Technology Assessment Program

Definition of Treatment-Resistant Depression in the Medicare Population

Final
Technology Assessment
Project ID: PSYT0816
Date: February 9, 2018

Technology Assessment Program–Technical Assessment

Project ID: PSYT0816

Definition of Treatment-Resistant Depression in the Medicare Population

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.:
HHSA290201500011I_HHSA29032006T

Publication Date:

Investigators:
Bradley N. Gaynes, M.D., M.P.H.
Gary Asher, M.D., M.P.H.
Gerald Gartlehner, M.D., M.P.H.
Valerie Hoffman, Ph.D.
Josh Green, B.A.
Erin Boland, M.P.H.
Linda Lux, M.P.A.
Rachel Palmieri Weber, Ph.D.
Charli Randolph, B.A.
Carla Bann, Ph.D.
Emmanuel Coker-Schwimmer, M.P.H.
Meera Viswanathan, Ph.D.
Kathleen N. Lohr, Ph.D., M.Phil., M.A.

Purpose of Review

To review the current definitions of treatment-resistant depression (TRD), to assess how closely current TRD treatment studies fit the most common definition, and to suggest how to improve TRD treatment research.

Key Messages

- TRD is commonly defined as a failure of treatment to produce response or remission for patients after two or more treatment attempts of adequate dose and duration, but no clear consensus exists about this definition.
- TRD definitions in treatment studies do not closely match the definition above; only 17 percent of studies do so.
- To improve TRD treatment research, experts should standardize the number of prior treatment failures and specify the adequacy of both dose and duration. In addition, they should identify the core outcome measures to be used in such research.

Preface

The Agency for Healthcare Research and Quality (AHRQ), through its Evidence-based Practice Centers (EPCs), sponsors the development of evidence reports and technology assessments to assist public- and private-sector organizations in their efforts to improve the quality of health care in the United States. The Centers for Medicare & Medicaid Services (CMS) requested this report from the EPC Program at the Agency for Healthcare Research and Quality (AHRQ). AHRQ assigned this report to the following EPC: RTI International - University of North Carolina at Chapel Hill (RTI-UNC) Evidence-based Practice Center (Contract Number: HHSA290201500011I_HHSA29032006T).

The reports and assessments provide organizations with comprehensive, evidence-based information on common medical conditions and new health care technologies and strategies. They also identify research gaps in the selected scientific area, identify methodological and scientific weaknesses, suggest research needs, and move the field forward through an unbiased, evidence-based assessment of the available literature. The EPCs systematically review the relevant scientific literature on topics assigned to them by AHRQ and conduct additional analyses when appropriate prior to developing their reports and assessments.

To bring the broadest range of experts into the development of evidence reports and health technology assessments, AHRQ encourages the EPCs to form partnerships and enter into collaborations with other medical and research organizations. The EPCs work with these partner organizations to ensure that the evidence reports and technology assessments they produce will become building blocks for health care quality improvement projects throughout the Nation. The reports undergo peer review and public comment prior to their release as a final report.

AHRQ expects that the EPC evidence reports and technology assessments, when appropriate, will inform individual health plans, providers, and purchasers as well as the health care system as a whole by providing important information to help improve health care quality.

If you have comments on this evidence report, they may be sent by mail to the Task Order Officer named below at: Agency for Healthcare Research and Quality, 5600 Fishers Lane, Rockville, MD 20857, or by email to epc@ahrq.hhs.gov

Gopal Khanna, M.B.A.
Director
Agency for Healthcare Research and Quality

Stephanie Chang, M.D., M.P.H.
Director
Evidence-based Practice Center Program
Center for Evidence and Practice Improvement
Agency for Healthcare Research and Quality

Elise Berliner, Ph.D.
Director, Technology Assessment Program
Center for Evidence and Practice Improvement
Agency for Healthcare Research and Quality

Arlene Bierman, M.D., M.S.
Director
Center for Evidence and Practice Improvement
Agency for Healthcare Research and Quality

Aysegul Gozu, M.D., M.P.H.
Task Order Officer
Center for Evidence and Practice Improvement
Agency for Healthcare Research and Quality

A draft of this technology assessment was distributed solely for the purpose of public and peer review. This final report does not represent and should not be construed to represent an AHRQ determination or policy.

This report is based on research conducted by the RTI-University of North Carolina at Chapel Hill (RTI-UNC) Evidence-based Practice Center under contract to the Agency for Healthcare Research and Quality (AHRQ), Rockville, MD (Contract No. HHSA290201500011I_HHSA29032006T). The findings and conclusions in this document are those of the author(s) who are responsible for its contents; the findings and conclusions do not necessarily represent the views of AHRQ and CMS. No statement in this article should be construed as an official position of AHRQ, CMS, or the U.S. Department of Health and Human Services.

None of the investigators has any affiliations or financial involvement related to the material presented in this report.

Persons using assistive technology may not be able to fully access information in this report. For assistance, contact epc@ahrq.hhs.gov.

Suggested Citation: Gaynes BN, Asher G, Gartlehner G, Hoffman V, Green J, Boland J, Lux L, Weber RP, Randolph C, Bann C, Coker-Schwimmer E, Viswanathan M, Lohr KN. 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.

Acknowledgments

The authors gratefully acknowledge the following individuals for their contributions to this project and deeply appreciate their considerable support, commitment, and contributions. From the Agency for Healthcare Research and Quality: Aysegul Gozu, M.D., M.P.H., our task order officer, and Elise Berliner, Ph.D., director of the Technology Assessment Program at the Center for Evidence and Practice Improvement. From the Center for Medicare & Medicaid Services' Coverage and Analysis Group: Susan M. Miller, M.D.; Rosemarie Hakim, Ph.D.; Linda Gousis, J.D.; and Lori Ashby, M.A., director of the Division of Medical and Surgical Services. From the RTI International–University of North Carolina Evidence-based Practice Center (EPC) staff: Carol Woodell, B.S.P.H., EPC project manager; Rachel Palmieri Weber, Ph.D., EPC project manager; Meera Viswanathan, Ph.D., EPC project director; Dan Jonas, M.D., M.P.H., EPC co-director; Lynn Whitener, Dr.P.H., M.S.L.S., librarian; Sharon Barrell, M.A., editor; and Loraine Monroe, publications specialist.

Peer Reviewers

Prior to publication of the final evidence report, EPCs sought input from independent Peer Reviewers without financial conflicts of interest. However, the conclusions and synthesis of the scientific literature presented in this report does not necessarily represent the views of individual reviewers.

Peer Reviewers must disclose any financial conflicts of interest greater than $10,000 and any other relevant business or professional conflicts of interest. Because of their unique clinical or content expertise, individuals with potential non-financial conflicts may be retained. The Task Order Officer and the EPC work to balance, manage, or mitigate any potential non-financial conflicts of interest identified.

The list of Peer Reviewers follows:

Allen Doederlein
Depression and Bipolar Support Alliance
Chicago, IL

Matthew V. Rudorfer, M.D.
National Institute of Mental Health
Division of Services and Intervention Research
Bethesda, MD

Pasqualina Santaguida, M.D., Ph.D.
McMaster University
Hamilton, Ontario

Definition of Treatment-Resistant Depression in the Medicare Population

Structured Abstract

Objectives. To inform future discussions and decisions about how to define treatment-resistant depression (TRD) and specify the important outcomes measured in research studies, and to clarify how trials or observational studies might best be designed and conducted to inform clinical practice and health policy.

Data sources. To provide a comprehensive understanding of how experts and investigators have defined and studied TRD, we first performed a *narrative* review of relevant literature. We considered consensus statements, practice guidelines, government materials, and other literature published from 1/1/1995 through 8/18/2017, except for systematic reviews (limited to start 1/1/2005). Next, we performed a *systematic* review of published studies of TRD interventions (1/1/2005 through 8/18/2017) indexed in MEDLINE®, EMBASE, PsycINFO, and Cochrane Library.

Review methods. Trained personnel dually reviewed all titles and abstracts for eligibility. Studies marked for possible inclusion by either reviewer and those with inadequate abstracts underwent dual full-text review. Disagreements were resolved by consensus discussion. One member of the research team abstracted data; a senior investigator reviewed abstractions for accuracy and completeness.

Results. Our narrative review indicated that no consensus definition existed for TRD. We identified four basic definitions for TRD (3 for major depressive disorder [MDD]; 1 for bipolar disorder). Based on frequency of reporting in the literature, the most common TRD definition for MDD required a minimum of two prior treatment failures and confirmation of prior adequate dose and duration. The most common TRD definition for bipolar disorder required one prior treatment failure. For all TRD definitions, no clear consensus emerged on defining adequacy of either dose or duration. Little agreement exists about the best approach to diagnose TRD or the preferred outcome measure, although the Hamilton Depression Rating Scale was the most used. We found general agreement about minimizing bias by using randomization; studies have not focused on minimizing placebo effects. Evidence about the risk factors (e.g., age, sex, number of prior failed treatments, and length of current depressive episode) associated with TRD and data to assess potential prognostic factors were limited.

Only 17 percent of intervention studies enrolled study populations that met frequently specified criteria for TRD. Most studies (88%) were randomized controlled trials; all studies applied some exclusion criteria to limit potential confounders. Depressive outcomes and clinical global impressions were commonly measured; functional impairment and quality-of-life tools were rarely used.

Conclusions. No agreed-upon definition of TRD exists; although experts may converge on two as the best number of prior treatment failures, they do not agree on definitions for adequacy of either dose or duration or outcomes measures. Critical to advancing TRD research are two key steps: (1) developing a consensus definition of TRD that addresses how best to specify the

number of prior treatment failures and the adequacy of dose and duration; and (2) identifying a core package of outcome measures that can be applied in a standardized manner. Our recommendations about stronger approaches to designing and conducting TRD research will foster better evidence to translate into clearer guidelines for treating patients with this serious condition.

Contents

List of Tables

List of Figures

List of Appendixes

Evidence Summary

Introduction

Patients with either major depressive disorder (MDD) or bipolar disorder can manifest depressive episodes. In 2015, 6.7 percent of adults in the United States (16.1 million people) experienced a depressive episode in the past year,[1] with the great majority being a part of MDD.[2] Treatment for MDD can be inadequate because either patients do not seek it or the care they receive is substandard.[3] Even for patients receiving adequate treatment, only 30 percent (3% of patients with MDD) reach the treatment goal of full recovery or remission. The remaining 70 percent of MDD patients will either respond without remission (about 20%) or not respond at all (50%).[4]

Patients whose depressive disorder does not respond satisfactorily to adequate treatment clearly have harder-to-treat depression,[5] generally referred to as treatment-resistant depression (TRD). TRD is a complex phenomenon influenced by variety in depressive subtypes, psychiatric comorbidity, and coexisting medical illnesses.[6] It poses a common, challenging presentation to psychiatric and primary care clinicians.[7]

Although TRD episodes are most commonly associated with MDD, they are also seen in the depressed phase of bipolar disorder. More than 30 percent of those suffering from bipolar disorder and receiving treatment do not experience sustained remission of depressive symptoms.[8]

TRD has substantial effects on patients, their families, communities, and society at large. TRD represents the highest direct and indirect medical costs among those with MDD.[9] Individuals with TRD are twice as likely to be hospitalized; the cost of this hospitalization is more than six times the mean total cost for depressed patients who are not treatment resistant.[10] TRD can nearly double both direct and indirect 2-year employer medical expenditures relative to expenditures for patients whose MDD responds to treatment.[11]

TRD is especially relevant for Medicare beneficiaries. Mood disorders (mainly MDD and bipolar disorder) are the second leading cause of disability in Medicare patients under the age of 65.[12] Depression in the elderly is associated with suicide more than at any other age;[13] adults 65 or older constitute 16 percent of all suicide deaths.[14] The decrease in average life expectancy for those with depressive illness, including Medicare beneficiaries, is 7 to 11 years.[15] Depression is a major predictor of the onset of stroke, diabetes, and heart disease;[16] it raises patients' risk of developing coronary heart disease[16] and the risk of dying from a heart attack nearly threefold.[17]

No universally accepted operational definition of TRD exists.[18] Definitional dilemmas limit the ability of systematic reviewers or other experts to synthesize information and generalize the TRD findings to the array of patient populations encountered in daily practice. Moreover, varying conceptualizations of TRD have made translating research findings or systematic reviews into clinical practice guidelines challenging and inconsistent. Indeed, guideline definitions of TRD differ, agreement on what constitutes prior treatment adequacy is lacking, and recommended "next step" interventions can diverge.[19-23]

This systematic review was proposed as a large Technology Assessment by the Centers for Medicare & Medicaid Services and conducted for the Agency for Healthcare Research and Quality (AHRQ). We reviewed definitional and other aspects of TRD in clinical research. One aim is to inform future discussions and decisions about how to define the condition and specify the important outcomes measured in research studies. A second aim is to clarify how researchers might best design and conduct trials or observational to guide clinical practice and health policy.

Methods

To provide a comprehensive and broad understanding of how various experts and investigators have defined and studied TRD, we examined 11 Key Questions (KQs) (Table A). The review was structured to address two categories of questions: a Narrative Review answering general questions, which was more qualitative and less structured than a systematic review and therefore used a more contextual narrative review approach, and a Systematic Review, which addressed specific study design questions.

Table A. Key Questions

Narrative Review: KQs 1–5 (general questions)	Systematic Review: KQs 6–11 (specific study design questions)
1. What definitions of TRD appear in these sources and do definitions converge on a best one?	6. What are the inclusion criteria for patients in these studies, specifically concerning patient characteristics, prior treatments, and diagnostic characteristics?
2. What methods do investigators use to diagnose this condition in clinical research, and does a consensus exist about the best ways to reach a clear diagnosis?	7. How do these criteria compare or contrast with definitions encountered in the narrative review?
3. What measures (i.e., endpoints or outcomes) exist to determine the success or failure of treatment in TRD studies; what clinical focus do they represent (e.g., severity); what psychometric and other properties do they have?	8. What were primary characteristics of included studies, such as design, run-in or wash-out periods, and length?
4. What research designs do investigators use in TRD studies and does any consensus exist about best approaches to minimize bias and placebo effects and other elements of study design (e.g., length)?	9. How were included studies designed to account for TRD risk factors identified in the narrative review?
5. What are the risk factors for TRD?	10. What are relationships between risk factors and results of included TRD studies?
	11. What variables or information did included studies report (e.g., patient outcomes, time to relapse, treatment adherence, attrition, and use of health care resources)?

KQ = Key Question; TRD = treatment-resistant depression.

Table B provides an abbreviated list of PICOTS (populations, interventions, comparators, outcomes, time frames, and settings); Table 1 in the main report documents detailed inclusion/exclusion criteria.

We set eligible dates of English-language publications to focus on TRD treatments approved by the U.S. Food and Drug Administration, yield a reasonably comprehensive evidence base, and reflect contemporary approaches to TRD. We searched the published literature from 1/1/1995 through 8/18/2017. For the narrative KQs, we sought literature published since 1/1/95, except for systematic reviews, for which the start date was 1/1/2005. For the systematic KQs we set the publication date for intervention studies as 1/1/2005.

We followed standard procedures for systematic literature searches specified in the AHRQ *Methods Guide for Effectiveness and Comparative Effectiveness Reviews.*[24] Although KQs 1 through 5 were narrative questions (which often turn up for systematic reviews), we nevertheless applied standard methods for the title/abstract and full-text review as much as possible to meet the typical global standards for this step. We searched for publications indexed in MEDLINE®, EMBASE, PsycINFO, and Cochrane Library; studies had to use TRD definitions with depressive diagnoses consistent with the *Diagnostic and Statistical Manual of Mental Disorders,* Fourth edition[25] or 5th edition.[26] Other materials included consensus statements, clinical practice guidelines, and relevant government reports; website sources were Clinicaltrials.gov, Guideline.gov (AHRQ's National Guidelines Clearinghouse), HSRProj (Health Services Research Projects in Progress database), and UpToDate®. To address the different types of questions for the two reviews (Table A), we sorted searches as follows: KQs 1 through 5 considered all sources noted above (including systematic reviews) except for individual intervention trials; only intervention trials were eligible for KQs 6 through 11.

We dually reviewed all titles and abstracts for eligibility. Studies marked for possible inclusion underwent full-text review, also done as dual reviews. Disagreements were resolved by consensus discussion.

One member of the research team abstracted data; another (senior) investigator reviewed the abstraction for accuracy and completeness. For KQ 10 involving additional (regression) analyses (explained in the main report), we assessed risk of bias with appropriate tools: for randomized controlled trials (RCTs), the Cochrane Risk of Bias tool;[27] for nonrandomized trials and observational studies, the Newcastle-Ottawa Scale.[28]

The draft Technology Assessment was peer-reviewed and posted for public comment.

Table B. Inclusion criteria (abbreviated) for literature searches

PICOTS	Inclusion Criteria
Population	All adult populations (≥18 years old) identified as having a primary diagnosis of depression (including MDD and bipolar disorder) who have had a depressive episode and have not responded to treatment(s) (the least stringent definition of TRD). The depressive episode must be part of an MDD or a bipolar disorder.
Interventions	Any pharmacologic intervention tested as a treatment for TRD as a primary therapy or as an augmentation agent to an existing primary therapy. Any nonpharmacologic device or procedure tested as a treatment for TRD as a primary therapy or as augmentation to an existing primary therapy. Psychotherapy (i.e., cognitive behavioral therapy, third wave cognitive behavioral therapy, psychodynamic therapies, and integrative therapies) Complementary and alternative medicine (CAM) interventions and formal exercise programs.
Comparators	All those above in studies with concurrent control groups or control groups from an interrupted time series or pre/post studies with interrupted time series
Outcomes	Benefits that are reported as primary endpoints (or outcomes) Reduction in suicidal ideation or suicide attempts Quality of life Response to treatment Remission Change in depressive severity Functional capacity (physical and cognitive functioning measured by validated scales) Speed of remission Speed of response Intervention durability (rates or counts of recurrence of a depressive episode for those who have remitted) Adverse events from the intervention identified as either critical or important for decisionmaking Serious adverse events per Food and Drug Administration definition (rates or counts) Overall adverse events (rates or counts) Treatment discontinuations attributed to adverse events (rates or counts)
Timing	Any study duration
Setting	Studies in very highly developed countries[29]
Study Designs	For KQs 1–5: Consensus statements, guidelines, or other materials, and systematic reviews For KQs 6–11: Randomized, or prospective nonrandomized, or observational studies (including concurrent controls and interrupted time series)

Results

We included 235 articles: 38 for KQs 1 through 5 and 197 for KQs 6 through 11. For KQs 6 through 11 articles, we identified 163 unique studies: 142 RCTs (88%), 4 nonrandomized trials (2%), 13 observational studies (8%), 2 case controls (1%), and 1 interrupted time series (<1%). The results reported below focus on only the "key points" drawn from our syntheses; the findings are reported in detail in two chapters of the main report.

Narrative Review

Key Question 1: Definitions of Treatment-Resistant Depression

- We identified four categories of TRD definitions, distinguished primarily by number of previous failed antidepressant treatment attempts (Table C). Identified this way, individuals either did or did not have TRD.
- We identified five TRD staging models (which delineate the degree of resistance along a spectrum), but only limited research addressed reliability and validity. These models appeared to be equally valid for documenting treatment failure in depressed patients, but their applicability and feasibility in clinical practice are unclear (Table D). Identified this way, people could have TRD along a spectrum of severity.
- No consensus exists on the best TRD definition. However, the majority of systematic reviews and guidelines or consensus statements reported that the commonly used definitions were based on treatment of patients whose depression failed to respond (a decrease in depressive severity of at least half) or did not go into remission (complete recovery as measured by a score on a depressive severity instrument below a threshold) following two or more treatment attempts of an adequate dose and duration.
- Experts do not agree on how to define adequate dose and adequate duration, although the minimum duration cited is typically 4 weeks.

Table C. Four categories of definitions of treatment-resistant depression by number of treatment failures and components of defintion

Number of Treatment Failures	Type of Publication on TRD Treatments, Date	Defines Nonresponse or Lack of Remission	Specifies Current Episode?	Defines Adequate Dose?	Defines Adequate Duration (weeks)?
1 or more	Seminal article on defining TRD, 1996[30]	+	--	+	≥6
	SR - pharmacologic, 2007[31]	+/- -	+/- -	+/- -	≥4 to ≥8
	SR - lamotrigine augmentation, 2010[32]	--	--	--	4
	SR - psychotherapy, 2011[33]	-	--	+/- -	≥6
	Review defining TRD, 2014[18]	-	--	+	6 to 8
	SR - rTMS, 2015[34]	-	--	--	--
	SR - rTMS, 2015[35]	-	+	--	4 to 6
	SR - predictors of nonresponse, 2016[36]	-	--	--	--
2 or more	Seminal article on definition of TRD, 2001[37]	+	+	+	≥4
	SR - pharmacologic treatments, 2007[31]	++/-	+/- -	++/-	4 to 6
	ICER coverage policy analysis, 2012[38]	+	+	-	--
	SR - lithium or atypical antipsychotics, 2013[39]	-	+/- -	--	≥4
	SR - rTMS, 2014[40]	+	+/- -	+	≥4
	SR - nonpharmacological, 2014[41]	+/- -	+/- -	+	≥4
	Australian/New Zealand Clinical Practice Guideline, 2015[42]	-	--	+	≥4 to 8
	SR - pharmacologic and somatic, 2016[43]	--	+/- -	--	--
	VA/DoD Clinical Practice Guideline, 2016[23]	-	--	+	≥4 to 6
3 or more	ICSI Adult Depression in Primary Care Guideline, 2016[21]	+	--	--	--
For bipolar TRD: 1 or more	SR – nonpharmacological, 2014[41]	+	--	--	10 to 12
	Australian/New Zealand Clinical Practice Guideline, 2015[42]	--	--	+	≥3

Legend: + = definition was provided; - = definition was not provided; ++/- = more studies in review provided definition; +/- - = more studies in the review did not provide definition. ICER = Institute for Clinical and Economic Review; ICSI = Institute for Clinical Systems Improvement; MDD = major depressive disorder; SR = systematic review; TRD = treatment-resistant depression; VA/DoD = Veterans Administration/Department of Defense.

Table D. Staging models for treatment-resistant depression to define the spectrum of illness

Models Authors and Year of Publication	How is Severity Scored?	Is Failure Defined?	Specify Current Episode?	Define Adequate Dose?	Define Adequate Duration?	Staging Schema	Predictive Validity and Reliability Tested? Other Comments
Antidepressant Treatment History Form[44]	Sum score based on points per treatment	-	+	+	≥4 weeks	5 stages (0 to 5)	Predictive validity confirmed in 3 prospective ECT studies; reliability good in two studies. Does not correspond with number of treatment failures; does not count psychotherapy in failed trials
Thase and Rush Staging Model (TRSM)[5, 31, 44, 45]	Stages, with higher-numbered stages indicating a greater degree of treatment resistance	+	-	-	≥4 weeks	5 stages (1 to 5)	Predictive value has not been systematically assessed; reliability has not been tested. Stage II corresponds with 2 treatment failures; considers number of classes of ADs that have failed to provide a response but not psychotherapy
European Staging Model[44, 46]	Number of weeks with treatment resistance	+	+	-	Varies by nonre-sponder, TRD, and CRD	3 categories: nonresponder, TRD, and CRD	Predictive value has not been systematically assessed; reliability has not been tested All TRD stages are consistent with 2 treatment failures
Massachusetts General Hospital Staging model (MGH-s)[44, 45]	Points based on number of prior failures	+	-	+	≥6 weeks	3 stages based on number of AD failures and 3 points for ECT failure	Retrospective chart review showed an association between higher MGH-s score and worse outcome; a retrospective study showed the MGH-s model better predicted nonremission than the TRSM. Reliability for these models was not tested. No direct correspondence with ≥2 treatment failures.
Maudsley Staging Model[44, 46]	Points per number of prior attempts, duration, symptoms severity, augmentation use, ECT	+	+	+	Varies by intervention	Points based on duration, symptom severity, number of treatment failures, augmentation, ECT	Only tool with prospective testing showing good validity; 2 studies showed the MSM score predicted future nonresponse significantly better than the TRSM. Reliability not tested. No direct correspondence with ≥2 treatment failures

Legend: + = definition was provided; - = definition was not provided; ++/- = more studies in review provided definition; +/- - = more studies in the review did not provide definition. AD = antidepressant; CRD = chronic resistant depression; ECT = electroconvulsive therapy**;** RCT = randomized controlled trial; rTMS = repetitive transcranial magnetic stimulation; TRD = treatment-resistant depression.

Key Question 2: Preferred Diagnostic Tools

- No consensus exists about the preferred approach for diagnosing TRD.
  - The literature emphasizes that a diagnosis of a major depressive episode as a part of MDD or bipolar disorder can be made through a standard clinical evaluation (based on the Diagnostic and Statistical Manual) of Mental Disorders, International Classification of Diseases, or Research Diagnostic Criteria) or a through a more structured clinical assessment (Mini International Neuropsychiatric Interview, Structured Clinical Interview for the Diagnostic and Statistical Manual, or Schedule for Affective Disorders and Schizophrenia), but no preferred approach exists.
  - Diagnosing TRD further entails collecting a careful history before treatments (e.g., the number of prior pharmacologic attempts of adequate dose and duration that did not produce remission) or administering a structured, staging tool (Antidepressant Treatment Response Questionnaire, Thase Rush Staging Model, Massachusetts General Hospital Staging Model, or the Maudsley Staging Model) to confirm treatment resistance, but no preferred approach exists, the evidence base is limited, and careful history has not been compared directly with a structured tool.
- The medical setting has no influence on choice of diagnostic tool, although some issues of feasibility arise.

Key Question 3: Preferred Outcome Measures to Determine Success or Failure

- No consensus exists about the best outcome measure to use for TRD.
- The three main categories of outcome measures—depression-specific measures, general psychiatric status measures, and functional scales—have both patient-reported and clinician-administered versions available.
  - The most common depression-specific measure is the Hamilton Depression Rating Scale (HAM-D). Remission (complete recovery as measured by a score below a threshold) is the preferred endpoint regardless of tool.
  - General psychiatric status measures were infrequently described; most commonly reported was the Clinical Global Impression scale.
  - Various functional scales have been reported, but no one scale is the most frequently used.
- Most measures have adequate psychometric properties (e.g., reliability and validity) for measuring depressive outcomes.
- The minimum clinically important difference has been variably defined for many of these outcome measures, but none is a consensus preference.

Key Question 4: Preferred Study Designs for Treatment-Resistant Depression Treatment Intervention Studies

- Most investigators and expert groups preferred randomized designs over nonexperimental ones as a means of minimizing bias.
- Most available literature did not address, or apparently achieve consensus about, designs that might minimize placebo effects.
- No consensus exists about the appropriate or necessary length of trials or other studies of TRD. A study length of “at least 6 weeks” was often recommended.

- Studies also recommended using whole structured clinical interviews to diagnosis depression, because these full assessments could better confirm the MDD (or bipolar) diagnosis and clarify psychiatric comorbidity, seen as a key potential confounder in TRD treatment trials.
- Getting patients to an adequate dose of a given medication may take a few weeks; for that reason, 6 weeks of adequate dosing may require a trial length longer than 6 weeks.
- Compliance and consideration of prior psychotherapy use are important to assess and control for in analyses.

Key Question 5: Risk Factors for Treatment-Resistant Depression

- Evidence about what risk factors are associated with a TRD diagnosis is quite limited.
- Several components of depression (disease severity, duration of current episode, number of previous hospitalizations, and number of failed antidepressant trials) appeared to be associated with increased risk of TRD.
- The sociodemographic variables of age (older) and marital status (divorced or widowed) increased the risk of TRD.
- Coexisting anxiety symptoms, anxiety disorders, and personality disorders were associated with TRD.
- Some other clinical characteristics (such as having melancholic features, suicidality) were associated with greater risk of TRD.

Systematic Review

We divided interventions into four categories:

- Brain stimulation treatments (BST): electroconvulsive therapy, repetitive transcranial magnetic stimulation, vagal nerve stimulation, and deep brain stimulation (73 studies);
- Pharmacotherapy (71 studies);
- Psychotherapy (11 studies); and
- Complementary and alternative medicine (CAM) therapies and exercise (8 studies).

Key Question 6: Inclusion Criteria for Intervention Studies

- Confirmation of TRD for study entry was often poorly described.
- The HAM-D and the Montgomery–Åsberg Depression Rating Scale were commonly used to set minimum depressive severity thresholds for study entry; most studies involved patients with moderately severe depression.
- Studies were inconsistent about the necessary duration of prior treatment attempts for study entry.
- Most studies required at least one, and often two, prior failed treatment attempts of adequate therapy.
- Several patient characteristics were rarely considered for study entry: duration of depressive symptoms, prior depressive relapses, prior treatment intolerance, prior augmentation or combination therapy, prior psychotherapy, and suicidality.

Key Question 7: Inclusion Criteria Compared with Definitions of Treatment-Resistant Depression

- Inclusion criteria as specified by the eligible TRD studies did not closely align with the definition(s) of TRD identified in the narrative review.
- Although the most common definition of TRD from our narrative review involved a minimum of two failed prior adequate antidepressant studies, the most common definition in included studies was a minimum of one failed trial (49%) (only 38% required a minimum of two failed trials).
- Of all 163 studies, 77 percent considered in their selection criteria whether the patient had been treated previously with an adequate dose; 58 percent systematically confirmed that the dose was adequate by specifying dosage levels through interview, questionnaire, or other formal clarifications.
- Of all 163 studies, 82 percent considered in their selection criteria whether prior treatments were of adequate duration; 72 percent systematically confirmed that the duration was adequate ($\geq$ 4 weeks of treatment).
- Thirty-three percent of the studies set inclusion criteria based on stage of TRD using a staging model.
- Seventeen percent of studies had *all* the most commonly described criteria for TRD: a minimum of two prior treatment failures, confirmation that a dose was adequate, and confirmation that duration was 4 weeks or longer.

Key Question 8: Characteristics (Design) of Included Studies

- Most studies had RCT designs (88%); this rate did not differ by intervention type.
- Few studies had run-in periods (21%) or wash-out periods (23%).
- Study duration varied across studies, ranging from less than 2 weeks to more than 4 years; the majority of BST studies (the most common intervention type) lasted $\leq$2 months (62%).

Key Question 9: Controlling for Potential Confounders

- A considerable majority (88%) used randomization as a means to control for potential confounders.
- All studies applied some exclusion criteria to limit potential confounders. Number of prior failed treatments, psychiatric and medical comorbidities, and severity of disease were the most commonly applied restriction factors to achieve homogeneous study populations.
- Several studies (20%) stratified analyses by potential confounders. Generally, these factors were age, sex or gender, number of prior failed treatments, and duration of current depressive episode.
- Of 20 nonrandomized studies, only eight reported statistical techniques to control for potential confounding.

Key Question 10: Addressing Risk Factors and Their Relationship to Outcomes

- For most risk factors that might influence treatment response, data were either insufficient for regression analyses or reflected no statistically significant impact on study results.
- In a comparison of pharmacotherapy with pharmacotherapy plus augmentation with a second medication, multivariable analyses indicated that female sex had a significant effect on discontinuation; studies with 60 percent or more female participants had statistically significantly higher discontinuation rates because of adverse events (ratio of odds ratios = 2.81; 95% confidence interval, 1.04 to 7.59) than studies with fewer than 60 percent females.
- A smaller placebo response was associated with a statistically significantly larger treatment effect for response (p=0.027), remission (p=0.001), and discontinuation because of adverse events (p=0.010). Study duration did not have an impact on placebo response.

Key Question 11: Key Study Outcomes

- The two most common outcome measures used to assess depression were the HAM-D and the Montgomery–Åsberg Depression Rating Scale.
- Assessment of manic outcomes was rare.
- The Clinical Global Impression scale was the most common general psychiatric outcome reported, nearly always in pharmacology studies and less than half the time in BST studies.
- Functional impairment and quality-of-life outcomes were infrequently reported.
- Other than in psychotherapy studies, adherence to treatment was not commonly reported.
- Overall attrition was a frequently reported outcome, but specific reasons for attrition (e.g., adverse events or lack or efficacy) were less often described.
- Disability status, time to relapse, and use of health care services were very rarely reported.

Discussion

Our narrative review indicated that no consensus definition existed for TRD. We identified four basic definitions for TRD (three for MDD, one for bipolar disorder). The most common TRD definition for MDD requires a minimum of two prior treatment failures and confirmation of prior adequate dose and duration. No clear consensus emerged on how to define adequacy of either dose or duration. We identified little consensus about the best tools to diagnose TRD or measure its outcome. We saw some agreement on the benefit of minimizing bias by randomization; most of the literature did not address, or apparently achieve consensus about, designs that might minimize placebo effects. Evidence identifying risk factors for TRD and data to assess potential prognostic factors were limited.

Our systematic review indicated that inclusion criteria as specified by the eligible TRD trials or observational studies generally did not closely align with TRD definitions from the narrative review. Only 17 percent of studies reported “two prior treatment failures and confirmation of prior adequate dose and duration.” Most studies (88%) were RCTs, and all applied some exclusion criteria to limit potential confounders. Depressive outcomes were the frequently

reported endpoints; clinical global impressions were also often assessed. Functional impairment and quality of life tools were infrequently used.

Findings in Relationship to What Is Already Known

The variability in the definitions and conceptualization of TRD (from our narrative review) is consistent with other reports from the past decade identifying the lack of any standard, systematic definition of TRD.[18, 31, 44, 47] Indeed, defining the adequacy of dose and duration, clarifying failure (as remission or response and after what length of time), and determining whether TRD requires the prior use of different classes of antidepressants are all variably defined and implemented. Taken together, the available literature highlights the resulting difficulty in synthesizing information across trials or other types of studies or documents. This characteristic of the evidence base also underscores the problems of translating research findings into guidelines for selecting better treatment options for patients with TRD.

Our systematic review highlighted some key findings not previously described. The mismatch between the most common number of treatment failures (at least two) and what most recent literature has assessed (at least one failure) was stark. Also, the failure of inclusion criteria in recent TRD studies to confirm systematically both adequate dose (58%) and duration (72%) has not previously been described, nor has the finding that only 17 percent of recent intervention studies are consistent with the most common definition of TRD. These results highlight another concern about how to compare and synthesize data across treatment studies.

Finally, despite the substantial morbidity associated with TRD, the relative infrequency of use of patient-oriented outcomes such as functional impairment and quality-of-life measures in considering the benefits of TRD treatment was newly demonstrated. So too was the infrequent measurement of both adherence to treatment or health care services use.

Implications for Clinical and Policy Decisionmaking

This current state of evidence underscores the challenges facing clinicians. Effective treatments exist, but because of the variability in TRD definitions and study populations, determining to which patients the results apply is difficult. Similarly, the state of the evidence poses challenges for policymakers. Officials, at both Centers for Medicare & Medicaid Services and other public agencies or private-sector organizations, must be confident that two main assumptions are being met. The first is that the population of patients with TRD is being consistently and systematically defined; the second is that meaningful and comparable outcomes of importance to both patients and clinicians are being monitored. Neither is consistently reported in the literature, limiting translation of this treatment information into actual care.

The high level of morbidity associated with TRD is clear. For adequate clinical and policy decisionmaking about TRD patients, however, a widely agreed-upon definition of the condition that addresses how to best determine the number of prior treatment failures and the adequacy of dose and duration is critical. Some means of systematically monitoring this for TRD on a large scale (e.g., a treatment registry using common data elements in an electronic medical record) could substantially help clarify which criteria best define TRD, what the course of illness is, and how interventions might affect that course.

Limitations of this Technology Assessment

Comparative Effectiveness Review Process

The primary challenge of this process was the broad, comprehensive, and inclusive nature of topic, which combined a narrative review (for five KQs) and a systematic one (for six KQs). Given how variable the definitions of TRD are in the literature, we needed to cast a wide net for both published and gray literature to assemble the proper universe of sources that could be managed within a reasonable amount of time and resources. We addressed this challenge by focusing the 11 KQs and the time periods for the literature searches to reflect current conceptualizations of TRD.

The Evidence Base

The primary limitation of this evidence base is the heterogeneity of TRD definitions encountered in both reviews (KQs 1 through 5 and KQs 6 through 11). With no agreed-upon definition of TRD and no consensus on very important outcomes, determining to what population clinical trials results apply is difficult. This heterogeneity will prevent others from synthesizing or combining data, even for the more common TRD interventions such as brain stimulation technologies or medications, to translate findings into clinical practice recommendations.

Furthermore, data were insufficient to reliably assess prognostic factors that can predict outcomes of TRD treatment.

Research Recommendations

We propose several steps to address existing evidence gaps and substantially improve the study and treatment of patients with TRD.

Reducing the heterogeneity of how TRD patient populations are defined is a necessary first step. Perhaps the most critical task is to reach agreement on a standardized, systematic, and feasible definition of TRD. Such a definition should clearly specify the number of prior treatment attempts, what an adequate dose is, and what an adequate duration is. At the very least, the minimum number of past failed therapy attempts should be two. Systematic confirmation of adequacy of prior treatment attempts is a necessary part of this "definitional" step.

Systematic, standardized accounting for potential confounders is also crucial. The factors that must be considered include the following: depressive severity, duration of current episode, prior treatment intolerance, prior augmentation or combination therapy, prior psychotherapy, and psychiatric comorbidities. Other socio-cultural factors that may influence the course of TRD (e.g., gender and age) may also be relevant and deserve further study, as may genetic factors.[42] Randomization can account for some measured and unmeasured confounders in larger trials, but the smaller RCTs that we identified had imbalances in baseline characteristics and rarely adjusted for such differences. Moreover, nonrandomized TRD studies adjusted for potential confounders less than half of the time

Agreement on a package of outcome measures to be administered in a standard way should be strongly encouraged. The field would benefit from an evidence-informed, multi-stakeholder consensus process to develop a core outcomes set for TRD, potentially something similar to the Outcome Measures in Rheumatology (OMERACT) process in rheumatology (https://www.omeract.org/). Of particular importance is including one measure of depressive severity, one measure of general psychiatric status, one measure of functional impairment or

quality of life, and one measure of adherence to medications or other interventions. Another key outcome to consider includes a measure of suicidality (to assess for reduction in suicidality). Common use of measures will allow for better comparisons among trials; it should improve our ability to combine studies for meta-analyses. Patient-reported instruments may be preferred because they are more feasible, generally speaking, and more patient centered than clinician-reported instruments.

Researchers and clinicians should come to consensus on a standard length of treatment. The key is to provide enough time for patients to receive an adequate dose and duration of the intervention. Given the chronicity of TRD and the time to reach an adequate dose and length of treatment, at least 2 months is the bare minimum for studies to be conducted. Also, the increasing risk of relapse with greater levels of resistance[7] argues for the need to demonstrate longer term efficacy for the more resistant patients, suggesting a need for even longer trials for the more severely resistant. For this latter group, who may have a more chronic and persistent course, different treatments might be required than for those whose depressive episodes are more episodic.

Whether either run-in stages or wash-out periods affect the efficacy or effectiveness of TRD treatments remains unclear. Comparative trials should examine this issue to clarify whether investigators should use one or the other in designing their trials.

We found only a very few studies of interventions other than pharmacological or BST interventions (that is, psychotherapies and CAM or exercise as remedies for TRD). This gap reduced the evidence base relevant for patients who prefer to avoid, or for whom it would be inappropriate to try, pharmacological agents or more invasive procedures. Consideration of less-studied interventions could help inform patient decisions about options and improve the level of shared or informed decisionmaking.

Trials or robust types of observational studies to test the *effectiveness* of all such interventions in real-world settings are necessary. Targeting only efficacy (via RCTs) may produce information for clinicians, patients, or policymakers that cannot easily be applied in "ordinary," every-day circumstances.

To allow for better assessment of quality in TRD, publications of RCTs need to adhere to Consolidated Standards of Reporting Trials specifications for reporting.[48] Similarly, publications of nonrandomized controlled trials or observational studies should adhere to Strengthening the Reporting of Observational Studies in Epidemiology (STROBE).[49] Documenting all steps in such investigations, reporting on all planned outcomes, and otherwise ensuring complete transparency for this work are critical actions in adding to the professional literature. These steps would help ensure a consistent definition of TRD and its reported outcomes.

Finally, TRD needs to be monitored, consistently, systematically, and on a large scale. For instance, a treatment registry using common data elements could substantially help clarify the criteria that best define TRD, what the course of illness is, and how interventions might affect that course. Coordination between different specific treatment registries that already exist (e.g., the vagal nerve stimulation registry required by the U.S. Food and Drug Administration,[50, 51] and the transcranial magnetic stimulation registry recently launched by Neurostar[52]) and that have been suggested (e.g., a ketamine registry[53]) would be a necessary step. Data quality would be a key challenge for such an enterprise.

Conclusions

We encountered substantial diversity at every stage of research on TRD interventions. Of particular concern was the lack of consensus about various elements of even a TRD diagnosis and appropriate inclusion or exclusion criteria. Additionally, little or no agreement about important outcomes and how to assess them hampered analysis. An extensive set of recommendations about additional and more robust approaches to the design and conduct of this research will foster better evidence to translate into clearer guidelines for treating patients with this serious condition.

Background and Objectives

Clinical and Epidemiological Issues

Depressive episodes can be seen in patients with either major depressive disorder (MDD) or bipolar disorder. In 2015, 6.6 percent of adults in the United States experienced a depressive episode in the past year.[1] The bulk of these episodes are part of MDD, experienced by more than 13 million U.S. residents each year.[2] Of these individuals, one-half seek help for this condition; one in five of those seeking help receive adequate acute-phase treatment.[3] Even for patients receiving adequate treatment, only 30 percent (i.e., 3% of patients with MDD) reach the treatment goal of full recovery or remission.[4]

The remaining 70 percent of MDD patients will either respond without remission (about 20%) or not respond at all (50%).[4] Patients whose depressive disorder does not respond satisfactorily to adequate treatment clearly have harder-to-treat depression,[5] which is generally (albeit not uniformly) referred to as treatment-resistant depression (TRD). Although often broadly defined this way, TRD is a complex phenomenon that is influenced by heterogeneity in depressive subtypes, psychiatric comorbidity, and coexisting medical illnesses.[6] It poses a common, challenging presentation to psychiatric and primary care clinicians,[7] one in which decision-making is guided by both available data and patient preference.

Although TRD is most commonly associated with MDD, treatment-resistant depressive episodes are also seen in the depressed phase of bipolar disorder. Bipolar disorder affects 2.6 percent of the U.S. adult population each year.[54] Much like MDD, bipolar depression can be treatment resistant, and TRD can sometimes result from a failure to properly diagnose a patient presenting with depression as in fact suffering from bipolar disorder. More than 30 percent of those suffering from bipolar disorder and receiving treatment do not experience sustained remission of depressive symptoms.[8] Even among those who do achieve recovery for lengthy periods, depressive relapses are common; more than 20 percent of individuals with successfully treated bipolar depression will experience a depressive relapse within a year.[8]

TRD has substantial effects on patients and major impacts on families, communities, and society at large, most of which have been described for MDD patients. TRD represents the highest direct and indirect medical costs among those with MDD.[9] These costs increase with the severity of TRD.[55] Individuals with TRD are twice as likely to be hospitalized; the cost of this hospitalization is more than six times the mean total cost for depressed patients who are not treatment resistant.[10] TRD can nearly double both direct and indirect 2-year employer medical expenditures relative to expenditures for patients whose MDD responds to treatment ($35,500 for those with TRD and $18,600 for those with MDD).[11]

TRD is especially relevant for Medicare beneficiaries, for whom unsuccessfully treated depression has harmful sequelae. Mood disorders, which consist primarily of MDD and bipolar disorder, are the second leading cause of disability in Medicare patients under the age of 65.[12] Furthermore, depression in the elderly is more associated with suicide than at any other age;[13] although adults 65 or older make up 12 percent of the population, they constitute 16 percent of all suicide deaths.[14] Indeed, the decrease in average life expectancy for those with depressive illness, including Medicare beneficiaries, is 7 to 11 years, similar to that in elderly smokers.[15]

Finally, depression is a major predictor of the onset of stroke, diabetes, and heart disease.[16] Being depressed increases patients' risk of developing coronary heart disease,[16] and it raises the risk of dying from a heart attack nearly threefold.[17]

Rationale for Review

No universally accepted operational definition exists for TRD. Criteria for TRD have been variably defined in clinical research and practice,[18] reflecting numerous difficulties and controversies about its definition. These definitional dilemmas limit the ability of systematic reviewers or other experts to synthesize information and generalize the findings of many TRD studies to the array of patient populations encountered in daily practice. In addition, they restrict the application of clinical research findings to clinical practice, including community populations of TRD patients.

Moreover, these varying conceptualizations of TRD have made translating research findings or systematic reviews into clinical practice guidelines challenging and inconsistent. Treatment guidelines reflect this variability: their definitions of TRD differ, agreement on what constitutes prior treatment adequacy is lacking, and recommended "next step" interventions can diverge.[19-21, 23, 56]

The purpose of this report is to examine comprehensively the study design issues affecting both outcomes and bias in studies of TRD; we are not in this review determining outcomes associated with specific treatments of TRD. Our specific aims are two-fold: (1) to inform future discussions and decisions about how to define TRD and the important outcomes measured in research studies, and (2) to clarify how trials or observational studies might best be designed and conducted to guide clinical practice and health policy.

Key Questions

Narrative Review Questions

We first performed a narrative review of relevant literature to address Key Questions (KQs) 1 through 5. The narrative review questions were more qualitative and less structured than a systematic review; therefore, we used a more contextual narrative review approach. Our work is based on searches of consensus statements, guidelines, materials from the U.S. Food and Drug Administration, the U.S. National Institutes of Health (including the National Institute of Mental Health), and the U.S. Substance Abuse and Mental Health Services Administration; systematic reviews; and a review of UpToDate®, an evidence-based, peer-reviewed clinical information source. In addition, we used information from the Medicare Evidence Development & Coverage Advisory Committee panel meeting on April 27, 2016,[57] to augment our reporting on TRD definitions, study design issues, and the related topics.

The specific narrative KQs are:

1. What definitions of TRD are found in this literature? What consensus, if any, exists about the best definition(s) for this condition?
2. What methods do investigators use to diagnose this condition in clinical research? What consensus, if any, exists about the best measure(s) to use? Does the setting of the medical visit influence the choices that investigators make about the diagnostic tool they use?
3. What measures have been developed to determine the success and failure of treatment in clinical research studies of TRD?
    a. What consensus, if any, exists about the best measure(s) to investigate treatments for TRD? What are the main points of agreement about such measures?
    b. Are these measures physician reported or patient reported?

c. What are the psychometric properties of these measures? Is the minimum significant clinical difference defined for these measures?
d. Compare and contrast these measures in how they describe:
    i. Change in depression scores as measured by depression scales
    ii. Change in depressive symptomatology (e.g., sleep disorders, fatigue, weight change, cognition)
    iii. Change in measures of anhedonia
    iv. Change in measures of functional capacity (e.g., physical functioning, ability to care for self)
    v. Change in measures of quality of life
    vi. Change in measures of suicide ideation
    vii. Change in suicide attempts
    viii. Other

4. What types of research designs are used to study TRD?
    a. What consensus, if any, exists about the type of study design that best minimizes bias and the placebo effect in this field?
    b. If no consensus exists about study designs to accomplish these goals, what are the trends in study designs for assessing interventions for TRD? Do these trends reflect long-lasting (e.g., traditional) designs or short-lived, evolving, or newly emerging designs?
    c. What consensus, if any, exists about the appropriate length of a trial?
5. What are the risk factors for TRD?

Systematic Review Questions

From a systematic literature search for individual studies on TRD, we address KQs 6 through 11 with their subquestions as listed below.

6. What are the inclusion criteria for patients in these studies? Specify at least the factors listed below.
    a. Patient characteristics:
        i. Age
        ii. Type of depressive episode (unipolar, bipolar, psychotic, atypical, other)
        iii. Number of depression relapses and time to relapse
        iv. Psychiatric comorbidities
        v. Medical comorbidities (e.g., diabetes, cardiac disease, renal disease, dementia, and other cognitive abnormalities)
        vi. Suicidal ideation
        vii. Suicide attempts
        viii. Duration of symptoms
        ix. Screening tools used to make the diagnosis
        x. Diagnostic tools to confirm the diagnosis
    b. Prior treatments:
        i. The number, duration, dosage, or classes of antidepressants attempted for each trial of therapy
        ii. The number of failed trials of adequate therapy
        iii. The number of prior treatment trials that patients did not tolerate

iv. The use of augmentation and combination pharmacological therapies for each attempted treatment trial
v. The use of electroconvulsive therapy
vi. The use of psychotherapy

c. Diagnostic characteristics
i. The use of structured versus unstructured diagnostic assessments
ii. Scores on standardized and validated depression rating instruments
iii. Setting in which the diagnosis was made (i.e., primary care, generalized psychiatric setting, specialty psychiatric setting, other)

7. How do these inclusion criteria compare or contrast with the definition(s) of TRD noted in the narrative questions?
8. What were primary characteristics of included studies?
   a. What was the main design of each included study (e.g., randomized controlled trial with blinding; interrupted time series; use of placebo, wait-list, or sham procedure)?
   b. Were run-in or wash-out periods (or both) used in included studies? If so, how long were they?
   c. How long was each included study?
9. How were included studies designed to account for the risk factors for TRD (see (Narrative Review KQ 5)? If the following characteristics are not noted above as risk factors, how did included studies account for at least the following: age, sex, race, socioeconomic status, duration of symptoms, disease severity, coexisting medical and psychiatric conditions, and placebo effect?
10. What are relationships between risk factors and various results of included studies?
    a. Using regression analysis or other statistical techniques, determine whether the risk factors for Narrative Review KQ 5 and Systematic Review KQ 9 can be correlated with study results (i.e., the magnitude of treatment effects)?
    b. What is the influence of placebo response on the magnitude of treatment effects for different types of interventions?
    c. Does study duration moderate the influence of placebo response?
11. What variables or information did included studies report? Specifically:
    a. What measures are used to define end points in these TRD trials?
    b. In addition to the measures noted for Narrative Review KQ 3, did these studies record:
       i. Adherence to treatment
       ii. Attrition from care
       iii. Changes in patient-selected factors of importance (i.e., outcome measures identified by patient as important)
       iv. Changes in employment or disability status
       v. Changes in use of medical resources (e.g., hospitalizations, emergency room or physician visits)
       vi. Time to relapse

Organization of This Report

Apart from this introduction, this Technology Assessment report has the following chapters. Next is the Methods chapter, which covers all the steps used to address KQs 1 through 11. Following that, for ease of presentation and readability, we have split Results into two chapters;

the first concerns the narrative review KQs and the second addresses the systematic review KQs. We conclude with the Discussion chapter, which addresses findings, strengths, and limitations of the project and the evidence base, applicability, and similar topics for all 11 KQs. The appendices document the following methods or data: A, Search Strategy; B, List of Excluded Studies; C, Evidence Tables; and D, Risk of Bias Ratings.

Methods

This Technology Assessment is organized into sections addressing the Narrative Review Key Questions (KQs) (1 through 5) and the Systematic Review KQs (6 through 11) concerning treatment-resistant depression (TRD). Table 1 gives our selection (inclusion/exclusion) criteria based on PICOTS (Population, Interventions, Comparators, Outcomes, Timing, and Setting) and outlines our methods to answer the KQs.

Inclusion and Exclusion Criteria

Our population of interest is adults 18 years of age or older with depression who have not responded to treatment(s). The depressive illness can be part of either major depressive disorder (MDD) or bipolar disorder, but one of these diagnoses must be a primary diagnosis per the Diagnostic and Statistical Manual of Mental Disorders, fourth edition[25] or 5th edition.[26] For example, studies of patients who have schizophrenia with a secondary diagnosis of MDD or who have dysthymia would not be eligible for this report. If a study involved both eligible and ineligible patients and did not report data separately, we excluded that whole study. Populations with no evidence of treatment nonresponse (e.g., a study in which the absence of treatment response is not part of the selection criteria) were also not eligible.

Eligible interventions included those that have both been tested as a treatment targeting TRD in adults and identified by guidelines, consensus statements, the Medicare Evidence Development & Coverage Advisory Committee (MEDCAC) panel of April 27, 2016, or systematic reviews as alternatives for TRD treatment studies. These criteria ensure consideration of interventions with a minimum threshold amount of data addressing their effectiveness in TRD populations. Comparison groups include concurrent control groups (e.g., active, sham, or placebo) and a control group from an interrupted time series.

We required outcomes to have been identified in our previous comparative effectiveness work[58, 59] as the most meaningful to depression management decisionmaking. In that review, we had asked our Technical Expert Panel and Key Informants to rank the relative importance of these depression management outcomes following a process proposed by the GRADE Working Group.[60] We then had that panel anonymously rank the relative importance of outcomes using SurveyMonkey©. Participants used a 9-point Likert scale to rank outcomes into three categories: (1) critical for decisionmaking, (2) important but not critical for decisionmaking, and (3) of low importance for decisionmaking. They identified six outcomes as critical and five as important, and they supported the inclusion of an additional depressive outcome (change in depressive severity). For one of the adverse events outcomes, serious adverse events, we used the U.S. Food and Drug Administration (FDA) definition[61] and considered physical, psychological, and cognitive events. We required relevant studies for the current project to report on at least 1 of these 12 outcomes.

We carefully considered eligible dates of publications to focus the review on information most relevant to the current understanding of TRD and the interests of the Centers for Medicare & Medicaid Services. Given our interest in including information pertinent to FDA-approved treatments for TRD (including vagus nerve stimulation, approved in 2005[62]), we set our starting publication date for intervention studies (KQs 6 through 11) and systematic reviews (part of KQs 1 through 5) as January 1, 2005. For all other materials eligible for the KQ 1 through 5 narrative review (e.g., consensus statements and guidelines), we considered literature published since January 1, 1995, reflecting our sponsor's interest in comprehensively reviewing the variety of

conceptualizations of TRD. This beginning date also provides literature relevant to current definitions of TRD with diagnoses.

All study durations and all settings in very highly developed countries, according to the Human Development Index (using three dimensions: long and healthy life, knowledge, and a decent standard of living),[29] were eligible. Pre/post studies that did not use interrupted time series analyses were excluded, because potential confounding from multiple sources renders questionable the ability of these study designs to support causal inferences. We included English-language articles and excluded studies that had not been published fully in English, because their ability to provide meaningful information about the current understanding of TRD in a Medicare- or Medicare-related population is limited.

Table 1. Inclusion/exclusion criteria for studies of treatment-resistant depression

PICOTS	Inclusion Criteria	Exclusion Criteria
Population	All adult populations (≥18 years old) identified as having a primary diagnosis of depression (including MDD and bipolar disorder) who have not responded to treatment(s) (the least stringent definition of TRD). The depressive episode must be part of a major depressive disorder or a bipolar disorder per DSM-IV or -5.	Populations without a primary diagnosis of MDD or bipolar disorder are excluded, as well as those without evidence of treatment nonresponse.
Interventions	Any pharmacologic intervention[a] tested as a treatment for TRD as a primary therapy or as an augmentation agent to an existing primary therapy. • Antidepressants (e.g., SSRIs, SNRIs, TCAs, MAOIs, atypical agents) • Atypical antipsychotics • Anticonvulsants • Mood stabilizers • Psychostimulants • Agents approved by FDA for other indications but tested in TRD populations (e.g., ketamine, levothyroxine, clonidine) Any nonpharmacologic device or procedure tested as a treatment for TRD as a primary therapy or as augmentation to an existing primary therapy. • Devices (i.e., ECT, rTMS, VNS, DBS, MST, tDCS, LFMS, CES). For some of our analyses, we collapse this set of interventions into the category "brain stimulation therapies," or BST. Any nonpharmacologic intervention tested as a treatment for TRD as a primary therapy or as augmentation to an existing primary therapy. • CAM (i.e., acupuncture, meditation [e.g., mindfulness-based stress reduction], omega-3 fatty acids, S-adenosyl-L-methionine (SAMe), St. John's wort (*Hypericum perforatum*), light therapy, sleep deprivation) • Psychotherapy (i.e., CBT, third wave CBT, psychodynamic therapies, and integrative therapies) • Exercise (i.e., any formal exercise program)	Interventions not targeting TRD
Comparators	All comparative studies with concurrent control groups or control groups from an interrupted time series or pre/post studies with interrupted time series analyses (which require that data are collected at two or more time points before and after an intervention).	Pre/post studies where interrupted time-series analyses were not conducted

Table 1. Inclusion/exclusion criteria (continued)

PICOTS	Inclusion Criteria	Exclusion Criteria
Outcomes	Mental health outcomes identified in previous depression comparative effectiveness review work as either critical or important for decisionmaking: Benefits that are reported as primary endpoints (or outcomes) for a study: • Reduction in suicidal ideation or suicide attempts • Quality of life • Response to treatment • Remission • Change in depressive severity • Functional capacity (physical and cognitive functioning measured by validated scales) • Speed of remission • Speed of response • Intervention durability (rates or counts of recurrence of a depressive episode for those who have remitted) Adverse events from the intervention identified as either critical or important for decisionmaking: • Serious adverse events per FDA definition[b] (rates or counts) • Overall adverse events (rates or counts) • Treatment discontinuations attributed to adverse events (rates or counts)	None
Timing	Any study duration For KQs 1–5: For systematic reviews, publication date from 1/1/2005 to present; for other literature type or study designs, publication date from 1/1/95 to present For KQs 6–11: Literature publication date from 1/1/2005 to present	Literature published before these specific dates
Setting	Studies that took place in very highly developed countries[c]	Studies that took place in high, medium, or low human development countries
Study designs	For KQs 1–5: Consensus statements, guidelines, materials from CMS, SAMHSA, FDA, or NIH; UpToDate®; information from the 2016 MEDCAC panel meeting on the definition of TRD; and systematic review articles that (a) searched two or more literature databases, (b) included dual review of the literature and data abstraction, and (c) included quality or risk of bias assessments of included studies. For KQs 6–11: Randomized, or prospective nonrandomized, or observational studies (including concurrent controls and interrupted time series)	For KQs 1–5: Evidence not meeting inclusion criteria. Individual trials were *not* considered in this section. For KQs 6–11: Pre/post studies without interrupted time-series analyses Any studies without a control group
Language	English only	Studies not published in English

[a] Pharmacologic: SSRIs: citalopram, escitalopram, fluoxetine, fluvoxamine, paroxetine, sertraline; SNRIs: desvenlafaxine, duloxetine, levomilnacipran, mirtazapine, venlafaxine; NADRIs: bupropion; TCAs: amitriptyline, clomipramine, desipramine, doxepin, imipramine, maprotiline, nortriptyline; MAOIs: phenelzine, selegiline transdermal, tranylcypromine; 5HT Ras: nefazodone, trazodone, vilazodone, vortioxetine; atypical antipsychotics: cariprazine, quetiapine; NMDAs: ketamine; Other Pharmacologic for Combination or Augmentation: Atypical antipsychotics: aripiprazole, asenapine maleate, clozapine, iloperidone, lurasidone, olanzapine, paliperidone, quetiapine, risperidone, ziprasidone; Anticonvulsants: carbamazepine, lamotrigine, oxcarbazepine, topiramate, valproic acid; Psychostimulants: amphetamine-dextroamphetamine, armodafinil, dexmethylphenidate, dextroamphetamine, lisdexamfetamine, methyphenidate, modafinil; Mood stabilizers: lithium, divalproex; Other augmenters: bupropion, buspirone, clonidine, liothyronine, pindolol, pramipexole, triiodo-thyronine (T3).

[b] For serious adverse events, we use the FDA definition and will consider physical, psychological, and cognitive events.[61]

[c] "Very High" on Human Development Index: Andorra, Argentina, Australia, Austria, Bahrain, Belgium, Brunei Darussalam, Canada, Chile, Croatia, Cyprus, Czech Republic, Denmark, Estonia, Finland, France, Germany, Greece, Hong Kong China (SAR), Hungary, Iceland, Ireland, Israel, Italy, Japan, Korea (Republic of), Kuwait, Latvia, Liechtenstein, Lithuania, Luxembourg, Malta, Montenegro, Netherlands, New Zealand,

Norway, Poland, Portugal, Qatar, Saudi Arabia, Singapore, Slovakia, Slovenia, Spain, Sweden, Switzerland, United Arab Emirates, United Kingdom, United States.[29]

BST = brain stimulation treatment; CAM = complementary and alternative medicine therapies; CBT = cognitive behavior therapy; CES = cranial electrotherapy stimulation; CMS = Centers for Medicare & Medicaid Services; DBS = deep brain stimulation; DSM-IV = Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition; DSM-5 = Diagnostic and Statistical Manual of Mental Disorders, 5th Edition; ECT = electroconvulsive therapy; FDA = U.S. Food and Drug Administration; KQ = Key Question; LFMS = low field magnetic stimulation; MAOI = monoamine oxidase inhibitor; MDD = major depressive disorder; MEDCAC = Medicare Evidence Development &Coverage Advisory Committee; MST = magnetic seizure therapy; NIH = National Institutes of Health; PICOTS = population, intervention, comparator; outcome; timing, setting; rTMS = repetitive transcranial magnetic stimulation; SAMe = S-adenosyl-L-methionine; SAMHSA = Substance Abuse and Mental Health Services Administration; SNRI = serotonin-norepinephrine reuptake inhibitor; SSRI = selective serotonin reuptake inhibitor; TCA = tricyclic antidepressant; tDCS = transcranial direct stimulation; TRD = treatment-resistant depression; VNS = vagus nerve stimulation.

Searching for the Evidence: Literature Search Strategies to Identify Relevant Studies to Answer Key Questions

An experienced research librarian at the RTI International-University of North Carolina Evidence-based Practice Center developed the strategy for our comprehensive search of the literature. To ensure methodological quality, we followed standard procedures for systematic literature searches specified in the Agency for Healthcare Research and Quality (AHRQ) Effective Health Care Program *Methods Guide for Effectiveness and Comparative Effectiveness Reviews.*[24] Although KQs 1 through 5 were narrative questions (which often turn up for systematic reviews), we nevertheless applied standard methods for the title/abstract and full-text review as much as possible to meet the typical global standards for this step.

We systematically searched the published literature from January 1, 1995, to August 18, 2017, that is indexed in MEDLINE®, EMBASE, PsycINFO, and Cochrane Library and that addresses treatment of TRD in adults. The aim was to assemble literature relevant to current definitions of TRD with diagnoses consistent with definitions in *Diagnostic and Statistical Manual of Mental Disorders*, Fourth edition[25] and 5th edition.[26] As noted earlier, for consensus statements and guidelines that were not part of the published peer-reviewed literature we considered materials that had appeared since January 1, 1995. For systematic reviews that became part of the evidence base for some parts of the narrative KQs 1 through 5, and for intervention studies relevant for the systematic review KQs 6 through 11, we set our publication date for intervention studies as January 1, 2005. We also reviewed both reference lists of all systematic reviews that we included for KQs 1 through 5 and indexed protocols; the goal was to identify any relevant citations that our electronic searches might have missed.

In addition, we searched for consensus statements, management guidelines, and relevant government materials from various Federal agencies, specifically including the following: CMS (and MEDCAC), FDA, National Institutes Health, National Institute of Mental Health, and the Substance Abuse and Mental Health Services Administration. We abstracted Information relevant to KQs 1 through 5 from such sources. We also searched other Web sites such as Clinicaltrials.gov, Guideline.gov (AHRQ's National Guidelines Clearinghouse), HSRProj (Health Services Research Projects in Progress database), and UpToDate.

Further, the Evidence-based Practice Center Program's Scientific Resource Center contacted relevant stakeholders, including manufacturers of prescription medications and medical devices used to treat MDD, for scientific information packets (Supplemental Evidence and Data for Systematic reviews [SEADS]) that contained any unpublished information on the efficacy and/or safety of their products when used specifically to treat TRD. SEADS allow an opportunity for the intervention developers and distributors to provide the Evidence-based Practice Center with both published and unpublished data that they believe should be considered for the review.

Finally, a notice was also placed in the *Federal Register* requesting any relevant information on the use treatments for TRD.

Trained members of the research team dually reviewed all titles and abstracts for eligibility based on the pre-established inclusion/exclusion criteria presented in Table 1. Studies marked for possible inclusion by either reviewer underwent full-text review. Any study with inadequate information in the abstract also proceeded to full-text review. These researchers then dually reviewed each full-text article for inclusion or exclusion on the basis of the eligibility criteria. We documented reasons for exclusion at this stage; we also tagged those selected for inclusion with the relevant KQ that the article addressed. Disagreements about inclusion were resolved by consensus discussion.

Data Abstraction and Data Management

From included systematic reviews, consensus statements, guidelines, and other relevant materials, we abstracted relevant information (e.g., definitions of TRD, study designs, methods, measures, and risk factors) to answer Narrative Review KQs 1 through 5. From all included individual intervention studies, we abstracted relevant information to answer Systematic Review KQs 6 through 11 (see below).

These steps allowed us to catalogue and describe the available controlled studies. We tracked all literature screening results in the EndNote database. We also recorded the reason that each excluded full-text publication did not satisfy the eligibility criteria.

We abstracted data relevant to the KQs from any studies that met our inclusion criteria into a standardized template. One member of the research team collected the data, and another (senior) investigator reviewed the abstraction for accuracy and completeness.

Assessment of Risk of Bias of Included Studies

Two senior investigators independently assessed the risk of bias of only the individual studies included for KQ 10, because that is the only KQ for which we are assessing a causal relationship. We use risk of bias as a covariate in the regression analyses. Disagreements were resolved by discussion and consensus or by consulting an independent third party. We did not assess risk of bias of other studies as the report is a Technology Assessment.

For randomized controlled trials (RCTs), we used the Cochrane Risk of Bias tool.[27] Elements of risk of bias assessment for RCTs include, among others, randomization and allocation concealment, similarity of compared groups at baseline, masking of patients and study personnel, use of intent-to-treat analysis, and overall and differential loss to followup.

For nonrandomized trials and observational studies, we employed criteria outlined by the Newcastle-Ottawa Scale, a commonly used tool for assessing quality of nonrandomized studies.[28] Elements of this tool assess the comparability of baseline characteristics, the method of statistical adjustment for baseline confounding, and the assessment of outcomes.

Data Synthesis

For the Narrative Review KQs, we present summary text and a series of tables that address each KQ. For example, for KQ 1, the summary table documents the variability of the definitions of TRD used; this information let us identify where any consensus appears to lie. Similar to KQ 1, separate summary tables for KQ 2, KQ 3, and KQ 4 present the various methods used to diagnose TRD and the measures and study designs that investigators use in TRD research. Such

summary tables allow us to identify any consensus for these issues. Finally, for KQ 5, we report information on identified risk factors for TRD. For all these KQs, we have interpretative text summarizing the content of the tables. We did no quantitative analyses; rather, we provide a qualitative synthesis of what information in these tables mean as answers to these five KQs.

For the subquestions in KQ 2, KQ 3, and KQ 4, we developed separate summary tables that address the specific characteristics called out in these questions. Examples include diagnostic tools used in the different diagnostic settings, psychometric properties of measures used to determine efficacy or effectiveness, and study designs that have demonstrated effects on, for example, minimizing bias and placebo effects.

For this Narrative Review, we defined "consensus" in the following manner. "Consensus" reflects a "majority agreement" that can be arrived at in numerous ways. Such a "majority agreement" can be determined by guidelines or best practices, or consensus statements, or systematic reviews that represent state of the art consideration of the relevant topic, or other ways. Our charge was to identify a "consensus" in the TRD literature. Since there is no clearly identified group to formulate such an agreement, we believe that what is presented in consensus statements, clinical practice guidelines, government reports, and the peer-reviewed literature is a legitimate marker for agreement among clinical, research, and policy experts, especially if a majority of those citations appear to agree with each other over time. Hence, we consider such a majority count to indicate consensus.

For the Systematic Review KQs, we developed a similar series of tables addressing KQs 6 through 9 and KQ 11, again with summary text highlighting key findings based on the information in the various tables. For KQ 10 (regression or other statistical analysis), we first defined patient- and study-level covariates that might be relevant in examining correlations. Because we did not have access to individual patient data, we focused primarily on study-level characteristics (e.g., study design, study duration, risk of bias).

To avoid issues of ecological fallacy, whereby inferences about the nature of individuals are wrongly deduced from inferences for the group to which those individuals belong, we converted patient-level covariates to study-level covariates whenever possible (e.g., severity of depression to proportion of patients with severe depression in a study). To ensure consistency, we developed a data codebook and an analysis plan after we selected the covariates. We used Microsoft Excel and SAS software for data management, data cleaning, and graphical display of the data.

Regression or other statistical analyses focus on interventions for which we have at least 10 studies using a similar comparator intervention. [63]Our main focus is on interventions in general for TRD; we did not focus any of our primary analyses on intervention type (e.g., rTMS vs. psychopharmacologic). We combined interventions into categories (e.g., pharmacological interventions, behavioral interventions). We classified comparator interventions as inactive (e.g., placebo, waiting list, sham) or active. We also selected relevant outcome measures, focusing insofar as possible on patient-centered outcomes.

For computational ease, we focused on dichotomous outcomes (odds ratios). We also recalculated the direction of effect, if necessary, so that an odds ratio >1 indicates a beneficial effect and an odds ratio <1 indicates a harmful effect.

The impact of the study- and patient-level characteristics on treatment effects for each outcome were assessed using random effects meta-regression models. The models were fit using SAS PROC GLIMMIX with a binomial likelihood and logit link function. To quantify the impact of a characteristic on the treatment effect, we computed the ratio of odds ratios and associated 95 percent confidence intervals to compare the odds ratio of the intervention effect

(e.g., Brain Stimulation Therapy vs. control) for studies with the specified characteristic (e.g., included older adults) to the odds ratio for studies without the specified characteristic (e.g., did not include older adults). We began with bivariable analyses, comparing results for one characteristic at a time and then conducted multivariable analyses, including all of the characteristics that were significant in the bivariable analyses for the applicable outcome. Some study- and patient-level characteristics were excluded from the analyses owing to lack of variability across studies or high levels of missing data.

Assessing Applicability

Applicability of findings may vary substantially by the PICOTS. For that reason, we highlight how variability of PICOTS elements could influence applicability (i.e., generalizability or external validity). For example, a TRD definition may differ by population, such as whether the depressive episode is part of MDD or bipolar disorder. Also, a TRD definition relevant to specialty psychiatric settings may not be applicable (or feasible) in primary care settings. Furthermore, findings may differ depending on the definition of the primary outcome of interest (e.g., depression remission vs. improved function).

Results: Narrative Review Key Questions

Introduction

For ease of presentation and use of our findings, we have divided our results into two chapters, one focused on the five narrative questions and one (which follows) on the six systematic questions. In this chapter, we present our findings sequentially by Key Question (KQ)—namely KQs 1 through 5. Generally, we deal with treatment-resistant depression (TRD) first for its presence in patients with major depressive disorder (MDD) and then TRD in patients with bipolar disorder.

As noted in Methods, we present interpretative text and summary tables documenting our findings. All main sections are introduced by a set of bulleted key points.

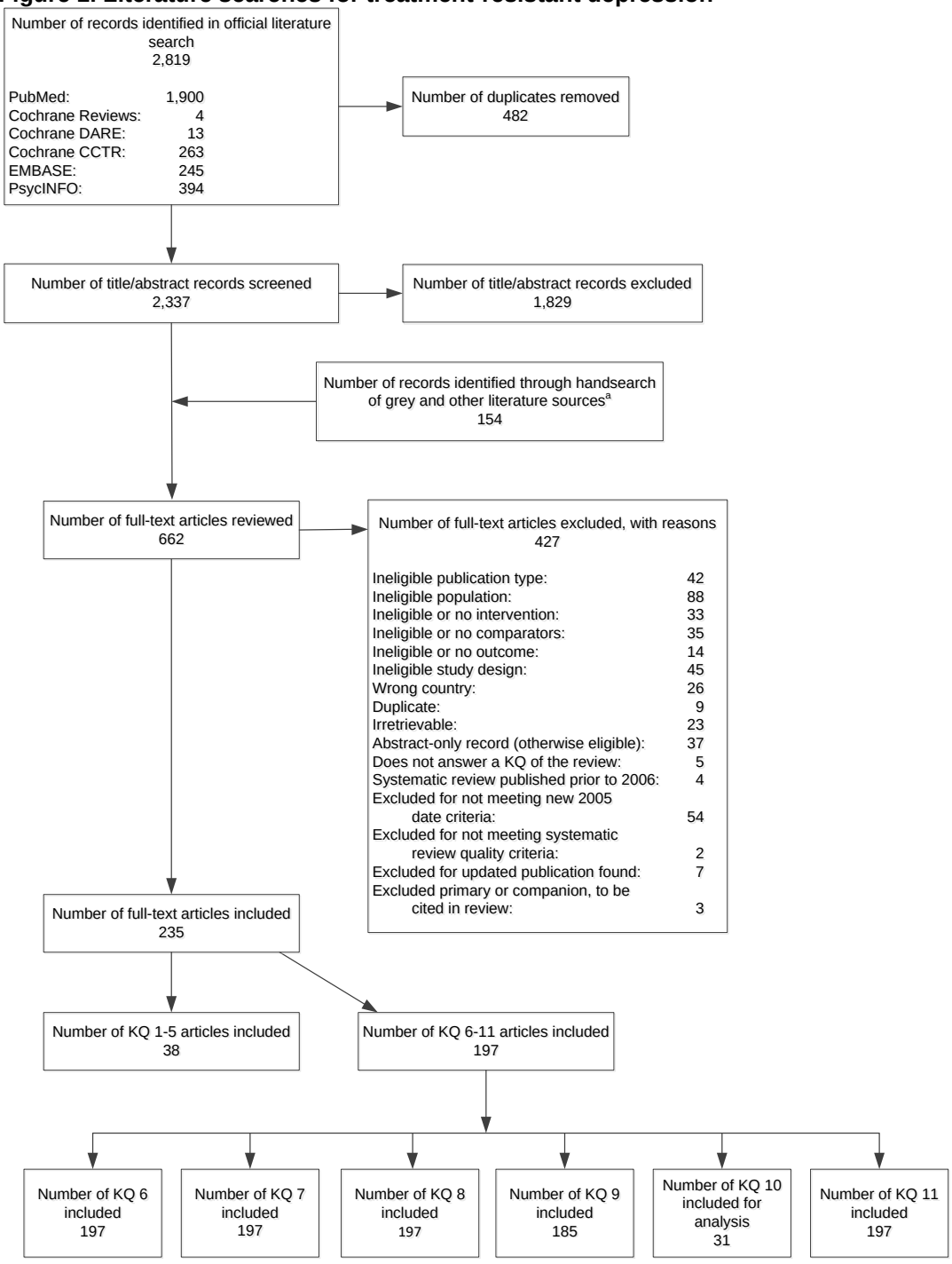
Results of Literature Searches

Figure 1 documents the literature searches for all KQs. It also records the yields from our efforts to identify other sources of information, including handsearching grey and other literature sources. We received materials containing 30 suggested articles via the SEADS process. Twenty-three were already included in our database. Seven were screened during title/abstract review. Two moved forward to full text review, resulting in one article added to the included pool of studies for this review.

We screened 2,337 potentially relevant articles. The two most common reasons for exclusion at the full-text level were ineligible population (89 exclusions) and ineligible study design (45 exclusions). Across all 11 KQs, we included 235 articles (38 for KQs 1 through 5, reported in this chapter, and 197 for KQs 6 through 11, reported in the next chapter). Appendix A presents the literature search strategies; Appendix B lists the articles excluded at the full-text stage of review.

For our analyses, we used three general categories of publications: systematic reviews, nonsystematic reviews, and guidelines or consensus statements.

Figure 1. Literature searches for treatment-resistant depression



[a]We handsearched 191 records from a search of ClinicalTrials.gov, 31 records from a search of the National Guideline Clearinghouse, 2 records from a search of HSRProj, 5 websites (NIMH, UpToDate, AHRQ EHC, SAMHSA, FDA), the 4/27/2016 MEDCAC Panel Proceedings, records from the TRD proposal and protocol, 30 records received from a SEADs/FRN request, and articles included for KQs 1 through 5 for additional relevant records. Relevant records were checked against the database for duplication and 154 relevant records were moved forward to dual full-text review.

AHRQ = Agency for Healthcare Research and Quality; CCTR = Central Register of Controlled Trials; DARE = Database of Abstracts of Reviews of Effects; EHC = Effective Health Care; EMBASE = Excerpta Medica Database; FDA = Food and Drug Administration; FRN = Future Research Needs; HSRProj = Health Services Research Projects in Progress; KQ = Key Question; MEDCAC = Medicare Evidence Development and Coverage Advisory; NIMH = National Institute of Mental Health; PsycINFO = Psychological Information Database; SAMHSA = Substance Abuse and Mental Health Services Administration; SEADs = Supplemental Evidence and Data Request.

Key Question 1: Definitions of Treatment-Resistant Depression in This Literature Base

Description of Included Studies

We identified 38 publications that directly addressed definitions of TRD (see Table 2). 14 publications were systematic reviews;[31-36, 39-41, 44, 58, 64-66] six publications were reviews that were not systematic.[5, 18, 30, 37, 45, 46] In addition, 18 publications reported on 13 different guidelines or consensus statements.[19, 21, 23, 38, 42, 67-79] All but two addressed only MDD; the exceptions addressed TRD in both MDD and bipolar disorder.[41, 42]

Table 2. Types and numbers of publications addressing definitions of TRD

Types of Publications Addressing TRD	Number of Publications	MDD Only	Both MDD and Bipolar
Systematic reviews	14	9	1
Non-systematic reviews	6	9	0
Guidelines and consensus statements	18	18	1
Total	38	36	2

MDD = major depressive disorder; TRD = treatment resistant depression.

We sort our findings for KQ 1 into three categories:

1. *Definitions of TRD*: We identify the various definitions of TRD and how they differ by key components such as numbers of depression treatment failures.
2. *Staging models of TRD (which classify patients by level of treatment resistance)*: We cover issues relating to adequacy of treatment duration, dosage, and other factors noted in the Introduction.
3. *Consensus statements or guidelines involving TRD*: We identify what best practices or guidelines appear to be highly relevant to definitions of TRD.

Key Points

1. TRD has been defined as both a dichotomous term (i.e., someone either has it or does not) and a continuous term that falls along a spectrum (i.e., people have different degrees, or stages, of severity).
2. Defined dichotomously, we identified four distinct definitions of TRD, distinguished primarily by the number of prior treatment failures.
3. Since 2005, definitions have emphasized failure to achieve remission as the preferred definition of treatment failure.
4. Nearly all definitions have addressed TRD as a part of MDD; the few definitions that have considered bipolar TRD have noted that TRD within bipolar disorder is a distinct entity from MDD.
5. The five TRD staging models we identified had only limited research addressing reliability and validity (particularly predictive validity); these models appeared to be equally valid for documenting treatment failure in depressed patients, but their applicability and feasibility in clinical practice are unclear.
6. No widely acknowledged consensus exists on the best definition of TRD. However, the majority of systematic reviews and guidelines or consensus statements reported that the commonly used definitions were based on patients whose depression failed to respond (a decrease in depressive severity of at least half) or did not go into remission (complete

recovery as measured by a score on a depressive severity instrument below a threshold) following two or more treatment attempts of an adequate dose and duration.

a. Whether the treatment attempts require different classes of antidepressants is not a settled matter.
b. Experts do not agree on how to define adequate dose and duration, although the minimum duration cited is 4 weeks.

Detailed Synthesis

TRD has no established definition. We identified two general approaches to defining it. The first and more common approach considers TRD as a dichotomous term—patients either have or do not have this diagnosis—based on whether they meet a set of threshold criteria. The second approach considers TRD an illness that falls along a spectrum, with different degrees, or stages, of severity. In this latter scheme, either patients do not have TRD at all or they have different degrees of TRD severity (e.g., Stage 1, Stage 2, as described later).

Use of one or the other of these basic definitions—and more importantly what specific elements are included in them—has challenged researchers, clinicians, and policymakers for years. Uppermost is the need to agree on a "proper" definition of TRD. Accordingly, below we discuss the evidence in three sections: the variety of dichotomous TRD definitions, the staging models of TRD, and what the consensus definition appears to be.

Dichotomous Definitions of Treatment-Resistant Depression

TRD is defined most commonly by the number of prior antidepressant failures of treating depression (in either MDD or, less often, bipolar disorder). These failures can range from a single treatment failure (relating to any drug) to three or more failures using three different classes of antidepressants.

Further, most definitions consider additional variables. One is how failure is defined (i.e., as an inadequate reduction in depressive severity, a lack of response [typically a 50% decrease in severity], or failure to remit [understood to mean complete recovery from a depressive episode and typically indicated by a reduction of depressive severity below a threshold]). Others include whether the failure occurred in the current episode, whether the individual received an adequate dose of the medication, and whether the duration of treatment was considered adequate. Definitions for these additional variables often differed as well.[41]

Of note, since 2005, a consensus has been developing that the proper definition of failure is the inability to achieve remission (sometimes "to remit"). Other agreement among experts appears to be that an adequate trial is, at a minimum, 4 weeks at an adequate dose,[31, 41] although some argue that 6 weeks should be the minimum. Most studies did not present information requirements for prior treatment length. Durations of psychotherapy treatments tended to be longer, six or more weeks. Agreement is less clear whether an adequate dose means a minimum effective therapeutic dose or a maximum tolerated dose.

We had initially identified seven currently used definitions of TRD, but we condensed them into four possible categories because some distinctions did not appear to be easy to explain or useful for primary care clinicians or researchers. Table 3 groups them according to the numbers

Table 3. Four categories of definitions of treatment-resistant depression by number of treatment failures

Number of Treatment Failures	Type of Publication on TRD Treatments, Date	Ways to Define Failure	Specify Current Episode?	Define Adequate Dose?	Define Adequate Duration?
1 or more	Seminal article on defining TRD, 2003[45]	Nonresponse (<25%), partial response (≥25% to <50%), response without remission (50% or greater symptom reduction)	During current episode	Within therapeutic range but conflicting dosages recommended for same drug by different authors	≥6 weeks
	International Workshop on "Present and Future of TMS: Safety and Ethical Guideline" Rossi et al., 2009[68]	Not described	Not described	Not described	Not described
	SR - lamotrigine augmentation, 2010[32]	Nonresponse (<50% decrease in HAM-D after 1 failed treatment or equivalent)	Not described	Not described	4 weeks
	SR - psychotherapy, 2011[33]	Nonresponse, partial response, or no remission (not further defined)	Not described	Most studies in the review provided the accepted dose range	≥6 weeks
	Nonsystematic review defining TRD, 2014[18]	Nonresponse, partial response, or no remission (not further defined)	Not described	Majority of studies in the review referred to adequacy as therapeutic levels	6 to 8 weeks
	SR - rTMS, 2015[35]	Nonresponse, inadequate response, insufficient response (not further defined)	During current episode	Not described	4 to 6 weeks
	SR - predictors of nonresponse, 2016[36]	Nonresponse, no remission (not further defined)	Not described	Not described	Not described
	SR and MA of olanzapine/fluoxetine combination, Luan et al., 2017[66]	Nonresponse, no remission (not further defined)	Not described	Not described	Not described

Table 3. Four categories of definitions of treatment-resistant depression by number of treatment failures (continued)

Number of Treatment Failures	Type of Publication on TRD Treatments, Date	Ways to Define Failure	Specify Current Episode?	Define Adequate Dose?	Define Adequate Duration?
2 or more	Seminal article on definition of TRD, 2001[37]	Nonresponse or lack of remission (not further defined)	During current episode	At least two thirds of the maximal PDR dose	≥4 weeks
	SR - pharmacologic treatments, 2007[31]	Majority of studies in the review using 2 or more failures incorporated nonresponse or remission in the definition	Majority of studies in the review did not specifically indicate whether the failed trials were during the current MDD episode or part of previous episodes as well	Majority of studies in this review refer to adequacy in a general manner: standard minimum effective doses, maximum tolerated doses within the therapeutic range, acceptable therapeutic doses	≥4 or ≥6 or ≥8 weeks
	Unipolar Depression Guideline Harter et al., 2010[70]	Nonresponse in terms of insufficient change in patients' symptoms	Not described	Not described	3 to 4 weeks
	Institute for Clinical and Economic Review (ICER) Coverage Policy Analysis, 2012[38]	Lack of clinically meaningful response or remission or lack of decrease in symptom severity in HAM-D of less than 5 points (a 3-point difference is considered clinically meaningful)	In the current episode	At least one of the treatment trials must have been administered at an adequate course of mono- or poly–drug therapy	Not described for TRD
	Report from European Medicines Agency consensus meeting in 2009—Improving outcome in TRD, Schlaepfer et al., 2012[69]	Failure to respond (not further defined)	Not described	Minimum dosage that will produce the expected effect or the maximum dosage within the therapeutic range that the patient can tolerate until the expected effect is achieved	4–6 weeks is considered an adequate trial period to see clinical response, although recent research suggests that longer periods (up to 8 or 12 weeks) may be needed to achieve remission
	SR - lithium or atypical antipsychotics, 2013[39]	Failure to respond (not further defined)	Few studies in the review specified in the current episode of depression	Not described	≥4 weeks for augmentation

Table 3. Four categories of definitions of treatment-resistant depression by number of treatment failures (continued)

Number of Treatment Failures	Type of Publication on TRD Treatments, Date	Ways to Define Failure	Specify Current Episode?	Define Adequate Dose?	Define Adequate Duration?
2 or more (continued)	SR - rTMS, 2014[40]	Defines nonresponse (≤50% improvement on HAM-D)	Few studies in the review specified in the current episode of depression	Variably defined by individual study authors in the review	≥4 weeks
	SR - nonpharmacological, 2014[41]	No standard definition of AD failure, but a variety of TRD staging tools provide different ways of assessing the adequacy of prior treatment so that the clinician can determine treatment failure	Few studies in the review specified in the current episode of depression	Most common definition: maximum tolerated dose	≥4 to 8 weeks
	Australian/New Zealand Clinical Practice Guideline, 2015[42]	Failure to respond (not further defined)	Not described	Maximal dosage (or blood level achieved)	≥4 weeks
	British Association for Psychopharmacology Guidelines for all depressive orders, revision based on 2012 consensus meeting, Cleare et al., 2015[79]	Lack of improvement (defined as at least a 20% to 30% reduction in HAM-D in different studies) at 4 and 6 weeks; only 20% and 10%, respectively, will go on to eventual response (≥50% improvement) at 8 weeks	Important to consider the current episode in the context of the overall history of depression and the nature of previous episodes when considering treatment options	Adequate treatment, defined as "recommended therapeutic dose"	4 to 8 weeks
	CANMAT Guidelines, 2016 Kennedy et al., 2016[74] Pharmacological treatments	Partial response (25%-49% reduction in symptom scores) or no response (<25% reduction)	Not described	Not described	2 to 4 weeks
	VA/DoD Clinical Practice Guideline, 2016[23]	Failure to respond (not further defined)	Not described	Appropriate dose titration and target dose range	≥4 to 6 weeks
	SR and MA of pharmacological and somatic interventions, Papadimitropoulou et al., 2017[65]	Nonresponse: MADRS scores from baseline less than 50% at study endpoint Nonremission mostly defined as a MADRS score greater than 7 at study endpoint	At least one failure in the current episode	Not described	Not described

Table 3. Four categories of definitions of treatment-resistant depression by number of treatment failures (continued)

Number of Treatment Failures	Type of Publication on TRD Treatments, Date	Ways to Define Failure	Specify Current Episode?	Define Adequate Dose?	Define Adequate Duration?
3 or more	ICSI Adult Depression in Primary Care Guideline, 2016[21]	Failure to achieve remission with an adequate trial of therapy and three different classes of ADs at adequate duration and dosage True treatment resistance was seen as occurring on a continuum, from failure to reach remission after an adequate trial of a single AD to failure to achieve remission despite several trials of ADs, augmentation strategies, electroconvulsive therapy (ECT), and psychotherapy	Not described for TRD	Not described for TRD	Not described for TRD
For bipolar TRD: 1 failure	Washington State Health Care Authority, 2014[41]	Definition couched as lack of significant reduction in score on a depression symptom scale rather than in terms of the number of treatment failures International Society for Bipolar Disorders recommends using no significant reduction in Montgomery-Åsberg Depression Rating Scale (MADRS) or Hamilton Rating Scale for Depression (HAM-D) score	Not described for bipolar TRD	Not described for bipolar TRD	Time frame required for an adequate trial of AD may need to be *longer* than with unipolar depression because of the greater natural fluctuation of the disease, which suggests that the clinician may need to observe a patient 2 to 4 weeks beyond the time frame usually considered adequate for an AD trial International Society for Bipolar Disorders recommends an ideal trial duration of 10 to 12 weeks

Table 3. Four categories of definitions of treatment-resistant depression by number of treatment failures (continued)

Number of Treatment Failures	Type of Publication on TRD Treatments, Date	Ways to Define Failure	Specify Current Episode?	Define Adequate Dose?	Define Adequate Duration?
For bipolar TRD: 1 failure (continued)	Australian/New Zealand Clinical Practice Guideline, 2015[42]	Failure to remit (not further defined)	Does not clarify but implies current episode	Bipolar I TRD: lithium (blood level 0.6–0.8 mMol/L) or two other adequate ongoing mood-stabilizing treatment, plus lamotrigine (50–200 mg/day) or with full dose quetiapine (≥ 600 mg/day) as monotherapy Bipolar II TRD: lithium (blood level 0.6–0.8 mMol/L) or two other adequate ongoing mood-stabilizing treatment, plus lamotrigine (50–200 mg/day) or with full dose quetiapine *(*defined for Bipolar II as 300–600 mg/day) as monotherapy	Does not clarify but implies as least 3 weeks for Bipolar I and II TRD

AD = antidepressant; ECT = electroconvulsive therapy**;** HAM-D = Hamilton Rating Scale For Depression; MA = meta-analysis; MADRS = Montgomery-Åsberg Depression Rating Scale; MDD = major depressive disorder; PDR = Physician's Desk Reference; RCT = randomized controlled trial; SR = systematic review; TRD = treatment-resistant depression; VA\DoD = Veterans Administration\Department of Defense.

of failures. Specifically, for failures of MDD therapies, these are (1) one for more failures; (2) two or more failures; and (3) three or more failures. For failures of bipolar I or II depression, TRD is defined as failure of one prior trial, although this failure has been variably identified as following a single antidepressant attempt of 10 to 12 weeks[41] or following a single adequate trial of lithium, or a mood-stabilizing medication and lamotrigine, or quetiapine monotherapy for at least 4 weeks.[42]

Within these four main categories, the various sources are listed in chronological order. The columns represent the four recommended components of a TRD definition. The separate rows and citations can include individual studies, systematic reviews, or guidelines and similar documents. No review compared the utility of applying one versus any other definition.

The applicability and feasibility of these staging tools in clinical practice are not known. The current evidence cannot yet confirm that staging of TRD is going to improve clinical practice.[44]

Components of Different Models for Staging Treatment-Resistant Depression

Five Basic Staging Models

Staging models acknowledge the dimensional nature of TRD, classifying patients along a spectrum according to their level of resistance to treatment. Several clinical variables, distinct from the components of the dichotomous definition, might affect the development or level of TRD. These variables include the duration of the episode, the depression subtype, depressive severity, and psychiatric or medical comorbidity.[31, 44]

We identified five staging models of TRD. Table 4 delineates key components of these models.

The Antidepressant Treatment History Form (ATHF)[80] is a scale extending from 1 to 5; it ranks medication resistance according to the adequacy of prior antidepressant treatment (as indicated by medication dose, treatment duration, and patient adherence). Prediction of treatment response is limited to studies with electroconvulsive therapy (ECT); three prospective studies showed an association between a high score on the ATHF and worse patient outcome. Reliability has been good in two studies.[44] This tool appears to be intended primarily for use in research settings.[41]

The Thase and Rush staging model (TRSM)[5] proposes a 5-part categorical scale in which patients are staged according to the number of classes of antidepressants that have failed to provide a response. Treatment resistance moves from more frequently used antidepressants (such as selective serotonin reuptake inhibitors or tricyclic antidepressants [TCAs]) to less frequently used drugs (e.g., monoamine oxidase inhibitors) or ECT. The predictive value of the TRSM has not been systematically assessed, and reliability has not been tested.[44]

Table 4. Staging models for treatment-resistent depression to define the spectrum of illness

Models Authors and Year of Publication	How Is Severity Scored?	How Is Failure Defined? Specify Current Episode?	Define Adequate Dose? Define Adequate Duration?	Staging Schema	Predictive Validity and Reliability Tested? Correspondence With ≥2 Treatment Failures Other Comments
Antidepressant Treatment History Form (ATHF) Ruhé et al., 2012[44]	Classification of 0 to 5 per treatment with a possible sum score for all treatments; ranks medication resistance according to adequacy of most potent previous trial	Not defined Yes	Yes, part of 5 stages on form Yes, part of 5 stages on form	5 stages • Stage 0: no treatment • Stage I: Any drug <4 weeks or less than minimum adequate daily dose; for ECT 1–3 sessions • Stage 2: Any drug ≥4 weeks at less than minimum adequate daily dose; for ECT 4–6 sessions • Stage 3: Any drug ≥4 weeks at minimum adequate daily dose; for ECT 7–9 unilateral sessions • Stage 4: Any drug ≥4 weeks at higher than minimum adequate daily dose; for ECT 10–12 unilateral/7–9 bilateral ECT sessions • Stage 5: Any drug at level 4 augmented with lithium ≥2 weeks; for ECT ≥13 unilateral/≥10 bilateral ECT sessions	Reliability studies done in the 90s showed good interrater reliability (intraclass correlation coefficient=0.94) Prediction of response limited to ECT, with the highest ATHF score associated with increased relapse time or lower response to ECT; however, this did not hold true for the total ATHF score. Does not correspond directly with 2 or more failures First developed in 1990, then updated in 1999; differentiates between MDD with or without psychotic features, but does not include psychotherapy treatments Used to determine the adequacy of prior pharmacotherapy and ECT trials (dosage and duration) but does not account for combination strategies Stages 1 and 2 indicate inadequate treatment

Table 4. Staging models for treatment-resistent depression to define the spectrum of illness (continued)

Models Authors and Year of Publication	How Is Severity Scored?	How Is Failure Defined? Specify Current Episode?	Define Adequate Dose? Define Adequate Duration?	Staging Schema	Predictive Validity and Reliability Tested? Correspondence With ≥2 Treatment Failures Other Comments
Thase and Rush Staging Model (TRSM) Thase et al., 1997[5] Fava et al., 2003[45] Berlim et al., 2007[31] Ruhé et al., 2012[44]	Categorized as a particular stage, with higher-numbered stages indicating a greater degree of treatment resistance	Failure to respond Not mentioned	No ≥4 weeks	5 stages • Stage I: Failure of at least one adequate trial of one major class of AD • Stage II: Stage 1 + failure of an adequate trial of an AD in a distinctly different class from Stage 1 • Stage III: Stage II plus failure of adequate trial of a TCA • Stage IV: Stage III plus failure of an adequate trial of an MAOI • Stage V: Stage IV plus failure of a course of bilateral ECT	The predictive value has not been systematically assessed and reliability has not been tested. Stage II corresponds with two treatment failures Looks at number of classes of ADs that have failed to provide a response; does not count psychotherapy in count of failed trials

Table 4. Staging models for treatment-resistent depression to define the spectrum of illness (continued)

Models Authors and Year of Publication	How Is Severity Scored?	How Is Failure Defined? Specify Current Episode?	Define Adequate Dose? Define Adequate Duration?	Staging Schema	Predictive Validity and Reliability Tested? Correspondence With ≥2 Treatment Failures Other Comments
European Staging Model Fekadu et al., 2009[46] Ruhé et al., 2012[44]	Determined by the number of weeks with treatment resistance to adequate dose of at least 2 different classes of ADs Treatment resistance for more than 12 months is designated as a distinct staging category: CRD	Poor response to a second (adequate) trial with a different class of AD (for 6–8 weeks); does not emphasize remission "Clinically relevant" TRD is a current episode of depressive disorder that has not benefited from at least two adequate trials of AD compounds of different mechanism of action"	Not found Nonresponder: 6–8 weeks TRD: From Level 1 of 12–16 weeks to Level 5 of 26 weeks to 1 year CRD: at least 12 months	Three general categories: Nonresponder: Nonresponse to 1 adequate trial of TCA, SSRI, MAOI, SNRI, or other AD, or ECT TRD: Resistance to 2 or more adequate AD trials of different classes • TRD1: 12–16 weeks • TRD2: 18–24 weeks • TRD3: 24–32 weeks • TRD4: 30–40 weeks • TRD5: 36 weeks–1 year CRD: Resistant to several AD trials, including augmentation strategy, for at least 12 months	The predictive value has not been systematically assessed, and reliability has not been tested All of the TRD stages are consistent with two treatment failures

Table 4. Staging models for treatment-resistent depression to define the spectrum of illness (continued)

Models Authors and Year of Publication	How Is Severity Scored?	How Is Failure Defined? Specify Current Episode?	Define Adequate Dose? Define Adequate Duration?	Staging Schema	Predictive Validity and Reliability Tested? Correspondence With ≥2 Treatment Failures Other Comments
Massachusetts General Hospital Staging model (MGH-s) Fava et al., 2003[45] Ruhé et al., 2012[44]	This tool provides points for the number of prior AD failures and considers other treatment-related factors to provide a continuous score. Higher scores indicate a greater degree of resistance to treatment	Failure to achieve remission (refers to "inadequate response" but defines it as HAM-D ≤7, which indicates remission) Not considered	Optimization per MGH or ATR Questionnaire ≥6 weeks	Stages: • Stage 1: Nonresponse to each adequate trial • Stage 2: Optimization of dose, duration, and augmentation/ combination • Stage 3: ECT increases overall score by 3 points Staging is primarily based on the number of AD medications used and gives a special weight for failure of treatment with ECT (i.e., score of 3)	A retrospective chart review showed an association between higher MGH-s score and worse outcome; reliability has not been studied In a retrospective comparative study, the MGH-s model better predicted nonremission compared with the TRSM. Reliability for these models was not reported. No direct correspondence with ≥2 treatment failures Considers both the number of failed trials and the intensity/optimization of each trial but does not make assumptions regarding a hierarchy of AD classes

Table 4. Staging models for treatment-resistent depression to define the spectrum of illness (continued)

Models Authors and Year of Publication	How Is Severity Scored?	How Is Failure Defined? Specify Current Episode?	Define Adequate Dose? Define Adequate Duration?	Staging Schema	Predictive Validity and Reliability Tested? Correspondence With ≥2 Treatment Failures Other Comments
Maudsley Staging Model (MSM) Fekadu et al., 2009[46] Ruhé et al., 2012[44]	Not dichotomous; gives points per number of prior attempts, duration, symptoms severity, augmentation use, ECT; single score can vary from 3 to 1	Failure to achieve remission (HAM-D_{21} ≤10) Yes	Maudsley Prescribing Guidelines for estimating minimum effective doses of ADs Augmenting agents: at least 6 weeks ECT: 8-session course Psychotherapy: unable to determine adequacy	Parameters include: • Duration (1–3 points) • Symptom severity (1–5) • Number of treatment failures (1–7) • Augmentation strategy use (0/1) • ECT use (0/1) Scores: • Mild (scores=3–6), • Moderate (scores=7–10) • Severe (scores=11–15)	Only tool with prospective testing showing good prospective validity. Reliability testing has not been reported. In two prospective studies, the MSM score predicted future nonresponse significantly better than the TRSM. The studies did not report reliability. No direct correspondence with ≥2 treatment failures

AD = antidepressant; ATHF = Antidepressant Treatment History Form; ATR = Antidepressant Treatment Response (Questionnaire); CRD = chronic resistant depression; ECT = electroconvulsive therapy: HAM-D = Hamilton Rating Scale for Depression; MAOI = monoamine oxidase inhibitors; MDD = major depressive disorder; MGH-s = Massachusetts General Hospital Staging; MSM = Maudsley Staging Model; SNRI = serotonin-norepinephrine reuptake inhibitors; SSRI = selective serotonin reuptake inhibitors; TCA = tricyclic antidepressant; TRD = treatment-resistant depression; TRSM = Thase and Rush staging model.

Two models are variations on TRSM. The European Staging Model[81] distinguishes between nonresponse, TRD, and chronic resistant depression. "Nonresponders" are patients who do not respond to one form of treatment; patients are considered treatment resistant after they have a poor response to a second trial with a different class of antidepressant. Further staging depends on the duration of treatment with adequate medication trials. TRD staging ranges from 1 to 5, with resistance beyond 12 months indicating a distinct category—chronic resistant depression. No studies tested the reliability or predictive utility or reliability of this model.

Another model related to the TRSM is the Massachusetts General Hospital Staging model (MGH-s). It is a function primarily of the number of prior antidepressant failures (with no hierarchy of antidepressant classes).[45] It also considers optimization of treatments, augmentation and combination strategies, and prior failed ECT. Patients receive certain points for these components. The MGH-s produces a continuous score, reflecting the level of treatment resistance. A retrospective chart review showed an association between a higher MGH-s score and worse outcome.[44] No study has assessed reliability.

The fifth staging model, the Maudsley Staging Method (MSM),[46] summarizes the TRD stage into a single score, ranging from 3 to 15. Like the other models, MSM considers the number of treatment failures, and it considers augmentation strategies and ECT treatment (as does the MGH-s). Unlike the TRSM, European Staging Model, and MGH-s, however, the MSM includes. two disease characteristics, namely, duration and symptom severity at baseline. This model is the only one with validity assessed with prospective data.[46] Higher scores for patients were associated with failure to achieve remission; the model correctly predicted treatment resistance in more than 85 percent of cases. Reliability testing has not been reported.

Predictive Validity of Staging Models

A recent systematic review compared the predictive utility and reliability of these models. The review noted an evolution from single antidepressant adequacy ratings toward a multidimensional and more continuous, scored staging model that also introduced TRD characteristics (severity and duration). The operationalization criteria improved, and the scoring of different treatment strategies (between/within class switching and augmentation/combination) changed as evidence accumulated. Over time, efforts to validate models improved slightly.

The review identified six studies that had examined the predictive utility of four models: ATHF, TRSM, MGH-s, and MSM. Comparative predictive utility information existed for three models. In a retrospective study, the MGH-s model predicted nonremission better than the TRSM.[44, 82] The comparative predictive utility evidence has been best assessed for the MSM. In two prospective studies,[44, 46] the MSM score predicted nonresponse significantly better than the TRSM.

Overall, predictive validity has been assessed best for the MSM model. Still, the evidence base is limited, and the superiority of one model over any other model for use in a clinical setting is uncertain. A recent review of these methods, however, reported that they appear equally valid for documenting treatment failure in depressed patients.[41]

Consensus Definition of Treatment-Resistant Depression

Determining Consensus

A consensus can be identified in several ways. One approach is to have the strongest evidence base identifying the preferred definition. The limited evidence base available to us,

however, precludes this approach. A second way is to have a preponderance of best practice guidelines or consensus statements clearly identify a preferred definition. Most of the available guidelines and consensus statements, however, clearly stated that no widely acknowledged consensus exists about a preferred best definition of TRD, thus precluding this approach too.

A third way is to indicate the approach most frequently reported in the literature or by the guidelines or consensus statements. This approach appears the most feasible given the current state of the evidence. Below, we present the most frequently used definitions employed in systematic reviews addressing TRD and the most frequently reported definitions in guidelines and consensus statements.

Finding Consensus in Systematic Reviews

Of the eight systematic reviews since 2005 that directly addressed TRD (Table 5), four defined it as two or more previous treatment failures,[31, 41, 43, 58] and four defined it as one or more previous treatment failures.[33, 35, 36, 83] Notably, those systematic reviews considering more invasive or expensive interventions (or both), such as ECT, repetitive transcranial magnetic stimulation (rTMS), or other nonpharmacologic interventions, tended to use the cut-off of two or more failures. By contrast, less invasive interventions, such as medications or psychotherapy, were more likely to use a more stringent cut-off of one or more failures.

Table 5. Definitions of treatment-resistant depression by number of treatment failures and level of consensus: Systematic reviews as source

Number of Treatment Failures	Focus of Systematic Review, Author, Date	Definition of Treatment-Resistant Depression	Consensus of Definition Specifically Stated as Consensus or Was the Most Frequently Used Definition
1 or more	Lamotrigine augmentation, Thomas et al., 2010[32]	After at least one failed AD trial; patients had not responded to at least a 4-week course of a recommended dose of an AD	Not addressed
	Psychotherapy, Trivedi et al., 2011[33]	If patients reported partial or no remission following treatment with an adequate AD dose for ≥6 weeks	No consensus Noted the significant heterogeneity in the definition of TRD as well as in the measures used to determine MDD
	rTMS, Zhang et al., 2015[35]	Failure to respond to at least one course of adequate treatment for MDD during the current illness episode	Not addressed Included studies in the meta-analysis had TRD definitions that ranged from failed one or more to four ADs with a duration ranging from 4 to 6 weeks
	Predictors of nonresponse, De Carlo et al., 2016[36]	Failure to respond/remit after at least one AD treatment in subjects with a primary diagnosis of MDD; noted that lack of efficacy of the first AD reliably identifies TRD subjects	No consensus
	Olanzapine/fluoxetine combination, Luan et al., 2017[66]	Failure of one or more documented historical AD treatments	No consensus Noted the significant heterogeneity in the definition of TRD as well as in the characteristics of patients, duration of treatment, definition of response, and dosage of fluoxetine and olanzapine.

Table 5. Definitions of treatment-resistant depression by number of treatment failures and level of consensus: Systematic reviews as source (continued)

Number of Treatment Failures	Focus of Systematic Review, Author, Date	Definition of Treatment-Resistant Depression	Consensus of Definition Specifically Stated as Consensus or Was the Most Frequently Used Definition
2 or more	Pharmacologic treatments, Berlim et al., 2007[31]	Six different definitions of TRD were identified. Studies varied from one previous failed AD trial (n=5) to at least two previous trials with medications from different classes (n=8). Also, two studies explicitly included augmenting strategies in their definitions of TRD. The majority of studies did not specifically indicate which AD trials were considered in their definitions of TRD (i.e., those administered only during the current MDE or those given as part of previous episodes too) General consensus (55.3% of studies) was at least two adequate trials of different classes	Yes, a consensus: clinically significant TRD was defined as an episode of major depression that has not improved after at least two adequate trials of different classes of ADs, reflected by the majority of the retrieved studies using this definition (n=26 or 55.3%)
	Lithium or atypical antipsychotics, Edwards et al., 2013[39]	Patients with unipolar MDD who have not responded adequately to two or more AD drugs in the current episode	Authors acknowledge "there is much uncertainty regarding what constitutes the definition of TRD and whether or not, for example, a patient with a failure to respond to two antidepressants from the same class could be defined as treatment resistant"
	Nonpharmacologic interventions for TRD, Gaynes et al., 2011[58] rTMS, Gaynes et al., 2014[40]	An episode of MDD for patients who have not recovered following two or more adequate AD medication treatments (at least 4 weeks at an adequate dose per authors), regardless of the class of AD used or whether the treatment failures were required to be in the current episode	Yes, a consensus: two or more treatment failures in the current episode. Recovery is remission. Noted that TRD is a complex phenomenon that encompasses the number of treatment failures, the adequacy of prior treatments, depressive severity, comorbidities (both psychiatric and medical), symptom subtypes, and chronicity
	Nonpharmacologic treatments for TRD* Washington State Health Care Authority, 2014[41]	No standard definition of AD failure, but a variety of TRD staging tools provide different ways of assessing the adequacy of prior treatment so that the clinician can determine treatment failure	Majority of studies enrolled patients whose depression had had ≥ 2 AD prior treatment attempts, or reported that patients depression had experienced ≥ 2 AD failures, or reported that the mean number of prior AD failures was > 2
	Pharmacologic and somatic interventions, Papdimitropoulou et al., 2017[65]	Failure to respond to ≥2 AD treatment regimens prescribed at adequate dose and duration.	Interventions (and comparators) included in review: SSRIs, SNRIs, TCAs, tetracyclic antidepressants, MAOIs, atypical antidepressants, antipsychotics, OFC, adjunctive use of lithium, triiodothyronine (T3), lamotrigine, ketamine, ECT, and rTMS

*Relevant Systematic Review that did not meet quality criteria but was part of the MEDCAC meeting materials.

AD = antidepressant; ECT = electroconvulsive therapy; MAOI = monoamine oxidase inhibitor; MDD = major depressive disorder; MDE = major depressive episode; OFC = olanzapine/fluoxetine combination; RCT = randomized controlled trial; rTMS = repetitive transcranial magnetic stimulation; SSRI = selective serotonin reuptake inhibitor; SNRI = serotonin-norepinephrine reuptake inhibitor; TCA = tricyclic antidepressant; TRD = treatment-resistant depression.

Three of the systematic reviews identified a similar consensus definition of MDD. It involved patients depression who had not achieved remission following two or more adequate antidepressant medication treatments (at least 4 weeks at an adequate dose per authors).[31, 40, 41, 58] The requirement for failure to occur following medication from two different antidepressant *classes*, as opposed to merely two different *antidepressants*, varies by systematic review.

Finding Consensus in Guidelines or Other Materials

Of the 13 guidelines or consensus statements that directly addressed TRD, 8 defined it as two or more previous treatment failures,[23, 38, 42, 67, 69, 70, 74, 79] 1 defined is as a single failure,[68] and 1 defined it as three or more failures.[21] Details are in Table 6, which is ordered by the number of required treatment failures and then chronologically by source.

Three guidelines did not provide a definition. One referred to TRD but never defined it.[19] One considered stages of treatment resistance (rather than a dichotomous definition) based on the number of prior treatment failures. For example, Stage 1 indicates failure to achieve response after one course of adequate treatment, and Stage 2 indicates failure to achieve response after two courses of adequate treatment, and so on.[64] One concluded that the concept of TRD should not be used.[72] The latter guideline had previously used a dichotomous definition of two or more failures. However, the authors explained that because of the absence of evidence indicating a natural distinction between patients with one or two treatment failures and those without, as well as the pejorative nature of the term “treatment-resistant depression” for patients, they recommended a model addressing inadequate response by considering sequenced treatment options.

Summary of Consensus Findings

In summary, the majority of systematic reviews and guidelines or consensus statements reported that the most commonly used definition is patients whose depression does not remit following two or more treatment attempts of an adequate dose and duration. We found no agreement as to whether the treatment attempts require different classes of antidepressants. Similarly, the literature produces no agreement of how to define adequate dose and duration, although minimum duration tends to be cited as 4 weeks.

Table 6. Definitions of treatment-resistant depression and level of consensus by number of treatment failures: Guidelines and consensus statements as source

Number of Treatment Failures as a Consensus	Focus of Guideline or Consensus Statement, Author, Date	Definition of Treatment-Resistant Depression	Define Failure	Current Episode?	Define Adequate Dose?	Define Adequate Duration?	Consensus?	Specifically Stated or Most Frequently Used Definition?
1 or more	International Workshop on "Present and Future of TMS: Safety and Ethical Guideline" Rossi et al., 2009[68]	Patients with medication-refractory unipolar depression who failed one good (but not more than one) pharmacological trial	Not found	Not found	Not found	Not found	Yes	Report is from a consensus conference for TRD about when rTMS should be offered
2 or more	World Federation of Societies of Biological Psychiatry Guidelines for Unipolar Depression Bauer et al., 2009[67]	Patients who remain depressed and do not achieve adequate relief and a satisfactory level of functioning even after two or more adequate courses of treatment. Having failed to improve after two adequately performed trials of AD drug; these no-responders are considered "treatment resistant."	Patient is not showing any improvement after 4 weeks of treatment with an AD drug at an appropriate dose	Not found	Not found	At least 6 weeks, and 8 to 10 weeks	No	Reports the most commonly used definition. Notes that there is no clear consensus which strategy should be favored for the non-responding patient since to date no rigorous trial with a randomized, double-blind design has been conducted to answer this question.
	Unipolar Depression Guideline Harter et al., 2010[70]	In therapy-resistant depression (where pharmacotherapy has been administered adequately, with at least two drugs, one after the other, at a sufficiently high dosage and given for a long enough time interval), patients should be offered appropriate psychotherapy	Not found	Not found	Not found	Not found	Yes	Proposed a definition for when psychotherapy should be offered

Table 6. Definitions of treatment-resistant depression and level of consensus by number of treatment failures: Guidelines and consensus statements as source (continued)

Number of Treatment Failures as a Consensus	Focus of Guideline or Consensus Statement, Author, Date	Definition of Treatment-Resistant Depression	Define Failure	Current Episode?	Define Adequate Dose?	Define Adequate Duration?	Consensus?	Specifically Stated or Most Frequently Used Definition?
	Institute for Clinical and Economic Review (ICER) Coverage Policy Analysis, 2012[38]	Notes that definitions of so-called "treatment-resistant" depression vary; this generally refers to patients with persistent depression after attempted management with two or more medications	Failure to evoke a clinically significant and lasting response	Not found	Not found	Not found	No	Reports the most commonly used definition
	Report from European Medicines Agency consensus meeting in 2009—Improving outcomes in TRD Schlaepfer et al., 2012[69]	CHMP has stated that a patient is considered to be therapy resistant when consecutive treatment with two antidepressants of different classes (different mechanism of action), used for a sufficient length of time and at an adequate dose, fail to induce an acceptable effect. CHMP also notes that the definition of TRD itself is not always consistent between studies or treatment guidelines, and a clear definition would go some way to refining treatment options.	Not found	Not found	Not defined and consensus from the wider psychiatric community is still required	Not defined and consensus from the wider psychiatric community is still required	Yes	Cites the CHMP (EMA) definition. Some staging models have been used to define TRD, but further clinical validation is needed. In addition, true pharmacological resistance needs to be distinguished from resistance attributable to ongoing somatic or psychosocial problems

Table 6. Definitions of treatment-resistant depression and level of consensus by number of treatment failures: Guidelines and consensus statements as source (continued)

Number of Treatment Failures as a Consensus	Focus of Guideline or Consensus Statement, Author, Date	Definition of Treatment-Resistant Depression	Define Failure	Current Episode?	Define Adequate Dose?	Define Adequate Duration?	Consensus?	Specifically Stated or Most Frequently Used Definition?
	British Association for Psycho-pharmacology Guidelines for all depressive orders, revision based on 2012 consensus meeting Cleare et al., 2015[79]	Most describe it as a failure to respond to two or more adequate AD treatment trials	Lack of improvement (defined as at least a 20% to 30% reduction in HAM-D in different studies) at 4 and 6 weeks; only 20% and 10%, respectively, will go on to eventual response (≥50% improvement) at 8 weeks	Important to consider the current episode in the context of the overall history of depression and the nature of previous episodes when considering treatment options	Adequate treatment, defined as "recommended therapeutic dose"	4–8 weeks	No	Reports the most commonly used definition. Notes that problems arise in defining what comprises an adequate treatment trial, which drugs are to be included and in taking account of psychological treatments.
	VA/DoD Clinical Practice Guidelines for Management of MDD, 2016[23]	Lack of full response despite at least two adequate treatment trials	Lack of full response to an adequate treatment trial	Not found	Not found	Not found	Yes	Guideline says there is consensus

Table 6. Definitions of treatment-resistant depression and level of consensus by number of treatment failures: Guidelines and consensus statements as source (continued)

Number of Treatment Failures as a Consensus	Focus of Guideline or Consensus Statement, Author, Date	Definition of Treatment-Resistant Depression	Define Failure	Current Episode?	Define Adequate Dose?	Define Adequate Duration?	Consensus?	Specifically Stated or Most Frequently Used Definition?
	CANMAT Guidelines, 2016 Kennedy et al., 2016[74] Pharmacological treatments Psychological treatments Parikh et al., 2016[75] Neuro-stimulation treatments Milev et al., 2016[76] CAM treatments Ravindran et al., 2016[77] Special populations: youth, women, and the elderly MacQueen et al., 2016[78]	Notes that the most commonly employed definition is inadequate response to 2 or more AD drugs	Inadequate response (e.g., 25%-49% reduction in symptom scores) or no response (e.g., <25% reduction)	Not found	Not found	Not found	No	Reports the most commonly used definition Notes that the commonly applied definition does not take into account adjunctive strategies and does not differentiate between patients who have had partial response versus those who have had no response

Table 6. Definitions of treatment-resistant depression and level of consensus by number of treatment failures: Guidelines and consensus statements as source (continued)

Number of Treatment Failures as a Consensus	Focus of Guideline or Consensus Statement, Author, Date	Definition of Treatment-Resistant Depression	Define Failure	Current Episode?	Define Adequate Dose?	Define Adequate Duration?	Consensus?	Specifically Stated or Most Frequently Used Definition?
2 or more (MDD: ≥2; Bipolar I: ≥2 [specific treatments specified], Bipolar II: ≥2 [specific treatments specified])	Australian and New Zealand clinical practice guidelines for mood disorders Malhi et al., 2015[42]	MDD: Lack of improvement following adequate trials of two or more ADs Bipolar I depression: Failure to reach remission with adequately dosed lithium or to other adequate ongoing mood-stabilizing treatment, plus lamotrigine or with full-dose quetiapine as monotherapy Bipolar II depression: Failure to reach remission with adequately dosed lithium or other adequate ongoing mood-stabilizing treatment, plus lamotrigine or with full-dose quetiapine as monotherapy	Failure to reach remission with adequate dose	Not found	Not found	6 weeks of treatment	Ye	Provides several definitions depending on the underlying mood disorder

Table 6. Definitions of treatment-resistant depression and level of consensus by number of treatment failures: Guidelines and consensus statements as source (continued)

Number of Treatment Failures as a Consensus	Focus of Guideline or Consensus Statement, Author, Date	Definition of Treatment-Resistant Depression	Define Failure	Current Episode?	Define Adequate Dose?	Define Adequate Duration?	Consensus?	Specifically Stated or Most Frequently Used Definition?
3 or more, 3 different classes	ICSI, Guidelines for adult depression in primary care Trangle et al., 2016[21]	Defines true treatment resistance as failure to achieve remission with an adequate trial of therapy and three different classes of AD drugs at adequate duration and dosage	Failure to achieve remission	Not found	Not found	Not found	No	Identifies a definition for primary care clinicians "True treatment resistance is seen as occurring on a continuum, from failure to reach remission after an adequate trial of a single [AD drug] to failure to achieve remission despite several trials of [AD drugs] augmentation strategies, ECT and psychotherapy."
Did not specifically address	American Psychiatric Association guideline for the treatment of MDD Gelenberg et al., 2010[19]	Frequently uses the term "treatment-resistant" but never defines it; refers to "next steps" in treatment	Not found	For rTMS, FDA says individuals with MDD who have not had a satisfactory response to at least one AD trial in the current episode of illness	Not found	Not addressed for TRD; generally, adequate treatment with an AD medication for at least 4–6 weeks For psycho-therapy, a few months	No	Not found

Table 6. Definitions of treatment-resistant depression and level of consensus by number of treatment failures: Guidelines and consensus statements as source (continued)

Number of Treatment Failures as a Consensus	Focus of Guideline or Consensus Statement, Author, Date	Definition of Treatment-Resistant Depression	Define Failure	Current Episode?	Define Adequate Dose?	Define Adequate Duration?	Consensus?	Specifically Stated or Most Frequently Used Definition?
	NICE Depression Guidance, 2009[72] NICE VNS Guidance, 2009[73] NICE rTMS Guidance 2015[71] Various guidelines concerning depression, use of vagal nerve stimulation, and use of rTMS	Did not define Earlier NICE guidelines had referred to TRD defined as depression that had not responded adequately to two courses of AD drugs (of adequate dose and length) The current guideline groups preferred to approach the problem of inadequate response by considering sequenced treatment options rather than by a category of patient	Does not clearly define; refers to Inadequate response, which could reflect both lack of response and lack of remission, and considers both patient and clinician perspectives	Not found	Not found	Not found	No. NICE eschews use of the term "treatment-resistant depression"	Previous versions of the guidelines stated that the most commonly used definition was nonresponse to two or more adequate trials of ADs. Due to this definition not including psychotherapeutic treatment, non-antidepressant augmenting agents and psychosocial factors, the new guidelines use a sequence of treatment options rather than number of failures.

Table 6. Definitions of treatment-resistant depression and level of consensus by number of treatment failures: Guidelines and consensus statements as source (continued)

Number of Treatment Failures as a Consensus	Focus of Guideline or Consensus Statement, Author, Date	Definition of Treatment-Resistant Depression	Define Failure	Current Episode?	Define Adequate Dose?	Define Adequate Duration?	Consensus?	Specifically Stated or Most Frequently Used Definition?
	Unipolar depression Ontario Health Association, 2016[64]	Considers stages of treatment resistance (e.g., Stage 1 indicates failure to achieve response after one course of adequate treatment; Stage 2 indicates failure to achieve response after two courses of adequate treatment)	Cannot achieve remission	Cites FDA definition for rTMS: Treatment of adult patients with unipolar depression whose current episode did not respond to one adequate dose of AD medication	Not found	Long enough to take effect, at least 2 weeks or 10 sessions	No	Does not identify the most commonly used definition Notes that definition of adequate response ranges from failure to achieve response to failure to achieve full symptom remission and that most experts agree that inadequate response is the failure to achieve full symptom remission

AD = antidepressant; CAM = complementary and alternative medicine; CANMAT = Canadian Network for Mood and Anxiety Treatments; CHMP = Committee for Medicinal Products for Human Use; ECT = electroconvulsive therapy; EMA = European Medicines Agency; FDA = Food and Drug Administration; HAM-D = Hamilton Depression Rating Scale; ICER = Institute for Clinical and Economic Review; ICSI = Institute for Clinical Systems Improvement; MDD = major depressive disorder; NICE = National Institute for Health and Clinical Excellence; rTMS = repetitive transcranial magnetic stimulation; TMS = transcranial magnetic stimulation; TRD = treatment-resistant depression; VA/DoD = Department of Veterans Affairs and Department of Defense; VNS = vagus nerve stimulation.

Key Question 2: Diagnostic Tools to Identify Treatment-Resistant Depression in Clinical Research

Drawing from the same sources used in KQ 1, we address three questions below:

- What methods do investigators use to diagnose this condition in clinical research?
- What consensus, if any, exists about the best measure(s) to use?
- Does the setting of the medical visit influence the choices that investigators make about the diagnostic tool they use?

Key Points

1. Methods used to diagnose TRD in clinical research
   a. Emphasize careful, structured clinical confirmation of a current major depressive episode as a part of MDD or bipolar disorder based on the Diagnostic and Statistical Manual of Mental Disorders (DSM), International Classification of Diseases (ICD), or Research Diagnostic Criteria (RDC) criteria; this diagnosis could be through a standard clinical evaluation or a through a more-structured clinical assessment (Mini International Neuropsychiatric Interview, Structured Clinical Interview for the DSM, or Schedule for Affective Disorders and Schizophrenia)
   b. Further entail collecting a careful history before treatments (e.g., the number of prior pharmacologic attempts of adequate dose and duration that did not produce remission) or administering a structured, staging tool (Antidepressant Treatment History Form [ATHF], TRSM, MGH-s, or MSM) to confirm treatment resistance.
   c. Differ by how structured the assessment is (from standard clinical assessment to highly structured research tool); this factor affects feasibility
   d. Have a limited evidence base for validity and reliability
2. No consensus on or any preferred tool exists for making the diagnosis for various reasons:
   a. No validation of the standard clinical assessment of TRD (as compared with a more in-depth evaluation)
   b. Limited evidence base for structured tools
   c. Available tools (from standard clinical assessment to highly structured research tool) appear equally valid for diagnosing TRD
   d. No direct comparison of careful history with structured tool
   e. The clinical setting (e.g., psychiatric vs. primary care) has no influence on choice of diagnostic tool, although some issues of feasibility arise; no direct comparison exists of a careful history versus use of a structured tool.

Detailed Synthesis

Table 7 shows the variety of methods that clinical researchers have used to diagnose TRD. We relied for this analyses on all types of source publications. Some publications focused on TRD without clarifying how the investigators defined the condition. We discuss below the publications or materials that did provide some information about approaches to diagnosing TRD.

Table 7. Diagnostic approaches to treatment-resistant depression

Topic; Authors and Year of Publication Topic	Methods Used to Diagnose Treatment-Resistant Depression	Comments
Diagnosis and definition of TRD Fava et al., 2003[45]	Recommended using clinician-rated instruments, ideally a structured clinical interview Prospective assessment better than retrospective. No preferred specific instruments for diagnosing TRD prospectively. If retrospective, recommend either the clinician-rated ATHF, the clinician-rated HATH, or the self-rated ATRQ. Of these three instruments, only one, the ATHF, has been empirically validated via prospective treatment outcome reports.	Not a systematic review
SR of RCTs on "the meaning of TRD" Berlim et al., 2007[31]	Clinical confirmation by mental health professional using DSM, ICD, or RDC criteria; structured interviews involving MINI, SCID, or SADS Most studies did not describe how they confirmed the degree of treatment resistance; 10 of 46 studies (22%) described how they determined resistance, and 4 of 46 (9%) used a formal tool	Emphasis is on use for research setting
Bauer et al., 2009[67] World Federation of Societies of Biological Psychiatry Guidelines for Unipolar Depression	None listed; emphasizes thorough clinical assessment	Does not directly address TRD issues for these components
Maudsley Staging Method (MSM) Fekadu et al., 2009[46]	ICD-10 or DSM-IV, plus MSM to determine resistance	Notes that a key shortcoming of staging models is reliance on a single criterion, mainly treatment response. This is the explanation of the MSM.
Consensus Statement from the International Workshop on "Present and Future of TMS: Safety and Ethical Guideline," Siena, Italy Rossi et al., 2009[68]	None listed	Consensus statement about using rTMS
American Psychiatric Association guideline Gelenberg et al., 2010[19]	Does not directly address diagnosis of TRD; emphasizes diagnosis with thorough clinical examination based on standard guidelines (DSM)	A guideline for MDD, not TRD, but description of next step management approaches to depression overlaps some with TRD
Guidelines on unipolar depression Harter et al., 2010[70]	Does not directly address diagnosis of TRD; emphasizes diagnosis with thorough clinical examination based on standard guidelines (ICD)	None
SR on lamotrigine augmentation in MDD Thomas et al., 2010[32]	MDD diagnosed by DSM, ICD, or RDC criteria; did not specify how treatment resistance was determined, accepted what study authors reported as TRD	The SR included only one RCT. Subjects were inpatients who became outpatients.

Table 7. Diagnostic approaches to treatment-resistant depression (continued)

Topic; Authors and Year of Publication	Methods Used to Diagnose Treatment-Resistant Depression	Comments
SR on TRD Gaynes et al., 2011[58]; rTMS article from SR TRD Gaynes et al., 2014[40]	Clinical diagnosis of MDD with failure to achieve remission after 2 AD treatments; no specific diagnostic tools described	SR of psychopharmacologic and nonpsychopharmacologic interventions
SR of psychotherapy Trivedi et al., 2011[33]	Clinical assessment of MDD per DSM and HAM-D_{17} ≥14, or HAM-D_{17} ≥16, or MDD per Schedule for Affective Disorder, or BDI ≥15 + 1 failed AD trial (defined as HAM-D_{17} ≥14, or HAM-D17 ≥11, or HAM-D_{17} ≥8, or BDI ≥9; or MDD on Structured Clinical Interview for DSM or on Revised Clinical Interview Schedule)	Significant heterogeneity in the definition of TRD and the measures used to determine MDD Involved patients from both a psychiatric and a medical clinic, with an emphasis on providing information relevant to primary care patients
ICER Coverage Policy Analysis, 2012[38]	Clinical diagnosis per DSM of MDD that persists after two or more AD treatment attempts	Coverage policy compared rTMS with ECT
Report from European Medicines Agency consensus meeting in 2009—Improving outcome in TRD Schlaepfer et al., 2012[69]	Emphasizes clinical confirmation of MDD, describes a variety of TRD definitions, noting that CHMP (of the EMA) has stated that a patient is considered to be therapy resistant when consecutive treatment with two ADs of different classes (different mechanism of action), used for a sufficient length of time and at an adequate dose, fail to induce an acceptable effect	Mentions that often the first depression treatment is in primary care and the patient does not get to a psychiatrist until TRD Another important question is the definition of an adequate AD trial, defined as an appropriate drug given in a dosage and duration sufficient to produce a response
SR on staging methods for TRD Ruhé et al., 2012[44]	Five staging models were considered: ATHF, TRSM, ESM, MGH-s, and MSM.	Emphasizes the need for careful assessment to include assessing psychiatric comorbidity
SR on lithium or atypical antipsychotics in management of TRD Edwards et al., 2013[39]	Used DSM for clinical diagnoses and defined as failure to respond to two or more ADs in the current episode of depression	SR of lithium or antipsychotics
Review of literature on defining TRD Trevino et al., 2014[18]	A comprehensive diagnostic evaluation for MDD, possibly including a standardized measure, used before a patient is classified as having TRD	Authors distinguished between resistance and pseudo-resistance
Nonpharmacologic treatments for TRD Washington State Health Care Authority, 2014[41]	Emphasizes clinical assessment and confirmation of failure to remit Notes that several formal staging systems have been developed for systematically quantifying treatment resistance in terms of not only the number of prior failures but also whether previous treatment was adequate. These include the Antidepressant Treatment History Form (ATHF), the Maudsley Staging Method (MSM), the Massachusetts General Hospital Scale, and the Thase and Rush Scale.	Notes scores for TRD in staging models lower in primary care settings, suggesting a lower likelihood of TRD there SR of nonpharmacologic interventions for TRD (does not meet quality criteria)
Australian and New Zealand Clinical Practice Guidelines for Mood Disorders Malhi et al., 2015[42]	Implies clinical diagnosis using accepted diagnostic criteria (DSM, ICD) + failure to respond to at least one course of adequate treatment for MDD during the current illness episode	Implies importance of clinical diagnosis

Table 7. Diagnostic approaches to treatment-resistant depression (continued)

Topic; Authors and Year of Publication	Methods Used to Diagnose Treatment-Resistant Depression	Comments
British Association for Psychopharmacology Guidelines for all depressive orders, revision based on 2012 consensus meeting Cleare et al., 2015[79]	Implied clinical diagnosis using accepted diagnostic criteria (DSM, ICD), plus clinical assessment to determine degree of resistance	Guideline for psychopharmacology in MDD directly addresses TRD
NICE rTMS Guidance, 2015[71] NICE Depression Guidance, 2010[72] NICE VNS Guidance, 2009[73]	Emphasizes clinical assessment and monitoring	Does not support use of TRD term
SR and meta-analysis of use or TMS Zhang et al., 2015[35]	Diagnosis of adult MDD based on the DSM-IV, DSM-III, or DSM-III-R or the ICD-9 or ICD-10 criteria, plus either a HAM-D or MADRS score exceeding a threshold	None
CANMAT Guidelines, 2016: Kennedy et al., 2016[74] Pharmacological Treatments Parikh et al., 2016[75] Psychological Treatments Milev et al., 2016[76] Neurostimulation Treatments Ravindran et al., 2016[77] Complementary and Alternative Medicine Treatments MacQueen et al., 2016[78] Special Populations: Youth, Women, and the Elderly	Implied importance of clinical diagnosis by experts plus (most commonly) inadequate response to trials of two or more AD drugs	Did not directly address diagnosis for TRD
SR of predictors of nonresponse De Carlo et al., 2016[36]	A primary MDD diagnosis according to DSM or ICD criteria, plus a failure of at least one previous AD trial	The association between presence of psychiatric comorbidities and increased risk of TRD suggests the importance of a thorough psychiatric assessment to include comorbid anxiety, anxiety disorders, substance use disorders, and personality disorders
SR on rTMS for unipolar depression Ontario Health Association, 2016[64]	Implied clinical assessment of MDD, plus not achieving remission after at least one course of AD treatment	Compared rTMS and ECT
ICSI, Adult Depression in Primary Care (Guideline) Tranger et al., 2016[21]	Clinical assessment per DSM plus failure to achieve remission with an adequate trial of therapy and three different classes of AD drugs at adequate duration and dose	Referral or co-management with mental health specialty clinician if patient has inadequate treatment response; inadequate treatment not defined
Clinical practice guidelines for management of MDD VA/DoD, 2016[23]	Emphasized clinical assessment plus at least two adequate treatment trials and lack of full response to each	Consider referral to mental health specialist if more severe symptoms or if no remission after 8 to 12 weeks of second treatment using a first-line AD drug

Table 7. Diagnostic approaches to treatment-resistant depression (continued)

Topic; Authors and Year of Publication	Methods Used to Diagnose Treatment-Resistant Depression	Comments
SR and MA of olanzapine/fluoxetine combination Luan et al., 2017 (#2603)	Accepted individual study authors determination that patients were diagnosed with TRD	None
SR of pharmacologic and somatic interventions Papadimitropoulou et al., 2017[65]	Did not clarify any tool; required adult MDD patient who failed to respond to ≥2 AD treatment regimens prescribed at adequate dose and duration	Used network meta-analysis

AD = antidepressant; ATHF = Antidepressant Treatment History Form; ATRQ = Antidepressant Treatment Response Questionnaire; BDI = Beck Depression Inventory; CANMAT = Canadian Network for Mood and Anxiety Treatments; CER = comparative effectiveness review; CHMP = Committee for Medicinal Products for Human Use; DSM = Diagnostic and Statistical Manual of Mental Disorders; ECT = electroconvulsive therapy; EMA = European Medicines Agency; ESM = European Staging Model; HAM-D = Hamilton Rating Scale for Depression; HATH = Harvard Antidepressant Treatment History; ICD = International Classification of Diseases; ICER = Institute for Clinical and Economic Review; MA = meta-analysis; MADRS = Montgomery-Åsberg Depression Rating Scale; MDD = major depressive disorder; MGH-s = Massachusetts General Hospital Staging model; MINI = Mini International Neuropsychiatric Interview; MSM = Maudsley Staging Model; NICE = National Institute of Health and Clinical Evidence; RCT = randomized controlled trial; RDC = Research Diagnostic Criteria; rTMS = repetitive transcranial magnetic stimulation; SADS = Schedule for Affective Disorders and Schizophrenia;; SCID = Structured Clinical Interview for the DSM; SR = systematic review; TRD = treatment-resistant depression; TRSM = Thase and Rush Staging Model; VA/DoD = Department of Veterans Affairs and Department of Defense.

Methods Used to Diagnose Treatment-Resistant Depression in Clinical Research

Diagnosing TRD is a three-step process: (1) confirmation of MDD (or, less commonly, a bipolar disorder diagnosis), (2) subsequent determination of a degree of resistance meeting threshold criteria for TRD definition, and (3) confirmation that the patient is currently depressed. The tools used to diagnose TRD in clinical research involve the same approaches as outlined for KQ 1 (see Table 7).

For confirmation of an MDD, the literature emphasizes careful, structured clinical assessment and diagnosis of MDD. Nearly all reviews used a clinical confirmation based on widely accepted diagnostic criteria (DSM, ICD, or RDC) or a structured diagnostic assessment (Mini International Neuropsychiatric Interview, Structured Clinical Interview for the DSM, or Schedule for Affective Disorders and Schizophrenia). A single review addressing psychotherapy for TRD accepted a Hamilton Rating Scale for Depression-17 item version ($HAM\text{-}D_{17}$) ≥16 or a Beck Depression Inventory ≥9 for a diagnosis.[33] All guidelines and consensus statements described clinical confirmation using diagnostic criteria.

Determining whether the depressive illness met criteria for TRD involved two main steps. The first entails collecting a careful history before treatments (e.g., the number of prior pharmacologic attempts of adequate dose and duration that did not produce remission). The second or alternative tactic involves administering a structured, staging tool (ATHF, TRSM, MGH-s, or MSM) to assess the spectrum of resistance. The systematic and nonsystematic reviews generally did not describe or did not use formal instruments to clarify treatment resistance, although some reviews recommended using a structured instrument to confirm.[44, 45]

Confirming that a patient was currently depressed in systematic or nonsystematic reviews involved primarily scoring above a particular threshold on a validated depression monitoring tool, such as the HAM-D or the Montgomery-Åsberg Depression Rating Scale (MADRS). In consensus statements or guidelines, clinical confirmation of a current depressive episode was based on standard clinical assessments.

As noted in KQ 1, the evidence base for validity and reliability of these *diagnostic* tools is limited.

Consensus on Best Measure for Diagnosing Treatment-Resistant Depression in Clinical Research

No consensus exists on the best measure for diagnosing TRD; the limited evidence does not provide much guidance. As noted in KQ 1, the MSM has evidence supporting its prospective predictive validity in TRD populations in general;[46] the ATHF has evidence of prospective predictive validity but only in populations having received ECT.[44] The MGH-s has retrospective chart review evidence of an association between higher resistance and worse outcomes.[44]

A systematic review of the staging models described for KQ 1 noted that, despite validation of the MSM, further investigation of the reliability and predictive utility of TRD staging models and additional disease characteristics is required.[44] Correct staging of TRD might improve generalizability of results from clinical studies and improve delivery of care to TRD patients. Limitations of the current validation studies are small sample sizes, the use of chart review methodology, and the potential nongeneralizability of their findings to less severe or outpatient populations.[44]

At present, no staging model for TRD seems to be both applicable in clinical practice and valid for research purposes; using such models would facilitate the generalization of research findings from studies addressing TRD and next-step strategies. An effective staging model must be user-friendly for busy clinicians, clear, and able to predict the likelihood of remission in an objective manner. Such a model would help clinicians plan treatment, inform their patients adequately, and judge the merits of new therapies for TRD.

Influence of Setting on Choice of Diagnostic Tool

The available evidence does not indicate a preferred or recommended tool for diagnosing TRD in primary care settings. One systematic review looking at the use of psychotherapy in TRD considered trials in both psychiatric and primary care settings, with an aim of providing evidence for primary care clinicians and patients.[33] The authors did not, however, distinguish between diagnosing TRD in those two settings. A second TRD systematic review noted that scores for TRD in staging models were lower in primary care settings, suggesting a lower likelihood of TRD there.[41]

Two guidelines addressed management of TRD in primary care settings.[21, 69] Both indicated that TRD can be confirmed in primary care or mental health settings but that mental health settings may be indicated for subsequent TRD management.

Key Question 3: Success or Failure of Treatment in Clinical Studies of Treatment-Resistant Depression

This is a complex question for this chapter. It entails synthesizing information on the measures that researchers might use to determine the success and failure of treatment for TRD patients. Of particular interest is assessing the severity of depression. Among the questions are whether experts agree about such measures, whether they are reported by patients or clinicians (typically, physicians), and whether we can document their psychometric properties and find information about minimally significant clinical differences. Also of concern is whether and, if

so, how well such measures describe a wide array of clinical, quality-of-life, or behavioral variables.

Key Points

1. No consensus exists regarding the best measures to use.
   a. Tools have not been created specifically for TRD measurement.
   b. Both patient-reported and clinician-administered measures are available for each category of outcomes; we found no stated preference for one type over the other, although patient-reported tools are more feasible to use.
   c. The most commonly reported depression-specific measure is the HAM-D. Although we detected no strong consensus on a preferred instrument, experts appear to agree that the preferred outcome is remission (complete recovery as measured by a score below a threshold) using a standardized and validated measure (regardless of the tool).
   d. General psychiatric status measures were infrequently described; most commonly reported was the Clinical Global Impression (CGI).
   e. Various functionality scales have been reported, but none is the most commonly used.
2. Most measures have adequate psychometric properties.
   a. All depressive-specific measures have been validated and have acceptable psychometric properties.
   b. For general psychiatric measures, the CGI has clinical utility and has been validated, but disagreement exists about its degree of validity.
   c. Functional impairment tools have been validated and have acceptable psychometric properties.
3. The minimum significant clinical difference (minimally clinically important difference, or MCID) has been defined for many of these measures. Experts disagree about which measure is preferred and, for a specific measure, which difference best accounts for a minimum clinically significant change.

Detailed Synthesis

Tables 8, 9, and 10 show the variety of depressive-specific, general psychiatric, and functioning or quality-of-life measures, respectively, described in our eligible sources. This review is not meant to be comprehensive, and it reflects what reviews generally report about the instruments. For example, we did not assess the evidence base for the range of scores reported on particular instruments (e.g., what is a mild vs. moderate vs. severe depressive severity) or the quality of the evidence base for the instruments' psychometric characteristics. Rather, we list what prior summaries or articles have reported. Below, we review the outcomes identified in our search for the three main issues for this question.

Assessing Depression (or Treatment-Resistant Depression) and Severity of Symptoms

Table 8 documents several key variables or descriptors for assessing depression, including severity, in TRD patients. (The distinction here is with KQ 2, which is concerned more with diagnosis per se.) In all, we identified seven different types. These methods are grouped, first, as all patient-reported measures (e.g., Beck Depression Inventory [BDI], with two versions, through

Table 8. Measures to test depressive severity in treatment-resistant depression

Brief Description	Physician or Patient Reported Scale Scores	Psychometric Properties	Minimally Important Clinical Differences	Comments
BDI Beck Depression Inventory Assesses severity of depressive symptoms; not appropriate for diagnosis 21 items Two major revisions since 1961 origin: BDI-IA in 1979 and the BDI-II, in 1996, which replaced 4 items to align with DSM-IV MDD[84]	Patient reported BDI-II Scale: • 0–13: no depression • 14–19: mild depression • 20–28: moderate depression • 29–63: severe depression[85, 86]	BDI: Convergent validity—with clinical ratings: 0.72; with HAM-D: 0.73[87] BDI (based on 4 studies) • Sensitivity: 0.85 (95% CI, 0.79 to 0.90) • Specificity: 0.83 (95% CI, 0.70 to 0.91)[72] BDI-II: Internal consistency: 0.9 (range 0.83 to 0.96)[85] BDI-II: Retest reliability: ranged from 0.73 to 0.96[85]	BDI-II: Change of >5 is clinically significant, although smaller changes should be considered for MCID[88] (anchor based) BDI-II sensitive to change in depression in cross-cultural studies (anchor based): • 5-point = MCID • 10–19 points = moderate change • ≥20 points to a large change[86]	High reliability, capacity to discriminate between depressed and nondepressed subjects Improved concurrent, content, and structural validity BDI-II: endorsed by NICE for use in primary care for measuring baseline depression severity and responsiveness to treatment[89] BDI and MADRS similar in differentiating between different Axis-I diagnoses and sensitivity to change during antidepressive treatment (MADRS-S focuses on core depressive symptoms, whereas BDI is more sensitive to maladaptive personality traits)[90] BDI and MADRS highly intercorrelated (*r*=0.869), note tested in samples of psychiatric patients with prominent psychiatric symptomatology and has not been evaluated in those with milder symptoms[90]
CES-D Center for Epidemiological Study—Depression Scale Assesses severity of depressive symptoms in community populations; not appropriate for diagnosis Developed by NIMH in 1977 20-item tool most commonly used, although other versions range from 4 to 16 items	Patient self-report Score: Scores range from 0 to 60, with high scores indicating greater depressive symptoms • ≥16 is cutoff for clinical depression[91] • Optimal cutoff score for clinically relevant depression: 22[92]	Based on 8 studies: Sensitivity: 0.84 (95% CI, 0.78 to 0.89) Specificity: 0.74 (95% CI, 0.65 to 0.81)[72] High internal consistency: coefficient α: range from 0.85 in general population to 0.90 in a psychiatric population[84] Moderate correlation (0.49) between CES-D and clinical interview ratings of depression[93]	None reported[94]	None

Table 8. Measures to test depressive severity in treatment-resistant depression (continued)

Brief Description	Physician or Patient Reported Scale Scores	Psychometric Properties	Minimally Important Clinical Differences	Comments
GDS Geriatric Depression Scale Screening test for depression in elderly patients Long version: 30 items; short version: 15 items	Patient self-report Score: • 0: no depression • 30: severe depression for long form, 15 for short form Scores • 1–10 normal • ≥11 possible depression or ≥14 avoids false-positives[95] GDS website: • 0–9 = normal, • 10–19 = mild depression, • 20–30 = severe depression Short form: • >5 suggests depression • >10 indicates highly likely depression Other studies of medical patients suggest cutoffs at 5–7[96-100]	High internal consistency (reliability): Cronbach's α: 0.94[101] but others have reported 0.87 for long form and 9.76 for short form[102] From multiple studies: • Long form: range from 0.69–0.99[103] • Short form: 0.74–0.86[100, 102] Short form: Using a cutoff score of 6 differentiates between depressed and nondepressed elderly primary care patients: • Sensitivity = 0.81 • Specificity = 0.75[100] Based on analysis of 11 studies: • Sensitivity = 0.87 (95% CI, 0.80 to 0.91) • Specificity = 0.75 (95% CI, 0.69 to 0.80)[72] Meta-analysis of primary care patients for diagnostic accuracy: Long form: • Sensitivity = 77.4% • Specificity = 65.4% Short form: • Sensitivity = 81.3% • Specificity = 78.4% Percentage correctly identified: • GDS30: 71.2% • GDS15: 77.6% a significant difference (chi^2= 24.8; p<0.0001)[104] Correlation between GDS and HAM-D: 0.83[84]	None reported[94]	Distinguishes symptoms of depression and dementia

Table 8. Measures to test depressive severity in treatment-resistant depression (continued)

Brief Description	Physician or Patient Reported Scale Scores	Psychometric Properties	Minimally Important Clinical Differences	Comments
HAM-D Hamilton Rating Scale for Depression Assesses symptom severity; Not designed as a diagnostic instrument but is used to assess efficacy of treatment Multiple versions: Originally 21 items; Reduced to 17 when 4 items dropped owing to lack of construct validity. 17-item version most commonly used (maximum score is 54) Subsequently expanded to 24 or 28 items	Clinician rating Reported ranges of severity vary. For 17-item version: • 0–6: no depression • 7–17: mild depression • 18–24: moderate depression • 24: severe depression[105] or: • <8: none • 8–13: mild • 14–19: moderate • 20–25: severe • >25: very severe Score ranges for other versions of the HAM-D: • HAM-D_{21}:0 to 64 • HAM-D_{24}: 0 to 75[58]	Validity: HAM-D_{17} with MADRS[105] Convergent validity: Adequate Discriminant validity: Adequate.[106] HAM-D score significantly correlated with each of the 8 SF-36 subscales[107] Good internal, interrater, and retest reliability estimates for the overall scale across many studies, but weak interrater and retest coefficients at the item level. Convergent, discriminant, and predictive validity were good (other studies found coefficients ranging from 0.83 to 0.94)[108] Interrater reliability of the scale proved consistent, exceeding 0.85[109]	In MDD population, MCID (distribution based): • HAM-D_{17} of ≥27.1% ± 25.7% change • HAM-D_{21} of ≥27% ± 25.1% change • HAM-D_{24} of 28% ± 25.2% change[110]	Requires a trained rater with sufficient knowledge of the instrument and the symptoms of the depressive syndrome[108, 111] Main limitations of HAM-D_{17} include: • failure to include all symptom domains of MDD • presence of items measuring different constructs uneven weight attributed to different symptom domains[105]
MADRS Montgomery-Åsberg Depression Rating Scale 10 items	Clinician reported Score • ≤10 = no depression (or remission) • >30 (or sometimes 35) = severe depression[105] MADRS ≤ 5 equals complete or symptom-free remission (CGI-S = 1) MADRS ≤11 equals remission (CGI-S <2)[94]	Very high internal consistency High correlation with HAM-D Did not differ according to sensitivity to change during antidepressant treatment	MCID: Ranges from 1.6 to 1.9 (distribution based, using the standard error of measurement)[112]	A HAM-D_{17} score of 7 corresponded to an 8 or 9 on the MADRS; the MADRS is said to be superior to the HAM-D_{17} in the conduct of clinical trials[113]

Table 8. Measures to test depressive severity in treatment-resistant depression (continued)

Brief Description	Physician or Patient Reported Scale Scores	Psychometric Properties	Minimally Important Clinical Differences	Comments
PHQ, PHQ-9 Patient Health Questionnaire Screening tool, by itself not appropriate for diagnosis Based on PRIME-MD and DSM MDD diagnosis Used in a primary care setting, designed both to indicate the likelihood of a depressive episode and to characterize the severity of depression	Patient reported Scale: 1–4: minimal depression 5–9: mild depression 10–14: moderate depression 15–19: moderately severe depression 20–27: severe depression[114]	From 11 studies, threshold of ≥10: Sensitivity: 0.82 (95% CI, 0.77 to 0.86) Specificity: 0.83 (95% CI, 0.76 to 0.88)[72, 115] Validity: Comparable to the larger, clinician-administered screening instrument PRIME-MD[116] Internal consistency: Cronbach's α: 0.89 and 0.86[117] Based on 11 studies: Sensitivity: 0.82 (95% CI, 0.77 to 0.86) Specificity: 0.83 (95% CI, 0.76 to 0.88)[72] In a sample of 7% prevalence of major depression, interviews with mental health providers demonstrated a positive predictive value ranging from 31% for a PHQ-9 cutoff of 9 to 51% for a cutoff of 15[117]	Individual change: estimated as 2 standard errors of measurement—5 points on the 0 to 27 point PHQ-9 scale[118] (distribution based)	Endorsed by the VA and NICE

Table 8. Measures to test depressive severity in treatment-resistant depression (continued)

Brief Description	Physician or Patient Reported Scale Scores	Psychometric Properties	Minimally Important Clinical Differences	Comments
QIDS-CR16, QIDS-SR16 Quick Inventory of Depressive Symptomatology Based on DSM MDD diagnosis	Clinician rated Patient self-report Score: • 1–5: no depression • 6–10: mild • 11–15: moderate • 16–20: severe • 21–27: very severe	Limited study of sensitivity/specificity; ≥13 (in primary care medical patients): Sensitivity: 76.5% Specificity: 81.8%[119] Cronbach's α: • SR: 0.69 to 0.89 • CR: 0.65 to 0.87[120] Correlated moderately to highly with several depression severity scales: Of 7 pooled studies, QIDS-SR16: correlated with the HAM-D_{17} (r = 0.76, CI 0.69 to 0.81) Convergent validity with the QIDS-CR16[120]	PGI-I minimally improved = QIDS-SR of ≥28.5% ± 28.7% change[110] (distribution-based)	None

BDI = Beck Depression Inventory; CES-D = Center for Epidemiological Study—Depression Scale; CGI = Clinical Global Impression; CI = confidence interval; DSM = Diagnostic and Statistical Manual; GDS = Geriatric Depression Scale; HAM-D = Hamilton Rating Scale For Depression; MADRS = Montgomery-Åsberg Depression Rating Scale; MCID = minimally clinically important difference; MDD = major depressive disorder; NICE = National Institute of Health and Care Excellence (UK); NIMH = National Institute of Mental Health; PGI-I = Patient Global Impression of Improvement; PHQ = Patient Health Questionnaire; PRIME-MD = Primary Care Evaluation of Mental Disorders; QIDS = Quick Inventory of Depressive Symptomatology; VA = Department of Veterans Affairs.

the Patient Health Questionnaire. These are followed by measures that are either clinician-reported (i.e., HAM-D and the MADRS) or can be both clinician reported and patient self-report (Quick Inventory of Depressive Symptomatology [QIDS-CR and QIDS-SR, respectively]).

More questionnaires or measures are patient self-report than otherwise. Several of these have more than one version (typically a traditional "long" version and one or more versions with fewer items). For example, the clinician-reported HAM-D has five versions that differ by numbers of items (17-, 21-, 24-, 25-, and 28 -item versions), of which the HAM-D_{17} is generally the most frequently applied.

Some have a history dating back two decades or more and thus have long histories of use for several purposes but usually not for making a definitive clinical diagnosis. Some assess depression severity, some are used as screening tools, some are based on a DSM definition (usually DSM-IV); and some are described explicitly as not intended to be diagnostic. All are suitable for use in trials or nonexperimental studies for the purposes intended (measuring severity, evaluating outcomes).

Information describing the scales and their psychometric properties is generally reported in the literature in some detail, although such data are often spread over multiple articles. Less is known about MCIDs; when such data have been reported at all, different investigators based their MCIDs on different methods (i.e., anchor based or distribution based). Generally, the reliability and validity statistics (i.e., psychometric properties) for most of these measures are sufficient for group comparisons; some are sufficiently high to permit comparisons of individuals. In the literature we reviewed, no psychometric assessments were assessed in a TRD population.

Although we identified a variety of potential tools for assessing severity or outcomes, investigators and other experts generally agreed that remission was the treatment goal. However this might be measured by a standardized and validated tool, this was the clear preferred outcome.

Assessing General Psychiatric Illness and Severity

Two other questionnaires—the CGI and the Brief Psychiatric Rating Scale (BPRS) —qualify as appropriate and adequate measures of the presence or severity of psychiatric illness broadly defined. Both are clinician rated. Table 7 provides information on these, although neither can be said to be "better" than the other. The CGI has two versions: CGI-I relates to improvement (reflecting change in status from an initial assessment), and the CGI-S reflects the severity of the psychiatric condition (reflecting status at each assessment). Both have acceptable psychometric properties. The BPRS can also be used to evaluate suicidality. Scores from the BPRS have been calibrated against those from the CGI, presumably allowing some cross-walk between studies using one or the other.

Table 9. Two measures to assess general psychiatric illness severity

Brief Description	Physician or Patient Reported Scale Scores	Psychometric Properties	Minimally Important Clinical Differences
BPRS Brief Psychiatric Rating Scale	Clinician rated Score: Depressive mood: 1: not present 2: very mild 3: mild 4: moderate 5: moderately severe 6. severe 7: extremely severe Same scores for suicidality	Adequate psychometric properties[121, 490-4] CGI approximately corresponded to BPRS total score • Mildly ill: 31 • Moderately ill: 41 • Markedly ill: 53 Minimally improved score of 53, CGI score associated with BPRS reductions of 24, 27% and percentage BPRS reductions of 24, 27, and 30% at weeks 1, 2, and 4, respectively. Corresponding numbers for a CGI rating of "much improved": 44, 53, and 58%[122]	CGI-I of "minimally improved" = BPRS reduction of 24% at week 1, 27% at week 2, and 30% at week 4 (anchor based) CGI rating of "much improved" = BPRS reduction of 44% (week 1), 53% (week 2), and 58% (week 4) (anchor based)[122]
CGI Clinical Global Impression Scale Assesses clinician's impression of patient's illness severity; used before and after treatment	Clinician rated Scores range from • 1 = not ill at all • to • 7 = among the most extremely ill[123]	Improvement ratings strongly related to both concurrent severity of illness and changes in severity of illness ratings from baseline Both CGI ratings positively correlated with both self-report and clinician-administered measures of social anxiety, depression, impairments, and quality of life.	CGI-I of "minimally improved" is considered the clinical gold standard for minimum clinically important difference

BPRS = Brief Psychiatric Rating Scale; CGI = Clinical Global Impression Scale; CGI-I = Clinical Global Impression Scale-Improvement;

Assessing Functional Impairment in Treatment-Resistant Depression

The last main concern for KQ 3 involves how to assess functional impairment in TRD patients. Functional impairment tends to be broadly defined: enjoyment of or satisfaction with life, impairments in a wide array of daily activities or relationships, several domains of health status (which can be collapsed into self-reports of physical or mental health), or overall levels of

disability. Of the five measures available for this task, four are self-reported and one is clinician reported.

The four patient-reported measures generally have appropriately strong psychometric properties, and one or more produce scores that are calibrated against scores of other measures (chiefly those for evaluating severity of depression). Only two, however, have reported MCIDs specifically for MDD (the Quality of Life Enjoyment and Satisfaction Questionnaire and the SDS).

For the clinician-rated tool, the DSM-IV had advised use of the Global Assessment of Functioning (GAF) scale assessing a patient's overall psychological functioning (Axis V). The American Psychiatric Association, however, discontinued use of the GAF in the DSM-5, and the American Psychiatric Association now suggests that clinicians use the World Health Organization Disability Assessment Schedule (WHODAS 2.0) as a measure of disability.[124]

Examining Minimal Clinically Important Differences

Most measures have adequate psychometric properties. All depressive-specific measures have been validated and have acceptable psychometric properties as reflected in measures of internal consistency, convergent validity, test-retest reliability, and correlation with other instruments to assess depressive severity (Table 6). The idea of an MCID is increasingly salient for addressing these issues. (It is also referred to as "clinically relevant," "clinically significant," or "minimum important difference.") The concept refers to the minimum change in a measurable outcome in which the patient or clinician perceives a difference because of a therapy or intervention. It has evolved as a practical means of giving clinical relevance to changes in standardized instrument scores when no gold standard of meaningful change exists.[125]

Two types of techniques, an anchor-based approach and a distribution-based method, have been used to calculate the MCID. Anchor-based methods use a measure with established or face-value clinical meaning such as the CGI-S to "anchor" scores on the measure of interest; distribution-based methods generally use the statistical characteristics of the sample such as the standard deviation to separate "signal" from "noise."[125]

In TRD, this concept applies most directly to depressive-specific measures. Although MCIDs have been defined for various depressive measures, no clear agreement exists about what constitutes a minimum clinically significant difference or how to determine it. For example, some users define such a difference as a numerical change in a score reflected by a standardized mean difference (e.g., HAM-D difference of 7 points).[126] Other researchers define it by how a change on a depressive measure scale (e.g., HAM-D) aligns with a clinician's perception of improvement (e.g., CGI).[126] Yet others suggest that a patient's perception of clinical improvement is the preferred standard.[127] Finally, some report this difference as a percentage change in score, with varying percentages reported as clinically meaningful.[94]

General psychiatric measures can also be used as a measure of minimum significant clinical difference (Table 7). The CGI-I, for example, has such an assessment built into its scale; a CGI-I rating of 3 indicates that the patient has "minimally improved," as contrasted with either "no change" or "moderately improved," This tool often functions as the gold standard against which other measures of a minimal degree of clinical improvement are compared. The Patient Global Impression of Improvement (PGI-I), a patient analogue to the CGI-I, has a similar range and minimal improvement score (e.g., PGI-I = 3 indicates "a little better"). The BPRS does not have a directly analogous score, but percentage of change in BPRS at a particular time (e.g., 24% change at week 1, 27% change at week 2, or 30% change at week 4) has been associated with a

CGI-I of 3.[122] The discrete psychometric properties of these instruments have been less studied, because they often are used as the clinical reference standard against which other measures are compared. Their accuracy is supported primarily by their consistency with each other.[122]

For functionality and quality-of-life measures, psychometric properties appear adequate (Table 9). We could not find reports of MCIDs for GAF, SF-36 Health Survey, or WHODAS tools. We did identify MCID measures for the SDS (~4 points for the overall score and 1 point for each individual item score)[128] and for the Q-LES-Q (mean percentage score change of 10.69).[129]

Table 10. Five measures to assess functional impairment in treatment-resistant depression

Brief Description	Physician or Patient Reported Scale Scores	Psychometric Properties	Minimally Important Clinical Differences
GAF Global Assessment of Functioning Scale	Clinician reported Scale: • 0–100 • ≤50 = severe symptoms and/or psychosocial dysfunction • 51–60 = moderate scores, • 61–70 = mild scores • ≥71 = absent or only transient symptoms and/or minimal dysfunction[130]	Not well validated[131] 1 validation study identified[132]	None identified for MDD
Q-LES-Q Quality of Life Enjoyment and Satisfaction Questionnaire Assesses the degree of enjoyment and satisfaction in the past week Regular form: 93 items Short form: 16 items (adults only)	Patient self-report Scale: 1 to 5 Maximum total score across 14 items for short form =70 Higher scores indicate greater enjoyment and satisfaction with specific items in instrument[133, 134]	Short form significantly correlated with the CGI-S[135] High test-retest reliability, sensitivity, and responsiveness: Item correlations with total score: 0.41–0.81[135] • Sensitivity: 80% • Specificity: 100% • Internal consistency and test-retest coefficients: 0.9 and 0.93[135] Remission scores of 6 on the MADRS correlate with the Q-LES-Q SF score of 58+/-10%[136]	A “minimally improved” assessment on the CGI-I scale = the mean % of the maximum Q-LES-Q short form score change of 10.69 overall (anchor based)[129]

Table 10. Five measures to assess functional impairment in treatment-resistant depression (continued)

Brief Description	Physician or Patient Reported Scale Scores	Psychometric Properties	Minimally Important Clinical Differences
SDS Sheehan Disability Scale Assesses functional impairment in work/school, social, and family life 3 items using a 10-point visual analog scale	Patient self-report Scale: • 0 = unimpaired • 1–3 = mild impairment • 4–6= moderate impairment • 7–9= marked impairment • 10= extreme impairment Maximum summed total measure = 30 highly impaired[133, 137] Patients who score 5 or greater on any of the 3 scales are associated with significant functional impairment[121]	Sensitivity: 83% Specificity: 69%[121]	~4 points for total score and 1–2 points for an item score (anchor based)[128]
SF-36 Health Survey[138] 8 scales that can be combined into two summary measures: physical health and mental health 36 generic, self-reported items organized into 8 domains	Patient self-report Each scale is directly transformed into a 0–100 scale, each question carrying equal weight. The higher the score the less disability (0= maximum disability and 100 = disability).	Validated Version 2.0 of the instrument was released in 1996 A shorter version, the 12-Item Short Form Health Survey (SF-12) was developed for the MOS, a multiyear study of patients with chronic conditions	None identified for MDD
WHODAS World Health Organization Disability Assessment Schedule Featured in the new DSM-5, for use in initial patient interview and for monitoring treatment progress. Not intended to be used as the sole basis for a diagnosis Adult self-administered version has 36 items; a shorter version has 12 items	Patient-self report	WHODAS 2 (12 item): Reported to be a reliable, valid, and useful tool for assessing overall disability in primary care patients with depression; showed adequate internal consistency ($\alpha = 0.89$) and construct validity because is significantly associated with quality of life and depression severity (convergent validity) and able to discriminate between patients on sick leave and those who are working[124]	None identified for MDD

CGI-I scale = Clinical Global Impression-Improvement; CGI-S = Clinical Global Impression-Severity; GAF = Global Assessment of Functioning Scale; MADRS = Montgomery-Åsberg Depression Rating Scale; MDD = major depressive disorder; MOS = Medical Outcomes Study; Q-LES-Q = Quality of Life Enjoyment and Satisfaction Questionnaire; Q-LES-Q SF = Quality of Life Enjoyment and Satisfaction Questionnaire Short form; SDS = Sheehan Disability Scale; SF-12 = 12-Item Short Form Health Survey; SF-36 = 36-Item Short Form Health Survey; WHODAS = World Health Organization Disability Assessment Schedule.

Key Question 4: Types of Research Designs to Study Treatment-Resistant Depression

Key Points

1. Most investigators and expert groups preferred randomized designs over nonexperimental ones as a means of minimizing bias.
2. Most of the available literature did not address, or apparently achieve consensus about, designs that might minimize placebo effects.
3. No consensus was seen about the appropriate or necessary length of trials or other studies of TRD. A study length of "at least 6 weeks" was often recommended; however, experts often noted that longer trials or studies would be preferable.
4. Studies also recommended using whole structured clinical interviews to diagnosis depression, because these full assessments could better confirm the MDD (or bipolar) diagnosis and clarify psychiatric comorbidity, seen as a key potential confounder in TRD treatment trials.
5. Getting patients to an adequate dose of a given medication may take a few weeks; for that reason, 6 weeks of adequate dosing may require a trial length longer than 6 weeks.
6. Compliance and consideration of prior psychotherapy use are important to assess and control for in analyses.

Detailed Synthesis

Types of Research Designs Used in Various Types of Studies or Projects

Table 11 records recommendations about research designs from numerous expert panels, groups doing systematic reviews (with or without meta-analyses), and research teams examining definitions of TRD or conducting trials or other studies of treatments for TRD. The table lists sources in chronological order.

Table 11. Summary of research design information

Topic; Authors and Year of Publication	Research Designs Used	Consensus on Study Design to Minimize Bias	Consensus on Study Design to Minimize Placebo Effect	Consensus on Length of Antidepressant Trials
Definition and epidemiology of TRD Fava, 1996[30]	RCTs (type unspecified)	Preferred randomized design	Did not address	Recommended study duration: at least 6 weeks
Sequential strategies for AD nonresponders Thase and Rush, 1997[5]	RCTs (double blind) RCTs (open label) RCTs (type unspecified) NRCTs Prospective cohort studies Retrospective cohort studies Cohort studies (type unspecified)	Preferred randomized design	Did not address	Recommended study duration: at least 4 weeks

Table 11. Summary of research design information (continued)

Topic; Authors and Year of Publication	Research Designs Used	Consensus on Study Design to Minimize Bias	Consensus on Study Design to Minimize Placebo Effect	Consensus on Length of Antidepressant Trials
Definition and meaning of TRD Sackeim, 2001[37]	Did not address	Preferred randomized design	Did not address	Noted that little consensus exists and that some suggest 8 weeks or longer
Diagnosis and definition of TRD Fava, 2003[45]	Refers to "studies" but does not specify types	Did not address	Did not address	Recommended study duration: at least 6 weeks, maybe longer
SR of RCTs—what is the meaning of TRD Berlim and Turecki, 2007[31]	RCTs (type unspecified)	Preferred randomized design Recommended use of prospective study designs and validated instruments	Did not address	Recommended study duration: at least 6 weeks Noted that 4 weeks was likely too short and that ideal duration could be even longer than 6 weeks
Maudsley Staging Method Fekadu et al., 2009[46]	RCTs (type unspecified) Retrospective cohort studies	Did not address	Did not address	Recommended study duration: at least 6 weeks
World Federation of Societies of Biological Psychiatry guidelines for unipolar depression Bauer et al., 2009[67]	RCTs (double blind) RCTs (open label) RCTs (type unspecified)	Preferred randomized design	Did not address	Recommended study duration: at least 6 weeks
NICE VNS Guidance, 2009[73] NICE Depression Guidance, 2010[72] NICE rTMS Guidance, 2015[71]	RCTs (double blind) RCTs (open label) RCTs (type unspecified) Prospective cohort studies SRs and meta-analyses	Preferred meta-analysis or randomized design	Did not address, although mentioned that some clinical trials had a placebo effect	Did not directly address TRD
Consensus Statement from the International Workshop on "Present and Future of TMS: Safety and Ethical Guideline", Siena, Italy Rossi et al., 2009[68]	RCTs (type unspecified) NOTE: Did not focus on specific study designs or other issues such as blinding or use of cohorts in this review	Preferred randomized design	Did not address	Did not address
American Psychiatric Association guideline MDD Gelenberg, 2010[19]	Double-blind RCTs, single-blind RCT, open-label RCT, trials—type not specified, prospective cohort, retrospective chart review, cohort not specified, case-control, meta-analysis, systematic review	Preferred randomized design	Did not address	Recommended 4–8 weeks, consistent with a duration of at least 6 weeks

Table 11. Summary of research design information (continued)

Topic; Authors and Year of Publication	Research Designs Used	Consensus on Study Design to Minimize Bias	Consensus on Study Design to Minimize Placebo Effect	Consensus on Length of Antidepressant Trials
Guidelines on treatment of unipolar depression Harter et al., 2010[70]	RCTs (type unspecified)	Preferred randomized design	Did not address	Recommended 4–8 weeks, consistent with a duration of at least 6 weeks
SR on use of lamotrigine in MDD Thomas, Nandhra, and Jayaraman, 2010[32]	RCTs (double blind) RCTs (single blind) RCTs (open label) RCTs (type unspecified) CRTs NRCTs Prospective cohort studies Retrospective cohort studies Cohort studies (type not specified) Case-control studies Interrupted time series	Preferred randomized design	Did not address	4 weeks
SR on TRD therapies Gaynes et al., 2011[58] Article on use of rTMS article for TRD Gaynes et al., 2014[40]	RCTs (double blind), RCTs (single blind), RCTs (open label), RCTs (type unspecified), CRTs NRCTs Prospective cohort studies, retrospective cohort studies Cohort studies (type unspecified) Case-control studies SRs and meta-analysis	Preferred meta-analysis or randomized design	Did not address	Did not address for medication
SR on psycho-therapy Trivedi, Nieuwsma, and Williams, 2011[33]	RCTs (double blind) RCTs (single blind) RCTs (open label) RCTs (type unspecified) CRTs	Preferred randomized design	Did not address	Recommended study duration should be at least 6 weeks
ICER Coverage Policy Analysis, 2012[38]	RCTs (double blind) RCTs (type unspecified) Prospective cohort studies	Preferred randomized design	Did not address	Did not address
SR on staging methods for TRD Ruhé et al., 2012[44]	Did not specify; included RCTs, prospective cohorts, and retrospective chart reviews	Preferred randomized design Recommended use of prospective study designs and validated instruments	Did not address	Recommended study duration: at least 6 weeks

Table 11. Summary of research design information (continued)

Topic; Authors and Year of Publication	Research Designs Used	Consensus on Study Design to Minimize Bias	Consensus on Study Design to Minimize Placebo Effect	Consensus on Length of Antidepressant Trials
Report from European Medicines Agency consensus meeting in 2009—Improving outcome in TRD Schlaepfer et al., 2012[69]	RCTs (double blind) RCTs (single blind) RCTs (type unspecified) Meta-analyses	Preferred meta-analyses or randomized design	Did not address	Noted that 4 to 6 weeks is considered to be an adequate trial period to see clinical response, although recent research suggests that longer periods (up to 8 or 12 weeks) may be needed to achieve remission
SR on use of lithium or atypical antipsychotics in managing TRD Edwards et al., 2013[39]	RCTs (double blind) RCTs (single blind) RCTs (open label) RCTs (type unspecified) CRTs	Preferred randomized design	Did not address	Recommended 4 weeks of treatment
Review of literature on definitions of TRD Trevino et al., 2014[18]	Did not clarify	Did not address	Did not address	Recommended study duration: at least 6 weeks
Nonpharma-cological treatments or TRD Washington State Health Care Authority, 2014[41]	RCTs (double blind) RCTs (single blind) RCTs (open label) RCTs (type unspecified) NRCTs Prospective cohort studies Retrospective cohort studies SRs and meta-analysis	Preferred meta-analyses or randomized design Adequate dosage of medication should be at maximum tolerated doses (according to prescription recommendations) Recommended that treatment compliance should be assessed	Did not address	Recommended study duration: at least 6 weeks Noted that standard AD trial of ≥6 weeks is most common
Australian and New Zealand clinical practice guidelines for mood disorders Malhi et al., 2015[42]	RCTs (double blind) RCTs (type unspecified)	Preferred randomized design	Did not address	Did not directly address TRD Noted that before altering any treatment, allowing a trial of appropriate duration, usually 3 weeks at adequate dosage, is important
British Association for Psychopharmacology Guidelines for all depressive orders, revision based on 2012 consensus meeting Cleare et al., 2015[79]	RCTs (type unspecified), SR, and meta-analyses	Preferred meta-analysis or randomized design	Did not address	Recommended study duration: at least 4 to 6 weeks of treatment

Table 11. Summary of research design information (continued)

Topic; Authors and Year of Publication	Research Designs Used	Consensus on Study Design to Minimize Bias	Consensus on Study Design to Minimize Placebo Effect	Consensus on Length of Antidepressant Trials
Predictors of response for use of rTMS Silverstein et al., 2015[34]	Studies of any design (i.e., experimental and observational) even when a study had no placebo comparison group	Did not address	Did not address	Did not address
SR and meta-analysis of use of rTMS Zhang et al., 2015[35]	RCTs (double blind) RCTs (single blind) RCTs (open label) RCTs (type unspecified) CRTs	Preferred randomized design	Did not address	Allowed study authors to define, but 8 of 10 studies used ≥ 6 weeks as definition of adequate length
CANMAT Guidelines, 2016 Kennedy et al., 2016[74] Pharmacological treatments Parikh et al., 2016[75] Psychological Treatments Milev et al., 2016[76] Neuro-stimulation treatments Ravindran et al., 2016[77] CAM treatments MacQueen et al., 2016[78] Special populations: youth, women, and the elderly	RCTs (open label) RCTs (type unspecified) SRs and meta-analyses	Preferred randomized design	Did not address	Did not directly address for TRD Implied that at least 4 weeks of adequate treatment are needed
De Carlo, Calati, and Serretti, 2016[36] Predictors of nonresponse	RCTs (double blind) RCTs (open label) RCTs (type unspecified) NRCTs Prospective cohort studies Retrospective cohort studies Cohort studies (type unspecified)	Preferred randomized design	Did not address	Did not address
SR on use of rTMS in unipolar depression Ontario Health Association, 2016[64]	RCTs (double blind) RCTs (single blind) RCTs (open label) RCTs (type unspecified) CRTs	Preferred randomized design	Did not address	Did not address

Table 11. Summary of research design information (continued)

Topic; Authors and Year of Publication	Research Designs Used	Consensus on Study Design to Minimize Bias	Consensus on Study Design to Minimize Placebo Effect	Consensus on Length of Antidepressant Trials
ICSI guideline on adult depression in primary care Trangle et al., 2016[21]	Placebo-controlled studies, other randomized controlled trials (types unspecified) open label, cohort (type unspecified), meta-analyses, systematic reviews	Preferred randomized design	Did not address	Did not address
Clinical practice guidelines for management of MDD VA/DoD, 2016[23]	RCTs (double blind) RCTs (type unspecified) Cohort studies (type not specified SRs and meta-analyses	Preferred meta-analyses or randomized design	Did not address	Recommended at least 4 to 6 weeks of treatment
SR and MA of olanzapine/ fluoxetine combination Luan et al., 2017[66]	RCTs (type unspecified)	Preferred randomized design	Did not address	Did not address
SR of pharmacologic and somatic interventions Papadimitro-poulou et al., 2017[65]	RCTs (triple blind) RCTs (double blind) RCTs (single blind) RCTs (type unspecified) SRs and meta-analyses	Preferred meta-analyses or randomized design	Did not address	Dependent upon intervention, ketamine shows superior efficacy at 2 weeks, others at 4, 6, and 8 weeks

AD = antidepressant; CAM = complementary and alternative medicine; CANMAT = Canadian Network for Mood and Anxiety Treatments; CRT = cluster randomized trial; ICSI = Institute for Clinical Systems Improvement; MA = meta-analysis;; MDD = major depressive disorder; NICE = National Institute of Health and Care Excellence; NRCT = nonrandomized controlled trial; RCT = randomized controlled trial; rTMS = repetitive Transcranial Magnetic Stimulation; SR = systematic review; TRD = treatment-resistant depression; VA/DoD = Department of Veterans Affairs and Department of Defense; VNS = vagus nerve stimulation.

As with other KQs, a couple of the sources in this table are not ones that met our standard inclusion and exclusion criteria specified in the Methods chapter, but we have included them because they were part of materials made available at the April 2016 MEDCAC meeting. In all, the evidence base for KQ 4 includes 36 sources: systematic reviews, nonsystematic reviews, clinical practice guidelines, and other statements from professional societies or regulatory agencies.

Consensus About Study Design to Minimize Bias and Placebo Effect

Researchers have used a considerable range of study designs with variably defined study components to examine treatments for TRD or assemble evidence about those interventions (see Table 10). In the midst of this diversity, we encountered a consistent recommendation that the standard use of operationalized TRD definitions and consistent use of validated tools would improve the quality of the evidence base.

Few publications directly addressed the issue, but we detected a general consensus that a prospective, randomized trial design using certain study components best helped minimize bias. Randomized trials of various types, including in a few cases cluster randomized trials or nonrandomized controlled trials, were the preferred study design by the available systematic

reviews,[31-36, 39-41, 44, 58, 64-66] the nonsystematic reviews,[5, 18, 30, 37, 45, 46] and the guidelines or consensus statements.[19, 21, 23, 38, 42, 67-79]

Some investigators or experts extended their inclusion criteria to accept various nonexperimental (observational) designs of reasonable strength, such as prospective cohort studies. A handful included observational studies that could possibly have been subject to bias from confounding factors. About seven sources included systematic reviews as a research design; some also included meta-analyses.

Similarly, virtually all sources that commented on a preferred study design to minimize bias reflected a consensus that they preferred randomized designs. A few sources specified a preference for use of meta-analyses, including network (i.e., indirect) meta-analyses to enable combining and examining data from trials with both active and inactive comparators.[43]

The literature also emphasized other key study design components to minimize the role of bias. Systematic review of staging methods recommended the use of prospective study designs and validated instruments, where possible.[31, 44] Studies also recommended using whole structured clinical interviews to diagnosis depression,[31, 44] because these full assessments could better confirm the MDD (or bipolar) diagnosis and clarify psychiatric comorbidity, seen as a key potential confounder in TRD treatment trials.[36]

Adequate dosage of medication should be at maximum tolerated doses (according to prescription recommendations),[31, 41] further standardizing the definition. In general, systematic reviews,[31, 41] nonsystematic reviews,[18, 30, 44-46] and guidelines or consensus statements[19, 67, 69, 70] recommended that study duration with adequate dosing should be at least 6 weeks. They also appeared to prefer that remission, operationally defined using a validated instrument, is the preferred outcome.[31, 44] Systematic reviews emphasized that both compliance and consideration of prior psychotherapy use are important to assess and control for in analyses[31, 41, 44]

At the same time, these sources clearly appreciated that adding the above components risked the feasibility and applicability of these management strategies in real-world settings.[44]

We found nothing mentioned specifically about reducing placebo effect.

Trends on Best Study Design to Assess TRD Interventions

Current trends show increased use of traditional study designs. This preference reflects the variability of definitions and study design components and the limited evidence base for standardized, validated instruments to confirm TRD. In the literature we reviewed, we identified no newly emerging designs, although the systematic review for KQs 6 through 11 (next chapter) generally can identify the most current study designs.

Consensus on Appropriate Trial Length

Finally, as reported elsewhere, most sources that addressed the issue of trial length for antidepressant medication studies generally took the stance of trial duration being “at least 6 weeks” of a treatment at an adequate dose.[18, 19, 30, 31, 41, 44-46, 67, 69, 70] Of note, getting patients to an adequate dose of a given medication may take a few weeks; for that reason, 6 weeks of adequate dosing may produce a trial length longer than 6 weeks.

Some groups were comfortable with 4 to 6 weeks’ duration for a drug study; some advocated for or advised longer trials (e.g., 8 to 10 weeks). Yet others commented that the trials needed to provide adequate dosages of the medications in question. Almost one-third of the sources did not deal with this issue at all, however.

Key Question 5: Risk Factors for Treatment-Resistant Depression

We reviewed all included publications for any reference to risk factors for TRD and abstracted relevant information. These risk factors can be sorted into four main categories: (1) risk factors reflected in TRD definition and staging, (2) sociodemographic risk factors, (3) psychiatric and medical comorbidity, and (4) other clinical variables. We sorted each of those categories by whether the publications were systematic reviews, ostensibly providing the highest quality of evidence; nonsystematic reviews; or guideline or consensus statements. Finally, having TRD can be represented by two different outcomes: failure for a patient with depression either to achieve remission with or to respond to an intervention (because the latter also indicates failure to remit). The literature has reported it both ways. Accordingly, within each cell in the tables below, we identify what is reported about predictors for remission and response, respectively.

Key Points

1. Although many risk factors are posited for TRD, evidence addressing risk factors for TRD is quite limited.
2. Several components of the TRD definition (disease severity, duration of current episode, number of previous hospitalizations, and number of failed antidepressant trials) appeared to be associated with greater risk of TRD. For example, the probability of responding to an antidepressant declines by a factor of approximately 15 percent to 20 percent for each prior failed drug treatment.
3. The sociodemographic variables of age (older) and marital status (divorced/widowed) increased the risk of TRD.
4. Coexisting anxious symptoms, anxiety disorders, and personality disorders were associated with TRD.
5. Some other clinical characteristics (such as having melancholic features, suicidality) were associated with greater risk of TRD.

Detailed Synthesis

Tables 12 through15 summarize risk factors associated with TRD. The evidence base, hampered by various definitions of TRD and potential risk factors, is quite limited. In all, the evidence base for KQ 5 includes six separate sources. One recent systematic review provided the most comprehensive summary of risk factors for TRD.[36]

Table 12 reflects the evidence associating the different components of TRD and the presence of TRD itself. Key components of the definition—disease severity, duration of current episode, number of previous hospitalizations, and number of failed antidepressant trials—appeared to be associated with an increased risk of TRD. One guideline noted that the probability of responding to an antidepressant declines by a factor of approximately 15 percent to 20 percent for each prior failed drug treatment.[67] Failure of a particular class of antidepressant, however, did not appear to be associated with TRD risk. Of note, use of higher doses of antidepressant medication in prior treatment trials was associated with the highest risk of having TRD; however, the need for higher doses is likely a consequence of having TRD rather than a predictor of TRD.

Table 12. Components of definitions of treatment-resistant depression

Topic Authors and Year of Publication Type of Source	Depressive Disease Severity	Duration of Current Episode	Number of Previous Hospitali-zations	Number of Prior (Failed) Treatments	No Treatment Response During First 2 Treatments	Class of Previous Antidepres-sant(s)	Dose of Previous Antidepres-sant Treatment(s)
Definition and meaning of TRD Sackeim, 2001[37] Non-systematic review	Not addressed	For remission: Greater duration of current episode associated with lower remission rates For response: Greater duration of current episode associated lower response rates	Not addressed	For remission: Not addressed For response: Greater number of failed AD treatments associated with lower response rates	Not addressed	Not addressed	Not addressed
World Federation of Societies of Biological Psychiatry Guidelines for Unipolar Depression Bauer et al., 2009[67] Guideline	Not addressed	Not addressed	Not addressed	For remission: Greater number of prior AD attempts associated with decreased remission rates For response: Greater number of prior AD attempts associated with decreased response rates	Not addressed	Not addressed	Not addressed

Table 12. Components of definitions of treatment-resistant depression (continued)

Topic Authors and Year of Publication Type Source	Depressive Disease Severity	Duration of Current Episode	Number of Previous Hospitali-zations	Number of Prior (Failed) Treatments	No Treatment Response During First 2 Treatments	Class of Previous Antidepres-sant(s)	Dose of Previous Antidepres-sant Treatment(s)
Predictors of nonresponse De Carlo, Calati, and Serretti, 2016[36] Systematic review	For remission: Higher baseline severity of depression associated with lower remission rates illness For response: No association with depressive severity	For remission: Evidence not available For response: Longer duration of current episode associated with lower response rates	For remission: Evidence not available For response: Greater number of prior hospitali-zations associated with lower response rates	Not addressed	Not addressed	For remission: Mixed results For response: Not addressed	For remission: Weak evidence suggesting higher doses associated with lower remission rates, but high doses may be a consequence of TRD rather than a predictor For response: Higher dose associated with lower response rates, but high doses may be a consequence of TRD rather than a predictor

AD = antidepressant; TRD = treatment-resistant depression.

Table 13 shows the limited evidence regarding associations between sociodemographic risk factors and TRD. Both age, reflected as a continuous variable or as a cut-off of 40 years or older, and being divorced or widowed were associated with a greater risk of TRD. Neither sex nor gender nor race appeared to be clear risk factors.

Table 14 reviews the information regarding medical and psychiatric comorbidities. The limited evidence showed no clear association between medical comorbidities or chronic pain, respectively, and TRD. Particular psychiatric comorbidities, however, were predictors of TRD. Specifically, coexisting anxious symptoms and comorbid anxiety disorders, substance abuse disorders, and personality disorders each were identified as a risk factor for TRD.

Table 15 reports the information regarding other clinical risk factors for TRD. The limited evidence showed two clinical variables were risk factors for TRD as measured by failure to respond: having melancholic features and having suicidality. One review that looked at risk factors for not responding to rTMS indicated that particular genetic polymorphisms appeared associated with the risk of not responding to rTMS.

Table 13. Demographics and related risk factors for treatment-resistant depression

Topic Authors and Year of Publication Type	Age	Female Sex	Marital Status	Nonwhite Race or Ethnicity
De Carlo, Calati, and Serretti, 2016[36] Predictors of nonresponse Systematic review	For remission: No association between older age and lower remission rates For response: Older age associated with lower rate of response	For remission: No association found between sex and lower rates of remission For response: No association found between sex and lower rates of response	For remission: Being divorced/widowed associated with lower rates of remission For response: No association found between marital status and lower rates of response	For remission: Inconclusive evidence For response: Inconclusive evidence
SR on rTMS Ontario Health Association, 2016[64] Systematic review	For remission: Age ≥40 years associated with lower remission rates after rTMS treatments For response: Not addressed	Not addressed	Not addressed	Not addressed

rTMS = repetitive transcranial magnetic stimulation; SR = systematic review

Table 14. Medical and psychological comorbidities as risk factors for treatment-resistant depression

Topic Authors and Year of Publication Type	Coexisting Medical Comorbidities	Coexisting Psychiatric Comorbidities	Chronic Pain
Definition and Epidemiology of TRD	For remission: Not addressed	For remission: Not addressed	Not addressed
Fava and Davidson, 1996[30] Nonsystematic review	For response: While frequently cited as a potential risk factor, there is no clear evidence that medical comorbidity is associated with decreased response rates	For response: Substance abuse and even moderate consumption of alcohol have been associated with poorer response to antidepressant treatment Comorbid personality disorders have also been associated with decreased response rates in some but not all studies	
Predictors of Nonresponse De Carlo, Calati, and Serretti, 2016[36] Systematic review	For remission: No association between comorbid medical diagnoses and reduced remission rates For response: No association between comorbid medical diagnoses and response rates	For remission: Anxious symptoms associated with lower remission rates, whereas anxiety disorders did not show a clear association A personality disorder diagnosis was also associated with lower remission rates For response: Anxiety symptoms and anxiety disorders were associated with lower response rates A personality disorder was associated with a lower response rate	For remission: No association between pain and reduced remission rates For response: No association between pain and reduced response rates

TRD = treatment-resistant depression.

Table 15. Relationship between other clinical characteristics and treatment-resistant depression

Topic Authors and Year of Publication Type or Topic	MDD Onset Before Age 20	Family History of Depressive Disorder	Melancholic Features	Suicidal Risk or Behavior	Other Clinical Characteristics
Predictors of response for rTMS Silverstein et al., 2015[34] Systematic review	Not addressed	Not addressed	Not addressed	Not addressed	For remission: Not addressed For response: Association of genetic polymorphisms with decreased rate of response following rTMS treatment
Predictors of nonresponse De Carlo, Calati, and Serretti, 2016[36] Systematic review	For remission: No association between onset of first episode and lower remission rates For response: No association between onset of first episode and lower response rates	For remission: No association between family history of mood disorders and lower remission rates For response: No association between family history of mood disorders and lower response rates	For remission: No association between melancholic features and lower remission rates For response: Having melancholic features increased the likelihood of nonresponse	For remission: No association between increased suicidality and lower remission rates For response: Increased suicidality was associated with lower response rates	Not addressed

MDD = major depressive disorder; rTMS = repetitive transcranial magnetic stimulation.

Key Question 6: Patient Characteristics, Approaches to Prior Treatments as Inclusion Criteria, and Elements of Diagnostic Assessments

Description of Included Studies

Our searches identified 163 unique studies (in 197 publications) of interventions in TRD populations. We divided interventions into four categories: (1) brain stimulation treatments (BST), which included electroconvulsive therapy (ECT), repetitive transcranial magnetic stimulation (rTMS), vagal nerve stimulation, and deep brain stimulation (73 studies); (2) pharmacotherapy, including ketamine (71 studies); (3) psychotherapy (11 studies); and (4) complementary and alternative medicine (CAM) therapies and exercise (8 studies).

Below we report on patient and study characteristics overall and by treatment-specific categories. Certain characteristics may change based on study intervention (e.g., participants in a study of ECT may be systematically different from those in a trial of psychotherapy). We give tables with counts for most subsections below; notable exceptions and trends not captured in tables are presented in the text.

Key Points

1. Confirmation of TRD for study entry was often poorly described.
2. The HAM-D and the Montgomery–Åsberg Depression Rating Scale were commonly used to set minimum depressive severity thresholds for study entry; most studies involved patients with moderately severe depression.
3. Studies were inconsistent about the necessary duration of prior treatment attempts for study entry.
4. Most studies required at least one and often two prior failed treatment attempts of adequate therapy.
5. Several patient characteristics were rarely considered for study entry: duration of depressive symptoms, prior depressive relapses, prior treatment intolerance, prior augmentation or combination therapy, prior psychotherapy, and suicidality.

Detailed Synthesis

Patient Characteristics Related to Inclusion or Exclusion

Age

Nearly all studies included participants ages 18 years or older, although four studies (pharmacology only) limited enrollment to participants 60 years or older (Table 16). Fifty studies (31%) used a maximum cutoff of 65 years of age. Forty studies (25%) did not report age criteria; 38 studies (23%) did not report an age limit for inclusion. Most studies reported a mean age between 50 and 60 years.

Table 16. Number of studies reporting maximum age for study enrollment

Age	BST	Psychotherapy	Pharmacology	CAM/Exercise
Maximum				
60	1	0	0	2
65	12	5	30	3
70	9	0	1	1
75	4	2	6	0
80	1	0	3	1
85	3	0	1	0
No maximum age	15	3	19	1
No age requirements reported	28	1	11	0

BST = brain stimulation therapy; CAM = complementary and alternative medicine.

Type of Depressive Episode

Of 163 studies, 154 (95%) included *any* patients with unipolar MDD. Forty-five (28%) studies included patients with both unipolar and bipolar disease (37 BST, 6 pharmacotherapy, 2 CAM/exercise). Nine (6%) studies included only patients with bipolar disease (2 BST, 6 pharmacology, 1 CAM/exercise).

Most studies excluded patients with psychotic depression; rarely were other specific types of depression considered (Table 17). Three studies included patients with double depression (MDD plus antecedent dysthymia). Two studies of bipolar depression excluded patients with rapid cycling disease. One trial excluded patients with seasonal affective disorder and depression secondary to a medical condition. One trial reported rates of dysthymia and schizoaffective disorder.

Table 17. Number of studies considering depressive episode type for study inclusion

Type of Depressive Episode	BST	Psychotherapy	Pharmacology	CAM/Exercise
Psychotic				
Inclusion	0	0	1	0
Exclusion	34	11	50	7
Just reported	10	0	4	0
Not considered	29	0	16	1
Atypical				
Inclusion	0	0	1	0
Exclusion	2	0	1	0
Just reported	0	0	7	0
Not considered	71	11	62	8
Chronic				
Inclusion	3	5	3	0
Exclusion	1	1	0	0
Just reported	5	1	11	0
Not considered	71	11	62	8
Melancholic				
Inclusion	0	0	0	0
Exclusion	0	0	0	0
Just reported	4	0	10	0
Not considered	69	11	57	8
Catatonic				
Inclusion	0	0	0	0
Exclusion	0	0	1	0
Just reported	0	0	1	0
Not considered	73	11	69	8
Postpartum				
Inclusion	0	0	0	0
Exclusion	0	0	3	0
Just reported	0	0	2	0
Not considered	73	11	66	8

BST = brain stimulation therapy; CAM = complementary and alternative medicine.

Number of Depression Relapses and Time to Relapse

Only a single trial considered number of depression relapses, a specific subtype of failed treatment where a patient has a depressive episode that remits only briefly and then returns before full recovery occurs, as an inclusion criterion. No study either considered time to relapse as an inclusion criterion or reported on it.

Psychiatric Comorbidities

Most studies did not name psychiatric comorbidities as inclusion criteria, although some studies did report on the number of included participants with other psychiatric disease. More frequently, investigators used coexisting psychiatric illnesses as exclusion criteria; substance abuse was the most frequently noted exclusion (Table 18). Many studies named specific disorders for exclusion; some studies more generally excluded any type of psychiatric comorbidity or any other Axis I disorders that were present. Fewer than three studies listed any Axis II disorders generally or attention deficit/hyperactivity disorder specifically as exclusion criteria (not included in table).

Table 18. Number of studies considering comorbid psychiatric diagnoses as exclusion criteria

Other Psychiatric Diagnoses	BST	Psychotherapy	Pharmacology	CAM/Exercise
Studies with any general or specific psychiatric comorbidity exclusion	49	11	57	8
Any psychiatric comorbidity	5	0	2	2
Any general Axis I disorder	6	0	13	0
Anxiety disorders	6	1	8	1
PTSD	9	1	12	1
Substance abuse	42	10	46	7
Eating disorder	3	1	8	0
OCD	4	1	10	0
Psychotic disorders	18	4	23	1
Any personality disorder	18	4	23	1

BST = brain stimulation therapy; CAM = complementary and alternative medicine; OCD = obsessive-compulsive disorder; PTSD = post-traumatic stress disorder.

Medical Comorbidities

Many studies (n=66) potentially excluded participants at the discretion of the investigators, most often because of serious or unstable medical conditions that would have limited participation or involvement in study procedures; examples include ECT or exercise (Table 19). Infrequently cited exclusion criteria (e.g., fewer than 5 studies each) included stroke or transient ischemic attack, multiple sclerosis, Huntington's chorea, cranial mass, orthopedic injuries including head trauma, and glaucoma (not listed in table).

Table 19. Number of studies considering comorbid medical diagnoses as exclusion criteria

Other Medical Diagnoses	BST	Psychotherapy	Pharmacology	CAM/Exercise
Studies with any general or specific medical comorbidity exclusion	58	7	49	8
Unspecified serious or unstable medical conditions	27	2	31	6
Unspecified neurological disorders	28	1	3	1
Seizure disorders	22	0	12	3
Other cognitive abnormalities	31	6	22	0
Cardiac disease	7	0	8	0
Renal disease	0	0	2	0
Diabetes	2	0	0	0
Pregnancy	8	0	21	1
Abnormal laboratory test results	0	0	12	3
Pacemakers or metal implants	13	0	1	0

BST = brain stimulation therapy; CAM = complementary and alternative medicine.

Suicidal Ideation and Suicide Attempts

Only one trial specifically included patients with acute suicidality. Fifty-eight studies excluded patients with current suicidal ideation, and 11 studies excluded patients with recent suicide attempts. Seven studies reported only the presence of suicidal ideation among participants; 15 studies reported on prior suicide attempts among participants (Table 20).

Table 20. Number of studies considering suicidal ideation and prior suicide attempts as inclusion or exclusion criteria or reporting on these events

Suicide Ideation or Attempts	BST	Psychotherapy	Pharmacology	CAM/Exercise
Ideation				
Inclusion	0	0	0	0
Exclusion	18	5	31	4
Just reported	0	0	7	0
Not considered	55	6	33	4
Attempts				
Inclusion	0	0	1	0
Exclusion	7	1	3	0
Just reported	4	3	8	0
Not considered	32	7	59	8

BST = brain stimulation therapy; CAM = complementary and alternative medicine.

Duration of Symptoms

Most studies (n=126) did not consider duration of symptoms for study inclusion or exclusion. As shown in Table 21, 22 studies (14%) required a minimum symptom duration for inclusion; eight studies (5%) defined both minimum and maximum durations for inclusion. The mean minimum duration of symptoms ranged between 2 months (pharmacology) and 12 months (BST and psychotherapy). The mean maximum duration of symptoms ranged between 12 and 18 months, with a notable exception for BST interventions (51 months).

Table 21. Number of studies specifying required symptom duration for study inclusion

Symptom Duration	BST	Psychotherapy	Pharmacology	CAM/Exerciser
Minimum but no maximum	11	4	6	1
Maximum but no minimum	2	1	1	0
Both minimum and maximum	2	0	5	1
Not considered	57	6	57	6

BST = brain stimulation therapy; CAM = complementary and alternative medicine

Instruments to Make a Diagnosis of Depression and Rate Severity

Across all studies, research teams used a total of 12 tools to rate depression severity; the HAM-D was most frequently used (Table 22). Although investigators did not use any of these tools to make a *formal* diagnosis of depression, they often used them to set thresholds for study inclusion (e.g., HAM-D > 20). Additionally, researchers used many of these tools to measure baseline characteristics and to measure study outcomes.

Diagnostic Tools to Confirm a Diagnosis of Depression

In addition to unstructured clinical assessments, investigators used three structured diagnostic tools to confirm the diagnosis of depression: the Diagnostic and Statistical Manual of Mental Disorders (DSM) checklist, the Mini International Neuropsychiatric Interview (MINI), and the Structured Clinical Interview for DSM (SCID) (Table 23). Some studies combined more than one method (e.g., unstructured clinical assessment plus MINI).

Table 22. Number of studies using depression screening instruments for study inclusion

Screening Instrument	BST	Psychotherapy	Pharmacology	CAM/Exercise
HAM-D (all)	38	7	29	4
$HAM\text{-}D_{17}$	22	4	22	4
$HAM\text{-}D_{21}$	10	2	7	0
$HAM\text{-}D_{24}$	5	1	0	0
$HAM\text{-}D_{28}$	1	0	0	0
IDS/QIDS (all)	0	0	12	2
$IDS\text{-}C_{30}$	0	0	4	0
$QIDS\text{-}C_{16}$	0	0	4	0
$QIDS\text{-}SR_{16}$	0	0	4	2
BDI	0	4	0	0
MADRS	14	0	16	1
CGI (all)	5	1	8	0
CGI-S	5	1	4	0
CGI-I	0	0	4	0
GAF	2	1	0	0
No screening tool	20	1	14	1

BDI = Beck Depression Inventory; BST = brain stimulation therapy; CAM = complementary and alternative medicine; CGI = Clinical Global Impression Scale (S= severity, I = improvement); GAF = Global Assessment of Functioning Scale); HAM-D = Hamilton Rating Scale for Depression; IDS = Inventory of Depressive Symptomatology (C = clinician rated, SR = self-rated); MADRS = Montgomery–Åsberg Depression Rating Scale; QIDS = Quick Inventory of Depressive Symptomatology (C = clinician rated, SR = self-rated).

Table 23. Numbers of studies using tools to confirm depression diagnosis

Tools to Confirm Diagnoses	BST	Psychotherapy	Pharmacology	CAM/Exercise
Structured DSM checklist	0	0	6	0
MINI	21	2	14	0
SCID	17	4	24	5
Unstructured clinical assessment	36	5	26	3

BST = brain stimulation therapy; CAM = complementary and alternative medicine; DSM = Diagnostic and Statistical Manual; MINI = Mini International Neuropsychiatric Interview; SCID = Structured Clinical Interview for DSM.

Few studies (n=57, 35%) used a standardized definition of TRD to confirm the diagnosis. (These instruments were described in the previous results chapter.) The Antidepressant Treatment History Form (ATHF), which confirms adequate dose and duration, was the most commonly used (Table 24), followed by the Thase and Rush Staging Model (TRSM). The Antidepressant Treatment Response Questionnaire (ATRQ), which systematically clarifies prior pharmacologic use but does not formally stage a depression episode, was reported in eight studies (2 BST and 6 pharmacology). None of the studies we identified for this part of the Technology Assessment used the European Staging Method, the Massachusetts General Hospital staging method (MGH-s), or the Maudsley Staging Method (MSM).

Table 24. Numbers of studies using standardized definitions of treatment-resistant depression to confirm diagnoses

Sources of Standardized Definitions	BST	Psychotherapy	Pharmacology	CAM/Exercise
ATHF	15	1	13	1
ATRQ	2	0	6	0
TRSM	15	1	3	0

ATHF = Antidepressant Treatment History Form; ATRQ = Antidepressant Treatment Response Questionnaire; BST = brain stimulation therapy; CAM = complementary and alternative medicine; TRSM = Thase and Rush Staging Model

Prior Treatments as Inclusion Criteria

Reporting of Prior Treatments

Of the 163 studies, nearly all (n=157) reported some inclusion criteria concerning prior antidepressant treatments for study entry. Slightly more than two-thirds of studies (n=108) listed a defined duration of prior treatment for study inclusion; by contrast, about 15 percent of studies (n=23) reported prior treatment attempts for study inclusion as *adequate* (Table 25). Eighteen studies excluded participants with a predefined number of prior failed treatment attempts.

Table 25. Number of studies using reported duration of prior treatment attempts for study inclusion

Length of prior treatment attempts	BST	Psychotherapy	Pharmacology	CAM/Exercise
<4 weeks	2	0	0	0
4 weeks	17	1	16	0
5–7 weeks	24	3	24	1
≥8 weeks	2	4	9	5
Adequate	10	0	12	1
Just reported	1	0	4	0
Not considered	17	3	6	1

BST = brain stimulation therapy; CAM = complementary and alternative medicine.

Of all 163 studies, 126 (77%) specified a predetermined antidepressant dosage for study inclusion (Table 26), and 99 studies (61%) required both a prespecified prior treatment duration and antidepressant dosage for study inclusion. Finally, 4 studies (3%) required a prespecified dosage but did not consider length of prior treatment attempts.

Table 26. Number of studies using reported duration and dosage of prior treatment attempts for study inclusion

Length of prior treatment attempts	Dosage of prior treatment considered	Dosage of prior treatment not considered
<4 weeks	2	0
4 weeks	33	1
5–7 weeks	46	6
≥8 weeks	18	2
Adequate	23	0
Just reported	2	3
Not considered	2	25
Total	126	37

The most frequent class of antidepressants required for study inclusion comprised the second-generation antidepressants. Monoamine oxidase inhibitors (MAOI) and atypical antipsychotics were the most frequently excluded (Table 27). Often, studies either reported only the antidepressants that participants had used previously or did not report prior treatment medications at all. Table 27 documents the number of studies reporting medication classes that participants had used and whether specific medication classes were only reported or were considered as inclusion or exclusion criteria. These findings are further split by the four main intervention categories (i.e., BST, psychotherapy, pharmacology, and CAM/exercise).

Table 27. Number of studies using various classes of antidepressants attempted for treatment before study inclusion

Trial Intervention and Drug Classes	Inclusion	Exclusion	Only Reported	Not Considered
BST				
SSRI	7	1	19	46
SNRI	6	1	17	49
NDRI	0	0	7	66
TCA	7	0	20	46
MAOI	2	1	8	62
5-HT receptor antagonist	1	0	6	66
Atypical antipsychotics	0	1	13	59
NMDA	0	0	3	70
Anticonvulsants	0	3	8	62
Psychostimulants	0	0	5	68
Mood stabilizers	1	3	13	56
Psychotherapy				
SSRI	2	0	3	6
SNRI	2	0	2	7
NDRI	2	0	1	8
TCA	2	0	0	9
MAOI	1	0	1	9
5-HT receptor antagonist	1	0	1	9
Atypical antipsychotics	0	0	1	10
NMDA	0	0	1	10
Anticonvulsants	0	0	1	10
Psychostimulants	0	0	1	10
Mood stabilizers	1	1	1	8
Pharmacotherapy				
SSRI	20	2	14	35
SNRI	8	2	15	46
NDRI	2	1	14	54
TCA	5	3	11	52
MAOI	1	10	8	52
5-HT receptor antagonist	1	1	9	60
Atypical antipsychotics	0	13	7	51
NMDA	0	4	4	63
Anticonvulsants	2	6	4	59
Psychostimulants	0	4	5	62
Mood stabilizers	4	6	6	55
CAM/Exercise				
SSRI	4	0	1	3
SNRI	0	1	1	6
NDRI	0	1	1	6
TCA	0	1	1	6
MAOI	0	1	0	7
5-HT receptor antagonist	0	1	0	7
Atypical antipsychotics	0	1	0	7
NMDA	0	1	0	7
Anticonvulsants	0	1	0	7
Psychostimulants	0	1	0	7
Mood stabilizers	0	1	0	7

5-HT = 5-hydroxytryptamine; BST = brain stimulation therapy; CAM = complementary and alternative medicine; MAOI = monoamine oxidase inhibitor; NDRI = norepinephrine-dopamine reuptake inhibitor; NMDA = N-methyl D-aspartate; SNRI = serotonin-norepinephrine reuptake inhibitor; SSRI, selective serotonin reuptake inhibitor; TCA = tricyclic antidepressant

Number of Failed Attempts of Adequate Therapy

Some investigators required a specified number of failed attempts of adequate therapy as another inclusion criterion. "Adequate" was defined in a variety of ways; these could include using, for instance, prespecified duration of treatment, prespecified dosage of antidepressant, a combination of those two criteria, or some other common definition such as that applied in the ATHF. (Table 26 also describes studies in terms of duration and dose.)

The number of failed attempts ranged between one and four (Table 28)—generally either one or two (varying by intervention type). Eighteen studies (6 BST, 9 pharmacology, 3 CAM/Exercise) *prespecified* a maximum number of treatment failures for study inclusion. No trial required more than four failed treatment attempts for study entry; seven studies in all did not consider this criterion.

Table 28. Number of studies requiring a failed attempt of adequate therapy for study inclusion

Number of Failed Attempts	BST	Psychotherapy	Pharmacology	CAM/Exercise
1	21	8	47	4
2	35	3	21	3
3	4	0	2	0
4	8	0	0	0
Not considered	5	0	1	1

BST = brain stimulation therapy; CAM = complementary and alternative medicine.

Prior Treatment Intolerance

Few studies (n=6) used patient intolerance of prior treatment as an inclusion criterion (4 BST stimulation, 2 pharmacology). Four studies used prior intolerance as an exclusion criterion (1 BST, 3 pharmacology). Additionally, four studies reported the prevalence of prior treatment intolerance among study participants without using it as an inclusion criterion (2 BST stimulation, 2 pharmacology).

Use of Augmentation and Combination Pharmacological Therapies

Few studies considered use of prior augmentation and combination therapies for either inclusion (n=7) or exclusion (n=3) criteria. Only 13 studies reported whether participants had previously used an augmentation or combination regimen.

Use of Electroconvulsive Therapy and Psychotherapy

A minority of studies considered ECT (a type of BST) (n=57, or 35%, for BST or pharmacotherapy interventions) (Table 29). An even smaller minority included psychotherapy (n=28, or 17%, across all interventions) among its inclusion criteria.

Table 29. Number of studies considering use of electroconvulsive therapy or psychotherapy for study inclusion

Type of Therapy	BST	Psychotherapy	Pharmacotherapy	CAM/Exercise
ECT				
Inclusion	4	0	1	0
Exclusion	17	0	15	0
Just reported	15	0	5	0
Not considered	37	11	50	8
Psychotherapy				
Inclusion	1	1	0	0
Exclusion	1	8	9	2
Just reported	1	0	5	0
Not considered	70	2	57	6

BST, brain stimulation therapies; CAM, complementary and alternative medicine; ECT = electroconvulsive therapy.

Diagnostic Characteristics

Diagnostic Assessments

The majority of studies used structured diagnostic assessments (n=97, 60%); these included a DSM checklist, SCID, or MINI (Table 30). Unstructured assessments were primarily clinical assessments, often described in the study methods as "… met DSM criteria for MDD." However, study authors often did not state clearly whether investigators used these tools to confirm the diagnosis of MDD for study entry or to confirm the presence of TRD.

Table 30. Number of studies reporting structured or unstructured diagnostic assessments

Type of Assessment	BST	Psychotherapy	Pharmacology	CAM/Exercise
Structured	37	8	47	5
Unstructured	36	3	24	3

BST = brain stimulation therapy; CAM = complementary and alternative medicine.

Scores on Standardized and Validated Depression Rating Instruments

Investigators used a variety of depression rating instruments either to limit study enrollment or to track study outcomes (see Tables 22 and 23 above). For studies that measured or monitored depression severity using a validated instrument (Table 31), the majority included participants with moderate depression (91 of 120 studies). Table 32 gives the thresholds for mild, moderate, and severe depression according to different instruments.

Table 31. Mean depression severity rating in studies using validated depression-rating instruments

Depression Severity Level	BST	Psychotherapy	Pharmacology	CAM/Exercise
Mild	2	4	5	1
Moderate	39	5	41	6
Severe	9	0	8	0

BST = brain stimulation therapies; CAM = complementary and alternative medicine.

Table 32. Severity cut points for commonly used depression rating instruments

Instrument	Mild	Moderate	Severe
BDI	≤ 18	18–29	≥ 30
HAM-D_{17}	≤ 13	14–19	≥ 20
HAM-D_{21}	≤ 15	16–22	≥ 23
HAM-D_{24}	≤ 18	19–26	≥ 27
IDS-C_{30}	≤ 23	24–36	≥ 37
MADRS	≤ 19	20–34	≥ 35
QIDS-SR_{16}	≤ 10	11–15	≥ 16
QIDS-C_{16}	≤ 10	11–15	≥ 16

BDI = Beck Depression Inventory; HAM-D = Hamilton Rating Scale for Depression; IDS = Inventory of Depressive Symptomatology (C = clinician rated); MADRS = Montgomery–Åsberg Depression Rating Scale; QIDS = Quick Inventory of Depressive Symptomatology (C = clinician rated, SR = self-rated).

Study Setting

Most studies enrolled from or were conducted in outpatient settings (n=105, 64%) (Table 33). A small minority (n=26; 16%) were conducted exclusively in inpatient settings, which were typically psychiatric wards. A substantial number of studies (n=83, 51%) did not clearly describe the study setting or did not report the setting at all.

Table 33. Clinical settings in which participants were enrolled or treated

Enrollment or Study Setting	BST	Psychotherapy	Pharmacology	CAM/Exercise
Primary care clinic	0	3	0	0
Psychiatric clinic	14	1	11	3
Primary care + psychiatric clinics	0	2	6	0
Unspecified outpatient clinic	31	4	25	5
Inpatient setting	12	1	13	0
Inpatient + any outpatient clinic	9	0	7	0
Setting not reported	7	0	9	0

BST = brain stimulation therapies; CAM = complementary or alternative medicine.

Key Question 7: Comparison of Inclusion Criteria With Definition of Treatment-Resistant Depression from Narrative Questions

This KQ assesses how well the inclusion criteria from eligible intervention studies (reported in KQ 6) match current definitions of TRD (from KQ 1 in the previous chapter). Based on the 38 publications from KQ 1 (14 systematic reviews, 6 nonsystematic but relevant reviews, and 18 guidelines or consensus statements), we have identified the following four key variables defining TRD:

1. **Minimum number of treatment failures.** In KQ 1, a minimum of two treatment failures was the most common definition.
2. **Prior adequate treatment dose.** In KQ 1, this criterion ranged from "minimum effective dose" to "maximum tolerated dose."
3. **Prior adequate treatment duration.** In KQ 1, the most common duration was either at least 4 or at least 6 weeks. We selected at least 4 weeks as a threshold for adequate in considering the findings from the systematic reviews.
4. **Formal staging of TRD** using a staging system described in KQ 1.

Against the key variables noted above, we assessed how well inclusion criteria for the 151 unique intervention studies reflect what previous observers or investigators had considered to be

TRD. Six of the studies we have reviewed for this chapter involved patients with bipolar disorder alone; for reasons of simplicity, we have combined these studies with the MDD studies.

With regard to adequacy of dose or duration (variables 2 and 3 above), we assessed not merely whether the trial authors stated that they had considered prior treatment dose (or duration); we also determined whether they *specifically* indicated how they had confirmed that information, as noted below.

For adequate dose, we first identified whether studies stated that they had restricted eligibility to patients who had received an adequate prior dose as part of their inclusion criteria (i.e., that an adequate dose was *considered* in determining eligibility). If so, we subsequently identified whether the study had systematically *confirmed* this dose by specifying dosage levels through interview, questionnaire (e.g., ATRQ), or other formal clarification. We considered a statement that eligibility criteria required a minimum therapeutic dose as stated by product labeling to indicate confirmation.

Similarly, for adequate duration, we first identified whether the studies stated they *considered* this criterion by restricting eligibility to those patients with a prior adequate duration; if that were the case for these studies, we then determined whether they *confirmed* the duration by clarifying that patients previously received what KQ 1 had indicated was an adequate dose. In KQ 1, approximately one-half of the eligible reviews and guidelines identified a minimum of 4 weeks of treatment; the other half identified the minimum as 6 weeks. We define an adequate dose here as 4 weeks because one of the primary tools to confirm adequacy of dose and duration, the ATHF, required at least 4 weeks to be considered adequate duration.

Key Points

1. Inclusion criteria as specified by the eligible TRD studies did not closely align with the definition(s) of TRD identified in the narrative review.
2. Although the most common definition of TRD involves a minimum of two failed prior adequate antidepressant studies, the most common minimum number of treatment failures used as an inclusion criteria for TRD was one (49%) (only 38 percent required a minimum of two failed trials).
3. Of all 163 studies, 77 percent considered in their selection criteria whether the patient had been treated previously with an adequate dose; 58 percent systematically confirmed that the dose was adequate by specifying dosage levels through interview, questionnaire, or other formal clarification.
4. Of all 163 studies, 82 percent considered in their selection criteria whether prior treatments were of adequate duration; 72 percent systematically confirmed that the duration was adequate (≥4 weeks of treatment).
5. Thirty-three percent of all studies set inclusion criteria based on stage of TRD using a formal staging model.
6. Seventeen percent of studies had all the most commonly described criteria for TRD: a minimum of two prior treatment failures, confirmation that a dose was adequate, and confirmation that duration was 4 weeks or longer.

Detailed Synthesis

For this KQ, we analyzed the match between inclusion criteria of the 163 studies examined in this chapter with the key components of TRD definitions identified in KQ 1 (the narrative results

chapter). To reflect better overall how well the TRD studies match TRD definitions, we did not sort by type of intervention; rather, we considered treatment studies as a whole.

As reported in Table 34, we first sorted by the number of minimum prior treatment failures that investigators used to indicate TRD in their studies. Per KQ 1, this criterion specified a minimum of one, two, or three prior treatment failures (the rows identified on the far left); we added a row for four prior treatment attempts because some intervention studies used that criterion. We also specified a row for studies that did not consider prior treatment failures at all. Then, for each of these categories for prior failed treatment attempts, we recorded the numbers of studies that (a) considered (identified) adequate dosage as a selection criterion and (b) confirmed that dose by specifying dosage levels through interview, questionnaire (e.g., ATRQ) or other formal clarification (e.g., ATHF or TSRM stage of 3 or greater). Subsequent to that analysis, we documented whether investigators identified adequate duration as a selection criterion for their studies and whether they confirmed duration—in this case as at least 4 weeks. Finally, we recorded the number of studies that formally staged the TRD using a specific staging model.

Table 34. Numbers of studies of treatment-resistant depression considering or confirming key inclusion criteria for defining the diagnosis

Minimum Prior Treatment Failures	Number of Studies Using This Minimum for Inclusion	Adequate Dosage Considered	Adequate Dosage Confirmed	Adequate Duration Considered	Adequate Duration Confirmed	TRD Staged
1	80	69	59	71	67	22
2	62	47	30	52	45	25
3	6	5	3	5	2	3
≥4	8	4	2	5	3	3
Not considered	7	1	1	1	0	1

TRD = treatment-resistant depression.

Minimum Number of Prior Treatment Failures

The most common minimum number of treatment failures used as an inclusion criterion for TRD was one (see Table 33) (i.e., used by 80 of 163 studies [49%]). The criterion specifying a minimum of at least two treatment failures (the most commonly described TRD definition) was used by 62 studies (38%). Of note, as indicated in Table 28 in KQ 6, pharmacology studies were most likely to use a minimum of one failed trial (59%, or 47/80), while BST was most likely to use a minimum of two (56%, or 35/62). Only a small number had either three, or four or more as a minimum for prior treatment failures (6/163, 4%, and 8/163 (5%). Finally, seven studies (4%) did not consider the number of prior treatment failures explicitly in their selection criteria.

Adequate Dose Considered and Confirmed

Overall, 77 percent of all studies (125/163) stated explicitly that they considered adequate dose as part of their selection criteria. This percentage did not substantially differ by the minimum number of prior failures. For studies requiring one trial failure, 86 percent considered this criterion; for those requiring a minimum of two failures, 76 percent (47/62) considered adequate dose as part of their selection criteria.

By contrast, however, fewer confirmed the dose by specifying discrete dosage levels through interview or other means of other formal clarification. Overall, 58 percent (86/163) systematically confirmed that the dose was adequate. Dose confirmation was most likely for those with a minimum of one prior treatment failure (i.e., 74 percent [59/80]); for studies requiring two treatment failures, the figure was 48 percent (30/62).

Studies that did not consider the number of prior treatment failures in selection criteria only rarely considered adequate dosage for eligibility.

Adequate Duration Considered and Confirmed

Overall (Table 34), 82 percent of studies (133/163) considered adequate duration as part of their inclusion criteria. Those studies with a minimum requirement of one treatment failure most commonly considered adequate duration (89%, or 71/80); those requiring two failures considered adequate duration 84 percent of the time (52/62).

Fewer studies systematically confirmed duration. Overall, 72 percent (117/163) confirmed that the duration was adequate (≥4 weeks of treatment). Such confirmation was mostly likely for those with a minimum of one prior treatment failure (84 percent [67/80]); for studies requiring at least two treatment failures, the figure was 73 percent (45/62).

Formal Staging of Treatment-Resistant Depression

Inclusion criteria requiring formal staging of TRD using an identified model to determine eligibility were not common (Table 34). Only 33 percent of all studies (53/163) set inclusion criteria based on stage of TRD using a formal model. Confirmation was most likely for those with a minimum of two prior treatment failures (40%; 25/62) and less common for those with a minimum of one prior treatment failure (28%; 22/80).

Number of Studies Enrolling TRD Sample With Most Commonly Reported Components of the TRD Definition

We reviewed the eligible studies to determine those that met the most commonly reported part of the TRD definition: a minimum of two prior treatment failures, a confirmed adequate dose, and a confirmed adequate duration of treatment (≥4 weeks). Only 17% of studies (27/163) specified and confirmed through eligibility criteria that their population had these three most common components of the current TRD definition.

When we loosened the benchmark so that study inclusion criteria merely needed to state that adequacy of dose and duration was *considered* (not systematically confirmed), only 25% of studies (40/163) met that mark.

Key Question 8: Main Study Designs, Approaches for Run-In or Wash-Out Periods, and Study Durations

Description of Included Studies

As reported earlier for KQ 6, we had 163 unique studies of interventions in TRD populations, analyzed in four broad categories: BST (including ECT, rTMS, vagal nerve stimulation, and deep brain stimulation [73 studies]), pharmacotherapy (71 studies), psychotherapy (11 studies), and CAM therapies and exercise (8 studies). To address this KQ, we focused on designs of the studies, ways that investigators applied either run-in or wash-out periods, and length of studies. The three tables below document counts and percentages overall and within each intervention category; notable exceptions and trends not captured in tables are presented in the text.

Key Points

1. Most studies had RCT designs (88%); this rate did not differ by intervention type. Few psychotherapy RCTs were double blind (1 of 10).
2. Few studies had run-in periods (21%) or wash-out periods (23%).
   a. In the majority of the 35 studies (63%) with a run-in stage, it consisted of an active medication period.
   b. In the majority of the 37 studies (68%) with a wash-out stage, it consisted of a medication-free period.
   c. Only one CAM/exercise trial included a run-in period; no CAM/exercise trials included a wash-out period.
3. Study duration varied across studies, ranging from less than 2 weeks to more than 4 years.
   a. The majority of the BST studies lasted 2 months or less (62%).
   b. The average duration of the psychotherapy studies was longer than the length of other intervention studies; 36 percent lasted 1 year or longer.

Detailed Synthesis

Study Characteristics

Study Design

Of the 163 unique TRD studies, 73 were of BST, 11 of psychotherapy, 71 of pharmacology, and 8 of CAM or exercise interventions. Overall (Table 35), a majority of the studies had RCT designs (88%). None was a cluster RCT. Study design was mostly consistent across intervention types. The vast majority of the BST, psychotherapy, and pharmacology studies were RCTs (89%, 91%, and 85%, respectively); all CAM/exercise studies had RCT designs. Although a majority or plurality of BST, pharmacology, and CAM/exercise RCTs were double-blind, only one of the ten psychotherapy RCTs was; more than one-half of the psychotherapy RCTs were single-blind (6 of 10).

Table 35. Numbers of studies by study design and intervention type

Study Design	Total n (%)	BST n (%)	Psychotherapy n (%)	Pharmacology n (%)	CAM/Exercise n (%)
RCT studies	143 (88%)	65 (89%)	10 (91%)	60 (85%)	8 (100%)
Double-blind	97 (68%)	55 (85%)	1 (10%)	37 (62%)	4 (50%)
Single-blind	27 (19%)	8 (12%)	6 (60%)	10 (17%)	3 (38%)
Open label	19 (13%)	2 (3%)	3 (30%)	13 (22%)	1 (13%)
Cluster	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Non-RCT studies	20 (12%)	8 (11%)	1 (1.4%)	11 (15%)	0 (0%)
Nonrandomized controlled	4 (25%)	2 (25%)	0 (0%)	2 (18%)	0 (0%)
Prospective cohort	4 (25%)	2 (25%)	1 (100%)	1 (9%)	0 (0%)
Retrospective cohort	9 (45%)	1 (13%)	0 (0%)	8 (73%)	0 (0%)
Case-control	2 (10%)	2 (25%)	0 (0%)	0 (0%)	0 (0%)
Interrupted time series	1 (5%)	1 (13%)	0 (0%)	0 (0%)	0 (0%)
Total	163 (100%)	73 (100%)	11 (100%)	71 (100%)	8 (100%)

BST = brain stimulation therapies; CAM = complementary and alternative medicine; n = number; RCT = randomized controlled trial.

Run-In and Wash-Out Periods

Of 163 studies, 35 (17%) included run-in periods before randomization (Table 36). The types of run-in periods included use of placebo medication (n=1), active medication (n=22), stable medication (n=9), and no treatment (n=3) phases. The use of run-in periods varied by type of intervention; 11 percent, 30 percent, 27 percent, and 13 percent of BST, psychotherapy, pharmacology, and CAM/exercise studies, respectively, included a run-in period before randomization. The goals were to help screen out noncompliant patients, mitigate the effects of a placebo response, ensure that enrolled participants were stable enough to participate in the study, or some combination of these objectives.

Table 36. Numbers of studies with run-in and wash-out periods by intervention type

Use of Run-In and Wash-Out Periods	Total n (%)	BST n (%)	Psychotherapy n (%)	Pharmacology n (%)	CAM/ Exercise n (%)
Total	163 (100%)	73 (100%)	11 (100%)	71 (100%)	8 (100%)
Run-in	35 (17%)	8 (11%)	4 (30%)	22 (27%)	1 (13%)
Placebo	1 (<1%)	1 (1%)	0 (0%)	0 (0%)	0 (0%)
Active medication	22 (13%)	3 (4%)	2 (20%)	17 (23%)	0 (0%)
Stable medication	9 (2%)	3 (4%)	2 (10%)	3 (0%)	1 (13%)
No treatment	3 (2%)	1 (1%)	0 (0%)	2 (3.1%)	0 (0%)
No run-in	128 (83%)	65 (89%)	8 (70%)	54 (73%)	7 (88%)
Total	163 (100%)	73 (100%)	11 (100%)	71 (100%)	8 (100%)
Wash-out	37 (23%)	13 (18%)	1 (9%)	23 (32%)	0 (0%)
Medication-free	25 (15%)	9 (12%)	0 (0%)	16 (23%)	0 (0%)
Taper	10 (7%)	4 (6%)	1 (9%)	5 (7%)	0 (0%)
Immediate discontinuation	2 (1%)	0 (0%)	0 (0%)	2 (3%)	0 (0%)
No wash-out	126 (77%)	60 (82%)	10 (91%)	48 (68%)	8 (100%)

BST = brain stimulation therapies; CAM = complementary and alternative medicine; n = number.

Of these 163 studies, 37 (23%) required a wash-out period before randomization. Most (n=25) required stopping some existing medications; some (n=10) required a medication taper; and a few (n=2) required immediate discontinuation of medication. Whereas 32 percent of pharmacology studies included a wash-out period, only 18 percent of BST studies and 9 percent of psychotherapy studies did so. No CAM/exercise trial included a wash-out period.

Study Duration

The length of included studies ranged from less than 2 weeks (n=5) to more than 4 years (n=9). Taking all studies into consideration (Table 37), the highest proportion ranged from more than 1 month to 2 months (36%); the next commonly used lengths of studies were more than 2 months to 3 months (14%) and more than 2 weeks to 1 month (12%). Twelve studies did not report study duration.

Study duration varied by some intervention types. At one end of the spectrum, 61 percent of BST studies lasted 2 months or less, whereas at the other end, more than 36 percent of psychotherapy studies lasted more than 1 year.

Table 37. Numbers of studies by study duration and intervention type

Duration of Studies	Total n (%)	BST n (%)	Psychotherapy n (%)	Pharmacology n (%)	CAM/ Exercise n (%)
Total	163 (100%)	73 (100%)	11 (100%)	71 (100%)	8 (100%)
≤ 2 weeks	5 (3%)	1 (1%)	0 (0%)	4 (6%)	0 (0%)
<2 weeks to 1 month	20 (12%)	16 (22%)	0 (0%)	4 (6%)	0 (0%)
>1 month to 2 months	59 (36%)	28 (38%)	1 (9%)	27 (37%)	3 (4%)
>2 months to 3 months	23 (14%)	6 (8%)	2 (18%)	11 (15%)	4 (6%)
>3 months to 4 months	15 (9%)	3 (4%)	2 (18%)	10 (14%)	0 (0%)
>4 months to 6 months	8 (5%)	4 (6%)	1 (9%)	3 (4%)	0 (0%)
>6 months to 8 months	3 (2%)	0 (0%)	1 (9%)	1 (1%)	1 (1%)
>6 months to 1 year	1 (<1%)	1 (1%)	0 (0%)	0 (0%)	0 (0%)
>1 year to 2 years	5 (2%)	2 (3%)	2 (18%)	1 (1%)	0 (0%)
>2 years to 3 years	2 (1%)	2 (3%)	0 (0%)	0 (0%)	0 (0%)
>3 years to 4 years	1 (<1%)	0 (0%)	1 (9%)	0 (0%)	0 (0%)
>4 years	9 (6%)	5 (7%)	1 (9%)	3 (4%)	0 (0%)
Not reported	12 (7%)	5 (7%)	0 (0%)	7 (10%)	0 (0%)

BST = brain stimulation therapies; CAM = complementary and alternative medicine; n = number.

Key Question 9: Risk Factors or Other Patient Characteristics Specifically for Treatment-Resistant Depression

Concerns With Risk or Prognostic Factors

An unequal distribution of risk or prognostic factors at baseline can lead to selection bias and confounding in controlled studies. Because in any given study the same risk or prognostic factors can act as a confounder for one outcome variable but not for another, in the following sections we refer to them collectively as *potential confounders*.

Table 38. Risk and prognostic factors that can act as potential confounders

Risk or Prognostic Factor
Age
Chronic pain
Class(es) of previous antidepressant(s)
Medical comorbidities
Psychiatric conditions
Disease severity
Dose of previous antidepressant
Duration of current episode
Family history of depressive disorder
History of bipolar disorder
Interferon or glucocorticoid treatment
Marital status
Melancholic features
Number of previous hospitalizations
Number of prior (failed) treatments
Onset of disease before age 20
Race and ethnicity
Severe, sudden depression during past 3 years
Sex or gender
Socioeconomic status
Suicidal ideation or attempts

Table 38 presents risk and prognostic factors of TRD that could act as potential confounders. We developed this list from an analysis of published systematic reviews and guidelines (KQ 5, in the previous chapter) and a discussion with Centers for Medicare & Medicaid Services (CMS) staff.

Methodologically and statistically, analysts can use four main approaches to minimize the effect of potential confounders: (1) randomization, (2) restriction, (3) stratification, and (4) statistical adjustment.

Randomization refers to allocating individuals (or clusters of individuals) to treatment groups “at random.” The advantage of randomization is that known and unknown potential confounders will be distributed equally across treatment groups if the sample size of a study is large enough.

Randomization, however, cannot guarantee the absence of confounding, particularly in studies with small sample sizes.

Restriction refers to selectively including patients for a study who have similar risk or prognostic factors. Restriction is usually achieved through inclusion or exclusion criteria that define study populations. Restriction leads to homogenous study populations, but it also limits the generalizability (or applicability) of results to populations excluded from a particular study. Restriction cannot control for unknown confounders.

Stratification refers to analyzing data statistically within certain categories. In studies this is usually achieved by subgroup analyses, which should be defined a priori. Like restriction, stratification cannot control for unknown confounders.

Statistical adjustment refers to using statistical methods (usually regression analyses) that control for the unequal distribution of potential confounders across treatment groups. Statistical approaches other than regression analyses are propensity score matching and inverse probability weighting.

In the following sections, we first provide an overview of the designs of the 163 included studies. We then examine to what extent studies employed each of the four strategies to minimize potential confounding.

Key Points

1. A considerable majority of studies (88%) used randomization as a means to control for potential confounders.
2. All studies applied some exclusion criteria that limited potential confounders. Number of prior failed treatments, psychiatric and medical comorbidities, and severity of disease were the most commonly applied restriction factors to achieve homogeneous study populations.
3. Several studies (20%) stratified analyses by potential confounders. Generally, these factors were age, sex, or gender; number of prior failed treatments; and duration of current depressive episode.
4. Of 20 nonrandomized studies, only 8 reported statistical techniques to control for potential confounding.

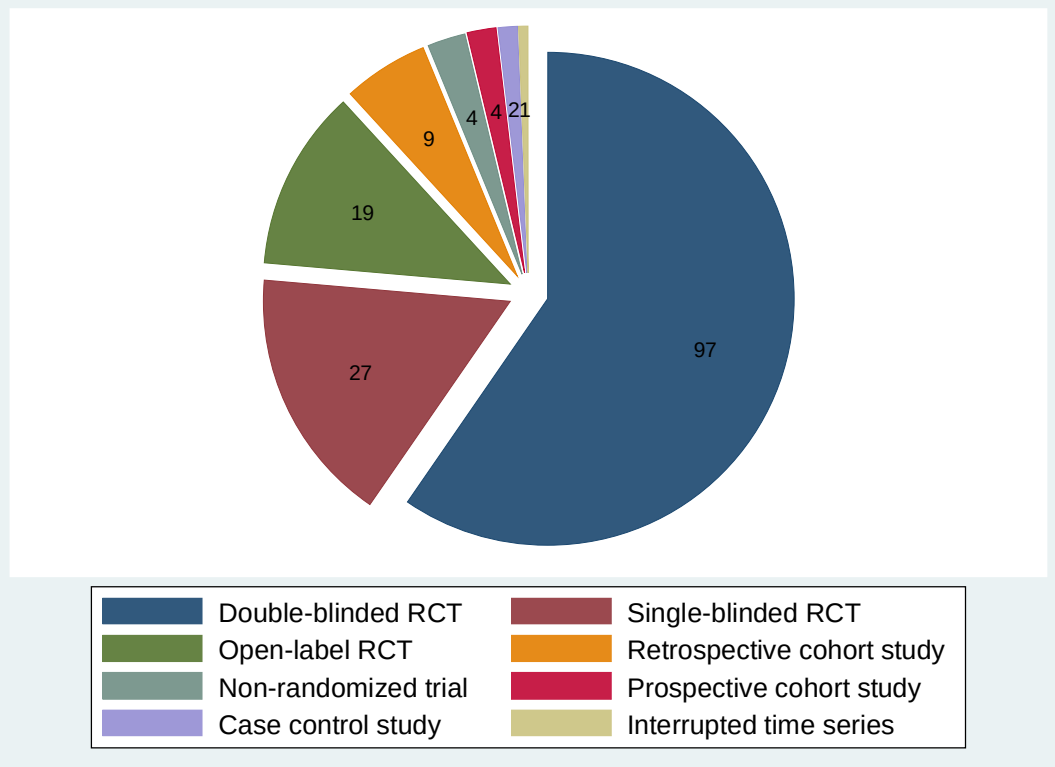
Detailed Synthesis

Overview of Studies of Treatment-Resistant Depression

For this KQ, we included 163 studies that met our eligibility criteria, providing data on BST, pharmacologic therapies including ketamine, psychological interventions, and CAM or exercise interventions. Figure 2 presents an overview of their methodological designs; of these, 143 studies were RCTs of various sorts; some were nonrandomized trials, and a handful were prospective or retrospective cohort studies, case-control studies, or interrupted time series.

Sample sizes ranged from 5 to 3,052 patients; the median sample size was 60 participants.

Figure 2. Overview of study designs and number of studies for treatment-resistant depression



RCT = randomized controlled trial.

Randomization

Of 163 studies, 143 (88%) used randomization as a means to control for potential confounders. Of this evidence base, 97 (68%) were double blinded, 27 (19%) were single blinded, and 19 (13%) were open label.

Critical appraisal of the randomization methods revealed that in only three cases were randomization methods clearly inadequate. These studies used alternation or nonrandom number tables as assignment methods. Randomization methods were adequate in 74 RCTs (51%); in 66 RCTs (48%), the randomization methods were not reported adequately.

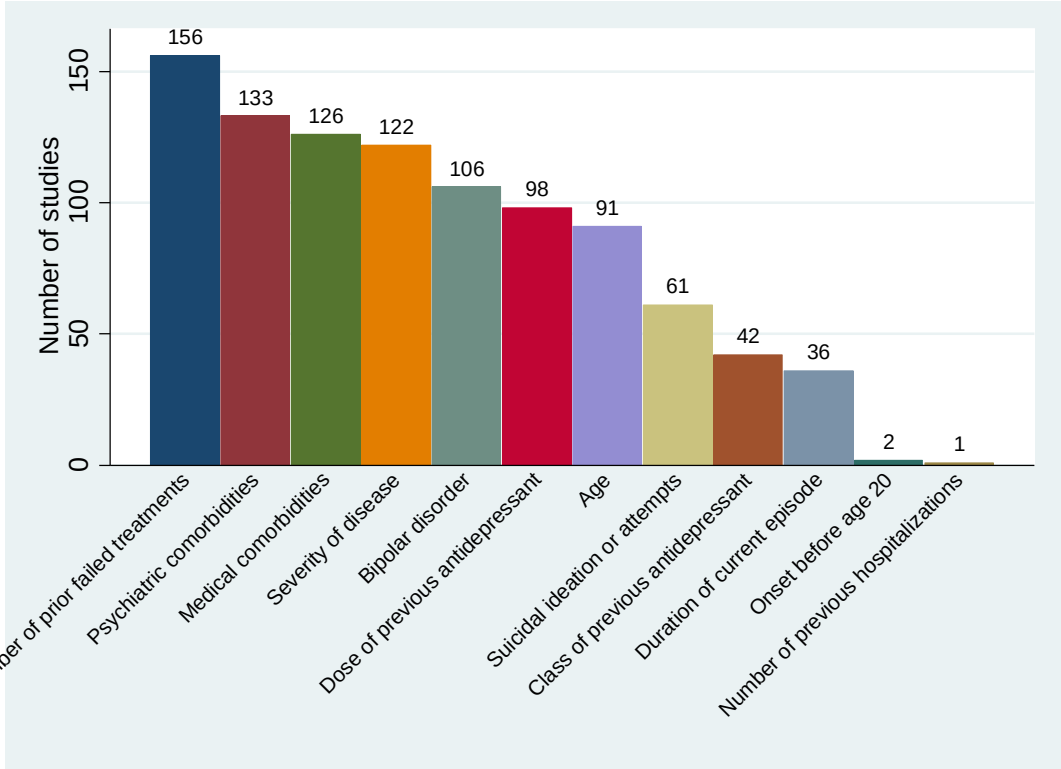
An assessment of the patient characteristics at baseline revealed that in 28 RCTs (20% of all RCTs), potential confounders were not distributed equally across treatment groups after randomization. Most commonly, these studies had substantial differences between treatment groups in age, baseline severity of depression, duration of illness, or length of the current episode. Ideally, authors would explore such differences by adjusting statistically in their analyses. Only 2 of these 27 RCTs, however, reported results that adjusted statistically for differences in baseline characteristics.

Restriction

All studies applied inclusion and exclusion criteria to achieve homogeneous populations. We assessed to what extent studies used potential confounders from Table 37 as inclusion or exclusion criteria to determine their study populations.

Figure 3 presents the numbers of studies that applied such criteria representing these potential confounders. Most studies used more than one of these restriction factors. Therefore, the sum of studies in the figure is larger than the number of included studies for this KQ.

Figure 3. Numbers of studies that used various potential confounders as criteria for inclusion or exclusion of potential study participants



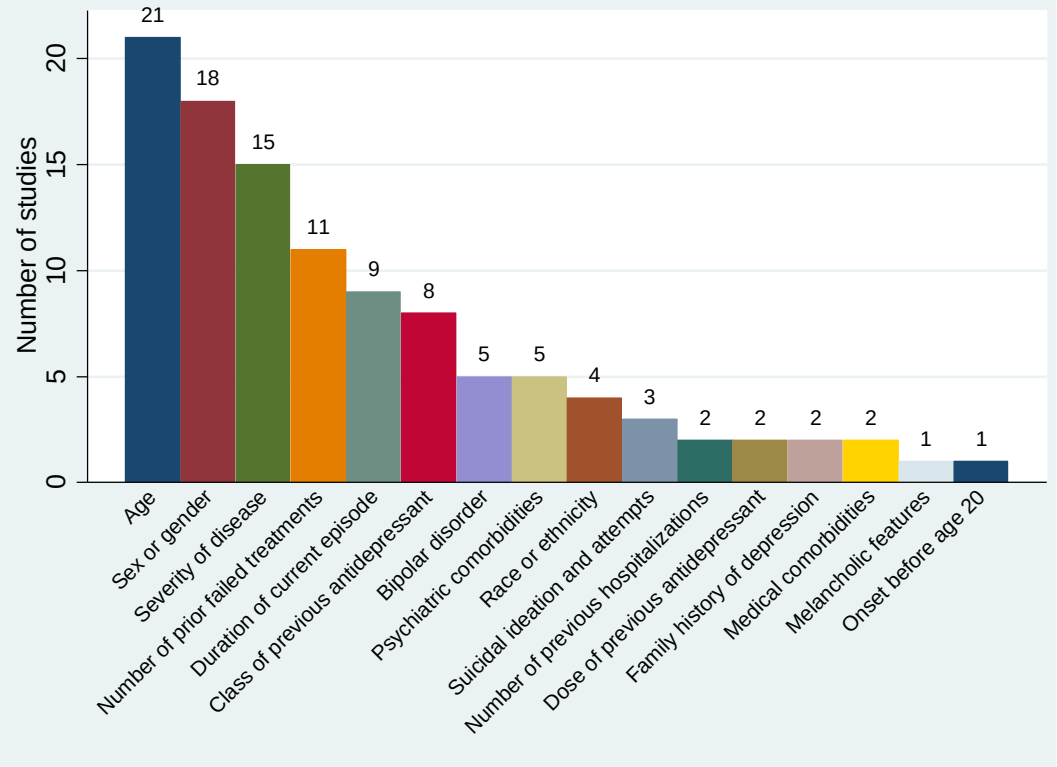
The two most common potential confounders were number of prior failed treatments for TRD and psychiatric comorbidities. The next two frequently encountered confounders were medical comorbidities and severity of disease.

Although marital or socioeconomic status, race or ethnicity, and sex or gender are factors that could act as potential confounders and are listed in Table 37, using them as criteria to determine inclusion or exclusion would not be ethical. Consequently, no study used these factors as inclusion or exclusion criteria. Furthermore, no study restricted its population by chronic pain, family history of depression, melancholic features, onset of severe or sudden depression during the past 3 years, or interferon and glucocorticoid treatment.

Stratification

Several studies stratified the population and conducted one or more subgroup analyses. Figure 4 presents the number of studies that conducted such analyses stratifying their patient populations by one of the potential confounders listed in Table 37. No study conducted subgroup analyses on chronic pain, socioeconomic or marital status, onset of severe or sudden depression during the past 3 years, or interferon and glucocorticoid treatment.

Figure 4. Number of studies that conducted subgroup analyses stratifying by various potential confounders



Statistical Adjustment

Nineteen studies did not use randomization as a means to minimize the effect of potential confounders. Of these, eight studies reported statistical techniques to control for potential confounding. Seven studies used statistical adjustment of confounders during analyses—adjusting, for example, for baseline differences in severity of disease or age; the eighth study used propensity score matching to minimize the effect of potential confounders. Eleven studies did not report any statistical techniques to control for potential confounders.

Key Question 10: What are relationships between risk factors or placebo response on results of studies?

Risk factors for disease can sometimes also act as prognostic factors for response, remission, discontinuation of treatment, or other measures of treatment effectiveness. In general, risk factors are characteristics that are associated with *causing a specific condition*. Prognostic factors are characteristics that influence the outcome in patients who *already have the condition*. For example, a history of previous depressive episodes is a risk factor for developing another depressive episode in the future but is also a prognostic factor for treatment response. Patients with previous depressive episodes, on average, do not respond as well to treatments as patients with no previous depressive episodes. By contrast, female sex is a risk factor for depression but not a prognostic factor for treatment effectiveness of antidepressants. Men and women,

generally, respond equally well to antidepressant treatments and have similar risks for most adverse events.

Similar to patient characteristics, study characteristics can also have an impact on the magnitude of treatment effects in studies. For example, studies with high risk of bias or studies funded by the industry tend to show larger treatment effects than studies with low risk of bias or publicly funded studies.[139, 140]

This KQ explores whether patient-level risk factors or specific study-level characteristics have an impact on results of studies in patients with TRD. It also assesses the impact of placebo response on treatment effects and the relationship between placebo response and study duration.

Table 39 presents patient- and study-level factors that we took into consideration for this KQ. For patient-level factors, we developed this list from an analysis of published systematic reviews and guidelines (KQ 5) and a discussion with CMS staff. For study-level characteristics we relied on available methods research.

Table 39. Potential prognostic factors for treatment-resistant depression treatment success

Potential Prognostic Factors	Variables in Regression Model
Age	*Lack of variation of data*
Age 65 or older	Inclusion of older adults in study
Bipolar disorder	Proportion of patients with bipolar disorder
Coexisting psychiatric comorbidities	*Lack of data*
Coexisting medical comorbidities	*Lack of data*
Duration of current depressive symptoms	*Lack of variation of data*
Female sex	Proportion of female patients
Funding source	Funding source
Number of prior (failed) treatments	*Lack of data*
Onset of depression before age 20	*Lack of data*
Race or ethnicity	Proportion of nonwhite patients
Severity of depression	Proportion of patients with severe depression
Socioeconomic status	*Lack of data*
Risk of bias	Risk of bias
Study design	*Lack of variation of data*
Study duration	Study duration
Study sample size	Study sample size

Because we did not have access to individual patient data of included studies, we converted patient-level factors to study-level factors. For example, we converted “severity of depression” to “proportion of patients with severe depression”. In several instances (e.g., age or duration of current depressive symptoms), the available studies did not provide enough data or the variation of data was too low to make such conversions in a meaningful way. Table 38 also presents the variables that we were able to use in the regression model; here, we italicize those variables that we could not use.

Key Points

Forty-two studies provided data on two comparisons of interest, namely rTMS compared with sham rTMS (25 studies) and pharmacologic treatments compared with pharmacologic treatments plus augmentation (17 studies).

1. For most risk factors that might influence treatment response, data were either insufficient for regression analyses or reflected no statistically significant impact on study results.

2. In a comparison of pharmacotherapy with pharmacotherapy plus augmentation with a second medication, multivariable analyses indicated that female sex had a significant effect on discontinuation; studies with 60 percent or more female participants had statistically significantly higher discontinuation rates because of adverse events (ratio of odds ratios [ROR] 2.81; 95% confidence interval [CI] 1.04 to 7.59) than studies with fewer than 60 percent females.
3. A smaller placebo response was associated with a statistically significantly larger treatment effect for response (p=0.027), remission (p=0.001), and discontinuation because of adverse events (p=0.010). Study duration did not have an impact on placebo response.

Description of Included Studies

Out of 163 unique studies of interventions in TRD populations, 42 comparisons were similar enough with respect to interventions and control interventions to warrant regression analyses to assess the impact of patient- and study-level characteristics on study results. These studies addressed two treatment categories: (1) rTMS versus sham rTMS (25 studies) and (2) pharmacotherapy versus pharmacotherapy plus augmentation (17 studies). Our outcomes of interest were response, remission, relapse, overall risk of adverse events, risk of serious adverse events, discontinuation because of adverse events, and suicidal ideation and attempts. Because of lack of data, we could not assess relapse and suicidal ideation and attempts.

Detailed Synthesis

Relationships Between Risk Factors and Results of Included Studies

Out of the 17 variables of interest presented in Table 38, data were sufficient to conduct regression analyses on eight potential prognostic factors for studies comparing rTMS with sham rTMS: depression severity, funding source, proportion of participants 65 years or older, proportion of female participants, proportion of participants with bipolar disease, risk of bias, sample size, and study duration.

Table 40 presents results of bivariable and multivariable regression analyses. The outcome measure is the ROR, which compares the treatment effect (odds ratio) of studies with a specific potential prognostic factor (covariate) with the treatment effect of studies without this covariate. For example, in Table 39 the ROR for response for the proportion of female participants in the bivariable analysis is 2.55 (95% CI, 1.16 to 5.64). This can be interpreted that studies with more than 60 percent of female participants had, on average, odds ratios of response that were twice as large as studies with fewer than 60 percent female participants.

In bivariable analyses, several statistically significant differences emerged (bolded in Table 39). Studies with sample sizes larger than 60 participants had statistically significantly smaller response (ROR, 0.38; 95% CI, 0.17 to 0.84) and remission rates (ROR, 0.14; 95% CI, 0.03 to 0.62) than studies with fewer than 60 participants.

Table 40. Results of bivariable and multivariable regression analyses of potential prognostic factors for studies comparing rTMS with sham rTMS

Potential Prognostic Factor	Response ROR (95% CI)	Remission ROR (95% CI)	Risk of Serious Adverse Events ROR (95% CI)	Discontinuation Because of Adverse Events ROR (95% CI)
Bivariable analyses				
Risk of bias (high vs. low or moderate risk of bias)	0.69 (0.16, 2.99)	1.61 (0.14, 18.66)	0.98 (0.06, 17.73)	Nonestimable[a]
Funding source (public vs. industry funding)	1.58 (0.71, 3.52)	2.72 (0.87, 8.54)	1.72 (0.20, 14.92)	Nonestimable[a]
Proportion of patients with severe depression (severe vs. mild or moderate)	2.32 (0.74, 7.23)	4.16 (0.69, 24.99)	0.85 (0.14, 5.05)	0.91 (0.14, 5.75)
Inclusion of older adults (studies with patients 65 years or older vs. studies without)	1.09 (0.30, 3.97)	0.71 (0.03, 14.52)	Nonestimable[a]	Nonestimable[a]
Study sample size (n ≥60 participants vs.<60 participants)[b]	**0.38 (0.17, 0.84)**	**0.14 (0.03, 0.62)**	0.39 (0.03, 5.51)	1.43 (0.16, 12.51)
Study duration (≥6 weeks vs. <6 weeks)[b]	0.56 (0.24, 1.30)	0.44 (0.09, 2.01)	0.45 (0.03, 6.38)	1.67 (0.19, 14.66)
Proportion of female participants (≥60% vs. <60%)[b]	**2.55 (1.16, 5.64)**	**5.78 (1.52, 21.92)**	1.08 (0.13, 9.15)	1.62 (0.21, 12.42)
Proportion of participants with bipolar disorder (≥15% vs.<15%)[b]	**3.26 (1.27, 8.39)**	Nonestimable[a]	Nonestimable[a]	Nonestimable[a]
Multivariable analysis				
Study sample size (n ≥60 participants vs.<60 participants)[b]	0.68 (0.13, 3.52)	0.27 (0.03, 2.33)	NA	NA
Proportion of female participants (≥60% vs. <60%)[b]	1.85 (0.50, 6.86)	2.31 (0.34, 15.56)	NA	NA
Proportion of participants with bipolar disorder (≥15% vs.<15%)[b]	1.57 (0.29, 8.37)	NA	NA	NA

[a] Nonestimable indicates that the model could not be estimated because of a lack of variation for the characteristic among studies reporting the outcome.

[b] The variable was dichotomized around the median.

CI = confidence interval; NA = not applicable; ROR = ratio of odds ratios; rTMS = repetitive transcranial magnetic stimulation; vs. = versus.

Studies with 60 percent or more of female participants had statistically significantly larger response (ROR, 2.55; 95% CI, 1.16 to 5.64) and remission (ROR, 5.78; 95% CI, 1.52 to 21.92) rates than studies with fewer than 60 percent females.

Likewise, studies with 15 percent or more patients with bipolar disorder had statistically significantly larger response rates (ROR, 3.26; 95% CI, 1.27 to 8.39) than studies with fewer than 15 percent of patients with bipolar disorder.

In multivariable analyses, however, none of these factors remained statistically significant (Table 40).

For the comparison of pharmacotherapy with pharmacotherapy plus augmentation, data were sufficient to conduct regression analyses on four potential prognostic factors: depression severity, proportion of female participants, sample size, and study duration.

Table 41 presents results of bivariable and multivariable regression analyses. In bivariable analyses, three statistically significant differences emerged (bolded in Table 40). Studies that included patients with severe depression had statistically significantly higher rates of discontinuation because of adverse events (ROR, 13.33; 95% CI, 1.32 to134.15) than studies that included patients with mild or moderate depression.

Second, studies with a study duration of 6 weeks or longer reported significantly lower response rates than studies shorter than 6 weeks (ROR, 0.36; 95% CI, 0.14 to 0.93). Third, studies with 60 percent or more female participants had statistically significantly higher discontinuation rates because of adverse events (ROR, 3.71; 95% CI, 1.56 to 8.84) than studies with fewer than 60 percent females. In multivariable analyses, this was the only factor that remained statistically significant after adjusting for other covariates (ROR, 2.81; 95% CI, 1.04 to 7.59; Table 41). These findings, however, need to be interpreted cautiously.

Table 41. Results of bivariable and multivariable regression analyses of potential prognostic factors for studies comparing pharmacotherapy with pharmacotherapy plus augmentation

Potential Prognostic Factor	Response ROR (95% CI)	Remission ROR (95% CI)	Risk of Serious Adverse Events ROR (95% CI)	Discontinuation Because of Adverse Events ROR (95% CI)
Bivariable analyses				
Proportion of patients with severe depression (severe vs. mild or moderate)	0.96 (0.57, 1.59)	1.10 (0.65, 1.84)	0.54 (0.02, 15.80)	**13.33 (1.32, 134.15)**
Study sample size (n ≥60 participants vs. <60 participants)[a]	0.45 (0.15, 1.34)	0.75 (0.23, 2.48)	Nonestimable[b]	2.94 (0.55, 15.71)
Study duration (≥6 weeks vs. <6 weeks)[b]	**0.36 (0.14, 0.93)**	0.52 (0.17, 1.62)	0.51 (0.08, 3.30)	2.20 (0.70, 6.93)
Proportion of female participants (≥60% vs. <60%)[a]	0.93 (0.61, 1.42)	0.88 (0.58, 1.32)	0.72 (0.15, 3.46)	**3.71 (1.56, 8.84)**
Multivariable analysis				
Proportion of patients with severe depression (severe vs. mild or moderate)	NA	NA	NA	10.23 (0.96, 108.51)
Proportion of female participants (≥60% vs. <60%)[a]	NA	NA	NA	**2.81 (1.04, 7.59)**

[a] The variable was dichotomized around the median.

[b] Nonestimable indicates that the model could not be estimated because of a lack of variation for the characteristic among studies reporting the outcome.

CI = confidence interval; NA = not applicable; ROR = ratio of odds ratios; vs. = versus.

The Influence of Placebo Response on the Magnitude of Treatment Effect

Only the body of evidence comparing rTMS with sham rTMS provided data to address this question. We conducted regression models to explore the impact of placebo response on estimates of treatment effect of rTMS. We tested whether the effect of treatment varied according to the placebo response in the study. Among rTMS studies, we found a significant variation in treatment effect according to levels of placebo response. The smaller the placebo response, the larger the treatment effect; and the larger the placebo response, the smaller the treatment effect. This finding was true for response (p=0.027), remission (p=0.001), and discontinuation because of adverse events (p=0.010). No significant variation was found across levels of placebo response for the risk of serious adverse events (p=0.369).

Overall, these findings are not surprising because outcome measures in controlled trials are mathematically essentially the difference of effects or the ratio of events between intervention and control groups. Given that the treatment effect in the intervention group is the sum of a placebo effect and an actual treatment effect, it is easier to achieve a larger difference between treatment and control groups if the placebo effect is small.

Study Duration as a Moderator of Placebo Response

In cases where we observed significant variation of treatment effect by placebo response interaction was significant, we tested whether study duration moderated the impact of placebo response on the treatment effect, using a three-way interaction between treatment, proportion of placebo responders, and study duration. The relationship between placebo response and treatment effect did not vary by study duration for any of the outcomes.

Key Question 11: Variables or Information Used to Define Endpoints

This final systematic review KQ is intended to summarize our findings from the 163 included studies about the wide array of variables or other information that the investigators used in defining their outcomes or endpoints. We drew on measures examined in the narrative review KQ 3 for some of these analyses.

In addition, we determined whether any studies recorded information about the following: occurrence of adverse events; attrition from care attributed to either adverse events or lack of efficacy; time to relapse; adherence to treatment; changes in any factors that patients might have deemed salient; changes in employment or disability status; and changes in the use of health care resources, such as hospital admissions, emergency room use, or physician visits. We note that the numbers of studies are not necessarily mutually exclusive (i.e., they will not sum to 163) because investigators often used more than one of these instruments or interviews in a single study.

We first list key points below and then present a detailed synthesis of our analyses. The latter focuses on the following: (a) endpoints, such as outcomes that are specific to depression, those reflecting general psychiatric status, and those identifying functional impairment or quality of life, and (b) additional outcomes specifically requested by CMS. For each, we indicate the frequency with which these various measures are reported in the eligible trials or other studies. As with previous questions, we report these analyses in terms of studies investigating BST, psychotherapy, pharmacology, and CAM/exercise interventions.

Key Points

1. The two most common outcome measures used to assess depression were the HAM-D and the MADRS. The HAM-D was the most common depression instrument used across all interventions, and the MADRS was used in one-half of the pharmacology studies.
2. Assessment of manic outcomes was rare.
3. The CGI scale was the most common general psychiatric outcome reported, nearly always in pharmacology studies and slightly less than half the time in BST studies.
4. Functional impairment and quality-of-life outcomes were infrequently reported.
5. Adverse events were commonly reported for BST and pharmacology studies but not in either psychotherapy or CAM/exercise studies.
6. Other than in psychotherapy studies, adherence to treatment was not commonly measured.
7. Overall attrition was a commonly reported outcome, but specific attributions of attrition (e.g., to adverse events or lack or efficacy) were less commonly described.
8. Disability status, time to relapse, and use of health care services were very rarely reported.

Detailed Synthesis

Common Clinical Endpoints and Outcomes

We encountered myriad endpoints in the included studies. We classified them mainly as those specific to depression, those specific to mania, those relevant for psychiatric conditions broadly defined, and those more widely applied for assessing quality of life or functional impairment. The measures that studies most commonly used or reported on were the following (in alphabetical order, spelled out for ease of reference and with typical acronyms by which they are usually known):

- The Beck Depression Index (BDI), in several versions
- The Brief Psychiatric Rating Scale (BPRS)
- The Clinical Global Impression (CGI) scales—CGI-I for improvement and CGI-S for severity
- The General Assessment of Functioning (GAF)
- The Hamilton Rating Scale for Depression (HAM-D), in several versions depending on the number of items used
- The Inventory of Depressive Symptomatology (IDS) or the Quick IDS (QIDS), in two versions relating to self-rated (SR) or clinician rated (C)
- The Montgomery-Åsberg Depressive Rating Scale (MADRS)
- The Sheehan Disability Scale (SDS)
- The Short Form Health Surveys (SF), in two main versions depending on the number of items used (SF-36 or SF-12)
- The Young Mania Rating Scale (YMRS)

In addition to these measures, research teams sometimes used other questionnaires or interview assessment tools as endpoints. These appeared, however, in fewer than five studies and we do not report further on them. They included the following (also in alphabetical order, with acronyms when the instruments are reasonably well known by them):

- Apathy Evaluation Scale
- Center for Epidemiologic Studies Depression Scale (CES-D)
- Generalized Anxiety Disorder scale (GAD-7)
- Geriatric Depression Scale
- Melancholia Scale
- Mini–Mental State Examination (MMSE)
- Multidimensional Assessment of Fatigue
- Patient Health Questionnaire (PHQ-9)
- Quality of Life Enjoyment and Satisfaction Questionnaire (Q-LES-Q)
- Structured Clinical Interview for Depression (SCID)
- Symptom Checklist-90-revised (SCL-90-R)

In Table 42 we report frequencies of use of the measures listed first (above) for the four main categories of interventions.

Table 42. Numbers of studies using common measures of endpoints, by type of intervention for treatment-resistant depression

Measures	BST	Psychotherapy	Pharmacology	CAM/Exercise
Depression-Specific Measures				
HAM-D (all)	63 (86.3%)	8 (72.7%)	43 (60.6%)	7 (87.5%)
HAM-D_6	3 (4.1%)	0 (0%)	0 (0%)	1 (12.5%)
HAM-D_{17}	35 (47.9%)	5 (45.5%)	35 (49.3%)	6 (75.0%)
HAM-D_{21}	11 (15.1%)	2 (18.2%)	8 (11.3%)	0 (0%)
HAM-D_{24}	8 (11.0%)	1 (9.1%)	0 (0%)	0 (0%)
HAM-D_{28}	6 (8.2%)	0 (0%)	0 (0%)	0 (0%)
MADRS	28 (38.4%)	1 (10.0%)	35 (49.3%)	1 (12.5%)
BDI (all)	27 (37.0%)	7 (63.6%)	5 (7.0%)	2 (25.0%)
BDI	16 (21.9%)	4 (36.4%)	4 (5.6%)	0 (0%)
BDI-II	8 (11.0%)	3 (27.3%)	0 (0%)	2 (25.0%)
BDI-SF	3 (4.1%)	0 (0%)	1 (1.4%)	0 (0%)
IDS/QIDS	12 (16.4%)	0 (0%)	17 (23.9%)	3 (37.5%)
IDS-SR_{30}	5 (6.8%)	0 (0%)	4 (5.6%)	0 (0%)
IDS-C_{30}	5 (6.8%)	0 (0%)	0 (0%)	1 (12.5%)
QIDS-SR_{16}	2 (2.7%)	0 (0%)	12 (16.9%)	1 (12.5%)
QIDS-C_{16}	0 (0%)	0 (0%)	1 (1.4%)	1 (12.5%)
Mania-Specific Measure				
YMRS	5 (6.8%)	0 (0%)	2 (2.8%)	0 (0%)
General Psychiatric Measures				
CGI (all)	32 (43.8%)	3 (27.3%)	64 (90.1%)	4 (50.0%)
CGI-I	20 (27.4%)	1 (9.1%)	33 (46.5%)	3 (37.5%)
CGI-S	12 (16.4%)	2 (18.2%)	31 (43.7%)	1 (12.5%)
BPRS	7 (9.6%)	0 (0%)	7 (9.9%)	0 (0%)
Functional Impairment or Quality-of-Life Measures				
SF (all)	1 (1.4%)	1 (9.1%)	6 (8.5%)	1 (12.5%)
SF-12	0 (0%)	1 (9.1%)	4 (5.6%)	0 (0%)
SF-36	1 (1.4%)	0 (0%)	2 (2.8%)	1 (12.5%)
GAF	6 (8.2%)	0 (0%)	2 (2.8%)	2 (25.0%)
SDS	1 (1.4%)	0 (0%)	7 (9.9%)	0 (0%)
Q-LES-Q	0 (0%)	0 (0%)	3 (4.2%)	1 (12.5)

BDI = Beck Depression Inventory; BPRS = Brief Psychiatric Rating Scale; BST = brain stimulation therapies, CAM = complementary and alternative medicine; CGI = Clinical Global Impression; CGI-I or -S = Clinical Global Impression-Improvement or -Severity; GAF = Global Assessment of Functioning; HAM-D = Hamilton Depressive Rating Scale; IDS = Inventory of Depressive Symptomatology (C=clinician–rated, SR = self-rated); MADRS = Montgomery-Åsberg Depressive Rating Scale; QIDS = Quick Inventory of Depressive Symptomatology (C=clinician–rated, SR = self-rated); Q-LES-Q = Quality of Life Enjoyment and Satisfaction Questionnaire; SDS = Sheehan Disability Scale; SF = Short Form; SF-12 = Short Form-12 item version; SF-36 = Short Form-36 item version; YMRS = Young Mania Rating Scale.

Among the depression-specific measures, the HAM-D was the most commonly used measure in studies of all four types of interventions. In total, it was applied in 121 studies. The BDI (in its various versions) and MADRS were also frequently used for BST studies and MADRS for pharmacology studies as well. MADRS was used in 65 studies in all and BDI in 41.

Among the general psychiatric assessments or questionnaires, the CGI was the most used measure, again essentially only for BST or pharmacology studies. For the two general psychiatric measures, the CGI was by far the more commonly used (103 studies in all). Among the quality-of-life or functional impairment measures, the Short Form (SF) measures, the GAF, and the SDS appeared in similar numbers of studies (between seven and nine studies) across the intervention types. Assessment of mania-specific symptoms in TRD was rare.

Additional Outcomes of Interest

We also report below other outcomes of interest reported in TRD studies (Table 43). We sort the table into six main outcomes in these studies: (1) adverse events (which may have been

collected actively, passively, or by a process not clearly described) and related adverse event categories; (2) attrition (with two subcategories of how such attribution was explained); (3) treatment adherence, (4) change in disability status, (5) use of health care resources, and (6) time to relapse. No studies reported change in employment.

Table 43. Numbers of studies reporting on other outcomes or endpoints of interest, by type of intervention for treatment-resistant depression

Additional Outcomes of Interest Measures	BST	Psychotherapy	Pharmacology	CAM/Exercise
Adverse events rates	57 (78.1%)	3 (27.3%)	58 (81.7%)	3 (37.5%)
Active	18 (24.7%)	2 (18.2%)	25 (35.2%)	0 (0%)
Passive	26 (35.6%)	1 (9.1%)	10 (14.1%)	0 (0%)
Unclear	14 (19.2%)	0 (0%)	23 (32.4%)	3 (37.5%)
Serious adverse events rates	29 (39.7%)	2 (18.2%)	41 (57.7%)	0 (0%)
Overall adverse event rates	36 (49.3%)	3 (27.30%)	37 (52.1%)	2 (25.0%)
Attrition overall	48 (65.8%)	11 (100%)	48 (75%)	5 (62.5%)
Attrition attributed to adverse events	28 (38.4%)	2 (18.2%)	45 (70.3%)	2 (25.0%)
Attrition attributed to lack of efficacy	17 23.3%)	1 (9.1%)	22 (34.4%)	1 (12.5%)
Adherence to treatment	20 (27.4%)	5 (45.5%)	12 (16.9%)	1 (12.5%)
Change in disability status	1 (1.4%)	1 (9.1%)	8 (11.3%)	0 (0%)
Use of health care resources	2 (2.7%)	2 (18.2%)	1 (1.4%)	0 (0%)
Time to relapse	7 (9.6%)	0 (0%)	3 (4.2%)	0 (0%)

BST = brain stimulation therapies; CAM = complementary and alternative medicine.

By a wide margin, adverse events (including serious adverse events and overall rates) were reported more often for BST and pharmacology studies. Overall attrition was commonly reported for all interventions, but rates of attrition for either problems of adverse events or lack of efficacy were more frequently reported by BST and drug studies (consistent with the closer monitoring of adverse events for these interventions).

Adherence to treatment was infrequently reported in TRD studies, most commonly in psychotherapy studies (five of 11 studies) but less commonly with other intervention types.

Relatively few studies of any type of intervention reported on the other variables. Eight pharmacology studies reported on change in disability status, and seven BST studies provided information on time to relapse. Otherwise, very few gave findings about use of health care services.

Discussion

This chapter brings together the two parts of this Technology Assessment on treatment-resistant depression (TRD)—the Narrative Review Key Questions (KQs) 1 through 5 and the Systematic Review KQs 6 through 11. Findings for the former draw on (a) literature searches (one of which was systematic) of appropriate databases and (b) gray literature and materials such as various clinical practice guidelines, consensus statements, and other information found on three main websites. For these former topics, we drew on a total of 38 publications, of which 14 were systematic reviews, 6 were non-systematic reviews, and 18 were guidelines or consensus statement.

Findings for the latter represent syntheses that follow systematic literature searches in accord with standard methods for the Evidence-based Practice Center program of the Agency for Healthcare Research and Quality. For these latter topics, we drew on a total of 163 studies (in 197 publications), of which 143 were randomized controlled trials (RCTs) of various designs, 4 were nonrandomized trials, 13 were observational studies, 2 were case controls and 1 was an interrupted time series.

A core theme throughout our syntheses is the degree to which clinicians, researchers, and other experts agree on a wide array of issues in investigating TRD. The variability in definitions of TRD, the different definitions of successful outcomes (e.g., response vs. remission), and the large number of potential risk factors substantially hampered our ability to summarize and synthesize this evidence base.

In the remainder of this chapter, we first present our key findings from both parts of the report. We then discuss core findings in terms of what is already known in the clinical and research realms and literature and explore the clinical and policymaking implications of our work. Additionally, we examine the applicability (i.e., generalizability) of our findings for patient populations, TRD interventions, and ways to measure important outcomes in trials or other studies.

Later sections of this chapter examine the limitations of both the literature itself and our ability to synthesize it adequately to address matters of interest for the Centers for Medicare & Medicaid Services (CMS). Finally, we offer a set of recommendations about needed research to address both gaps in the evidence base and common problems or drawbacks of studies to date.

Key Findings

Narrative Review: Definition of Treatment-Resistant Depression

Applying Definitions or Diagnostic Tools

No consensus definitions exist for TRD. Available definitions are anchored primarily by consideration of three key variables: number of prior treatment failures (the primary consideration), adequacy of prior treatment doses, and adequacy of prior treatment duration. These definitions address TRD mainly as a part of major depressive disorder (MDD); in contrast, within bipolar disorder, TRD definitions have centered on one prior treatment failure.

For these three variables, the most commonly used definition is a continuing depressive episode following at least two prior antidepressants treatments of at least 4 or 6 weeks of an adequate dose; as to defining adequacy, descriptions range from a minimum effective dose to a maximum tolerated dose.

Five staging models for TRD are in use, but they have only a limited evidence base supporting their validity. These models appear equally valid for documenting treatment failure in depressed patients in intervention studies, but their applicability and feasibility in clinical practice are unclear.

Consensus is also lacking about the "best" tool for diagnosing TRD in clinical research; neither is there a "most commonly used" tool. Diagnostic tools used to identify TRD emphasize careful clinical assessment for MDD, but they differ in how structured the assessment is, ranging from standard clinical assessment to a highly structured research tool. Moreover, the evidence base for validity and feasibility of such tools or instruments is limited, and the accuracy of a careful history (more feasible) and a structured tool has never been directly compared. As noted above, the limited evidence base suggests that staging models are equally valid for diagnosing TRD. Outside of feasibility (which affects use of diagnostic tools in all locations to some degree), setting does not appear to substantially influence the choice of which tool to use.

Measuring Outcomes or Endpoints of Studies and Observational Studies

Similarly, we could find no consensus about the best measure(s) to determine success or failure in TRD studies. Outcomes have consisted primarily of depression-specific measures; we did find agreement that remission is the preferred outcome regardless of which tool is used. Investigators have also assessed general psychiatric status, functional impairment, and various domains of quality of life, but patient-oriented outcomes are rarely assessed as primary outcomes.

Both patient-reported and clinician-administered measures are available for each category; we found no stated preference for one type over the other, although patient-reported tools are more feasible to use. The available measures appear to have adequate psychometric properties. However, the degree of validity of the general psychiatric measure Clinical Global Impression (which has two variants) is unclear. The minimally clinically important difference (MCID) has been defined for many of these measures; nevertheless, no clear consensus about a preferred definition for MCIDs has emerged.

Minimizing Bias and Determining Appropriate Research Study Duration

We uncovered some agreement about how best to minimize bias in research studies. Most investigators and expert groups over the past decade preferred randomized designs over nonexperimental ones. Most of the available literature did not address, or apparently achieve consensus about, designs that might minimize placebo effects. Although 6 weeks was a frequently recommended minimum study length, we found no agreement on a preferred duration of trials or observational studies, although experts frequently recommend studies longer than 4 to 6 weeks.

Addressing Risk Factors

Evidence addressing risk factors for TRD was quite limited. Several components of the TRD definition (disease severity, duration of current episode, number of previous hospitalizations, and number of failed antidepressant trials) appeared to be associated with greater risk of TRD. Coexisting anxious symptoms, anxiety disorders, and personality disorders were related to higher

risk of TRD as well, as were specific clinical characteristics such as having melancholic features and suicidality (suicide ideation or attempts).

Systematic Review: Current Clinical Trials and Observational Studies of Treatment-Resistant Depression

Classifying the Main Focus of Trials or Observational Studies

The large majority of trials or observational studies investigating TRD focused on either brain stimulation therapies (BST) or pharmacotherapy interventions. BST approaches included electroconvulsive therapy and repetitive transcranial magnetic stimulation (rTMS). Many fewer studies dealt with either psychotherapy interventions or complementary and alternative medicine (CAM) or exercise (which we generally combined). They were conducted primarily in adult, nongeriatric patient populations (i.e., 18 years of age or older, but relatively few of age 65 or older); study patients tended to have moderate depressive severity as measured by a variety of standardized instruments (e.g., Hamilton Depression Rating Scale [HAM-D]).

Specifying Inclusion or Exclusion Criteria for Study Entry

Aspects of specifying inclusion or exclusion criteria for study entry were rather variable. Confirmation of prior MDD diagnosis and current TRD for study entry were often poorly described. The HAM-D and the Montgomery–Åsberg Depression Rating Scale were the most commonly used instruments to set thresholds for study entry or to measure study outcomes.

Moreover, we found little consistency among studies for the necessary number (or types) of prior treatment attempts for study entry; most studies required at least one, and sometimes two, prior failed treatment attempts of adequate therapy (as described above). Several different patient characteristics were only rarely considered for study entry; these typically "historical" characteristics included duration of depressive symptoms, prior depressive relapses, prior treatment intolerance, prior augmentation or combination therapy, prior psychotherapy, or suicidality.

This variability in study inclusion criteria reflects the variability in TRD definitions reported above for the narrative review. Indeed, inclusion criteria as specified by the eligible TRD trials or observational studies generally did not closely align with TRD definitions identified in the narrative review. For example, although the most common definition of TRD we found in the narrative review involved a minimum of two failed prior treatment attempts with adequate use of an antidepressant, only 38 percent of studies we identified met that definition. Moreover, only 49 percent of studies we identified required at least a single such failed attempt. BST studies were more likely to require a minimum of two or more failed treatment attempts, whereas pharmacologic studies were more likely to require a minimum of at least one failed attempt.

Investigators also did not systematically confirm that other key parts of a TRD diagnosis were part of their inclusion criteria. For example, 77 percent of studies considered adequate dose in their selection criteria, and only 58 percent systematically confirmed that the dose was adequate. Similarly, 82 percent of all studies considered in their selection criteria whether prior treatments were of an adequate duration; of those, only 72% systematically confirmed that the length of such earlier treatment attempts was adequate (≥4 weeks of therapy).

We note that these criteria were not strict. While our narrative review led to considering a minimum adequate treatment duration to be 4 weeks, this duration may be too short, because it barely gives antidepressants (which take approximately 4 weeks to demonstrate a clinical

response at a given dose) enough time for a clear response to be observed. Similarly, our criterion for adequate dose, reflecting what was reported in the literature, set the bar relatively low at what would be considered a minimum therapeutic dose. Such a dose does not reflect what would occur in clinical situations, where dosage would be increased to a moderate level after an initial failure to achieve a robust response.

Even with this relatively lenient working definition, only 17% of studies (27/163) specified and confirmed through eligibility criteria that their population had these three most common components of the current TRD definition: a minimum of two prior treatment failures, a confirmed adequate dose, and a confirmed adequate duration of treatment (≥4 weeks). Relaxing our definition of adequacy only slightly improved this rate. When we relaxed our definition of adequacy to be that a study's inclusion criteria merely *considered* adequacy of dose and duration (but did not systematically define and confirm), only 25% of studies (40/163) had inclusion criteria that met this mark.

Designing Studies

The majority of all the studies from our systematic review evidence base (88%) had a randomized controlled design. A very few were nonrandomized trials; the remainder were an array of observational studies.

Few of the clinical trials had run-in periods (21%) for screening out medication nonadherence, mitigating the effects of a placebo response, and/or ensuring that enrolled participants were stable enough to participate in the study. Similarly, few had wash-out periods (23%) to ensure that participants did not have certain TRD-related medications in their systems before the study began. Study duration differed markedly across these various types of studies. It ranged from less than 2 weeks to more than 4 years; more than one-half lasted 2 months or less.

Controlling for Potential Confounders

The great majority of all these studies (88%) used randomization to control for potential confounders. All studies applied some exclusion criteria that limited potential confounders. Severity of disease, number of prior failed treatments, psychiatric and medical comorbidities, and bipolar disease were the most commonly applied restriction factors to achieve homogeneous study populations.

Several studies stratified analyses by potential confounders, most commonly age, sex or gender, number of prior failed treatments, and duration of current depressive episode. However, as noted above, some of these factors (e.g., severity of disease, number of prior failed treatments, and duration of current episode) were neither consistently nor systematically described. Of 19 nonrandomized studies, only eight sets of investigators reported using any statistical techniques to control for potential confounding.

Addressing Risk Factors and Their Relationships to Outcomes

There was limited information on whether patient-level risk factors or specific study-level characteristics had an impact on results of studies in patients with TRD. In studies comparing rTMS with sham rTMS, bivariable analyses indicated that sample size, female sex, and bipolar disorder showed a statistically significant impact on measures of treatment effects. However, in multivariable analyses, however, none of these variables remained statistically significant.

In studies comparing pharmacotherapy with pharmacotherapy plus augmentation, bivariable analyses indicated that depression severity, female sex, and study duration rendered a statistically

significant impact on measures of treatment effects. In multivariable analyses, however, only the effect of female sex on discontinuation because of adverse events remained statistically significant. Studies with 60 percent or more female participants had statistically significantly higher discontinuation rates because of adverse events (ROR 2.81; 95% CI 1.04 to 7.59) than studies with fewer than 60 percent females.

A smaller placebo response was associated with a statistically significantly larger treatment effect regarding response (p=0.027), remission (p=0.001), and discontinuation because of adverse events (p=0.010). Study duration did not have an impact on placebo response.

Identifying Key Outcomes of Studies

Depressive symptoms were the most commonly reported endpoint across all types of studies; sometimes these are denoted as the severity of depression because of the nature of the measures used to assess them. Most often, the HAM-D was used in these studies, but often the Montgomery–Åsberg Depression Rating Scale was encountered in this body of evidence. The Clinical Global Impression was the most common general psychiatric outcome reported, most often in pharmacology studies. Functional impairment and quality-of-life outcomes were infrequently reported; usually one of four well-known instruments was used, but we saw no consistent patterns for these endpoint measures.

Adverse events were commonly reported for BST and pharmacology studies, but they were not often reported in psychotherapy and CAM/exercise studies. Overall rate of attrition was a commonly reported phenomenon, but specific attributions of attrition (e.g., because of adverse events or lack or efficacy) were less commonly made.

Adherence to treatment was not commonly reported. Disability status, time to relapse, and use of health care services were mentioned or documented only very rarely.

Findings in Relationship to What Is Already Known

The variability in the definitions and conceptualization of TRD (from our narrative review) is quite consistent with other reports from the past decade identifying the lack of any standard, systematic definition of TRD.[18, 31, 44, 47] Taken all together, the available literature highlights the resulting difficulty in synthesizing information across studies (or other types of studies or documents). This characteristic of the evidence base also underscores the problems of translating research findings into guidelines for selecting better treatment options for patients with TRD. Key challenges include the lack of agreement on what constitutes adequacy of dose and duration of prior treatments, what the preferred definition of treatment success or failure is, and how best to confirm TRD both in clinical research and in clinical care.

What the narrative review newly highlights is how the great variability in information from systematic reviews and influential nonsystematic reviews is reflected in the great variability of guidelines and consensus statements about managing patients with TRD. Although a minimum of two prior treatment failures appears most commonly in both systematic reviews and guidelines or consensus statements, defining the adequacy of dose and duration, clarifying failure (as remission or response, and after what length of time), and determining whether TRD requires the prior use of different classes of antidepressants are all variably defined and implemented. This lack of agreement complicates both developing and administering patient management guidelines.

Our systematic review highlighted some key findings not previously described. The mismatch between the most common number of treatment failures (at least two) and what most

of the recent literature has assessed (at least one) was stark. Also, the failure of inclusion criteria in recent TRD studies to confirm systematically both adequate dose (58%) and duration (72%) has not been described previously, nor has the finding that only 17 percent of recent intervention studies are consistent with the most common definition of TRD. These results highlight another concern about how to compare and synthesize data across treatment studies.

Finally, despite the substantial morbidity associated with TRD, the relative infrequency of use of patient-oriented outcomes such as functional impairment and quality-of-life measures in considering the benefits of TRD treatment was newly demonstrated. So, too, was the infrequent measurement of both adherence to treatment and health care services use.

Implications for Clinical and Policy Decisionmaking

This current state of evidence underscores the challenges facing clinicians. A substantial body of literature addresses TRD, but the absence of standard, systematized identification and management approaches in the TRD database makes it difficult to translate this evidence efficiently to establish clinical practice guidelines for care. The greatest concern is the heterogeneity in identifying TRD per se and in how (and when) to determine treatment failure. Effective treatments exist, but because of this variability, determining to which TRD patients the results apply is difficult.

Similarly, the state of the evidence poses challenges for policymakers. Officials at CMS, other public agencies, and private-sector organizations must be confident that two main assumptions are being met: first, that the population of patients with TRD is being consistently and systematically defined, and second, that meaningful and comparable outcomes of importance to both patients and clinicians are being monitored. Neither is consistently reported in the literature, limiting translation of this treatment information into actual care. Given the high levels of significant health effects within the TRD population, the administration of effective interventions is vital.

Despite this variability, we see several important implications from the available evidence base. The high level of morbidity associated with TRD is clear, and the literature suggests that, at a minimum, TRD can be understood as two or more prior treatment failures of an adequate treatment dose (at least minimally effective) and an adequate treatment duration (approximately 4 or more weeks of treatment). For adequate clinical and policy decisionmaking about TRD patients, however, a widely agreed-upon definition of the condition that addresses how to determine the number of prior treatment failures and the adequacy of dose and duration is critical. Some means of consistently and systematically monitoring TRD on a large scale (e.g., a treatment registry using common data elements in an electronic medical record) could substantially help clarify which criteria best define TRD, what the course of illness is, and how interventions might affect that course.

Limitations of this Technology Assessment

Comparative Effectiveness Review Process

The primary challenge of this process was the broad, comprehensive, and inclusive nature of the topic, which combined a narrative review (for five KQs) and a systematic one (for six KQs). Given how variable the definitions of TRD are in the literature, we needed to cast a wide net for both published and gray literature to assemble the proper universe of sources that could be managed within a reasonable amount of time and resources. This requirement produced

important information but also a considerable amount of materials that proved to be of little or no utility.

Once we identified each article or item from either the peer-reviewed or gray literature, we abstracted large amounts of information. Then, we tried to harmonize information across articles and other sources. However, these data items were not always directly comparable. Critical examples include the following: some articles might say one failure of a prior treatment attempt, others might say one or more, and yet others might say one to three, which sometimes precluded organizing studies into easily recognizable, meaningful, and comparable categories; moreover, some articles systematically define a treatment duration, but others do not.

We addressed this challenge by focusing the 11 KQs and the time periods for the literature searches to reflect the current conceptualization of TRD. For example, we limited our systematic review search to the past 10 years to focus on the contemporary understanding of this serious condition.

Also, some questions proved particularly challenging. For example, KQ 3 has, as part of the question, a consideration of psychometric properties and what amount to MCIDs. These are extremely important concepts, but the scope of this particular technology assessment did not accommodate a comprehensive (systematic) literature search for information about these topics. Accordingly, the information reported was what has been reported in a general (but not systematic) search for relevant reviews or summaries.

The Evidence Base

The primary limitation of this evidence base is the heterogeneity of TRD definitions encountered in both the narrative review of TRD definitions and the systematic review of the studies. In the narrative review, although most study authors used a definition of one or more prior treatment failures, they did not, overall, yield any agreement about a correct measure. More importantly, the definitions of an adequate dose and duration varied considerably. This unevenness was reflected in the inconsistency of inclusion criteria of even the systematically identified clinical trials selected for the systematic review, where even the most frequent count of prior adequate trial failures—two—was not the most common *minimum* used in studies. Agreed-upon definitions of adequacy of dose and duration of prior studies were also absent; this gap perhaps led to the infrequency of systematically and consistently confirming these parts of the definition in clinical trials. Findings also do not account for the high level of heterogeneity of patients who have TRD; we encountered considerable variability in such basic parameters as how many interventions were tried in the past, the types of symptoms and their duration and severity, and presence of numerous coexisting physical or mental health conditions.

As a result, at the core of the limitations in this evidence base is that with no agreed-upon definition of TRD and no consensus on very important outcomes, determining to what population clinical trials results apply is difficult. This heterogeneity will prevent others from synthesizing or combining data, even for the more common TRD interventions such as brain stimulation technologies or medications, to translate findings into clinical practice recommendations.

Research Recommendations

We propose several steps to address existing evidence gaps and substantially improve the study and treatment of TRD. Many of these steps speak to the clinicians and researchers who

work with TRD patients; others are aimed more at organizations that support this type of research. In either case, the points are pertinent to all audiences.

Reducing the heterogeneity of how TRD patient populations are defined is a necessary first step to improving the evidence base. Perhaps the most critical step is to reach agreement on a standardized, systematic, and feasible definition of TRD. Such a definition should clearly specify the number of prior treatment attempts, what an adequate dose is, and what an adequate duration is. At the very least, the minimum number of past failed therapy attempts should be two. Systematic confirmation of adequacy is a necessary part of this "definitional" step.

Systematic, standardized accounting for potential confounders is also crucial. The factors that must be considered include, at a minimum, the following: depressive severity, duration of current episode, prior treatment intolerance, prior augmentation or combination therapy, prior psychotherapy, and psychiatric comorbidities. Other socio-cultural factors that may influence the course of TRD (e.g., gender and age) may also be relevant and deserve further study, as may genetic factors.[42] Randomization can account for some measured and unmeasured confounders in larger trials, but the smaller RCTs that we identified, which had imbalances in baseline characteristics, rarely adjusted for such differences. Moreover, nonrandomized TRD studies adjusted for potential confounders less than one-half the time, for example. Acting on how best to deal with confounders is essential to improving this evidence base.

Agreement on a core package of outcome measures to be administered in a standard manner should be strongly encouraged. The field would benefit from an evidence-informed, multistakeholder consensus process to develop a core outcome set for TRD, potentially something similar to the Outcome Measures in Rheumatology process in rheumatology (https://www.omeract.org/). Of particular importance is including one measure of depressive severity, one measure of general psychiatric status, one measure of functional impairment or quality of life, and one measure of adherence to medications or other interventions. Another key outcome to consider includes a measure of suicidality (to assess for reduction in suicidality). Common use of measures will allow for better comparisons among trials; doing so should improve our ability to combine studies for meta-analyses. Patient-reported instruments may be preferred because they are more feasible, generally speaking, and more patient centered than clinician-reported instruments; examples of such self-report scales addressing functional impairment include the Sheehan Disability Scale.[128]

Researchers and clinicians should come to consensus on a standard length of treatment. The key is to provide enough time for patients to receive an adequate dose and duration of the intervention. Given the chronicity of TRD and the time to reach an adequate dose and length of treatment, at least 2 months is the bare minimum for studies to be conducted. Also, the increasing risk of relapse with greater levels of resistance[7] argues for the need to demonstrate longer-term efficacy for the more-resistant patients, suggesting a need for even longer trials for the more severely resistant. For this latter group, who may have a more chronic and persistent course, different treatments might be required than for those whose depressive episodes are more episodic.

Whether either run-in stages or wash-out periods affect the efficacy or effectiveness of TRD treatments remains unclear. Comparative trials should examine this issue to clarify whether investigators should use one or the other in designing their trials.

We found (and were able to include) only a very few studies of interventions other than pharmacological or BST interventions (that is, psychotherapies and CAM or exercise as remedies for TRD). This gap reduced the evidence base relevant for patients who prefer to avoid,

or for whom it would be inappropriate to try, pharmacological agents or more invasive procedures. Consideration of less-studied interventions could help inform patient decisions about options and improve the level of shared or informed decisionmaking.

Trials or other robust types of observational studies to test the *effectiveness* of all such interventions in real-world settings are necessary. Targeting only efficacy (via RCTs) may produce information for clinicians, patients, or policymakers that cannot easily be applied in "ordinary," every-day circumstances. At the same time, intervention trials should not be limited to such comparative effectiveness methods. Indeed, some experts suggest taking a broader perspective—one that would have the field also pursue basic efficacy research (e.g., "does this new intervention work compared to placebo?") and translational strategies (e.g., "how do we disseminate this intervention that we know works into real-world settings?") to improve clinicians' and policymakers' understanding of TRD. Among the ideas along these lines are that intervention studies should indeed include more modalities (e.g., psychotherapies), but also examine "experimental therapeutics"[141] and a translational approach that may incorporate designs that go beyond traditional trials and feature more patient engagement.

To allow for better assessment of quality, publications of RCTs need to adhere to Consolidated Standards of Reporting Trials (CONSORT) specifications for reporting.[48] Similarly, publications of nonrandomized controlled trials or observational studies should adhere to Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) guidelines.[49] Documenting all steps in such investigations, reporting on all planned outcomes, and otherwise ensuring complete transparency for this work are critical actions in adding to the professional literature. These steps would help ensure a consistent definition of TRD and its reported outcomes.

Finally, TRD needs to be monitored consistently, systematically, and on a large scale. For instance, a treatment registry making use of common data elements in an electronic medical record could substantially help clarify which criteria best define TRD, what the course of illness is, and how interventions might affect that course. Coordination between different specific treatment registries that already exist (e.g., the vagal nerve stimulation registry required by the FDA,[50, 51] and the transcranial magnetic stimulation registry recently launched by Neurostar[52]) and have been suggested (e.g., a ketamine registry[53]) would be a necessary step. Data quality would be a key challenge for such an enterprise.

Conclusions

Our basic assignment from the Agency for Healthcare Research and Quality and CMS was to consider a wide array of "study design" issues relating to trials (or observational studies) of TRD therapies. Across all 11 of the complex topics addressed in this Technology Assessment, we encountered substantial diversity at every stage of research on TRD interventions. Of particular concern was the lack of consensus about various elements of even a TRD diagnosis and appropriate inclusion or exclusion criteria. Additionally, little or no agreement about important outcomes and how to assess them hampered analysis. An extensive set of recommendations about more and more-robust approaches to the design and conduct of this research will foster better evidence to translate into clearer guidelines for treating patients with this serious condition.

References

1. Center for Behavioral Health Statistics and Quality. Results from the 2015 National Survey on Drug Use and Health: Detailed Tables. Rockville, MD: Substance Abuse and Mental Health Services Administration; 2016. http://www.samhsa.gov/data/sites/default/files/NSDUH-DetTabs-2015/NSDUH-DetTabs-2015/NSDUH-DetTabs-2015.pdf. Accessed on September 23, 2016.

2. Kessler RC; Berglund P; Demler O; Jin R; Koretz D; Merikangas KR; Rush AJ; Walters EE; Wang PS. The epidemiology of major depressive disorder: results from the National Comorbidity Survey Replication (NCS-R). JAMA. 2003 Jun 18;289(23):3095-105. PMID: 12813115.

3. Wang PS; Lane M; Olfson M; Pincus HA; Wells KB; Kessler RC. Twelve-month use of mental health services in the United States: results from the National Comorbidity Survey Replication. Arch Gen Psychiatry. 2005 Jun;62(6):629-40. PMID: 15939840.

4. Trivedi MH; Rush AJ; Wisniewski SR; Nierenberg AA; Warden D; Ritz L; Norquist G; Howland RH; Lebowitz B; McGrath PJ; Shores-Wilson K; Biggs MM; Balasubramani GK; Fava M. Evaluation of outcomes with citalopram for depression using measurement-based care in STAR*D: implications for clinical practice. Am J Psychiatry. 2006 Jan;163(1):28-40. PMID: 16390886.

5. Thase ME; Rush AJ. When at first you don't succeed: sequential strategies for antidepressant nonresponders. J Clin Psychiatry. 1997;58 Suppl 13:23-9. PMID: 9402916.

6. Berlim MT; Fleck MP; Turecki G. Current trends in the assessment and somatic treatment of resistant/refractory major depression: an overview. Ann Med. 2008;40(2):149-59. PMID: 18293145.

7. Rush AJ; Trivedi MH; Wisniewski SR; Nierenberg AA; Stewart JW; Warden D; Niederehe G; Thase ME; Lavori PW; Lebowitz BD; McGrath PJ; Rosenbaum JF; Sackeim HA; Kupfer DJ; Luther J; Fava M. Acute and longer-term outcomes in depressed outpatients requiring one or several treatment steps: A STAR*D report. Am J Psychiatry. 2006 Nov;163(11):1905-17. doi: DOI 10.1176/appi.ajp.163.11.1905. PMID: WOS:000241669900014.

8. Perlis RH; Ostacher MJ; Patel JK; Marangell LB; Zhang H; Wisniewski SR; Ketter TA; Miklowitz DJ; Otto MW; Gyulai L; Reilly-Harrington NA; Nierenberg AA; Sachs GS; Thase ME. Predictors of recurrence in bipolar disorder: primary outcomes from the Systematic Treatment Enhancement Program for Bipolar Disorder (STEP-BD). Am J Psychiatry. 2006 Feb;163(2):217-24. doi: 10.1176/appi.ajp.163.2.217. PMID: 16449474.

9. Gibson TB; Jing Y; Smith Carls G; Kim E; Bagalman JE; Burton WN; Tran QV; Pikalov A; Goetzel RZ. Cost burden of treatment resistance in patients with depression. Am J Manag Care. 2010 May;16(5):370-7. PMID: 20469957.

10. Crown WH; Finkelstein S; Berndt ER; Ling D; Poret AW; Rush AJ; Russell JM. The impact of treatment-resistant depression on health care utilization and costs. J Clin Psychiatry. 2002 Nov;63(11):963-71. PMID: 12444808.

11. Ivanova JI; Birnbaum HG; Kidolezi Y; Subramanian G; Khan SA; Stensland MD. Direct and indirect costs of employees with treatment-resistant and non-treatment-resistant major depressive disorder. Curr Med Res Opin. 2010 Oct;26(10):2475-84. doi: 10.1185/03007995.2010.517716. PMID: 20825269.

12. Social Security Administration. SSI Annual Statistical Report, 2011. Washington, DC: Social Security Administration, Office of Retirement and Disability Policy, Office of Research, Evaluation, and Statistics; 2012. http://www.ssa.gov/policy/docs/statcomps/ssi_asr/2011/ssi_asr11.pdf (Table 6, P. 25). Accessed on 27 May, 2016.

13. Fiske A; Wetherell JL; Gatz M. Depression in older adults. Annu Rev Clin Psychol. 2009;5:363-89. doi: 10.1146/annurev.clinpsy.032408.153621. PMID: 19327033.

14. Center for Medicare Advocacy. Medicare and Mental Health. Washington, DC: Center for Medicare Advocacy; n.d. http://www.medicareadvocacy.org/medicare-and-mental-health/. Accessed on 27 May, 2016.

15. Chesney E; Goodwin GM; Fazel S. Risks of all-cause and suicide mortality in mental disorders: a meta-review. World Psychiatry. 2014 Jun;13(2):153-60. doi: 10.1002/wps.20128. PMID: 24890068.

16. National Heart, Lung, and Blood Institute. Who Is at Risk for Heart Disease? Bethesda, MD: National Institutes of Health; 2014. http://www.nhlbi.nih.gov/health/health-topics/topics/hdw/atrisk. Accessed on 27 May, 2016.

17. Khawaja IS; Westermeyer JJ; Gajwani P; Feinstein RE. Depression and coronary artery disease: the association, mechanisms, and therapeutic implications. Psychiatry (Edgmont). 2009 Jan;6(1):38-51. PMID: 19724742.

18. Trevino K; McClintock SM; McDonald Fischer N; Vora A; Husain MM. Defining treatment-resistant depression: a comprehensive review of the literature. Ann Clin Psychiatry. 2014 Aug;26(3):222-32. PMID: 25166485.

19. Association AP. Practice Guideline for the Treatment of Patients With Major Depressive Disorder. Third ed. Washington, DC: American Psychiatric Association; 2010.

20. Lam RW; Kennedy SH; Grigoriadis S; McIntyre RS; Milev R; Ramasubbu R; Parikh SV; Patten SB; Ravindran AV; Canadian Network for M; Anxiety T. Canadian Network for Mood and Anxiety Treatments (CANMAT) clinical guidelines for the management of major depressive disorder in adults. III. Pharmacotherapy. J Affect Disord. 2009 Oct;117 Suppl 1:S26-43. doi: 10.1016/j.jad.2009.06.041. PMID: 19674794.

21. Institute for Clinical Systems Improvement. Depression, Adult in Primary Care. Bloomington, MN Institute for Clinical Systems Improvement 2016. http://www.icsi.org/guidelines__more/catalog_guidelines_and_more/catalog_guidelines/catalog_behavioral_health_guidelines/depression. Accessed on 27 May, 2016.

22. National Institute for Health and Clinical Excellence (NICE). Depression. The treatment and management of depression in adults. Manchester, England: National Institute for Health and Clinical Excellence; 2004.

23. Department of Veterans Affairs; Department of Defense. VA/DoD Clinical Practice Guideline for the Management of Major Depressive Disorder. Washington, DC: Department of Veterans Affairs and Department of Defense; 2016. http://www.healthquality.va.gov/guidelines/MH/mdd/MDDFullFinal5192016.pdf. Accessed on 27 May, 2016.

24. Agency for Healthcare Research and Quality. Methods Guide for Effectiveness and Comparative Effectiveness Reviews. AHRQ Publication No. 10(14)-EHC063-EF. Rockville, MD: Agency for Healthcare Research and Quality; January 2014. www.effectivehealthcare.ahrq.gov/methodsguide.cfm.

25. American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders. Fourth ed. Washington, DC: American Psychiatric Association; 1994.

26. American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders (5th edition; DSM-5). Arlington, VA: American Psychiatric Association; 2013.

27. Higgins JP; Altman DG; Gotzsche PC; Juni P; Moher D; Oxman AD; Savovic J; Schulz KF; Weeks L; Sterne JA; Cochrane Bias Methods G; Cochrane Statistical Methods G. The Cochrane Collaboration's tool for assessing risk of bias in randomised trials.

BMJ. 2011;343:d5928. doi: 10.1136/bmj.d5928. PMID: 22008217.

28. Wells GA; Shea B; O'Connell D; Peterson J; Welch V; Losos M; Tugwell P. The Newcastle-Ottawa Scale (NOS) for assessing the quality of nonrandomised studies in meta-analyses [webpage on the Internet]. Ottawa, ON: Ottawa Hospital Research Institute; 2011. http://www.ohri.ca/programs/clinical_epidemiology/oxford.asp. Accessed on 5 February, 2013.

29. United Nations Development Programme. Human Development Report 2015: Work for Human Development. New York, NY: United Nations Development Programme; 2015. http://hdr.undp.org/sites/default/files/2015_human_development_report.pdf. Accessed on May 22, 2017.

30. Fava M; Davidson KG. Definition and epidemiology of treatment-resistant depression. Psychiatr Clin North Am. 1996 Jun;19(2):179-200. PMID: 8827185.

31. Berlim MT; Turecki G. What is the meaning of treatment resistant/refractory major depression (TRD)? A systematic review of current randomized trials. Eur Neuropsychopharmacol. 2007 Nov;17(11):696-707. PMID: 17521891.

32. Thomas SP; Nandhra HS; Jayaraman A. Systematic review of lamotrigine augmentation of treatment resistant unipolar depression (TRD). J Ment Health. 2010 Apr;19(2):168-75. doi: 10.3109/09638230903469269 [doi]. PMID: 20433324.

33. Trivedi RB; Nieuwsma JA; Williams JW. Examination of the utility of psychotherapy for patients with treatment resistant depression: a systematic review (Structured abstract). J Gen Intern Med. 2011(6):643-50. PMID: DARE-12011004035.

34. Silverstein WK; Noda Y; Barr MS; Vila-Rodriguez F; Rajji TK; Fitzgerald PB; Downar J; Mulsant BH; Vigod S; Daskalakis ZJ; Blumberger DM. Neurobiological predictors of response to dorsolateral prefrontal cortex repetitive transcranial magnetic stimulation in depression: a systematic review. Depress Anxiety. 2015;32(12):871-91. PMID: 26382227.

35. Zhang YQ; Zhu D; Zhou XY; Liu YY; Qin B; Ren GP; Xie P. Bilateral repetitive transcranial magnetic stimulation for treatment-resistant depression: A systematic review and meta-analysis of randomized controlled trials. Braz J Med Biol Res. 2015;48(3):198-206.

36. De Carlo V; Calati R; Serretti A. Socio-demographic and clinical predictors of non-response/non-remission in treatment resistant depressed patients: A systematic review. Psychiatry Res. 2016;240:421-30. doi: 10.1016/j.psychres.2016.04.034. PMID: 2016-28410-067.

37. Sackeim HA. The definition and meaning of treatment-resistant depression. J Clin Psychiatry. 2001;62 Suppl 16:10-7. PMID: 11480879.

38. New England Comparative Effectiveness Public Advisory Council. Treatment Options for Treatment-Resistant Depression: Repetitive Transcranial Magnetic Stimulation. Boston, MA: Institute for Clinical and Economic Review; 2012. http://cepac.icer-review.org/wp-content/uploads/2011/04/rTMS-Coverage-Policy-Analysis.pdf. Accessed on 31 May, 2016.

39. Edwards SJ; Hamilton V; Nherera L; Trevor N. Lithium or an atypical antipsychotic drug in the management of treatment-resistant depression: a systematic review and economic evaluation. Health Technol Assess. 2013 Nov;17(54):1-190. doi: 10.3310/hta17540 [doi]. PMID: 24284258.

40. Gaynes BN; Lloyd SW; Lux L; Gartlehner G; Hansen RA; Brode S; Jonas DE; Swinson Evans T; Viswanathan M; Lohr KN. Repetitive transcranial magnetic stimulation for treatment-resistant depression: a systematic review and meta-analysis. J Clin Psychiatry. 2014 May;75(5):477-89. doi: 10.4088/JCP.13r08815. PMID: 24922485.

41. Washington State Health Care Authority. Nonpharmacologic Treatments for Treatment-Resistant Depression. 2014.

42. Malhi GS; Bassett D; Boyce P; Bryant R; Fitzgerald PB; Fritz K; Hopwood M; Lyndon B; Mulder R; Murray G; Porter R; Singh AB. Royal Australian and New Zealand College of Psychiatrists clinical practice guidelines for mood disorders. Aust N Z J Psychiatry. 2015 Dec;49(12):1087-206. doi: 10.1177/0004867415617657. PMID: 26643054.

43. Papadimitropoulou K; Vossen C; Karabis A; Donatti C; Kubitz N. Comparative Efficacy and Tolerability of Pharmacological and Somatic Interventions in Adult Patients with Treatment-Resistant Depression: A Systematic Review and Network Meta-analysis. Curr Med Res Opin. 2016 Dec 30:1-27. doi: 10.1080/03007995.2016.1277201. PMID: 28035869.

44. Ruhe HG; van Rooijen G; Spijker J; Peeters FP; Schene AH. Staging methods for treatment resistant depression. A systematic review. J Affect Disord. 2012 Mar;137(1-3):35-45. doi: 10.1016/j.jad.2011.02.020. PMID: 21435727.

45. Fava M. Diagnosis and definition of treatment-resistant depression. Biol Psychiatry. 2003;53(8):649-59.

46. Fekadu A; Wooderson S; Donaldson C; Markopoulou K; Masterson B; Poon L; Cleare AJ. A multidimensional tool to quantify treatment resistance in depression: the Maudsley staging method. J Clin Psychiatry. 2009 Feb;70(2):177-84. PMID: 19192471.

47. European Medicines Agency. Guideline on clinical investigation of medicinal products in the treatment of depression. London: European Medicines Agency; 2013. http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2013/05/WC500143770.pdf. Accessed on January 8, 2018.

48. Boutron I; Altman DG; Moher D; Schulz KF; Ravaud P; Group CN. CONSORT Statement for Randomized Trials of Nonpharmacologic Treatments: A 2017 Update and a CONSORT Extension for Nonpharmacologic Trial Abstracts. Ann Intern Med. 2017 Jul 04;167(1):40-7. doi: 10.7326/M17-0046. PMID: 28630973.

49. von Elm E; Altman DG; Egger M; Pocock SJ; Gotzsche PC; Vandenbroucke JP; Initiative S. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. J Clin Epidemiol. 2008 Apr;61(4):344-9. doi: 10.1016/j.jclinepi.2007.11.008. PMID: 18313558.

50. Naumann J; Grebe J; Kaifel S; Weinert T; Sadaghiani C; Huber R. Effects of hyperthermic baths on depression, sleep and heart rate variability in patients with depressive disorder: a randomized clinical pilot trial. BMC Complement Altern Med. 2017 Mar 28;17(1):172. doi: 10.1186/s12906-017-1676-5. PMID: 28351399.

51. Aaronson ST; Sears P; Ruvuna F; Bunker M; Conway CR; Dougherty DD; Reimherr FW; Schwartz TL; Zajecka JM. A 5-year observational study of patients with treatment-resistant depression treated with vagus nerve stimulation or treatment as usual: comparison of response, remission, and suicidality. Am J Psychiatry. 2017 Jul 01;174(7):640-8. doi: 10.1176/appi.ajp.2017.16010034. PMID: 28359201.

52. NeuroStar Advanced Therapy. About NeuroStar® TMS Therapy. Malvern, PA: NeuroStar Advanced Therapy; 2017. https://neurostar.com/neurostar-tms-depression-treatment/. Accessed on August 2, 2017.

53. Singh J; Singh A; Kar SK; Pahuja E. Early augmentation response with low-frequency repetitive transcranial magnetic stimulation in treatment resistant depression. Clinical Psychopharmacology and Neuroscience. 2017;15(2):197-8. doi: 10.9758/cpn.2017.15.2.197. PMID: 2017-25469-017.

54. Kessler RC; Chiu WT; Demler O; Merikangas KR; Walters EE. Prevalence, severity, and comorbidity of 12-month DSM-IV disorders in the National Comorbidity Survey Replication. Arch Gen Psychiatry. 2005 Jun;62(6):617-27. doi: 10.1001/archpsyc.62.6.617. PMID: 15939839.

55. Russell JM; Hawkins K; Ozminkowski RJ; Orsini L; Crown WH; Kennedy S; Finkelstein S; Berndt E; Rush AJ. The cost consequences of treatment-resistant depression. J Clin Psychiatry. 2004 Mar;65(3):341-7. PMID: 15096073.

56. National Institute for Health and Clinical Excellence (NICE). Depression. The treatment and management of depression in adults. Manchester, England: National Institute for Health and Clinical Excellence; 2004.

57. Centers for Medicare & Medicaid Services. MEDCAC Meeting 4/27/2016 - Treatment Resistant Depression. Baltimore, MD: Centers for Medicare & Medicaid Services; 2016. https://www.cms.gov/medicare-coverage-database/details/medcac-meeting-details.aspx?MEDCACId=71&bc=AAAIAAAAAAAAAAAA%3d%3d&. Accessed 21 November 2016.

58. Gaynes BN; Lux L; Lloyd S; Hansen RA; Gartlehner G; Thieda P; Jonas D; Brode S; Swinson Evans T; Crotty K; Viswanathan M; KN L. Nonpharmacologic Interventions for Treatment-Resistant Depression in Adults Comparative Effectiveness Review No. 33. (Prepared by RTI International-University of North Carolina under Contract No. 290-02-0016I, TO #2.) AHRQ Publication No. 11-EHC056-EF. Rockville, MD: Agency for Healthcare Research and Quality; Sep 2011. http://www.effectivehealthcare.ahrq.gov/reports/final.cfm. PMID: 22091472.

59. Gartlehner G; Gaynes B; Amick H; Asher G; Morgan LC; Coker-Schwimmer E; Forneris C; Boland E; Lux L; Gaylord S; Bann C; Pierl CB; Lohr K. Nonpharmacological Versus Pharmacological Treatments for Adult Patients with Major Depressive Disorder Comparative Effectiveness Review No. 161. (Prepared by the RTI-UNC Evidence-based Practice Center under Contract No. 290-2012-00008I.) AHRQ Publication No. 15(16)-EHC031-EF. Rockville, MD: Agency for Healthcare Research and Quality; December 2015. www.effectivehealthcare.ahrq.gov/reports/final.cfm. PMID: 26764438.

60. Guyatt GH; Oxman AD; Kunz R; Atkins D; Brozek J; Vist G; Alderson P; Glasziou P; Falck-Ytter Y; Schunemann HJ. GRADE guidelines: 2. Framing the question and deciding on important outcomes. J Clin Epidemiol. 2011 Apr;64(4):395-400. doi: 10.1016/j.jclinepi.2010.09.012. PMID: 21194891.

61. U.S. Federal Drug Administration. CFR - Code of Federal Regulations Title 21 Section 312.32 IND Safety Reporting. Silver Spring, MD: U.S. Federal Drug Administration; 2016. https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfCFR/CFRSearch.cfm?fr=312.32. Accessed 21 November 2016.

62. U.S. Federal Drug Administration. Summary of Safety and Effectiveness Data. Silver Spring, MD: U.S. Food and Drug Administration; 2005. www.accessdata.fda.gov/cdrh_docs/pdf/P970003S050b.pdf. Accessed on March 20, 2017.

63. Fu R; Gartlehner G; Grant M; Shamliyan T; Sedrakyan A; Wilt TJ; Griffith L; Oremus M; Raina P; Ismaila A; Santaguida P; Lau J; Trikalinos TA. Conducting quantitative synthesis when comparing medical interventions: AHRQ and the Effective Health Care Program. Methods Guide for Effectiveness and Comparative Effectiveness Reviews [Internet]. Rockville, MD: Agency for Healthcare Research and Quality; 2010.

64. Repetitive transcranial magnetic stimulation for treatment-resistant depression: A systematic review and meta-analysis of randomized controlled trials. Ontario Health Technology Assessment Series. 2016;16(5):1-66.

65. Papadimitropoulou K; Vossen C; Karabis A; Donatti C; Kubitz N. Comparative efficacy

and tolerability of pharmacological and somatic interventions in adult patients with treatment-resistant depression: a systematic review and network meta-analysis. Curr Med Res Opin. 2017;33(4):701-11. doi: 10.1080/03007995.2016.1277201.

66. Luan S; Wan H; Wang S; Li H; Zhang B. Efficacy and safety of olanzapine/fluoxetine combination in the treatment of treatmentresistant depression: A meta-analysis of randomized controlled trials. Neuropsychiatr Dis Treat. 2017;13((Luan S.; Wan H.; Wang S.) Department of Mental Health, China-Japan Union Hospital of Jilin University, Changchun, China):609-20. doi: 10.2147/ndt.s127453.

67. Bauer M; Whybrow PC; Angst J; Versiani M; Moller HJ. World Federation of Societies of Biological Psychiatry (WFSBP) Guidelines for Biological Treatment of Unipolar Depressive Disorders, Part 1: Acute and continuation treatment of major depressive disorder. World J Biol Psychiatry. 2002 Jan;3(1):5-43. PMID: 12479086.

68. Rossi S; Hallett M; Rossini PM; Pascual-Leone A. Safety, ethical considerations, and application guidelines for the use of transcranial magnetic stimulation in clinical practice and research. Clin Neurophysiol. 2009 Dec;120(12):2008-39. doi: S1388-2457(09)00519-7 [pii]; 10.1016/j.clinph.2009.08.016 [doi]. PMID: 19833552.

69. Schlaepfer TE; Ågren H; Monteleone P; Gasto C; Pitchot W; Rouillon F; Nutt DJ; Kasper S. The hidden third: Improving outcome in treatment-resistant depression. Journal of Psychopharmacology. 2012;26(5):587-602.

70. Harter M; Klesse C; Bermejo I; Schneider F; Berger M. Unipolar depression: diagnostic and therapeutic recommendations from the current S3/National Clinical Practice Guideline. Dtsch Arztebl Int. 2010 Oct;107(40):700-8. doi: 10.3238/arztebl.2010.0700. PMID: 21031129.

71. National Institute for Health and Care Excellence. Repetititve transcranial magnetic stimulation for depression [Clinical Gudelines]. London: National Institute for Health and Care Excellence; 2015. http://nice.org.uk/guidance/ipg542. Accessed on May 22, 2017.

72. National Institute for Health and Clinical Excellence (NICE). Depression. The treatment and management of depression in adults. 2010(Revised October 2009).

73. National Institute for Health and Care Excellence. Vagus nerve stimulation for treatment-resistant depression. London: National Institute for Health and Care Excellence; 2009. http://nice.org.uk/guidance/ipg330. Accessed on May 22, 2017.

74. Kennedy SH; Lam RW; McIntyre RS; Tourjman SV; Bhat V; Blier P; Hasnain M; Jollant F; Levitt AJ; MacQueen GM; McInerney SJ; McIntosh D; Milev RV; Muller DJ; Parikh SV; Pearson NL; Ravindran AV; Uher R; Group CDW. Canadian Network for Mood and Anxiety Treatments (CANMAT) 2016 Clinical Guidelines for the Management of Adults with Major Depressive Disorder: Section 3. Pharmacological Treatments. Can J Psychiatry. 2016 Sep;61(9):540-60. doi: 10.1177/0706743716659417. PMID: 27486148.

75. Parikh SV; Quilty LC; Ravitz P; Rosenbluth M; Pavlova B; Grigoriadis S; Velyvis V; Kennedy SH; Lam RW; MacQueen GM; Milev RV; Ravindran AV; Uher R; Canmat Depression Work Group. Canadian Network for Mood and Anxiety Treatments (CANMAT) 2016 Clinical Guidelines for the Management of Adults with Major Depressive Disorder: Section 2. Psychological Treatments. Can J Psychiatry. 2016 Sep;61(9):524-39. doi: 10.1177/0706743716659418. PMID: 27486150.

76. Milev RV; Giacobbe P; Kennedy SH; Blumberger DM; Daskalakis ZJ; Downar J; Modirrousta M; Patry S; Vila-Rodriguez F; Lam RW; MacQueen GM; Parikh SV; Ravindran AV; Group CDW. Canadian Network for Mood and Anxiety Treatments (CANMAT) 2016 Clinical Guidelines for the Management of Adults with Major

Depressive Disorder: Section 4. Neurostimulation Treatments. Can J Psychiatry. 2016 Sep;61(9):561-75. doi: 10.1177/0706743716660033. PMID: 27486154.

77. Ravindran AV; Balneaves LG; Faulkner G; Ortiz A; McIntosh D; Morehouse RL; Ravindran L; Yatham LN; Kennedy SH; Lam RW; MacQueen GM; Milev RV; Parikh SV; Canmat Depression Work Group. Canadian Network for Mood and Anxiety Treatments (CANMAT) 2016 Clinical Guidelines for the Management of Adults with Major Depressive Disorder: Section 5. Complementary and Alternative Medicine Treatments. Can J Psychiatry. 2016 Sep;61(9):576-87. doi: 10.1177/0706743716660290. PMID: 27486153.

78. MacQueen GM; Frey BN; Ismail Z; Jaworska N; Steiner M; Lieshout RJ; Kennedy SH; Lam RW; Milev RV; Parikh SV; Ravindran AV; Group CDW. Canadian Network for Mood and Anxiety Treatments (CANMAT) 2016 Clinical Guidelines for the Management of Adults with Major Depressive Disorder: Section 6. Special Populations: Youth, Women, and the Elderly. Can J Psychiatry. 2016 Sep;61(9):588-603. doi: 10.1177/0706743716659276. PMID: 27486149.

79. Cleare A; Pariante CM; Young AH; Anderson IM; Christmas D; Cowen PJ; Dickens C; Ferrier IN; Geddes J; Gilbody S; Haddad PM; Katona C; Lewis G; Malizia A; McAllister-Williams RH; Ramchandani P; Scott J; Taylor D; Uher R; Members of the Consensus M. Evidence-based guidelines for treating depressive disorders with antidepressants: A revision of the 2008 British Association for Psychopharmacology guidelines. J Psychopharmacol. 2015 May;29(5):459-525. doi: 10.1177/0269881115581093. PMID: 25969470.

80. Oquendo MA; Baca-Garcia E; Kartachov A; Khait V; Campbell CE; Richards M; Sackeim HA; Prudic J; Mann JJ. A computer algorithm for calculating the adequacy of antidepressant treatment in unipolar and bipolar depression. J Clin Psychiatry. 2003 Jul;64(7):825-33. PMID: 12934985.

81. Souery D; Amsterdam J; de Montigny C; Lecrubier Y; Montgomery S; Lipp O; Racagni G; Zohar J; Mendlewicz J. Treatment resistant depression: methodological overview and operational criteria. Eur Neuropsychopharmacol. 1999 Jan;9(1-2):83-91. PMID: 10082232.

82. Petersen T; Papakostas GI; Posternak MA; Kant A; Guyker WM; Iosifescu DV; Yeung AS; Nierenberg AA; Fava M. Empirical testing of two models for staging antidepressant treatment resistance. J Clin Psychopharmacol. 2005 Aug;25(4):336-41. doi: 00004714-200508000-00008 [pii]. PMID: 16012276.

83. McClintock SM; Cullum M; Husain MM; Rush AJ; Knapp RG; Mueller M; Petrides G; Sampson S; Kellner CH. Evaluation of the Effects od Severe Depression on Global Cognitive Function and Memory. CNS Spectr. 2010 May;15(5):304-13. PMID: 20448521.

84. Smarr KL; Keefer AL. Measures of depression and depressive symptoms: Beck Depression Inventory-II (BDI-II), Center for Epidemiologic Studies Depression Scale (CES-D), Geriatric Depression Scale (GDS), Hospital Anxiety and Depression Scale (HADS), and Patient Health Questionnaire-9 (PHQ-9). Arthritis Care Res. 2011 Nov;63(S11):S454-66. doi: 10.1002/acr.20556. PMID: 22588766.

85. Wang Y; Gorenstein C. Psychometric properties of the Beck Depression Inventory-II: a comprehensive review. Revista Brasileira de Psiquiatria. 2013;35:416-31.

86. Hiroe T; Kojima M; Yamamoto I; Nojima S; Kinoshita Y; Hashimoto N; Watanabe N; Maeda T; Furukawa TA. Gradations of clinical severity and sensitivity to change assessed with the Beck Depression Inventory-II in Japanese patients with depression. Psychiatry Res. 2005 Jun 30;135(3):229-35. doi: 10.1016/j.psychres.2004.03.014. PMID: 15996749.

87. Beck AT; Steer RA; Garbin MG. Psychometric properties of the Beck Depression Inventory: twenty-five years of evaluation. Clin Psychol Rev. 1988;8(1):77-100.

88. Dworkin RH; Turk DC; Wyrwich KW; Beaton D; Cleeland CS; Farrar JT; Haythornthwaite JA; Jensen MP; Kerns RD; Ader DN; Brandenburg N; Burke LB; Cella D; Chandler J; Cowan P; Dimitrova R; Dionne R; Hertz S; Jadad AR; Katz NP; Kehlet H; Kramer LD; Manning DC; McCormick C; McDermott MP; McQuay HJ; Patel S; Porter L; Quessy S; Rappaport BA; Rauschkolb C; Revicki DA; Rothman M; Schmader KE; Stacey BR; Stauffer JW; von Stein T; White RE; Witter J; Zavisic S. Interpreting the clinical importance of treatment outcomes in chronic pain clinical trials: IMMPACT recommendations. J Pain. 2008 Feb;9(2):105-21. doi: 10.1016/j.jpain.2007.09.005. PMID: 18055266.

89. National Institute for Health and Care Excellence. Depression in adults. London: National Institute for Health and Care Excellence; 2011. https://www.nice.org.uk/guidance/qs8/chapter/quality-statement-1-assessment. Accessed on July 11, 2017.

90. Svanborg P; Asberg M. A comparison between the Beck Depression Inventory (BDI) and the self-rating version of the Montgomery Asberg Depression Rating Scale (MADRS). J Affect Disord. 2001 May;64(2-3):203-16. doi: Doi 10.1016/S0165-0327(00)00242-1. PMID: WOS:000168604400010.

91. Lewinsohn PM; Seeley JR; Roberts RE; Allen NB. Center for Epidemiological Studies-Depression Scale (CES-D) as a screening instrument for depression among community-residing older adults. Psychol Aging. 1997;12:277- 87.

92. Haringsma R; Engels GI; Beekman AT; Spinhoven P. The criterion validity of the Center for Epidemiological Studies Depression Scale (CES-D) in a sample of self-referred elders with depressive symptomatology. Int J Geriatr Psychiatry. 2004 Jun;19(6):558-63. doi: 10.1002/gps.1130. PMID: 15211536.

93. Orme JG; Reis J; Herz EJ. Factorial and discriminant validity of the Center for Epidemiological Studies Depression (CES-D) scale. J Clin Psychol. 1986 Jan;42(1):28-33. PMID: 3950011.

94. Masson SC; Tejani AM. Minimum clinically important differences identified for commonly used depression rating scales. J Clin Epidemiol. 2013 Jul;66(7):805-7. doi: 10.1016/j.jclinepi.2013.01.010. PMID: WOS:000320426800018.

95. Brink TL; Yesavage JA; Lum O; Heersema P; Adey MB; Rose TL. Screening tests for geriatric depression. Clin Gerontol. 1982;1:37-44.

96. Marcus RN; McQuade RD; Carson WH; Hennicken D; Fava M; Simon JS; Trivedi MH; Thase ME; Berman RM. The efficacy and safety of aripiprazole as adjunctive therapy in major depressive disorder: a second multicenter, randomized, double-blind, placebo-controlled study. J Clin Psychopharmacol. 2008 Apr;28(2):156-65. doi: 10.1097/JCP.0b013e31816774f9. PMID: 18344725.

97. Cullum S; Tucker S; Todd C; Brayne C. Screening for depression in older medical inpatients. Int J Geriatr Psychiatry. 2006 May;21(5):469-76. doi: 10.1002/gps.1497. PMID: 16676293.

98. Bijl D; van Marwijk HW; Ader HJ; Beekman AT; de Haan M. Test-characteristics of the GDS-15 in screening for major depression in elderly patients in clinical practice. Clin Gerontol. 2005;29:1-9. doi: 10.1300/J018v29n01_01

99. Weintraub D; Oehlberg KA; Katz IR; Stern MB. Test characteristics of the 15-item geriatric depression scale and Hamilton depression rating scale in Parkinson disease. Am J Geriatr Psychiatry. 2006 Feb;14(2):169-75. doi: 10.1097/01.JGP.0000192488.66049.4b. PMID: 16473982.

100. Friedman B; Heisel MJ; Delavan RL. Psychometric properties of the 15-item

geriatric depression scale in functionally impaired, cognitively intact, community-dwelling elderly primary care patients. J Am Geriatr Soc. 2005 Sep;53(9):1570-6. doi: 10.1111/j.1532-5415.2005.53461.x. PMID: 16137289.

101. Yesavage JA; Brink TL; Rose TL; Lum O; Huang V; Adey M; Leirer VO. Development and validation of a geriatric depression screening scale - a preliminary-report. J Psychiatr Res. 1983;17(1):37-49. doi: Doi 10.1016/0022-3956(82)90033-4. PMID: WOS:A1983QQ74400004.

102. van Marwijk HW; Wallace P; de Bock GH; Hermans J; Kaptein AA; Mulder JD. Evaluation of the feasibility, reliability and diagnostic value of shortened versions of the geriatric depression scale. Br J Gen Pract. 1995 Apr;45(393):195-9. PMID: 7612321.

103. Lopez MN; Quan NM; Carvajal PM. A psychometric study of the Geriatric Depression Scale. Eur J Psychol Assess. 2010;26(1):55-60. doi: 10.1027/1015-5759/a000008. PMID: WOS:000274049400008.

104. Mitchell AJ; Bird V; Rizzo M; Meader N. Diagnostic validity and added value of the geriatric depression scale for depression in primary care: a meta-analysis of GDS(30) and GDS(15). J Affect Disord. 2010 Sep;125(1-3):10-7. doi: 10.1016/j.jad.2009.08.019. PMID: WOS:000281377100002.

105. Cusin C; Yang H; Yeung A; Fava M. Chapter 2. Rating scales for depression. In: Baer L, Blais MA, eds. Handbook of Clinical Rating Scales and Assessment in Psychiatry and Mental Health. Current Clinical Psychology. New York: Humana Press; 2010.

106. American Thoracic Society (ATS). Hamilton Rating Scale for Depression (HRSD). New York, NY: American Thoracic Society; 2013. http://www.thoracic.org/assemblies/srn/questionaires/hrsd.php. Accessed on July 3, 2017.

107. Price RB; Shungu DC; Mao X; Nestadt P; Kelly C; Collins KA; Murrough JW; Charney DS; Mathew SJ. Amino acid neurotransmitters assessed by proton magnetic resonance spectroscopy: relationship to treatment resistance in major depressive disorder. Biol Psychiatry. 2009 May 1;65(9):792-800. doi: 10.1016/j.biopsych.2008.10.025. PMID: 19058788.

108. Bagby RM; Ryder AG; Schuller DR; Marshall MB. The Hamilton Depression Rating Scale: has the gold standard become a lead weight? Am J Psychiatry. 2004 Dec;161(12):2163-77. doi: 10.1176/appi.ajp.161.12.2163. PMID: 15569884.

109. Lopez-Pina JA; Sanchez-Meca J; Rosa-Alcazar AI. The Hamilton Rating Scale for Depression: a meta-analytic reliability generalization study. Int J Clin Health Psychol. 2009 Jan;9(1):143-59. PMID: WOS:000262363900010.

110. Rush AJ; Trivedi MH; Ibrahim HM; Carmody TJ; Arnow B; Klein DN; Markowitz JC; Ninan PT; Kornstein S; Manber R; Thase ME; Kocsis JH; Keller MB. 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. Biol Psychiatry. 2003 Sep 01;54(5):573-83. PMID: 12946886.

111. Williams JBW. Standardizing the Hamilton Depression Rating Scale: past, present, and future. Eur Arch Psychiatry Clin Neurosci. 2001;251(S2):6-12. doi: 10.1007/bf03035120.

112. Duru G; Fantino B. The clinical relevance of changes in the Montgomery-Asberg Depression Rating Scale using the minimum clinically important difference approach. Curr Med Res Opin. 2008 May;24(5):1329-35. doi: 10.1185/030079908X291958. PMID: 18377706.

113. Carmody TJ; Rush AJ; Bernstein I; Warden D; Brannan S; Burnham D; Woo A; Trivedi MH. The Montgomery Asberg and the Hamilton ratings of depression: a comparison of measures. Eur Neuropsychopharmacol. 2006

Dec;16(8):601-11. doi: 10.1016/j.euroneuro.2006.04.008. PMID: 16769204.

114. Substance Abuse and Mental Health Services Administration. Patient Health Quesionnaire (PHQ-9) Rockville, MD: SAMHSA; 2005. https://www.integration.samhsa.gov/images/res/PHQ%20-%20Questions.pdf. Accessed on May 21, 2017.

115. Maurer DM. Screening for depression. Am Fam Physician. 2012 Jan 15;85(2):139-44. PMID: 22335214.

116. Spitzer RL; Kroenke K; Williams JB. Validation and utility of a self-report version of PRIME-MD: the PHQ primary care study. Primary Care Evaluation of Mental Disorders. Patient Health Questionnaire. JAMA. 1999 Nov 10;282(18):1737-44. PMID: 10568646.

117. Kroenke K; Spitzer RL; Williams JB. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med. 2001 Sep;16(9):606-13. PMID: 11556941.

118. Lowe B; Unutzer J; Callahan CM; Perkins AJ; Kroenke K. Monitoring depression treatment outcomes with the patient health questionnaire-9. Med Care. 2004 Dec;42(12):1194-201. PMID: 15550799.

119. Lamoureux BE; Linardatos E; Fresco DM; Bartko D; Logue E; Milo L. Using the QIDS-SR16 to identify major depressive disorder in primary care medical patients. Behav Ther. 2010 Sep;41(3):423-31. PMID: WOS:000278892500015.

120. Reilly TJ; MacGillivray SA; Reid IC; Cameron IM. Psychometric properties of the 16-item Quick Inventory of Depressive Symptomatology: a systematic review and meta-analysis. J Psychiatr Res. 2015 Jan;60:132-40. doi: 10.1016/j.jpsychires.2014.09.008. PMID: 25300442.

121. Rush AJ; Pincus HA; First MB; Blacker D; Endicott J; Keith SJ; Phillips KA; Ryan ND; Smither J, G. R.; Tsuang MT; Widiger TA; Zarin DA. Handbook of Psychiatric Measures. Washington, DC: American Psychiatric Association; 2000.

122. Leucht S; Kane JM; Kissling W; Hamann J; Etschel E; Engel R. Clinical implications of Brief Psychiatric Rating Scale scores. Br J Psychiatry. 2005 Oct;187:366-71. doi: 10.1192/bjp.187.4.366. PMID: 16199797.

123. Guy W. Clinical Global Impressions (CGI) scale. In: Rush AJ, Pincus HA, First MB, Blacker D, Endicott J, Keith SJ, et al., eds. Handbook of psychiatric measures. 1st ed. Washington, DC: American Psychiatric Association; 2000:100-2.

124. Luciano JV; Ayuso-Mateos JL; Fernandez A; Serrano-Blanco A; Roca M; Haro JM. Psychometric properties of the twelve item World Health Organization Disability Assessment Schedule II (WHO-DAS II) in Spanish primary care patients with a first major depressive episode. J Affect Disord. 2010 Feb;121(1-2):52-8. doi: 10.1016/j.jad.2009.05.008. PMID: 19464735.

125. Hermes ED; Sokoloff D; Stroup TS; Rosenheck RA. Minimum clinically important difference in the Positive and Negative Syndrome Scale with data from the Clinical Antipsychotic Trials of Intervention Effectiveness (CATIE). J Clin Psychiatry. 2012 Apr;73(4):526-32. doi: 10.4088/JCP.11m07162. PMID: 22579152.

126. Moncrieff J; Kirsch I. Empirically derived criteria cast doubt on the clinical significance of antidepressant-placebo differences. Contemp Clin Trials. 2015 Jul;43:60-2. doi: 10.1016/j.cct.2015.05.005. PMID: 25979317.

127. Button KS; Kounali D; Thomas L; Wiles NJ; Peters TJ; Welton NJ; Ades AE; Lewis G. Minimal clinically important difference on the Beck Depression Inventory--II according to the patient's perspective. Psychol Med. 2015 Nov;45(15):3269-79. doi: 10.1017/S0033291715001270. PMID: 26165748.

128. Sheehan KH; Sheehan DV. Assessing treatment effects in clinical trials with the discan metric of the Sheehan Disability Scale. Int Clin Psychopharmacol. 2008

Mar;23(2):70-83. doi: 10.1097/YIC.0b013e3282f2b4d6. PMID: 18301121.

129. Endicott J; Paulsson B; Gustafsson U; Schioler H; Hassan M. Quetiapine monotherapy in the treatment of depressive episodes of bipolar I and II disorder: Improvements in quality of life and quality of sleep. J Affect Disord. 2008 Dec;111(2-3):306-19. doi: 10.1016/j.jad.2008.06.019. PMID: 18774180.

130. Holtzheimer PE; Kelley ME; Gross RE; Filkowski MM; Garlow SJ; Barrocas A; Wint D; Craighead MC; Kozarsky J; Chismar R; Moreines JL; Mewes K; Posse PR; Gutman DA; Mayberg HS. Subcallosal cingulate deep brain stimulation for treatment-resistant unipolar and bipolar depression. Arch Gen Psychiatry. 2012 Feb;69(2):150-8. doi: 10.1001/archgenpsychiatry.2011.1456. PMID: 22213770.

131. Aas IH. Global Assessment of Functioning (GAF): properties and frontier of current knowledge. Ann Gen Psychiatry. 2010 May 07;9:20. doi: 10.1186/1744-859X-9-20. PMID: 20459646.

132. Jones SH; Thornicroft G; Coffey M; Dunn G. A brief mental health outcome scale-reliability and validity of the Global Assessment of Functioning (GAF). Br J Psychiatry. 1995 May;166(5):654-9. PMID: 7620753.

133. Lam RW; Michalak EE; Swinson RP. Assessment scales in depression, mania and anxiety. London: Taylor and Francis; 2005.

134. Endicott J; Nee J; Harrison W; Blumenthal R. Quality of Life Enjoyment and Satisfaction Questionnaire: a new measure. Psychopharmacol Bull. 1993;29(2):321-6. PMID: 8290681.

135. Stevanovic D. Quality of Life Enjoyment and Satisfaction Questionnaire-short form for quality of life assessments in clinical practice: a psychometric study. J Psychiatr Ment Health Nurs. 2011 Oct;18(8):744-50. doi: 10.1111/j.1365-2850.2011.01735.x. PMID: 21896118.

136. Demyttenaere K; Andersen HF; Reines EH. Impact of escitalopram treatment on Quality of Life Enjoyment and Satisfaction Quesionnaire scores in major depressive disorder and generalized anxiety disorder. Int Clin Psychopharmacol. 2008;23:276-86.

137. Sheehan DV; Harnett-Sheehan K; Raj BA. The measurement of disability. Int Clin Psychopharmacol. 1996 Jun;11(Suppl 3):89-95. PMID: 8923116.

138. Optum, Inc. SF-36 Health Survey. Eden Prairie, MN: Optum Inc; n.d. https://campaign.optum.com/content/optum/en/optum-outcomes/what-we-do/health-surveys/sf-36v2-health-survey.html. Accessed on December 4, 2016.

139. Savovic J; Jones HE; Altman DG; Harris RJ; Juni P; Pildal J; Als-Nielsen B; Balk EM; Gluud C; Gluud LL; Ioannidis JP; Schulz KF; Beynon R; Welton NJ; Wood L; Moher D; Deeks JJ; Sterne JA. Influence of reported study design characteristics on intervention effect estimates from randomized, controlled trials. Ann Intern Med. 2012 Sep 18;157(6):429-38. doi: 10.7326/0003-4819-157-6-201209180-00537. PMID: 22945832.

140. Lundh A; Sismondo S; Lexchin J; Busuioc OA; Bero L. Industry sponsorship and research outcome. Cochrane Database Syst Rev. 2012 Dec 12;12:MR000033. doi: 10.1002/14651858.MR000033.pub2. PMID: 23235689.

141. Quiroz JA; Singh J; Gould TD; Denicoff KD; Zarate CA; Manji HK. Emerging experimental therapeutics for bipolar disorder: clues from the molecular pathophysiology. Mol Psychiatry. 2004 Aug;9(8):756-76. doi: 10.1038/sj.mp.4001521. PMID: 15136795.