Review of Department of Veterans Affairs Monograph on Potential Therapeutic Effects of Service and Emotional Support Dogs on Veterans with Post Traumatic Stress Disorder

Committee on Review of Department of Veterans Affairs Monograph on Potential Therapeutic Effects of Service and Emotional Support Dogs on Veterans with Posttraumatic Stress Disorder

Institute for Laboratory Animal Research

Division on Earth and Life Studies

A Consensus Study Report of
The National Academies of
SCIENCES • ENGINEERING • MEDICINE

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

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

This activity was supported by a contract between the National Academy of Sciences and the Department of Veterans Affairs under Contract No. 36C24E20C0006. Any opinions, findings, conclusions, or recommendations expressed in this publication do not necessarily reflect the views of any organization or agency that provided support for the project.

Digital Object Identifier: https://doi.org/10.17226/25827

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

Printed in the United States of America

Suggested citation: National Academies of Sciences, Engineering, and Medicine. 2021. *Review of Department of Veterans Affairs Monograph on Potential Therapeutic Effects of Service and Emotional Support Dogs on Veterans with Post Traumatic Stress Disorder*. Washington, DC: The National Academies Press. https://doi.org/10.17226/25827.

The National Academies of
SCIENCES • ENGINEERING • MEDICINE

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

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

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

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

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

The National Academies of
SCIENCES • ENGINEERING • MEDICINE

Consensus Study Reports published by the National Academies of Sciences, Engineering, and Medicine document the evidence-based consensus on the study's statement of task by an authoring committee of experts. Reports typically include findings, conclusions, and recommendations based on information gathered by the committee and the committee's deliberations. Each report has been subjected to a rigorous and independent peer-review process and it represents the position of the National Academies on the statement of task.

Proceedings published by the National Academies of Sciences, Engineering, and Medicine chronicle the presentations and discussions at a workshop, symposium, or other event convened by the National Academies. The statements and opinions contained in proceedings are those of the participants and are not endorsed by other participants, the planning committee, or the National Academies.

For information about other products and activities of the National Academies, please visit www.nationalacademies.org/about/whatwedo.

COMMITTEE ON REVIEW OF DEPARTMENT OF VETERANS AFFAIRS MONOGRAPH ON POTENTIAL THERAPEUTIC EFFECTS OF SERVICE AND EMOTIONAL SUPPORT DOGS ON VETERANS WITH POSTTRAUMATIC STRESS DISORDER

Members

DAVID B. ALLISON, *Chair*, Indiana University, Bloomington
MARY BURCH, American Kennel Club, New York, NY
JALPA DOSHI, University of Pennsylvania, Philadelphia
CHERYL GISCOMBE, University of North Carolina at Chapel Hill
ERINN HADE, The Ohio State University, Columbus
STEWART HILLIARD, U.S. Air Force, Lackland Air Force Base, TX
EVAN MAYO-WILSON, Indiana University, Bloomington
LORI ZOELLNER, University of Washington, Seattle

Staff

ANDREA HODGSON, Study Director, Board on Life Sciences
AUDREY THÉVENON, Program Officer, Board on Life Sciences
TERESA SYLVINA, Board Director, Institute for Laboratory Animal Research
KOSSANA YOUNG, Senior Program Assistant, Board on Life Sciences
ROBERTA WEDGE, Senior Program Officer, Board on Health Care Services

INSTITUTE FOR LABORATORY ANIMAL RESEARCH COUNCIL

MARGARET S. LANDI, *Chair*, GlaxoSmithKline, Brentford, United Kingdom
CORY BRAYTON, Johns Hopkins University, Baltimore, Maryland
ROBERT DYSKO, University of Michigan Medical School, Ann Arbor
ANNE MAGLIA, University of Massachusetts, Lowell
SUZAN MURRAY, Smithsonian Conservation Biology Institute, Front Royal, Virginia
BARBARA NATTERSON HOROWITZ, University of California, Los Angeles
GUY PALMER, Washington State University, Pullman
ROSALIND ROLLAND, New England Aquarium, Boston, Massachusetts

Staff

TERESA SYLVINA, Director

Reviewers

This Consensus Study Report was reviewed in draft form by individuals chosen for their diverse perspectives and technical expertise. The purpose of this independent review is to provide candid and critical comments that will assist the National Academies of Sciences, Engineering, and Medicine in making each published report as sound as possible and to ensure that it meets the institutional standards for quality, objectivity, evidence, and responsiveness to the charge. The review comments and draft manuscript remain confidential to protect the integrity of the process.

We thank the following individuals for their review of this report:

MARISA DOMINO, University of North Carolina at Chapel Hill
BRIAN HEALY, Harvard Medical School
PAUL MUNDELL, Consultant
ELSPETH CAMERON RITCHIE, MedStar Washington Hospital Center
IAN SALDANHA, Brown University
LANCE WALLER, Emory University

Although the reviewers listed above provided many constructive comments and suggestions, they were not asked to endorse the conclusions or recommendations of this report, nor did they see the final draft before its release. The review of this report was overseen by **ALICIA CARRIQUIRY,** Iowa State University, and **KATHRYN BASHAM,** Smith College School for Social Work. They were responsible for making certain that an independent examination of this report was carried out in accordance with the standards of the National Academies and that all review comments were carefully considered. Responsibility for the final content rests entirely with the authoring committee and the National Academies.

Contents

Summary

Posttraumatic stress disorder (PTSD) is a leading cause of impairment in quality of life and functioning among Veterans. Service dogs[1] (SERVs) have been promoted as a potential intervention for Veterans with PTSD; however, research supporting their effectiveness is limited. The National Defense Authorization Act of 2010 (NDAA), directed the Department of Veterans Affairs (VA) to conduct a study to assess the potential therapeutic and economic effects of using SERVs for the treatment or rehabilitation of Veterans with physical or mental injuries or disabilities (which was defined to include PSTD). However, the study design described in the eventual monograph (reviewed by this committee) specifically compares the effectiveness of SERVs to emotional support dogs (EMOTs). The NDAA also mandated the VA to have the study reviewed by a committee of the National Academies of Sciences, Engineering, and Medicine. This report serves as the committee's review of the VA's draft monograph addressing the completeness and accuracy of reporting; the rigorousness of the study design, conduct, and data analysis; and the scientific validity of the conclusions presented within the draft monograph (the complete Statement of Task can be found in Chapter 1). To address whether the draft monograph was comprehensive, the committee assessed the draft monograph's adherence to the Consolidated Standards of Reporting Trials (CONSORT) guidelines,[2] which describe the minimum information needed to understand how a trial was conducted and what was found. In addition to the CONSORT guidelines, the committee identified a number of issues related to the design, statistical analyses planned and conducted, and the interpretation of the study results.

GENERAL ASSESSMENT

While the committee identified many elements of the report that it believes merit change, the vast majority of those elements are addressable with some rewriting, additional information, or additional statistical analyses. The committee recognizes that several of the issues identified with respect to the design of the study cannot be altered now that the study is complete. However, these can generally be conceptualized and addressed within the draft monograph as limitations rather than considered fatal flaws. Thus, on balance, the draft monograph describes a study of an important topic that appears to have been well executed; however, it is limited in the specific conclusions it can potentially support. The committee believes that with some non-trivial revisions the study results can be well analyzed, interpreted, and reported upon within the draft monograph.

ETHICAL CONSIDERATIONS

A strength of this draft monograph was the manner in which the authors reported addressing ethical and animal welfare issues. The authors appear to have followed the ethical guidelines that are

[1]Under the Americans with Disabilities Act (ADA), service animals are defined as dogs that are individually trained to do work or perform tasks for people with disabilities. Service dogs are working animals, not pets. The work or task a dog has been trained to provide must be directly related to the person's disability. Whereas, dogs whose sole function is to provide comfort or emotional support do not qualify as service animals under the ADA. These dogs are considered emotional support animals and they provide comfort to help relieve a symptom of a person's disability. They have not been trained to perform specific tasks related to the person's disability. In theory they provide a benefit by being present.

[2]See http://www.consort-statement.org (accessed July 1, 2020).

commonly recommended for research and ensured that both human and animal participants were protected from undue risk and harm throughout the study.

COMPLETENESS OF REPORTING

The committee followed the CONSORT guidelines checklist reviewing how each element was addressed in the draft monograph. When reports of clinical trials do not describe the items in these guidelines, it may be difficult or impossible to assess risk of bias adequately, interpret study results, or apply the results to policy and practice decisions. The committee identified several areas in which the draft monograph does not adhere to CONSORT guidelines and it suggests that the authors provide the missing information and utilize the CONSORT guidelines when revising the draft monograph.

STATISTICAL METHODS

The committee reviewed how the statistical methods are described and presented within the draft monograph and also commented on the appropriateness of the methods for the study itself.

Intent to Treat Analyses and Design Characteristics

As noted in the study protocol, analyses were to be carried out for both the intent to treat (ITT) and the per-protocol population (PP); however, the protocol does not specify a priori which analysis (the ITT or PP) will be considered the primary inferential analysis. The ITT analysis of the full randomized sample is considered the gold standard for inference in superiority trials. Unfortunately, the ITT analyses of the primary and secondary outcomes are not included in the draft monograph. The vast majority of presented results are devoted to the subset of participants who were successfully paired with a dog, the PP cohort. It is unclear in the protocol and the draft monograph if the intention was to deem the intervention a success if all primary outcome measures were in favor of the SERV intervention (which would define co-primary outcomes), or if the trial would be a success if any one of the outcome measures were in favor of the experimental group (multiple primary outcomes).

Interpretation of Clinical Significance and Effect Sizes

The study results should be interpreted with respect to their magnitude and precision, emphasizing the pre-specified primary outcomes. The interpretation of overall trial results should be balanced with respect to primary, secondary, and exploratory outcomes and analyses, instead of relying on the conclusions that emphasize only those with a statistically "significant" result, particularly given the number of secondary and exploratory analyses performed. The unadjusted primary and secondary endpoints should be provided at baseline, at the time of dog pairing, any intermediate time points, and at 18 months, for all randomized participants, along with the estimated difference between groups in each measure and 95% confidence interval (CI) at 18 months.

Missing Data and Loss to Follow-Up

In the protocol and design paper, the study team set out to describe missing data and the patterns by treatment assignment group, and to employ multiple imputation ("when needed"). However, the overall description of the multiple imputation procedure is lacking. Detail is needed as to what covariates were included in the missing data model, in what software the missing data model was estimated, the number of imputed datasets this included, and the analysis methods employed following multiple imputation.

INTERPRETATION

The committee considered aspects of the draft monograph that relate to the ability to appropriately interpret the study results. This included considering the study design and how to improve the clarity of the draft monograph. Overall, the conclusions are not sufficiently supported by the evidence presented because they do not address important limitations of the study.

Interpretation of Within-Group Change

The study design limits the range of possible interpretations because it did not include a no treatment condition. It is conceivable that Veterans would have improved in the absence of the dogs, and Veterans might have improved to an even greater extent than they improved in the study. A no treatment comparator would have been appropriate because in clinical trial design it is customary to establish that one intervention is better than nothing before comparing two similar interventions. The authors appropriately acknowledge the importance of the lack of a control group (lines 2494-2496), though this does not translate well into the interpretation of the study findings regarding within-group change. It is important to point out that within-group change is not evidence of efficacy or effectiveness.

Interpretation of Potential Equivalence

The absence of a difference between groups is not evidence of equivalence. Because the study was designed and powered to detect the *superiority* of the SERV intervention over the EMOT intervention, the results can only be interpreted as failing to reject the hypothesis that SERV was more effective than EMOT. Evidence of non-zero effects can only come from the between conditions (not within condition) tests and lack of statistically significant evidence of differential effectiveness should not be mistaken for evidence of a lack of differential effectiveness.

Non-Masked Lack of Equipoise Design Limitations

The concept of equipoise refers to whether providers and participants in a trial have equivalent beliefs and feelings about the conditions to which the participants may be assigned such that any differences in beliefs and feelings do not provide an alternative explanation for the effects of treatment assignment. This is particularly a problem in clinical trials where there is one clearly preferred intervention. Three ways of addressing equipoise in the context of a trial such as this study are available: enhanced study design elements; measurement of patient expectancies, preferences, and satisfaction; and careful discussion of study limitations. The draft monograph describes steps that were taken to account for this; however, the committee identified areas where this concept could have been better implemented or data that could have been collected to enable measurement of the expectancy between the groups.

Not Designed as a PTSD Treatment Trial

The authors should clarify that the clinical trial was not designed to test a primary intervention in the treatment of PTSD. Throughout the discussion, a naïve reader would assume that it indeed was a well-done PTSD treatment trial. If the focus was on PTSD, interviewer-administered PTSD assessment would have been the primary outcome measure, a threshold of PTSD severity would have been specified as an eligibility criterion, and assessment would have been prior to the implementation of the intervention (a proximal baseline measured) and assessed multiple times throughout the trial to develop a trajectory line. Furthermore, concurrent PTSD treatment would have either been controlled in the study design or systematically measured, reported, and statistically adjusted for in the analyses. The latter was possible but not done. Accordingly, it is not possible to conclude that the intervention influenced PTSD symptoms if ongoing PTSD treatment occurred during the intervention (especially, for patients receiving evidence-

based interventions), and it was not measured. Throughout the discussion, it should be clear that the trial focused on improving disability functioning and quality of life.

Clarifying the Differences Between SERV and EMOT Intervention Groups

There were vastly different amounts of face-to-face instruction on dog handling and ownership given to Veterans in the two groups. Careful wording of the interpretation and noting of this limitation is needed. For instance, the draft monograph focuses the reader's attention on the highly trained SERV versus EMOT dimension, leading the reader to believe that (1) no other potentially significant differences existed in the treatments given the two different groups and (2) potential differences in outcome variables for Veterans in SERV versus EMOT groups would be due strictly to the psychological (for the Veterans) dimension of SERV-trained dog versus EMOT-trained dog. The draft monograph would benefit from revision designed to focus the reader's attention on the effect of all aspects of being in SERV versus EMOT intervention groups rather than the effect of living solely with a SERV versus an EMOT for 18 months.

Fidelity and Protocol Adherence

In many ways, the Contract of Statement of Work (SOW; provided to the committee as additional information) has a much better discussion of both the SERV and EMOT interventions than the draft monograph. The final monograph should have a much more detailed description for both interventions, potentially using content directly from the SOW. Although both descriptions need more detail, attention should be paid to enhancing the description of training for EMOTs as the control condition. The draft monograph notes that dogs in both groups had basic obedience training (line 1625) and both had to pass the American Kennel Club (AKC) Canine Good Citizen test, though it does not elaborate on the specifics of the test. As written, the EMOT intervention description leaves the naïve reader questioning what skills these dogs were trained in and how well they performed them. More specific information needs to be presented in the manuscript about markers of intervention fidelity for both SERV and EMOT interventions over the course of the trial. The authors are encouraged to think strategically about the key components of fidelity, data they potentially have that addresses fidelity, and report related analyses when describing the interventions. The importance of this information is well recognized by the authors, yet, the draft monograph does not present the analysis of these data.

Symptom Worsening, Avoidance Symptoms, and Safety Behaviors

There is an extensive discussion regarding symptom worsening, avoidance symptoms, and safety behaviors. This should be revised to reflect indices measured and reported in the trial, results analyzed and presented, and more careful theoretical and empirical understanding of fear conditioning and avoidance under the consultation and editing of a cognitive behaviorally oriented clinical psychologist. This likely means substantial cutting of this discussion section or additional post hoc analyses added to the manuscript (with the appropriate acknowledgment within the text that these were post hoc).

Addressing the Clarity and Consistency of the Use of Construct Terms

The committee noted challenges with accuracy, congruency, consistency, and reliability of the major concepts of interest, measurement tools, and outcome variables in the study. Throughout the draft monograph, there is a need for greater consistency with terminology (e.g., disability functioning, quality of life, depressive symptoms, suicidal ideation and behavior, and anger symptoms). Overall, inconsistent and incongruent terminology for key study constructs, as well as areas of insufficient clarity regarding reliability and patient-illness specificity, can result in confusion and inaccurate interpretation of study findings.

ECONOMIC ANALYSIS

The committee raised concerns that partial reporting of the economic analyses in this monograph could be misleading. Thus, the committee recommends that the investigators either (1) revise this draft monograph to include a comprehensive account of all economic outcomes, analyses, and results or (2) include all the economic outcomes, analyses, and results in the second planned monograph. The goal of whatever approach the authors choose is to avoid an incomplete monograph draft and potentially incorrect interpretation.

1

Introduction

Posttraumatic stress disorder (PTSD) is a leading cause of impairment in quality of life and functioning among Veterans. Service dogs[1] (SERVs) have been promoted as a potential intervention for Veterans with PTSD; however, research supporting their effectiveness is limited. The National Defense Authorization Act (NDAA)[2] directed the Department of Veterans Affairs (VA) to conduct a study to assess the potential therapeutic and economic benefit of using SERVs for the treatment or rehabilitation of Veterans with physical or mental injuries or disabilities. However, the study design described in the eventual monograph (reviewed by this committee) specifically compares the effectiveness of SERVs to emotional support dogs (EMOTs). The text within the NDAA of 2010, specifically defines physical or mental injuries to include PTSD.

> "The Secretary shall conduct a scientifically valid research study of the costs and benefits associated with the use of service dogs for the treatment or rehabilitation of Veterans with physical or mental injuries or disabilities. The matters studied shall include the following:
> (1) The therapeutic benefits to such Veterans, including the quality of life benefits reported by the Veterans partaking in the study.
> (2) The economic benefits of using service dogs for the treatment or rehabilitation of such Veterans, including—
> (A) savings on health care costs, including savings related to reductions in hospitalization and reductions in the use of prescription drugs; and
> (B) productivity and employment gains for the Veterans."

This same act mandated that the VA have the study reviewed by a National Academies of Sciences, Engineering, and Medicine committee. In response to these requests, in 2011 the VA began a 3-year longitudinal study to assess whether the provision of a SERV combined with usual care improved mental health in Veterans with PTSD. In this initial study the control group received no dog and usual care. This study encountered several challenges including difficultly in recruiting participants, as reported by the investigators.[3]

Following this effort, the VA revised the study plan and in 2014 a "longitudinal, randomized, intent-to-treat, two-arm, parallel design, multicenter clinical trial was conducted at three VA medical centers: Atlanta VA Healthcare System (Decatur, GA; Site 508), Iowa City VA Healthcare System (Iowa City, IA; Site 584) and the VA Portland Healthcare System (Portland, OR; Site 648)" (quoted from the

[1]Under the Americans with Disabilities Act (ADA), service animals are defined as dogs that are individually trained to do work or perform tasks for people with disabilities. Service dogs are working animals, not pets. The work or task a dog has been trained to provide must be directly related to the person's disability. Whereas, dogs whose sole function is to provide comfort or emotional support do not qualify as service animals under the ADA. These dogs are considered emotional support animals and they provide comfort to help relieve a symptom of a person's disability. They have not been trained to perform specific tasks related to the person's disability. In theory they provide a benefit by being present.

[2]National Defense Authorization Act for Fiscal Year 2010. 2009. Pub. L. No. 111-84, 123 Stat. 2190.

[3]VA presentation to the committee, April 2, 2020.

draft monograph). This revised plan intended to compare the effects of providing either a SERV or an EMOT. The authors expanded the study to include EMOTs because the primary question of interest was whether the benefits of SERVs go beyond the basic human-animal bond. In 2017, the VA published a paper describing the rationale and design of the trial (Saunders et al., 2017). In 2020, in accordance with the NDAA of 2010, the VA submitted a draft monograph reporting the results of the clinical trial titled "Performance and Results of Post Traumatic Stress Disorder—Service Dog Study" to the National Academies for independent review by an expert committee. The present report serves as the committee's review of the VA's draft monograph[4] addressing the completeness and accuracy of reporting, the rigorousness of the study design, conduct, and data analysis, and the scientific validity of the conclusions presented within the draft monograph (the full Statement of Task is provided in Box 1-1). It is important to note that this committee was tasked with reviewing only the specific trial and the outcomes presented in the VA's draft monograph and not with reviewing the use of SERVs writ large or the wider fields of PTSD and clinical trials. However, the committee does draw on its knowledge and expertise within those wider domains to inform its review.

BOX 1-1 Statement of Task

In response to a request from the U.S. Department of Veterans Affairs (VA), the National Academies of Sciences, Engineering, and Medicine will appoint an ad hoc committee to conduct a review of the Department of Veterans Affairs Monograph on Potential Therapeutic Effects of Service and Emotional Support Dogs on Veterans with Post-Traumatic Stress Disorder. The committee will prepare a consensus report that critiques the draft monograph and addresses the following questions:

- Are the research design and methods well documented, scientifically rigorous, and reasonable approaches to answer the research questions?
- Does the data analysis systematically apply appropriate statistical and sound reasoning techniques to evaluate the data on the therapeutic outcomes of service dogs and emotional support dogs for Veterans with PTSD?
- Do the findings thoroughly report the data analysis and provide factual and objective answers to the research questions?
- Do the findings present original scholarship and discuss principal outcomes of primary research with reliable credibility in a factual and objective way in relation to the research question and existing knowledge?
- Does the draft monograph provide a coherent and cohesive written account and description of the main messages that are important to communicate?
- Does the draft monograph provide clear, appropriate, and accurate graphics of the research results?
- What other significant improvements, if any, might be made in the draft monograph?

The consensus report will be subject to the National Academies' external peer-review process. Co-authors of the monograph will respond to the consensus report and submit a revised monograph for a second round of review and consensus reporting by the committee. If necessary, iterative cycles of response from authors to consensus reporting will continue until the committee determines that the report is consistent with accepted scientific principles and is suitable for publication. All subsequent rounds of consensus reporting will be subject to review by the National Academies.

At the conclusion of the project, a statement of completion of review will be provided to the VA, indicating that the final version of the monograph has been reviewed for consistency with accepted scientific principles and is suitable for publication that the VA can incorporate in the monograph when published. The National Academies review does not guarantee acceptance of the monograph for publication.

[4]This draft monograph is the main reference in this report. Any mention of line numbers throughout this report refers to the document received from the VA and provided to the committee for review. This draft and all subsequent drafts will be made available to the public upon completion of the committee's review.

Committee's Approach to the Task and Organization of the Report

To address the Statement of Task, the committee held a public meeting on April 2, 2020, to discuss the task with VA representatives. Following this, the committee held several virtual meetings to discuss the draft monograph, create a plan to address the task, and come to consensus. These discussions were informed by individual review of the draft monograph and all the associated documents provided by the VA for the committee to review (Appendix B has a list of all documents received), referencing the relevant literature, and discussing the standards of clinical trial design, conduct, analysis, and reporting relevant to the specifics of the VA's trial. To address whether the draft monograph was comprehensive, the committee assessed its adherence to the Consolidated Standards of Reporting Trials (CONSORT) guidelines (Grant et al., 2018; Moher et al., 2010), a set of reporting guidelines that describes the minimum information needed to understand how a trial was conducted and what was found. In addition to the CONSORT guidelines, the committee identified a number of issues related to the design, statistical analysis planned and conducted, and the interpretation of results.

Chapter 2 of this report provides a general assessment of the draft monograph and a high-level overview of issues identified by the committee. Chapter 3 discusses the ethical considerations outlined in the draft monograph, a clear strength of the document and the study.[5] In Chapter 4, the committee follows the CONSORT guidelines to review the material reported in the draft monograph. Chapter 5 discusses how the statistical methods are described and presented within the draft monograph and also the appropriateness of the methods for the study itself. Chapter 6 describes issues related to the ability to interpret the draft monograph and the study—covering limitations of the study that should qualify findings and conclusions from the study and how they are presented within the draft monograph. Finally, in Chapter 7, this report concludes with a review of the economic analysis presented in the draft monograph. Table 1-1 provides a mapping exercise of the committee's report and the Statement of Task.

TABLE 1-1 Mapping the Statement of Task to the Report Chapters

Element of Statement of Task	Report Chapter with Relevant Content
Are the research design and methods well documented, scientifically rigorous, and reasonable approaches to answer the research questions?	Chapters 4, 5, 6, and 7
Does the data analysis systematically apply appropriate statistical and sound reasoning techniques to evaluate the data on the therapeutic outcomes of service dogs and emotional support dogs for Veterans with PTSD?	Chapter 5
Do the findings thoroughly report the data analysis and provide factual and objective answers to the research questions?	Chapter 6
Do the findings present original scholarship and discuss principal outcomes of primary research with reliable credibility in a factual and objective way in relation to the research question and existing knowledge?	Chapter 6
Does the draft monograph provide a coherent and cohesive written account and description of the main messages that are important to communicate?	Chapter 4
Does the draft monograph provide clear, appropriate, and accurate graphics of the research results?	Chapters 4, 5, and 6
What other significant improvements, if any, might be made in the draft monograph?	Chapter 3 and 7

[5]For clarity, throughout this report, "the study" will always refer to the clinical trial conducted by the VA. Similarly, reference to "the authors" refers to the VA authors of the draft monograph.

REFERENCES

Grant, S., E. Mayo-Wilson, P. Montgomery, G. Macdonald, S. Michie, S. Hopewell, D. Moher, J. L. Aber, D. Altman, K. Bhui, A. Booth, D. Clark, P. Craig, M. Eisner, M. W. Fraser, F. Gardner, S. Grant, L. Hedges, S. Hollon, S. Hopewell, R. Kaplan, P. Kaufmann, S. Konstantopoulos, G. Macdonald, E. Mayo-Wilson, K. McLeroy, S. Michie, B. Mittman, D. Moher, P. Montgomery, A. Nezu, L. Sherman, E. Sonuga-Barke, J. Thomas, G. VandenBos, E. Waters, R. West, J. Yaffe, and on behalf of the CONSORT-SPI Group. 2018. CONSORT-SPI 2018 Explanation and Elaboration: Guidance for reporting social and psychological intervention trials. *Trials* 19(1):406.

Moher, D., S. Hopewell, K. F. Schulz, V. Montori, P. C. Gøtzsche, P. J. Devereaux, D. Elbourne, M. Egger, and D. G. Altman. 2010. CONSORT 2010 Explanation and Elaboration: Updated guidelines for reporting parallel group randomised trials. *BMJ* 340:c869.

Saunders, G. H., K. Biswas, T. Serpi, S. McGovern, S. Groer, E. M. Stock, K. M. Magruder, D. Storzbach, K. Skelton, T. Abrams, M. McCranie, J. Richerson, P. A. Dorn, G. D. Huang, and M. T. Fallon. 2017. Design and challenges for a randomized, multi-site clinical trial comparing the use of service dogs and emotional support dogs in veterans with post-traumatic stress disorder (PTSD). *Contemporary Clinical Trials* 62:105-113. doi: 10.1016/j.cct.2017.08.017.

VA (Department of Veterans Affairs). In Press. *Draft Monograph on Performance and Results of Post Traumatic Stress Disorder-Service Dog Study*. Washington, DC: Department of Veterans Affairs.

2

General Assessment

The review committee considered the strengths and weaknesses of the study and the draft monograph describing it in the following major categories:

- Exposition of the rationale for the study;
- Design of the study;
- Execution of the study;
- Analysis of the study data;
- Reporting of the study and its results;
- Interpretation of the results.

While the committee identified many elements of the report that it believes merit change, the vast majority of those elements are addressable with some rewriting, additional information, or additional statistical analyses. The committee recognizes that several of the issues identified with respect to the design of the study cannot be altered now that the study is complete. However, these can generally be conceptualized and addressed within the draft monograph as limitations rather than considered fatal flaws. That is, these design issues limit the conclusions that can be drawn from the study results, and may require an alteration in some of the interpretive statements in the current report, but they do not invalidate the study. Thus, on balance, the draft monograph describes a study of an important topic that appears to have been well executed; however, it is limited in the specific conclusions it can potentially support. The committee believes that with some non-trivial revisions the study results can be well analyzed, interpreted, and reported on within the draft monograph.

At the design level, the greatest limitation is the single control condition (provision of an emotional support dog [EMOT]) to which the treatment condition (provision of the service dog [SERV]) is compared. While this control condition does indeed control for being provided with a dog, it does not control for all other non-specific elements of the treatment condition. Additionally, by not including a no treatment control, the study is not capable of supporting conclusions of efficacy in any absolute sense, but only of differential efficacy, or lack thereof, between the two treatment conditions. Interpretation would have been more straightforward had the provision of a SERV produced clearly statistically significant results on major outcomes in terms of superiority to the control condition. Unfortunately, given the largely null results, it is difficult to determine whether the two treatments are equally effective, the two treatments are equally ineffective, or a type II error has been made (i.e., the null hypothesis was not rejected even though it is false). Again, this design choice is not a fatal flaw, but it does limit what can be concluded from the study. It will be important for the study authors to revise the reporting in the draft monograph to more accurately reflect the limitations of interpretation imposed by the control condition, as the current interpretation extends beyond what the study design can support.

Virtually all other elements raised in this report are either strengths (e.g., the treatment of both the human subjects and the canine participants appears to have been done in a manner commensurate with high ethical standards) or should, in principle, be fixable. Examples of fixable elements include the handling of analyses involving missing data, the need to include a true intent to treat analysis, more detail in much of the reporting to conform with the Consolidated Standards of Reporting Trials guidelines and other modern standards of reporting, avoidance of overreaching statements and removal of extraneous and

potentially inaccurate information from the introduction and discussion, and greater standardization of some language.

One particular example of a fixable element that necessitates early discussion for the sake of clarity relates to the use of the terms SERV and EMOT throughout the draft monograph. The draft monograph uses the acronyms SERV and EMOT to denote interventions that include either a SERV or an EMOT, respectively. However, at times these acronyms are used to strictly refer to the dog type, rather than the entire intervention. This can give the reader a misleading impression that the two treatment groups are exactly the same, but for this one variable. However, the groups differ in more ways (e.g., time spent with a trainer; breed, sex, and personality of the dog—all of which are elaborated on later in this report) and therefore the intervention is more than the dog but rather the entire set of variables specific to each group. The authors should clarify the use of the terms so that the reader is continually reminded to consider the totality of the interactions with the contractors (the non-governmental personnel who provided and trained the dogs and instructed the Veterans in dog handling) and with the researchers and VA staff (who administered evaluations) as the intervention and not just the specific role the dog was expected to play. For the purposes of this report, the committee uses the acronyms in conjunction with the terms "intervention group" to provide an example without introducing new terms that could cause confusion between this report and the draft monograph.

The committee looks forward to revision in which these important, but correctable, issues are addressed.

3

Ethical Considerations

Adhering to ethical guidelines is a critical component of research. A strength of the draft monograph was the manner in which the authors reported addressing ethical and animal welfare issues. This chapter reviews the discussion of the ethical measures taken during the study, detailed in the draft monograph, and, when needed, identifies areas where the draft monograph could be strengthened. It is worth noting that the committee did not review individual consent forms or the implementation of the study protocol.

HUMAN PARTICIPANTS

The researchers appear to have met reasonable ethical standards for the human aspects of the study. Institutional review board (IRB) approval was obtained through the VA Central IRB (protocol #13-54), Veterans provided their written informed consent to participate in the study, and the authors sought to maintain confidentiality. The privacy of participants in the study was protected in part by limiting the contractors' discretion with respect to what they could say to Veterans about removal of dogs during the study, or ask Veterans for the purpose of dog pairing. There were a number of provisions pertaining to the prevention of conflict of interest situations in which either the contractor or the Veteran could benefit from participation in the study. Protected health information and individually identifying information of Veterans were also safeguarded. The rights of Veterans, including the right to join or withdraw from the study at any time, were protected, and care was taken to ensure the procedures involving the dogs did not put the Veterans or family members at unreasonable risk of harm related to the disclosure of confidential information or injury from the dogs (e.g., bites). Procedures were included to address any problems by removing dogs from the home or providing additional training as needed.

CANINE PARTICIPANTS

The researchers also reported meeting the reasonable ethical standards for the dog-related aspects of the study. The Institutional Animal Care and Use Committee (IACUC) approvals were obtained (Atlanta protocol #V001-14 and Iowa City protocol #1490201, note that IACUC was considered unnecessary by the Portland VA IACUC) for animal welfare oversight. The purpose of the study to compare interventions involving the provision of service dogs (SERVs) and emotional support dogs was described clearly. The dogs used for the study were acquired by the VA in what appears to be a lawful and customary manner (the federal government is a large consumer of working dogs and procures hundreds every year). While VA policy states that the VA will never take responsibility for or possession of a SERV, an exception was made for this study given that it was mandated in Section 1077 of the National Defense Authorization Act of 2010. The Statement of Work (SOW) used to procure the dogs was clear and detailed both in terms of the factors used to evaluate individual dogs presented to the VA for purchase, and also the factors used to evaluate contractors providing the dogs.

Throughout the study, it appears that dogs were provided with humane care and healthful conditions. Dog welfare in Veteran homes was monitored through veterinary and home visits. Observations during the study suggested that a satisfactory amount of bonding between Veterans and dogs had occurred and that there were no signs that dogs were anxious or stressed. Dogs were

discontinued from the program if they had health or behavior problems or the Veteran was unable to care for them in an appropriate, safe manner. All Veterans in the study received education pertaining to the basic care of dogs.

Finally, disposition of the dogs after the study concluded was provided for by making the dogs the property of those Veterans who desired to keep their dogs. Those dogs that were not retained by Veterans were returned to the contractors for disposition, at which time the VA's responsibility for and liability related to the dog ceased.

CONCLUSIONS

The review committee identified three issues related to ethics that could improve the draft monograph: (1) there is a need for more information about the training that was provided to substitute caregivers of the dogs, (2) the VA might consider revisions to clarify what happened to the dogs that were not matches with any Veteran participating in the study (e.g., were they placed in homes by the contractors?) and (3) the language related to liability is not needed in the manuscript.

Overall, the authors appear to have followed the ethical guidelines that are commonly recommended for research and ensured that both human and animal participants were protected from undue risk and harm throughout the study.

4

Completeness of Reporting

The draft monograph includes some but not all of the minimum information needed to understand what was done and what was found. The Consolidated Standards of Reporting Trials (CONSORT) statement recommends the minimum information to include in a trial report, as described in the CONSORT 2010 Explanation and Elaboration (Moher et al., 2010), in the extension for social and psychological interventions (Grant et al., 2018), and the extension for harms (Ioannidis et al., 2004), which are relevant to this study. When reports of clinical trials do not describe the items in these guidelines, it may be difficult or impossible to assess risk of bias adequately, interpret study results, or apply the results to policy and practice decisions. The committee identified several areas in which the draft monograph does not adhere to CONSORT guidelines and suggests that the authors provide the missing information and utilize the CONSORT guidelines when revising the draft monograph.

The headings and subheadings of the rest of the chapter align with the elements of the CONSORT checklist and are followed by a discussion of the issue identified by the committee and suggested changes to improve the reporting completeness.

Title

1a. Identification as a randomized trial in the title.

The title of the draft monograph should identify the study as a randomized trial. A more descriptive title could also describe the population, the interventions compared, and the primary outcome. For example, "A randomized trial of differential effectiveness of service dog placement versus emotional support dog placement to improve quality of life for Veterans with PTSD."

Abstract and Executive Summary

The abstract and the executive summary do not include the minimum recommended information. For example, the trial registration number, financial support, the number of participants included in the analysis, or the magnitude and precision of the result for the primary outcome should be added. The abstract should include the information recommended in the CONSORT extension for social and psychological interventions (CONSORT-SPI; Grant et al., 2018) and in the below excerpted list from the CONSORT Explanation and Elaboration document (Hopewell et al., 2008):

- Description of the trial design (e.g., parallel, cluster, non-inferiority).
- Eligibility criteria for participants and the settings where the data were collected. When applicable, eligibility criteria for the setting of intervention delivery and the eligibility criteria for the persons who delivered the interventions.
- Interventions intended for each group.
- Specific objective or hypothesis. If pre-specified, how the intervention was hypothesized to work.
- Clearly defined primary outcome for this report.
- How participants were allocated to interventions.

- Who was aware of intervention assignment after allocation (e.g., participants, providers, those assessing outcomes), and how any masking was done.
- Number randomized to each group.
- Trial status.
- Extent to which interventions were actually delivered by providers and taken up by participants as planned.
- Number analyzed in each group.
- For the primary outcome, a result for each group and the estimated effect size and its precision.
- Important harms (adverse events or side effects).
- General interpretation of the results.
- Registration number and name of trial register.
- Source of funding.

The abstract and the executive summary emphasize "positive" results. The study was designed to test the superiority of the service dog (SERV) treatment group over the emotional support dog (EMOT) treatment group on the primary outcome using an intent to treat (ITT) analysis. The abstract and the executive summary focus on a secondary outcome, a per-protocol analysis (which does not properly account for missing data), and on the within-group comparisons rather than the between-group comparisons, which is also not appropriate. Instead, the abstract and the executive summary should provide a balanced interpretation of the results and give a more complete description of the outcomes and limitations. The conclusions in the abstract should focus on the pre-specified primary outcome and analysis (i.e., this study found no evidence of important differences in quality of life, and it found no consistent evidence of clinically and statistically important differences between SERV and EMOT treatment groups on secondary outcome measures). Because the study did not include a no treatment or usual care comparator, conclusions concerning the effectiveness or ineffectiveness of both interventions should be revised. The authors could simply remove these conclusions, or they could add that such conclusions are based on post hoc, exploratory analyses and show association, not causation (see Table 2 in Campbell and Stanley, 1959). That is, one cannot conclude from the within-group change that either intervention had an "impact" beyond what might have happened in the absence of the interventions because it is possible that participants in both groups would have done better or worse without dog placements. This is discussed further in Chapter 6.

In the abstract and the executive summary, as in the rest of the draft monograph, the conclusions are not supported by the results. The conclusions overstate the importance of a single observed difference, on the self-report measure of posttraumatic stress disorder (PTSD), the clinical importance of which has not been justified. Issues with the interpretation of results are described further in Chapter 6 and changes should be incorporated in the abstract and the executive summary.

Although not described elsewhere in the draft monograph, the executive summary of the draft monograph alludes to a follow-up study on line 41 that would examine whether the widening trends for suicidal behavior and other mental health outcomes continue past the timeframe of the current trial. The authors should consider removal of text mentioning this potential future study unless they opt to add additional information regarding the potential trial in the body of the draft monograph or if the addition of the ITT (discussed later in this report) warrants further discussion. If the authors choose to propose to follow-up this cohort to evaluate secondary outcomes with "positive" results at the end of the trial, then the authors should also add that additional assessments and analyses limited to secondary outcomes with "positive" results at the end of the first phase of the trial could produce biased estimates. Extending the follow-up to include all of the pre-specified outcomes, and interpreting long-term results in light of the totality of the evidence while emphasizing the pre-specified primary outcome, would be less prone to bias. The statement on line 2414 of the draft monograph should also be edited to avoid implying that the

study team detected a significant reduction in suicidal ideation or behavior to use more conservative language that more accurately reflects the patterns observed.

Introduction

2a. Scientific background and explanation of rationale.

The scientific background provided in the introduction of the draft monograph should provide a clear rationale for the study and what was known prior to the start of the trial. One mechanism for accomplishing this is a systematic review of the literature prior to the start of a clinical trial. It is unclear if a systematic review was conducted prior to planning the trial; however, doing so could have provided an explanation for why certain studies are highlighted in the introduction text and others in Table D in the draft monograph. This could also explain how the trialists used previous evidence to arrive at their conclusions and study design (see, e.g., Robinson and Goodman, 2011). The authors should clarify whether or not a systematic review was conducted prior to this study, and if one was not done then it should be stated as a limitation within the draft monograph.

The introduction is overly long and contains unnecessary information, especially on pages 12-18. For example, information on experimental drugs that are not in routine use are not needed (moreover, d-cycloserine has normally been tested as an adjunct, not as a stand-alone treatment), and paragraphs about the domestication of wolves are not required to understand the methods and results of this trial. Thus, the introduction should be shortened by deleting several sections and revising substantially the remaining content. If mentioned at all, other treatments for PTSD need fuller discussion, particularly the unknown potential for physical and psychological harms. Multiple citations in the introduction misrepresent the studies cited, and the introduction emphasizes positive results from previous studies instead of providing a balanced description of what is known. Throughout the abstract and the introduction, prior evidence is generalized inappropriately.

The introduction should specifically cite and discuss the evidence and rationale for conducting a randomized superiority trial comparing the SERV intervention group with the EMOT intervention group. It would be appropriate to include qualifications regarding the similarity of study settings, populations, imprecision, limitations in study design, potential confounders, and risk of reporting bias. Also, it would be helpful to present quantitative estimates from the findings of key studies (rather than just a brief mention of the direction [i.e., positive or negative] of the association). Early in the introduction, the study by Magruder and Yeager is referenced (lines 351-354), along with point estimates to quantify effects; however, these are noted to be the odds, rather than the odds ratios, and the associated confidence intervals are not included to quantify variability. Furthermore, the draft monograph references Allen and Blascovich (1996), which appears to be relevant, but neither the research question nor the results are described in the draft monograph. A more detailed discussion of the methods, findings, and limitations of seminal studies involving SERVs and PTSD (or perhaps other mental health problems) such as O'Haire and Rodriquez (2018) would be helpful for the readers to understand prior to getting to the results of this trial. Because the general effects of dog ownership are shared in both the EMOT and SERV intervention groups, the inclusion of content in the introduction about the benefits of pet ownership does not demonstrate the need to conduct a trial comparing two types of dog placement. Furthermore, the assertion that pets cause "faster recovery of the cardiovascular and immune systems from stressful events" cites a study with a small sample of healthy participants who took a math test in their own homes, with or without a pet present. This basic behavioral research study does not inform clinical conclusions regarding the treatment of physical and mental health problems. Raina et al. (1999) conducted a study of healthy older adults; however, it does not have direct implications for the treatment of PTSD and it did not find a statistically significant association between pet ownership and psychological well-being as the draft monograph implies. Additionally, the causal claims about the health benefits of pet ownership exaggerate the strength of the evidence. For example, the association between pet ownership and mortality is an example of confounding used in epidemiology textbooks; people who own pets might be healthier than

people who do not own pets *before they own pets*, but the draft monograph does not consider the limitations of such studies, such as bias and (residual) confounding. Some claims about "impact" and "results" refer to cross-sectional studies that could not demonstrate temporal relationships (i.e., non-causal language concerning "associations" would be appropriate, with many qualifications). To support the assertion that the "physical presence of a dog helps fill the human need for attention and emotional intimacy," the draft monograph cites an essay rather than an empirical study. To support the assertion that dogs reduce stress, the draft monograph cites a qualitative study of 12 participants rather than a study that could support causal inferences. The study by Cornell and Brown (2011) is not included in the reference list, so the committee was unable to assess it. Rather than focus on the putative benefits of pet ownership, the introduction should set up the primary hypotheses regarding the potential effects of SERV placement compared with EMOT placement on quality of life. The authors could consider noting that there is some history of interest and inquiry into the use of animals in the military, as discussed in a special issue of the *Army Medical Department Journal* (Ritchie and Amaker, 2012), being careful to note that the articles it contains are not studies about their effectiveness.

The introduction should describe how the intervention might work. That is, the introduction should provide a conceptual or mechanistic framework illustrating or explicating how placing dogs with service training might improve outcomes compared with placing EMOTs. Discussion of the purported mechanism of action would strengthen the draft monograph. Specifically, the introduction could explain how SERV placement might improve quality of life more than EMOT placement using a logic model or clarify the salient differences between the two interventions, including a rationale for why specifically each SERV behavior (light, block, etc.) was chosen, as the committee notes was done for the sweep command on lines 2343-2345, indicating it was included at the request of a subset of recipients in the "pilot" (specifically, by women in the study with PTSD as a result of sexual trauma). For example, Valentine et al. (1993) was cited but that study was restricted to people who were deaf or hard of hearing, and the relevance of those study participants to the VA's study is not described. The citation is misleading because the use of SERVs to address physical needs is very different from their use for psychological needs. The fact that people with hearing loss feel safer in the community with a SERV does not mean that people with PTSD will feel safer in the community with a SERV. The presumed mechanisms of action are different. Throughout the report, statements about how SERVs could improve mobility and quality of life for this population are not adequately supported by the studies cited.

The introduction also does not describe the equipoise[1] required to ethically and scientifically conduct a randomized clinical trial (RCT). For example, some potential disadvantages of dog ownership are mentioned in the discussion section of the draft monograph but not the introduction.

Lastly, the goals, design, and results of the "pilot" study the VA conducted are not described in sufficient detail to interpret the findings. For example, it is unclear whether the study was designed as a "pilot" study or whether the study was designed with a different objective. It is unclear whether that study was randomized, and it is unclear whether the results demonstrate that participants would not have accepted randomization to SERV placement compared with no intervention. The number of participants enrolled and number or type of adverse outcomes per group are relevant to the draft monograph. The section of the draft monograph about lessons learned from the "pilot" study includes conclusions about the obligations of organizations and causal inferences that are not supported by data provided in the draft monograph or cited in other reports. The purpose, design, and results of that study should be described following relevant guidelines (see Eldridge et al., 2016a) or a more detailed report should be referenced. Notably, a pilot study should evaluate specific design issues for larger trials (see, e.g., Eldridge et al., 2016b), and it is unclear what issues the "pilot" study was meant to evaluate that were referred to in the draft monograph.

[1]The concept of equipoise is whether providers and participants in a trial have equivalent beliefs and feelings about the conditions to which the participants may be assigned such that any differences in beliefs and feelings do not provide an alternative explanation for the effects of treatment assignment.

2b. Specific objectives or hypotheses. If the trial addresses both harms and benefits, the introduction should so state.

Overall, the introduction implies that the study objectives were different from the planned study objectives. Specifically, the introduction implies that the study sought to evaluate the general effects of dog placement for people with PTSD. In fact, the study was designed to evaluate the specific effects of SERV placement *compared with* EMOT placement on quality of life. Notably, the pre-post difference within each group does not represent a causal effect in a randomized trial. As the Saunders et al. (2017) protocol says,

> While a standard-of-care control group may be scientifically justified, the control intervention selected was provision of an emotional support dog. This was chosen because the study aims to determine whether provision of a service dog, and the specific tasks it can perform, is beneficial to Veterans with PTSD. This is a significant challenge because it is not known whether and to what degree, the benefits of a service dog arise from factors other than performing the tasks it is trained to provide; the dynamics of a living animal need to be considered. Therefore, one necessary control involves the impacts of pet ownership, which as noted above, have been shown to enhance psychological and social well-being.

Therefore, the introduction should be revised to clarify that the study was designed to assess whether an intervention that included SERVs improved quality of life compared with an intervention that included EMOTs. The effect of interest was the between-group difference. Later, the draft monograph describes effects on safety behaviors and worsening; however, there are no outcomes or results presented in the draft monograph that support those conclusions. If that information were to be added, then the introduction should describe whether the study was designed to assess safety behaviors, worsening, or other harms.

Methods

3a. Description of trial design (such as parallel, factorial) including allocation ratio.

The draft monograph should completely state the trial design. The draft monograph refers to the Saunders et al. (2017) protocol and does not provide a full description of the trial study design. Readers might infer from the protocol that the trial was a superiority trial given the following passage,

> The primary aim of this RCT is to determine whether overall functioning and quality of life of Veterans with PTSD are improved by the provision of service dogs relative to provision of emotional support dogs. It was hypothesized that given the special training of service dogs to handle tasks that may benefit Veterans with PTSD, they would provide greater improvements than emotional support dogs. Secondary aims are to compare the impact of service and emotional support dogs on mental health outcomes, health care utilization and costs, and employment and productivity.

If this inference is correct, the draft monograph should identify the trial as a parallel superiority trial with a 1:1 allocation ratio.

3b. Important changes to methods after trial commencement (such as eligibility criteria), with reasons.

Additionally, the committee identified a need for clarity regarding the document referred to by the authors as the "protocol" document. Throughout, the draft monograph should indicate if the authors are referring to content from the study registration on ClinicalTrials.gov, the Saunders et al. (2017) protocol publication, or the original Statement of Work.

The draft monograph should describe all changes to the methods after trial commencement, as was done in line 1099 of the draft monograph. Notably, the inclusion/exclusion criteria in the ClinicalTrials.gov registration, protocol, and the draft monograph are discrepant. The draft monograph should include a section describing deviations from the trial registration and protocol, including why and when changes were made (e.g., what proportion of participants had been enrolled, what proportion had completed assessments).

The draft monograph suggests that participants were assigned to treatment before pairing with a dog for the purpose of using multiple assessments in the analysis, which is not the reason stated in the protocol. Randomization long before receiving the intervention would be expected to increase post-randomization drop out (e.g., participants would normally be assigned after a run-in period, not before a run-in[2]). Moreover, the decision for "each person to be their own control" is (1) inconsistent with the primary goal of a parallel trial (the pre-post difference within each group does not represent the causal effect in a randomized trial) and (2) post hoc, which could be potentially influenced by the observed results. The protocol describes the time between randomization and receiving the intervention as a pragmatic issue, "Following the dog-matching interview the vendor [contractor] (only) is unblinded to the type of dog the participant is to receive, so that selection and training of the dog can begin." The draft monograph should state the a priori rationale for the design, including the description and analysis of this period (e.g., line 844). The authors should identify any other reasons or use of this period as exploratory and post hoc.

4a. Eligibility criteria for participants. When applicable, eligibility criteria for settings and those delivering the interventions.

Inclusion and exclusion criteria were specified in the draft monograph, though the criteria do not entirely match with those specified on ClinicalTrials.gov or in the Saunders et al. (2017) paper. For the primary outcome measures, no eligibility score cut-offs were used, failing to ensure that there was pre-intervention impairment and room for improvement on mobility and quality of life. Similarly, if the trial was intended to be a PTSD treatment trial, a cut-off score for PTSD severity would likely have been specified beyond the presence of diagnosis to also reduce floor effects. This is discussed in more detail in Chapter 6.

4b. Settings and locations where the data were collected.

Information is provided on contractors and sites; however, information is needed on the standardization procedures and reliability assessments within and across sites in terms of dog training objectives and training actually received, as well as in terms of training of the interviewers using interviewer-based measures (e.g., Clinician Administered PTSD Scale for DSM-5 [CAPS-5], Columbia-Suicide Severity Rating Scale [C-SSRS]). This is discussed in more detail in Chapter 6.

[2]The committee notes that the period of time between randomization and receipt of the intervention is neither a true "run-in" nor "wash-out" phase, but these terms come close to approximating this design element to other clinical trials.

Participants were all receiving VA services prior to and during the study, which could influence their outcomes before and after dog pairing, but concurrent mental health services are not described in the draft monograph. It should describe the context in which these interventions took place, including the prevalence of specific concurrent intervention in each of the randomized groups.

5. The interventions for each group with sufficient details to allow replication, including how and when they were actually administered.

The committee was concerned about how the two treatment groups might have differed in ways other than having a trained SERV with specific PTSD-relevant skills compared to having a trained dog without these PTSD-relevant skills. If the dogs in one treatment group were on average more difficult to train or live with than dogs in the other treatment group, then this difference might have mediated the effect of the treatment, leading to an inability to properly interpret the effect of the treatment group. For instance, based on an analysis of rankings by veterinarians, Stafford (1996) found that German Shepherds were perceived to be more aggressive than Labrador and Golden Retrievers (Stafford, 1996; see also Van den Berg et al., 2010). In a study examining the validity of the stranger-directed aggression subscale of the CBARQ™ (Canine Behavioral Assessment and Research Questionnaire), German Shepherds, Labrador Retrievers, and Golden Retrievers exhibited different levels of stranger-directed aggression on three of the ten subscale items (van den Berg et al., 2010). In another survey study, male dogs differed from female dogs on every one of 10 behavioral characteristics, including aggression dominance, territorial aggression, aggression toward other dogs, trainability, and affection demand (Hart and Hart, 2005; see also Hart and Serpell, 1995). Therefore, if male dogs, or a particular breed, predominated in one treatment group, then this systematic difference could have mediated the effects of the SERV versus the EMOT intervention and compromise interpretation.

The manner in which individual male and female Labrador Retrievers, Golden Retrievers, German Shepherds, and Labrador-Golden crosses were assigned to treatment groups is not entirely clear in the draft monograph, and it is not evident that precautions were taken to ensure that the dogs did not vary systematically across the two groups in terms of sex, breed, or other characteristics aside from the type of training they received and the amount and manner of instruction Veterans received (up to 2 days of instruction on dog handling at their homes for Veterans in the EMOT intervention group versus up to 2 weeks of instruction at the vendor site for Veterans in the SERV intervention group). Although the authors say that they strove to make the contribution of dogs from each of three contractors roughly equal at the three test sites (lines 1319-1322), the exact distribution of dogs by sex, breed, site, and supplying contractor in EMOT versus SERV groups is not presented. Therefore, a table containing these data should be included in the draft monograph, and treated in the discussion. These changes would substantially improve the discussion, address the concerns of readers attentive to such issues, and inform future research that may use similar manipulations and design.

5a. Extent to which interventions were actually delivered by providers and taken up by participants as planned.

This is discussed in more detail in Chapter 6 in a section discussing issues with fidelity and protocol adherence.

5b. Where other informational materials about delivering the intervention can be accessed.

The draft monograph does not describe whether and how materials needed to deliver the interventions can be accessed. The draft monograph should state whether the research materials (e.g., training manuals, questionnaires), data, and statistical code used to produce the results are available and how they can be accessed (Taichman et al., 2017). Ideally, these items would be provided along with ancillary materials (e.g., protocol) on a public repository at or before the time of the report's publication.

Consistent with best practices for conducting rigorous and reproducible science (IOM, 2015; NASEM, 2019; NIH, 2020), the investigators could post these items in an open repository.[3,4] Alternatively, the authors could add these materials to the repository provided to the committee with review material and include a link to this in the draft monograph.[5]

If the investigators cannot or choose not to share research materials, data, and code, then the draft monograph should include a section explaining why each item is unavailable. For example, software or training materials might be copyrighted by other organizations; however, there may be no legal or ethical obstacles to sharing statistical code written by the investigators. While it might not be possible to share identifiable participant data, it might be possible to share de-identified data (e.g., using perturbation or other methods).

5c. When applicable, how intervention providers were assigned to each group.

The draft monograph should describe how dogs were selected to be SERVs or EMOTs. Each participant is assigned a unique dog, with dogs differing not only on whether they have been trained as a SERV versus an EMOT, but also on other characteristics such as breed and sex, discussed earlier in this report, and personality traits. If the contractors were given the procedural latitude to assign dogs perceived to be more easily trainable to the SERV intervention group, and comparatively less trainable animals to the EMOT intervention group, Veterans would have been assigned to dogs differing across treatment groups in their training, and to dogs that also systematically vary in other characteristics. Thus, any effect of being assigned to receive a SERV or an EMOT could arise because of dog-related factors other than the dog's training when thinking broadly about the entire package as the intervention. This is akin to the concept of packet randomization (Pavela et al., 2015). Explaining the process by which dogs were selected for each of the treatment groups merits discussion within the text of the draft monograph.

6a. Completely defined pre-specified outcomes, including how and when they were assessed.

The draft monograph should completely define the study outcomes using five elements (domain, specific measure, metric, method of aggregation, time point) (Mayo-Wilson et al., 2017; Zarin et al., 2017). That is, the draft monograph lists the names of measures, but an outcome has other elements. For each measure, was the total used or were subscales analyzed separately? Did the investigators plan to compare the value at the end or change in value? Were they assessed categorically or continuously? Why were data collected at multiple time points—were all time points analyzed?

6b. Any changes to trial outcomes after the trial commenced, with reasons.

The draft monograph does not describe changes to the outcomes after trial commencement. Several changes are evident from reviewing the registration on ClinicalTrials.gov, in the Saunders et al. (2017) protocol publication, and in the original Statement of Work. For example, the time points do not match, the number of outcomes differ because some subscales were used in the final analysis rather than the total scores, the prioritization of outcomes are not the same, an unregistered measure appears in the final report, and the draft monograph introduces a category of "tertiary" outcomes that does not appear in the registration or protocol. There are inconsistencies related to how the primary, secondary, and tertiary measures are described and how they will be analyzed, and at which time points the primary outcome measures were evaluated. For example, suicidal ideation (C-SSRS) was not specified as a secondary outcome in the protocol and anger (dimensions of anger reactions) was omitted from the

[3]See https://vivli.org (accessed July 1, 2020).
[4]See https://osf.io (accessed July 1, 2020).
[5]See https://www.research.va.gov/programs/animal_research/ptsdstudy.cfm (accessed July 1, 2020).

ClinicalTrials.gov registration. CAPS-5 is included in the secondary outcome section, though not specified as a secondary outcome on ClinicalTrials.gov or in the protocol. Also, "healthcare utilization and costs from VA administrative datasets" is not included in Saunders et al. (2017) or the draft monograph. In addition, pre-specified levels of reliable change or clinically meaningful change were neither present in the Saunders et al. (2017) protocol nor the ClinicalTrials.gov registration. Harms outcomes mentioned in the draft monograph are not described in the registration and protocol.

The draft monograph should describe all outcomes that were assessed, including all pre-specified outcomes and any other outcomes that were not registered or reported in the published protocol. It should also state which outcomes were added or modified after the trial commenced and all changes to the categorization/importance of outcomes (e.g., downgrading from primary to secondary).

Lastly, methods for missing data in the protocol and the methods section are not consistent with the main analysis, which is a per-protocol analysis. The published trial protocol says that "Analyses of all outcome measures will use an intent-to-treat (ITT) population as well as a per-protocol population (PP) which is defined as the population of participants who are paired with a dog using their initial randomization assignment." The results in the draft monograph focus on the PP rather than the ITT population, and the draft monograph draws conclusions that do not account for uncertainty due to missing data. The committee's concerns about changes to the planned methods for handling missing data are addressed in more detail in Chapter 5.

6. Harms: Clarify how harms-related information was collected (mode of data collection, timing, attribution methods, intensity of ascertainment, and harms-related monitoring and stopping rules, if pertinent).

Methods to assess and analyze harms are not reported clearly. Moreover, methods to assess harms reference Medical Dictionary for Regulatory Activities (MedDRA) but are not consistent with the International Council for Harmonisation (ICH) of Technical Requirements for Pharmaceuticals for Human Use and regulatory guidelines and some classifications are confusing. For example, why was "anxiety" classified as a serious adverse outcome (e.g., did it require hospitalization)?

The draft monograph should state how and when non-systematic harms were collected, and how non-systematic harms were coded and by whom. For example, did the investigators ask about specific harms at certain visits or all visits? Were there open-ended questions as well as specific prompts? If assessed systematically, the draft monograph should state how worsening and safety behaviors were assessed.

7a. How sample size was determined.

The draft monograph should provide more information concerning the sample size justification. Further justification is needed of anticipated effect sizes or effect sizes of interest, the effect contrasts of primary interest (i.e., the difference between randomized groups at 18 months), or the assumed correlation across time points if utilized. There is a discrepancy between different iterations of the protocol[6] (primary outcomes) and draft monograph (all measures) concerning which measures were used for power analysis. This concern is covered in more detail in Chapter 5.

7b. When applicable, explanation of any interim analyses and stopping guidelines.

The draft monograph should describe interim analyses or stopping guidelines. If there were interim analyses or stopping guidelines, the draft monograph should describe which analyses were

[6]Meaning the registration on ClinicalTrials.gov, the Saunders et al. (2017) protocol publication, or the original Statement of Work.

conducted and when, as well as state whether any methods were used to account for multiple testing, and if so, which. If there were no interim analyses or stopping guidelines, the manuscript should state this.

8a. Method used to generate the random allocation sequence.

Randomization was generated through a computer program, with the assignment of an intervention group made separately (stratified) by site and balanced, with the same number of individuals set to receive a SERV or an EMOT, for a set number of allocations (blocked). The committee did not identify deficiencies in relation to this topic.

8b. Type of randomization and details of any restriction (such as blocking and block size).

As described in the draft monograph and study protocol paper, participants were recruited, assessed for eligibility and then randomized. The study design protocol utilized several key steps to ensure efficient and random allocation to each treatment group. For instance, randomization was performed centrally through a computer telephone randomization system. The randomization scheme was stratified by site, and developed via computer program. Clarification is needed on whether the randomization scheme had varied, permuted, block sizes.

9. Mechanism used to implement the random allocation sequence, describing any steps taken to conceal the sequence until interventions were assigned.

The allocation was concealed from the local study team and the participant, until the participant was paired with a dog. The randomization allocation was implemented through a telephone randomization system. The committee did not identify deficiencies in relation to this topic.

10. Who generated the random allocation sequence, who enrolled participants, and who assigned participants to interventions.

Clarification on who developed the study random allocation scheme and who enrolled the participants is needed. The protocol implies that the study team enrolled participants and facilitated the random allocation via this telephone system; however, clarification is needed as to who developed the scheme and how this remained concealed before dog pairing.

11a. Who was aware of intervention assignment after allocation (e.g., participants, providers, those assessing outcomes), and how any masking was done.

Participants and providers were aware of group assignment; however, the draft monograph does not state whether assessors were masked[7] to group assignment, or how masking was accomplished, if done. Assessor-rated outcomes include C-SSRS, CAPS-5, dog-related measures, and health care utilization information. The draft monograph also has some confusing language explaining the timing of masking the assignment group and the timing of randomization that requires clarification. For example, line 58 states there was a minimum 3-month observation period and then randomized (meaning the assignment group would be unknown during the observation period); whereas line 954 states after randomization, the observation period occurred and after which the condition was revealed to a participant and line 1325 forward is obscure as to when the type of dog assigned to a participant was revealed: during end observation period, at clinic clearing visit, or at home clearing visit.

[7]For the purposes of this report, the committee will use the term "masked" rather than "blinded" to refer to the concealment of information for the sake of conducting the trial.

11b. If relevant, description of the similarity of interventions.

[Intentionally left blank – no committee comments]

12a. Statistical methods used to compare group outcomes. How missing data were handled, with details of any imputation method.

More information is needed in the draft monograph concerning the statistical methods. For example, there is insufficient information on the handling of the time variable, as time varies widely between assessment points, baseline, randomization, and pairing across participants. Methods for additional analyses, such as subgroup analyses and tertiary or exploratory analyses, including health care utilization analyses, are also not described. Chapter 5 addresses both the reporting and the conduct of the statistical methods in greater detail.

12b. Methods for additional analyses, such as subgroup analyses, adjusted analyses, and process evaluations.

Chapter 5 provides a detailed discussion of the methods of analysis.

12. Harms: Describe plans for presenting and analyzing information on harms (including coding, handling recurrent events, specification of timing issues, handling of continuous measures, and any statistical analyses).

The draft monograph does not include sufficient information about the analysis of potential harms, or concerning how potential harms were selected for inclusion. The draft monograph should include a comprehensive account of how potential harms were assessed, analyzed, and how they were selected for inclusion. Were any potential harms assessed systematically (i.e., in the same manner for all participants)? How were non-systematically assessed harms collected? What information was collected and analyzed about each potential harm? Were rates or hazards calculated, or only cumulative incidence? Did the coder(s) have any training or experience using MedDRA or other systems for coding analyzing harms? Were harms analyzed at higher levels above the preferred terms? Did other harms occur that are not mentioned in the draft monograph—if so, how were the harms in the draft monograph selected for inclusion?

Results

13a. For each group, the numbers randomly assigned, receiving the intended intervention, and analyzed for the outcomes. Where possible, the number approached, screened, and eligible prior to random assignment, with reasons for non-enrollment.

The flowchart in the draft monograph (see Figure B) is missing information and should be replaced with a CONSORT-SPI flow diagram (Grant et al., 2018). Standard formatting should be included, such as numbers approached, screened, randomized, offered the intervention, receiving the intervention, assessed at each time point, and included in the analysis, along with reasons for exclusion at each step.

13b. For each group, losses and exclusions after randomization, together with reasons.

The draft monograph includes incorrect statements about the effects of missing data. Specifically, post-randomization but pre-pairing withdrawals might be unrelated to the effects of the interventions per se, as the authors note, but these withdrawals could be related to participant characteristics and

expectations, and they could introduce selection bias. The authors should include text to assess the missingness mechanisms and assessment of assumptions made in each analysis (missing completely at random, missing at random, missing not at random) and the associated limitations on interpretation. The committee suggests that the authors choose an approach best suited for observed missingness patterns. For instance, a full description of the reasons given for why participants discontinue participation, if these participants had more severe outcomes, and what their other treatments during this time may have been would be informative. All statements that missing data would not bias the effect estimate should be removed or labeled as speculation. This topic is explored in more detail in Chapter 5.

13. Harms: Describe for each arm the participant withdrawals that are due to harms and their experiences with the allocated treatment.

The draft monograph should indicate the number and timing of withdrawals because of harms.

14a. Dates defining the periods of recruitment and follow-up.

Dates of recruitment are not stated. The specific months (and years) of recruitment should be stated.

14b. Why the trial ended or was stopped.

One might assume that the trial ended because it met the planned recruitment goals. However, the specific reason for ending the trial should be stated.

15. A table showing baseline characteristics for each group. Include socioeconomic variables where applicable.

Baseline characteristics are not reported for the randomized groups. Table N in the draft monograph includes only those participants who were paired with a dog. No description of the full randomized cohort is provided. Table N also includes categories of sparse data; consideration should be given to collapsing some of these for wider distribution out of concern for potentially and inadvertently identifying participants. Table P in the draft monograph (incidence of *Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition* [DSM-5] psychiatric diagnosis at screening) is difficult to read and could be simplified, for example, by removing the two redundant columns of "does not meet" and collapsing rows into larger overarching categories (e.g., any anxiety disorder). Tables N and P should be organized and revised to improve readability.

The table of baseline characteristics in the draft monograph focuses on statistical significance rather than the magnitude and importance of differences between groups. Although investigators commonly test for statistically significant differences in baseline values, CONSORT guidelines advise against doing so. They state "Unfortunately significance tests of baseline differences are still common…. Tests of baseline differences are not necessarily wrong, just illogical under the premise that one has truly randomly assigned subjects to treatments. Such hypothesis testing is superfluous and can mislead investigators and their readers" (Moher et al., 2010).

Instead, potential differences in confounders at baseline should be described without respect to statistical significance. P-values in Table N in the draft monograph should be deleted, and conclusions concerning different p-values at baseline should be removed as they are not meaningful and could be misleading (Harvey, 2018; Moher et al., 2010). Missing in Table N are percentages of DSM-5 Criterion A trauma types and concomitant therapies for PTSD at the time of study. Baseline characteristics should be reported for both the randomized population (ITT) and the PP population.

In general, greater attention could be paid to ensure that the numbers included in the tables match the numbers in the text, the usage of symbols (such as an asterisk) to denote significance should be consistent across the manuscript, and the manuscript should be checked for consistency of all numbers (e.g., line 1834 attempts to provide a breakdown of the number of Veterans in the EMOT and SERV intervention groups among a group of 10 participants, but it refers to only 5 in the EMOT group and 3 in the SERV group).

16. For each group, number included in each analysis and whether the analysis was by original assigned groups.

Although it seems that the number of participants included in each analysis probably differs, it is unclear how many participants are included in each analysis. This should be reported for each result (e.g., in the column or row of a table, as appropriate) in addition to the flowchart (see Figure B in the draft monograph).

16. Harms: Provide the denominators for analyses on harms.

The draft monograph includes no data to support conclusions about potential harms, such as "the study results found no evidence that either dog type worsened functioning or impeded extinction of fear conditioning by acting as a safety behavior." If this was assessed during the trial, then numerical results (including mean differences or relative risks and confidence intervals and the number of participants analyzed, proportions with numerators and denominators, etc.) should be reported for harms outcomes.

17. For each outcome, results for each group, and the estimated effect size and its precision (such as 95% confidence interval). Indicate availability of trial data.

Rather than emphasizing the statistical significance of effects, the draft monograph should emphasize magnitude and precision of effects, which it currently does so in only a few cases. The ITT results estimated in accordance with the ITT principle for primary or secondary outcomes should be presented in the draft monograph. Results for several pre-specified outcomes are not reported in the manuscript, or are not reported fully (e.g., health care utilization, costs, employment, productivity).

All pre-specified outcomes should be identified in the draft monograph and the methods for analysis and the corresponding results should be reported following CONSORT guidelines. If the authors decide not to include results for selected pre-specified outcomes in this monograph (e.g., because those results will be included in a second monograph), then this monograph should state that the results for those outcomes will be included in a second monograph. The current draft includes incomplete reporting of relevant methods and results for planned outcomes, which is undesirable.

As described in item 5b, the draft monograph should describe how readers can access the data and code, or at a minimum, include a data availability statement.

17b. For binary outcomes, presentation of both absolute and relative effect sizes is recommended.

The draft monograph should include the relative and absolute differences between groups, not only the within-group changes.

18. Results of any other analyses performed, including subgroup analyses, adjusted analyses, and process evaluations, distinguishing pre-specified from exploratory.

Additional analyses are addressed in more detail in Chapter 5.

19. All important harms or unintended effects in each group (for specific guidance, see CONSORT for Harms [Ioannidis et al., 2004]).

As described above, it is unclear whether and how potential harms were assessed. The draft monograph should include numerical results related to claims about harms, which are currently not reported.

Discussion

20. Trial limitations, addressing sources of potential bias, imprecision, and, if relevant, multiplicity of analyses.

Chapter 6 provides a thorough discussion of the trial's limitations. That chapter covers issues related to the study design, conduct, reporting, and the discussion of the study results that impact the interpretation of the study and how they should be addressed within the draft monograph as limitations of the trial. Examples of the topics discussed within the chapter include within-group change, claims about equivalence of the interventions, limitations related to the period of time between randomization and receiving the intervention, lack of equipoise in the design, as well as other topics that will improve the clarity of the draft monograph.

20. Harms: Provide a balanced discussion of benefits and harms, with emphasis on study limitations, generalizability, and other sources of information on harms.

The discussion about safety behaviors is misleading. As noted elsewhere, no data are presented in the draft monograph that would confirm or disconfirm the hypothesis that SERVs reinforce safety behaviors. Moreover, the draft monograph concentrates on a narrowly focused cognitive model of PTSD that is centered on fear conditioning and extinction, and the description of related neuroscientific issues is both scientifically questionable and irrelevant to the goals and results of the clinical trial.

21. Generalizability (external validity, applicability) of the trial findings.

Generalizability of the findings is limited by the lack of intervention fidelity reporting in the draft monograph; not allowing for understanding of the necessity of specific training parameters of the SERVs or EMOTs for comparison to other trials or future implementation. This is discussed in more detail in Chapter 6.

The trial was also not set up to be a PTSD clinical trial, where independent evaluators would assess interviewer-rated PTSD severity at multiple critical time points, including immediately prior to randomization, with the interviews recorded and checked for reliability over time and across sites. Importantly, the impact of VA or outside ongoing PTSD treatment would either have been mitigated by inclusion/exclusion criteria (e.g., stabilization of psychotropic medications, non-trauma-focused psychotherapy only allowed, or no ongoing trauma-focused cognitive behavioral therapy for a discrete period) or utilization carefully assessed and incorporated into the analytic plan. As analyzed, the other ongoing interventions cannot be ruled out as the potential cause of any observed PTSD symptom change. This is discussed in more detail in Chapter 6. These factors limit generalizability to the broader PTSD clinical trials literature and limit the inferences that can be made specifically about the SERV and EMOT interventions.

22. Interpretation consistent with results, balancing benefits and harms, and considering other relevant evidence.

The committee identified several issues relating to the interpretation of the results of the draft monograph that need to be addressed. This is discussed in more detail in the subsequent chapters of the report in order to provide sufficient information to the authors for revision. In short, the conclusions are not sufficiently supported by the evidence presented in the draft monograph because they do not address important limitations of the study. In particular, the authors should consider the following topics (as elaborated on in the subsequent chapters of this report): missing data and the differences in missing data between groups; within-group change is not evidence of effectiveness; no evidence of equivalence; interpreting the results with respect to their magnitude and precision, emphasizing the pre-specified primary outcome; focus on the planned analysis and time points; and consideration of clinical importance.

Important Information

23. Registration number and name of trial registry.

The draft monograph includes a trial registration number, but the trial registration is not up to date. The ClinicalTrials.gov record should have been updated by June 2020 (i.e., 1 year after the primary completion date) to include the protocol and the results for all primary and secondary outcomes.

24. Where the full trial protocol can be accessed, if available.

The draft monograph references a published protocol, but the protocol is not presented in detail because it does not include a statistical analysis plan, though one is referenced. Differences between the protocol and the draft monograph are partly attributable to inadequate reporting in the protocol (i.e., outcomes were not defined following Standard Protocol Items: Recommendations for Interventional Trials [SPIRIT] guidelines [SPIRIT, 2020]). The draft monograph should state that the protocol is available but that it does not include complete outcome definitions or methods (e.g., for handling missing data). The statistical analysis plan was not pre-specified, and the draft monograph should describe it as a limitation that the methods for analysis were not described a priori (e.g., following relevant guidance [Gamble et al., 2017]).

25. Sources of funding and other support, role of funders. Declaration of any other potential interests.

Role of funder could be further specified, particularly in regard to any reporting requirements and oversight prior to scientific peer review.

26a. Any involvement of the intervention developer in the design, conduct, analysis, or reporting of the trial.

[Intentionally left blank – no committee comments]

26b. Other stakeholder involvement in trial design, conduct, or analyses

[Intentionally left blank – no committee comments]

26c. Incentives offered as part of the trial.

[Intentionally left blank – no committee comments]

CONCLUSIONS

The authors should ensure that the draft monograph adheres to relevant reporting standards. Many academic journals require that authors submit completed CONSORT checklists with trial reports for publication. The discussion in this chapter was structured to identify aspects of the draft monograph that do not adhere to CONSORT 2010 guidelines, and relevant extensions, with suggestions for addressing these issues. The committee will review the revised monograph (iteration 2) for completeness by identifying where the revised monograph does and does not adhere to the CONSORT recommendations. Therefore, the committee would recommend that the authors complete a CONSORT-SPI checklist (Grant et al., 2018) with the revised monograph. Carrying out this exercise will better enable the authors to review their own draft for completeness and while the authors do not need to make their completed checklist public, journals require this for submission with clinical trial reports.

REFERENCES

Allen, K., and J. Blascovich. 1996. The value of service dogs for people with severe ambulatory disabilities: A randomized controlled trial. *JAMA* 275:1001-1006.

Campbell, D. T., and J. C. Stanley. 1959. *Experimental and Quasi-Experimental Designs for Research*. Ravenio Books.

Eldridge, S. M., C. L. Chan, M. J. Campbell, C. M. Bond, S. Hopewell, L. Thabane, and G. A. Lancaster. 2016a. CONSORT 2010 statement: Extension to randomised pilot and feasibility trials. *BMJ* 355:i5239.

Eldridge, S. M., G. A. Lancaster, M. J. Campbell, L. Thabane, S. Hopewell, C. L. Coleman, and C. M. Bond. 2016b. Defining feasibility and pilot studies in preparation for randomised controlled trials: Development of a conceptual framework. *PLoS One* 11(3):e0150205.

Gamble, C., A. Krishan, D. Stocken, S. Lewis, E. Juszczak, C. Doré, P. R. Williamson, D. G. Altman, A. Montgomery, P. Lim, J. Berlin, S. Senn, S. Day, Y. Barbachano, and E. Loder. 2017. Guidelines for the content of statistical analysis plans in clinical trials. *JAMA* 318(23):2337-2343.

Grant, S., E. Mayo-Wilson, P. Montgomery, G. Macdonald, S. Michie, S. Hopewell, D. Moher, J. L. Aber, D. Altman, K. Bhui, A. Booth, D. Clark, P. Craig, M. Eisner, M. W. Fraser, F. Gardner, S. Grant, L. Hedges, S. Hollon, S. Hopewell, R. Kaplan, P. Kaufmann, S. Konstantopoulos, G. Macdonald, E. Mayo-Wilson, K. McLeroy, S. Michie, B. Mittman, D. Moher, P. Montgomery, A. Nezu, L. Sherman, E. Sonuga-Barke, J. Thomas, G. VandenBos, E. Waters, R. West, J. Yaffe, and on behalf of the CONSORT-SPI Group. 2018. CONSORT-SPI 2018 Explanation and Elaboration: Guidance for reporting social and psychological intervention trials. *Trials* 19(1):406.

Hart, B. L., and L. Hart. 2005. Breed-specific profiles of Canine (Canis familiaris) behavior. In *Current Issues and Research in Veterinary Behavioral Medicine*, edited by D. Mills et al. Papers Presented at the 5th International Veterinary Behavior Meeting, West Lafayette, IN: Purdue University Press. Pp. 107-113.

Hart, B. L., and J. A. Serpell. 1995. Analyzing breed and gender differences in behaviour. In *The Domestic Dog: Its Evolution Behaviour and Interactions with People*. Cambridge, UK: Cambridge University Press. Pp. 66-67.

Harvey, L. A. 2018. Statistical testing for baseline differences between randomised groups is not meaningful. *Spinal Cord* 56:919. https://doi.org/10.1038/s41393-018-0203-y.

Hopewell, S., M. Clarke, D. Moher, E. Wager, P. Middleton, D. G. Altman, K. F. Schulz, and the CONSORT Group. 2008. CONSORT for reporting randomized controlled trials in journal and conference abstracts: Explanation and elaboration. *PLOS Medicine* 5(1):e20.

Ioannidis, J. P. A., S. J. W. Evans, P. C. Gøtzsche, R. T. O'Neill, D. G. Altman, K. Schulz, and D. Moher. 2004. Better reporting of harms in randomized trials: An extension of the consort statement. *Annals of Internal Medicine* 141(10):781-788.

IOM (Institute of Medicine). 2015. *Sharing Clinical Trial Data: Maximizing Benefits, Minimizing Risk*. Washington, DC: The National Academies Press. https://doi.org/10.17226/18998.

Mayo-Wilson, E., N. Fusco, T. Li, H. Hong, J. K. Canner, and K. Dickersin. 2017. Multiple outcomes and analyses in clinical trials create challenges for interpretation and research synthesis. *Journal of Clinical Epidemiology* 86:39-50.

Moher, D., S. Hopewell, K. F. Schulz, V. Montori, P. C. Gøtzsche, P. J. Devereaux, D. Elbourne, M. Egger, and D. G. Altman. 2010. CONSORT 2010 explanation and elaboration: Updated guidelines for reporting parallel group randomised trials. *BMJ* 340:c869.

NASEM (National Academies of Sciences, Engineering, and Medicine). 2019. Reproducibility and Replicability in Science. Washington, DC: The National Academies Press. https://doi.org/10.17226/25303.

NIH (National Institutes of Health). 2020. NIH data sharing information. https://www.ncbi.nlm.nih.gov/pubmed (accessed September 10, 2020).

O'Haire, M. E., and K. E. Rodriguez. 2018. Preliminary efficacy of service dogs as a complementary treatment for posttraumatic stress disorder in military members and Veterans. *Journal of Consulting and Clinical Psychology* 86(2):179-188.

Pavela, G., H. Wiener, K. R. Fontaine, D. A. Fields, J. D. Voss, and D. B. Allison. 2015. Packet randomized experiments for eliminating classes of confounders. *European Journal of Clinical Investigation* 45(1):45-55.

Raina, P., D. Waltner-Toews, B. Bonnett, C. Woodward, and T. Abernathy. 1999. Influence of companion animals on the physical and psychological health of older people: An analysis of a one-year longitudinal study. *Journal of the American Geriatrics Society* 47(3):323-329.

Ritchie, E. C., and R. Amaker (Senior Editors). The therapeutic use of canines in Army medicine. *The AMEDD Journal* April 2012.

Robinson, K. A., and S. N. Goodman. 2011. A systematic examination of the citation of prior research in reports of randomized, controlled trials. *Annals of Internal Medicine* 154(1):50-55.

Saunders, G. H., K. Biswas, T. Serpi, S. McGovern, S. Groer, E. M. Stock, K. M. Magruder, D. Storzbach, K. Skelton, T. Abrams, M. McCranie, J. Richerson, P. A. Dorn, G. D. Huang, and M. T. Fallon. 2017. Design and challenges for a randomized, multi-site clinical trial comparing the use of service dogs and emotional support dogs in veterans with post-traumatic stress disorder (PTSD). *Contemporary Clinical Trials* 62:105-113. doi: 10.1016/j.cct.2017.08.017.

SPIRIT (Standard Protocol Items: Recommendations for Interventional Trials). 2020. Outcomes. https://www.spirit-statement.org/outcomes-2 (accessed July 1, 2020).

Stafford, K. J. 1966. Opinions of veterinarians regarding aggression in different breeds of dogs. *New Zealand Veterinary Journal* 44:138-141.

Taichman, D. B., P. Sahni, A. Pinborg, L. Peiperl, C. Laine, A. James, S.-T. Hong, A. Haileamlak, L. Gollogly, F. Godlee, F. A. Frizelle, F. Florenzano, J. M. Drazen, H. Bauchner, C. Baethge, and J. Backus. 2017. Data sharing statements for clinical trials: A requirement of the international committee of medical journal editors. *PLOS Medicine* 14(6):e1002315.

VA (Department of Veterans Affairs). In Press. *Draft Monograph on Performance and Results of Post Traumatic Stress Disorder-Service Dog Study*. Washington, DC: Department of Veterans Affairs.

Valentine, D. P., M. Kiddoo, and B. LaFleur. 1993. Psychosocial implications of service dog ownership for people who have mobility or hearing impairments. Philadelphia, PA: Haworth Press.

Van den Berg, S. M., H. C. M. Heuven, L. Van den Berg, D. L. Duffy, and J. A. Serpell. 2010. Evaluation of the C-BARQ as a measure of stranger-directed aggression in three common dog breeds. *Applied Animal Behaviour Science* 124:136-141.

Zarin, D. A., T. Tse, R. J. Williams, and T. Rajakannan. 2017. Update on trial registration 11 years after the ICMJE policy was established. *New England Journal of Medicine* 376(4):383-391.

5

Statistical Methods

This chapter reviews the inclusion, presentation, and discussion of the statistical methods used to assess the results of the clinical trial. This chapter covers issues related to the study population used in the statistical analyses, study design characteristics and planned analyses, the analyses presented and their interpretation, the need to discuss the clinical significance of observations and effect sizes, missing data and the implications for interpretation, sensitivity analysis, and exploratory analyses. In each section, the committee articulates the specific issue and provides a recommendation and rationale for needing to resolve the issue.

INTENT TO TREAT ANALYSES ARE ABSENT

As noted in the study protocol, analyses were to be carried out for both the intent to treat (ITT) population and the per-protocol population (PP). The protocol does not specify a priori which analysis (the ITT or PP) will be considered the primary inferential analysis; however, the ITT analysis, analyzing all consented participants as randomized, is considered the gold standard for inference in superiority trials (Ranganathan et al., 2016; Schulz et al., 2010). The ITT aims to avoid selection bias by employing the benefits of randomization while the PP analysis might include selection bias that results from post-randomization exclusions or withdrawals. Unfortunately, the ITT analyses of the primary and secondary outcomes are not included in the draft monograph. The vast majority of presented results are devoted to the subset of participants who were successfully paired with a dog, the PP cohort. Initial demographic and health history information for all those randomized (the ITT cohort) should be included for the reader's assessment. To reflect the intended randomized design, the ITT analyses are critical as the first step to formal evaluation of the randomized intervention. As will be discussed further in this chapter, the ITT analyses should be accompanied by a thorough description of those that may be missing the primary outcome(s), with evaluation if these participant outcomes can reasonably be considered missing completely at random or missing at random, conditional on any covariates that may account for the missingness (Bell et al., 2014; Little et al., 2012). Subsequent analyses of the PP may help to inform questions surrounding comparison of those who followed the intended protocol, or as a sensitivity analysis. However, conclusions drawn from the analyses in the PP population should account for uncertainty due to missing data in this cohort. It should be noted that the descriptive summaries of randomized participants' characteristics need not be evaluated by a statistical test (as was done in Table N in the draft monograph).

DESIGN CHARACTERISTICS AND PLANNED ANALYSIS

The authors have provided some information regarding the background and design of the randomized trial, including through a peer-reviewed paper (Saunders et al., 2017). Given the study protocol and design paper, the experimental design is established to detect differences between randomized groups for three primary outcomes. However, in the design there was no accommodation for having two primary outcomes (World Health Organization Disability Assessment Schedule 2.0 [WHODAS 2.0] and the Veterans RAND 12 Item Health Survey [VR-12]), which were operationalized by three outcome measures (WHODAS 2.0, VR-12 Physical Component Score, VR-12 Mental

Component Score). It is unclear in the protocol and the draft monograph if the intention was to deem the intervention a success if all primary outcome measures were in favor of the service dog (SERV) intervention (which would define co-primary outcomes), or if the trial would be a success if any one of the outcome measures were in favor of the experimental group (multiple primary outcomes). The inclusion of multiple primary or co-primary outcomes has varying implications on the Type I error rate of the design and subsequent interpretation of results (FDA, 2017). Clarification on the intention of the investigators (co-primary or multiple primary outcomes) needs to be included. Adjustment or comment on multiplicity issues surrounding these designations should be included in the methods and results interpretation. Moreover, the trial protocol details that the data and safety monitoring committee (DSMC) will "receive analyses of primary and secondary outcome measures on a routine basis" (p. 74, Protocol provided to the committee; see Appendix C for more information). However, it does not appear that interim analyses were carried out, or that interim monitoring to evaluate efficacy was incorporated into the study design to avoid inflation of the Type I error. The draft monograph (lines 1441-1442) details that the DSMC "decided that study safety data, rather than results of interim analysis" would be used for study monitoring. It would be helpful to explicitly note if any interim analysis of primary or secondary outcomes, within or between randomized groups, was performed, or if this was a change from the trial protocol. If interim testing was carried out, the committee encourages discussion of the interpretation of these analyses and any planned stopping rules or adaptations.

Furthermore, more detailed information should be provided in the draft monograph for the rationale of study sample size and the effect sizes used in the trial design as minimally important effects. The draft monograph (lines 952-954) suggests that effect sizes were informed by previous preliminary work. More discussion of the limitations of this work to inform the current trial would be welcome. Furthermore, line 957 states that the intended sample size accounts for (non-differential) drop out of 25%; however, this appears to be 35%, because 110 participants per group were planned when 82 were required with no drop out. In line with the discussion of missing data, it would be beneficial to include a comment on this planned rate of drop out.

The trial protocol and design paper describe the intended outcomes as the "between group difference" (p. 69, Protocol provided to the committee; Saunders et al., 2017, Table 2) for the establishment of sample size and power. However, more precise language around the effect of interest is important to include throughout the manuscript. For instance, it is critical to say explicitly at which time point the primary outcome measures are intended to be compared between randomized groups (presumably 18 months), and to clearly define secondary outcomes (including the primary measures at earlier time points), subgroup analyses, or sensitivity analyses. Primary analyses should align with the study design, and sample size justification. If changes were required or suggested during the course of the trial (perhaps by the DSMC), these changes should be described and the rationale given. For instance, a rationale should be provided for the adjustment of covariates (site, gender, baseline score) in outcome models. These covariates were not pre-specified in the protocol, but mentioned in the Saunders et al. (2017) publication. While adjustment for stratification factors is not required, it is generally recommended because these factors are part of the overall trial design and should improve precision in estimates. Justification for why gender should be included in outcome model specification is missing. Similarly, the draft monograph does not state how time was included in each model for the estimation of treatment effects. The draft monograph should specify the mean, standard deviation, and range of time between randomization and pairing. It appears that this window is fairly large and varied from 3 months to 12 months or more in some cases. Therefore, being clear in how time was parameterized (as a continuous covariate or a categorical one) in each model will aid in interpretation.

PRESENTED ANALYSES AND INTERPRETATION

To aid interpretation by readers and reviewers, model estimates comparing randomized groups at the intended primary outcome time point (18 months) should be provided. It appears that the effect estimates provided in the draft monograph are given at time zero, rather than at 18 months. The

committee recommends that the unadjusted primary and secondary outcomes be reported at baseline, at the time of dog pairing, any intermediate time points, and at 18 months for all randomized participants, along with the estimated difference between groups in each measure and its 95% confidence interval (CI) at 18 months. The authors may also consider estimating the unadjusted difference at 18 months in each outcome for the change from baseline (the difference between the groups of measures at baseline—the measure at 18 months and 95% CI). Virtually no estimations of effects size and their precision are presented. These are also missing for binary outcome data.

INTERPRETATION OF CLINICAL SIGNIFICANCE AND EFFECT SIZES

The results should be interpreted with respect to their magnitude and precision, emphasizing the pre-specified primary outcomes. The interpretation of overall trial results should be balanced with respect to primary, secondary, and exploratory outcomes and analyses, instead of relying on the conclusions that emphasize only those with a statistically "significant" result, particularly given the number of secondary and exploratory analyses performed. Throughout the draft monograph, there is a strong reliance on p-values and "statistical significance," rather than the interpretation of estimated effect sizes and precision. This is particularly concerning given the substantial missing data. Furthermore, substantial superfluous use of p-values in testing baseline demographic and outcome measure characteristics are presented (e.g., Tables N, O, and P in the draft monograph); however, as previously described, these are not meaningful (Harvey, 2018). Throughout the abstract and discussion, there is considerable discussion of potential differences between the effects of the SERVs versus the emotional support dogs (e.g., lines 913-924). Although the interest in this potential difference was the underpinning of the study design, ultimately there was only *one* statistically significant difference between the two intervention groups (self-reported PTSD [posttraumatic stress disorder] [PCL-5], lines 1995-1998), and that difference was not robust across models; the manuscript incorrectly suggests that the results of the sensitivity analyses are consistent, which is true only if both sets of results are interpreted as no compelling evidence that the interventions had differential effects.

The PCL-5 finding needs to be put in appropriate context of differential effect sizes and clinically meaningful differences. The authors are encouraged to review psychometric papers on the PCL-5 by Blevins et al. (2015), Bovin et al. (2016), and Wortmann et al. (2016). Contemporary thinking of reliable or meaningful clinical change information is presented in Wortmann et al. (2016). Wortmann et al. (2016) utilized a score of < 24 on the PCL-5 in their sample for clinically significant change; using two standard deviations below baseline at the start of the trial. Using their 24 cut-off or two standard deviations from baseline in the current trial, it is likely that only a small portion of those in the study made clinically meaningful changes in their self-reported PTSD symptoms (see Table DD in the draft monograph). There are other methods to calculate reliable change (RC; see Wortmann et al., 2016), which could be considered. Five and 10 points (lines 1200-1203) are not substantiated via empirical literature, were not specified or justified a priori in the protocol, and do not follow clinically meaningful change literature. Note scores of 31 to 33 (see Table DD in the draft monograph) indicate the presence of diagnosable PTSD (Bovin et al., 2016), arguing that in the aggregate the sample is within clinical symptom levels at the end of the intervention period. Ultimately, the precision of the findings (e.g., effect sizes, CIs) needs to be reported throughout. These are Consolidated Standards of Reporting Trials reporting requirements and requirements of many, if not all, of the journals that publish clinical trials (as discussed in Chapter 4). Clinically meaningful change needs to be appropriately measured, reported, and discussed within the draft monograph (see reliable change literature, Jacobson and Traux [1991]). Finally, the large amount of discussion about potential treatment group effects regarding PTSD almost obscures the major finding, which is that there were no reliable differences observed across almost all primary and secondary outcome measures after being provided with dogs. The committee recommends the reorienting of the discussion section to primary outcome measures and the major finding of the trial.

In general, the committee found the organization of some of the tables difficult to understand. Some tables appear to give only F-statistics and p-values, without an accompanying effect measure

(estimate). The committee suggests revising the tables to improve readability, including simple estimates of effect that would be understandable to non-statistician scientific audiences. The authors may want to consult the following references: Durlak, 2009; Marfo and Okyere, 2019; McLeod et al., 2016; Pek and Flora, 2019; and Vacha-Haase and Thompson, 2004.

MISSING DATA AND LOSS TO FOLLOW-UP

In the protocol and Saunders et al. (2017) paper, the study team set out to describe missing data and the patterns by treatment assignment group, and to employ multiple imputation ("when needed"). At the design stage, the study team acknowledges the potential of substantial drop out (accounting for 25%-35% attrition), presumably in part because randomization was performed far ahead of treatment implementation. As a consequence, the ITT analysis that was specified at the design stage was not followed despite substantial missing outcome data. A clear analysis strategy to address this limitation is not presented.

Furthermore, in the draft monograph, missing data have been described for participants who were paired with a dog (see Table RR in the draft monograph); however, these comparisons between those who completed follow-up and those who did not following pairing include only outcomes. There is no description of participant characteristics between completers and non-completers, in the subcohort who were paired. Similarly, there is no description of demographics or baseline assessment of those who were randomized and who did not successfully pair with a dog compared to those who were paired with a dog. These summaries are critical to providing readers, including those who intend to use these data for making policy, the opportunity to gauge differential drop out prior to active treatment. Particularly, these descriptions would help justify if these data are missing at random, given the study covariates. The authors should comment on their justification for the assumption that missing at random is reasonable. While multiple imputation is suggested as a means to address issues of missing outcome data, it appears that this was done primarily on the PP cohort. Failure to fully describe and attempt to account for the uncertainty due to missing data will limit the interpretation of study results. Additional analyses and discussion of the authors' assumptions and limitations for the ITT cohort are requested.

The overall description of the multiple imputation procedure is lacking. Detail is needed as to what covariates were included in the missing data model, what software the missing data model was estimated in, the number of imputed datasets this included, and the analysis methods employed following multiple imputation (Little et al., 2012).

SENSITIVITY ANALYSIS

The authors have conducted several additional analyses, which one may consider sensitivity analyses. The primary analysis model was planned to be a linear repeated measures model (i.e., a repeated measures analysis of covariance [ANCOVA]) to accommodate all time points under the study. The authors supplement this analysis with a linear mixed effects model, including a random intercept for each patient. The committee suggests that the linear mixed effects model should be the primary analysis model over the repeated measures ANCOVA. The linear mixed effects model allows all individuals randomized to the trial (regardless of missing data) to be included, under the assumption that data are missing at random, conditional on observed covariates. Further analyses that also account for characteristics related to missingness would bolster the reliability, interpretability, and confidence in the results from models with missing observations. The authors do follow the initially proposed models with those that incorporate multiply imputed data. While the committee encourages these methods as sensitivity analyses, further detail is needed as to what covariates were included in the missing data model, what software the missing data model was estimated in, the number of imputed datasets this included, and the analysis methods employed following multiple imputation.

SUBGROUP OR EXPLORATORY ANALYSES

Across most measures, it appears that there may be differential effects of site, baseline score, and gender. While these covariates were included in the models, it is not clear that they were examined for potential interactions with the intervention and time. Moreover, a subgroup or stratification factor for analyses could include severity or impairment of baseline scores. Given the high rate of co-occurrence, baseline comorbid major depressive disorder should be examined as a potential effect modifier. If upon updated analysis there continues to be a lack of clinically meaningful effects at the primary time points of interest between randomized intervention groups, subgroup analysis may not be warranted given the potential for inflation of the Type I error. The authors should refer to the following references for further guidance: Koch and Framke, 2014; Tanniou et al., 2014.

REFERENCES

Bell, M. L., M. Fiero, N. J. Horton, and C.-H. Hsu. 2014. Handling missing data in RCTs: A review of the top medical journals. *BMC Medical Research Methodology* 14(1):118.

Blevins, C. A., F. W. Weathers, M. T. Davis, T. K. Witte, and J. L. Domino. 2015. The posttraumatic stress disorder checklist for DSM-5 (PCL-5): Development and initial psychometric evaluation. *Journal of Traumatic Stress* 28:489-498.

Bovin, M. J., B. P. Marx, F. W. Weathers, M. W. Gallagher, P. Rodriguez, P. P. Schnurr, and T. M. Keane. 2016. Psychometric properties of the PTSD checklist for Diagnostic and Statistical Manual of Mental Disorders–Fifth Edition (PCL-5) in veterans. *Psychological Assessment* 28(11):1379.

Durlak, J. A. 2009. How to select, calculate, and interpret effect sizes. *Journal of Pediatric Psychology* 34(9)917-928. https://doi.org/10.1093/jpepsy/jsp004.

FDA (Food and Drug Administration). 2017. Multiple endpoints in clinical trials guidance for industry. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/multiple-endpoints-clinical-trials-guidance-industry (accessed September 21, 2020).

Harvey, L. A. 2018. Statistical testing for baseline differences between randomised groups is not meaningful. *Spinal Cord* 56:919. https://doi.org/10.1038/s41393-018-0203-y.

Jacobson, N. S., and P. Truax. 1991. Clinical significance: a statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology* 59:12-19.

Koch, A., and T. Framke. 2014. Reliably basing conclusions on subgroups of randomized clinical trials. *Journal of Biopharmaceutical Statistics* 24(1):42-57.

Little, R. J., R. D'Agostino, M. L. Cohen, K. Dickersin, S. S. Emerson, J. T. Farrar, C. Frangakis, J. W. Hogan, G. Molenberghs, S. A. Murphy, J. D. Neaton, A. Rotnitzky, D. Scharfstein, W. J. Shih, J. P. Siegel, and H. Stern. 2012. The prevention and treatment of missing data in clinical trials. *New England Journal of Medicine* 367(14):1355-1360.

Marfo, P., and G. A. Okyere. 2019. The accuracy of effect-size estimates under normals and contaminated normals in meta-analysis. *Heliyon* 5(6):e01838. doi: 10.1016/j.heliyon.2019.e01838.

McLeod, L. D., J. C. Cappelleri, and R. D. Hays. 2016. Best (but oft-forgotten) practices: Expressing and interpreting associations and effect sizes in clinical outcome assessments. *The American Journal of Clinical Nutrition* 103:685-693.

Pek, J., and D. B. Flora. 2018. Reporting effect sizes in original psychological research: A discussion and tutorial. *Psychological Methods* 23(2):208-225. http://dx.doi.org/10.1037/met0000126.

Ranganathan, P., C. Pramesh, and R. Aggarwal. 2016. Common pitfalls in statistical analysis: Absolute risk reduction, relative risk reduction, and number needed to treat. *Perspectives in Clinical Research* 7(1):51-53.

Saunders, G. H., K. Biswas, T. Serpi, S. McGovern, S. Groer, E. M. Stock, K. M. Magruder, D. Storzbach, K. Skelton, T. Abrams, M. McCranie, J. Richerson, P. A. Dorn, G. D. Huang, and M. T. Fallon. 2017. Design and challenges for a randomized, multi-site clinical trial comparing the

use of service dogs and emotional support dogs in veterans with post-traumatic stress disorder (PTSD). *Contemporary Clinical Trials* 62:105-113. doi: 10.1016/j.cct.2017.08.017.

Schulz, K. F., D. G. Altman, and D. Moher. 2010. CONSORT 2010 statement: Updated guidelines for reporting parallel group randomised trials. *BMJ* 340:c332.

Tanniou, J., I. v. d. Tweel, S. Teerenstra, and K. C. B. Roes. 2014. Level of evidence for promising subgroup findings in an overall non-significant trial. *Statistical Methods in Medical Research* 25(5):2193-2213.

VA (Department of Veterans Affairs). In Press. *Draft Monograph on Performance and Results of Post Traumatic Stress Disorder-Service Dog Study*. Washington, DC: Department of Veterans Affairs.

Vacha-Haase, T., and B. Thompson. 2004. How to estimate and interpret various effect sizes. *Journal of Counseling Psychology* 51(4):473-481. https://doi.org/10.1037/0022-0167.51.4.473.

Wortmann, J. H., A. H. Jordan, F. W. Weathers, P. A. Resick, K. A. Dondanville, B. Hall-Clark, and J. Mintz. 2016. Psychometric analysis of the PTSD Checklist-5 (PCL-5) among treatment-seeking military service members. *Psychological Assessment 28*(11):1392.

6

Interpretations and Limitations

This chapter will review in more detail aspects of the draft monograph that relate to the ability to appropriately interpret the study results. It identifies issues with the study design, conduct, reporting, and the discussion of the results, as well as the authors' descriptions and the style used to present the information. The draft monograph contains minimal discussion of the limitations of the study making it difficult to properly qualify the outcomes and findings presented. A greater clarity is needed to distinguish between the authors' interpretation of the outcomes and what the study actually demonstrates.

STUDY DESIGN LIMITATIONS AND THEIR IMPACT ON INTERPRETATION

The conclusions overstate confidence in the evidence provided because they do not address important limitations of the study. These conclusions should be better qualified because of the amount of missing data (discussed in Chapter 5), differences in missing data between the groups (discussed in Chapter 5), and study design limitations discussed in this section.

Interpretation of Within-Group Change

Within-group change is not evidence of efficacy or effectiveness. The draft monograph interprets pre-post changes within each group as evidence that both interventions caused the within-group changes. The study did not include a no treatment condition, so it is possible that Veterans would have improved in the absence of the dogs, and Veterans might have improved to an even greater extent than they improved in the study. For example, in the draft monograph Table PP (line 2221) could be deleted. Because there is no rigorous evidence that either intervention is better than doing nothing prior to this trial, the lack of a no treatment condition was bound to make interpretation of this trial difficult. That is, the absence of a difference provides neither evidence that both interventions are effective nor evidence that neither intervention is effective. The presence of a difference between groups would also be difficult to interpret; for example, if the emotional support dog (EMOT) intervention worsens symptoms of posttraumatic stress disorder (PTSD), then evidence that the service dog (SERV) intervention is better than the EMOT intervention would not demonstrate that the SERV intervention is better than doing nothing. A no treatment comparator would have been appropriate because in clinical trial design it is customary to establish that one intervention is better than nothing before comparing two similar interventions. The authors appropriately acknowledge the importance of the lack of a control group (lines 2494-2496), though this does not translate well into the interpretation of the study findings regarding within-group change. Careful review of the manuscript to mitigate the interpretation of effects of the interventions should take place (e.g., line 2334).

Interpretation of Potential Equivalence

The absence of a difference between groups is not evidence of equivalence. Because the study was designed and powered to detect the *superiority* of the SERV intervention over the EMOT intervention, the results can only be interpreted as failing to reject the hypothesis that SERV was more effective than EMOT. This is a particularly common mistake in the interpretation of results from

randomized controlled trials. The mistake involves conflating no statistically significant difference in effectiveness with "equally effective." In fact, such evidence is just as compatible with a conclusion of "equally ineffective" as with a conclusion of "equally effective." For a discussion of this topic in a different context (e.g., childhood obesity interventions), see "error #8" in Brown et al. (2018). Careful review of the manuscript for this type of error is needed (e.g., "equally effective" wording). For example, on line 1063, the draft monograph says that

> the comparison group selected for the study was provision of an EMOT rather than a no intervention control group. This was done because it was important to be able to differentiate benefits derived from the special characteristics of a service dog (specific tasks and public access rights) from those derived simply having a dog as a companion (the human animal bond).

Unfortunately, the interpretation of results is not consistent with this design; for example, the abstract (line 34) says, "This research illustrates the positive impact of both SERVs and EMOTs on Veterans with PTSD as many of the outcomes showed improvement over time for each group." To determine whether either intervention was better than nothing, the study would have needed to include a no treatment comparator. To determine whether both interventions are equally effective or ineffective, an equivalence trial would have been necessary (in which the equivalency margins would have been pre-specified and *much* narrower than the superiority margin used in this trial, leading to a much larger sample size). The draft monograph of this randomized clinical trial (RCT) contains statements making such inflations (e.g., line 2300). These statements should be corrected. Evidence of non-zero effects can come from only the between conditions (not within conditions) tests and a lack of statistically significant evidence of differential effectiveness should not be mistaken for evidence of a lack of differential effectiveness.

Limitations Resulting from Delay Between Randomization and Receipt of Intervention

The authors suggest that the phase of the study between randomization and receipt of intervention adequately controls and tests for symptom improvement prior to the onset of the intervention (e.g., line 2324), but any post-randomization differences in drop out or symptom changes could bias these results. The failure to re-assess trial eligibility criteria for patients when they actually receive the intervention alters the typical "run-in" observations phase of a clinical trial design. Typically, patients who are no longer eligible after a run-in are not randomized. Although in the aggregate patients did not improve in this study (e.g., line 2529), there may have been some patients who made clinically meaningful changes in the protracted phase (after 3 months up to some unspecified range) and their data were included in the analysis. Because data are presented in aggregate, potentially including randomized ineligible (less severe) participants, this possibility is impossible to assess. Notably, it does not appear that the investigators re-assessed interviewer-administered PTSD prior to randomization or at intervention. This should be stated as a limitation when interpreting these data because PTSD severity prior to the start of the intervention is not known and therefore the potential change induced by the intervention is a challenge to estimate. Self-reported PTSD often mirrors general psychopathology and patients often fail to make distinctions between trauma-related distress and general distress; limiting the utility of only self-report PTSD measures. For these reasons, statements that symptom improvement are the result of the intervention without mention of these limitations are overstated and do not address important threats to causal inference.

Non-Masked Lack of Equipoise Design Limitations

The concept of equipoise refers to whether providers and participants in a trial have equivalent beliefs and feelings about the conditions to which the participants may be assigned such that any differences in beliefs and feelings do not provide an alternative explanation for the effects of treatment

assignment. This is particularly a problem in clinical trials where there is one clearly preferred intervention. This is likely the case when comparing EMOT and SERV intervention groups. Conceived this way, the issue is one of expectancy and effects of treatment assignment. In a pharmaceutical RCT, this would typically be dealt with by a double-masked, placebo-controlled design.

In unable-to-be-masked interventions such as the interventions studied in this RCT, such procedures are not possible. Instead, three ways of addressing this concern are available: (1) enhanced study design elements; (2) measurement of patient expectancies, preferences, and satisfaction; and (3) careful discussion of study limitations. First, the concern may be addressed through design by minimizing (though usually not strictly eliminating) such potential non-equivalencies. The committee believes that the inclusion of a control group in which a dog is provided (i.e., the provision of an EMOT) effectively controls for the non-specific effects of merely having a dog and for the non-specific effects of receiving some intervention that has some validity. However, it does not control for overall contact with other humans in the setting of treatment provision, the types of activities that the participants were required to do, or the credibility of the intervention. While this does not invalidate the trial, it does limit the conclusions that can be drawn.

Another design feature in non-masked trials that is commonly employed is the use of masked assessors, where the primary outcomes of the trial are repeatedly measured by reliably trained assessors. Considerable care should be given to the standardized training, masking, and inter-rater reliability across assessors and sites in most clinical trials. This design feature helps mitigate patient self-report biases commonly implicated when assumptions of equipoise are not met. While it is unclear if assessors were masked to the grouping of the Veteran, failure to include this design feature should be noted as a limitation.

A second way to address the lack of equipoise is through analysis by measuring and showing that expectancies and feelings do not differ between groups, are not related to outcomes, or by statistically adjusting for them if they do differ. No measurement of equipoise is presented in the draft monograph, despite interventions being non-masked. Patient preference, intervention credibility, or patient expectancy—all commonly used measures to address equipoise—were not measured. Another measure commonly employed at the end of these types of clinical trials is intervention satisfaction. It does not appear that a well-validated measure of satisfaction was given, though there may be some single items that could serve as a proxy. Taken together, the interventions could not be compared for equivalency nor could they be controlled for in statistical analyses to assess their possible effects. Patient preference effects are often implicated in differential patient drop out in clinical trial interventions. Higher drop out from the EMOT intervention group than the SERV intervention group is consistent with differences in Veteran preferences and expectancies for these interventions. Ultimately, the potential unmeasured role of patient preference and expectancy should be discussed when addressing differential drop out and any observed intervention effects, and as a limitation of the trial. Though not required, the authors could consider conducting additional analyses to potentially address differential attrition by determining whether successful placement (defined by retention in the study and successful placement with a dog, rather than based on any of the measured outcomes) could be predicted by baseline covariates.

Third, a lack of equipoise can be handled in part with careful discussion. This recognizes the limitations of the design, considers the extent to which such limitations may mitigate conclusions of the intervention effects, and shares such considerations with the reader of the draft monograph. Issues raised above regarding the interpretation of the pre-active intervention phase, the lack of masked assessment, differences between EMOT and SERV interventions, and failure to measure patient expectancies all point to design choices that cannot rule out inherent problems with ultimate interpretation of observed effects. If clinically meaningful, significant finds are made, it simply cannot be argued that the *specialized training of the SERVs alone* accounts for any observed intervention effects. The authors are encouraged to make sure that the interpretation of the results concentrates on the entire manipulation given to the two treatment groups. The entire manipulation would compare encompass traveling to a remote location for pairing with the dog, being trained 1-2 weeks, being given a more highly trained dog, and living with a SERV-trained dog for 18 months versus staying home for pairing with the dog, being trained only 1-2

days, being given a less highly trained dog, and living with an EMOT for 18 months. These limitations impact both the mechanistic interpretation of the interventions and reflect a threat to causal inference.

Not Designed as a PTSD Treatment Trial

The authors need to be clear with the reader that the clinical trial was not designed to test a primary intervention in the treatment of PTSD. Throughout the discussion, a naïve reader would assume that it indeed was a well-done PTSD treatment trial. If the focus was on PTSD, masked interviewer-administered PTSD assessment would have been the primary outcome measure, a threshold of PTSD severity would have been specified as an eligibility criterion, and assessment would have been prior to the implementation of the intervention (a proximal baseline measured) and assessed multiple times throughout the trial to develop a trajectory line. Standardized training to a criterion of reliability across sites and inter-rater reliability within and across sites for the Clinician Administered PTSD Scale for DSM-5 (CAPS-5) would have been reported, and masking of interviewers explained. As is reported, the quality of the CAPS-5 assessments presented in the draft monograph are not strong, potentially lacking masking of interviewers (at least this element is unclear within the draft monograph), training to a specified criterion, and inter-rater reliability reported within and across sites, and the reader should not be encouraged to believe that gold standard measurement occurred in this trial (lines 1130 and 2325). Even though solid psychometrically validated measures were used, PTSD was assessed at a substandard level and predominantly relied on self-reported symptoms, which can be affected by patient biases (e.g., lack of equipoise) and may reflect general distress rather than PTSD per se. Furthermore, concurrent PTSD treatment (e.g., selective serotonin reuptake inhibitors, cognitive processing therapy, prolonged exposure, etc.) would have either been controlled in the study design or systematically measured, reported, and statistically adjusted for in the analyses. The latter was possible but not done. Accordingly, it is not possible to conclude that the intervention influenced PTSD symptoms if ongoing PTSD treatment occurred during the intervention (especially, for patients receiving evidence-based interventions), and it was not measured. Throughout the discussion, it should be clear that the trial focused on improving disability functioning and quality of life. These limitations impact the generalizability of the study to the broader PTSD clinical trials literature and reflect a threat to causal inference that the SERV intervention and/or EMOT intervention altered PTSD symptoms.

IMPROVING THE CLARITY OF THE DRAFT MONOGRAPH

In general, the committee suggests the use of clearer language throughout the draft monograph to enhance the interpretation and avoid conjecture. One suggested strategy is for the authors to include the addition of a statement of the primary response (in accordance with the original study design), then add a statement of the authors' interpretation of what each finding suggests, as well as indicate the limitations of the trial design that could contribute to the finding. For example, the authors could consider constructing the text to read "We were unable to reject our primary hypothesis of X…. However, while not demonstrated, we interpret this lack of change to mean X…. We recognize that ongoing mental health treatment could have accounted for this."

In addition, the draft monograph is quite ambiguous about the findings related to the main objectives and hypotheses of the trial. For instance, the authors largely focus on the one positive finding related to self-reported PTSD symptoms without an explicit statement acknowledging that no statistical differences were found between the study groups on the majority of primary, secondary, and tertiary outcomes. The authors directly and quickly move on to state, "The main finding of this study is that both dog types had a beneficial effect on participants' functioning and disability." This was not an a priori hypothesis specified in the protocol and the authors fail to acknowledge the limitations that this finding could be a result of regression to the mean. Similarly, the conclusion section begins with the sentence, "This research illustrates the positive impact of both SERVs and EMOTs on Veterans with PTSD as

many of the outcomes showed improvement over time for each group." Hence, it gives the perception that the results are being cherry picked to provide a positive spin to the study findings.

The committee strongly recommends that the draft monograph should first explicitly acknowledge the null findings on the primary, secondary, and tertiary outcomes (except severity of self-reported PTSD symptoms). In other words, the draft monograph should clearly communicate the findings on what it was originally designed to examine. Only then should it talk about improvements on many outcomes observed in both study groups and clearly acknowledge that these results could be due to regression to the mean.

Clarifying the Differences Between SERV and EMOT Intervention Groups

Vastly different amounts of face-to-face instruction on dog handling and ownership were given to Veterans in the two groups. Veterans in the SERV intervention group received 1-2 weeks of training at the contractor site. Veterans in the EMOT intervention group received 1-2 days training at their homes (lines 1380-1383, 1658-1672). The committee contends that there is a substantial difference in how much one can teach a neophyte dog handler or utilizer in 1-2 weeks versus 1-2 days, and the differing degrees of instruction could potentially have affected the results of the study. If, for instance, Veterans in the SERV intervention group on the whole became more comfortable and competent in handling, controlling, and commanding their dogs than Veterans in the EMOT intervention group, this might have affected how the Veterans in the two groups responded to having a dog (line 854). In lines 2514-2515, the authors do not address the consideration of differing amounts of training given to the Veterans in the SERV intervention group versus those in the EMOT intervention group while closely examining other ways in which the experience of having a dog might have been affected by unintentional treatment group differences.

Careful wording of the interpretation and noting of these limitations are needed. For instance, the draft monograph focuses the reader's attention on the highly trained SERV versus EMOT dimension (lines 1050-1053), leading the reader to believe that (1) no other potentially significant differences existed in the treatments given the two different groups; and (2) potential differences in outcome variables for Veterans in SERV versus EMOT intervention groups would be due to strictly the psychological (for the Veterans) dimension of SERV versus EMOT. Often times, the use of SERV and EMOT in the draft monograph imply the dog type rather than the entirety of the intervention that they represent. Examples include the conclusion on lines 2566-2567 that attributes the significant advantage of Veterans in the SERV intervention group over those in the EMOT intervention group on self-reported PTSD scores to the advantage of the dog type in reduction of PTSD symptoms; and the treatment of this topic in lines 2492-2499 of the discussion, which remains focused on the SERV versus EMOT dimension rather than acknowledging that the experiences of Veterans in the SERV versus EMOT intervention groups varied systematically in at least one other way—amount of dog-related instruction (and potentially in the ways discussed in Chapter 4). These sections of the draft monograph would benefit from revision designed to focus the reader's attention on the effect of all aspects of being in SERV versus EMOT intervention groups rather than the effect solely of living with a SERV versus an EMOT for 18 months.

Fidelity and Protocol Adherence

Replicability and fidelity of each intervention are tenets of a strong intervention trial (see Foa and Meadows, 1997). The information in the draft monograph needs many more training details for future replicability (lines 1387-1401). In most intervention trials, there would be a general intervention manual that all sites follow, even if specific training procedures varied across sites. In many ways, the Contract of Statement of Work (SOW) has a much better discussion of both the SERV and EMOT interventions than the draft monograph. The revised monograph should have a much more detailed description for both interventions, potentially using content directly from the SOW. Although both descriptions need more details, attention should be paid to enhancing the description of training for EMOTs as the control

condition. As written, the EMOT intervention description leaves the naïve reader questioning what skills these dogs were trained in and how well they performed them. The draft monograph notes that dogs in both groups had basic obedience training (line 1625) and both had to pass the American Kennel Club (AKC) Canine Good Citizen test, though it does not elaborate on the specifics of the test. Additionally, the draft monograph notes that EMOTs were tested on the AKC Community Canine Test, but again does not elaborate on the test specifics (which include waits under control; walks on a loose leash [does not pull]; walks through a crowd; walks by distraction dogs; sits in a small group [3 people with dogs]; allows someone to approach and pet; follows instructions to "leave it" and "down or sit stay" at a distance; comes when called; waits for handler to enter/exit doorway).

More specific information needs to be presented in the manuscript about markers of intervention fidelity for both SERV and EMOT interventions over the course of the trial. This simply means, did the dogs and owners do what they were supposed to do? This includes the implementation of training received by the owners, services performed by the dogs, and the match between the dog and the owner. Without fidelity assessment in the intervention phase itself and reported fidelity analyses, the trial is limited to making conclusions solely about the provision of an EMOT versus a SERV rather than any services the dog performed during the intervention phase. Analysis of fidelity indices should be presented in the method section (lines 1387-1401). Within the draft monograph, there are likely some, albeit limited and not masked, intervention fidelity data that could be included in the method section. This could include analysis of scores on the DOGe Quiz, analyses of specific shared (Q#2-16) and disparate intervention Post-Pairing Evaluations (e.g., Q#27), mask coding of a subset of Dog Visit Reports or Form 21: Dog Related Questions. Form 21 seems particularly relevant for pulling together specific items for fidelity analysis or mask coding including Q#2 satisfaction. This analysis should also include the examination of site differences. It is critical for the investigators to present data that show the EMOTs were as well behaved as SERVs throughout the intervention period and that SERVs continued to perform the tasks they were trained to do through the intervention period. The authors are encouraged to think strategically about the key components of fidelity, data they potentially have that address fidelity, and report related analyses when describing the interventions. The authors should also acknowledge as a limitation in the discussion the fact that standard intervention fidelity procedures such as the digital recording of the intervention, random sampling during the intervention period, and using an objective, outside rater code for the fidelity were not employed and thus mitigate conclusions about whether the intended differences between the intervention arms were indeed observed during the intervention.

Intervention fidelity reporting could be augmented with a well-done qualitative analysis of open-ended questions (e.g., Bazeley, 2013, and Creswell and Creswell, 2018). Qualitative data, addressing fidelity, were collected but are not reported in the draft monograph. It is not clear how the authors plan to use the data from the open-ended qualitative questions. The authors acknowledge the importance of this information in the protocol, stating that, "Feedback from study subjects overall is that the key elements of how the dogs help with PTSD related challenges are not being captured by quantitative measures. For this reason, we have added some open-ended questions to the post evaluation (e.g., Service Dog/Emotional Support Dog Post-Pairing Evaluation)." This is further specified in Saunders et al. (2013, p. 53) stating that

> At the 18-month visit, an exit interview is conducted by trained interviewer during which the participant is asked whether he/she wants to keep the dog, the reasons why, the positive and negative aspects of having a dog, the ways in which the dog helped with symptoms of PTSD, the specific service dog tasks used and the frequency with which each was used, what other tasks participants would have liked the dog to be trained to do, the ways in which the dog has impacted HRQoL (Health-related quality of life), ways in which the dog has influenced interpersonal relationships and whether the participant thinks others would say their dog has helped them.

Thus, the importance of this information is well recognized by the authors. Yet, the draft monograph does not present the analysis of these data—not reporting materials clearly specified a priori in the protocol. The committee encourages the authors to conduct qualitative analysis as described a priori and include it in the draft monograph. Regardless, the authors should clearly describe their plan for addressing this discrepancy between reporting and the protocol in the revised monograph.

Symptom Worsening, Avoidance Symptoms, and Safety Behaviors

The draft monograph presents an extensive discussion regarding symptom worsening, avoidance symptoms, and safety behaviors (e.g., lines 2325-2394). This discussion should be revised to reflect indices measured and reported in the trial, results analyzed and presented, and more careful theoretical and empirical understanding of fear conditioning and avoidance under the consultation and editing of a cognitive behaviorally oriented clinical psychologist. This likely means substantial cutting of this discussion section or additional post hoc analyses added to the manuscript (with the appropriate acknowledgment within the next text that these were post hoc).

In a standard, well-conducted RCT, one gets an unbiased estimate of the average treatment effect in those randomized, not the individual effect; it cannot rule out that some benefit greatly or others benefit less unless additional heterogeneity analysis are conducted. Within the PTSD literature, there are established methods to quantify and analyze symptom worsening. See Jayawickreme et al. (2014) and Foa et al. (2002) for potential measurement and reporting of reliable symptom worsening. Differential drop out and adverse outcomes, compared between interventions, are also indices that can speak to these concerns. Data on the aggregate do not provide evidence that worsening did not occur and conclusions based on aggregate data need to be tempered or modified (lines 2381-2387, 2391-2394). Also, non-significant findings should not be interpreted (World Health Organization Disability Assessment Schedule 2.0, CAPS-5, lines 2381-2387). If the authors desire to discuss symptom worsening, analysis of it (not in the aggregate) should be presented.

There is an extensive discussion of avoidance, with details describing the authors' strong interest in understanding the role of avoidance in the use of SERVs. Yet, avoidance subscale PTSD data are not analyzed in the results section, a psychometrically validated measure of avoidance was not included, nor did the study assess safety behaviors. There are psychometrically sound measures of avoidance in the literature (e.g., van Minnen and Hagenaar, 2010). It is typical within PTSD trials to report analyses across symptom clusters (i.e., re-experiencing, avoidance, hyperarousal, cognition/mood). Reporting observations without data (lines 2384-2385) is not sufficient. Analyses to support the observations referenced in lines 2384-2385 could be reported post hoc, using appropriate post hoc controls. Then the authors would be in a firmer position discussing the role of avoidance in the discussion section of the draft monograph. Even if this analysis is included, it could not address whether SERVs actually performed tasks like sweeping, blocking, or turning on the lights for their Veterans during the intervention phase. This speaks to previous concerns about the fidelity of the interventions because it is not clear that any of these potentially avoidance and safety signal relevant dog tasks routinely occurred during the intervention phase. Without this evidence of these occurring, it is very difficult to argue that these tasks were helpful or harmful for the Veterans. If the role of avoidance is discussed, the above limitations should be noted in the same paragraphs of the revised monograph.

Safety behaviors themselves were not measured. The theoretical discussion and review of this relevant literature in exposure therapies could be strengthened. The authors make a statement about the focus of concern being the fear response (e.g., lines 2347-2350) but fail to highlight that the concern is broader than this. Specifically, it is about whether choosing to not reduce avoidance perpetuates PTSD symptoms or, just as critically, leaves a patient vulnerable to relapse, especially when the safety cue is absent (e.g., the dog dies). The section on related neurobiology and theory could either be updated (e.g., Foa and Kozak [1986] is an old theoretical article), improved, or removed (e.g., lines 2356-2360), as some of the statements are not consistent with the empirical literature (i.e., fear activation is necessary for fear reduction, line 2356) or are not conceptually accurate. Moreover, the text in this section of the draft

monograph could be interpreted as narrowly presenting one version of a neurocognitive model of PTSD, focusing largely on fear conditioning and extinction. A key point that is missing in the draft monograph is that safety behaviors do not give a participant *the opportunity to disconfirm maladaptive beliefs* and might reinforce safety seeking. This should be clearly stated at the onset of the discussion of avoidance and safety behaviors. There are no data in this draft monograph that would confirm or disconfirm the hypothesis that SERVs reinforce safety behaviors. The authors should consider removing content in this section; if safety behaviors were not assessed, then the draft monograph should say so and draw *very* limited conclusions about them.

Addressing the Clarity and Consistency of the Use of Construct Terms

The committee noted challenges with accuracy, congruency, consistency, and reliability of the major concepts of interest, measurement tools, and outcome variables in the study. Throughout the draft monograph, there is a need for greater consistency with terminology (e.g., disability functioning, quality of life, depressive symptoms, suicidal ideation and behavior, and anger symptoms). The constructs or domains assessed in the study should be consistent with the intention of instrument developers, except when explicitly noted and justified. For example, the construct of PTSD is described inconsistently (e.g., PTSD symptoms, PTSD severity) making it difficult to understand if the authors are referring to the presence of symptoms or where these symptoms are on a severity scale. Both suicidal ideation and suicidal behavior were measured in the study. Yet, in several areas of the draft monograph, the authors inconsistently refer to the construct as only suicidality or suicidal ideation. Depression is a term consistently noted in the draft monograph. However, depressive symptoms may be a more appropriate term, given the instrument used to operationalize this construct. Furthermore, the psychometric quality of measurement instruments should be discussed.

Measures of intervention characteristics related to the feasibility of the study (e.g., preference, credibility, expectancy, and satisfaction) are not reported in the draft monograph. If assessed, discussion of these measures should be included in the draft monograph. The committee had overarching concerns about the assessment of PTSD. Although all participants were required to sustain mental health care at the VA, the investigators did not incorporate any data regarding VA PTSD-related clinical care received during the trial (e.g., type, dosage, duration), clinician-reported data, or masked assessment of assessor-rated PTSD at the multiple assessment points during the trial. Furthermore, the committee noted that the measure used to assess work productivity, which was general, did not measure productivity losses specifically due to the disease of interest (i.e., PTSD or mental health). Although there is a more specific version of this instrument that is particular to disease status, this version was not used in the study. This should be noted as a limitation of the study. Overall, inconsistent and incongruent terminology for key study constructs, as well as areas of insufficient clarity regarding reliability and patient illness specificity, can result in confusion and an inaccurate interpretation of the study findings. Therefore, the committee suggests that the authors revise the draft monograph in response to these concerns by appropriately addressing these elements as limitations of the study.

Assessing Anomalies in Reporting Results

The committee recommends that the authors of the draft monograph share the dataset, code, and research materials in a permanent repository, as described elsewhere. At a minimum, the authors should be prepared to facilitate inspections of the published results, such as the granularity-related inconsistency of means test (Brown and Heathers, 2017).

REFERENCES

Bazeley, P. 2013. *Qualitative Data Analysis: Practical Strategies*. Thousand Oaks, CA: SAGE Publications Ltd.

Brown, A. W., D. G. Altman, T. Baranowski, J. Martin Bland, J. A. Dawson, N. V. Dhurandhar, S. Dowla, K. R. Fontaine, A. Gelman, S. B. Heymsfield, W. Jayawardene, S. W. Keith, T. K. Kyle, E. Loken, J. M. Oakes, J. Stevens, D. M. Thomas, and D. B. Allison. 2019. Childhood obesity intervention studies: A narrative review and guide for investigators, authors, editors, reviewers, journalists, and readers to guard against exaggerated effectiveness claims. *Obesity Reviews* 20:1523-1541.

Brown, N. J. L., and J. A. J. Heathers. 2017. The GRIM test: A simple technique detects numerous anomalies in the reporting of results in psychology. *Social Psychological and Personality Science* 8(4):363-369. https://doi.org/10.1177/1948550616673876.

Creswell, J. W., and J. D. Creswell. 2018. *Qualitative, Quantitative, and Mixed Methods Approaches*. Thousand Oaks, CA: SAGE Publications Ltd.

Foa, E. B., and M. J. Kozak. 1986. Emotional processing of fear: Exposure to corrective information. *Psychological Bulletin* 99(1):20-35.

Foa, E. B., and E. A. Meadows. 1997. Psychosocial treatments for posttraumatic stress disorder: A critical review. *Annual Review of Psychology* 48(1):449-480.

Foa, E. B., L. A. Zoellner, N. C. Feeny, E. A. Hembree, and J. Alvarez-Conrad. 2002. Does imaginal exposure exacerbate PTSD symptoms? *Journal of Consulting and Clinical Psychology* 70(4):1022-1028.

Jayawickreme, N., S. P. Cahill, D. S. Riggs, S. A. Rauch, P. A. Resick, B. O. Rothbaum, and E. B. Foa. 2014. Primum non nocere (first do no harm): Symptom worsening and improvement in female assault victims after prolonged exposure for PTSD. *Depression and Anxiety* 31(5):412-419.

Saunders, G. H., K. Biswas, T. Serpi, S. McGovern, S. Groer, E. M. Stock, K. M. Magruder, D. Storzbach, K. Skelton, T. Abrams, M. McCranie, J. Richerson, P. A. Dorn, G. D. Huang, and M. T. Fallon. 2017. Design and challenges for a randomized, multi-site clinical trial comparing the use of service dogs and emotional support dogs in veterans with post-traumatic stress disorder (PTSD). *Contemporary Clinical Trials* 62:105-113. doi: 10.1016/j.cct.2017.08.017.

VA (Department of Veterans Affairs). In Press. *Draft Monograph on Performance and Results of Post Traumatic Stress Disorder-Service Dog Study*. Washington, DC: Department of Veterans Affairs.

van Minnen, A., and M. A. Hagenaars. 2010. Avoidance behaviour of patients with posttraumatic stress disorder. Initial development of a questionnaire, psychometric properties and treatment sensitivity. *Journal of Behavior Therapy and Experimental Psychiatry* 41(3):191-198.

7

Economic Analysis

The committee raised a number of concerns regarding the economic analyses presented in the draft monograph. The first concern relates to a mismatch between what was specified in the protocol and what was reported in the draft monograph. It does not include all of the economic outcomes specified in the protocol and in the study registration, and the draft monograph introduces new outcomes and analyses that were not planned. Furthermore, the committee raised concerns that partial reporting of the economic analyses in the draft monograph could be misleading. Thus, the committee recommends that the investigators either (1) revise this monograph to include a comprehensive account of all economic outcomes, analyses, and results or (2) include all of the economic outcomes, analyses, and results in the second planned monograph. If results for the pre-specified economic outcomes (i.e., those in the study registration and protocol) are not included in this monograph, then this monograph should list the planned outcomes in the methods and note that they will be reported elsewhere. The goal of whatever approach the authors choose is to avoid an incomplete monograph draft and potentially incorrect interpretation. The rest of this chapter provides more detailed comments to further illustrate these points.

The committee noted several inconsistences between the protocol and the draft monograph. For instance, in the protocol the authors describe a plan to estimate inference using generalized linear models (GLMs) for the Work Productivity and Activity Impairment (WPAI)-related outcomes to allow more flexibility in the distribution of the error terms and outcomes modeled. However, in the draft monograph, linear (i.e., ordinary least squares) regressions were described. The GLMs may be more appropriate depending on the distribution of the outcomes but why the authors switched to linear regressions has not been justified in the draft monograph. Also, it is not clear that two analyses were estimated for productivity per the protocol, where the first analysis focused on work productivity and to assign a value of 0% productivity to any individual who is not employed and the second analysis focused on overall productivity and to use % productivity at work if the respondent was formally employed and % productivity for activities of daily living (ADLs) if the respondent was not formally employed. If these analyses were performed, they should be explained in detail. Furthermore, it is suggested that the methods section of the draft monograph include the details on how the overall work productivity and ADL outcomes were derived from the WPAI V2.0 for clarity for readers not familiar with the WPAI instrument. The committee noted that these were well described in the protocol but do not appear in the draft monograph.

The second concern raised by the committee relates to the lack of details and justification for statistical techniques and model specification, including the variable selection for conducting the economic analyses in the protocol or in the draft monograph. For instance, it is unclear what models were estimated for the non–Department of Veterans Affairs (VA) health care utilization. The type of family and link function for the models is not specified in the text or as a footnote to each table in the draft monograph for the economic outcomes. Also, given the differential rate of drop out and missingness, a more thorough assessment of the mechanism of missingness accounting for differences in economic variables among those with complete data and incomplete data is important before selecting the method to account for missing/censored cost data (Glick et al., 2015). Also, the rationale for the selection of the variables (or lack thereof) as covariates in the regressions for the economic outcomes analyses is not explained and is also problematic because it appears that the analyses of clinical outcomes controlled for center, gender etc., whereas the economic analyses did not. Similarly, time is specified as a dummy

variable in the economic analyses versus as a continuous variable in the clinical outcomes analyses. More details about the analyses performed, and justification for those analyses, would be needed to understand the economic analyses.

The third concern raised by the committee is that the economic outcomes were not used to determine the study sample size and that the protocol does not provide a power analysis for the economic outcomes. Hence, it is unclear if the economic analyses are statistically powered for making appropriate inferences about the economic outcomes. This is important because given the right skewness of economic outcomes data the sample size required is typically higher than that required for clinical outcomes. It is typically recommended to calculate the statistical power available for economic outcomes at the sample size planned based on the clinical outcomes (Glick et al., 2015). Currently, the reader of the draft monograph has no way of knowing if the study is statistically powered for identifying differences in economic outcomes. For instance, the primary analysis for work productivity is estimated only among those who were employed and includes only approximately 25 patients in each group. Hence, it does not appear that this study was sufficiently powered for economic outcomes and thus that should be stated as a limitation.

The fourth limitation of the study is that critical economic measures were not collected in the trial, including health-related quality of life and intervention costs. An economic evaluation will typically include a utility instrument such as the EuroQol-5 Dimension or Health Utilities Index, which measures generic quality of life. Such an instrument is important to express the cost-effectiveness of an intervention in terms of the most widely recommended metric of incremental costs per quality-adjusted life years saved that permit comparisons across different interventions within and between different health problems. In an economic evaluation, it is also important to capture the costs of the intervention itself in addition to other medical costs that are likely to be offset (or incurred) by the intervention. However, neither the protocol nor the draft monograph mention or capture any costs associated with the two interventions (service dogs and emotional support dogs). Without these costs the economic evaluation will be incomplete. The gold standard for economic evaluations is a societal perspective to generating cost estimates. This approach would include not just services used by the study participants but also all societal costs that may differ between participants in the two arms. As indicated above, even for the intervention costs this would include time to acquire and train the dogs multiplied by the value of the time of the trainers (which would not be zero under any circumstances even using volunteers, but would be valued at the appropriate market wages, if available, or a standardized wage), time costs related to on-boarding a new dog, related transportation costs, veterinary care, as well as differences in all types of outcomes for study participants, such as medications used and health care visits. The outcomes captured cover only a small fraction of these inputs. While some economic evaluations do use a more limited set of outcomes, it is typically because they adopt another costing perspective than societal costs; if, for example, the decision maker for the selection of treatment is a clinic, an agency perspective might be incorporated. While it would have been ideal if the study team had collected input quantities and costs during the study period, there are methods that have been proven fairly accurate to impute these inputs post hoc. Finally, the non-VA outpatient and inpatient care questionnaire collected information on all use of non-VA services. However, it did not ask the reason for the inpatient, emergency room, or outpatient visit. Hence, the team will not be able to assess the mental health–related non-VA health care utilization and costs, which may be more likely to be impacted by the intervention. The committee noted that in the protocol the team specifically described examining mental health–related VA care using the administrative datasets but did not plan to collect and analyze non-VA care in a similar manner.

Finally, the draft monograph includes an incomplete account of health care costs. The trial registration and the study protocol state that information on VA health care utilization will be collected from VA administrative datasets. This information and related analyses are not provided in the draft monograph and are critical for the interpretation of the main study findings. Instead the draft monograph presents only the non-VA, self-reported health care utilization data. However, the use of non-VA care is typically substantially lower than the use of VA care among Veterans, especially for mental health concerns such as posttraumatic stress disorder as required in this study. Hence, it is important to look at

the differences in overall (VA plus non-VA) health care utilization and costs between the two study groups to get an accurate picture of the economic impact of the intervention. The current lack of differences on the non-VA-administered health care cannot be interpreted as a lack of overall difference in health care use (and costs). This issue can be addressed by presenting all of the results for the planned outcomes (including VA and non-VA health care utilization and costs and employment and productivity) so that the interpretation of findings may provide a balanced interpretation given the totality of the results.

REFERENCE

Glick, H. A., J. A. Doshi, S. S. Sonnad, and D. Polsky. 2015. *Economic Evaluation in Clinical Trials, Second Edition*. Oxford, UK: Oxford University Press.

8

Glossary

Confounders: Common causes of exposure and of an outcome, in which the variable is causally associated with the exposure and causally or non-causally associated with the outcome and does not lie in the causal pathway between exposure and outcome.

Mediator: An intermediate variable, which is caused by the exposure and in turn causes the outcome.

Outcomes: An event in a person used to assess a treatment's effect. A fully defined outcome includes five elements: (1) outcome domain, (2) specific measure, (3) specific metric, (4) method of aggregation, and (5) time point.

Result: A numerical contrast between a treatment and a comparison group (e.g., relative risk, mean difference).

Sensitivity analysis: Analysis that assesses the impact of varying key assumptions made, such as the statistical model, the assumptions made regarding missing data or protocol deviations, or in the definition of outcomes, on the overall study conclusions.

Subgroup analysis: Analysis of the exposure and outcome relationship within a subset of the cohort, such as by gender or recruitment site.

Appendix A

Committee Biographies

David B. Allison, *Chair*, is the Dean and a Distinguished Professor in the School of Public Health at Indiana University Bloomington. He received his Ph.D. from Hofstra University in 1990. He then completed a post-doctoral fellowship at the Johns Hopkins University School of Medicine and a second post-doctoral fellowship at the National Institutes of Health (NIH)-funded New York Obesity Research Center at St. Luke's/Roosevelt Hospital Center. He was a research scientist at the New York Obesity Research Center and an Associate Professor of medical psychology at the Columbia University Vagelos College of Physicians and Surgeons until 2001. He became the Dean and a Provost Professor at the Indiana University Bloomington School of Public Health in 2017. Prior he was a Distinguished Professor, a Quetelet Endowed Professor, and the Director of the NIH-funded Nutrition Obesity Research Center at the University of Alabama at Birmingham. Dr. Allison is a member of the National Academy of Medicine.

Mary Burch is one of less than 100 Certified Applied Animal Behaviorists in the United States who routinely answers questions about problems related to a dog's behavior. Dr. Burch is the Director of the American Kennel Club (AKC) Family Dog program and in this capacity she designs and oversees programs that are implemented on a national scale for pet dog owners and dog trainers. Examples of these programs are the AKC Community Canine and Urban Canine Good Citizen programs that were designed and implemented by Dr. Burch, and the AKC S.T.A.R. Puppy program (a widely utilized puppy training program). The development of each of these programs has required knowledge and expertise in animal behavior, training, and canine development. Other programs developed and implemented nationwide by Dr. Burch include the AKC Trick Dog, AKC Therapy Dog, AKC FIT DOG, and AKC Temperament Test. She also participates in interviews about training and behavior. Dr. Burch is a member of the American Service Dog Access Coalition (ASDAC) committee that is working with airlines to decrease fraud with regard to self-trained service dogs. As a part of an ASDAC project, she reviewed the tests for 24 service dog organizations and developed a test that could be used for testing self-trained dogs (whose owners will take them to an evaluator for testing).

Jalpa Doshi is a Professor at the University of Pennsylvania and a Senior Fellow at the Leonard Davis Institute of Health Economics. She is also the Director of the Economic Evaluations Unit of the Center for Evidence-based Practice and the Director of Value Based Insurance Design Initiatives at the Center for Health Incentives and Behavioral Economics. Dr. Doshi received her Ph.D. in pharmaceutical health services research from the University of Maryland, Baltimore. Her work applies health economics, outcomes research, and policy methods to address issues related to pharmaceutical access, costs, outcomes, and value. She has extensive experience working with data from administrative claims, electronic medical records, surveys, registries, and clinical trials. She co-authored *Economic Evaluation in Clinical Trials* (Oxford University Press), the first book dedicated entirely to this topic. Her research has received widespread attention from the media including *The New York Times* and *The Wall Street Journal* and has informed policies of private insurers and government programs. Her work has been recognized by numerous prestigious awards from multiple national and international organizations. She currently serves as an Associate Editor of the *Health Economics* journal. Dr. Doshi serves on the Board of Directors of the International Society for Pharmacoeconomics and Outcomes Research.

Cheryl Giscombe is currently the Levine Family Associate Professor of Quality of Life, Health Promotion and Wellness, at the University of North Carolina (UNC) at Chapel Hill School of Nursing. Dr. Giscombe is an expert in psychiatry and clinical psychology, specifically relevant to the diagnosis and treatment of posttraumatic stress disorder and mental, social, and psychosocial functions and health. She has extensive experience working with Veterans with mental health disorders and offers a unique and necessary nursing prospective. For the past 15 years, her research has received consistent federal and foundation funding from the National Institutes of Health (NIH), the Substance Abuse and Mental Health Services Administration, the Health Resources and Services Administration, the Robert Wood Johnson Foundation, and the Josiah Macy Jr. Foundation. Dr. Giscombe's research incorporates sociohistorical and biopsychosocial perspectives to investigate how stress and coping strategies contribute to stress-related psychological and physical health outcomes. Dr. Giscombe is dually trained in nursing and psychology. She completed a B.A. in psychology from North Carolina Central University and a B.S.N. from Stony Brook University in New York. She earned her M.A. and Ph.D. in social and health psychology from Stony Brook University and an M.S.N. from the psychiatric-mental health nurse practitioner/clinical nurse specialist program at UNC at Chapel Hill. Dr. Giscombe was selected as a "Leader in the Field" by the American Psychological Association when she was awarded the Carolyn Payton Early Career Award. Dr. Giscombe is also a Harvard Macy Institute Art Museum-Based Health Professions Education Fellow, and she is currently the lead Principal Investigator on an NIH R01 grant to implement a mindfulness-based stress management intervention to reduce chronic illness risk.

Erinn Hade is an Associate Professor in the Departments of Biomedical Informatics and Obstetrics and Gynecology and a Program Leader for Population Studies in the Center for Biostatistics at The Ohio State University. Dr. Hade is a biostatistician expert with experience in population studies and multi-center trials. Dr. Hade received her Ph.D. in public health and biostatistics from The Ohio State University and her M.S. from the University of Washington. During her academic career, she has developed a strong collaborative record as a biostatistician and maintained a line of primary research that complements the methodologic challenges in these collaborations. Dr. Hade's methodologic interests include work in the design and inference from randomized and non-randomized population-based trials, the design of pragmatic trials, and in statistical methods for adaptive trial designs.

Stewart Hilliard is an expert on breeding, psychometric testing, and the training and development of working dogs, especially military working dogs. Dr. Hilliard is the author of a number of scientific papers in reputable psychological journals. In 1983, Dr. Hilliard received his B.A. in psychology and in 1997 he received his Ph.D. in behavioral neuroscience. Dr. Hilliard has worked in various research and development and operational capacities for the Department of Defense Military Working Dog program since 1997, including as the Chief of the Military Working Dog Course at the 341st Training Squadron, Lackland Air Force Base, Texas. In this role, he managed the basic training of most of the patrol and substance detector dogs supplied to all branches of the U.S. armed forces. Dr. Hilliard also served as the Chief of Military Working Dog Logistics and Procurement at the 341st. He directed the testing and procurement of all of the dogs purchased for the Military Working Dog Course and the Specialized Search Dog Course. Dr. Hilliard is currently the Chief of the Military Working Dog Breeding Program at the 341st Training Squadron.

Evan Mayo-Wilson investigates methods for conducting, reporting, and synthesizing health and social intervention research. His primary area of interest is in ways to increase transparency and reproducibility, such as trial registration and data sharing. Dr. Mayo-Wilson received his B.A. from Columbia University in psychology; his M.S. from the University of Oxford Department of Social Policy and Intervention; his M.P.A. from the University of Pennsylvania; and his D.Phil. (Ph.D.) in philosophy from the University of Oxford Department of Social Policy and Intervention. He is currently an Associate Professor in the Department of Epidemiology and Biostatistics at the School of Public Health at Indiana University

Bloomington. Prior to his teaching career, Dr. Mayo-Wilson worked as an Associate Scientist in the Department of Epidemiology at the Johns Hopkins Bloomberg School of Public Health.

Lori Zoellner is a Professor in the Department of Psychology at the University of Washington. Dr. Zoellner has a long track record in psychotherapy and pharmacotherapy clinical trials and considerable experience in treating posttraumatic stress disorder (PTSD) and training others in its treatment. An important part of these clinical trials is the understanding of underlying therapeutic process mechanisms and the examination of biopsychosocial factors, phenotypic markers of PTSD, and treatment response. Dr. Zoellner's work throughout her career has included the integration across animal to human models of understanding stress, anxiety, and fear. Her current work is characterized by examining memory processing, fear conditioning, fear generalization, avoidance, reward learning, and extinction learning and how these processes map onto prevention and treatment, especially using cognitive behavioral therapy for PTSD. Prior to her clinical research, Dr. Zoellner received her B.A. in psychology and sociology from Rice University and her M.A. and Ph.D. in clinical psychology from the University of California, Los Angeles.

Appendix B

Acronyms and Abbreviations

ADA	Americans with Disabilities Act
ADL	activity of daily living
ANCOVA	analysis of covariance
C-SSRS	Columbia-Suicide Severity Rating Scale
CAPS-5	Clinician Administered PTSD Scale for DSM-5
CBARQ	Canine Behavioral Assessment and Research Questionnaire
CI	confidence interval
CONSORT	Consolidated Standards of Reporting Trials
CONSORT-SPI	CONSORT Extension for Social and Psychological Intervention
DSM-5	*Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition*
DSMC	Data and Safety Monitoring Committee
EMOT	emotional support dog
ESA	emotional support animal
GLM	generalized linear model
IACUC	Institutional Animal Care and Use Committee
ICH	International Council for Harmonisation
IRB	institutional review board
ITT	intent to treat
MedDRA	Medical Dictionary for Regulatory Activities
NDAA	National Defense Authorization Act
PCL-5	self-reported PTSD
PP	per-protocol population
PTSD	posttraumatic stress disorder
RCT	randomized clinical trial
SERV	service dog
SOW	Statement of Work
SPIRIT	Standard Protocol Items: Recommendations for Interventional Trials
VA	Department of Veterans Affairs
VR-12	Veterans RAND 12 Item Health Survey

WHODAS 2.0	World Health Organization Disability Assessment Schedule 2.0
WPAI	Work Productivity and Activity Impairment

Appendix C

Documents Reviewed by the Committee

In order to complete its review, the committee was provided with several documents that are listed below and are available through the National Academies' Public Access File. In order to review these documents, please email paro@nas.edu for more information.

1. The draft monograph titled Performance and Results of Post Traumatic Stress Disorder-Service Dog Study.
2. A link to the Department of Veterans Affairs (VA's) website that houses:
    a. Contract of Statement of Work for procuring dogs for the study.
    b. The complete set of study forms used to conduct the study.
    c. The dog care course and assessment taken by Veterans prior to receiving a dog.
3. The study protocol.
4. Three additional text documents providing written responses to committee questions following the public meeting.
5. The presentation provided by the VA at the public meeting.