitions was to enable the measures to be extracted and calculated consistently using data from EHRs and other sources and used for multiple purposes, including clinical care and research.

The registries used the standardized definitions to add the harmonized measures to their respective registry platforms. The registries compared the measure definitions to the registry data dictionaries and identified and addressed questions related to the measure definitions, availability of data, and practice workflow. Once the measures were added to the registry platforms, they appeared in the registry dashboard views at the participating pilot sites, and pilot sites were able to view the continually-updated measures throughout the data collection period (March 15, 2020 – March 15, 2021).

More information on the registry configuration process, the pilot sites, and the resulting display of the harmonized outcome measures on the registry dashboards is available in the Report on Registry Configuration published by AHRQ in November 2020.[19]

Findings

Calculation of the harmonized outcome measures within the PRIME Registry and PsychPRO was feasible, but technical and operational barriers needed to be overcome to ensure that relevant data were available and that the measures were meaningful to clinicians. Broadly, these barriers can be grouped into four categories: availability of data in EHR systems; structure of data in EHR systems; collection and documentation of the PHQ-9; and COVID-19 related disruptions in care.

Availability of Measure Data Elements

As noted above, the PRIME Registry and PsychPRO extracted data from participating sites' EHR systems and other data collection systems. Calculation of the measures was limited by the availability of the measure data elements within these systems. While a key goal of this project was to create infrastructure to calculate the harmonized outcome measures in registries, some necessary data could not be captured consistently for two measures, as discussed below.

Adverse Events

Both registries were designed primarily to support quality improvement activities (rather than research) and rely on secondary use of data collected for routine clinical care. As such, many sites participated without seeking IRB approval, and patients were not required to provide informed consent. Any change in the registry data collection that resulted in the need for IRB approval and possibly informed consent at the site level would have introduced substantial burden and reduced the sustainability of the registries. Thus, the registries indicated that it was critical to only request data that were captured as part of providing routine care for patients with depression.

The intent of the adverse events measure was to capture all adverse events related to depression treatment. While some adverse events were documented in patients' medical records, it was possible that patients experienced other side effects that they either did not discuss with their clinicians or which were not noted because they were not significant enough to result in treatment changes. The harmonization workgroup recommended "use of a brief, publicly available, validated measurement tool to capture adverse events" as a way to supplement the data found in the medical record.[11] Specifically, the group suggested the Frequency, Intensity, and Burden Side Effects Rating Scale (FIBSER).[20] The FIBSER is a short, three-item patient-reported questionnaire that documents the frequency, intensity, and burden of side effects.

Through discussions with the registries and pilot sites, the project team learned that no sites currently used the FIBSER to capture adverse events. While there is evidence of use of the FIBSER in research settings, there is little to no evidence in the peer-reviewed literature of use of the FIBSER in routine clinical care. Because the FIBSER was not routinely used and would have been added for the purposes of this study only, both registries expressed concerns that use of the FIBSER would require IRB approval of this study as human subjects research and possibly require patient informed consent. This would have introduced substantial burden for participating sites. Because of this concern, the FIBSER was not collected through the registries, and FIBSER scores were unavailable to support calculation of the adverse events measure. Further work is needed to explore the utility of the FIBSER in routine clinical care.

This project also assessed the feasibility of calculating the adverse events measure using structured data in the health system setting (see Chapter 3) and using data extracted from clinical notes (see Chapter 5).

Death From Suicide

The registries noted concerns about the availability of the necessary data to calculate the death from suicide measure. Death may not be recorded in the EHR, and even when the date of death was recorded, information on the cause of death often was not available. In addition, in cases of suicide, a different cause of death may be listed because of the perceived stigma of suicide. While death from suicide is important to measure, it is equally important to ensure that all necessary data are captured and deaths are not underreported. Because of the practical challenges of systematic ascertainment and the limitations of mortality data in EHRs, the death from suicide measure was not included in the registry data analysis. Further work is needed to improve the documentation of mortality in EHRs, possibly through linkages to other data sources.

Structure of Data in EHR Systems

While some data are not recorded routinely in EHRs (e.g., cause of death), other data are recorded in formats that make extraction and analysis of the data difficult. Specifically, some data that are relevant for calculation of the harmonized outcome measures may be found in unstructured clinical notes, as opposed to in structured fields. The registries raised this concern in relation to the suicide ideation and behavior measure, which requires data on nonfatal suicide attempts/suicide attempt behaviors, planning/preparatory acts, and active suicide ideation. While structured codes for these concepts exist, a 2015 review of EHR data from primary care practices found that only 3 percent of patients with documentation of suicide ideation in unstructured clinical notes had a corresponding ICD-9 code, and only 5 percent of patients who indicated suicide ideation on the PHQ-9 (item 9) had a corresponding ICD-9 code. For suicide attempt, only 19 percent of patients with a suicide attempt documented in the notes had a corresponding ICD-9 code.[21] The findings from the 2015 review may not be generalizable to all sites participating in this pilot, particularly the PsychPRO sites, but they do support the concerns expressed by the registries about inconsistency in the use of structured codes for documentation of suicide ideation and behavior.

The suicide ideation and behavior measure also may be calculated using a patient’s response to item 9 of the PHQ-9 (‘Thoughts that you would be better off dead, or thoughts of hurting yourself in some way?’). This approach was used to calculate the measure in the patient registries’ dashboards for this pilot project. However, there were challenges to the use of this approach as well, notably the issue of extracting item-level scores from the pilot practices’ EHR systems. Item-level scores were available for patients who completed the PHQ-9 through the registry PRO platforms, but for patients who completed the PHQ-9 in an office setting or using other tools, documentation of the item 9 responses required the pilot sites to set up an appropriately named custom field in the EHR and modify their workflow to enter the item 9 data in the field. For the pilot project, this was addressed through training and ongoing communication with the sites throughout the data collection period. However, this issue may be more challenging as practices adopt the harmonized measures outside of the framework and support of the pilot study.

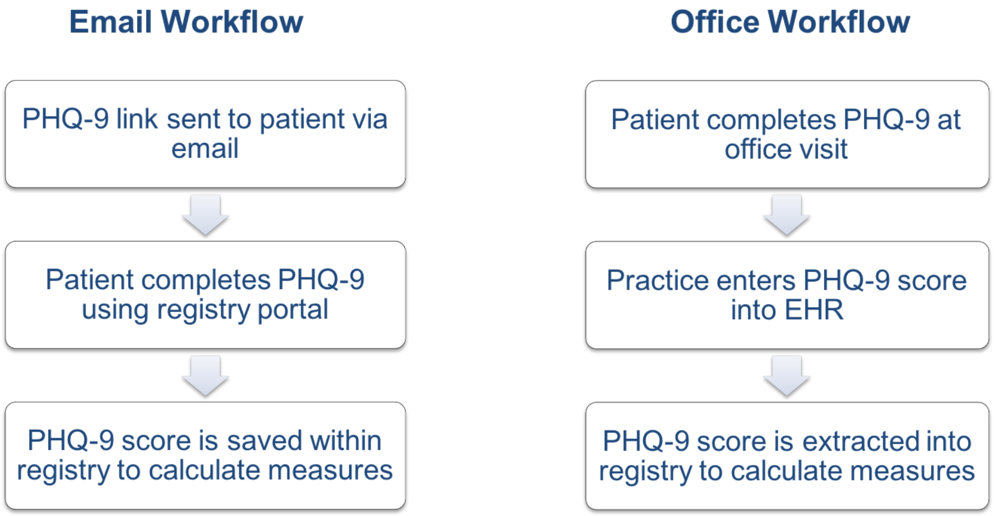
Further work to compare the data on suicide ideation and behavior recorded in structured form to the data recorded in clinical notes at a diverse set of clinical practice settings would be useful in identifying the scope of these challenges. In addition, future implementations of the harmonized measures may benefit from using natural language processing (NLP) approaches to extract suicide ideation and behavior concepts from clinical notes; these approaches are discussed further in Chapter 5.

Finally, information on adverse effects of treatment may be found in clinical notes, as opposed to structured fields in the EHR, particularly for side effects that are burdensome to patients but not significant enough to require changes in treatment. Future implementations of the harmonized measures may benefit from using NLP approaches to extract data relevant to this measure (see Chapter 5).

Collection and Documentation of the PHQ-9

Four of the six harmonized measures – remission, response, recurrence, and suicide ideation – rely on the PHQ-9, making collection and documentation of the PHQ-9 total score and item-level scores a central component of this project. Practices participating in this project used two workflows to collect and document the PHQ-9, as shown in Figure 2.

Figure 2. Workflows for collection and documentation of the PHQ-9


Email Workflow
Office Workflow
PHQ-9 link sent to patient via email
Patient completes PHQ-9 using registry portal
PHQ-9 score is saved within registry to calculate measures
Patient completes PHQ-9 at office visit
Practice enters PHQ-9 score into EHR
PHQ-9 score is extracted into registry to calculate measures

The project encountered two barriers to the consistent collection and documentation of the PHQ-9: completion of the PHQ-9 remotely (outside of an office visit); and documentation of the PHQ-9. These are discussed further below.

Completion of the PHQ-9 Remotely

A critical component of standardization of outcome measures is standardization of the timeframe for measurement. Comparisons of outcome measures across data sources is most meaningful when outcomes are measured on consistent schedules. For the depression outcome measures, this requires completion of the PHQ-9 at regular intervals of, at minimum, every 6 months. Completion of the PHQ-9 more frequently is encouraged to support measurement-based care.

The PsychPRO sites participating in this project used an existing workflow to capture the PHQ-9, in which patients received a link to complete the PHQ-9 ahead of a scheduled visit. The PRIME Registry developed a workflow in which eligible patients received a link to complete the PHQ-9 approximately 6 to 8 weeks after an office visit. This follow-up PHQ-9 was sent directly to the patient via email, with the goal of enabling clinicians to monitor a patient's depression symptoms (and possibly response to treatment) without requiring another office visit. Patients who did not respond to the initial email with a link to complete the PHQ-9 received follow-up email reminders.

This workflow was designed to allow patients to complete the PHQ-9 directly through the registry patient portal, and it relied on patients' willingness to complete the PHQ-9 outside of an office visit. However, many patients did not respond to the emailed invitation to complete the PHQ-9. Approximately 6 percent of patients who received the email link completed the PHQ-9 during the data collection period. All patients who opened the email completed the PHQ-9, suggesting that the issue was not due to difficulty with the registry patient portal or the format of the PHQ-9. Other possible explanations include use of outdated email addresses or lack of awareness about the importance of completing the PHQ-9. Future implementations of the harmonized measures must consider how to improve response rates for PHQ-9s captured remotely.

Documentation of the PHQ-9

Extraction of the PHQ-9 score data presented some challenges during this project. Sites participating in the pilot administered the PHQ-9 in several ways; some sites used the registry patient portal, some captured the PHQ-9 through their EHR, and some captured the PHQ-9 on paper and scanned a copy of the instrument into their EHR. The registries had ready access to the PHQ-9s completed through the patient portals. For other PHQ-9s, the data must be extracted from the EHR to support calculation of the harmonized measures. In PsychPRO, registry sites used a custom field within the EHR to document the PHQ-9 score, and the registry was able to extract the summary score for measurement purposes. Documentation of the item 9 responses was a new requirement for this project, and the pilot sites needed to set up an appropriately named custom field in the EHR and modify their workflow to enter the item 9 data in the field. Extraction of the data was feasible technically; the larger challenge was the implementation of a consistent workflow at the practice site to ensure that the PHQ-9 summary score and item 9 scores were documented in the appropriate fields for each patient.

A second challenge related to documentation of the PHQ-9 was the use of different modalities at the practice level and at the individual patient level. At the outset of this project, the workflows were designed with the expectation that practices would use one modality (e.g., collection through the EHR, collection on paper and entered into the EHR, or collection through the registry patient portal). Throughout the project, it became clear that many practices used multiple approaches to collect the PHQ-9, even from the same patient. For example, a patient may have completed a PHQ-9 on paper during an office visit, then completed a PHQ-9 sent via email as a follow-up, and then completed another PHQ-9 on paper at a subsequent office visit. Future implementations of the harmonized measures should plan to collect the PHQ-9 using multiple modalities and to extract the PHQ-9 data from multiple locations (e.g., registries, EHR structured fields, EHR unstructured notes, standalone PRO systems).

Finally, it should be noted that the PsychPRO and PRIME Registry sites that elected to participate in this project were either already using the PHQ-9 as part of providing routine clinical care or interested in doing so as part of a broader effort to improve care and outcomes for patients with depression. As a result, adoption of the PHQ-9 was not identified as a barrier to calculation of the measures in this project. However, future implementations of the measures in other settings may encounter challenges around PHQ-9 adoption with practices that use other validated instruments, such as the Geriatric Depression Scale[22] or Quick Inventory of Depressive Symptomatology (QIDS),[23] rather than the PHQ-9. While the measure definitions allow for use of other instruments provided a crosswalk is available, very few crosswalks are currently available. Further work is needed in this area to develop crosswalks to link scores for the PHQ-9 and other relevant instruments.

More information on these barriers can be found in the Report on Registry Configuration published by AHRQ in November 2020.[19]

Disruptions in Care Due to COVID-19

In addition to the challenges noted above, the project encountered unexpected challenges due to the COVID-19 pandemic and the resulting disruptions in routine care. Data collection for this project began on March 15, 2020, and the pilot practices almost immediately began transitioning to telehealth or, in some cases, closing temporarily. Routine visits were postponed for many patients, as COVID-19 and other urgent care took priority. Telehealth use has remained widespread for some types of primary care visits and for many mental health visits. The disruptions related to COVID-19 may have affected this project in several ways. First, the number of office visits at many participating practices dropped substantially starting in mid-March and remained low through multiple waves of COVID-19 outbreaks in different locations over the summer. This may have affected collection of the baseline PHQ-9, resulting in lower numbers for the pilot data analysis (see Chapter 4). In addition, it is unclear whether other changes related to COVID-19 (e.g., telehealth instead of in-person, changes in the practice workflow to reduce infection risk) affected communication with patients about the importance of completing the PHQ-9 remotely to support depression care.

Assessment of Value and Burden

Some clinicians participating in this effort provided feedback at the conclusion of the data collection period. These clinicians noted that the PHQ-9 and harmonized measures are useful tools to assess and monitor patients individually, while the aggregate, practice-level measure calculations were less useful from a clinical decision-making perspective. Future implementations may benefit from focusing on presentation of the individual-level measure results and PHQ-9 scores to clinicians. In addition, clinicians noted the need for flexibility in obtaining and documenting PHQ-9 scores (in person, via phone, remotely).

Lessons Learned

Calculation of the harmonized outcome measures in two patient registries produced several lessons learned. First, some data that were critical for calculating the harmonized outcome measures were recorded routinely and available through EHRs, but only as unstructured text found in clinical notes. Future implementations of the harmonized outcome measures should plan to create structured EHR fields and associated process changes to capture the information, wherever possible. The use of robust NLP approaches to extract relevant data from notes and convert it into structured variables may be necessary in some settings. This will require additional time and resources to validate and deploy NLP tools, but the resulting data will provide a more complete view of patient outcomes over time. Work in other tasks on this project demonstrated that extraction of relevant data (PHQ-9 scores, adverse effects of treatment, suicide ideation and behavior) from unstructured text using NLP tools is feasible (Chapter 5). The findings from that work should be applied to future implementations of the measures.

A second challenge encountered in this project was the low completion rate for PHQ-9s sent to patients via email. More work is needed to understand why patients do not respond to emailed links to the PHQ-9 and how to improve response rates. Future implementations of the measures may benefit from testing trusted, reliable ways of delivering the PHQ-9 so that patients recognize and trust the email message and understand that their clinician wants them to complete the assessment. This may take multiple forms, including (1) improving tools to communicate with patients about the importance of completing PHQ-9, such as revising the text that is sent with the PHQ-9 link; (2) improving methods for contacting patients, such as offering a text option or adjusting the frequency of emails and reminders; and (3) considering other ways to encourage participation, such as development of a patient view in the registry platform to allow patients to track their progress. Collaboration with patients and/or patient organizations would also be beneficial to discuss barriers, identify possible solutions, and improve messaging.

Finally, future implementations of the harmonized measures should plan to collect the PHQ-9 using multiple modalities and to extract the PHQ-9 data from multiple locations (e.g., registries, EHR structured fields, EHR unstructured notes, standalone PRO systems). Clinicians use multiple tools to collect the PHQ-9, and collection and documentation practices may vary at the level of an individual patient. As an example, an individual patient may have PHQ-9 scores entered into the EHR as free text, attached as a scanned document, and entered directly through a patient portal. While this project likely encountered increased variability in PHQ-9 documentation practices due to COVID-19 related disruptions in care, it is likely that variability in documentation will persist. Future implementation efforts should plan to identify all possible locations for PHQ-9 scores and extract the scores whenever possible.

3. Calculation of Harmonized Depression Outcome Measures in a Health System

Introduction

The harmonized depression outcome measures are intended for use in research and clinical practice settings. Collection and calculation of the measures in a clinical practice setting is important to ensure that the necessary data are captured and documented at the point of care, so the data are available for clinicians providing patient care and for use in research, quality measurement, and population health efforts. This task tested the feasibility of calculating and displaying a subset of the harmonized outcome measures[11] within the standard clinician workflow using a Substitutable Medical Apps, Reusable Technology (SMART) on Fast Healthcare Interoperability Resource (FHIR) application ('app').

The objectives of this task were to design and develop a SMART on FHIR app to calculate and display the harmonized outcome measures, deploy the app within a health system that includes both primary care and mental health care practices, and assess the value and burden of collecting and calculating the harmonized measures within routine clinical practice settings. The project team also assessed technical and operational barriers and identified lessons learned that could inform future implementations of the harmonized depression outcome measures in clinical practice settings. The methods, results, barriers, and lessons learned are discussed in the following sections.

Design and Development of SMART on FHIR App

The purpose of the Major Depression Outcomes app is to combine clinical information with PROs to give clinicians a 'snapshot' view of an individual patient's longitudinal depression symptoms, treatments, and outcomes. A key goal of the app is to enable clinicians to view the outcome measures within the EHR, without the burden of having to log in to a separate system. The app uses the FHIR standard to retrieve data from the EHR system and leverages the tools developed under a prior federally-funded project to support the electronic capture and exchange of PRO data using the FHIR standard.[24, 25]

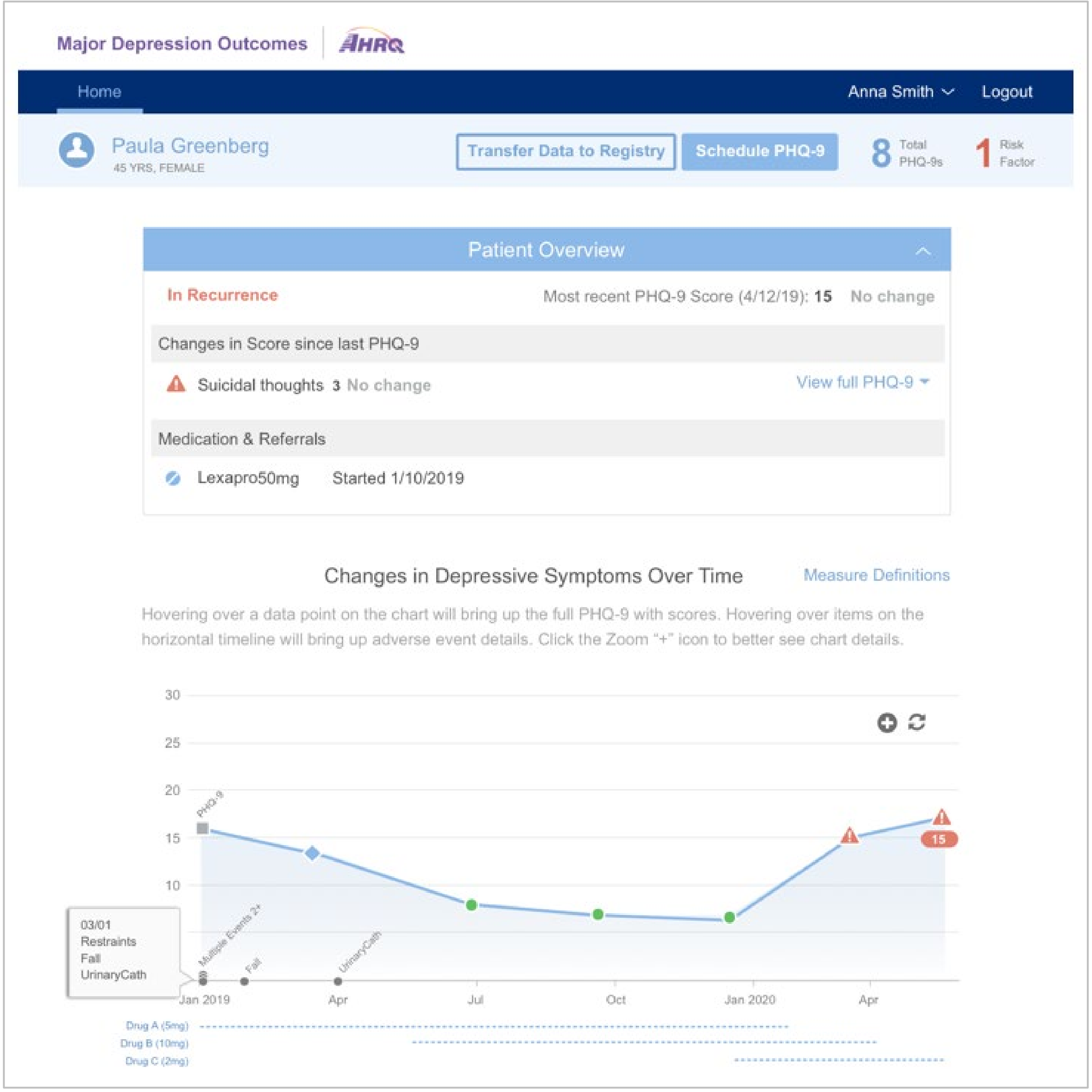
The OM1 team designed the app in consultation with Elimu Informatics. Elimu developed and tested the app and provided technical support for the implementation of the app at Baystate Health. The project team sought feedback on the app design from co-investigators on this project, other clinical and technical subject matter experts, the Stakeholder Panel for this project, and AHRQ. More information on the app technical requirements, design, data flow, development and testing process, and implementation steps can be found in the App Implementation Guide (see complete list of project artifacts in Appendix B).

The app comprises three components. The primary component is the outcome measures dashboard. Two optional supporting services are available to facilitate collection of the PHQ-9 and transferring data to registries.

Outcome Measures Dashboard

The outcome measures dashboard is a platform to calculate and visualize depression treatment and outcome measures for an individual patient over time. The dashboard launches from an individual patient's record in the EHR. Upon launch of the app, the app uses FHIR application programming interfaces (APIs) to call relevant data, including PHQ-9 scores from the EHR or other data collection systems, medications for depression, and any events of interest (e.g., adverse events, other diagnoses). The relevant data are displayed in the dashboard view within the EHR (Figure 3).

Figure 3. View of Major Depression Outcomes App as populated with sample data



Paula Greenberg
45 YRS, FEMALE

In Recurrence

Most recent PHQ-9 Score (4/12/19): **15** No change

Changes in Score since last PHQ-9

Suicidal thoughts **3** No change

Medication & Referrals

Lexapro50mg Started 1/10/2019

Changes in Depressive Symptoms Over Time

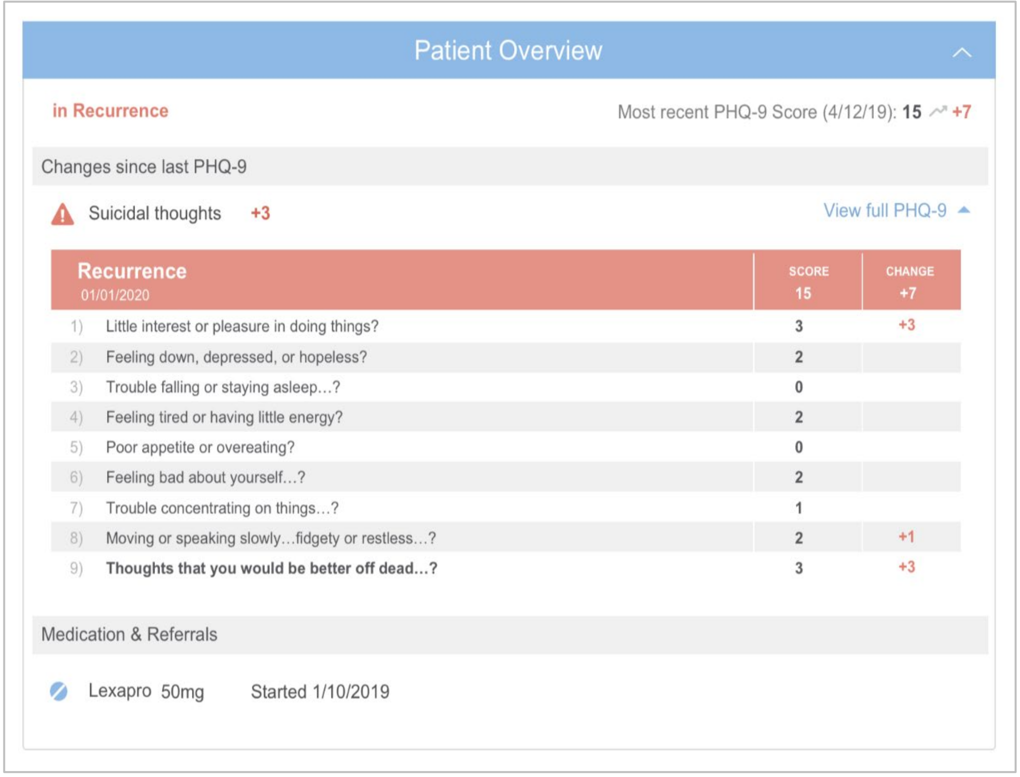
Hovering over a data point on the chart will bring up the full PHQ-9 with scores. Hovering over items on the horizontal timeline will bring up adverse event details. Click the Zoom "+" icon to better see chart details.

03/01
Restraints
Fall
UrinaryCath

The dashboard organizes relevant data into three sections. In the Patient Demographics section, the patient name, age, and sex are displayed on the left. The number of completed PHQ-9s and any indications of suicide ideation on the PHQ-9 are displayed on the right with the Risk Factor label.

Directly below the header bar is the Patient Overview section. This section includes information on whether the patient is responding to treatment, in remission, or has experienced a recurrence, as defined using the harmonized outcome measures (Appendix A). On the right, the most recent PHQ-9 score, the most recent PHQ-9 completion date, and the change from the prior score are displayed. Suicidal thoughts, as measured by item 9 of the PHQ-9, is displayed in this section, along with an option to view the most recent, complete PHQ-9. Clicking on this option displays the item-level scores for the most recent PHQ-9 and changes from the previous PHQ-9 at the item level (Figure 4). Finally, the Medication & Referrals section shows any depression-related medication prescriptions for the patient and referrals to therapy, where available. The medication name, dose, and start date and stop date (if available) are displayed.

Figure 4. View of the most recent PHQ-9 within the app as populated with sample data



Patient Overview

in Recurrence

Most recent PHQ-9 Score (4/12/19): **15** +7

Changes since last PHQ-9

Suicidal thoughts +3

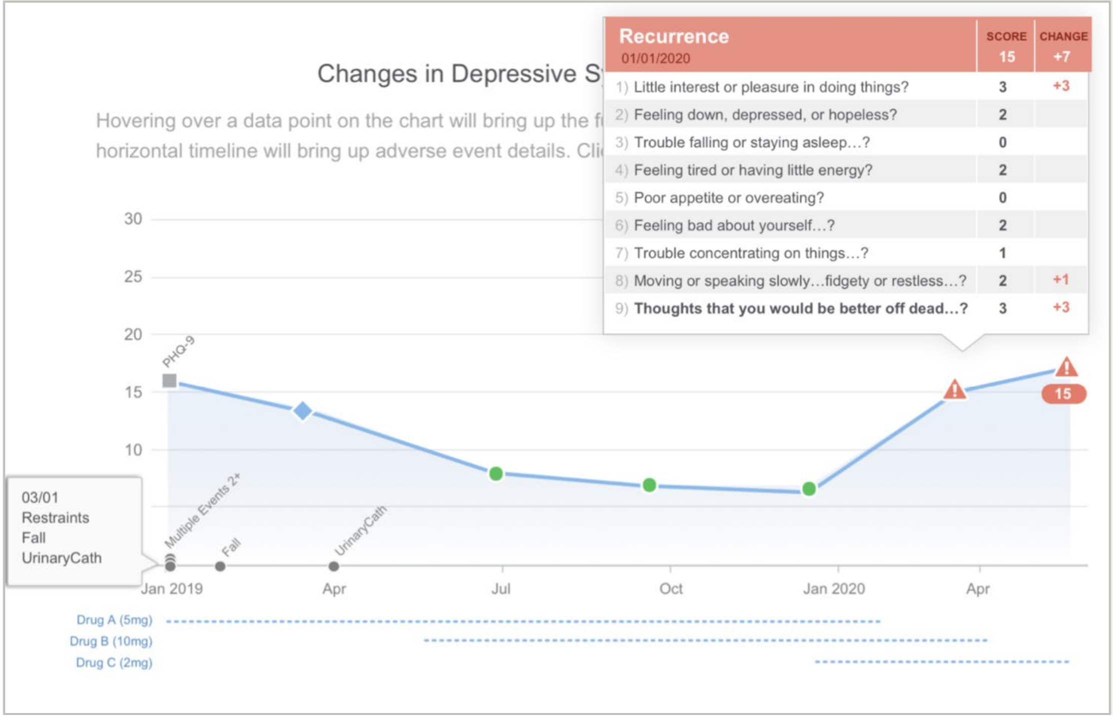
View full PHQ-9

Recurrence 01/01/2020	SCORE 15	CHANGE +7
1) Little interest or pleasure in doing things?	3	+3
2) Feeling down, depressed, or hopeless?	2	
3) Trouble falling or staying asleep…?	0	
4) Feeling tired or having little energy?	2	
5) Poor appetite or overeating?	0	
6) Feeling bad about yourself…?	2	
7) Trouble concentrating on things…?	1	
8) Moving or speaking slowly…fidgety or restless…?	2	+1
9) **Thoughts that you would be better off dead…?**	3	+3

Medication & Referrals

Lexapro 50mg Started 1/10/2019

Figure 5. Previous PHQ-9 displayed on the graph as populated with sample data



Changes in Depressive S

Hovering over a data point on the chart will bring up the f horizontal timeline will bring up adverse event details. Cli

Recurrence 01/01/2020	SCORE 15	CHANGE +7
1) Little interest or pleasure in doing things?	3	+3
2) Feeling down, depressed, or hopeless?	2	
3) Trouble falling or staying asleep…?	0	
4) Feeling tired or having little energy?	2	
5) Poor appetite or overeating?	0	
6) Feeling bad about yourself…?	2	
7) Trouble concentrating on things…?	1	
8) Moving or speaking slowly…fidgety or restless…?	2	+1
9) **Thoughts that you would be better off dead…?**	3	+3

03/01
Restraints
Fall
UrinaryCath

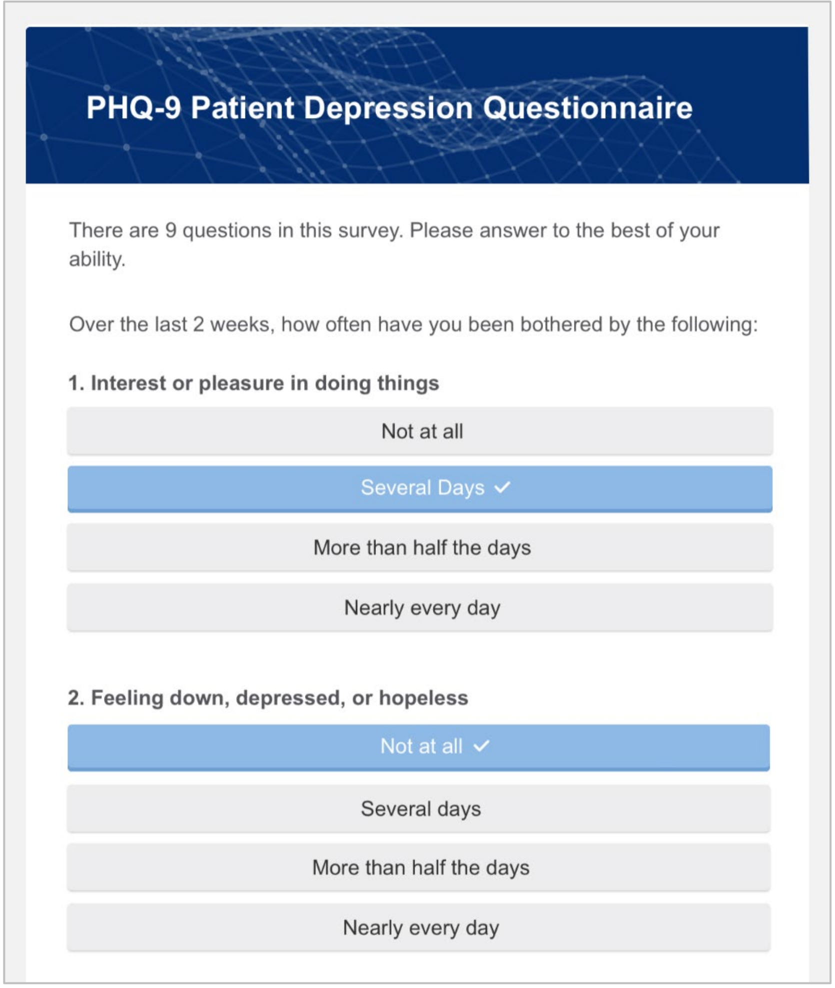
The third section of the dashboard is Changes in Depressive Symptoms over Time. This section presents a graph of PHQ-9 scores from the past year. The app calculates the outcome measures using the standardized definitions and displays the measures on the graph using symbols and colors (e.g., green circle for remission). Medications are shown below the graph in alignment with the period for which the medication was prescribed. The full PHQ-9 can be displayed for any PHQ-9 shown on the graph by hovering over the score icon (Figure 5).

All clinical data accessed by the app are documented and stored elsewhere in the EHR. PHQ-9 scores are stored in the app's FHIR server to improve app performance. No data may be edited in this module.

PHQ-9 Service

The app includes an optional PHQ-9 service that can be used to collect the PHQ-9 directly from patients at scheduled intervals. This functionality is important to support consistent collection of the PHQ-9 in clinical settings where tools to capture the PHQ-9 remotely do not exist. Using this supporting service, clinicians can send a PHQ-9 to a patient for completion on an ad-hoc basis or on a regular schedule.

Figure 6. Patient view in PHQ-9 service



PHQ-9 Patient Depression Questionnaire

There are 9 questions in this survey. Please answer to the best of your ability.

Over the last 2 weeks, how often have you been bothered by the following:

1. Interest or pleasure in doing things

2. Feeling down, depressed, or hopeless

To access the PHQ-9 scheduling tool, clinicians launch the dashboard from the EHR and click the 'Schedule PHQ-9' button (see Figure 3). This launches a scheduling interface through which clinicians can schedule a one-time PHQ-9 (ad-hoc scheduling) or set up a regular schedule (e.g., monthly) for PHQ-9s for the individual patient. Clinicians can also modify existing PHQ-9 schedules by accessing the scheduling feature through the dashboard. At the scheduled date and time, the PHQ-9 service sends a text message to the patient with a link to complete the PHQ-9 using a mobile device (Figure 6).

Patients are able to complete the PHQ-9 and click 'submit' to transmit their results to the clinician. Text message reminders are sent when a patient does not complete the PHQ-9 within 48 hours of the original contact and again within 1 week. Submitted PHQ-9 scores, including item level responses, are stored in a HAPI FHIR server that is accessible to the clinicians' EHR system. Clinicians can then view the PHQ-9 responses using the Outcome Measures Dashboard module of the app, as described above.

Registry Reporting Service

A secondary objective of the Major Depression Outcomes app is to support exchange of data on the harmonized outcome measures with other data collection efforts, such as patient registries. To accomplish this objective, the app includes an optional registry reporting service that sends outcome measures data to a designated recipient. The recipient is designated as part of the implementation of the registry reporting service, and data are sent at an individual patient level. The registry reporting service is integrated with the outcome measures dashboard and can be launched by clicking the 'Transfer Data to Registry' button on the dashboard (Figure 3). Upon launch, the service transforms the following resources from FHIR DSTU2 or STU3 into FHIR R4 and packages them into a FHIR R4 Bundle:

- Patient
- QuestionnaireResponse (PHQ-9 responses stored within the app's FHIR STU 3 server)
- Condition (depression-related diagnoses from the EHR)
- MedicationRequest (medications used for treatment)
- Procedure (reason is depression)
- AllergyIntolerance (substance is a depression medication)

The service sends the Bundle to the registry's API endpoint. The service only sends resources that have been updated since the last submission to the registry, and submissions are tracked using an Observation resource in the app's FHIR server. By using FHIR R4 resources, the registry reporting service provides a standards-based approach to transferring data that could be used by multiple registries or other data collection efforts.

Research Study To Assess SMART on FHIR App

Following app development and testing, a research study was launched to assess the feasibility and value of calculating and displaying the harmonized outcomes measures in the app in the primary care and mental health settings. The objectives of the pilot study were to:

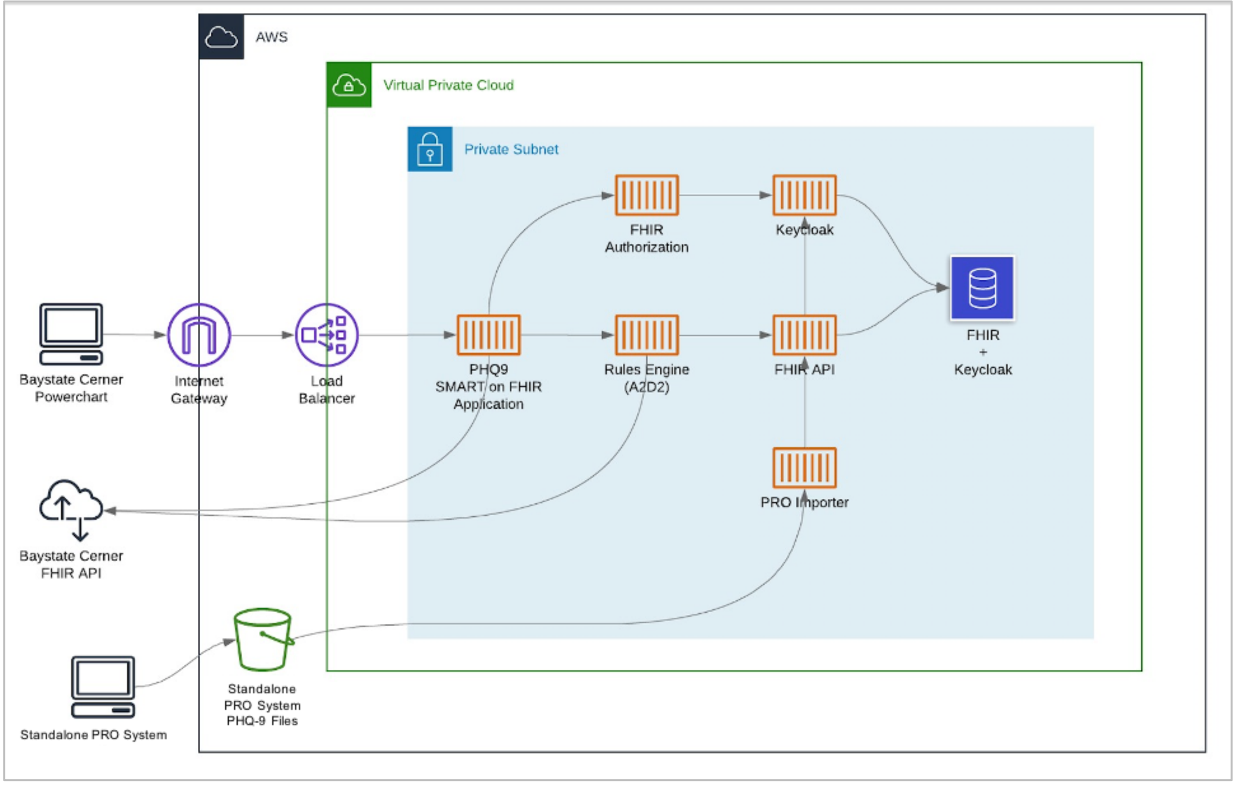
- Demonstrate that it is feasible technically and operationally to use a SMART on FHIR app to extract the relevant data, including PHQ-9 scores, from existing clinical systems, calculate the outcome measures, and display the measure results in the clinician's workflow.
- Assess the burden of calculating the measures and the value of providing the measure results.

Methods

Deployment of App Within a Health System

Baystate deployed the outcome measures dashboard of the Major Depression Outcomes SMART on FHIR app as part of this research study. The dashboard was deployed within an Amazon Web Services (AWS) environment and integrated with the Cerner EHR platform in use at Baystate. Figure 7 depicts the app deployment at Baystate.

Figure 7. App deployment at Baystate

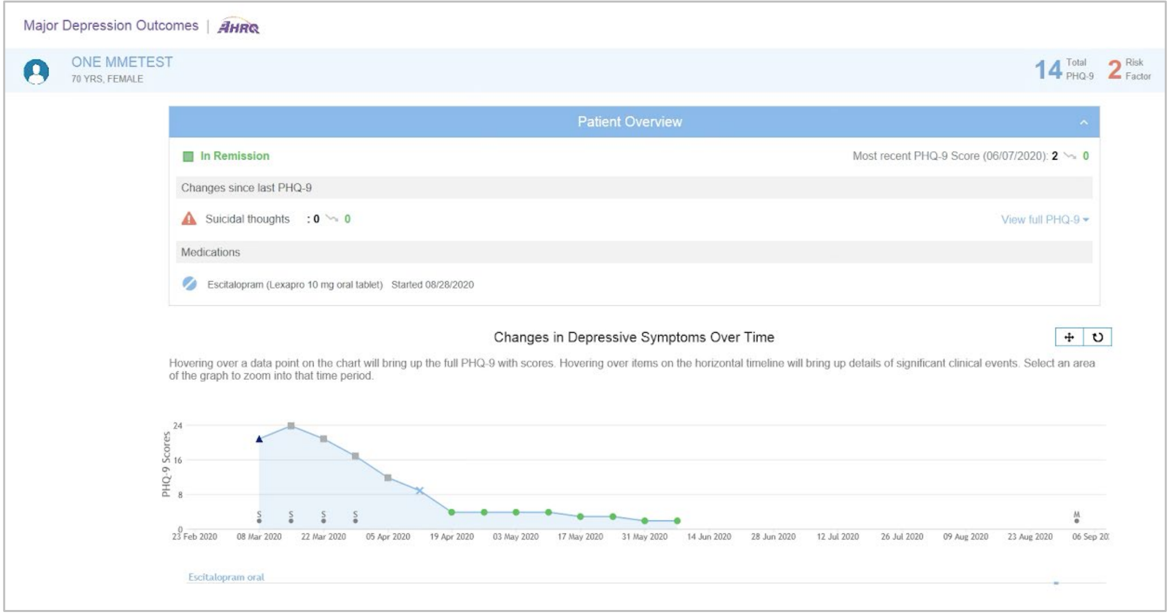


Clinical data relevant to the harmonized outcome measures are accessed from the Cerner EHR system using the FHIR standard and displayed in the app. For the PHQ-9, the app accesses scores stored in Cerner as well as PHQ-9s completed through a standalone PRO system. A custom API was created to support integration of the standalone PRO system. The app as viewed within the Cerner test environment at Baystate is shown in Figure 8.

Assessment of Burden and Value

A longitudinal, observational study was designed to assess the value and burden of calculating the harmonized outcome measures and viewing the results using the Major Depression Outcomes app. Baseline data on patient characteristics, treatments, and symptoms were combined with longitudinal data on outcomes during the study timeframe for display in the app. All data were collected from the institution's EHR and PRO systems. The app launched from within the patient's record in the EHR, aggregated EHR and PRO data, calculated the outcome measure results, and returned the results to the EHR so that they were viewable within the clinician's workflow.

Figure 8. Major Depression Outcomes App, as Viewed in Baystate Health test environment



Major Depression Outcomes | AHRQ

ONE MMETEST
70 YRS, FEMALE

14 Total PHQ-9 2 Risk Factor

Patient Overview

🟩 In Remission

Most recent PHQ-9 Score (06/07/2020): 2 0

Changes since last PHQ-9

Suicidal thoughts : 0 0

View full PHQ-9

Medications

Escitalopram (Lexapro 10 mg oral tablet) Started 08/28/2020

Changes in Depressive Symptoms Over Time

Hovering over a data point on the chart will bring up the full PHQ-9 with scores. Hovering over items on the horizontal timeline will bring up details of significant clinical events. Select an area of the graph to zoom into that time period.

Five clinicians representing different practice sites and settings (primary care, outpatient mental health) participated in this research study. Planned enrollment was 50 patients (10 patients per clinician). Patients were eligible for this study if they were at least 18 years of age, diagnosed with major depressive disorder or dysthymia, and willing and able to provide informed consent. There were no exclusion criteria.

At the conclusion of the data collection period, participating clinicians were asked to complete a brief web-based survey (Appendix C). The survey contained 10 questions assessing three domains: usability of the app, burden of using the app, and overall value of the measures as displayed in the app for informing patient care and improving engagement with patients. No patient data were included in the assessment phase. The Baystate IRB and Western IRB reviewed and approved the study, and the survey was approved under the Paperwork Reduction Act (Office of Management and Budget Control Number 0935-0249).

Findings

Deployment of App Within a Health System

The study demonstrated that it is technically and operationally feasible to use an open-source app to calculate and display the outcome measures in the clinician's workflow. The Major Depression Outcomes SMART on FHIR app was deployed within Baystate's Cerner platform, and clinicians participating in the study were able to launch the app from an individual patient's record and view the outcome measures dashboard. However, several technical and operational challenges were encountered, resulting in delays in the deployment of the app and patient enrollment.

First, Baystate requires completion of an information security review prior to use of new applications, such as web applications, mobile apps, and SMART on FHIR apps, to ensure compliance with the Health Insurance Portability and Accountability Act (HIPAA) and safeguard protected health information (PHI). The review encompasses questions about the application purpose, necessary IT support, data flow, and use cases. Multiple questions were raised during the security review for the SMART on FHIR app, particularly about potential security risks, authentication procedures, and data access, and a series of meetings was necessary to address questions from all relevant parties. As a result, the information security review required several months, delaying the start of app deployment work.

App deployment also required more time than anticipated due to resource issues and technical challenges. The app received information security approval in May 2020, but many necessary Baystate resources for app deployment had been reassigned to COVID-19 related planning and tasks at that point. App deployment would typically require allocation of a software engineer and at least some dedicated time from an infrastructure engineer. Due to COVID-19 related priorities, the Baystate IT group was only able to partially allocate a software engineer and only after the regional surges in COVID-19 cases being managed at the institution allowed for the resource to be assigned. This resulted in further delays, and the OM1 and Elimu project teams provided additional technical support to facilitate the app deployment. The app was deployed successfully in the production environment on September 1, 2020, approximately six months later than planned.

Following app deployment, a challenge related to collection and documentation of PHQ-9 scores was identified. At the outset of this project, Baystate was in the process of implementing a standalone PRO system with the capability to collect the PHQ-9 from patients remotely on a consistent schedule. To ensure that the app fit within the clinical workflow, a custom API was created to fetch the PHQ-9 data from the standalone PRO system and display the scores in the app. After the launch of the app, the project team learned through discussions with the participating clinicians that PHQ-9s for some patients were entered directly into the Cerner EHR, rather than captured through the standalone PRO system. Furthermore, it was possible for patients to have PHQ-9s only in the EHR, only in the standalone PRO system, or in both locations. This variation may be attributed to disruptions in care and delays in the roll-out of the standalone PRO system due to COVID-19. However, this variation also reflects a larger challenge encountered across the tasks in this project – namely, that wide variation exists in how PHQ-9s are captured and documented, even at the level of an individual patient, and a variety of approaches must be used to identify all PHQ-9 scores. To address this challenge in the context of the Baystate implementation, the app was modified to access and display PHQ-9s recorded as observations in the EHR system.

The IRB review and approval of the research study by Baystate's local IRB also required substantially more time than anticipated. While all data elements that are displayed in the app are collected from information routinely recorded in the EHR or the standalone PRO system, clinician use of the app represents a novel approach to viewing information on depression treatments and outcomes. As such, the research study was submitted for IRB review and approval, and patients were asked to provide consent for their data to be presented to clinicians within the context of the app. Several rounds of review and responses and additional clarifications about SMART on FHIR apps in general and this app in particular were required before the Baystate IRB approved the study protocol in September 2020.

Finally, to accommodate the requirement for patient informed consent, the app was made accessible only for clinicians who are participating in the pilot study. These clinicians identified eligible patients and discussed the possibility of participating with them. Interested patients then needed to provide informed consent before clinicians could use the app to view that patient's data. Due to the COVID-19 pandemic, patient visits were virtual, and consent was obtained over the phone by a member of the research team. This created an enrollment challenge, as many patients did not respond to phone calls from the research team.

Assessment of Value and Burden

Clinicians completed a brief survey on usage of the app at the conclusion of the pilot study. Five clinicians completed the survey and indicated that they found it to be straightforward to launch and use the app and that the training was sufficient. Clinicians reported spending, on average, 1-2 minutes using the app (per launch of the app). When asked whether they used the information presented in the app to inform decisions about patient care, three clinicians responded with ‘somewhat,’ one clinician responded with ‘yes’, and one clinician responded with ‘no.’ Clinicians identified response as the most useful measure, followed by remission.

Lessons Learned

Once deployed, the Major Depression Outcomes SMART on FHIR app proved to be highly useful in displaying a visual and summary view of longitudinal patient characteristics, treatments and standardized outcomes in a clear and actionable way. However, deployment and use of a SMART on FHIR app to calculate and display the harmonized outcome measures in clinical practice settings was challenging due to a combination of technical and operational barriers.

First, SMART on FHIR apps are relatively new technologies, and the information security review and IRB review processes were not necessarily designed to account for these types of technologies. The documentation for the application security review, for example, included many questions about authentication that were not relevant to this project, since the app launches through the EHR using the existing security and validation layers. Future efforts to implement the app may consider addressing questions related to information security proactively, perhaps by providing more general information about SMART on FHIR apps and authorization using the FHIR standard to relevant stakeholders.

Further implementations should also plan for adequate time from IT resources with relevant expertise. The SMART on FHIR app was built to be extensible to various infrastructures so it could be implemented in a wide range of care settings. However, this flexibility means that many decisions must be made during the deployment process. This requires a high level of working knowledge of how to implement a SMART on FHIR app within the specific infrastructure. Ideally, an app deployment team would include a software engineer and an infrastructure engineer.

Finally, discussions with clinicians to identify all possible pathways for completion and documentation of the PHQ-9 are important to ensure that the app is configured to display all relevant data.

4. Use of Harmonized Depression Outcome Measures for Research

Introduction

Consistent collection of the harmonized outcome measures in clinical practice settings would yield a robust data infrastructure that could be used to support patient-centered outcomes research in depression. While other tasks in this project focused on calculation of the harmonized measures in primary care and mental health care settings, this task assessed the feasibility of using the harmonized outcome measures data extracted from the patient registries for research purposes. The objectives of this task were to assess the availability of data necessary to conduct research on depression treatment and outcomes, evaluate the registry data quality, and design and conduct a pilot data analysis using data from the PRIME Registry and PsychPRO. The methods, results, and lessons learned are discussed in the following sections.

Assessment of Registry Data Availability

The purpose of this effort was to assess the availability of the data necessary to conduct patient-centered outcomes research using the harmonized outcome measures as calculated in the PRIME Registry and PsychPRO. The work described in Chapter 2 focused on ensuring that the data elements necessary to calculate the harmonized outcome measures were available in the two registries. In addition, data on the characteristics of the patient, the disease course, and treatments are necessary to conduct robust research studies using the registry data. Because the registries rely on data extracted from EHRs and other data collections systems, it is important to understand the types of data that are recorded routinely in the different care settings and any factors related to the process of providing care that may influence the interpretation of the data.

Methods

For this effort, a comprehensive list of variables that would be desirable in a study examining depression treatment and outcomes was created, drawing on the published Outcome Measures Framework for depression[11] as well as other sources.[15, 21-33] Each registry provided information on whether variables were feasible to extract, the difficulty of extraction, and any contextual information that should be considered when interpreting the study findings.

Findings

Variables describing patient characteristics are largely present and readily extracted in both registries, as shown in Table 1 below. Notable areas of difference include race/ethnicity, socioeconomic status, pregnancy, and PHQ-9 scores. Data on race and ethnicity are available for the majority (83%) of patients in the PRIME Registry. However, PsychPRO indicated that there is wide variability in the completion rate and coding of these data in EHRs in the mental health care setting. The lack of standardization and the low emphasis on collecting these data make it infeasible to extract race and ethnicity from the registry for research purposes. PRIME also includes data on social determinants of health for the majority of its patients, but these data are not currently captured in structured forms in PsychPRO.

Table 1. Demographics/patient characteristics

1 = Readily available; 2 = Available with moderate effort; 3 = Difficult to extract

Variable	Definition	PRIME: Feasible To Extract?	PRIME: Difficulty Level	PsychPRO: Feasible To Extract?	PsychPRO: Difficulty Level
Sex	Patient's sex at birth	Yes	1	Yes	1
Age	Age in years at the index date	Yes	1	Yes	1
Race	Patient's race	Yes	2	No	N/A
Ethnicity	Patient's ethnicity	Yes	2	No	N/A
Family history of depression or other mental health disorder	Whether patient has a family history of depression or other mental health disorder (ICD-10 code Z81.8)	Yes	1	Yes	1
Socioeconomic status	Socioeconomic status	Yes	2	No	N/A
Pregnant	Pregnant at the time of the index date	Yes	2	Yes	1
Postpartum status	Childbirth within the 12 months prior to the index date	No	N/A	No	N/A
Insurance type	Insurance type at baseline	Yes	3	Yes	3
Regional division	Regional division at baseline	Yes	1	Yes	2
PHQ-9 score category at baseline	Patient's PHQ-9 total score on the index date	Yes	1	Yes	3
PHQ-9 total score at baseline	Patient's PHQ-9 total score on the index date	Yes	1	Yes	3

Pregnancy and postpartum status are relevant when considering postpartum depression. In the PRIME Registry, information on pregnancy may not be available in structured form for patients who are receiving prenatal care at a different practice, and information on postpartum status is unlikely to be available. In PsychPRO, availability of these variables varies depending on how frequently clinicians document these codes. Coding for pregnancy is likely to occur in several scenarios, as described below:

- Medicaid programs have presumptive eligibility for pregnant women, so it may be helpful in establishing eligibility to code for the presence of pregnancy.
- In Medicare Advantage programs, the capitation rate for an individual patient is calculated using the conditions listed historically on claims. Including a code for pregnancy would result in a higher capitation rate for that individual patient.
- Additional counseling and medical education may be necessary due to the impact of the pregnancy on pre-existing mental illnesses, substance use disorders, and/or medication use.

Variation also existed at the level of difficulty of extracting the PHQ-9 scores. In the PRIME Registry, patients complete the PHQ-9 through the registry patient portal, and scores are calculated and saved directly in the registry, making this process relatively straightforward. In contrast, PsychPRO captures PHQ-9s entered by patients through the registry patient portal as well as entered in the EHR. Extraction of PHQ-9 scores from EHRs is complex, as variation exists across practices regarding documentation efforts as well as EHR system templates for entering these data. In practice, most EHR systems do not have a standardized field to capture these data. Many clinicians make notations in clinical notes, meaning the scores must be extracted based on key words or phrases defined in consultation with the clinicians. Other clinicians/practices have adapted existing EHR fields for documenting patient 'test' results.

These fields are mainly used to document medical test results (e.g., HbA1c, blood pressure). These variations create complexity and the need for careful mapping and review of the data when extracting PHQ-9 scores for research purposes.

While the registries indicated that family history of depression is readily available, it is important to note that data availability will vary based on how frequently clinicians use these codes, which are typically not relevant for billing purposes. Both registries also indicated that insurance type (Medicaid, Medicare, commercial, other) is difficult to extract. These data are derived from a non-standardized, unstructured data element. The information varies across EHR systems and may represent the name of the plan, insurance company name, plan type, or other information for system-specific billing purposes. While some entries may be grouped using a coding algorithm, other entries may not contain sufficient information to group them into the specified categories.

Both registries reported that variables on comorbidities and medical history can be extracted and grouped into relevant categories, as shown in Table 2 below.

Table 2. Disease characteristics at baseline

1 = Readily available; 2 = Available with moderate effort; 3 = Difficult to extract

Variable	Definition	PRIME: Feasible To Extract?	PRIME: Difficulty Level	PsychPRO: Feasible To Extract?	PsychPRO: Difficulty Level
Comorbidities/ Medical history	Presence or history of: - Anxiety disorders (any) - Bipolar disorder - Schizophrenia - Sleep disorder including insomnia - Alcohol/drug dependence - Coronary artery disease - Malignancy (any)	Yes	2	Yes	3

PsychPRO noted that medical comorbidities (e.g., coronary artery disease, malignancy) may be unavailable in many cases, as psychiatrists and behavioral health clinicians often do not have access to valid and reliable information about a patient's medical diagnoses.

Table 3 presents variables related to disease course. The registries reported similar findings in terms of data availability and difficulty of extraction, with two exceptions. First, PsychPRO noted that the challenges discussed above related to the PHQ-9 also apply to newly diagnosed depression. Second, PsychPRO reported a higher level of extraction difficulty for depressive severity at baseline. Depressive severity at diagnosis (baseline) is defined here using the PHQ-9 scores. However, a diagnosis of major depressive disorder does not require the use of the PHQ-9, and psychiatrists and other behavioral health clinicians often make this diagnosis by clinical judgement or using an assessment instrument other than the PHQ-9.

Both registries reported that duration of symptoms and prior history of depression/previous relapses would be difficult to capture using structured data alone. Patients who received care at the practice for a long period may have prior diagnosis codes, but, for many patients, data on these variables would be found in the clinical notes. Regarding prior treatments, both registries noted that these data are limited to treatments provided by the patients' current practice/clinicians. Treatments received outside of the current practice/clinicians would not be captured in the registries; for the PRIME Registry, this also affects the availability of data on therapy received from clinicians outside the practice.

Lastly, as discussed in Chapter 2, suicide ideation and behavior may only be documented in clinical notes, without associated structured codes. Both registries noted that these data are readily extracted using standard codes, but the quality of these data may not be robust.

Table 3. Disease course

1 = Readily available; 2 = Available with moderate effort; 3 = Difficult to extract

Variable	Definition	PRIME: Feasible to Extract?	PRIME: Difficulty Level	PsychPRO: Feasible to Extract?	PsychPRO: Difficulty Level
Newly diagnosed depression	PHQ-9 >9 or ICD-9/10 code for depression and no previous record of an ICD-9/10 code for depression or of a PHQ-9 >9	Yes	1	Yes	3
Depressive severity at diagnosis	Patient's PHQ-9 total score on the diagnosis date: - None or minimal depression (0-4) - Mild depression (5-9) - Moderate depression (10-14) - Moderately severe depression (15-19) - Severe depression (20+)	Yes	1	Yes	3
Duration of symptoms	Duration of depressive symptoms	No	N/A	No	N/A
Previous relapses	Prior history of depression	No	N/A	No	N/A
Prior treatments	Whether patient has received any treatment (such as antidepressant treatment, psychotherapy, ECT, TMS, VNS) for depression on or 12 months prior to the index date	Yes	2	Yes	2
Type of prior treatments	Yes/No for each of the treatments observed on or after the index date: - Psychotherapy - Pharmacotherapy - TMS - VNS	Yes	2	Yes	2
Lab tests: Thyroid function	Patient had a thyroid function test on or in the 12 months before or on the index date	Yes	1	Yes	1
Prior history of suicidality	Selection of 'several days', 'more than half the days' or 'nearly every day' option on previously completed PHQ-9 item 9 OR based on ICD-9/10/SNOMED codes	Yes	1	Yes	1

The availability of data on treatments of interest is presented in Table 4. Both registries reported that information on type of treatment is available with moderate effort, noting again that treatment information is limited to treatments provided through the current practice/clinicians. Information on alternative therapies is not available through structured data, although some relevant information may be found in clinical notes.

Table 4. Treatments

1 = Readily available; 2 = Available with moderate effort; 3 = Difficult to extract

Variable	Definition	PRIME: Feasible To Extract?	PRIME: Difficulty Level	PsychPRO: Feasible To Extract?	PsychPRO: Difficulty Level
Type of treatment	Yes/No for each of the treatments observed on or after the index date: - Psychotherapy - Pharmacotherapy - TMS - VNS	Yes	2	Yes	2
Type of pharmacotherapy	Yes/No for each of the pharmacotherapies observed on or after the index date: - SSRIs - SNRIs - Atypical antidepressants - Tricyclic antidepressants - MAOIs - SARIs - NDRIs - Esketamine	Yes	2	Yes	2
Alternative therapies	Any alternative therapies	No	N/A	No	N/A
Referral(s) for treatment	Any referrals for mental health services	Yes	1	Yes	1

Variables related to collection and documentation of the PHQ-9 post-baseline are presented in Table 5. These variables are readily available in the PRIME Registry. The challenges associated with extracting these data from EHR systems for PsychPRO are discussed above.

Table 5. PHQ-9 Post-baseline
1 = Readily available; 2 = Available with moderate effort; 3 = Difficult to extract

Variable	Definition	PRIME: Feasible To Extract?	PRIME: Difficulty Level	PsychPRO: Feasible To Extract?	PsychPRO: Difficulty Level
PHQ-9 at 6 months post-index date	Patient's PHQ-9 total score 6 months (+/- 60 days) post-index date. If multiple assessments were administered, the last PHQ-9 total score occurring during the 6-month window will be selected	Yes	1	Yes	3
PHQ-9 at 12 months post-index date	Patient's PHQ-9 total score 12 months (+/- 60 days) post-index date. If multiple assessments were administered, the last PHQ-9 total score occurring during the 12-month window will be selected	Yes	1	Yes	3
Number of PHQ-9s	The number of completed PHQ-9 assessments post-index date	Yes	1	Yes	3

The harmonized depression outcome measures are presented in Table 6. As with the PHQ-9 scores post-baseline, the registries report different levels of extraction difficulty for the outcome measures driven by the different approaches to capturing PHQ-9 scores. Both registries reported that death from suicide is unlikely to be captured in the practices' EHR systems. While a date of death may be available, cause of death is unlikely to be captured in structured form.

Table 6. Outcome measures
1 = Readily available; 2 = Available with moderate effort; 3 = Difficult to extract

Variable	Definition	PRIME: Feasible To Extract?	PRIME: Difficulty Level	PsychPRO: Feasible To Extract?	PsychPRO: Difficulty Level
Death from suicide	Whether a patient died from suicide	No	N/A	Yes	3
Improvement in depressive symptoms -- remission at 6 months	Patient's PHQ-9 total score 12 months (+/- 60 days) post-index date. If multiple assessments were administered, the last PHQ-9 total score occurring during the 12-month window will be selected	Yes	1	Yes	3
Improvement in depressive symptoms -- remission at 12 months	Whether a patient with the baseline PHQ-9 score > 9 demonstrates remission, defined as a PHQ-9 score at 6 months of < 5	Yes	1	Yes	3
Improvement in depressive symptoms -- response at 6 months	Whether a patient with the baseline PHQ-9 score > 9 demonstrates a response to treatment, defined as a PHQ-9 score at 6 months that is reduced by 50% or greater from the baseline PHQ-9 score	Yes	1	Yes	3
Improvement in depressive symptoms -- response at 12 months	Whether a patient with the baseline PHQ-9 score > 9 demonstrates a response to treatment, defined as a PHQ-9 score at 12 months that is reduced by 50% or greater from the baseline PHQ-9 score	Yes	1	Yes	3
Worsening in depressive symptoms – recurrence at 6 months	Whether a patient with the baseline PHQ-9 score > 9 who demonstrates remission (defined as a PHQ-9 score < 5 at any point post-baseline and before 6 months (+/- 60 days) post-index date) of at least two months' duration and subsequently experiences a recurrence of a depressive episode (defined as a PHQ-9 score > 9 OR hospitalization for depression or suicidality)	Yes	1	Yes	3

Variable	Definition	PRIME: Feasible To Extract?	PRIME: Difficulty Level	PsychPRO: Feasible To Extract?	PsychPRO: Difficulty Level
Worsening in depressive symptoms – recurrence at 12 months	Whether a patient with the baseline PHQ-9 score > 9 demonstrates remission, defined as a PHQ-9 score at 12 months of < 5 at any point post-baseline and before 12 months (+/- 60 days) post-index date) of at least two months' duration and subsequently experiences a recurrence of a depressive episode (defined as a PHQ-9 score > 9 OR hospitalization for depression or suicidality)	Yes	1	Yes	3
Worsening in depressive symptoms – change in category at 6 months	Whether or not a patient's PHQ-9 score at twelve months falls into a worse category of depression than their baseline PHQ-9. (e.g. if a patient has a score falling into the mild category at baseline and the moderate category on the 12-month PHQ-9, they would be considered in the worsening category)	Yes	1	Yes	3
Worsening in depressive symptoms – change in category at 12 months	Whether or not a patient's PHQ-9 score at twelve months falls into a worse category of depression than their baseline PHQ-9. (e.g. if a patient has a score falling into the mild category at baseline and the moderate category on the 12-month PHQ-9, they would be considered in the worsening category)	Yes	1	Yes	3
Suicidality post-baseline	Selection of 'several days', 'more than half the days' or 'nearly every day' option on PHQ-9 item 9 OR based on ICD-9/10/SNOMED codes	Yes	1	Yes	3

The final set of variables, presented in Table 7, describes healthcare resource utilization. For these variables, both registries reported that the data are limited to visits captured at the practice. Patients may receive care in multiple settings, and linkage to claims data would be necessary to provide a more complete view of overall healthcare resource utilization, including emergency room (ER) visits and hospitalizations. An area of emerging interest due to the COVID-19 pandemic is telehealth. The PRIME Registry reported that it is feasible to distinguish between telehealth and in-person office visits using the registry data. In PsychPRO, it is difficult currently to identify telehealth visits for mental health since there are no separate codes for telehealth. Instead, clinicians use the same E/M CPT code numbers for those services that are approved for telehealth. Clinicians may indicate that the work was done remotely by a modifier or a place of service code, but this is not required by all payers. Quality of these data is likely to be problematic due to variation in coding across clinicians/practices and recent changes to coding rules.

Table 7. Healthcare resource utilization
1 = Readily available; 2 = Available with moderate effort; 3 = Difficult to extract

Variable	Definition	PRIME: Feasible to Extract?	PRIME: Difficulty Level	PsychPRO: Feasible to Extract?	PsychPRO: Difficulty Level
Any outpatient visits in 12 months post-index (mental health)	Whether patient had any outpatient (mental health) visit in the 12-month period post-index date	Yes	1	Yes	2
Outpatient visits (mental health)	Defined as the number of outpatient (mental health) visits	Yes	1	Yes	2
Any outpatient visits in 12 months post-index (all other cause)	Whether patient had any outpatient (all other cause) visit in the 12-month period post-index date	Yes	1	No	N/A
Outpatient visits (all other cause)	Defined as the number of outpatient (all other cause) visits	Yes	1	No	N/A
Any telehealth visits in 12 months post-index (mental health)	Whether patient had any telehealth visit in the 12-month period post-index date	Yes	1	No	N/A
Telehealth visits (mental health)	Defined as the number of telehealth visits	Yes	1	No	N/A
Any telehealth visits in 12 months post-index (all other cause)	Whether patient had any telehealth (all other cause) visit in the 12-month period post-index date	Yes	1	No	N/A
Telehealth visits (all other cause)	Defined as the number of telehealth (all other cause) visits	Yes	1	No	N/A
Any ER/Hospitalization visits in 12 months post-index (mental health)	Whether patient had any ER/Hospitalization (mental health) visit in the 12-month period post-index date	No	N/A	No	N/A
ER/Hospitalizations (mental health)	Defined as the number of ER visits or hospitalizations (mental health)	No	N/A	No	N/A
Date of ER/Hospitalization (mental health)	Date of patient's ER visit or hospitalization (mental health)	No	N/A	No	N/A
Any ER/Hospitalization visits in 12 months post-index (all other cause)	Whether patient had any ER/Hospitalization visit (all other cause) in the 12-month period post-index date	No	N/A	No	N/A
ER/Hospitalizations (all other cause)	Defined as the number of ER visits or hospitalizations (all other cause)	No	N/A	No	N/A

In summary, the assessment of data availability indicates that, despite some gaps, it is feasible to conduct patient-centered outcomes research using the harmonized outcome measure definitions and data from the participating registries. Some descriptive variables that are important for adequately describing patients with depression and that help to provide context around their course of disease are unavailable in the structured data captured in one or both registries. These variables include: race, ethnicity, pregnancy/postpartum status over time, depression diagnosis date/duration of depression, severity of depression at diagnosis, prescription of alternative therapies, and hospitalization/ER visit information. In addition, deaths are not well-recorded in ambulatory EHR systems, and suicide as a cause of death is known to not be reliably recorded. Thus, although deaths due to suicide is one of the harmonized outcome measures, it is not feasible to calculate that measure for research purposes due to the systematic difficulties in capturing such information. The registries reflect routine clinical practice and documentation in the family medicine and mental health care settings, and these gaps reflect the nature of the data sources (EHRs, PRO measures) used to populate the registries.

Evaluation of Registry Data Quality

The purpose of this effort was to assess the quality of the data that are captured and transferred to the participating patient registries. Understanding data quality is critical to support identification of suitable use cases for the data (research, quality measurement, population health, etc.) and appropriate interpretation of analyses using the data.

Methods

The data quality evaluation conducted for this project focused on the PHQ-9 data collected through the registry patient portal and the data extracted from EHRs to populate the registry, with a particular emphasis on the types of data that are necessary to calculate the harmonized outcome measures for depression. The PRIME Registry and PsychPRO conduct data quality evaluations on a routine basis, and these evaluations were expanded to address the goals of this data quality evaluation. Findings were reported to the project team and are summarized below.

Findings

PHQ-9 Data

The PRIME Registry uses a patient portal to collect the PHQ-9. Several measures are used to ensure that the PHQ-9 data captured through the patient portal are accurate and that data quality is maintained. First, patients who are asked to complete the PHQ-9 as part of this project via the registry patient portal are assigned a unique login. This ensures that the PHQ-9 score is linked to the correct patient within the registry. Second, the responses to the PHQ-9 questionnaire are captured using radio buttons, such that each radio button corresponds to a valid PHQ-9 response. The use of radio buttons ensures that patients can select only one response for each question in the PHQ-9. There are four standard responses for each of the PHQ-9 questions: Not at all, Several days, More than half the days, and Nearly every day. Each standard response has a numeric value: 0, 1, 2, and 3 respectively. Each response to a PHQ-9 question is saved in an incomplete state until the patient submits the PHQ-9, at which point the data become available in the registry. Responses to the PHQ-9 are stored as JSON files in the registry database.

The PHQ-9 score is not calculated until the patient submits their responses, and there is no minimum number of required responses for submission or score calculation. The total PHQ-9 score is calculated by summing the numeric values of each of the selected standard responses for all nine questions, where the range of possible scores is 0-27. To limit duplicate scores, patients are limited to one active PHQ-9 request at a time, and each request is associated with a specific encounter. Patients who receive a second PHQ-9 request before submission of the prior PHQ-9 will no longer be able to submit responses for the prior PHQ-9. Additionally, PHQ-9 requests expire after 6 months if they are not completed and submitted by the patient.

PsychPRO also uses a patient portal through which patients can complete PROs, including the PHQ-9. Each patient is assigned a unique login to the patient portal. Once logged in to the patient portal, patients can see any assigned PHQ-9s that are available for them to complete. A clinician will assign a PHQ-9 that is associated with an upcoming encounter to a patient. The patient is able to complete that PHQ-9 up to the time of the appointment and through the appointment, at which point the PHQ-9 expires. If the encounter is cancelled prior to the patient taking the PHQ-9, that associated PHQ-9 will also be cancelled.

The responses the patient can select from for each of the questions in the PHQ-9 are limited using radio buttons, such that the patient can only select one response for each of the questions and the response is a valid response to the PHQ-9. The responses to each of the PHQ-9 questions are stored as a numeric value in the registry database, as discussed above. Patients have until the conclusion of the encounter associated with the PHQ-9 to edit their responses, after which their responses become locked. If the patient has provided at least one response to the PHQ-9, the score will be saved, so patients do not have to explicitly click a submit button in order to save their responses. After seven of the nine questions have saved responses, a score will be calculated.

EHR Data

Both the PRIME Registry and PsychPRO work with a registry technical vendor that is responsible for extracting data from participating sites' EHR systems and populating the registry. The registries work with the same technical vendor, and the data quality steps described below are the same for each registry.

Both registries have instantiated data quality processes to ensure that any EHR data collected for a patient via the registry technical vendor is accurate and complete. The vendor works directly with each practice to map data from EHR fields to the registry data elements as part of the onboarding process. As a first step, the practice identifies all quality and other registry measures that they would like to use. Once the mapping for those measures is completed, the vendor extracts data from the EHR, populates the registry, and calculates the measures. Practices are then asked to review the measure counts. Any areas of concern are reviewed at the individual patient level. This is an iterative process that continues until all issues are resolved. Registry data are refreshed on a monthly basis, and practices are also asked to review their measure counts regularly to ensure accuracy. As data are extracted for the registries, several data quality checks are completed to verify that there are no data anomalies in the EHR data. Data quality checks include syntax validation, mapping validation, lookup check, file count validation, and extract status validation in the different sources. Additionally, the registry technical vendor has numeric validation for PHQ-9 scores that verifies that the score is within the specified range.

In addition to the data quality checks conducted at the individual practice level, CMS performs an annual audit of select data from the registries that requires a CMS Data Validation Execution Report, and in more detailed cases, corrective action documentation.

As a final step, data from the practices are curated and standardized for inclusion in the finalized registry dataset. Patient unique identifiers are generated, the data are checked for completeness and correctness, the dataset is delimited, PHI is removed, variables of interest are derived, and the data are processed and displayed on the registry dashboard.

In summary, both registries have implemented data quality checks and have processes in place for regular assessments of data quality. Ongoing evaluation of data quality and prompt resolution of any identified issues are critical given that both registries are approved by CMS under the Qualified Clinical Data Registry (QCDR) program. The existing processes and data quality checks are sufficient to ensure the accuracy of data mapped from the EHR to the patient registries and to identify anomalies in the data that may be indicative of a documentation issue at the practice level. However, some data necessary for calculating the harmonized depression outcome measures are not recorded routinely using structured data (e.g., adverse effects of treatment, suicide ideation and behavior) or may be missing from the EHR altogether due to the clinician's inability to access the information (e.g., ER visits, cause of death) or the nature of documentation in the care setting (e.g., some social determinants of health). These issues are discussed in detail in the preceding section on data availability and should be considered when assessing the quality of the registry data in the context of a specific use case.

Pilot Data Analysis

The purpose of the pilot data analysis was to demonstrate the feasibility of using the harmonized outcome measures, as captured in the two participating patient registries, for conducting patient-centered outcomes research in depression. As discussed in Chapter 1, patient registries collect large observational data sets that could be used to address important questions about depression diagnosis, treatment, and outcomes in real-world care settings. A key goal of implementing the harmonized outcome measures in patient registries is to create research data infrastructure and facilitate the linkages, aggregations, and comparisons of data across registries for research purposes. This analysis was designed to assess the feasibility of conducting an analysis using the harmonized outcome measures in the PRIME Registry and PsychPRO.

The objectives of the analysis were:

- To describe patients with major depressive disorder receiving care in the family medicine and mental health care setting in terms of:
  - Demographics
  - Severity of symptoms (as measured by the PHQ-9)
  - Number of fully and partially completed PHQ-9s, and
  - Type of treatment
- To provide summary statistics for patients with major depressive disorder receiving care in the family medicine and mental health care setting, such as:
  - Remission
  - Response
  - Recurrence
  - Healthcare resource utilization

Methods

The study was a longitudinal, multi-center observational feasibility study that included data on eligible patients with a diagnosis of major depressive disorder. Retrospective data on previous disease status and patient characteristics were collected and combined with longitudinal data from these data sources on outcomes during the study timeframe. All data were collected from practice-level EHRs, patient portals, and other existing data sources, as needed.

The study collected data from the PRIME Registry and PsychPRO. A total of 21 sites participating in the registries (11 from PRIME and 10 from PsychPRO) were recruited to participate in this study. Planned enrollment was 200 patients. Patients were eligible for this study if they were at least 18 years of age and diagnosed with major depressive disorder or dysthymia. There were no exclusion criteria. All eligible patients at participating sites were included. The study was reviewed by IRBs for each registry and determined to not be human subjects research.

Sites participating in the registries used an existing process to extract data from their EHRs and send data to the registry on a regular basis. In addition, the registry data dictionaries were compared with the outcome measure definitions (Appendix A), and some additional data elements were extracted from participating site EHRs (if routinely documented) to support calculation of the outcome measures. These additional data elements include death, cause of death, suicide ideation and behavior, and adverse events related to depression treatment. Participating sites also captured the PHQ-9 at regular intervals during office/telehealth visits or through the registry patient portals.

The duration of ongoing prospective data collection for the purposes of this study was approximately 12 months.

Findings

The analysis demonstrated that calculation of the harmonized outcome measures in two registries representing different care settings is feasible. Results of the analysis are presented in a manuscript that will be submitted for publication separately.

Lessons Learned

The PRIME Registry and PsychPRO contain a wealth of information that can be used to study depression treatment and outcomes, but there are also gaps in the information captured in the registries. Further work is needed to expand data infrastructure to capture the contextual information necessary to characterize patients with depression and to improve linkages across data sources, so a more complete picture of an individual's treatment and outcomes is available. These needs are discussed in the prioritized research agenda (Chapter 5). In addition, more robust tools are needed to extract data from unstructured clinical notes so the full range of outcomes can be captured. Finally, it is important to note that variations in clinical practice and documentation at the practice and individual level and across care settings may lead to variation in the quality and completeness of the registry data, which may in turn affect the study findings and interpretations. The data should be viewed considering these limitations.

5. Resources To Support Future Use of Harmonized Depression Outcome Measures

Introduction

This project demonstrated that it is feasible to calculate the harmonized depression outcome measures across care settings, display the results to clinicians to support individual patient management and population health, and use the outcome measures data to support patient-centered outcomes research. The value of the harmonized measures would be increased by implementations of the measures in additional care settings and additional patient registries and other observational studies. The purpose of this task was to create a set of resources to facilitate future use of the harmonized depression outcome measures and harmonized outcome measures developed for other condition areas.[9, 34]

The objectives of this task were to (1) identify high-priority research questions that could be addressed using the harmonized depression outcomes measures; (2) provide a toolkit to assist registries interested in sharing data with external researchers to address new research questions; (3) assess the feasibility of using NLP tools to improve the utility of routinely recorded clinical data for calculating harmonized measures for research and other purposes; and (4) develop tools to facilitate the implementation of harmonized outcome measures in EHRs. These efforts and the resulting tools are described further below.

Prioritized Research Agenda

The goal of this task was to develop and prioritize a list of research topics related to depression care that can be addressed using the harmonized depression outcome measures. Barriers to using the harmonized outcome measures to address the prioritized research topics were also identified. The methods and findings are summarized below.

The project team conducted a horizon scan of the literature focusing on systematic reviews, research agendas, treatment guidelines, and prioritization research published within the past five years. In addition, the team conducted searches of relevant websites (AHRQ, American Psychiatric Association, ABFM, American College of Physicians, Patient-Centered Outcomes Research Institute [PCORI], and the United States Preventive Services Task Force) to identify documents that may not have been indexed on PubMed.

Eight research themes were identified through the review of the research gaps or future research sections of the documents identified in the search: treatment effectiveness; variation across care settings; screening, diagnosis, and prevention; treatment-resistant depression; impact of race, ethnicity, culture, and other factors on outcomes; depression and comorbidities; perinatal and postpartum depression; and suicidality.[26-38]

Specific patient-centered topics and questions of interest that could be addressed using the harmonized outcome measures in depression were then developed with reference to the previously reviewed research priorities and evidence gaps. Participants in the Stakeholder Panel were also asked to contribute priority research topics during the quarterly meeting, and those suggestions were incorporated into the list of potential questions. The Stakeholder Panel discussed the questions and provided feedback through a series of virtual meetings in 2020. Key points that emerged from the Panel discussion were as follows:

- Many of the questions relate to screening and diagnosis or effectiveness of treatment. For example, the questions in the Depression and Comorbidities theme largely relate to the effectiveness of treatment in different subpopulations. Reorganization of the questions into four overarching themes would reduce redundancy and improve the clarity of the research agenda.

- The role of measurement-based care should be incorporated into this research framework. When examining treatment effectiveness, it is critical to understand the framework in which care was provided – in other words, were the principles of measurement-based care followed?
- It is possible to answer some of these questions immediately using existing data infrastructure and methods. Other questions, however, require the development of new research methods or the expansion of data infrastructure. Prioritization therefore must reflect both the importance of the question and the interim steps necessary to address the question.

Based on the Stakeholder Panel's feedback, the questions were reorganized into four overarching themes: screening, diagnosis, and prevention; social determinants of health; treatment effectiveness; and variation across care settings. Many questions were revised to be more specific and to reflect the feasibility of addressing the question as well as sub-questions of interest. Specific barriers to addressing the questions were also identified. These barriers broadly fell into two categories: data infrastructure gaps and lack of methods frameworks. For some questions, the necessary data are not currently captured systematically, making it challenging to address the questions using observational data sources. For example, information on social determinants of health is often not documented or is only included in unstructured fields in EHRs, making it difficult to extract these data and examine their impact on depression screening, diagnosis, treatment, and outcomes. Further work is needed to expand the existing data infrastructure – meaning data captured through EHRs, patient registries, and other data collection systems – to capture these data in consistent, usable formats so they can be easily extracted for research purposes and used in the clinical workflow.

For other questions, new methods are needed. For example, there is interest in using patient-reported tools to capture side effects related to treatment, but existing tools have not been adopted in clinical practice. As a result, there are questions about how often to capture this information and how to use the information to inform clinical decision-making. Further work is needed to create a framework for the systematic collection and use of patient-reported side effect information in the clinical workflow and for research purposes.

While some patient-centered outcomes research (PCOR) questions can be addressed using the current data infrastructure and harmonized outcome measures, other important PCOR questions must be addressed after data infrastructure has been created and/or new methods have been developed. To reflect this, the research agenda organizes questions into short-term PCOR questions, long-term PCOR questions, data infrastructure questions, and methods development questions. A roadmap was also created to show the process of addressing the PCOR priorities. The roadmap reflects an iterative process of addressing barriers to answering specific questions, refining questions, and ultimately addressing the full range of complex questions that need to be answered in order to improve depression outcomes.

More information, including the research agenda and roadmap, can be found in the separate report, "A Prioritized Research Agenda for Using the Harmonized Outcome Measures to Support Patient-Centered Outcomes Research in Depression" (see complete list of project artifacts in Appendix B).

Data Use & Governance Toolkit

The purpose of this task was to develop a toolkit describing current best practices and providing practical information to assist registries interested in sharing data with external researchers.

The contents of the toolkit were developed based on review of the literature, existing registry practices, interviews with registries, and input from key stakeholders involved in the sharing of registry data. While some information in this toolkit may be relevant in other countries, this toolkit focuses on best practices for sharing data within the United States. Considerations related to data sharing differ across registries depending on the type of registry, registry purpose, funding source(s), and other factors; as such, this toolkit describes general best practices and considerations rather than providing specific recommendations.

The toolkit is organized into three sections: "Preparing to Share Data," "Governance," and "Procedures for Reviewing and Responding to Data Requests." The section on "Preparing to Share Data" discusses the role of appropriate legal rights to further share the data and the need to follow all applicable ethical regulations. Registries should also prepare for data sharing activities by ensuring data are maintained appropriately and developing policies and procedures for governance and data sharing.

The "Governance" section describes the role of governance in data sharing and outlines key governance tasks, including defining and staffing relevant oversight bodies; developing a data request process; reviewing data requests; and overseeing access to data by the requesting party. Governance structures vary based on the scope of data shared and registry resources.

Lastly, the section on "Procedures for Reviewing and Responding to Data Requests" discusses the operational steps involved in sharing data. Policies and procedures for sharing data may depend on what types of data are available for sharing and with whom the data can be shared. Many registries develop a data request form for external researchers interested in using registry data. When reviewing requests, registries may consider whether the request aligns with the registry's mission/purpose, the feasibility and merit of the proposed research, the qualifications of the requestor, and the necessary ethical and regulatory approvals, as well as administrative factors such as costs and timelines. Registries may require researchers to sign a data use agreement or other such contract to clearly define the terms and conditions of data use before providing access to the data in a secure manner.

The toolkit concludes with a list of resources and appendices with supporting materials that registries may find helpful.

The toolkit can be found in a separate document, "Data Use and Governance Toolkit" (see complete list of project artifacts in Appendix B).

Use of Natural Language Processing for Extracting Measures Data

Some data that are necessary for calculation of the harmonized outcome measures are not available in structured form, as noted above. The purpose of this task was to determine the feasibility of extracting relevant data, including suicide ideation and behavior, PHQ-9 scores, adverse effects of treatment, and psychiatric comorbidities from clinical notes.

Methods

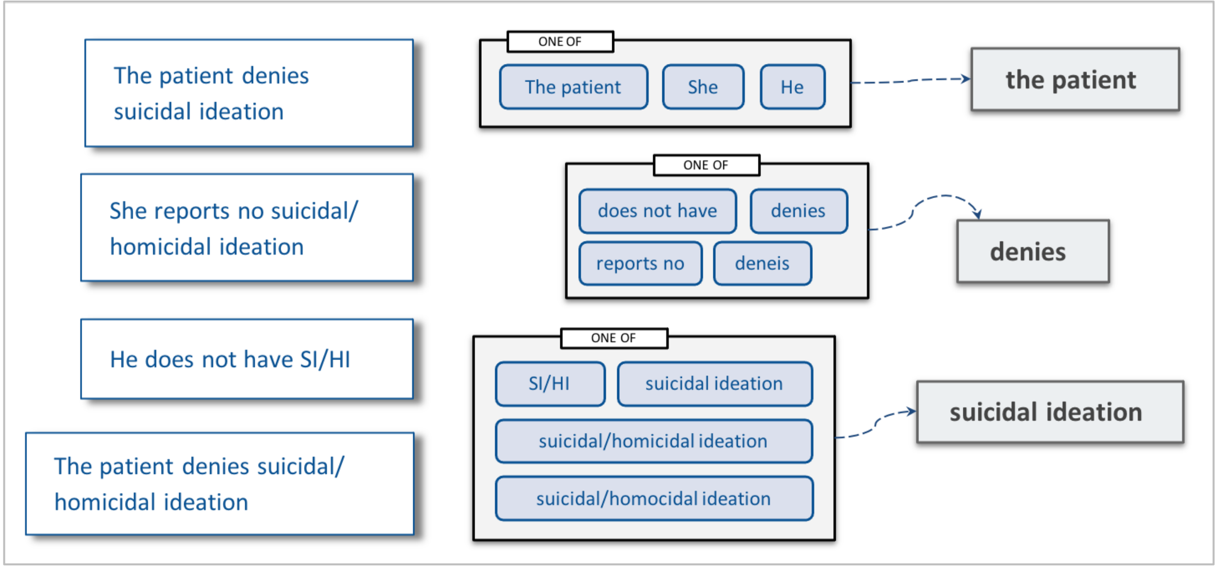
Data for this study were drawn from the OM1 Real-World Data Cloud (OM1, Inc, Boston, MA, USA). All data were de-identified, and the study was reviewed and approved by the Advarra IRB. Three cohorts were created for this study. The suicide ideation and behavior cohort was restricted to patients with at least one clinical note that mentioned suicide. The PHQ-9 cohort was restricted to patients with at least one clinical note that mentioned the PHQ-9. Eligibility for the PHQ-9 cohort was not restricted based on diagnosis of depression, both because the PHQ-9 may be used as a screening tool for depression and because other studies have shown that patients with elevated PHQ-9 scores sometimes lack a structured diagnosis code for depression. The adverse effects of treatment cohort was restricted to patients with a diagnosis of depression who had at least one mention of a selective serotonin reuptake inhibitor (SSRI), while the psychiatric comorbidities cohort was restricted to patients with a diagnosis of depression.

Two approaches were used to extract concepts related to suicide ideation and behavior. First, the project team selected and configured two open-source NLP packages (SpaCy[39] and cTAKES[40]) to extract suicide ideation and behavior from unstructured clinical notes. Concepts identified for extraction were:

- Presence of suicide ideation (including PHQ-9 item 9 scores)
    a. Frequency over the past 2 weeks (once, several days, more than half the days, nearly every day)
    b. Date
- Negation of suicide ideation
- Noted suicide attempt

Second, the project team used a custom approach to extract the same concepts for suicide ideation and behavior from unstructured clinical notes. This approach employed standard methods focused on identifying the linguistic patterns that are used to record relevant data in clinical notes. Examples of the linguistic patterns for recording suicide ideation and behavior are shown in Figure 9 below.

Figure 9. Linguistic patterns for recording suicide ideation and behavior in clinical notes


The patient denies suicidal ideation
She reports no suicidal/ homicidal ideation
He does not have SI/HI
The patient denies suicidal/ homicidal ideation
ONE OF
The patient
She
He
the patient
ONE OF
does not have
denies
reports no
deneis
denies
ONE OF
SI/HI
suicidal ideation
suicidal/homicidal ideation
suicidal/homocidal ideation
suicidal ideation

Results from both approaches were validated using subject matter experts, and the approaches were compared. Based on the results of the comparison, the custom approach was used to extract PHQ-9 scores and associated dates, adverse effects of treatment, and psychiatric comorbidities.

Findings

The presence of suicide ideation and behavior was extracted with low precision using the open-source NLP programs (Table 8).

Table 8. Precision and recall using open-source natural language processing packages

Package	Concept	Precision	Recall
SpaCy	Presence of suicide ideation and behavior	0.14	0.60
	Negation of suicide ideation and behavior	0.96	0.83
cTAKES	Presence of suicide ideation and behavior	0.08	0.64
	Negation of suicide ideation and behavior	0.99	0.28

False positives were the primary issue with these packages. Many notes that contained text about suicide ideation and behavior were categorized as positive responses even though the note did not indicate that the patient currently was experiencing suicide ideation and behavior (e.g., prior history of suicide ideation, family history of suicide, medication label text referencing suicide). These packages also miscategorized negations of suicide ideation and behavior, particularly when the negation was not directly next to the text about suicide ideation. Several examples of the types of phrases that were miscategorized using these packages are presented in Table 9.

Table 9. Examples of miscategorized note text

Note Text	Open-Source Package Categorization	Expected Categorization	Possible Explanations
9. Thoughts that you would be better off dead or of hurting yourself in some way-0	Neither positive response nor negation	Negation	Does not recognize PHQ-9 question and resulting score
Call immediately if this occurs and certainly call if you ever have any suicidal thoughts.	Positive response	Neither positive response nor negation	Does not recognize text about potential side effects / warnings
Have you been having thoughts about killing yourself? No	Neither positive response nor negation	Negation	Difficulty with question and answer format
SI: Denied	Positive response	Negation	Recognizes use of colon as indicating a new concept
Denies- depression, anxiety,panic attacks,suicidal ideations,homicidal ideations,	Positive response	Negation	The ‘denies’ is too far from the phrase ‘suicidal ideations’
Cymbalta - caused suicidal ideations.	Positive response	Neither positive response nor negation	Does not recognize that this is most likely indicating past suicidal ideations
Deneid weight or appetite changes, suicidal or homicidal ideations, never saw psych, but used to go to therapy	Positive response	Negation	Misspelling of ‘deneid’ may have resulting in misclassification

The utility of open-source NLP packages may have been limited by the use of non-standard English to record the relevant concepts in clinical notes. For example, suicide ideation and behavior were documented as part of lists of signs and symptoms in some cases, meaning the negation was not directly next to the phrase, ‘suicide ideation.’ In other cases, use of punctuation or other characters and misspellings may have led to misclassifications. Finally, the phrase ‘suicide ideation’ appeared in note text in several contexts (e.g., medication label text, family history), possibly leading to misclassifications. Similar examples of documentation using non-standard English were observed with extraction of the PHQ-9 and adverse effects of treatment (e.g., use of semi-structured templates with all questions and answer options for the PHQ-9 with sub-scores followed by the total score).

Based on these initial findings, the project team explored the use of custom approaches that allowed for more nuanced assessments of the language around the variable of interest. The custom model for suicide ideation and behavior resulted in improved precision, as shown in Table 10.

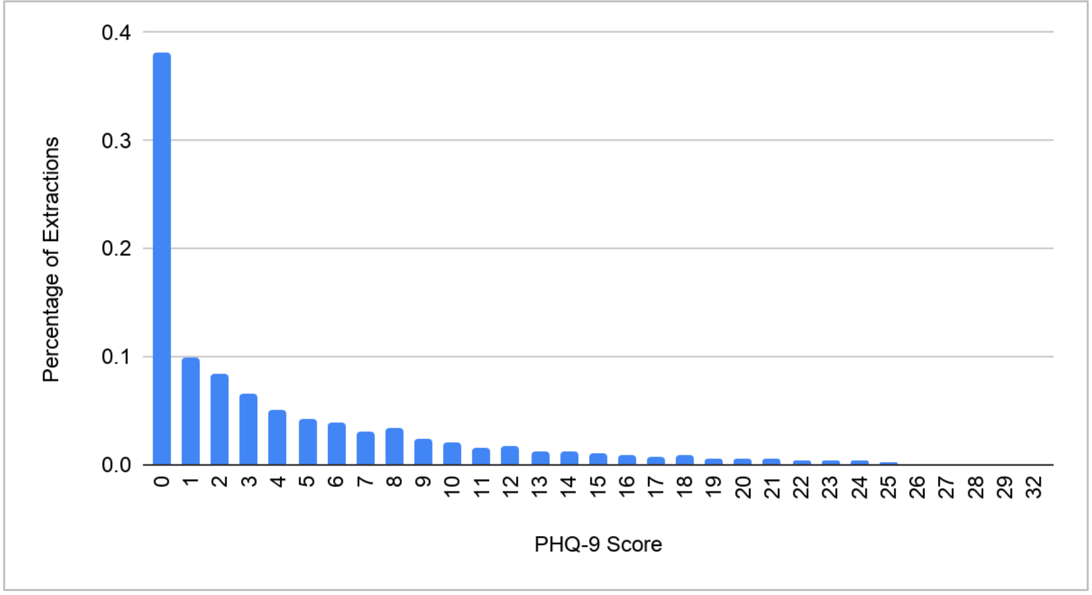
Table 10. Precision and recall using custom approaches

Approach	Concept	Precision	Recall
Custom Approach	Presence of suicide ideation and behavior	0.83	0.71
	Negation of suicide ideation and behavior	0.93	0.70

Using the inclusion criteria described above, 3.7 million patient notes were identified for inclusion in the suicide ideation and behavior cohort. Notes were categorized as negation of suicide ideation and behavior (e.g., ‘patient denies thoughts of suicide’), presence (e.g., ‘patient reports having thoughts of suicide’), neither negation nor presence (e.g., notes that include medication label text about suicide ideation), and unknown. Using the custom approach, 2,088,144 notes were categorized as negation, 25,255 notes as presence, 164,797 notes as neither, and 1,508,318 notes as unknown. Of the notes with a negation or presence, 1.2 percent were classified as presence and 98.8 percent as negation.

Based on the findings for suicide ideation and behavior, the custom approach was applied to the cohorts for PHQ-9 scores, adverse effects of treatment, and psychiatric comorbidities. For the PHQ-9 cohort, 1.1 million notes were identified for inclusion, and 735,267 PHQ-9 scores were extracted. Of these, 119,503 have a date other than the date of the encounter, and these dates were also extracted. The distribution of extracted PHQ-9 scores is presented in Figure 10 below.

Figure 10. Distribution of extracted PHQ-9 scores



For the adverse effects of treatment cohort, 2.6 million notes were identified for inclusion. Of these, 107,503 notes had at least one adverse effect mentioned in the same sentence or within seven words of the SSRI mention. A total of 36,844 adverse effects were extracted from 33,332 notes. Common adverse effects mentioned in the notes were weight gain, nausea, fatigue, sexual side effects, diarrhea, insomnia, headache, agitation, and suicidal thoughts.

For the psychiatric comorbidities cohort, 2.9 million notes were identified for inclusion. Notes were classified as affirmations, meaning the note indicated that the patient had a psychiatric comorbidity of interest; negations, meaning the note indicated that the patient did not have any comorbidities of interest; exclusions, meaning the note mentioned a comorbidity of interest in another context (e.g., mention of a questionnaire, such as the Generalized Anxiety Disorder-7 [GAD-7]); or unknown. Comorbidities of interest for this study were generalized anxiety disorder, obsessive-compulsive disorder, panic disorder, post-traumatic stress disorder, social anxiety disorder, substance use disorder, alcohol use disorder, bipolar disorder, schizophrenia, and insomnia. Of the included notes, 1,703,849 were classified as affirmations, 1,763 as negations, 130,724 as exclusions, and the remainder as unknown. Generalized anxiety disorder was mentioned most frequently, followed by bipolar disorder, anxiety disorder, and substance use disorder.

This effort demonstrated that extraction of suicide ideation and behavior, PHQ-9 scores and associated dates, adverse effects of treatment, and psychiatric comorbidities is feasible, and custom approaches produced higher precision results. Future work should focus on improving methods for extracting these concepts as well as other concepts necessary for characterizing patients with depression and understanding depression treatment and outcomes. Specific areas for future research include categorization of suicide ideation as active versus passive and further work on distinguishing suicide ideation from homicidal ideation in notes. In addition, further work to examine the agreement between data extracted from notes and other data in the patient's record (e.g., documentation of suicide ideation in notes vs. item 9 scores on the PHQ-9 vs. diagnosis codes for suicide ideation and behavior) would be beneficial.

FHIR Implementation Guide and Libraries

The purpose of this task was to facilitate the implementation of the harmonized depression outcome measures in EHRs using a standards-based approach. The project team produced two products. First, the Outcome Criteria Framework Implementation Guide presents an approach for defining outcome measures such that the measures can be applied across a range of use cases. The process described in the Implementation Guide was demonstrated through development of a FHIR Library containing a set of examples using the harmonized depression outcome measures. Both products are described further below.

Outcome Criteria Framework Implementation Guide

The Outcome Criteria Framework Implementation Guide defines a reproducible method and a formalism for representing condition-specific outcome definitions and criteria, such that the representations can be reused across different use cases, including quality measures, decision support tools, research studies, and routine clinical care. The Guide provides a reproducible framework for formalizing outcome measure definitions in a way that is aligned with healthcare interoperability standards, quality reporting processes, and the normal process of care, with the goals of minimizing data capture burden and increasing the prospects of data reuse.

The Outcome Criteria Framework Implementation Guide and the Project Scope Statement for the Implementation Guide are available publicly at the following locations:

- Implementation Guide: http://build.fhir.org/ig/HL7/fhir-outcome-criteria-framework-ig/

- Project Scope Statement: https://confluence.hl7.org/display/CQIWC/Outcome+Criteria+Framework+Implementation+Guide.

FHIR Library for Depression Outcome Measures

The process described in the Implementation Guide was used to create a core set of concrete definitions for the depression outcome measures, expressed in FHIR and based on QI Core and CQL, that could be leveraged by many different types of artifacts, such as quality measures or clinical decision support tools. The Library is available at the following location: https://github.com/HL7/fhir-outcome-criteria-framework-ig.

SMART on FHIR App

As discussed in Chapter 3, the Major Depression Outcomes SMART on FHIR app was created to facilitate the collection and calculation of the harmonized depression outcome measures within a health system. The primary objective of the app is to combine clinical information with PROs to provide clinicians with a dashboard showing an individual patient's depression symptoms, treatment, and outcomes. In addition to the dashboard, the app includes optional supporting services designed to facilitate collection of the PHQ-9 and transferring of data to registries.

The app is designed as an open-source, open-standards app that could be used in different EHR systems to collect and calculate the harmonized depression outcome measures. EHR system capabilities and implementations vary widely, and some modifications will likely be necessary to implement the app in new environments. The technical documentation provided with the app includes information relevant to implementing the app in widely used EHR systems. The app and technical documentation are available publicly in the SMART App Gallery (https://apps.smarthealthit.org/).

6. Implications for Future Use of Harmonized Outcome Measures

In this program, harmonized outcome measures for depression were calculated and used in two patient registries and one large health system. There were several findings and lessons learned that have implications for future implementations of the harmonized depression outcome measures and harmonized outcome measures developed for use in other clinical areas. This project demonstrated that it is feasible to calculate the harmonized depression outcome measures in a variety of care settings and that the data can then be used for research and measurement-based care purposes. However, technical and institutional barriers remain, and future implementations should be designed with these barriers in mind. The following sections summarize the key findings and lessons learned in this project and discuss implications for future implementations of harmonized outcome measures in depression and other clinical areas.

Key Findings and Lessons Learned

Feasibility of Calculating the Harmonized Measures

This project demonstrated that collection of the necessary data and calculation of the harmonized depression outcome measures is feasible in a variety of care settings. In the prior project, the harmonized narrative definitions were translated into standardized definitions that defined the initial population for measurement (e.g., all depression patients), the outcome focused population (patients who experienced the outcome of interest), and the data criteria and value sets for each outcome measure.[11,18] The goal of that effort was to define the measures with sufficient clarity such that they could be implemented consistently across data collection efforts. This effort tested the clarity of the definitions by implementing them in the family medicine, mental health, and health system settings. The definitions were interpreted consistently across care settings and implemented in the same manner by different organizations. This finding suggests that it is feasible to implement the definitions in other registries, research studies, health systems, and clinical practice settings, with an expectation that the measure results can be aggregated and compared.

While this project demonstrated that the measure definitions can be interpreted consistently, the measure results still should be considered in the context of the many factors that may have influenced the care provided, documentation of that care, and resulting outcomes. For example, the patients receiving care for depression in the mental health care and family medicine settings may differ in terms of disease severity, psychiatric comorbidities, access to mental health professionals, or other characteristics. As with any observational research, it is critical to identify these types of factors and consider their impact on outcomes when interpreting measure results from different settings. In fact, identification of the characteristics of the patient, disease, and clinician that influence patient outcomes is a key step in using the Outcome Measures Framework[7] to develop harmonized outcome measures. By providing a consistent framework for viewing the depression patient outcomes across care settings, clinicians, and patients, the harmonized measures create opportunities to identify differences in outcomes and embark on more focused investigations into the factors driving those differences, ideally leading to new insights into how to improve patient outcomes across care settings.

This project also demonstrated that most of the necessary data for calculation of the harmonized measures are recorded as part of routine clinical practice, albeit sometimes in unstructured text, making it feasible technically to extract the data and calculate the measures at the population level and at the individual patient level. It was also feasible to create standardized resources, such as the FHIR Implementation Guide, FHIR Library, and SMART on FHIR app, to support implementations of the measures. The implementations of the measures and the development of standardized resources identified some limitations in data availability, as noted in Chapters 2 and 4 of this report, particularly related to recording of PHQ-9 scores, adverse effects of treatment, suicide ideation and behavior, and cause of death. Of note, future implementations should plan to collect the PHQ-9 using multiple modalities and to extract the PHQ-9 data from multiple locations (e.g., registries, EHR structured fields, EHR unstructured notes, standalone PRO systems). Future implementations of the measures would also benefit from including data extracted from unstructured clinical notes and linkages to other data sources, such as administrative claims data, to generate a more robust view of patient treatments and outcomes over time.

Finally, this project demonstrated that it is feasible technically to capture the PHQ-9 at consistent intervals using email and text messages. However, more work is needed to identify and address the operational barriers that resulted in low response rates for PHQ-9s in this project. In particular, it is critical to understand how to implement the harmonized outcome measures with adequate messaging and resources to encourage patients to complete the PHQ-9 and to encourage clinicians to view and use the measures as part of clinical decision-making. Improving the workflow, resources, and messaging for the measures will be particularly important for implementations that seek to capture the full set of harmonized measures, including the Quality of Life Enjoyment and Satisfaction Questionnaire (Q-Les-Q) and the Work Productivity and Activity Impairment Questionnaire (WPAI).[41, 42]

Use of the Harmonized Measures for Research

A key goal of the harmonized measures is to create data infrastructure that can be used to link, compare, and aggregate data from multiple sources to support research. This project demonstrated that the harmonized depression outcome measures can be extracted from patient registries and used for research purposes. It was feasible to design a pilot data analysis that used the harmonized outcome measures and additional data from the patient registries to explore questions about depression treatment and outcomes. The data were captured through the registries over the course of the data collection period and then extracted and analyzed per a pre-defined statistical analysis plan. The Stakeholder Panel was also able to identify PCOR questions about depression that could be answered using the harmonized outcome measures. However, some limitations were noted, particularly related to the capture of patient and disease characteristics of interest. Future uses of the measures for research purposes may consider linking to other data sources (e.g., administrative claims data) to generate a more complete view of patients' treatments and outcomes. More work is also needed to improve methods and documentation for some critical aspects of depression care. For example, further work is needed to determine how best to capture information on adverse effects of treatment, how to capture and document social determinants of health, and how to document whether measurement-based care was used when providing care for a patient. Finally, future studies that use the harmonized outcome measures should consider extracting relevant data using NLP-based methods, as some relevant data are not recorded routinely using structured fields in the EHR.

Value of Implementing the Harmonized Measures

This project was designed to assess the value and burden of calculating the measures at the clinician level, health system level, and registry level. At the clinician level, the Major Depression Outcomes SMART on FHIR app proved to be a highly useful tool for visualizing and summarizing the longitudinal patient characteristics, treatments, and standardized outcomes in a clear and actionable way. Because the app was accessed directly from the EHR and relied on routinely recorded patient data, use of the app itself introduced very little burden for clinicians. Clinicians participating in this pilot project did experience increased burden due to the need to consent patients prior to use of the app; however, this requirement was necessary in the context of the research study and does not reflect the burden of using the app in routine clinical care. At the health system level, some effort was required to obtain the necessary approvals to implement the app, and IT resources were required for app deployment because the app is designed to be deployed within the health system's firewall.

At the registry level, implementation of the measures required programmatic support from registry staff and technology costs associated with modifying the data extraction and adding the measures to the registry dashboards. Registry pilot sites also experienced some burden in the form of ensuring the measures data were extracted and mapped correctly, implementing workflows where necessary to capture and review the PHQ-9, and revising existing workflows to document the PHQ-9 and item 9 scores so that they could be extracted for registry purposes. This project demonstrated that the outcome measures data can be used for research purposes. Ideally, this demonstration will lead to new opportunities for the registries to conduct funded research alone or in partnership with other organizations, thus producing revenue for the registries.

Impact of the COVID-19 Pandemic

While this project clearly demonstrated technical feasibility, the COVID-19 pandemic created obstacles to the demonstrations of operational feasibility, value, and burden. Specifically, the pandemic created unprecedented disruptions in clinical care across care settings, including substantially reduced office visits, a rapid switch to telehealth, and delayed care for routine and non-urgent appointments. While the effect on patient enrollment in the studies conducted under this project cannot be quantified directly, it is reasonable to conclude that enrollment was reduced due to fewer office visits, fewer opportunities to complete the baseline PHQ-9 in the pilot data analysis, and fewer opportunities to seek informed consent from patients for the study of the SMART on FHIR app. In addition, research has shown that U.S. adults experienced significantly higher levels of stress and anxiety in 2020;[43] this may have affected response rates for the PHQ-9 and willingness to enroll in the research study. Reduced enrollment and low response rates in turn led to a smaller sample size for pilot data analysis, thus minimizing the potential value of using the registry data for research purposes. In the app study, reduced enrollment led to fewer interactions with the app, potentially altering perceptions of its usability and value.

The pandemic also introduced significant stress for health care professionals and other relevant resources on multiple fronts. Health care professionals faced challenges related to providing care for COVID-19 patients, adopting new technologies and changing workflows to care for all other patients in a manner that reduced the risk of infection, and responding to financial uncertainties caused by reduced use of health care services by many people. Implementation of new workflows and new technologies was extremely difficult during this period, even with the goal of helping to provide care for patients with depression during the pandemic. Clinicians, staff, and IT resources had less time than anticipated to devote to this project, particularly during the initial roll-out of the measures in the spring and summer of 2020. Because the measures were not in use prior to the pandemic, it is difficult to draw any firm conclusions about how operational feasibility and assessments of value and burden may differ due to the pandemic. However, given the magnitude of the challenges facing health care professionals over the past year, it is reasonable to surmise that the pandemic negatively affected operational feasibility and possibly assessments of value and burden.

Future Research Needs

This project identified several areas for future research to support implementation and use of the harmonized depression outcome measures. First, many patients in this pilot project did not respond to emailed invitations to complete the PHQ-9, resulting in a lack of information on depression symptoms to calculate and make use of the harmonized outcome measures. Additional work to identify potential barriers to completion of the PHQ-9 and to develop communication and other tools to address these barriers would be valuable for increasing response rates for the PHQ-9 and possibly other PROs. Collaboration with patient organizations would be particularly beneficial to discuss barriers, identify possible solutions, and improve messaging. Second, more work is needed to identify or create methods for capturing adverse effects related to treatment. This project explored the use of structured and unstructured EHR data for this measure. Future efforts may explore how best to use PROs, such as the FIBSER, in clinical practice settings. Finally, more work is needed to improve collection of patient and clinician characteristics, such as social determinants of health and use of measurement-based care, that influence outcomes, so outcomes can be interpreted in the appropriate context. These issues are discussed in more detail in the Prioritized Research Agenda (Chapter 5).

Implications for Future Implementations

This project produced several findings that should inform future implementations of harmonized outcome measures in depression and other clinical areas, such as asthma,[10] atrial fibrillation,[9] lumbar spondylolisthesis, and non-small cell lung cancer. Lessons learned specific to implementation of the harmonized depression outcome measures are presented in earlier chapters of this report, while implications of this project's findings for implementation of harmonized outcome measures generally are discussed below.

Develop Flexible, Patient-Centered Plans for Collecting PROs

Collection of PROs is critical for understanding patient outcomes. PROs can provide insight into response to treatment, burden of illness, side effects related to treatment, disease progression over time, and other areas. The Outcome Measures Framework includes PROs as one of the five categories of outcome measures, recognizing their importance in understanding patient outcomes over time, and the harmonized measures developed under the prior project include recommendations for capturing PROs in each clinical area.

While there is broad recognition of the value of PROs, this project highlighted the challenges of collecting PROs in routine clinical care. Collection of PROs in the research setting has become more common in recent years, but collection of PROs in routine clinical practice is far from universal. Even when PROs are collected as part of routine care, variation exists in how the PROs are collected and documented. This variation could be viewed as a challenge to be addressed in future implementations. However, it is perhaps more helpful to view this variation as the result of clinicians using the most appropriate modality for an individual patient. PROs, ideally, are an important component of information sharing between doctors and patients. Tools to help patients complete PROs and to help clinicians view outcome measures based on PROs should support – not interfere with – the patient-doctor relationship. When viewed through this lens, it is clear that the challenge is not to reduce variation, but rather to provide flexibility in PRO modalities to accommodate the needs and preferences of diverse patient populations.

Future implementations of harmonized outcome measures should build on this lesson learned by developing flexible, patient-centered plans for collecting PROs using multiple modalities. Careful planning for how total scores and sub-scores (if relevant) will be documented and extracted for use in measure calculation (e.g., from structured fields and unstructured notes in

EHRs, standalone PRO systems) is essential, and planning should encompass clinical workflows, communication tools for clinicians and patients, and standards-based technological tools.

Use a Multipronged Approach To Capturing Data

The harmonized outcome measures produced in the prior project focused on data that are captured and recorded as part of routine clinical care, making the measures suitable for use in observational studies, such as patient registries, and in clinical care. Implementation of the depression measures showed that some data are recorded routinely in clinical notes, as opposed to in structured fields in the EHR. This challenge is not specific to depression and is likely to occur in other implementations of the harmonized outcome measures. For example, information on outcomes such as adverse effects of treatment, events of interest that occurred outside of the practice setting, and PRO scores may be more likely to be found in clinical notes. Information on the characteristics of the patient and the disease, such as social determinants of health, family history, and disease course, may also be more likely to be found in clinical notes.

Future implementations of harmonized outcome measures should develop a multi-pronged approach to capturing all relevant data for the calculation and interpretation of the outcome measures. This may include, for example, creation of structured fields in the EHR and associated workflows for documenting key data elements, use of NLP approaches to extract data from unstructured clinical notes, and linkage to other data sources, such as administrative claims data, to provide a more complete view of treatments and outcomes over time.

Build on Existing Standards-Based Resources

A broad goal of developing and implementing harmonized measures is to support the creation of data infrastructure that could serve as the foundation for learning health systems, population health management efforts, quality improvement initiatives, value-based care programs, and research studies. By capturing outcome measures consistently across care settings, stakeholders such as health care professionals, researchers, payers, and others will be able to reuse data for multiple purposes, thus reducing the burden of data collection while at the same time building connections to support the translation of research findings into practice and ultimately improving patient outcomes. In addition to the lessons learned noted above, this project created standards-based resources to support future implementations of harmonized outcome measures and to describe clearly the central role of standardized outcome measures in creating a data ecosystem that uses real-world data for multiple purposes. Of note, the Outcome Criteria Framework Implementation Guide provides a clear methodology for translating narrative outcome measure definitions into a core set of concrete definitions that leverage healthcare interoperability standards, such as FHIR, and that could be reused in the context of quality measurement, clinical decision support tools, population health management tools, and research – thereby connecting all components of a learning health system in a standards-based manner.

Future implementations of harmonized measures should refer to the Implementation Guide as a starting point for development of standards-based definitions for outcome measures. The FHIR Library for depression also provides a clear set of examples to guide future efforts.

7. Conclusions

Standardization of outcome measures across patient registries and routine clinical care is an important step toward creating robust, national-level data infrastructure that could serve as the foundation for learning health systems, quality improvement initiatives, and patient-centered outcomes research. This project demonstrated that it is feasible to calculate the harmonized outcome measures for depression in two patient registries and a health system setting, display the results to clinicians to support individual patient management and population health, and use the outcome measures data to support patient-centered outcomes research. This project also assessed the value and burden of capturing the measures in different care settings and created standards-based tools and other resources to support future implementations of harmonized outcome measures in depression and other clinical areas.

The findings from this project suggest that implementation of the harmonized outcome measures for depression is feasible across a variety of care settings. Future implementations of the measures in additional care settings and additional data collection efforts would increase the value of the existing data infrastructure for conducting patient-centered outcomes research. The findings and lessons learned from this project should serve as a roadmap to guide future implementations of harmonized outcome measures in depression and other clinical areas.

8. References

1. Gliklich RE, Leavy MB, Dreyer NA (sr eds). Registries for Evaluating Patient Outcomes: A User's Guide. 4th ed. (Prepared by L&M Policy Research, LLC, under Contract No. 290-2014-00004-C with partners OM1 and IQVIA) AHRQ Publication No. 19(20)-EHC020. Rockville, MD: Agency for Healthcare Research and Quality; September 2020. DOI: https://doi.org/10.23970/AHRQEPCREGISTRIES4. .

2. PRIME Registry. American Board of Family Medicine [Available from: https://primeregistry.org/.

3. PsychPRO. American Psychiatric Association [Available from: https://www.psychiatry.org/psychiatrists/registry.

4. 21st Century Cures Act, Public Law 114-255, 114th Cong., 2d sess. December 13, 2016. [Available from: https://www.gpo.gov/fdsys/pkg/PLAW-114publ255/pdf/PLAW-114publ255.pdf.

5. U.S. Food and Drug Administration. National Evaluation System for Health Technology (NEST) [Available from: https://nestcc.org/.

6. Centers for Medicare & Medicaid Services. Guidance for the Public, Industry, and CMS Staff. Coverage with Evidence Development. November 20, 2014 [Available from: https://www.cms.gov/medicare-coverage-database/details/medicare-coverage-document-details.aspx?MCDId=27.

7. Gliklich RE, Leavy MB, Karl J, et al. A framework for creating standardized outcome measures for patient registries. Journal of comparative effectiveness research. 2014;3(5):473-80.

8. Leavy MB, Schur C, Kassamali FQ, Johnson ME, Sabharwal R, Wallace P, Gliklich RE. Development of Harmonized Outcome Measures for Use in Patient Registries and Clinical Practice: Methods and Lessons Learned. Final Report. (Prepared by L&M Policy Research, LLC under Contract No. 290-2014-00004-C) AHRQ Publication No. 19-EHC008-EF. Rockville, MD: Agency for Healthcare Research and Quality; February 2019. DOI: https://doi.org/10.23970/AHRQEPCLIBRARYFINALREPORT.

9. Calkins H, Gliklich RE, Leavy MB, et al. Harmonized outcome measures for use in atrial fibrillation patient registries and clinical practice: Endorsed by the Heart Rhythm Society Board of Trustees. Heart Rhythm. 2019;16(1):e3-e16.

10. Gliklich RE, Castro M, Leavy MB, et al. Harmonized outcome measures for use in asthma patient registries and clinical practice. J Allergy Clin Immunol. 2019;144(3):671-81.e1.

11. Gliklich RE, Leavy MB, Cosgrove L, et al. Harmonized Outcome Measures for Use in Depression Patient Registries and Clinical Practice. Ann Intern Med. 2020;172(12):803-9.

12. Harbaugh RE, Devin C, Leavy MB, et al. Harmonized Outcome Measures for Use in Degenerative Lumbar Spondylolisthesis Patient Registries and Clinical Practice. J Neurosurg Spine. Published online March 19, 2021. doi: https://doi.org/10.3171/2020.9.SPINE20437.

13. Edelman MJ, Raymond DP, Owen DH, et al. Harmonized Outcome Measures for Use in Non-Small Cell Lung Cancer Patient Registries and Clinical Practice. J Natl Compr Canc Netw. Forthcoming 2021.

14. Brody DJ, Pratt LA, Hughes J. Prevalence of depression among adults aged 20 and over: United States, 2013–2016. NCHS Data Brief, no 303. Hyattsville, MD: National Center for Health Statistics. 2018.

15. Machado MO, Veronese N, Sanches M, et al. The association of depression and all-cause and cause-specific mortality: an umbrella review of systematic reviews and meta-analyses. BMC Med. 2018;16(1):112.

16. Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med. 2001;16(9):606-13.

17. Hamilton M. A rating scale for depression. J Neurol Neurosurg Psychiatry. 1960;23(1):56-62.

18. Gliklich RE, Leavy MB, Li F. Standardized Library of Depression Outcome Measures. Research White Paper. (Prepared by L&M Policy Research, LLC under Contract No. 290-2014-00004-C.) www.effectivehealthcare.ahrq.gov.

19. Leavy MB, Cooke D, Hajjar S, Bickelman E, Egan B, Clarke D, Gibson D, Casanova B, Gliklich R. Outcome Measure Harmonization and Data Infrastructure for Patient-Centered Outcomes Research in Depression. Report on Registry Configuration. (Prepared by OM1, Inc., with subcontractors American Board of Family Medicine and American Psychiatric Association under Contract No. 75Q80119C00005.) AHRQ Publication No. 21-EHC003. Rockville, MD: Agency for Healthcare Research and Quality; November 2020. DOI: https://doi.org/10.23970/AHRQEPCREGISTRYOUTCOME.

20. Wisniewski SR, Rush AJ, Balasubramani GK, et al. Self-rated global measure of the frequency, intensity, and burden of side effects. Journal of psychiatric practice. 2006;12(2):71-9.

21. Anderson HD, Pace WD, Brandt E, et al. Monitoring Suicidal Patients in Primary Care Using Electronic Health Records. The Journal of the American Board of Family Medicine. 2015;28:65-71.

22. Yesavage JA, Brink TL, Rose TL, et al. Development and validation of a geriatric depression screening scale: a preliminary report. J Psychiatr Res. 1982;17(1):37-49.

23. Rush AJ, Trivedi MH, Ibrahim HM, et al. The 16-Item Quick Inventory of Depressive Symptomatology (QIDS), clinician rating (QIDS-C), and self-report (QIDS-SR): a psychometric evaluation in patients with chronic major depression. Biological psychiatry. 2003;54(5):573-83.

24. HL7 FHIR. Patient Reported Outcomes FHIR Implementation Guide. [Available from: http://hl7.org/fhir/us/patient-reported-outcomes/2018Sep/index.html.

25. Garcia S, Zayas-Cabán T, Ladwa S, Donahue D, Rafiqi A, Nguyen C, et al. Advancing the Collection and Use of Patient-Reported Outcomes through Health Information Technology. Prepared by ESAC, Inc. for the Office of the National Coordinator for Health Information Technology under Contract No. HHSP233201500103I HHSP23337002T. March 2020. https://www.healthit.gov/sites/default/files/page/2020-03/ONCPROFinalReportFinal.pdf.

26. Cleare A, Pariante CM, Young AH, et al. Evidence-based guidelines for treating depressive disorders with antidepressants: A revision of the 2008 British Association for Psychopharmacology guidelines. J Psychopharmacol. 2015;29(5):459-525.

27. Cuijpers P, Quero S, Dowrick C, et al. Psychological Treatment of Depression in Primary Care: Recent Developments. Curr Psychiatry Rep. 2019;21(12):129.

28. Davies P, Ijaz S, Williams CJ, et al. Pharmacological interventions for treatment-resistant depression in adults. The Cochrane database of systematic reviews. 2019;12(12):Cd010557.

29. Dunlop BW. Evidence-Based Applications of Combination Psychotherapy and Pharmacotherapy for Depression. Focus (Am Psychiatr Publ). 2016;14(2):156-73.

30. Gaynes BN, Asher G, Gartlehner G, et al. Definition of Treatment-Resistant Depression in the Medicare Population. Technology Assessment Program. Project ID: PSYT0816. (Prepared by RTI–UNC Evidence-Based Practice Center under Contract No. HHSA290201500011I_HHSA29032006T). Rockville, MD: Agency for Healthcare Research and Quality. February 2018. http://www.ahrq.gov/clinic/epcix.htm.

31. Gaynes BN, Lux LJ, Lloyd SW, et al. AHRQ Comparative Effectiveness Reviews. Nonpharmacologic Interventions for Treatment-Resistant Depression in Adults. Rockville (MD): Agency for Healthcare Research and Quality (US); 2011.

32. Gong J, Simon GE, Liu S. Machine learning discovery of longitudinal patterns of depression and suicidal ideation. PLoS One. 2019;14(9):e0222665.

33. Machmutow K, Meister R, Jansen A, et al. Comparative effectiveness of continuation and maintenance treatments for persistent depressive disorder in adults. The Cochrane database of systematic reviews. 2019;5(5):Cd012855.

34. O'Connor E, Rossom RC, Henninger M, et al. Primary Care Screening for and Treatment of Depression in Pregnant and Postpartum Women: Evidence Report and Systematic Review for the US Preventive Services Task Force. Jama. 2016;315(4):388-406.

35. Qaseem A, Barry MJ, Kansagara D. Nonpharmacologic Versus Pharmacologic Treatment of Adult Patients With Major Depressive Disorder: A Clinical Practice Guideline From the American College of Physicians. Ann Intern Med. 2016;164(5):350-9.

36. Sebastianski M, Gates M, Gates A, et al. Evidence available for patient-identified priorities in depression research: results of 11 rapid responses. BMJ Open. 2019;9(6):e026847.

37. Siu AL, Force USPST. Screening for Depression in Children and Adolescents: US Preventive Services Task Force Recommendation Statement. Pediatrics. 2016;137(3):e20154467.

38. Siu AL, Force USPST, Bibbins-Domingo K, et al. Screening for Depression in Adults: US Preventive Services Task Force Recommendation Statement. JAMA. 2016;315(4):380-7.

39. spaCy [Available from: https://spacy.io/.

40. Apache cTAKES [Available from: https://ctakes.apache.org/.

41. Endicott J, Nee J, Harrison W, et al. Quality of Life Enjoyment and Satisfaction Questionnaire: a new measure. Psychopharmacology bulletin. 1993;29(2):321-6.

42. Reilly MC, Zbrozek AS, Dukes EM. The validity and reproducibility of a work productivity and activity impairment instrument. PharmacoEconomics. 1993;4(5):353-65.

43. McKnight-Eily LR, Okoro CA, Strine TW, et al. Racial and Ethnic Disparities in the Prevalence of Stress and Worry, Mental Health Conditions, and Increased Substance Use Among Adults During the COVID-19 Pandemic - United States, April and May 2020. MMWR Morbidity and mortality weekly report. 2021;70(5):162-6.

Appendix A. Harmonized Depression Outcome Measures

Table A-1. Harmonized outcome measures selected for pilot project[a]

OMF Category	Outcome Measure	Definition
Survival	Death from suicide	Patient with a diagnosis of major depression or dysthymia who died from suicide, reported in 12-month intervals. *This should be captured where feasible; however, it should be noted that this information may not be recorded accurately or available to all health care professionals.*
Clinical Response	Improvement in Depressive Symptoms—Response	Patients aged 18 or older with a diagnosis of major depression or dysthymia and an initial PHQ-9* score > 9 who demonstrates a response to treatment defined as a PHQ-9 score that is reduced by 50% or greater from the initial PHQ-9 score. *The PHQ-9, or another brief, publicly available, validated patient-reported instrument with empirically derived cutpoints equivalent to the PHQ-9 cutpoints for remission and response and for which an evidence-based crosswalk to the PHQ-9 exists, should be used to measure clinical response. Other measures may be used in addition for research or other purposes. Timeframe for measurement: ● 6 months (+/- 60 days) ● 12 months (+/- 60 days) *In some implementations, it would be beneficial to capture earlier responses and remissions and to obtain higher degrees of follow-up. Additional measurements outside of the windows listed above are recommended as supplemental measures.*
Clinical Response	Improvement in Depressive Symptoms—Remission	Patients aged 18 or older with a diagnosis of major depression or dysthymia and an initial Patient Health Questionnaire-9 (PHQ-9)* score > 9 who demonstrates remission defined as a PHQ-9 score < 5. *The PHQ-9, or another brief, publicly available, validated patient-reported instrument with empirically derived cutpoints equivalent to the PHQ-9 cutpoints for remission and response and for which an evidence-based crosswalk to the PHQ-9 exists, should be used to measure clinical response. Other measures may be used in addition for research or other purposes. Timeframe for measurement: ● 6 months (+/- 60 days) ● 12 months (+/- 60 days) *In some implementations, it would be beneficial to capture earlier responses and remissions and to obtain higher degrees of follow-up. Additional measurements outside of the windows listed above are recommended as supplemental measures.*

[a] *Gliklich RE, Leavy MB, Cosgrove L, Simon GE, Gaynes BN, Peterson LE, Olin B, Cole C, DePaulo JR, Jr., Wang P, Crowe CM, Cusin C, Nix M, Berliner E, Trivedi MH. Harmonized Outcome Measures for Use in Depression Patient Registries and Clinical Practice. Ann Intern Med. 2020;172(12):803-9. Epub 2020/05/19. doi: 10.7326/M19-3818. PubMed PMID: 32422056.*

OMF Category	**Outcome Measure**	**Definition**
Clinical Response	Worsening in Depressive Symptoms—Recurrence	Patients aged 18 or older with a diagnosis of major depression or dysthymia and an initial PHQ-9* > 9 who demonstrates remission (defined as a PHQ-9 score < 5) of at least two months' duration and subsequently experiences a recurrence of a depressive episode, defined as a 50% increase in PHQ-9 score or defined as a PHQ-9 score > 9 OR hospitalization for depression or suicidality.** *The PHQ-9, or another brief, publicly available, validated patient-reported instrument with empirically derived cutpoints equivalent to the PHQ-9 cutpoints for remission and response and for which an evidence-based crosswalk to the PHQ-9 exists, should be used to measure clinical response. Other measures may be used in addition for research or other purposes. **This definition was proposed by the workgroup. Data accruing from ongoing registries are needed to assess the feasibility of using this definition to capture recurrence. Timeframe for measurement: ● 6 months (+/- 60 days) ● 12 months (+/- 60 days) *In some implementations, it would be beneficial to capture earlier responses and remissions and to obtain higher degrees of follow-up. Additional measurements outside of the windows listed above are recommended as supplemental measures.*
Events of Interest	Suicide Ideation & Behavior	Selection of "several days," "more than half the days," or "nearly every day" option on PHQ-9 item 9 ("Thoughts that you would be better off dead or of hurting yourself in some way"). *Supplemental assessments of suicide ideation and behavior should be completed for patients who screen positive for suicide ideation on the PHQ-9 or when a clinician has concerns about suicidality. Supplemental assessments should be completed using an appropriate, brief, validated instrument, such as the Concise Health Risk Tracking (CHRT) scale. Includes nonfatal suicide attempts/suicide attempt behaviors, planning/preparatory acts, and active suicidal ideation.* Reported in 12-month intervals (in conjunction with the PHQ-9 suicide item).
Events of Interest	Adverse Events	Depression treatment-related adverse events. Use of a brief, publicly available, validated measurement tool to capture adverse events is recommended. Reported in 12-month intervals.

Appendix B. Artifacts Submitted Separately

1. SMART on FHIR App

The SMART on FHIR App source code and related technical documentation are available publicly in the SMART App Gallery (https://apps.smarthealthit.org/).

2. FHIR Implementation Guide

The Outcome Criteria Framework Implementation Guide and the Project Scope Statement for the Implementation Guide are available publicly at the following locations:

- Implementation Guide: http://build.fhir.org/ig/HL7/fhir-outcome-criteria-framework-ig/
- Project Scope Statement: https://confluence.hl7.org/display/CQIWC/Outcome+Criteria+Framework+Implementation+Guide.

3. FHIR Library for Depression Outcome Measures

The Library for the harmonized depression outcome measures is available publicly at the following location: https://github.com/HL7/fhir-outcome-criteria-framework-ig. The CQL and links to the appropriate value sets (as stored in the Value Set Authority Center [VSAC]) can be found here.

4. Data Use and Governance Toolkit

The Data Use and Governance Toolkit was submitted to AHRQ as a separate document on May 15, 2021.

5. Prioritized Research Agenda for Using the Harmonized Outcome Measures to Support Patient-Centered Outcomes Research in Depression

The Prioritized Research Agenda was submitted to AHRQ as a separate document on May 15, 2021.

6. Final Manuscripts

Two manuscripts were produced as part of this project. The first manuscript describes the development of a SMART on FHIR app (see Chapter 3), and the second describes the findings from the pilot data analysis (see Chapter 4). Both manuscripts were submitted to AHRQ as separate documents on May 15, 2021 and will be submitted for journal publication at the conclusion of the project.

Appendix C. Clinician Survey

Attachment C: Clinician Survey

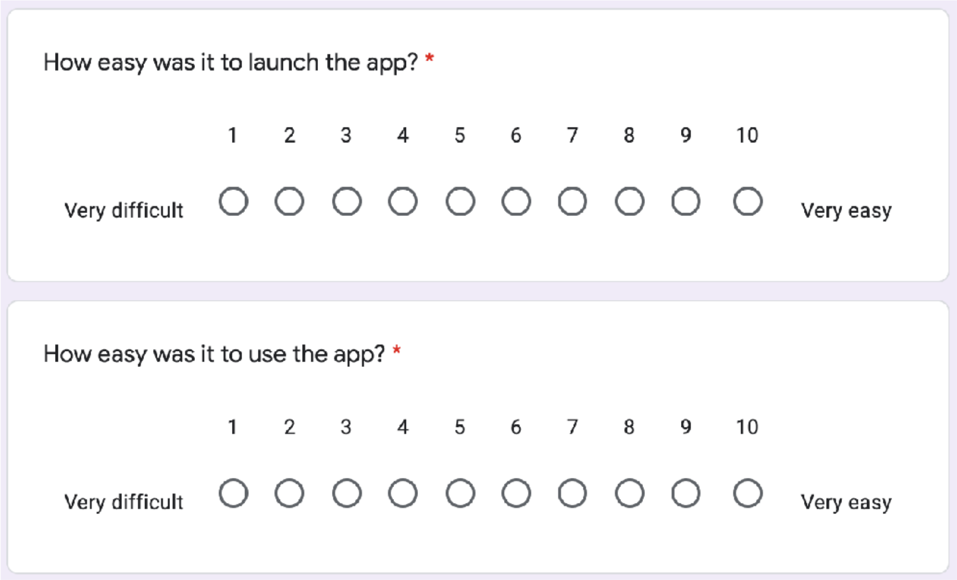
Form Approved
OMB No. 0935-0249
Exp. Date 01/31/2023

Clinician Survey on App Value and Feasibility

The purpose of this survey is to gather feedback on the Major Depression Outcomes app that you accessed as part of the 'Implementation of Harmonized Depression Outcome Measures in a Health System to Support Patient-Centered Outcomes Research' study (NCT04235712). The survey should take 5 minutes or less to complete, and your participation is voluntary.

Your responses to the survey questions will be used to assess the usefulness of the app and the harmonized depression outcome measures for informing clinical decision-making and the feasibility of using the app within your routine workflow. Your responses will be summarized in a report to the funding agency, the Agency for Healthcare Research and Quality, to help the agency understand the value of the app and the feasibility of implementation in other care settings. Information that could identify you will not be disclosed unless you have consented to that disclosure.

*Required

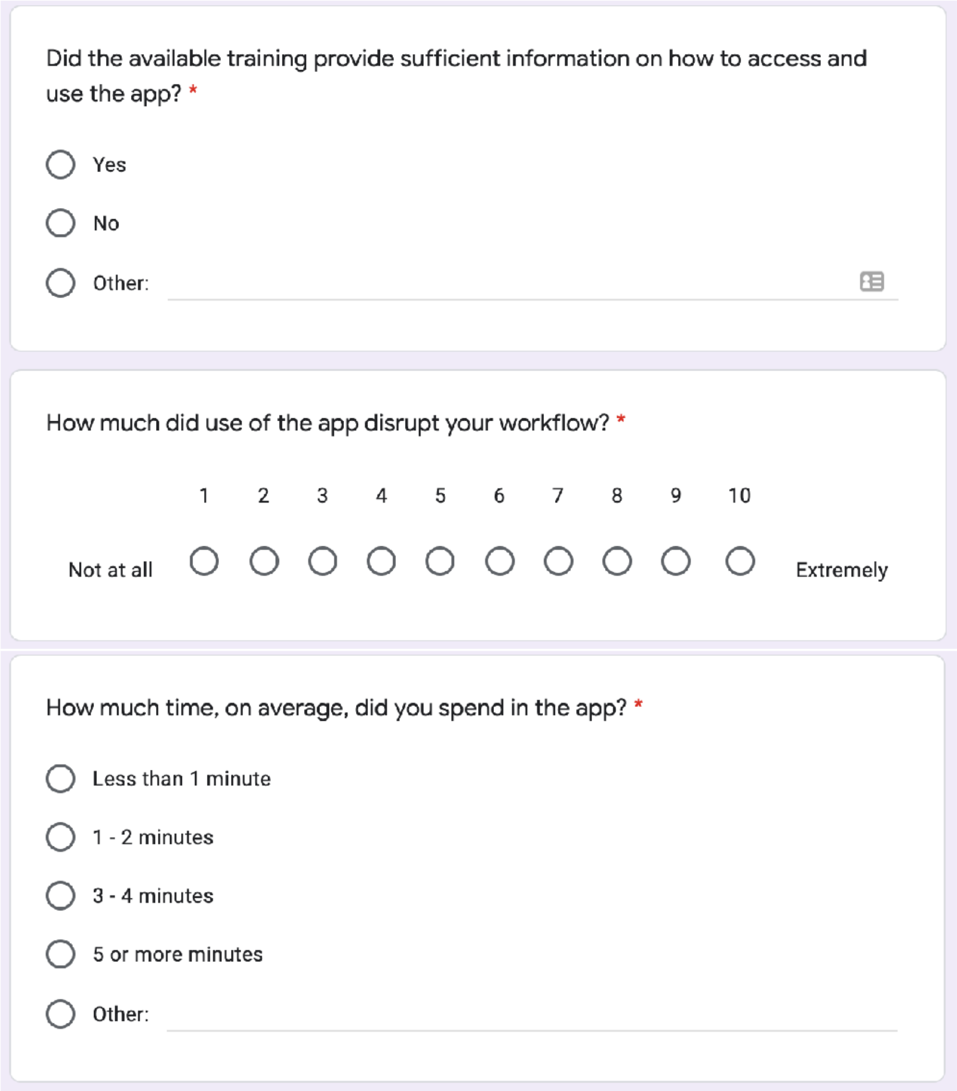


How easy was it to launch the app? *

	1	2	3	4	5	6	7	8	9	10	
Very difficult	○	○	○	○	○	○	○	○	○	○	Very easy

How easy was it to use the app? *

	1	2	3	4	5	6	7	8	9	10	
Very difficult	○	○	○	○	○	○	○	○	○	○	Very easy



Did the available training provide sufficient information on how to access and use the app? *

○ Yes

○ No

○ Other: ________________

How much did use of the app disrupt your workflow? *

	1	2	3	4	5	6	7	8	9	10	
Not at all	○	○	○	○	○	○	○	○	○	○	Extremely

How much time, on average, did you spend in the app? *

○ Less than 1 minute

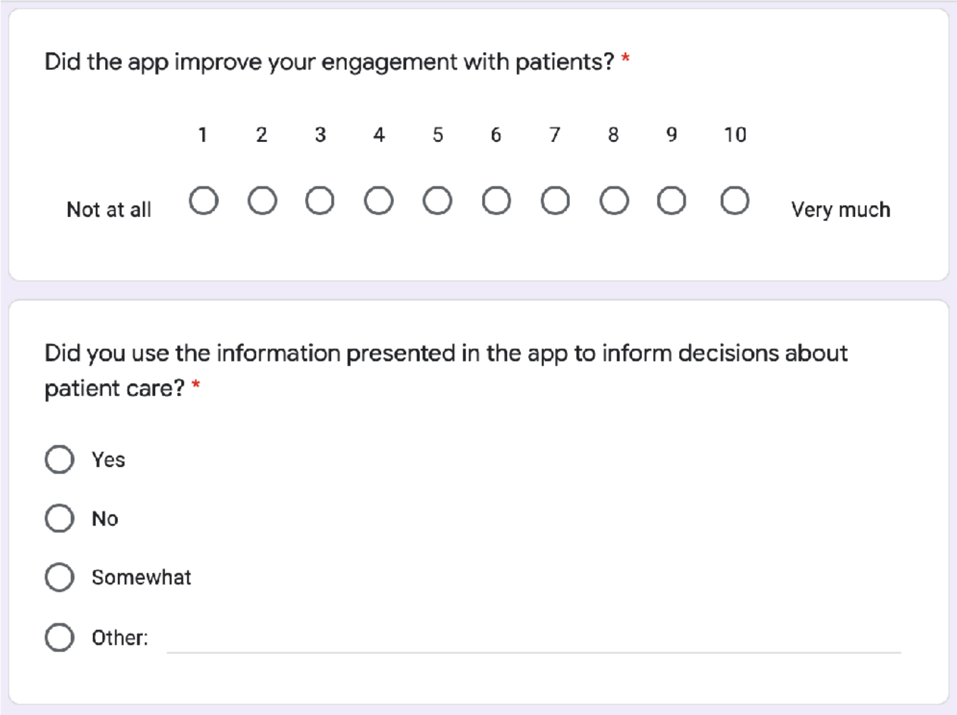
○ 1 - 2 minutes

○ 3 - 4 minutes

○ 5 or more minutes

○ Other: ________________

Attachment C: Clinician Survey



Did the app improve your engagement with patients? *

	1	2	3	4	5	6	7	8	9	10	
Not at all	○	○	○	○	○	○	○	○	○	○	Very much

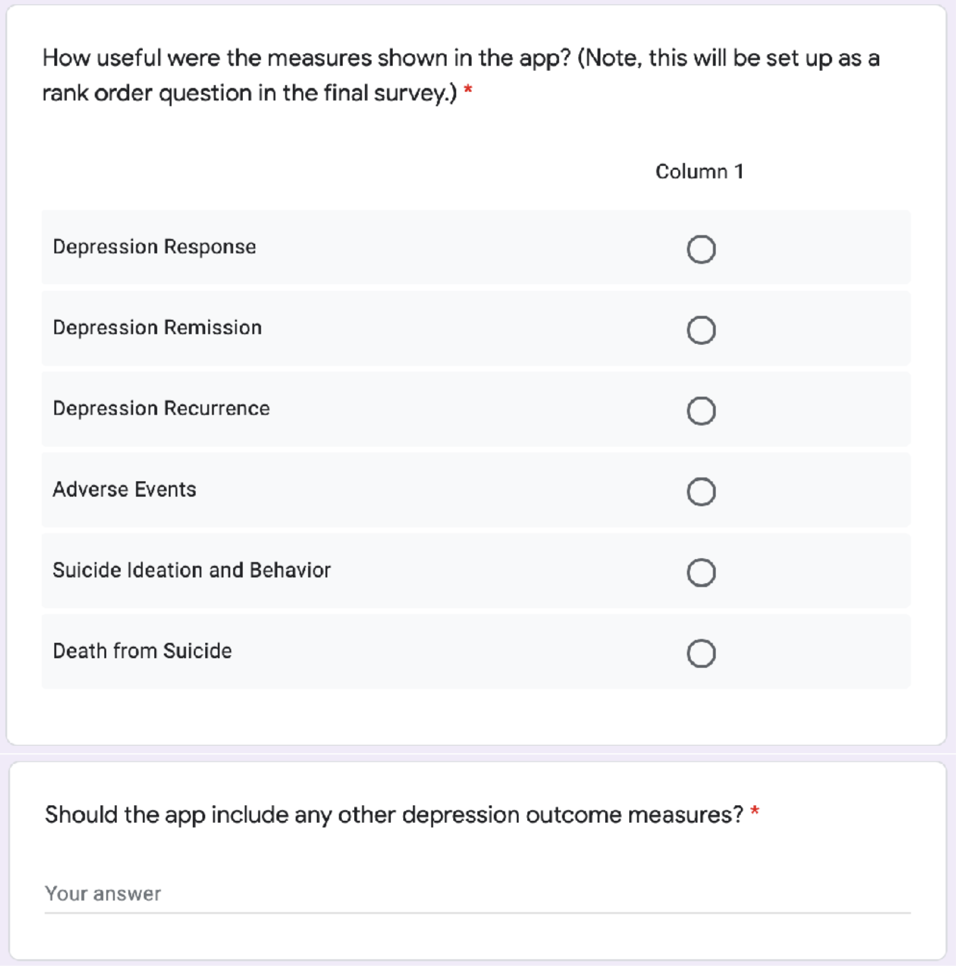
Did you use the information presented in the app to inform decisions about patient care? *

○ Yes

○ No

○ Somewhat

○ Other: ____________________

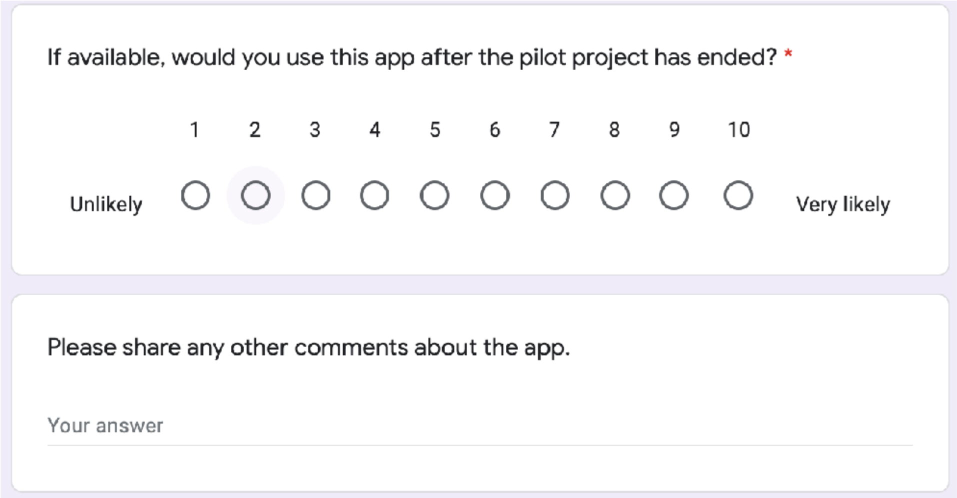


How useful were the measures shown in the app? (Note, this will be set up as a rank order question in the final survey.) *

	Column 1
Depression Response	○
Depression Remission	○
Depression Recurrence	○
Adverse Events	○
Suicide Ideation and Behavior	○
Death from Suicide	○

Should the app include any other depression outcome measures? *

Your answer



If available, would you use this app after the pilot project has ended? *

	1	2	3	4	5	6	7	8	9	10	
Unlikely	○	○	○	○	○	○	○	○	○	○	Very likely

Please share any other comments about the app.

Your answer

This survey is authorized under 42 U.S.C. 299a. The confidentiality of your responses to this survey is protected by Sections 944(c) and 308(d) of the Public Health Service Act [42 U.S.C. 299c-3(c) and 42 U.S.C. 242m(d)]. Information that could identify you will not be disclosed unless you have consented to that disclosure. Public reporting burden for this collection of information is estimated to average 5 minutes per response, the estimated time required to complete the survey. An agency may not conduct or sponsor, and a person is not required to respond to, a collection of information unless it displays a currently valid OMB control number. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to: AHRQ Reports Clearance Officer Attention: PRA, Paperwork Reduction Project (0935-0249) AHRQ, 5600 Fishers Lane, Room #07W42, Rockville, MD 20857