Evidence Synthesis Program

Neuroimaging and Neurophysiologic Biomarkers for Mental Health: An Evidence Map

October 2022

U.S. Department of Veterans Affairs
Veterans Health Administration
Health Services Research & Development Service

Recommended citation: Ullman K, Landsteiner A, Anthony M, et al. Neuroimaging and Neurophysiologic Biomarkers for Mental Health: An Evidence Map. Washington, DC: Evidence Synthesis Program, Health Services Research and Development Service, Office of Research and Development, Department of Veterans Affairs. VA ESP Project #09-009; 2022.

AUTHORS

Author roles, affiliations, and contributions to the present report (using the CRediT taxonomy) are summarized in the table below.

Author	Role and Affiliation	Report Contribution
Adrienne Landsteiner, PhD	Senior Scientist, Minneapolis ESP Center Minneapolis, MN	Conceptualization, Formal analysis, Investigation, Methodology, Visualization, Writing – original draft, review & editing
Kristen Ullman, MPH	Program Manager, Minneapolis ESP Center Minneapolis, MN	Conceptualization, Formal analysis, Investigation, Methodology, Project administration, Software, Visualization, Writing – original draft, review & editing
Catherine Sowerby, BA	Research Associate, Minneapolis ESP Center Minneapolis, MN	Formal analysis, Investigation, Visualization, Writing – original draft, review & editing
Caleb Kalinowski, MS	Research Associate, Minneapolis ESP Center Minneapolis, MN	Formal analysis, Investigation, Visualization, Writing – original draft, review & editing
Maylen Anthony, MPH	Research Associate, Minneapolis ESP Center Minneapolis, MN	Formal analysis, Investigation, Visualization, Writing – original draft, review & editing
Wei Duan-Porter, MD, PhD	Co-Director, Minneapolis ESP Center Minneapolis, MN	Conceptualization, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Visualization, Supervision, Writing – original draft, review & editing
Scott Sponheim, PhD	Staff Psychologist, Minneapolis VHA Minneapolis, MN Professor, Department of Psychiatry and Behavioral Sciences, University of Minnesota, Minneapolis, MN	Conceptualization, Methodology, Writing – review & editing
Michele Spoont, PhD	Clinical Research Psychologist and Core Investigator, Center for Care Delivery and Outcomes Research, Minneapolis VHA Minneapolis, MN National Center for PTSD Associate Professor, Departments of Medicine and Psychiatry, University of Minnesota Medical School Minneapolis, MN	Conceptualization, Methodology, Writing – review & editing
Kelvin Lim, MD	Director for Adult Mental Health Research, Dept of Psychiatry and	Conceptualization, Methodology, Writing – review & editing

Author	Role and Affiliation	Report Contribution
	Behavioral Science, University of Minneapolis Minneapolis, MN	
Jose Pardo, MD, PhD	Professor, Department of Psychiatry, University of Minneapolis Minneapolis, MN Director, Cognitive Neuroimaging Unit, Minneapolis VHA Minneapolis, MN	Conceptualization, Methodology, Writing – review & editing
Timothy J. Wilt, MD, MPH	Director, Minneapolis ESP Center Minneapolis, MN	Conceptualization, Funding acquisition, Methodology, Supervision, Writing – review & editing

This report was prepared by the Evidence Synthesis Program Center located at the **Minneapolis VA Health Care System,** directed by Timothy J. Wilt, MD, MPH and Wei Duan-Porter, MD, PhD and funded by the Department of Veterans Affairs, Veterans Health Administration, Health Services Research and Development.

PREFACE

The VA Evidence Synthesis Program (ESP) was established in 2007 to provide timely and accurate syntheses of targeted health care topics of importance to clinicians, managers, and policymakers as they work to improve the health and health care of Veterans. These reports help:

- Develop clinical policies informed by evidence;
- Implement effective services to improve patient outcomes and to support VA clinical practice guidelines and performance measures; and
- Set the direction for future research to address gaps in clinical knowledge.

The program comprises 4 ESP Centers across the US and a Coordinating Center located in Portland, Oregon. Center Directors are VA clinicians and recognized leaders in the field of evidence synthesis with close ties to the AHRQ Evidence-based Practice Center Program. The Coordinating Center was created to manage program operations, ensure methodological consistency and quality of products, interface with stakeholders, and address urgent evidence needs. To ensure responsiveness to the needs of decision-makers, the program is governed by a Steering Committee composed of health system leadership and researchers. The program solicits nominations for review topics several times a year via the program website.

The present report was developed in response to a request from the Office of Research and Development working group for the Commander John Scott Hannon Veterans Mental Health Care Improvement Act, Public Law 116-171, section 305 (SHA305). The scope was further developed with input from Operational Partners (below), the ESP Coordinating Center, and the review team. The ESP consulted several technical and content experts in designing the research questions and review methodology. In seeking broad expertise and perspectives, divergent and conflicting opinions are common and perceived as healthy scientific discourse that results in a thoughtful, relevant systematic review. Ultimately, however, research questions, design, methodologic approaches, and/or conclusions of the review may not necessarily represent the views of individual technical and content experts.

ACKNOWLEDGMENTS

The authors are grateful to the following individuals for their contributions to this project:

Operational Partners

Operational partners are system-level stakeholders who help ensure relevance of the review topic to the VA, contribute to the development of and approve final project scope and timeframe for completion, provide feedback on the draft report, and provide consultation on strategies for dissemination of the report to the field and relevant groups.

Stuart W. Hoffman, PhD
Senior Health Science Officer
Office of Research and Development (ORD)

Sumitra Muralidhar, PhD
Director, Million Veteran Program
Office of Research and Development (ORD)

Vetisha L. McClair, PhD
Health Science Officer
Clinical Science Research and Development (CSR&D)

Clifford Smith, PhD, ABPP
Director of Analytics, Innovations, and Collaborations
Office of Mental Health and Suicide Prevention (OMHSP)

Emily Hartwell, PhD
Clinical Psychologist
Office of Research and Development (ORD)
Office of Mental Health and Suicide Prevention (OMHSP)

Wendy Tenhula, PhD
Deputy Chief Research and Development Officer
Office of Research and Development (ORD)

Peer Reviewers

The Coordinating Center sought input from external peer reviewers to review the draft report and provide feedback on the objectives, scope, methods used, perception of bias, and omitted evidence (see Appendix C for disposition of comments). Peer reviewers must disclose any relevant financial or non-financial conflicts of interest. Because of their unique clinical or content expertise, individuals with potential conflicts may be retained. The Coordinating Center works to balance, manage, or mitigate any potential nonfinancial conflicts of interest identified.

ABBREVIATIONS TABLE

AHRQ EPC	Agency for Healthcare Research and Quality Evidence-based Practice Center
ASL	Arterial spin labeling
BDI	Beck Depression Inventory
CAPS	Clinician Administered PTSD Scale
DAISY	DistillerSR's Artificial Intelligence System
DTI	Diffusion tensor imaging
ECT	Electroconvulsive therapy
EEG	Evoked potentials and electroencephalogram
ESP	Evidence Synthesis Program
fMRI	Functional magnetic resonance imaging
GOSE	Glasgow Outcome Scale-Extended
HAM-A	Hamilton Anxiety Rating Scale
HAM-D	Hamilton Depression Rating Scale
KQ	Key Question
MADRS	Montgomery-Asberg Depression Rating Scale
MEG	Magnetoencephalography
MeSH	Medical Subject Headings
MINI	Mini-International Neuropsychiatric Interview
MRI	Magnetic resonance imaging
OCD	Obsessive compulsive disorder
PCL	PTSD Checklist
PET	Positron emission tomography
PHQ	Patient Health Questionnaire
PTSD	Posttraumatic stress disorder
RCT	Randomized controlled trial
rTMS	Repetitive transcranial magnetic stimulation
SCID	Structured Clinical Interview for DSM
SHA305	Commander John Scott Hannon Veterans Mental Health Care Improvement Act, Public Law 116-171, section 305
SPECT	Single photon emission computed tomography
SUD	Substance use disorder
TBI	Traumatic brain injury
TBS	Theta burst stimulation
tDCS	Transcranial direct current stimulation
US	United States
VA	Department of Veterans Affairs
Y-BOCS	Yale Brown Obsessive Compulsive Scale
YMRS	Young Mania Rating Scale

EXECUTIVE SUMMARY

Key Findings

- Many studies evaluated the use of structural or functional MRI in diagnosis and prognosis of depression, but there were important methodological concerns:
  - Nearly all diagnostic studies were cross-sectional, small in size, and included participants with variable past histories of symptoms and treatments.
  - Prognostic studies mostly focused on response to antidepressants, and were also generally small.
- A substantial number of studies used EEG for diagnosis and prognosis of depression, but these had similar methodological issues as MRI studies.
- Fewer studies examined bipolar disorder, PTSD, TBI, SUD, OCD, and anxiety disorders; they were mostly focused on diagnosis and were cross-sectional and small in size.
- 14 studies included US Veterans, addressing PTSD, TBI, and/or SUD:
  - All 11 diagnostic studies were cross-sectional, 2 prognostic studies were cohorts, and 1 was an RCT.

INTRODUCTION

Mental health conditions and traumatic brain injury (TBI) are common among Veterans and often negatively impact Veterans, their families, and their communities. The Department of Veterans Affairs (VA) devotes considerable resources to treating these conditions and improving diagnosis and treatment outcomes is an ongoing VA priority.

There have been substantial advancements in precision medicine, specifically the use of biomarkers and/or genetics in diagnosis, prognosis, and tailoring treatments for medical conditions. There are also several ongoing large-scale population-based studies to advance precision medicine, including the VA's Million Veterans Program. In the context of mental health, precision medicine has involved assessment of brain structure and functioning, as well as genetics and serum biomarkers. Despite advances in the development and availability of these tools, challenges to precision medicine for mental health conditions remain. These include complex and heterogeneous clinical phenotypes, high cost and technical difficulty of obtaining neuroimaging and neurophysiologic data, and differing assessments of symptoms and treatment response. Although these challenges have contributed to concerns about the reproducibility and validity of findings, results of more recent efforts to systematically collect and examine large neuroimaging datasets have yielded more promising results. Thus, future work in this area may yet produce insights that improve diagnosis and treatment outcomes in mental health.

This evidence review was requested by the VA Working Group to implement the Commander John Scott Hannon Veterans Mental Health Care Improvement Act (P.L. 116-171), Section 305: "Precision Medicine for Veterans Initiative" (SHA305). SHA305 tasks the VA with developing and implementing a precision medicine initiative focused on brain and mental health biomarkers.

To support the VA SHA305 Working Group, we conducted an evidence map to better understand characteristics of existing evidence on relationships between brain structure and functioning, and mental health conditions and TBI. An evidence map is well suited to address a broad scope covering multiple conditions and numerous neuroimaging and neurophysiological techniques, particularly when some of the evidence base may consist of more exploratory studies. An evidence map is also appropriate for meeting the overall goals of informing research policy and potential clinical demonstrations.

In this report, we provide descriptive information about the number and types of studies that address a wide range of neuroimaging and neurophysiologic assessments for diverse mental health conditions and TBI. We also highlight weaknesses and gaps in the evidence, as determined by the volume and characteristics of studies.

METHODS

Key Question (KQ)

KQ: What are the quantity, distribution, and characteristics of evidence assessing the accuracy and utility of neuroimaging and neurophysiologic biomarkers in the diagnosis and clinical management of the following conditions:

a) Depression
b) Anxiety
c) Posttraumatic stress disorder (PTSD)
d) Substance use disorder (SUD)
e) Bipolar disorder
f) Traumatic brain injury

Data Sources and Searches

We searched for peer-reviewed English language articles from January 2010 to April 2022 in MEDLINE and Embase. We used Medical Subject Headings (MeSH) and title/abstract terms for neuroimaging and neurophysiological tests and conditions of interest. We also searched websites for VA ESP and AHRQ EPC programs to identify relevant reviews.

Study Selection

Abstracts were screened with the assistance of DistillerSR's Artificial Intelligence System (DAISY) in 2 separate phases (see Methods section for full details).

For full-text review, we undertook 2 initial pilot rounds in which all reviewers separately determined eligibility for 10–15 articles in each round. We discussed articles to reach consensus on eligibility, with further clarification on operationalization of inclusion and exclusion criteria. Eligibility of remaining articles was determined by 1 reviewer, with ~50% of these also undergoing evaluation by a second reviewer.

Eligible populations included adults with at least 1 of the conditions of interest, as noted in KQ above. Eligible articles evaluated at least 1 neuroimaging or neurophysiological test of interest (*eg*, magnetic resonance imaging [MRI], including functional MRI [fMRI], and evoked potentials and electroencephalogram [EEG]) for diagnostic accuracy, clinical prognosis, and/or treatment response. Exclusion criteria included pediatric populations, evaluation of symptoms or cognitive functioning only in the context of neurodegenerative conditions or intracranial injury. We also excluded studies attempting to evaluate prognosis using exclusively cross-sectional data.

Data Abstraction and Assessment

We abstracted the following data from all eligible studies: population characteristics (*eg*, condition and method of diagnosis, sample size, demographic data (*eg*, mean or median age, proportion of women, focus on Veterans or combat exposure); neuroimaging test and/or EEG being evaluated (and genetic data if used); outcomes addressed (clinical diagnosis and/or prognosis); and study design (*eg*, cross-sectional or cohort, analytic methods used to assess diagnostic or prognostic accuracy). To verify accuracy of abstracted results, data from ~50% of articles were over-read by a second reviewer.

Quality Assessment and Summary of Results

We did not conduct formal quality assessment of eligible studies included in this report. We also did not undertake a formal synthesis of study results. Our results summaries are organized by the conditions of interest and focus on describing the characteristics of study populations, outcomes (clinical diagnosis, prognosis, and/or treatment response), and study designs (including analytic methods) of eligible studies.

RESULTS

Overview

From 50,989 unique search results, we identified 313 primary studies and 30 systematic reviews. At abstract screening, 47,586 results were excluded, with 54% of these based on low scores from a machine-learning algorithm (see Methods). Most eligible primary studies and systematic reviews addressed depression ($k = 236$, 69%), while fewer studies and reviews evaluated other conditions. Only 2 studies evaluated genetic data in addition to neuroimaging or neurophysiologic data. Three-quarters of primary studies used MRI-based imaging techniques ($k = 236$, 75%), while a fifth used EEG data ($k = 68$, 22%). For multiple conditions, there were none or few studies ($k \leq 5$) examining either diagnosis or prognosis.

Most primary studies had small sample sizes, with only 9 having more than 500 participants (range 555–4,541). Two-thirds of primary studies examined diagnosis ($k = 200$), 110 evaluated prognosis, and 3 addressed both diagnosis and prognosis. Most studies included young and middle-aged participants; only 5 studies had participants with mean ages of 65 or older and all of these addressed depression.

Depression

MRI-based Imaging Techniques (Structural and Functional MRI, DTI, and ASL)

Of 104 studies using MRI-based techniques to address diagnosis of depression, most used structural MRI ($k = 49$), fMRI ($k = 48$), or both ($k = 1$). A few studies used other MRI-based

techniques like diffusion tensor imaging (DTI, $k = 6$) and arterial spin labeling (ASL, $k = 2$). Most were cross-sectional ($k = 91$), while those remaining were cohort/longitudinal ($k = 13$). Three-quarters of studies used machine learning methods to develop models ($k = 75$). Nearly all studies assessed diagnostic model accuracy ($k = 100$) and sensitivity/specificity ($k = 93$). Three-quarters also undertook model validation ($k = 77$). Total sample sizes ranged 30–4541; half of the studies had n<100 ($k = 57$) and only 4 had $N > 1000$. Most included healthy controls ($k = 99$), while a quarter also had participants with bipolar disorder ($k = 26$). A third focused particularly on participants not on medications ($k = 34$). A fifth of studies included participants with their first episode of depression ($k = 19$). Most studies had substantial proportions of women ($k = 92$ with women >40%). Most participants were young and middle-aged; only 5 studies reported race. The most common study locations were China ($k = 57$) and the US ($k = 17$). The most frequent measures for determining diagnostic accuracy were standardized clinician assessments (*eg*, $k = 89$ studies used Hamilton Depression Rating Scale [HAM-D]). Clinician interviews were also used, including the Structured Clinical Interview for DSM (SCID; $k = 74$) and Mini-International Neuropsychiatric Interview (MINI; $k = 18$). Fewer studies used patient-reported measures such as the Beck Depression Inventory (BDI; $k = 21$).

Of 59 studies evaluating prognosis, most also used structural MRI ($k = 22$), fMRI ($k = 31$), or both ($k = 2$); few used DTI ($k = 5$). Nearly all studies examined treatment response ($k = 55$), most commonly to antidepressant therapy ($k = 36$). Fewer studies evaluated response to psychotherapy ($k = 6$), electroconvulsive therapy (ECT, $k = 9$), repetitive transcranial magnetic stimulation (rTMS, $k = 5$), transcranial direct current stimulation (tDCS, $k = 1$), theta burst stimulation (TBS, $k = 1$), or inpatient multi-modal treatment ($k = 1$). Two studies evaluated general trajectories over 2 years for middle-aged and older adults with depression. Twenty-two studies applied machine learning approaches and 34 validated predictive models. Most were cohorts/longitudinal observational studies ($k = 52$) and 7 were RCTs. A single study had total n>1000, while half had $N < 100$ ($k = 31$). Studies that focused on medication-free participants defined this variably, including those who had not received treatment for the current depressive episode or had undergone a washout period ($k = 24$). Others focused on treatment-resistant depression ($k = 11$). Only 2 studies distinguished participants in their first episode of depression. A third of studies included healthy controls ($k = 21$), and a few had participants with bipolar disorder ($k = 4$). Studies had relatively young participants, and women were well represented. Demographic information relating to race/ethnicity was reported in 9 studies. The most common locations were the US or Canada ($k = 21$) and China ($k = 12$).

EEG and Evoked Potentials

Of 24 studies evaluating EEG or evoked potentials for diagnosis of depression, most included healthy controls ($k = 23$) and were very small with total $N < 100$ ($k = 21$). All diagnostic studies were cross-sectional. Only 2 studies focused on participants in their first episode of depression. Study participants were young and middle-aged adults (mean age range 20–55), and more than half of studies had >40% women ($k = 17$). The most common study location was China ($k = 7$). Standardized clinician assessments (HAM-D and Montgomery-Asberg Depression Rating Scale [MADRS]) were the most frequently used diagnostic standard ($k = 14$). A majority of studies used machine learning methods ($k = 17$) and undertook model validation ($k = 20$).

Thirty studies examined prognosis in depression; most addressed response after antidepressant therapy ($k = 19$), while fewer evaluated rTMS ($k = 9$) and 1 study each examined acupuncture or

ketamine. Most were cohorts/longitudinal observational (k = 25) and a few used data from RCTs (k = 4). A third included medication-free participants (k = 11), and 8 focused on treatment-resistant participants (variably defined as not responding to sufficient course of antidepressants). No study included only participants with their first episode of depression. Six included participants who were healthy controls. The majority of studies had $N < 100$ (k = 22). Studies were most commonly conducted in the US or Canada (k = 13). Studies most commonly used standardized clinician assessments (HAM-D and MADRS) to define treatment response (k = 25). A third used machine learning (k = 9). Just under half undertook model validation (k = 12).

Other Neuroimaging Techniques (MEG, PET, and SPECT)

Eight eligible studies evaluated magnetoencephalography (MEG) for depression; 7 examined diagnosis and one addressed treatment response to antidepressants. Five studies also used MRI-based imaging techniques. All diagnostic studies had healthy controls as comparators, while one also included individuals with bipolar disorder. All were conducted in China or Taiwan and were very small (total N range 41–108). Participants were young (mean age range 30–37) and women were well represented (37–61% across studies). Six diagnostic studies were cross-sectional in design, and one was a longitudinal cohort. All studies used structured interviews as the gold standard, and 6 also used HAM-D as the standardized clinician assessment. The prognostic study on outcomes with antidepressants also used HAM-D to define response. Three studies used machine learning methods, and 6 validated models.

Four studies evaluated positron emission tomography (PET) for diagnosis (k = 2) or prognosis (k = 2) in depression. Three of these also used structural MRI to improve localization of PET data. Both diagnostic studies were cross-sectional and conducted in the US. Both prognostic studies occurred in Taiwan, with one being an RCT and the other an observational cohort. Studies were very small (N range 36–107 total participants) and included mostly young adults (mean age range 32–43). None of the studies used machine learning methods, and none conducted model validation.

Lastly, 3 eligible studies used single photon emission computed tomography (SPECT) for diagnosis (k = 1) or prognosis (k = 2). The diagnostic study was very large (N = 4,541), conducted in the US, used a structured clinical interview (MINI) as the gold standard, and undertook model validation. Both prognostic studies were conducted by 1 research group in France, evaluated response to rTMS, and also included participants with bipolar disorder. They had small samples (N = 33–58) and used patient-reported outcome (BDI) to determine response. None of the SPECT studies used a machine learning approach.

Bipolar Disorder

Forty-seven eligible studies evaluated diagnosis (k = 41) or prognosis (k = 6) for bipolar disorders. The majority also included participants with depression (k = 27 for diagnostic studies, and all prognostic studies). Nearly all studies examining diagnosis used MRI-based techniques (k = 24 with structural MRI, k = 19 with functional MRI, k = 3 with DTI, and k = 2 with ASL), with one of these also using magnetoencephalography (MEG). One study examined EEG for diagnosis. Half of studies included healthy controls (k = 23), and half were very small with total sample sizes less than 100 (k = 23). Only 3 studies had more than 250 participants (range 251–441). Most study participants were young adults, with only 2 studies having mean ages of 45 or

older. Most studies had at least 40% women (k = 44). Studies were conducted in different regions of the world, with most common locations being China (k = 15) and the US (k = 12).

Most diagnostic studies were cross-sectional in design (k = 31), while 3 were longitudinal (to confirm symptoms and diagnosis over 1–2 years). About half of diagnostic studies used machine learning methods (k = 25) and undertook model validation (k = 24). Less than half of studies used both structured clinical interviews (MINI and/or SCID) and standardized clinician assessments (Young Mania Rating Scale [YMRS]) as the diagnostic standard for bipolar disorder (k = 16). Another 18 studies used only structured interviews, and 3 used only YMRS. All prognostic studies were included above in results for depression.

Posttraumatic Stress Disorder

Thirty eligible articles evaluated PTSD, with the majority focusing on diagnosis (k = 24). Most used MRI-based techniques, including fMRI (k = 11), structural MRI (k = 7), both MRI and fMRI (k = 1), or fMRI and DTI (k = 1). Remaining studies evaluated PET (k = 1), SPECT (k = 2), MEG (k = 1), or EEG (k = 5). The majority were cross-sectional (k = 22), with fewer being longitudinal cohorts (k = 6) or RCT (k = 2). Most were small, with more than half having sample sizes <100 (k = 17). Sample sizes for remaining studies were 116–432 (k = 12) and 2,137 for 1 large database study. Studies were conducted mostly in the US or Canada (k = 18) and China (k = 7); a few were conducted in the Netherlands (k = 2), South Korea (k = 2), and Iran (k = 1). A third included US Veterans or active military (k = 10), with half of these including combat-exposed Veterans or active military (k = 5).

The most common assessments used for diagnostic standard included structured interviews (SCID, k = 12) and clinician assessments (Clinician Administered PTSD Scale [CAPS], k = 13). Many also used patient-reported outcome measures such as the PTSD Checklist (PCL, k = 7). Ten studies using machine learning methods, and 12 undertook model validation.

Studies in Veteran Populations

A total of 13 studies on PTSD were conducted in Veteran populations, the majority (k = 10) with US Veterans, 2 in combat-exposed Veterans from the Netherlands, and one in combat-exposed members of the Canadian Armed Forces. Of the 10 studies of US Veterans, nearly all evaluated diagnostic accuracy (k = 9). Half used MRI-based techniques (k = 5), while remaining used a variety of other methods (SPECT k = 1, MEG k = 2, EEG k = 2). Five studies included participants with co-occurring TBI. Diagnostic standards included SCID, CAPS, and patient-reported measures such as BDI, PCL, or Patient Health Questionnaire (PHQ). Six studies undertook model validation. Sample sizes ranged from 32–196. The single prognostic study used fMRI and clinical data from a small RCT evaluating response to an integrated psychotherapy to treat comorbid PTSD and alcohol use disorder.

Traumatic Brain Injury

Of 12 articles on TBI, most evaluated diagnosis (k = 10) and 2 reported on prognosis of disability. The majority used MRI-based techniques (k = 8), and fewer used EEG (k = 2) or SPECT (k = 2). One of the MRI studies also used MEG. Most were cross-sectional (k = 10), small in size (*eg*, k = 9 with N < 100), and included younger populations (k = 10 with mean age <45). Six studies included PTSD; all of these focused on diagnosis and were cross-sectional.

Both prognostic studies investigated predictive models for global disability at least 1 year after injury, measured using the Glasgow Outcome Scale-Extended (GOSE).

Studies in Veteran Populations

Seven studies included combat-exposed US Veteran populations. Most included participants with co-occurring PTSD and TBI (k = 5), and most evaluated MRI-based techniques (k = 5). One each evaluated EEG, SPECT, or MEG (this study also used MRI). All studies focused on diagnosis and were cross-sectional.

Substance Use Disorders (SUD)

Twenty studies addressed SUD, with 60% evaluating alcohol use disorder (k = 12) and the remaining studies focusing on cocaine use disorder (k = 3), opioid use disorder (k = 2), or methamphetamine use disorder (k = 3). The majority used structural and/or functional MRI (k = 12) or other MRI-based techniques (ASL, k = 2). Eight evaluated EEG or evoked potentials; no studies used other imaging techniques. About half focused on diagnosis (k = 9), while the rest reported on prediction of relapse (k = 6) or treatment response (k = 5). Most evaluated the accuracy of diagnostic or prognostic models (k = 16), and 40% undertook model validation (k = 8). Most studies were very small with total sample sizes <100 (k = 14); 1 study had a total sample size of 1,376. More than half used machine learning methods to develop models (k = 11). The most common locations were the US (k = 10) and China (k = 3).

Studies in Veteran Populations

We identified 3 studies that included US Veterans and all focused on prognosis. All three used structural MRI or fMRI techniques. One was an RCT including Veteran participants with comorbid PTSD and alcohol use disorder and evaluated improvement in PTSD symptoms with psychotherapy. The other 2 were cohort studies including both Veteran and civilian populations. One addressed predictors of relapse with treatment for alcohol use disorder, and the other examined relapse in methamphetamine use after inpatient treatment. None of these studies validated their predictive models.

Obsessive Compulsive Disorder and Anxiety Disorders

Obsessive Compulsive Disorder

Seventeen studies focused on diagnosis of OCD and all were cross-sectional. Two cohort studies evaluated prognosis. Overall, 12 studies used fMRI, 5 used structural MRI, and 1 each applied DTI and EEG. The most commonly used diagnostic standard was the Yale Brown Obsessive Compulsive Scale (Y-BOCS, k = 15), with 11 studies also using SCID. The 2 prognostic studies also used Y-BOCS to define treatment response to psychotherapy and antidepressants, respectively. Eight studies undertook model validation, and 7 used machine learning. All studies had sample sizes <200 and included young adults. Half had 16-40% women participants (k = 10), while 7 included 41–70% women. Most included healthy controls as the comparator (k = 14), and most were conducted in China (k = 14).

Anxiety Disorders

Four studies addressed diagnosis and 6 evaluated prognosis. All used either structural MRI (k = 3), or fMRI (k = 7). All diagnostic studies used the SCID and/or the Hamilton Anxiety Rating

Scale (HAM-A) as the diagnostic standards. Specific disorders examined were general anxiety disorder (k = 7), social anxiety disorder (k = 2), and panic disorder (k = 3). Three studies were cross-sectional, and those remaining were longitudinal cohorts. Most studies undertook model validation (k = 8). Four studies used machine learning. Sample sizes ranged 34–135 and included young adults with substantial representation of women. Most studies were conducted in the US (k = 6) or China (k = 2). Of prognostic studies, 4 addressed response to psychotherapy, 1 evaluated outcomes after antidepressant therapy, and 1 examined response to a computer-based behavioral intervention.

Systematic Reviews

Of 30 eligible systematic reviews**,** 17 addressed depression. Fewer reviews evaluated the remaining conditions: anxiety disorders (k = 3), bipolar disorders (k = 4), PTSD (k = 2), TBI (k = 3), or OCD (k = 1); no eligible reviews addressed SUD or reported on more than 1 condition. Most systematic reviews included MRI-based techniques (k = 16) or a number of neuroimaging or neurophysiologic data (k = 7). Fewer focused on EEG (k = 5), PET (k = 1) or SPECT (k = 1).

About half examined diagnosis (k = 16), 15 addressed response to treatment, and 3 evaluated change in symptoms or functioning. Four reviews reported on both diagnosis and prognosis**.** The number of studies included by reviews varied widely, ranging from 11–352.

DISCUSSION

Summary of Key Findings

To inform next steps for applying precision medicine to Veterans' healthcare and research, we conducted an evidence map of neuroimaging and neurophysiologic biomarkers in mental health and TBI. We identified 313 eligible primary studies and 30 eligible systematic reviews. The majority of the evidence addressed depression, while fewer studies and reviews examined other conditions of interest. Most primary studies used MRI-based neuroimaging techniques and a fifth employed EEG. Two-thirds of primary studies focused on diagnosis for conditions of interest, and nearly all of these were cross-sectional. Half of primary studies employed machine learning to analyze neuroimaging or neurophysiologic data and develop diagnostic or prognostic models. Primary studies generally included young and middle-aged adults, with only 5 studies having participants with mean ages ≥65. Studies were conducted in diverse locations around the world, with the most common being China, the US, or Canada; very few were conducted in more than 1 country. Overall, most of the evidence came from very small studies. Only 14 primary studies included US Veterans or active military service members; 12 addressed PTSD and/or TBI, and 2 evaluated SUD.

Key findings for primary studies include:

- Many studies evaluated structural or functional MRI for diagnosis and prognosis of depression, but there were important methodological concerns:

  - Nearly all diagnostic studies were cross-sectional, small in size, and included participants with variable past histories of symptoms and treatments.
  - Prognostic studies mostly focused on response to antidepressants, and were also generally small.
- A substantial number of studies used EEG for diagnosis and prognosis of depression, but these had similar methodological issues to MRI studies.
- Most studies on bipolar disorder were small and cross-sectional, included participants with depression, and focused on diagnosis.
- Studies evaluating PTSD were small and cross-sectional, and mainly used structural or functional MRI to address diagnosis.
- Studies examining TBI were small and cross-sectional, often included participants with co-occurring PTSD, and mainly used structural or functional MRI to address diagnosis.
- Studies on SUD used structural or functional MRI and EEG, most addressed alcohol use disorder, and half evaluated prediction of relapse or response to treatment.
- Studies on OCD and anxiety disorders were small and cross-sectional, mainly used structural or functional MRI, and focused on diagnosis.
- Fourteen studies included US Veterans, addressing PTSD, TBI, and/or SUD:
  - All 11 diagnostic studies were cross-sectional, 2 prognostic studies were cohorts, and 1 was an RCT.
- None evaluated prediction of adverse or side effects from treatments.

Implications for VA Policy

We found a large number of studies mainly using MRI-based techniques to evaluate diagnosis and prognosis for depression, but there were substantial methodological limitations. Additionally, none of the depression studies were conducted with US Veterans or military service members. Given that neuroimaging tests are costly and time consuming to conduct (and analyze), it is not clear that using such tests adds value in the clinical setting or that they could replace current standards for diagnosis of depression, which involve structured interviews and clinician assessments. Regarding prognosis, neuroimaging techniques may potentially aid in predicting early response and/or selection of appropriate therapies, but most studies included participants with variable histories of symptoms and past treatments. Only 2 studies focused on participants with their first episode of depression. Furthermore, no study evaluated prediction of adverse or side effects of treatments, whereas this is often an important factor in patient and clinician decisions to stop or switch antidepressants. There were fewer studies using EEG to examine depression and this evidence base has similar limitations as that evaluating MRI-based techniques. Thus, it is unclear how these data could be incorporated into current clinical practice to improve diagnosis or treatment selection and/or monitoring. Future systematic reviews focused on these techniques for diagnosis and/or prognosis in depression may also be needed to better characterize their potential utility for clinical care.

We found considerably less evidence addressing other mental health conditions and TBI, and fewer studies using other neuroimaging and neurophysiologic techniques. Although there were some studies on PTSD, TBI, and SUD that included US Veterans or military service members, overall these shared the same methodological limitations noted above. Therefore, it also appears premature to implement MRI (and other neuroimaging and neurophysiologic techniques) in the clinical diagnosis and treatment of these other conditions.

Future Research

While there are a large number of studies examining depression (using MRI or EEG), these were generally small and the majority used cross-sectional data to evaluate diagnosis. Additionally, participants often had variable trajectories of symptoms and treatments preceding data collection. These study design issues contribute to problems with replicability and validity of neuroimaging and neurophysiologic studies in mental health. Whereas most of the identified primary studies had less than 100 participants, current estimates are that thousands of individuals are needed to provide stable and valid results regarding important associations between neuroimaging findings and clinical phenotypes. Furthermore, it may be critical to use comparisons with age-standardized findings (developed from large populations) instead of data from small samples of age-matched controls. Additional considerations include the need for longitudinal data on symptoms and exposures, and transdiagnostic dimensional approaches in understanding clinical phenotypes. Having data before certain exposures may also be particularly important for studies evaluating PTSD and TBI.

The acquisition and analysis of (longitudinal) data from many individuals will likely require large ongoing investments in this research, as well as fundamental changes in research organization and incentives that currently promote competition and inhibit data sharing. Current projects that exemplify the level of resources, organization, and cooperation needed for such efforts include the Adolescent Brain Cognitive Development (ABCD) Study in the US and the UK Biobank.

Therefore, we recommend the following:

- Consider investment in larger studies (thousands of participants) to identify reproducible and precise associations between neuroimaging and neurophysiologic findings and mental health phenotypes.
- Conduct longitudinal studies with data on exposures, symptoms, and neuroimaging and neurophysiologic data over the life course.
- Consider transdiagnostic approaches for describing mental health phenotypes.
- Particularly for addressing Veterans' health and outcomes, develop longitudinal studies with initial data that precede combat and other service-related exposures.

Limitations

We sought to identify and describe the evidence for a broad range of neuroimaging and neurophysiologic tests used to evaluate the diagnosis and prognosis of a large number of mental health conditions and TBI. Therefore, we conducted an evidence map that provides descriptive information about research studies examining these questions and highlights gaps in the existing

evidence. Thus, we did not abstract detailed results for diagnostic or prognostic models using neuroimaging and/or neurophysiologic data. We also did not formally evaluate the quality of included primary studies or systematic reviews. Additionally, we employed machine-learning techniques to assist with the selection of relevant studies and reviews; it is possible that we may have missed some eligible studies. We also limited our search of the evidence to English-language studies and reviews.

Conclusions

Most existing evidence on neuroimaging and neurophysiologic data for mental health conditions evaluated MRI for diagnosis and prognosis in depression. In addition to the lack of evidence on other conditions or using other types of neuroimaging and neurophysiologic data, most existing studies were limited by small sample sizes and cross-sectional designs. These methodological concerns need to be addressed by future research using larger samples with longitudinal data. Existing evidence gaps and limitations indicate that it may be premature to apply neuroimaging and neurophysiologic tests to evaluate and treat mental health conditions and TBI in clinical settings.