International Journal of Methods in Psychiatric Research
Int. J. Methods Psychiatr. Res. 22(4): 267–275 (2013)
Published online in Wiley Online Library
(wileyonlinelibrary.com) **DOI:** 10.1002/mpr.1401

Design of the Army Study to Assess Risk and Resilience in Servicemembers (Army STARRS)

RONALD C. KESSLER,[1] LISA J. COLPE,[2] CAROL S. FULLERTON,[3] NANCY GEBLER,[4] JAMES A. NAIFEH,[3] MATTHEW K. NOCK,[5] NANCY A. SAMPSON,[1] MICHAEL SCHOENBAUM,[2] ALAN M. ZASLAVSKY,[1] MURRAY B. STEIN,[6,7] ROBERT J. URSANO[3] & STEVEN G. HEERINGA[4]

1 Department of Health Care Policy, Harvard Medical School, Boston, MA, USA
2 National Institute of Mental Health, Bethesda, MD, USA
3 Center for the Study of Traumatic Stress, Department of Psychiatry, Uniformed Services, University School of Medicine, Bethesda, MD, USA
4 Institute for Social Research, University of Michigan, Ann Arbor, MI, USA
5 Department of Psychology, Harvard University, Cambridge, MA, USA
6 Departments of Psychiatry and Family and Preventive Medicine, University of California San Diego, La Jolla, CA, USA
7 VA San Diego Healthcare System, San Diego, CA, USA

Key words
Suicide, mental disorders, US Army, epidemiologic research design, design effects, sample bias, sample weights, survey design efficiency, survey sampling

Correspondence
Ronald C. Kessler, Department of Health Care Policy, Harvard Medical School, Boston, MA, USA Telephone (+1) 617-432-3587 Fax (+1) 617-432-3588 Email: NCS@hcp.med.harvard.edu

Received 10 July 2013; accepted 15 July 2013

Abstract

The Army Study to Assess Risk and Resilience in Servicemembers (Army STARRS) is a multi-component epidemiological and neurobiological study designed to generate actionable evidence-based recommendations to reduce US Army suicides and increase basic knowledge about the determinants of suicidality. This report presents an overview of the designs of the six components of the Army STARRS. These include: an integrated analysis of the Historical Administrative Data Study (HADS) designed to provide data on significant administrative predictors of suicides among the more than 1.6 million soldiers on active duty in 2004–2009; retrospective case-control studies of suicide attempts and fatalities; separate large-scale cross-sectional studies of new soldiers (i.e. those just beginning Basic Combat Training [BCT], who completed self-administered questionnaires [SAQs] and neurocognitive tests and provided blood samples) and soldiers exclusive of those in BCT (who completed SAQs); a pre-post deployment study of soldiers in three Brigade Combat Teams about to deploy to Afghanistan (who completed SAQs and provided blood samples) followed multiple times after returning from deployment; and a platform for following up Army STARRS participants who have returned to civilian life. Department of Defense/Army administrative data records are linked with SAQ data to examine prospective associations between self-reports and subsequent suicidality. The presentation closes with a discussion of the methodological advantages of cross-component coordination.

Introduction

Suicide is the second leading cause of death among 25 to 44 year olds in the United States, claiming over 30,000 lives annually (Centers for Disease Control, 2005). For every completed suicide there are an estimated eight to 25 failed attempts (Goldsmith *et al.*, 2002). In 1999 the US Surgeon General issued a "call to action" on suicide prevention, saying "the nation must address suicide as a significant problem," and recommending enhanced "research to understand risk and protective factors." The US military also identified suicide as a major concern at around the same time and created several initiatives, including a successful Air Force Suicide Prevention Program (afspp.afms.mil). However, in contrast to the suicide rate in the United States remaining fairly level over the recent past, the suicide rate in the US Army doubled between 2003 and 2008 and has continued to rise since then.

Historically, the suicide rate in the US military has been below that in the civilian population, but it has climbed steadily since the beginning of the Iraq and Afghanistan conflicts to the point where suicide is now the second leading cause of death behind combat deaths (Armed Forces Health Surveillance Center, 2012) and has exceeded demographically matched civilian rates since 2008 (Kuehn, 2009). In fiscal year 2009, there were 160 recorded suicides in the Army. Of those, 79% were among soldiers who had deployed only once or had not deployed at all. Additionally, 60% of suicides were among first-term soldiers (http://www.army.mil/article/43038/army-releases-report-on-suicide-high-risk-behavior/). The rise in US Army suicides has persisted despite substantial efforts to publicize and encourage use of mental health services. Although many intervention programs are underway, success will require a better understanding of specific risk and protective factors in Army service. The profile associated with traditional individual risk factors (e.g. age, gender, presence of mental illness) may not generalize to Army personnel, all of whom are employed, selected for good health, and have health care available at no cost. Furthermore, the risk profile among military personnel might vary during different phases of duty and mission. Changing accession demographics (e.g. the number of new recruits with General Educational Development (GED) versus high school graduation, the number with conduct waivers) may also affect suicide risk, as suggested in a recent Institute of Medicine (IOM) report (Committee on Youth, 2006). Changes in barriers to care, both physical barriers (e.g. time and access to care) and community barriers (e.g. stigma, operational tempo, unit cohesion, leadership support), are also possible factors. More intangible risk and protective factors associated with war, such as increased feelings of patriotism and loyalty to one's unit might also be involved. Textured research on these and related issues is needed to identify modifiable risk and protective factors for suicidal behaviors and inform effective suicide risk and prevention strategies among Army servicemen and women.

The Department of the Army responded to these trends in 2008 by entering into an agreement with the National Institute of Mental Health (NIMH) to fund the Army Study to Assess Risk and Resilience in Servicemembers (Army STARRS; http://www.armystarrs.org), a multicomponent epidemiological and neurobiological study of risk and resilience factors for suicidality and its psychopathological correlates among Army personnel (Insel and McHugh,). The two overarching goals of Army STARRS are: to evaluate hypotheses about modifiable risk and resilience factors for suicidality that could be used to target effective preventive interventions for Army suicides; and to expand basic scientific understanding of psychosocial and neurobiological risk and resilience factors for suicidal behaviors and their psychopathological correlates. The Army STARRS samples could also be used as baselines for intervention implementations and evaluations in the future.

Army STARRS is supported under a Cooperative Agreement (U01) between NIMH and a consortia of scientific collaborators at the Uniformed Services University of the Health Sciences (USUHS; PI: Robert Ursano), the University of California San Diego (PI: Murray Stein), Harvard Medical School (Site PI: Ronald Kessler), and the University of Michigan (Site PI: Steven Heeringa) through the Henry M. Jackson Foundation. Additional U01 collaborating scientists and consultants come from the NIMH (Lisa Colpe; Michael Schoenbaum) and the Army (Kenneth Cox; Steven Cersovsky). The Army STARRS includes a number of coordinated component studies designed to facilitate non-experimental hypothesis generation and testing, intervention targeting, and intervention evaluation.

The first of these initiated by Army STARRS collaborators examined historical data in an integrated data system created by combining information obtained from a number of Army and Department of Defense (DoD) administrative databases on all soldiers who served in the Army between 2004 and 2009. Retrospective case-control studies of fatal suicides and non-fatal suicide attempts were then designed to provide preliminary quantitative data on risk and resilience factors as well as to create an opportunity to obtain qualitative data to help generate new hypotheses. A series of major multi-mode self-report surveys were then launched that included neurocognitive and genetic data linked both retrospectively and prospectively to

administrative data systems so as to examine patterns of association over time between risk-resilience factors and subsequent suicidal behaviors. Targeted sub-studies of high-risk soldiers and settings are now being planned that use these survey samples as sampling frames. Early component studies were designed to provide input into later component studies and subsequent interventions.

The current report presents a broad overview of the designs of these component Army STARRS studies. These studies were designed to create a coordinated whole to facilitate non-experimental hypothesis generation and testing, intervention targeting, and intervention evaluation. The individual study designs are for the most part conventional, but their coordination creates unique strengths. We discuss the synergistic effects of cross-component coordination after describing the individual studies.

Army STARRS component studies

The Historical Administrative Data Study (HADS)

The Army and DoD maintain over 200 different administrative data systems dealing with such diverse issues as certifications of training (Army Training and Requirements Resource System [ATRRS]), medical records (the Medical Data Repository [MDR] system), casualty reporting (Defense Manpower Data Center [DMDC/CASUALTY]) , and, importantly for our purposes, suicidal behaviors (DoD Suicide Event Report [DoDSER] system). While prior to the initiation of Army STARRS special-purpose efforts had been made to integrate some of these data systems, we felt that a great deal more could be learned about risk and resilience factors for suicides by linking the DoDSER with some of the other systems. We consequently established the Army STARRS Historical Administrative Data Study (HADS), an integrated administrative data file containing key elements from 38 different Army and DoD data systems for the over 1.6 million soldiers (Regular Army, Army Reserve, and National Guard) on active duty at some time during calendar years 2004–2009. The Army subsequently expanded the integrated data files to be continually updated for purposes of use in targeting future interventions. Analyses of the 2004–2009 HADS data are allowing Army STARRS collaborators to examine time trends in suicides, other types of deaths, and non-fatal injuries (suicidal and others) as well as to study a wide range of predictors of those outcomes. These analyses are being carried out for the most part using discrete-time survival analyses (Singer and Willett, 2003) with person-month the unit of analysis based on the roughly 51.1 million person-months in this data array (37.0 million Regular Army, 5.3 million activated Army Reserve, 8.8 million activated Army National Guard). Given the rarity of the outcomes under study, we are analyzing reduced samples consisting of all person-months with the outcomes of interest and a probability sub-sample of control person-months weighted by the inverse of their probability of selection and using a logistic link function to estimate coefficients.

In addition, individual-level data from the Army/DoD administrative data systems used to build the HADS are being linked to Army STARRS surveys for all consenting participants. Administrative data linked to the Army STARRS survey samples described later are allowing information from retrospective (to the surveys) administrative data, survey data, and, for some samples, neurocognitive and genetic data to be integrated to predict subsequent (prospective) outcomes identifiable in administrative records.

Soldier Health Outcomes Studies A and B (SHOS-A/B)

The Soldier Health Outcome Studies (SHOS) are retrospective case-control studies of soldiers who made non-fatal suicide attempts (SHOS-A) or were suicide fatalities (SHOS-B). While producing less definitive data than prospective studies on the predictors of suicidal behaviors, retrospective case-control studies like SHOS-A/B are useful because they provide rapid preliminary data on potentially important risk and resilience factors that can subsequently be evaluated more definitively in prospective naturalistic and intervention studies (Schlesselman, 1982).

SHOS-A cases are recruited from all patients in psychiatric inpatient units in five participating tertiary care medical facilities (Walter Reed National Military Medical Center, Washington, DC; Fort Bragg, NC; Fort Stewart, GA; Fort Lewis, WA; and Fort Hood, TX) who were admitted because of suicide attempts beginning November 2011. Cases provide written informed consent and then complete the same self-report survey as in our major survey of Army personnel (the All-Army Study [AAS], which is described later in this report), allowing case-control analyses to be carried out using all respondents from that survey as controls. In addition, a propensity score weight (Rosenbaum and Rubin, 1983) was developed based on case-control analysis of the first SHOS-A cases compared to AAS respondents to select a sub-sample of Regular Army AAS respondents as controls for more in-depth assessment, including an expanded version of the neurocognitive test battery used in the New Soldier Study (NSS) and collection of blood samples. Qualitative interviews based on the principles of *reason analysis* (Strauss, 1987) are also being administered to SHOS-A cases in an effort to uncover information about critical junctures in the progression to

attempts. We anticipate a final SHOS-A sample of 150 cases and 300 group-matched controls.

SHOS-B cases are selected by attempting to interview the next of kin and Army supervisors of all soldiers who committed suicide (as recorded in the DoDSER system) beginning March 2012 plus a group-matched sample of controls. As with SHOS-A, controls are being selected from Regular Army AAS respondents based on a propensity score weight (Rosenbaum and Rubin, 1983), but in this case the weight was developed based on analysis of the HADS (total Army) sample using predictors of suicide found in administrative records. As in SHOS-A, a qualitative component is included in SHOS-B to help uncover information about critical junctures in the progression to suicide. Hypotheses generated from analysis of these qualitative data are being evaluated prospectively whenever possible by expanding the assessments in the NSS and AAS. We anticipate a final SHOS-B sample of 150 cases and 300 group-matched controls.

The New Soldier Study (NSS)

Between 140,000 and 190,000 new soldiers enter the US Army each year, including those in the Regular Army (about half of enlistees), the US Army Reserve (USAR), and the US Army National Guard (USANG). The majority of these new soldiers begin their active duty by going through three months of Basic Combat Training (BCT), the main exceptions being those who enter the Army as officers and those who enter the USAR or USANG after leaving the Regular Army. One of the Army STARRS component studies is the NSS, a study that attempted to assess over 57,000 of these new soldiers in the two days after their arrival to report for BCT. NSS respondents were selected from three Army installations that provide BCT (Fort Benning, GA; Fort Jackson, SC; and Fort Leonard Wood, MO) with sample sizes proportional to the relative sizes of the cohorts across these sites. Continuous sampling throughout calendar years 2011–2012 was used to account for the fact that the composition of new soldiers changes across the year due to an influx of recent high school graduates in the summer and fall.

The NSS included group administration of a two-part self-administered questionnaire (SAQ) and neurocognitive tests along with the collection of blood samples obtained as part of a physical examination given to all new soldiers prior to the beginning of BCT. The contents of the neurocognitive tests are described in a series of reports in preparation. All these data were collected in Reception Battalion (RECBN), the period typically lasting several days when new soldiers are processed (physical exams; immunizations; eye exams; issuance of uniforms and, as needed, special stress-resistant eyeglasses; completion of various forms) before beginning BCT. A new RECBN cohort typically enters on the same day each week throughout the year, but breaks for holidays.

Army STARRS was assigned an auditorium that could hold between 200 and 300 people using laptops for SAQ self-administration at the RECBN site on each BCT installation. Each week, the Army Point-Of-Contact (POC) at each RECBN site selected a sample of new soldiers equal to the number of seats in the auditorium to attend the Army STARRS informed consent session. Army STARRS worked with the POCs to prevent systematic bias in selection procedures. Thirty-minute informed consent sessions explained study purposes, procedures, and protections against breach of confidentiality (separation of identifying information from survey and neurocognitive data; assignment of a study ID with linkage to identifying information maintained securely by the University of Michigan data collection team; contractual agreement that identifying information will not be provided to the Army; protocols to report only aggregate results and coarsen public use datasets to minimize risk of individual identification), emphasize the voluntary nature of participation (including the right to withdraw consent at a future date), and answer questions before seeking informed consent.

Written informed consent was then obtained from volunteers. Although not necessary for SAQ participation, all NSS respondents were additionally asked for consent to link their Army/DoD administrative records to their NSS responses and to participate in to-be-determined longitudinal follow-up data collections. Identifying information (name, birthday, Social Security number (SSN) for record linkage; telephone number, email, secondary contact information for longitudinal follow-up) was collected from consenting respondents separately from the SAQ and never merged with de-identified NSS data. These recruitment, consent, and data protection procedures were approved by the Human Subjects Committees of the Uniformed Services University of the Health Sciences for the Henry M. Jackson Foundation (the primary grantee), the Institute for Social Research at the University of Michigan (the organization implementing STARRS data collection), and all other collaborating organizations.

Blood samples were then obtained from volunteers as part of the RECBN physical examination. This is in addition to the normal blood samples taken from new soldiers during the reception phase of BCT. New soldiers participating in Army STARRS and consenting to provide a blood sample had one additional vial of blood drawn for Army STARRS. SAQ and neurocognitive test data were obtained in two 90-minute group-administered data

collection sessions over two successive days. Individual-level Army/DoD administrative data were subsequently linked to the de-identified individual-level NSS data for the subsample of NSS participants who provided written informed consent for this linkage. Sample sizes, response rates, weighting, and design effects are described in a separate report (Kessler *et al.*, 2013).

The All-Army Study (AAS)

The NSS focuses on a very small proportion of active duty soldiers; those about to enter BCT. Between 737,000 and 757,000 US Army soldiers have been on active duty at any point in time since the inception of Army STARRS data collection in early 2011 (576,000–581,000 Regular Army and 161,000–176,000 activated USAR or USANG), with only 8400–12,200 of these soldiers in BCT at any point in time during the same period. The AAS is a cross-sectional SAQ survey carried out throughout 2011–2012 in quarterly samples of active duty Army personnel exclusive of those in BCT or in Afghanistan. Unlike NSS, no neurocognitive test data or genetic data are collected in the AAS. In addition, the AAS SAQ is considerably shorter than the NSS SAQ (one 90-minute administration session in AAS compared to two in NSS).

The quarterly AAS replicates in Q1–2 2011 were representative of all soldiers stationed in the continental United States, while those in the remaining quarters added soldiers stationed elsewhere in the world other than a combat theater. Each quarterly replicate consisted of a stratified (by Army Command and location) probability sample of Army units (or, for large units, sub-units) selected without replacement with probabilities proportional to authorized unit strength from a sample frame of Unit Identification Codes to yield a representative time-space sample. The frame excluded civilian-only units and units of fewer than 30 soldiers (representing less than 2% of all Army personnel).

In addition to these quarterly replicates, the AAS was augmented to increase coverage in two ways. First, it was administered in Q2–3 2012 to a probability sample of soldiers stationed in Afghanistan who were surveyed in group-administered sessions while they were passing through Kuwait either leaving for or returning from their mid-tour leave. Unlike other AAS respondents, those stationed in Afghanistan were surveyed as individuals rather than as units, as individuals leave their units for mid-tour leave. It should be noted that mid-tour leave, despite the name, did not all occur in the middle of the deployed soldier's tour in Afghanistan but rather at random times between the second and second to last months of deployment, guaranteeing that the AAS was administered at relatively random times other than at the very beginning or very end of deployment.

Second, the AAS was administered in 2012–2013 to a supplemental sample of activated USAR and USANG units in the continental United States either just before or just after deployment to Afghanistan in order to correct for the fact that these kinds of units were excluded from the Regular Army AAS sample frame.

All personnel in each selected AAS unit (or sub-unit) were targeted for a group-administrated SAQ survey and were ordered to report to a group informed consent session similar to those described earlier for the NSS. The AAS respondents stationed in Afghanistan, in comparison, were selected using methods similar to those in the NSS (i.e. POCs selected a number of soldiers equal to the number that could be accommodated by the Army STARRS group administration setting to participate in the informed consent session). In each case, an informed consent presentation similar to that in NSS was group-administered to explain study purposes, procedures, and confidentiality protections, emphasize the voluntary nature of participation, and answer questions. Written informed consent was then obtained. As in the NSS, AAS respondents were additionally asked for consent to link their Army/DoD administrative records to their survey responses and to participate in to-be-determined longitudinal follow-up data collections. Identifying information (name, birthday, SSN for record linkage; telephone number, email, secondary contact information for longitudinal follow-up) was collected from consenting respondents separately from the survey and never merged with de-identified survey data. As with the NSS, these recruitment, consent, and data protection procedures were approved by the Human Subjects Committees of all collaborating organizations. Sample sizes, response rates, weighting, and design effects are described in a separate report (Kessler *et al.*, 2013).

The Pre-Post Deployment Study (PPDS)

NSS and AAS respondents are being tracked longitudinally through their administrative records, however outcomes are limited to those that are administratively recorded, such as suicide fatalities, non-fatal suicide attempts sufficiently severe to come to the attention of the military healthcare system, and mental disorders treated in the military healthcare system. A great many outcomes of interest will be missed by these administrative records, such as suicidal ideation and plans and onsets of mental disorders known to be powerful risk factors for suicides. In order to

address this problem, a number of targeted follow-up surveys are planned in the future; two of these are being implemented currently. The larger of these two is the Pre-Post Deployment Study (PPDS), a four-wave panel survey that collected baseline data (SAQ and blood samples) in Q1 2012 shortly before deployment to Afghanistan from 9421 soldiers in three Brigade Combat Teams. Follow-up data collections are scheduled for these same respondents three times after they return from deployment: within one month of their return (T1; SAQ and blood samples), two months after this first post-return assessment (T2; SAQ), and six months after the second post-return assessment (T3; web-based SAQ augmented with telephone interviews for SAQ non-respondents).

As with the AAS, all personnel in each selected PPDS unit were targeted for a group-administrated baseline SAQ and were ordered to report to a group informed consent session similar to the AAS to explain study purposes, procedures, and confidentiality protections, emphasize the voluntary nature of participation, and answer questions. Written informed consent was then obtained. As in the NSS, baseline PPDS respondents were additionally asked for consent to provide blood samples, to link their Army/ DoD administrative records to their survey responses, and to participate in future assessments. Identifying information was collected from consenting respondents separately from the survey and never merged with de-identified survey data. Similar informed consent procedures were used in the post-deployment data collections. These PPDS recruitment, consent, and data protection procedures were approved by the Human Subjects Committees of all collaborating organizations. Sample sizes, response rates, weighting, and design effects are described in a separate report (Kessler *et al.*, 2013).

The Pre-Post Separation Study (PPSS) platform

We noted earlier in the section on the NSS that between 140,000 and 190,000 new soldiers enter the US Army each year and that roughly half of these new soldiers are in the Regular Army. The other half are in the USAR and USANG. The number of Regular Army soldiers that leave the Army and return to civilian life each year is roughly comparable to the number that joins. The number of active duty soldiers that do not reenlist is likely to increase over the next few years as the Army downsizes due to the ends of the conflicts in Afghanistan and Iraq. Because this downsizing is occurring at a time of economic uncertainty in the civilian economy, the number of soldiers leaving the Army involuntarily (i.e. they want to reenlist but are not given an opportunity to do so) will increase as well. This will doubtlessly add to the stresses known to accompany the transition from military to civilian life (Wolpert, 2000; Hoge, 2010).

The T3 PPDS includes a number of questions about the transition from Army to civilian life, as we expect that as many as 1500 baseline PPDS respondents will have returned to civilian life at the time of the T3 survey. While few respondents from the NSS will have returned to civilian life by that time, a substantial number of AAS respondents have already done so and this number will increase over time. Concerns exist about the mental health of recently separated veterans (Ilgen *et al.*, 2012; Conner *et al.*, 2013). Indeed, a recent analysis of the US National Health Interview Survey for 1986 to 1994 found that even before the recent rise in the military suicide rate veterans were twice as likely to die by suicide as socio-demographically comparable non-veterans (Kaplan *et al.*, 2007). However, the vast majority of research on the mental health of veterans is based on analyses of Veterans Affairs (VA) treatment samples rather than on more broadly representative prospective epidemiological samples that follow soldiers through the transition from military to civilian life (Rosenheck and Fontana, 2007; Naragon-Gainey *et al.*, 2012). Based on this observation, we included a Pre-Post Separation Study (PPSS) component in Army STARRS. The PPSS is only included as a pilot during the first Army STARRS funding cycle in light of the fact that the number of AAS, PPDS, and especially NSS study respondents who return to civilian life during the first five years of Army STARRS funding will be comparatively small. The mixed-mode web-telephone survey design used in the T3 PPDS post-return survey will be applied in the PPSS. We think of the PPSS as a pilot for a much more ambitious program of long-term follow-up of the full Army STARRS sample in the coming years.

Coordination among component studies

As noted in the Introduction, the component Army STARRS studies were designed to create a coordinated whole that would facilitate hypothesis generation, non-experimental hypothesis testing, targeting of interventions designed to provide more definitive experimental tests of hypotheses about modifiable risk and resilience factors, and evaluations of such intervention. The advantages of coordination can be seen clearly in SHOS-A/B. Much previous research has been carried out using the retrospective case-control design to study risk and resilience factors for suicide (Cavanagh *et al.*, 2003; Dumais *et al.*, 2005) and non-fatal suicidal behaviors (Nock *et al.*, 2010; Bridge *et al.*, 2012), and the designs of Army STARRS

Int. J. Methods Psychiatr. Res. 22(4): 267–275 (2013). DOI: 10.1002/mpr

SHOS-A/B are consistent with those previous studies. However, the integration of SHOS-A/B into the larger Army STARRS initiative creates a unique opportunity to address the critical weakness of suboptimal control group selection that has plagued previous case-control studies in this area. Specifically, while it is well known that conclusions about risk and resilience factors depend critically on the control group selected in case-control studies (Schlesselman, 1982) and this sensitivity has been documented in previous analyses of suicide case-control research that varied the control groups (Brent *et al.*, 1988; Brent *et al.*, 1993), previous case-control studies of suicidal behaviors typically used control groups made up of healthy people randomly selected from the general population (but not matched on widely known risk factors such as the presence of a mental disorder) or convenience samples of psychiatric patients selected in an effort to control for the presence of a mental disorder (but not representative of the broader population of those at risk, many of whom commit suicide without ever seeking professional treatment for their emotional problems). As noted in the body of the paper, this weakness was addressed in SHOS-A/B by using propensity score matching methods (Rosenbaum and Rubin, 1983) to select probability samples of soldiers from the AAS as controls with an over-sampling of AAS respondents who reported suicidal ideation. This design refinement, which allows much more sensitive comparisons of cases and controls than in previous case-control studies of suicidal behaviors (Li *et al.*, 2011), would have been impossible in the absence of the HADS and AAS studies being carried out in parallel with SHOS-A/B.

In a similar way, the three large-scale Army STARRS SAQ data collections (NSS, AAS, PPDS) are enriched by being linked with administrative data and blood samples collected by Army phlebotomists. Prospective administrative data are of special importance in this regard, as they are allowing a wide range of analyses to be carried out that would otherwise have been impossible with cross-sectional survey data. For example, prospective administrative data linked to the NSS are currently being used to examine the extent to which an assessment of risk and resilience factors carried out at the very beginning of active duty can pinpoint new soldiers at elevated risk of suicidal behaviors and other serious adverse outcomes (e.g. serious injuries, victimization or perpetration of violent offenses) during the first two years of their Army service.

Discussion

This report has presented a brief overview of the designs of the component studies within the Army STARRS initiative. While this overview makes it clear that each of these component studies is a substantial undertaking in its own right, it is also important to recognize that all the component studies are strengthened by virtue of their integration with the others. This is true not only in that data elements are shared across component studies but also because the full set of studies taken together allows hypothesis generation and non-experimental testing to be combined with intervention targeting and, in time, intervention evaluation to address the full range of research questions raised in grappling with the problem of Army suicides. While this broad scope and integration do not guarantee that Army STARRS will succeed in achieving its substantive goals, they create a strong foundation for doing so.

Acknowledgments

On behalf of the Army STARRS Collaborators

Funding/Support

Army STARRS was sponsored by the Department of the Army and funded under cooperative agreement number U01MH087981 with the US Department of Health and Human Services, National Institutes of Health, National Institute of Mental Health (NIH/NIMH). The contents are solely the responsibility of the authors and do not necessarily represent the views of the Department of Health and Human Services, NIMH, the Department of the Army, or the Department of Defense.

Role of the Sponsors

As a cooperative agreement, scientists employed by NIMH (Colpe and Schoenbaum) and Army liaisons/consultants (COL Steven Cersovsky, MD, MPH USAPHC and Kenneth Cox, MD, MPH USAPHC) collaborated to develop the study protocol and data collection instruments, supervise data collection, plan and supervise data analyses, interpret results, and prepare reports. Although a draft of this manuscript was submitted to the Army and NIMH for review and comment prior to submission, this was with the understanding that comments would be no more than advisory.

Additional Contributions

The Army STARRS Team consists of Co-Principal Investigators: Robert J. Ursano, MD (Uniformed Services University of the Health Sciences) and Murray B. Stein, MD, MPH (University of California San Diego and VA San Diego Healthcare System); Site Principal Investigators: Steven Heeringa, PhD (University of Michigan) and Ronald C. Kessler, PhD (Harvard Medical School); NIMH

collaborating scientists: Lisa J. Colpe, PhD, MPH and Michael Schoenbaum, PhD; Army liaisons/consultants: COL Steven Cersovsky, MD, MPH (USAPHC) and Kenneth Cox, MD, MPH (USAPHC). Other team members: Pablo A. Aliaga, MA (Uniformed Services University of the Health Sciences); COL David M. Benedek, MD (Uniformed Services University of the Health Sciences); Susan Borja, PhD (National Institute of Mental Health); Gregory G. Brown, PhD (University of California San Diego); Laura Campbell-Sills, PhD (University of California San Diego); Catherine Dempsey, PhD, MPH (Uniformed Services University of the Health Sciences); Richard Frank, PhD (Harvard Medical School); Carol S. Fullerton, PhD (Uniformed Services University of the Health Sciences); Nancy Gebler, MA (University of Michigan); Joel Gelernter, MD (Yale University); Robert K. Gifford, PhD (Uniformed Services University of the Health Sciences); Stephen E. Gilman, ScD (Harvard School of Public Health); Marjan G. Holloway, PhD (Uniformed Services University of the Health Sciences); Paul E. Hurwitz, MPH (Uniformed Services University of the Health Sciences); Sonia Jain, PhD (University of California San Diego); Tzu-Cheg Kao, PhD (Uniformed Services University of the Health Sciences); Karestan C. Koenen, PhD (Columbia University); Lisa Lewandowski-Romps, PhD (University of Michigan); Holly Herberman Mash, PhD (Uniformed Services University of the Health Sciences); James E. McCarroll, PhD, MPH (Uniformed Services University of the Health Sciences); Katie A. McLaughlin, PhD (Harvard Medical School); James A. Naifeh, PhD (Uniformed Services University of the Health Sciences); Matthew K. Nock, PhD (Harvard University); Rema Raman, PhD (University of California San Diego); Nancy A. Sampson, BA (Harvard Medical School); LCDR Patcho Santiago, MD, MPH (Uniformed Services University of the Health Sciences); Michaelle Scanlon, MBA (National Institute of Mental Health); Jordan Smoller, MD, ScD (Harvard Medical School); Nadia Solovieff, PhD (Harvard Medical School); Michael L. Thomas, PhD (University of California San Diego); Christina Wassel, PhD (University of Pittsburgh); and Alan M. Zaslavsky, PhD (Harvard Medical School). The authors would also like to thank John Mann, Maria Oquendo, Barbara Stanley, Kelly Posner, and John Keilp for their contributions to the early stages of Army STARRS development.

Additional Information

A complete list of Army STARRS publications can be found at http://www.ARMYSTARRS.org.

Declaration of interest statement

In the past five years Kessler has been a consultant for Eli Lilly & Company, Glaxo, Inc., Integrated Benefits Institute, Ortho-McNeil Janssen Scientific Affairs, Pfizer Inc., Sanofi-Aventis Groupe, Shire US Inc., and Transcept Pharmaceuticals Inc. and has served on advisory boards for Johnson & Johnson. Kessler has had research support for his epidemiological studies over this time period from Eli Lilly & Company, EPI-Q, GlaxoSmithKline, Ortho-McNeil Janssen Scientific Affairs, Sanofi-Aventis Groupe, Shire US, Inc., and Walgreens Co. Kessler owns a 25% share in DataStat, Inc. Stein has in the last three years been a consultant for Healthcare Management Technologies and had research support for pharmacological imaging studies from Janssen. The remaining authors report no competing interests.

References

Armed Forces Health Surveillance Center (2012) Deaths while on active duty in the U.S. Armed Forces, 1990–2011. *Medical Surveillance Monthly Reports*, **19**(5), 2–5.

Brent D.A., Perper J.A., Goldstein C.E., Kolko D.J., Allan M.J., Allman C.J., Zelenak J.P. (1988) Risk factors for adolescent suicide. A comparison of adolescent suicide victims with suicidal inpatients. *Archives of General Psychiatry*, **45**(6), 581–588.

Brent D.A., Perper J.A., Moritz G., Allman C., Friend A., Roth C., Schweers J., Balach L., Baugher M. (1993) Psychiatric risk factors for adolescent suicide: a case-control study. *Journal of the American Academy of Child and Adolescent Psychiatry*, **32**(3), 521–529, DOI: 10.1097/00004583-199305000-00006

Bridge J.A., McBee-Strayer S.M., Cannon E.A., Sheftall A.H., Reynolds B., Campo J.V., Pajer K.A., Barbe R.P., Brent D.A. (2012) Impaired decision making in adolescent suicide attempters. *Journal of the American Academy of Child and Adolescent Psychiatry*, **51**(4), 394–403, DOI: 10.1016/j.jaac.2012.01.002

Cavanagh J.T., Carson A.J., Sharpe M., Lawrie S.M. (2003) Psychological autopsy studies of suicide: a systematic review. *Psychological Medicine*, **33**(3), 395–405, DOI: 10.1017/S0033291702006943

Centers for Disease Control (2005) *CDC Reports Latest Data on Suicide Behaviors, Risk Factors, and Prevention*, Atlanta, GA, Centers for Disease Control.

Committee on Youth, Population and Military Recruitment: Physical, Medical, and Mental Health Standards, National Research Council (2006) *Assessing Fitness for Military Enlistment: Physical, Medical, and Mental Health Standards*, Washington: DC, National Academy Press.

Conner K.R., Bohnert A.S., McCarthy J.F., Valenstein M., Bossarte R., Ignacio R., Lu N., Ilgen M.A. (2013) Mental disorder comorbidity and suicide among 2.96 million men receiving care in the veterans health administration health system. *Journal of Abnormal Psychology*, **122**(1), 256–263, DOI: 10.1037/a0030163

Dumais A., Lesage A.D., Alda M., Rouleau G., Dumont M., Chawky N., Roy M., Mann J.J., Benkelfat C., Turecki G. (2005) Risk factors for suicide completion in major depression: a case-control study of impulsive and aggressive behaviors in men. *American Journal of Psychiatry*, **162**(11), 2116–2124, DOI: 10.1176/appi.ajp.162.11.2116

Goldsmith S.K., Pellmar T.C., Kleinman A.M., Bunney W.E. eds. (2002) *Reducing Suicide: A National Imperative*, Washington: DC, National Academy Press.

Hoge C.W. (2010) *Once a Warrior Always a Warrior: Navigating the Transition from Combat to Home, Including Combat Stress, PTSD, and mTBI*, Guilford, CT: Globe Pequot Press.

Ilgen M.A., McCarthy J.F., Ignacio R.V., Bohnert A.S., Valenstein M., Blow F.C., Katz I.R. (2012) Psychopathology, Iraq and Afghanistan service, and suicide among Veterans Health Administration patients. *Journal of Consulting and Clinical Psychology*, **80**(3), 323–330, DOI: 10.1037/a0028266

Insel T.R. McHugh J.M. (submitted for publication) The U.S. Army-National Institute of Mental Health Army Study to Assess the Risk and Resilience in Servicemembers (Army STARRS): rapidly translating interventions to reduce suicide.

Kaplan M.S., Huguet N., McFarland B.H., Newsom J.T. (2007) Suicide among male veterans: a prospective population-based study. *Journal of Epidemiology and Community Health*, **61**(7), 619–624, DOI: 10.1136/jech.2006.054346

Kessler R.C., Heeringa S.G., Colpe L.J., Fullerton C. S., Gebler N., Hwang, I., Naifeh J.A., Nock M. K., Sampson N.A., Schoenbaum M., Zaslavsky A.M., Stein M.B., Ursano R.J. (2013) Response bias, weighting adjustments, and design effects in the Army Study to Assess Risk and Resilience in Servicemembers (Army STARRS). *International Journal of Methods in Psychiatric Research*, **22**(4), 288–302.

Kuehn B.M. (2009) Soldier suicide rates continue to rise: military, scientists work to stem the tide. *Journal of the American Medical Association*, **301**(11), 1111–1113, DOI: 10.1001/jama.2009.342

Li L., Shen C., Wu A.C., Li X. (2011) Propensity score-based sensitivity analysis method for uncontrolled confounding. *American Journal of Epidemiology*, **174**(3), 345–353, DOI: 10.1093/aje/kwr096

Naragon-Gainey K., Hoerster K.D., Malte C.A., Jakupcak M. (2012) Distress symptoms and high-risk behaviors prospectively associated with treatment use among returning veterans. *Psychiatric Services*, **63**(9), 942–944, DOI: 10.1176/appi.ps.201100349

Nock M.K., Park J.M., Finn C.T., Deliberto T.L., Dour H.J., Banaji M.R. (2010) Measuring the suicidal mind: implicit cognition predicts suicidal behavior. *Psychological Science*, **21**(4), 511–517, DOI: 10.1177/0956797610364762

Rosenbaum P.R., Rubin D.B. (1983) The central role of the propensity score in observational studies for causal effects. *Biometrika*, **70**(1), 41–55.

Rosenheck R.A., Fontana A.F. (2007) Recent trends In VA treatment of post-traumatic stress disorder and other mental disorders. *Health Affairs (Millwood)*, **26**(6), 1720–1727, DOI: 10.1377/hlthaff.26.6.1720

Schlesselman J.J. (1982) *Case-control Studies: Design, Conduct, Analysis*, New York: Oxford University Press.

Singer J.D., Willett J.B. (2003) *Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence*, New York: Oxford University Press.

Strauss A.L. (1987) *Qualitative Analysis for Social Scientists*, Cambridge: Cambridge University Press.

Wolpert D.S. (2000) Military retirement and the transition to civilian life. In Martin J. A., Rosen L.N., Sparacino L.R. (eds) *The Military Family: A Practice Guide for Human Services Providers*, pp 103–119, New York, Praeger.

International Journal of Methods in Psychiatric Research
Int. J. Methods Psychiatr. Res. 22(4): 276–287 (2013)
Published online in Wiley Online Library
(wileyonlinelibrary.com) **DOI:** 10.1002/mpr.1400

Field procedures in the Army Study to Assess Risk and Resilience in Servicemembers (Army STARRS)

STEVEN G. HEERINGA,[1] NANCY GEBLER,[1] LISA J. COLPE,[2] CAROL S. FULLERTON,[3] IRVING HWANG,[4] RONALD C. KESSLER,[4] JAMES A. NAIFEH,[3] MATTHEW K. NOCK,[5] NANCY A. SAMPSON,[4] MICHAEL SCHOENBAUM,[2] ALAN M. ZASLAVSKY,[4] MURRAY B. STEIN[6,7] & ROBERT J. URSANO[3]

1 University of Michigan, Institute for Social Research, Ann Arbor, MI, USA
2 National Institute of Mental Health, Bethesda, MD, USA
3 Center for the Study of Traumatic Stress, Department of Psychiatry, Uniformed Services University School of Medicine, Bethesda, MD, USA
4 Department of Health Care Policy, Harvard Medical School, Boston, MA, USA
5 Department of Psychology, Harvard University, Cambridge, MA, USA
6 Departments of Psychiatry and Family and Preventive Medicine, University of California San Diego, La Jolla, CA, USA
7 VA San Diego Healthcare System, San Diego, CA, USA

Key words
suicide, mental disorders, US Army, epidemiologic research design, design effects, sample bias, sample weights, survey design efficiency, survey sampling

Correspondence
Ronald C. Kessler, Department of Health Care Policy, Harvard Medical School, Boston, MA, USA. Telephone (+1) 617-432-3587 Fax (+1) 617-432-3588
Email: NCS@hcp.med.harvard.edu

Received 10 July 2013;
accepted 15 July 2013

Abstract

The Army Study to Assess Risk and Resilience in Servicemembers (Army STARRS) is a multi-component epidemiological and neurobiological study of unprecedented size and complexity designed to generate actionable evidence-based recommendations to reduce US Army suicides and increase basic knowledge about determinants of suicidality by carrying out coordinated component studies. A number of major logistical challenges were faced in implementing these studies. The current report presents an overview of the approaches taken to meet these challenges, with a special focus on the field procedures used to implement the component studies. As detailed in the paper, these challenges were addressed at the onset of the initiative by establishing an Executive Committee, a Data Coordination Center (the Survey Research Center [SRC] at the University of Michigan), and study-specific design and analysis teams that worked with staff on instrumentation and field procedures. SRC staff, in turn, worked with the Office of the Deputy Under Secretary of the Army (ODUSA) and local Army Points of Contact (POCs) to address logistical issues and facilitate data collection. These structures, coupled with careful fieldworker training, supervision, and piloting, contributed to the major Army STARRS data collection efforts having higher response rates than previous large-scale studies of comparable military samples.

Introduction

As described in more detail earlier in this issue (Kessler *et al.*, 2013a) and elsewhere (Ursano *et al.*, submitted for publication), the Army Study to Assess Risk and Resilience in Servicemembers (Army STARRS; http://www.armystarrs.org) is a multi-component epidemiological and neurobiological study of risk and resilience factors for suicidality and its psychopathological correlates in the US Army. Army STARRS is funded as a Cooperative Agreement between the US National Institute of Mental Health (NIMH) and a consortium of investigators supported jointly by the US Army and NIMH (Insel and McHugh, submitted for publication).

The earlier report by Kessler and colleagues in this issue detailed the fact that Army STARRS includes six coordinated component studies that were designed to interact with each other to facilitate non-experimental hypothesis generation and testing, intervention targeting, and intervention evaluation (Kessler *et al.*, 2013a). Each of these studies is a substantial undertaking in its own right. The unprecedented size, scope, and complexity of these six component studies created a number of logistical and coordination challenges for field implementation. The current report presents an overview of the approaches taken to meet these challenges, with a special focus on the complex field procedures involved in implementing the component studies. Data are also presented on sample sizes and projected response rates of the main Army STARRS surveys.

Organizational structure

The six components of Army STARRS include the following: (i) an Historical Administrative Data Study (HADS) of individual-level time series data from more than 50 million person-months in an integrated administrative database assembled from 38 different Army and Department of Defense (DoD) data systems for the more than 1.6 million soldiers who were on active duty in the US Army at any time between January 1, 2004 and December 31, 2009; (ii) parallel retrospective case–control studies of non-fatal suicides and suicide deaths that include in-depth interviews with soldiers (in the case of non-fatal attempts), Army supervisors, and next of kin (in the case of suicide deaths); (iii–v) three large-scale surveys of active duty Army personnel (the New Soldier Study [NSS] of soldiers at the beginning of Basic Combat Training [BCT]; the All-Army Study [AAS] of soldiers exclusive of those in BCT; and the Pre-Post Deployment Study [PPDS] of three Brigade Combat Teams initially assessed shortly before deploying to Afghanistan and then again three times after returning from deployment) that all included self-administered questionnaires (SAQs). The NSS and the Soldier Health Outcome Studies (SHOS-A) additionally administered neurocognitive tests, while the NSS and PPDS both obtained blood samples from a subset of respondents; (vi) Army STARRS is additionally carrying out a pilot study of the stresses and mental health problems associated with making the transition back to civilian life among soldiers who separate from military service. This pilot study is being implemented in preparation for future long-term follow-up studies of Army STARRS respondents after they separate from military service.

The logistical and coordination challenges of implementing these studies were addressed at the onset of the initiative by establishing an Executive Committee made up of the Co-Principal Investigators (Co-PIs) (Murray Stein, Robert Ursano), the site PIs (Steven Heeringa, Ronald Kessler) along with the Collaborating Scientists from NIMH (Lisa Colpe, Michael Schoenbaum) and US Army consultants (Steven Cervosky, Kenneth Cox) to provide overall supervision and direction. A Data Coordination Center was then established at the Survey Research Center (SRC), the Institute for Social Research, University of Michigan (www.src.isr.umich.edu) to implement the vast majority of Army STARRS data collections and to maintain the centralized Army STARRS data enclave that securely stores all Army STARRS analysis data and supports the computer servers used to carry out all substantive data analyses.

A study-specific design and analysis team was then established for each component Army STARRS study to develop instruments and work with SRC staff on design and field procedures. SRC staff, in turn, worked closely with the Office of the Deputy Under Secretary of the Army (ODUSA), under the auspices of which all Army STARRS data collections were carried out. In cases where data collection required access to local sites, the ODUSA worked with the Army to designate local Points of Contact (POC) who then coordinated with SRC to address logistical issues and facilitate data collection. The Training and Doctrine Command (TRADOC), Forces Command (FORSCOM), and for the Ary STARRS survey carried out in Kuwait the Army Medical Research and Materiel Command (MRMC), the US Central Command (USCENTCOM) and USCENTCOM's Joint Combat Casualty Research Team (JC2RT) were especially important in this regard. Additional coordination was provided by the Army Chaplain Corps, which provided support for the Army STARRS safety plan, and the Medical Command (MEDCOM), which provided Army medics for blood collection.

As noted earlier, all Army STARRS data are securely stored for analysis in the centralized Army STARRS SRC

data enclave. However, two specialized types of raw data are pre-processed elsewhere before being entered into the SRC data enclave for analysis. The first of these two involves the neurocognitive data collected in the NSS and in the case–control study of non-fatal suicide attempters. These data are scored at the University of Pennsylvania under the supervision of Army STARRS collaborator Rubin Gur prior to being transferred to SRC for inclusion in the data enclave for analysis. The second involves the blood samples collected in the NSS and PPDS. These samples are stored and pre-processed at the Rutgers University Cell and DNA Repository (RUCDR; http://www.rucdr.org). Genetic and other analyses performed on the stored blood samples are conducted either at RUCDR or other approved laboratories. All data derived from assays and tests performed on the stored blood are securely transferred to SRC for inclusion in the data enclave for analysis.

Instrumentation

Pilot work

Before turning to a discussion of field procedures, it is important to note that these procedures were constrained by a number of design requirements dictated by the results of an exhaustive review of the literature on risk and resilience factors for suicide and suicidal behaviors in the general population (Nock *et al.*, 2008) and the military (Gilman *et al.*, 2013). This review made it clear that suicidal behaviors develop through complex, multi-determined processes in which psychosocial and neurobiological factors combine to establish varying levels of risk that overlap for suicide and suicide attempts (Moscicki, 1999; Nock *et al.*, 2008). Rather than summarize the content of these reviews, we merely note for current purposes that the important predictors of suicidality documented in the review were many and varied. This meant that *detailed* assessments were required. In addition, *large samples* were required to achieve adequate statistical power to assess key hypotheses. Because of these requirements, it was necessary to make use of SAQs rather than interviewer-administered instruments.

Even though two-hour blocks of time were made available to Army STARRS to administer surveys (and two such sessions for new soldiers prior to beginning BCT), difficult decisions still had to be made in selecting short, efficient assessment batteries to assess all the constructs of interest to the research team. This problem was addressed by carrying out literature reviews of all instruments available to assess each construct of interest and then implementing extensive pilot studies to evaluate the psychometric properties of the instruments pinpointed in these literature reviews as top candidate measures. A number of methodological reports are either in preparation or, in two cases, completed (Thomas *et al.*, 2013; Kessler *et al.*, 2013c) to describe the results of those pilot studies and the psychometric properties of the final measures included in the Army STARRS component studies.

Another way to shorten assessment was to evaluate instrument skip logic carefully to make sure respondents were skipped out of survey sections as soon as the information needed to classify them on specific dimensions was obtained. This was especially important in the assessment of mental disorders, which took up a substantial part of the SAQs, and where it was possible to skip once it became clear that respondents failed to meet at least subthreshold criteria. Although the use of computer adaptive testing (CAT; Wainer, 2000) was carefully considered in this regard, CAT was rejected in the end due to our inability to carry out sufficiently large pilot studies to obtain stable test parameter estimates needed to guide CAT branching. Nonetheless, as noted next, the use of extensive skip logic in the SAQs had important implications for the preferred modes of data collection.

Data collection modes

The enormous size of the Army STARRS survey data collection effort led to the practical decision that SAQs be group-administered. However, the modes of data collection varied across these surveys. As described in a previous report in this issue (Kessler *et al.*, 2013a), the NSS was administered in three BCT facilities on a weekly basis over a period of two years, allowing SRC to establish a permanent data collection staff on these sites and to set up computer networks that allowed the SAQs to be computer-administered (CAI). Based on the success of this mode in the early NSS replicates, it was also used in administering the AAS at large installations and in the baseline and second follow-up wave of PPDS. However, it did not prove to be cost-effective to use CAI to administer the AAS in the many small installations, where the survey had to be carried out due to the logistical complications of transporting hardware for group survey administration. As a result, a paper-and-pencil-administered (PAPI) version of the AAS questionnaire was also developed.

Other Army STARRS component studies use a mix of data collection modes, including web-based CAI and telephone interviews. These modes are both used in assessing controls in the retrospective case–control study of non-fatal suicide attempts (SHOS-A) as well as in the third follow-up wave of the PPDS (T3), where web-based CAI is used initially and then telephone interviews are used to

assess respondents who fail to complete web-based SAQs. The need for these modes in T3 PPDS is that this survey is carried out approximately nine months after respondents return from deployment, by which time many of them are no longer assigned to the same unit. This means that these respondents have to be traced individually in order to administer the SAQs. This is most easily done with web-based surveys, but additional contact information (cell phones, social security numbers, contact information for next of kin who will know there whereabouts once they separate) was obtained from all baseline PPDS respondents for purposes of tracing them over time and for conducting telephone survey follow-ups of web survey non-respondents.

Multi-component assessment

As noted earlier, the initial literature review showed clearly that suicidal behaviors develop through complex, multi-determined processes. These processes are thought to involve psychosocial and neurobiological factors that combine to establish varying levels of risk (Moscicki, 1999; Nock *et al.*, 2008), with important factors including such diverse things as accumulating stressful life experiences that create risk of suicidality through processes partially mediated by biological pathways (McEwen, 2007) and modified by genetic susceptibilities (Krishnan *et al.*, 2007). The effects of these differential susceptibilities, in turn, are thought to be at least partially mediated by trait-temperament and environmental factors that are themselves jointly influenced by environmental and biological factors (Brent *et al.*, 2002; Caspi *et al.*, 2003; Higley and Linnoila, 1997; Kraemer *et al.*, 1997). There are formidable logistical challenges involved in sorting out these diverse influences that include the need for large longitudinal samples that assess a wide range of both biological and psychosocial variables and that provide opportunities for targeted intervention.

We were especially interested in having as much of this information as possible based on objective assessments due to concern about under-reporting in SAQs. Three approaches were used to do this, all of them having implications for field procedures. First, the Army and DoD administrative data systems provide an important source of independent (of self-report) information on environmental factors. Such information can be obtained at the level of the soldier's unit to define stressors to which the soldier was exposed by virtue of unit membership (e.g. numbers of unit members who recently died in combat, in non-duty-related injuries, or by suicide; numbers of unit members who recently had combat-related injuries, were reported to military authorities as victims of interpersonal or sexual violence, or had charges brought against them for being perpetrators of interpersonal or sexual violence). Individual-level data can also be obtained to characterize certain kinds of stressor exposures (e.g. information on such things as wages of the soldier being garnished due to financial debts, disciplinary actions, job performance ratings). Many other relevant administrative data systems exist as well, such as those providing data from electronic medical records, criminal justice records, and records of the Child and Protective Services system dealing with domestic violence issues. In order to obtain these data, though, it was necessary to obtain written informed consent from Army STARRS participants to access their administrative records. Field procedures for doing this are described later.

Second, the research team was especially interested in objectively measured neurocognitive factors that have been shown to predict suicidal behavior (Keilp *et al.*, 2008; Keilp *et al.*, 2001; Marzuk *et al.*, 2005). Administrative records provide access to a number of such tests that are administered to all soldiers in the Armed Services Vocational Aptitude Battery (ASVAB) in addition to the Armed Forces Qualification Test (AFQT) and General Technical (GT) Score. However, other neurocognitive factors have been shown to be robust predictors of suicide attempts that are not included in these test batteries, such as tests of poor decision-making, problem-solving, cognitive flexibility, and verbal fluency (Jollant *et al.*, 2005; Sadowski and Kelley, 1993). Recent work by members of the Army STARRS team has also shown that specific aspects of executive functioning associated with cognitive inflexibility or failure to adaptively adjust to changing demands distinguish depressed suicide attempters from non-attempters (Keilp *et al.*, 2001). In order to evaluate the effects of dimensions such as these it was necessary to administer special neuropsychological tests to Army STARRS respondents. However, it was logistically impossible to do this using the one-on-one administration methods traditionally used for such tests (i.e. one test administrator guiding one subject at a time through the test battery). As a result, special group-administered CAI neuropsychological software and test protocols were developed to administer these tests in conjunction with the Army STARRS surveys.

Third, the research team was quite interested in neurobiological predictors of suicidal behaviors. Although some predictors of this sort have been widely studied in clinical samples, much of this work uses protocols that could not realistically be used in broad-based epidemiological screening (e.g. post-mortem brain studies, Arango *et al.*,

1990; Arango *et al.*, 2001; Arango *et al.*, 1995; Boldrini *et al.*, 2004; Hsiung *et al.*, 2003; Mann *et al.*, 2000; Mann *et al.*, 1986; FDG-PET, Boldrini *et al.*, 2004; Parsey *et al.*, 2006). However, it was possible to obtain blood samples to study genetic effects related to suicidal behavior. Suicide appears to be partly heritable, as indicated by concordance being higher in monozygotic (MZ) than dizygotic (DZ) twins (Voracek and Loibl, 2007) and higher in the biological parents of adoptees who died by suicide than of other causes (Mann, 2003). However, the specific genes that contribute to vulnerability for suicide are unknown. This might be true because the numerous association studies on candidate genes have examined only a few candidate genes using a limited number of polymorphisms per gene (Anguelova *et al.*, 2003; Baldessarini and Hennen, 2004; Haghighi *et al.*, 2008; Mann *et al.*, 2001; Rujescu *et al.*, 2007; Uher and McGuffin, 2008; Zill *et al.*, 2004). Another important issue is that genes likely influence elements of the biological vulnerability for suicidal behavior rather than suicide or suicide attempt directly. For example, variants in the monoamine oxidase A gene have been associated with differences in aggression and impulsivity (Manuck *et al.*, 2000), while adverse childhood experiences (Huang *et al.*, 2004) and prenatal exposure to maternal smoking (Wakschlag *et al.*, 2009) have been shown to interact with genes to predict the development of antisocial behavior and greater impulsivity, both of which are risk factors for suicidal behavior in males (Huang *et al.*, 2004). Based on these considerations, it was hypothesized that more consistent results might be found in studies that distinguish suicide-related genes from the genes related to common associated major psychiatric disorders. In order to investigate this possibility, though, it was necessary to obtain blood samples from a large number of respondents. The field procedures used to do this are described later.

Fieldwork organization and procedures

Fieldwork organization

As noted earlier, most Army STARRS fieldwork is carried out by the professional SRC research staff. But there are exceptions. One exception is that AAS fieldwork in Kuwait could not be implemented by SRC staff due to restrictions on civilian travel to Kuwait. As a result, this fieldwork was carried out by Army staff using protocols developed by SRC with training and support provided by SRC. Army staff were also used to collect data from selected units located in Europe and in the Pacific Command (Korea, Hawaii and Alaska). A second exception is that subject recruitment and interviewing for the retrospective case–control study of non-fatal suicide attempts are carried out by research workers employed by the Army STARRS collaborators in the Department of Psychiatry at the Uniformed Services University of the Health Sciences (USUHS). These workers are physically located in psychiatric inpatient units in five participating tertiary care medical facilities (Walter Reed National Military Medical Center, Washington, DC; Fort Bragg, NC; Fort Stewart, GA; Fort Lewis, WA; and Fort Hood, TX), where they attempt to recruit soldiers admitted because of suicide attempts to participate in the case–control study of suicide attempts. Once obtaining written informed consent from these soldiers, the USUHS research workers administer surveys, carry out neuropsychological tests, and obtain blood samples. A third exception is that clinical interviewers employed by the Army STARRS collaborators at Harvard University conduct qualitative telephone interviews with suicide attempters. These interviews use a theoretically-guided form of qualitative interviewing designed to uncover information about critical junctures in the progression to suicide attempts and completions (Strauss, 1987). A final exception is that a clinical reappraisal study of the self-report assessment battery for DSM-IV disorders used in all of the large-scale Army STARRS SAQs (Kessler *et al.*, 2012) was carried out by clinical interviewers employed by the Army STARRS collaborators at USUHS. These interviews involve clinical reappraisal assessments administered by telephone with STARRS SAQ respondents using the Structured Clinical Interview for the Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition (SCID) (First *et al.*, 2002) as the clinical interview schedule.

Key challenges in fieldwork implementation

Formidable logistical challenges were faced in fielding the large-scale Army STARRS data collections. SRC was required to develop stand-alone, highly-secure, wireless computer networks that could efficiently serve up to 300 laptop computers on NSS and selected AAS Army installations. These networks had to be set up, broken down, and set up again twice each week at each installation where surveys were being carried out. The networks and laptops had to be shipped to new sites with each new quarterly sample replicate. SRC staff transported and set up laptops and network equipment for each data collection session, then packed and returned the equipment to the storage site established at the installation, and recharged laptop batteries between sessions at the storage sites. Where the target AAS sample unit was so small at a given site that it was not feasible or cost effective to set up computer

networks, PAPI was used for group self-administration of the AAS. Whole blood collection in the NSS and PPDS required the development of a special identification protocol involving use of wrist bands with ID numbers and bar codes to ensure respondent confidentiality and permit linking of blood samples with survey responses. Coordination was also required with Army phlebotomists.

Considerable travel and ongoing coordination were needed to establish mobile data collection facilities at the rotating set of data collection sites that were new in each replicate of the AAS. The SRC data collection teams that created this travelling set of facilities worked through the ODUSA to designate a local POC who worked directly with SRC to address logistical issues and facilitate data collection. An SRC Site Coordinator at each site worked with the POC to schedule sessions; obtain contact information for local Chaplains for the safety plan; and ensure access to storage space (for equipment and paper materials), tables and chairs, electrical outlets, etc. for each data collection session. Fitting Army STARRS data collections into the very busy schedules of Army units required a great deal of flexibility and creativity on the part of the SRC staff. Other challenges were faced in implementing the case–control studies of non-fatal suicide attempters and suicide deaths (contacting and recruiting next of kin; selecting, tracing, and recruiting control soldiers and supervisors).

Unit recruitment and logistical planning (including issuing operational orders) were carried out through the ODUSA, TRADOC, and FORSCOM. Contact with study units began with SRC staff briefing unit leaders on the purposes and importance of Army STARRS and then working with POCs to explain data collection requirements and develop local protocols to address logistical challenges. Additional coordination was needed with the Army Chaplain Corps, which provided support for the safety plan, and the MEDCOM, which provided Army medics for blood collection. The ODUSA maintained a travel team of high-ranking officers who accompanied the SRC data collection team to ensure Army support at each installation. The support and guidance provided by these Commands was essential to the success of Army STARRS data collection.

All soldiers selected for participation in Army STARRS were provided with an information sheet explaining the purposes of the study, providing answers to frequently asked questions, and giving a toll-free number for those who had additional questions. In the cases of the NSS, AAS, and PPDS, pre-designated respondents were additionally ordered to attend a group-administered 30-minute informed consent session that explained study purposes, procedures and confidentiality protections; emphasized the voluntary nature of participation (including the right to withdraw consent at a future date); and answered questions before seeking informed consent. SRC staff made the presentations at these sessions with in-person presentations made by ODUSA staff. Written informed consent was then obtained from volunteers. The Human Subjects Committees of the University of Michigan and USUHS (and for the Kuwait component, the Army Medical Research and Materiel Command) approved all recruitment, consent, and field procedures.

Fieldworker training

SRC hired and trained Site Coordinators and Group Session Proctors for the AAS, NSS, and PPDS data collections. SRC recruitment for the SCID clinical reappraisal study was conducted in person by the SRC field staff. SRC telephone recruitment for the SHOS-A/B case–control studies and telephone interviewing (for the case–control studies and for the third wave of PPDS follow-up interviews), in comparison, are being carried out by interviewers in the SRC Survey Services Laboratory in Ann Arbor, Michigan. Each professional SRC interviewer and fieldworker completes a General Interviewer Training (GIT) course before working on any project. Experienced workers additionally complete GIT refresher courses on a periodic basis. Each Site Coordinator, Group Session Proctor, and Interviewer who worked on Army STARRS also received 4–5 days of study-specific training and completed an Army STARRS certification test before beginning production work.

The USUHS clinical interviewers who administered the SCID interviews were trained by Michael First, a developer of the SCID, and were supervised by an experienced clinical research supervisor. All SCID interviews were digitally recorded with the permission of respondents for quality review purposes. The supervisor reviewed all written interviewer notes and selected recordings. Biweekly in-person clinical interviewer review meetings were held throughout the clinical calibration study field period to prevent interviewer drift. These meetings were chaired by the supervisor and attended remotely by the trainer.

The Harvard clinical interviewers who conduct the qualitative interviews with non-fatal suicide attempters were trained and supervised by Matthew Nock, a clinical psychologist with a long history of research on suicidal behaviors. Nock developed the interview schedule used in these assessments and also conducts some of the interviews. All these clinical interviews are audio-recorded with the permission of respondents and then transcribed and content analyzed using double-coding by independent

raters to establish inter-rater reliability. The coding system evolves in the course of content analysis. Interviewers/coders meet weekly with Nock to review results, address the problem of interviewer/coder drift, and discuss updates and revisions to the coding system.

Fieldwork quality control

As noted earlier, weekly or biweekly interviewer meetings and reviews of tape recorded interviews were used in the clinical reappraisal study (which has now ended) and continue to be used in the ongoing case–control studies to maintain quality control of data collection. In the case of the large-scale survey data collections, quality control procedures began with close monitoring by SRC staff of respondent selection procedures to avoid selectively recruiting respondents. CAI programs were then used in all data collections other than the AAS surveys implemented on small US installations and overseas to control skip logic. Completed CAI SAQs were sent electronically (by mail for PAPI and overseas SAQs) every night to the central Army STARRS Data Coordination Center at University of Michigan, allowing supervisors to monitor data flow and make quality control checks on a daily basis. In cases where problems were detected, rapid remediation efforts were undertaken. Despite these quality control steps, logistical problems occurred in a number of group administration sessions involving units of soldiers arriving late or having to leave early that led to incomplete surveys in a non-trivial proportion of cases. Computer hardware problems were also encountered in a small number of early SAQ sessions that resulted in loss of data. As these were relatively random occurrences, though, we addressed these data losses with the weighting procedures described later.

Sample sizes and response rates

Information on final sample sizes and response rates are available for the NSS and baseline PPDS. Only estimated projections are available, though, for the AAS and PPDS follow-up surveys. In the case of the AAS, while data collection has recently ended, with the late addition of activated Army Reserve (USAR) and National Guard (USANG) units in the continental United States that were either about to deploy to Afghanistan or about to separate from military service after returning from a deployment to Afghanistan, it will take some months to reconcile sample totals with population totals for these replicates. In the case of the PPDS, follow-up surveys are still in progress.

The numbers of respondents with substantially complete SAQ data are 50,765 in the NSS, 9421 in the baseline PPDS, and projected to be 35,372 in the AAS (Table 1). These numbers represent SAQ response rates of 88.8% in the NSS, 90.8% in the baseline PPDS, and 72.0% in the AAS replicates for which results are currently available. The numbers of SAQ respondents that additionally provided written informed consent and accurate identifying information to link their SAQ responses to their administrative data system (ADS) records are 39,132 in the NSS, 7425 in the baseline PPDS, and a projected 24,266 in the AAS. These numbers represent SAQ + ADS response rates of 68.5% in the NSS, 71.5% in the baseline PPDS, and a projected 49.8% in the AAS. Blood samples were also collected in the NSS and baseline PPDS, with some SAQ respondents providing blood samples but not ADS linkage and others providing ADS linkage but not blood samples. The numbers with complete SAQ data and blood samples (with or without ADS linkage) are 33,088 in the NSS and 7923 in the baseline PPDS (80.1–76.2% response rates), while the numbers with complete SAQ data, blood samples, and ADS linkage are 27,807 in the NSS and 6531 in the baseline PPDS (67.3–62.9% response rates).

A decomposition of reasons for incomplete response shows that even though all unit members in the AAS and PPDS are ordered to report to the Army STARRS informed consent sessions, 7.3% in the baseline PPDS and 23.5% in the AAS units for which results are currently available were absent due to conflicting duty assignments. However, the vast majority of those attending the informed consent sessions in both surveys (96.0–98.7%) consented to complete the SAQ and 98.0–99.2 of consenters completed the surveys. The situation is quite different in the NSS, where 100% of new soldiers selected for the survey attended the informed consent sessions (i.e. attendance at these sessions was made a part of the new soldier training schedule) and a similar fraction as in the AAS or PPDS consented to participate (97.7% versus 96.0–98.7%) but a smaller proportion of consenters completed the survey (91.0% versus 98.0–99.2%). Consent to link SAQ and administrative data was considerably higher among NSS (83.5%) and PPDS (84.0%) SAQ completers than AAS SAQ completers (72.4%). Based on these differences, the *cooperation* rates for the conjunction of SAQ completion and successful record linkage among soldiers attending informed consent sessions (i.e. response among those who were contacted and participated in the informed consent sessions) are lower in the NSS (68.5%) and AAS (65.1%) than the baseline PPDS.

Discussion

This paper has presented an overview of the field procedures in the Army STARRS, a multi-component initiative

Table 1. Interim sample dispositions in the Army STARRS surveys[1]

	New Soldier Study (Q1 2011–Q4 2012)	All Army Study (Q1–4 2011)	Baseline Pre-Post Deployment Study
I. Components of cooperation rate and response rate calculations			
Attending consent session (ACS)[2]	100.0%	76.5%	92.7%
Consent to complete survey (CCS/ACS)	97.7	96.0	98.7
Completion of survey (CS/CCS)[3]	91.0	98.0	99.2
Consent for linkage of ADS data among survey completers (CRL/CS)	83.5	72.4	84.0
Successful record linkage (ADS/[CS + CRL])[3]	92.3	95.6	93.8
II. Cooperation rates			
Survey	88.8	94.1	97.9
Survey + Consent for record linkage	74.2	68.1	82.2
Survey + ADS	68.5	65.1	77.2
Survey + blood	80.1	—	82.4
Survey + ADS + blood	67.3	—	67.9
III. Response rates			
Survey	88.8	72.0	90.8
Survey + Consent for record linkage	74.2	52.2	76.3
Survey + ADS	68.5	49.8	71.5
Survey + blood	80.1	—	76.2
Survey + ADS + blood	67.3	—	62.9
IV. Sample sizes			
Target sample	(57,152)	(49,128)[5]	(10,380)
Sample with completed surveys	(50,765)	(35,372)[5]	(9421)
Sample with completed surveys and ADS	(39,132)	(24,266)[5]	(7425)
Sample with completed surveys and blood	(33,088)	—	(7923)
Sample with completed surveys, ADS, and blood[4]	(27,807)	—	(6531)

[1]NSS dispositions are reported for calendar years 2011 and 2012. AAS dispositions are reported for replicates in calendar years 2011, as 2012 results are not yet finalized. PPDS dispositions are reported for the full pre-deployment PPDS sample.
[2]The AAS and PPDS target samples were all soldiers in designated units, allowing us to calculate the proportion of target respondents that attended the consent sessions. The NSS target samples, in comparison, were stipulated to be samples of new soldiers recruited on designated survey administration days in Reception Battalion to equal the numbers we could accommodate in the group survey administration settings established on the training bases. All new soldiers designated to be part of these target samples were designated as such by the Army Points of Contact (POCs) and marched to the Army STARRS group administration setting for the informed consent presentation before being asked to provide voluntary informed consent for the survey. Army STARRS data collection staff worked with Army POCs to guarantee that the target samples were not systematically biased. Based on these NSS recruitment procedures, the table stipulates that 100% of pre-designated respondents attended the NSS consent sessions.
[3]Failure to complete the survey was largely due to logistical problems with units arriving late or having to leave early from the group survey sessions, although some new soldiers were unable to complete the survey in the allocated 90-minute data collection period. Record linkage failure occurred when respondents who signed the informed consent form for record linkage either failed to provide linking information or provided information that did not match the information available in Army administrative records.
[4]Collection of blood did not begin until the fourth quarter of 2011 due to delays in IRB approval of this study component.
[5]The final sample sizes for the AAS are projected due to the numbers of respondents in the final replicates, which consist of activated USAR and USANG units, not yet being available at the time this report was prepared.

designed to investigate risk and resilience factors for suicidality and its psychopathological correlates among US Army personnel. We also presented preliminary information on sample sizes and response rates. The unprecedented size, scope, and complexity of Army STARRS created formidable challenges that, as described in the

body of the paper, we addressed by putting in place an organizational structure that provided coordination across component studies while allowing flexibility and creativity within studies and using the expertise of SRC to guide all data collection efforts. The fact that the component Army STARRS studies were carried out as part of a high-profile integrated research initiative helped promote cooperation, as indicated by the fact that the response rates in the AAS and baseline PPDS were a good deal higher than those in a number of other major military surveys (Bray *et al.*, 2006; Ryan *et al.*, 2007).

An issue of special importance in considering the response rates is that the Army STARRS studies, while *de-identified* (i.e. identifying information is kept separate from research data), are not *anonymous* (i.e. identifying information exists that can be linked to the research data of individual respondents) due to the fact that we are linking administrative data to survey responses and following respondents over time. This is in contrast to some other large military surveys, like the Department of Defense Survey of Health Related Behaviors Among Active Duty Military Personnel (DoD Health Behavior Surveys; Ryan *et al.*, 2007) and the Mental Health Surveillance Surveys in combat environments carried out by US Army Mental Health Advisory Teams (MHATs; Bliese *et al.*, 2011), which were purposefully designed to be anonymous in order to encourage complete and accurate reporting.

The rationale for anonymity in military surveys is compelling, based on the fact that meta-analyses strongly suggest that anonymity can influence survey reports of embarrassing behaviors both in the general population (Turner *et al.*, 1998) and of mental disorders in the military (Gadermann *et al.*, 2012). As a result, a strategic decision was made in fielding the Army STARRS to allow respondents to provide completely anonymous survey reports even though we needed identifying information for ADS linkage. This was done by creating a separate informed consent form for identifying information to link self-report data to other types of data and encouraging respondents to complete the SAQ even if they did not want to consent to ADS linkage.

It is noteworthy in this regard that the proportions of soldiers attending AAS and PPDS consent sessions who completed these surveys (94.1–98.0%) are similar to the proportions who participated in previous Army surveys that used completely anonymous surveys. For example, the cooperation rate in the most recently reported DoD Health Behavior Survey among soldiers attending consent-survey sessions was 95.8% (Bray *et al.*, 2009). The cooperation rate among soldiers attending consent-survey sessions in an earlier survey of pre-post deployment mental health of US Army soldiers and Marines deployed in Operation Iraqi Freedom (OIF) and Operation Enduring Freedom (OEF) in Afghanistan was 98% (Hoge *et al.*, 2004). Importantly, though, the proportions of eligible respondents who attended the consent-survey sessions in these earlier studies (80% of those described as "accessible for study," but only 64.7% of all unit members, in the DoD Health Behavior Survey; 58% in the OIF/OEF surveys) were considerably lower than in the Army STARRS PPDS (92.7%) and AAS (76.5%). This suggests that there was more self-selection of cooperative soldiers in these earlier surveys than in the AAS or PPDS, making it all the more striking that the SAQ cooperation rates in the AAS and PPDS were comparable to those in these earlier surveys.

An advantage of having access to SAQ data for soldiers who did not consent to ADS linkage is that comparisons can be made with the SAQ reports of soldiers that agreed to ADS linkage. With regard to objective variables reported in the SAQs (e.g. age, sex, education, rank, marital status), these comparisons allow us to examine the extent to which consent for ADS linkage was non-random. Data are also available in the SAQ on more subjective data, such as reports of being anxious, depressed, and suicidal. These reports might be biased by the knowledge that responses are not completely anonymous, making the comparison of results in the completely anonymous SAQs and the de-identified (but not completely anonymous) SAQS of considerable interest. This kind of comparison is the focus of a companion paper in this issue (Kessler *et al.*, 2013b).

Acknowledgments

On behalf of the Army STARRS Collaborators

Funding/Support: Army STARRS was sponsored by the Department of the Army and funded under cooperative agreement number U01MH087981 with the US Department of Health and Human Services, National Institutes of Health, National Institute of Mental Health (NIH/NIMH). The contents are solely the responsibility of the authors and do not necessarily represent the views of the Department of Health and Human Services, NIMH, the Department of the Army, or the Department of Defense.

Role of the Sponsors: As a cooperative agreement, scientists employed by NIMH (Colpe and Schoenbaum) and Army liaisons/consultants (COL Steven Cersovsky, MD, MPH USAPHC and Kenneth Cox, MD, MPH USAPHC) collaborated to develop the study protocol and data collection instruments, supervise data collection, plan and supervise data analyses, interpret results, and prepare reports. Although a draft of this manuscript was submitted to the Army and NIMH for review and comment prior to submission, this was with the understanding that comments would be no more than advisory.

Int. J. Methods Psychiatr. Res. 22(4): 276–287 (2013). DOI: 10.1002/mpr

Additional Contributions: The Army STARRS Team consists of Co-Principal Investigators: Robert J. Ursano, MD (Uniformed Services University of the Health Sciences) and Murray B. Stein, MD, MPH (University of California San Diego and VA San Diego Healthcare System); Site Principal Investigators: Steven Heeringa, PhD (University of Michigan) and Ronald C. Kessler, PhD (Harvard Medical School); NIMH collaborating scientists: Lisa J. Colpe, PhD, MPH and Michael Schoenbaum, PhD; Army liaisons/consultants: COL Steven Cersovsky, MD, MPH (USAPHC) and Kenneth Cox, MD, MPH (USAPHC). Other team members: Pablo A. Aliaga, MA (Uniformed Services University of the Health Sciences); COL David M. Benedek, MD (Uniformed Services University of the Health Sciences); Susan Borja, PhD (National Institute of Mental Health); Gregory G. Brown, PhD (University of California San Diego); Laura Campbell-Sills, PhD (University of California San Diego); Catherine Dempsey, PhD, MPH (Uniformed Services University of the Health Sciences); Richard Frank, PhD (Harvard Medical School); Carol S. Fullerton, PhD (Uniformed Services University of the Health Sciences); Nancy Gebler, MA (University of Michigan); Joel Gelernter, MD (Yale University); Robert K. Gifford, PhD (Uniformed Services University of the Health Sciences); Stephen E. Gilman, ScD (Harvard School of Public Health); Marjan G. Holloway, PhD (Uniformed Services University of the Health Sciences); Paul E. Hurwitz, MPH (Uniformed Services University of the Health Sciences); Sonia Jain, PhD (University of California San Diego); Tzu-Cheg Kao, PhD (Uniformed Services University of the Health Sciences); Karestan C. Koenen, PhD (Columbia University); Lisa Lewandowski-Romps, PhD (University of Michigan); Holly Herberman Mash, PhD (Uniformed Services University of the Health Sciences); James E. McCarroll, PhD, MPH (Uniformed Services University of the Health Sciences); Katie A. McLaughlin, PhD (Harvard Medical School); James A. Naifeh, PhD (Uniformed Services University of the Health Sciences); Matthew K. Nock, PhD (Harvard University); Rema Raman, PhD (University of California San Diego); Nancy A. Sampson, BA (Harvard Medical School); LCDR Patcho Santiago, MD, MPH (Uniformed Services University of the Health Sciences); Michaelle Scanlon, MBA (National Institute of Mental Health); Jordan Smoller, MD, ScD (Harvard Medical School); Nadia Solovieff, PhD (Harvard Medical School); Michael L. Thomas, PhD (University of California San Diego); Christina Wassel, PhD (University of Pittsburgh); and Alan M. Zaslavsky, PhD (Harvard Medical School).
Additional Information: A complete list of Army STARRS publications can be found at http://www.ARMYSTARRS.org.

Declaration of interest statement

In the past five years Kessler has been a consultant for Eli Lilly & Company, Glaxo, Inc., Integrated Benefits Institute, Ortho-McNeil Janssen Scientific Affairs, Pfizer Inc., Sanofi-Aventis Groupe, Shire US Inc., and Transcept Pharmaceuticals Inc. and has served on advisory boards for Johnson & Johnson. Kessler has had research support for his epidemiological studies over this time period from Eli Lilly & Company, EPI-Q, GlaxoSmithKline, Ortho-McNeil Janssen Scientific Affairs, Sanofi-Aventis Groupe, Shire US, Inc., and Walgreens Co. Kessler owns a 25% share in DataStat, Inc. Stein has in the last three years been a consultant for Healthcare Management Technologies and had research support for pharmacological imaging studies from Janssen. The remaining authors report no competing interests.

References

Anguelova M., Benkelfat C., Turecki G. (2003) A systematic review of association studies investigating genes coding for serotonin receptors and the serotonin transporter: II. Suicidal behavior. *Molecular Psychiatry*, **8**(7), 646–653, DOI: 10.1038/sj.mp.4001336.

Arango V., Ernsberger P., Marzuk P.M., Chen J.S., Tierney H., Stanley M., Reis D.J., Mann J.J. (1990) Autoradiographic demonstration of increased serotonin 5-HT2 and beta-adrenergic receptor binding sites in the brain of suicide victims. *Archives of General Psychiatry*, **47**(11), 1038–1047, DOI: 10.1001/archpsyc.1990.01810230054009.

Arango V., Underwood M.D., Boldrini M., Tamir H., Kassir S.A., Hsiung S., Chen J.J., Mann J.J. (2001) Serotonin 1A receptors, serotonin transporter binding and serotonin transporter mRNA expression in the brainstem of depressed suicide victims. *Neuropsychopharmacology*, **25**(6), 892–903, DOI: 10.1016/S0893-133X(01)00310-4.

Arango V., Underwood M.D., Gubbi A.V., Mann J.J. (1995) Localized alterations in pre- and postsynaptic serotonin binding sites in the ventrolateral prefrontal cortex of suicide victims. *Brain Research*, **688**(1–2), 121–133, DOI: 10.1016/0006-8993(95)00523-S.

Baldessarini R.J., Hennen J. (2004) Genetics of suicide: an overview. *Harvard Review of Psychiatry*, **12**(1), 1–13.

Bliese P.D., Thomas J.L., McGurk D., McBride S., Castro C.A. (2011) Mental health advisory teams: a proactive examination of mental health during combat deployments. *International Review of Psychiatry*, **23**(2), 127–134, DOI: 10.3109/09540261.2011.558834.

Boldrini M., Underwood M.D., Martini A., Kassir S. A., Mann J.J., Arango V. (2004) Distribution of serotonin-1A autoreceptors in the dorsal raphe nucleus of depressed suicide victims. ACNP 43rd Annual Meeting. San Juan, Puerto Rico.

Bray R.M., Hourani L.L., Olmsted K.L.R., Witt M., Brown J.M., Pemberton M.R., Marsden M.E., Marriott B., Scheffler S., Vandermaas-Peeler R., Weimer B., Calvin S., Bradshaw M., Close K., Hayden D. (2006) 2005 Department of Defense Survey of Health Related Behaviors Among Active Duty Military Personnel: A Component of the Defense Lifestyle Assessment Program (DLAP), Research Triangle Park, NC, Research Triangle Institute.

Bray R.M., Pemberton M.R., Hourani L.L., Witt M., Rae Olmsted K.L., Brown J.M., Weimer B.J., Lane M.E., Marsden M.E., Scheffler S.A.,

Vandermaas-Peeler R., Aspinwall K.R., Anderson E.M., Spagnola K., Close K.L., Gratton J.L., Calvin S.L., Bradshaw M.R. (2009) 2008 Department of Defense Survey of Health Related Behaviors Among Active Duty Military Personnel: A Component of the Defense Lifestyle Assessment Program (DLAP), Research Triangle Park, NC, Research Triangle Institute.

Brent D.A., Oquendo M., Birmaher B., Greenhill L., Kolko D., Stanley B., Zelazny J., Brodsky B., Bridge J., Ellis S., Salazar J.O., Mann J.J. (2002) Familial pathways to early-onset suicide attempt: risk for suicidal behavior in offspring of mood-disordered suicide attempters. *Archives of General Psychiatry*, **59**(9), 801–807, DOI: 10.1001/archpsyc.59.9.801.

Caspi A., Sugden K., Moffitt T.E., Taylor A., Craig I.W., Harrington H., McClay J., Mill J., Martin J., Braithwaite A., Poulton R. (2003) Influence of life stress on depression: moderation by a polymorphism in the 5-HTT gene. *Science*, **301**(5631), 386–389, DOI: 10.1126/science.1083968.

First M.B., Spitzer R.L., Gibbon M., Williams J. B.W. (2002) Structured Clinical Interview for DSM-IV Axis I Disorders, Research Version, Non-patient Edition (SCID-I/NP), New York, Biometrics Research, New York State Psychiatric Institute.

Gadermann A.M., Engel C.C., Naifeh J.A., Nock M.K., Petukhova M., Santiago P.N., Wu B., Zaslavsky A. M., Kessler R.C. (2012) Prevalence of DSM-IV major depression among U.S. military personnel: meta-analysis and simulation. *Military Medicine*, **177**(8 Suppl), 47–59.

Gilman S.E., Goldenberg M., Kessler R.C., McCarroll J.E., McLaughlin K.A., Peterson C., Schoenbaum M., Stanley B., Ursano R.J. (2013) Suicide among Soldiers: a review of psychological risk and protective factors. *Psychiatry*, **76**(2), 97–125.

Haghighi F., Bach-Mizrachi H., Huang Y.Y., Arango V., Shi S., Dwork A.J., Rosoklija G., Sheng H.T., Morozova I., Ju J., Russo J.J., Mann J.J. (2008) Genetic architecture of the human tryptophan hydroxylase 2 Gene: existence of neural isoforms and relevance for major depression. *Molecular Psychiatry*, **13**(8), 813–820, DOI: 10.1038/sj.mp.4002127.

Higley J.D., Linnoila M. (1997) Low central nervous system serotonergic activity is traitlike and correlates with impulsive behavior. A nonhuman primate model investigating genetic and environmental influences on neurotransmission. *Annals of the New York Academy of Sciences*, **836**, 39–56, DOI: 10.1111/j.1749-6632.1997.tb52354.x.

Hoge C.W., Castro C.A., Messer S.C., McGurk D., Cotting D.I., Koffman R.L. (2004) Combat duty in Iraq and Afghanistan, mental health problems, and barriers to care. *New England Journal of Medicine*, **351**(1), 13–22, DOI: 10.1056/NEJMoa040603.

Hsiung S., Adlersberg M., Arango V., Mann J.J., Tamir H., Liu K. (2003) Reduced 5-HT1A receptor signaling in brains of suicide victims: involvement of adenylyl cyclase, phosphatidylinositol 3-kinase, Akt and MAP kinase. *Journal of Neurochemistry*, **87**(1), 182–194, DOI: 10.1046/j.1471-4159.2003.01987.x.

Huang Y.Y., Cate S.P., Battistuzzi C., Oquendo M.A., Brent D., Mann J.J. (2004) An association between a functional polymorphism in the monoamine oxidase a gene promoter, impulsive traits and early abuse experiences. *Neuropsychopharmacology*, **29**(8), 1498–1505, DOI: 10.1038/sj.npp.1300455.

Insel T., McHugh J.M. (submitted for publication) The U.S. Army-National Institute of Mental Health Army Study to Assess Risk and Resilience in Servicemembers (Army STARRS): rapidly translating interventions to reduce suicide.

Jollant F., Bellivier F., Leboyer M., Astruc B., Torres S., Verdier R., Castelnau D., Malafosse A., Courtet P. (2005) Impaired decision making in suicide attempters. *American Journal of Psychiatry*, **162**(2), 304–310, DOI: 10.1176/appi.ajp.162.2.304.

Keilp J.G., Gorlyn M., Oquendo M.A., Burke A.K., Mann J.J. (2008) Attention deficit in depressed suicide attempters. *Psychiatry Research*, **159**(1–2), 7–17, DOI: 10.1016/j.psychres.2007.08.020.

Keilp J.G., Sackeim H.A., Brodsky B.S., Oquendo M.A., Malone K.M., Mann J.J. (2001) Neuropsychological dysfunction in depressed suicide attempters. *American Journal of Psychiatry*, **158**(5), 735–741, DOI: 10.1176/appi.ajp.158.5.735.

Kessler R.C., Calabrese J.R., Farley P.A., Gruber M.J., Jewell M.A., Katon W., Keck P.E., Nierenberg A.A., Sampson N.A., Shear M.K., Shillington A.C., Stein M.B., Thase M.E., Wittchen H.U. (2012) Composite International Diagnostic Interview screening scales for DSM-IV anxiety and mood disorders. *Psychological Medicine*, 1–13, DOI: 10.1017/S0033291712002334.

Kessler R.C., Colpe L.J., Fullerton C.S., Gebler N., Naifeh J.A., Nock M.K., Sampson N.A., Schoenbaum M., Zaslavsky A.M., Stein M.B., Ursano R.J., Heeringa S.G. (2013a) Design of the Army Study to Assess Risk and Resilience in Servicemembers (Army STARRS). *International Journal of Methods and Psychiatric Research*, **22**(4), 267–275.

Kessler R.C., Heeringa S.G., Colpe L.J., Fullerton C.S., Gebler N., Naifeh J.A., Nock M.K., Sampson N. A., Schoenbaum M., Zaslavsky A.M., Stein M.B., Ursano R.J. (2013b) Response bias, weighting adjustments, and design effects in the Army Study to Assess Risk and Resilience in Servicemembers (Army STARRS). *International Journal of Methods in Psychiatric Research*, **22**(4), 288–302.

Kessler R.C., Santiago P., Colpe L.J., Dempsey C.L., First M.H., Heeringa S. G., Stein M.B., Fullerton C.S., Gruber M.J., Naifeh J.A., Nock M.K., Sampson N.A., Schoenbaum M., Zaslavsky A.M., Ursano R.J. (2013c) Clinical reappraisal of the Composite International Diagnostic Interview Screening Scales (CIDI-SC) in the Army Study to Assess Risk and Resilience in Servicemembers (Army STARRS). *International Journal of Methods and Psychiatric Research*, **22**(4), 303–321.

Kraemer G.W., Schmidt D.E., Ebert M.H. (1997) The behavioral neurobiology of self-injurious behavior in rhesus monkeys. Current concepts and relations to impulsive behavior in humans. *Annals of the New York Academy of Sciences*, **836**, 12–38, DOI: 10.1111/j.1749-6632.1997.tb52353.x.

Krishnan V., Han M.H., Graham D.L., Berton O., Renthal W., Russo S.J., Laplant Q., Graham A., Lutter M., Lagace D.C., Ghose S., Reister R., Tannous P., Green T.A., Neve R.L., Chakravarty S., Kumar A., Eisch A.J., Self D. W., Lee F.S., Tamminga C.A., Cooper D.C., Gershenfeld H.K., Nestler E.J. (2007) Molecular adaptations underlying susceptibility and resistance to social defeat in brain reward regions. *Cell*, **131**(2), 391–404, DOI: 10.1016/j.cell.2007.09.018.

Mann J.J. (2003) Neurobiology of suicidal behaviour. *Nature Reviews Neuroscience*, **4**(10), 819–828, DOI: 10.1038/nrn1220.

Mann J.J., Brent D.A., Arango V. (2001) The neurobiology and genetics of suicide and attempted suicide: a focus on the serotonergic system. *Neuropsychopharmacology*, **24**(5), 467–477, DOI: 10.1016/S0893-133X(00)00228-1.

Mann J.J., Huang Y.Y., Underwood M.D., Kassir S.A., Oppenheim S., Kelly T.M., Dwork A.J., Arango V. (2000) A serotonin transporter gene promoter polymorphism (5-HTTLPR) and prefrontal cortical binding in major depression

and suicide. *Archives of General Psychiatry*, **57**(8), 729–738, DOI: 10.1001/archpsyc.57.8.729.

Mann J.J., Stanley M., McBride P.A., McEwen B.S. (1986) Increased serotonin2 and beta-adrenergic receptor binding in the frontal cortices of suicide victims. *Archives of General Psychiatry*, **43**(10), 954–959, DOI: 10.1001/archpsyc.1986.01800100048007.

Manuck S.B., Flory J.D., Ferrell R.E., Mann J.J., Muldoon M.F. (2000) A regulatory polymorphism of the monoamine oxidase-A gene may be associated with variability in aggression, impulsivity, and central nervous system serotonergic responsivity. *Psychiatry Research*, **95**(1), 9–23, DOI: 10.1016/S0165-1781(00)00162-1.

Marzuk P.M., Hartwell N., Leon A.C., Portera L. (2005) Executive functioning in depressed patients with suicidal ideation. *Acta Psychiatrica Scandinavica*, **112**(4), 294–301, DOI: 10.1111/j.1600-0447.2005.00585.x.

McEwen B.S. (2007) Physiology and neurobiology of stress and adaptation: central role of the brain. *Physiological Reviews*, **87**(3), 873–904, DOI: 10.1152/physrev.00041.2006.

Moscicki E.K. (1999) Epidemiology of suicide. In Jacobs D.G. (ed.) *The Harvard Medical School Guide to Suicide Assessment and Intervention*, pp 40–51, San Francisco, CA: Jossey-Bass.

Nock M.K., Borges G., Bromet E.J., Cha C.B., Kessler R.C., Lee S. (2008) Suicide and suicidal behavior. *Epidemiologic Reviews*, **30**(1), 133–154, DOI: 10.1093/epirev/mxn002.

Parsey R.V., Hastings R.S., Oquendo M.A., Huang Y.Y., Simpson N., Arcement J., Huang Y., Ogden R.T., Van Heertum R.L., Arango V., Mann J.J. (2006) Lower serotonin transporter binding potential in the human brain during major depressive episodes. *American Journal of Psychiatry*, **163**(1), 52–58, DOI: 10.1176/appi.ajp.163.1.52.

Rujescu D., Thalmeier A., Moller H.J., Bronisch T., Giegling I. (2007) Molecular genetic findings in suicidal behavior: what is beyond the serotonergic system? *Archives of Suicide Research*, **11**(1), 17–40, DOI: 10.1080/13811110600897317.

Ryan M.A., Smith T.C., Smith B., Amoroso P., Boyko E.J., Gray G.C., Gackstetter G.D., Riddle J. R., Wells T.S., Gumbs G., Corbeil T.E., Hooper T.I. (2007) Millennium Cohort: enrollment begins a 21-year contribution to understanding the impact of military service. *Journal of Clinical Epidemiology*, **60**(2), 181–191, DOI: 10.1016/j.jclinepi.2006.05.009.

Sadowski C., Kelley M.L. (1993) Social problem solving in suicidal adolescents. *Journal of Consulting and Clinical Psychology*, **61**(1), 121–127, DOI: 10.1037/0022-006X.61.1.121.

Strauss A.L. (1987) *Qualitative Analysis for Social Scientists*, Cambridge: Cambridge University Press.

Thomas M.L., Brown G.G., Gur R.C., Hansen J.A., Nock M.K., Heeringa S., Ursano R.J., Stein M.B. (2013) Parallel psychometric and cognitive modeling analyses of the Penn Face Memory Test in the Army Study to Assess Risk and Resilience in Servicemembers. *Journal of Clinical and Experimental Neuropsychology*, **35**(3), 225–245, DOI: 10.1080/13803395.2012.762974.

Turner C.F., Ku L., Rogers S.M., Lindberg L.D., Pleck J.H., Sonenstein F.L. (1998) Adolescent sexual behavior, drug use, and violence: increased reporting with computer survey technology. *Science*, **280**(5365), 867–873, DOI: 10.1126/science.280.5365.867.

Uher R., McGuffin P. (2008) The moderation by the serotonin transporter gene of environmental adversity in the aetiology of mental illness: review and methodological analysis. *Molecular Psychiatry*, **13**(2), 131–146, DOI: 10.1038/sj.mp.4002067.

Ursano R.J., Heeringa S., Stein M.B., Kessler R.C. (submitted for publication) The Army Study to Assess Risk and Resilience in Servicemembers (STARRS).

Voracek M., Loibl L.M. (2007) Genetics of suicide: a systematic review of twin studies. *Wiener Klinische Wochenschrift*, **119**(15–16), 463–475,

Wainer H. (2000) *Computerized Adaptive Testing: A Primer*, Second Edition, Mahwah, NJ: Erlbaum.

Wakschlag L.S., Kistner E.O., Pine D.S., Biesecker G., Pickett K.E., Skol A.D., Dukic V., Blair R.J., Leventhal B.L., Cox N.J., Burns J.L., Kasza K.E., Wright R.J., Cook E.H., Jr. (2009) Interaction of prenatal exposure to cigarettes and MAOA genotype in pathways to youth antisocial behavior. *Molecular Psychiatry*, **15**(9), 928–937, DOI: 10.1038/mp.2009.22.

Zill P., Buttner A., Eisenmenger W., Moller H.J., Bondy B., Ackenheil M. (2004) Single nucleotide polymorphism and haplotype analysis of a novel tryptophan hydroxylase isoform (TPH2) gene in suicide victims. *Biological Psychiatry*, **56**(8), 581–586, DOI: 10.1016/j.biopsych.2004.07.015.

International Journal of Methods in Psychiatric Research
Int. J. Methods Psychiatr. Res. 22(4): 288–302 (2013)
Published online in Wiley Online Library
(wileyonlinelibrary.com) **DOI:** 10.1002/mpr.1399

Response bias, weighting adjustments, and design effects in the Army Study to Assess Risk and Resilience in Servicemembers (Army STARRS)

RONALD C. KESSLER,[1] STEVEN G. HEERINGA,[2] LISA J. COLPE,[3] CAROL S. FULLERTON,[4] NANCY GEBLER,[2] IRVING HWANG,[1] JAMES A. NAIFEH,[4] MATTHEW K. NOCK,[5] NANCY A. SAMPSON,[1] MICHAEL SCHOENBAUM,[3] ALAN M. ZASLAVSKY,[1] MURRAY B. STEIN[6,7] & ROBERT J. URSANO[4]

1 Department of Health Care Policy, Harvard Medical School, Boston, MA, USA
2 University of Michigan, Institute for Social Research, Ann Arbor, MI, USA
3 National Institute of Mental Health, Bethesda, MD, USA
4 Center for the Study of Traumatic Stress, Department of Psychiatry, Uniformed Services University School of Medicine, Bethesda, MD, USA
5 Department of Psychology, Harvard University, Cambridge, MA, USA
6 Departments of Psychiatry and Family and Preventive Medicine, University of California San Diego, La Jolla, CA, USA
7 VA San Diego Healthcare System, San Diego, CA, USA

Key words
suicide, mental disorders, US Army, epidemiologic research design, design effects, sample bias, sample weights, survey design efficiency, survey sampling

Correspondence
Ronald C. Kessler, Department of Health Care Policy, Harvard Medical School, Boston, MA, USA.
Telephone (+1) 617-432-3587,
Fax (+1) 617-432-3588
Email: NCS@hcp.med.harvard.edu

Received 10 July 2013;
accepted 15 July 2013

Abstract

The Army Study to Assess Risk and Resilience in Servicemembers (Army STARRS) is a multi-component epidemiological and neurobiological study designed to generate actionable recommendations to reduce US Army suicides and increase knowledge about determinants of suicidality. Three Army STARRS component studies are large-scale surveys: one of new soldiers prior to beginning Basic Combat Training (BCT; $n = 50{,}765$ completed self-administered questionnaires); another of other soldiers exclusive of those in BCT ($n = 35{,}372$); and a third of three Brigade Combat Teams about to deploy to Afghanistan who are being followed multiple times after returning from deployment ($n = 9421$). Although the response rates in these surveys are quite good (72.0–90.8%), questions can be raised about sample biases in estimating prevalence of mental disorders and suicidality, the main outcomes of the surveys based on evidence that people in the general population with mental disorders are under-represented in community surveys. This paper presents the results of analyses designed to determine whether such bias exists in the Army STARRS surveys and, if so, to develop weights to correct for these biases. Data are also presented on sample inefficiencies introduced by weighting and sample clustering and on analyses of the trade-off between bias and efficiency in weight trimming. *Copyright © 2013 John Wiley & Sons, Ltd.*

Introduction

The Army Study to Assess Risk and Resilience in Servicemembers (Army STARRS; http://www.army-starrs.org) is a multi-component epidemiological and neurobiological study of risk and resilience factors for suicidality and its psychopathological correlates among US Army personnel (Ursano *et al.*, under review). One of these components, the Historical Administrative Data Study (HADS) is a study examining associations among information collected on all soldiers (2004–2009) using Army and Department of Defense (DoD) administrative data records to predict suicide outcomes. Two others are retrospective case–control studies of suicide attempts and fatalities. The other main component studies in Army STARRS are three large-scale surveys (Kessler *et al.*, 2013). One of these, the New Soldier Study (NSS), attempted to obtain information from self-administered neurocognitive tests and self-administered questionnaires (SAQs) in a representative sample of over 57,000 of new soldiers reporting for Basic Combat Training (BCT) (Heeringa *et al.*, 2013). The second, the All-Army Study (AAS), attempted to obtain SAQ information from a representative sample of nearly 50,000 soldiers other than those in BCT (Heeringa *et al.*, 2013). The third, the Pre-Post Deployment Study (PPDS), attempted to obtain SAQ information from all 10,380 members of three Brigade Combat Teams scheduled to deploy to Afghanistan shortly after the baseline PPDS was carried out (Heeringa *et al.*, 2013). The NSS and PPDS additionally attempted to collect blood samples from all respondents, while all three studies attempted to obtain informed consent from SAQ respondents to link their Army/ DoD administrative records with their self-report responses.

An important characteristic of the Army STARRS surveys is that identifying information is needed from SAQ respondents to link administrative records with SAQ data. Concerns can be raised about the absence of anonymity in this design, as some military researchers have suggested that lack of anonymity can lead to under-reporting of emotional problems in military surveys (Warner *et al.*, 2008; Warner *et al.*, 2007). A number of large military surveys, like the DoD Survey of Health Related Behaviors Among Active Duty Military Personnel (DoD Health Behavior Surveys; Ryan *et al.*, 2007) and the Mental Health Surveillance Surveys in combat environments carried out by US Army Mental Health Advisory Teams (MHATs; Bliese *et al.*, 2011), are administered anonymously based on this concern in an effort to encourage complete and accurate reporting.

A good deal of methodological research has been carried out on the effects of anonymity in surveys. One line of this research investigates the effects of experimentally manipulating perceived risk of disclosure of survey responses (Couper *et al.*, 2008, 2010). These studies find that only when risk of disclosure is virtually certain and the information in the survey is potentially damaging to the individual does risk of disclosure reduce survey response rates. Emphasizing the confidentiality of responses in identified surveys, however, has been shown consistently to increase survey response rates significantly (Edwards *et al.*, 2009). Based on this evidence, the informed consent sessions preceding the Army STARRS surveys were designed to be quite elaborate (30-minute group-based sessions) and presented detailed information on the tight security measures put in place to guarantee survey response confidentiality.

A second line of experimental research investigates the effects of anonymity on honesty of responding to sensitive questions among people who participate in surveys. The results of this research are mixed, with some studies showing that anonymity increases reports of embarrassing behaviors (Ong and Weiss, 2000; Werch, 1990) and others finding no such effects (Brink, 1995; Campbell and Waters, 1990). It is unclear why this variability exists, but it has been found even in studies examining the same types of behaviors (Begin *et al.*, 1979; Fidler and Kleinknecht, 1977). A broader experimental literature documents effects of "social distance" on reporting of potentially embarrassing behaviors even within anonymous surveys, with highest reported rates in self-administered surveys, lower rates in telephone surveys, and lowest rates in face-to-face surveys (Rogers *et al.*, 1998; Turner *et al.*, 1998).

Non-experimental studies have also been carried out on this issue. For example, a meta-analysis of studies designed to estimate prevalence of major depression in surveys of military samples found that anonymous surveys, all else equal, yielded higher prevalence estimates than confidential surveys that were not anonymous (Gadermann *et al.*, 2012). The most dramatic non-experimental evidence for such an effect came in a study of responses to the Post-Deployment Health Assessment (PDHA) in a sample of infantry soldiers returning from Iraq (Warner *et al.*, 2011). Completion of the PDHA is required of all soldiers returning from deployment. PDHA responses are neither anonymous nor confidential, as each soldier who completes a PDHA is required to have an in-person review of responses with a health care provider and to discuss deployment-related health problems reported in the survey and to allow the health care professional an opportunity to provide referrals for needed treatment (http://www.pdhealth.mil/dcs/dd_form_2796.asp). The effects of this lack of

confidentiality on PDHA responses were examined by administering a completely anonymous survey containing some of the same questions as the PDHA about emotional problems to a group of soldiers shortly after they completed the PDHA. Reported prevalence of depression was over three times as high in the anonymous survey as in the PDHA (7.0% versus 1.9%, $\chi^2_1 = 87.7$, $p < 0.001$), with similar differences found for a number of other reports, such as having symptoms of post-traumatic stress disorder (PTSD) (7.7% versus 3.3%, $\chi^2_1 = 48.9$, $p < 0.001$) and of having thoughts-concerns about losing control or hurting someone (8.6% versus 3.4%, $\chi^2_1 = 63.1$, $p < 0.001$).

A number of factors could be involved in the dramatic under-reporting of emotional problems in the PDHA, as respondents know with certainty that their responses will be reviewed in a meeting with a health professional. The situation is quite different, of course, in the Army STARRS surveys, where respondents are guaranteed that their self-reports will be used only for research purposes, that personally identifying information will never be linked to research data, that the identifying information they provide will be maintained securely by the civilian academic research team carrying out the study, and that this identifying information will never be shared with the Army. It is unclear whether lack of anonymity will affect reports of emotional problems in a situation of this sort.

In an effort to address the possibility of such an effect in the Army STARRS surveys, a strategic decision was made to allow Army STARRS survey respondents to provide completely anonymous survey reports. This was done by asking first for informed consent to complete the survey and then asking separately for identifying information to link survey data to administrative data. Importantly, the survey cooperation rates (i.e. the proportions of soldiers attending the consent sessions that agreed to complete the surveys) were comparable to those achieved in anonymous surveys of similar samples (Heeringa *et al.*, 2013). However, meaningful proportions of SAQ respondents in the three surveys chose not to provide identifying information: 22.9% in the NSS ($n = 11{,}633$), 31.4% in the AAS ($n = 11{,}106$), and 21.2% in the baseline PPDS ($n = 1996$). These respondents would presumably either not have completed the surveys or would have under-reported emotional problems in the surveys if the option for anonymous reporting was not provided.

Access to these anonymous surveys made it possible for us to compare the characteristics of soldiers who completed anonymous versus confidential (i.e. not anonymous) surveys. Furthermore, we had access not only to the Army/DoD administrative records of all respondents who completed confidential (i.e. non-anonymous surveys in which respondents provided identifying information for purposes of linking the SAQ responses to their administrative records) but also to a limited amount of de-identified individual-level administrative record data for all soldiers in the Army. The latter data were provided by the Army for purposes of sample post-stratification. We were able to use these data to make part-whole comparisons aimed at investigating basic differences between survey respondents who consented to administrative data linkage and all other soldiers (i.e. both those who did not complete the survey and those who completed the survey but did consent to provide the identifying data needed to link survey responses to administrative records). These comparisons were used to evaluate response bias in the Army STARRS surveys and to develop weighting adjustments designed to correct for these biases to the extent possible by adjusting for two types of differences: (i) differences between the anonymous survey sample and the de-identified survey sample in variables assessed in the survey; and (ii) differences between the de-identified survey sample and the population in variables available in the Army/DoD administrative records. The results of these analyses are presented in the current report. Data are also presented on sample inefficiencies introduced by weighting and time-space clustering and on analyses of the trade-off between bias and efficiency in weight trimming.

Data adjustments and processing

Sample clustering

The time-space clustering of observations in the NSS, AAS, and PPDS studies could lead to inefficiencies in estimation due to increases in the variances of statistics estimated from the survey data (Heeringa *et al.*, 2010). To obtain correct estimates of variances and associated inferences about the survey population, we used design-based methods of estimation (Wolter, 1985) that required us to define strata and within-stratum sampling error calculation units (SECUs) for each sample to characterize the sample design stratification and the time-space clustering of observations within strata. In the case of the NSS, this was done by beginning with the fact that each week between January 2011 and November 2012 NSS group-administered SAQ data collections were conducted with 200 to 400 new soldiers at each of three Army training installations shortly after they arrived for BCT. Both the implicit stratification of the sample by location and time and the "clustering effects" of weekly administrations to groups of incoming soldiers introduced complex design effects. (The weighting of observations, discussed later in the sub-section on case-level missing data, also contributes to

Int. J. Methods Psychiatr. Res. 22(4): 288–302 (2013). DOI: 10.1002/mpr

design effects.) The NSS "two SECU-per-stratum" sampling error calculation model for design-based variance was formed by first defining pseudo strata based on the training facility location of the survey and bi-weekly windows of time. Each of the weekly time-space clusters of respondents was defined as a separate SECU and two-week pairs of SECUs were combined at a specific BCT installation to define strata to capture the stratification influences on time-space clustering. The two-SECU coding approach, while not necessary, was chosen because of its flexibility in permitting design-based variance estimation under both the Taylor Series Linearization (TSL), Balanced Repeated Replication (BRR) and Jackknife Repeated Replication (JRR) methods. The same sampling error calculation model also permits analysts the option to use Bootstrap methods of inference for the complex sample of NSS observations.

The AAS, in comparison, was selected in quarterly replicates at the unit level stratified by Army command and unit size within command. Large units from substrata within commands (where computer-administered interviewing [CAI] was the data collection mode) were typically treated as pseudo-self representing (SR) units and split into two random SECU groups for variance estimation purposes. Splitting was done at the session level whenever possible and at the individual soldier level for units that were surveyed in a single session. Non-SR smaller units were usually paired with another similar unit within the same command and quarterly time period to create a sampling error stratum for variance estimation. Unit pairing was always carried out not only within command, but also within size stratum and survey mode (i.e. either CAI or paper-and-pencil interviewing [PAPI]) in order to allow data to be analyzed within meaningful subgroups of interest (e.g. United States Army Forces Command [FORSCOM]-only, CAI-only, etc.) while still maintaining the ability to perform design-based variance estimation.

The PPDS sample, finally, consisted of all soldiers in three Brigade Combat Teams scheduled to deploy to Afghanistan (and return) in the 2011–2012 time frame. Two of the three were Infantry Brigade Combat Teams (one light infantry, the other airborne), each consisting of six battalions (two infantry and one each of cavalry, fires, special troops, and support) and the third was a Stryker (mechanized infantry) Brigade Combat Team consisting of three infantry battalions, one artillery battalion, one support battalion, a number of separate companies (network support, military intelligence, engineer, anti-tank, and headquarters), and one cavalry squadron.

PPDS was designed as a "census" of all soldiers in these three Brigade Combat Teams. While the three Brigade Combat Teams in the PPDS were selected purposefully because of their deployment schedule, a design-based approach to PPDS estimation and inference serves to capture the influence of non-response and post-stratification weighting adjustments on the sampling error of statistics estimated from the PPDS data The design-based sampling error calculation model developed for the analysis of these data effectively treats the three Brigade Combat Teams as a sample from a super-population of all possible such units that underwent a similar deployment experience. A two SECU-per-stratum sampling error calculation model for PPDS design-based variance estimation was formed by first randomly creating strata of 50 to 100 soldiers within each of these units and then further randomly creating half-samples of soldiers within each of these strata to define SECUs. The two-SECU-per-stratum coding approach, as noted earlier, is not the only one that could have been used to estimate variances, but was used here because of its flexibility in allowing implementation of design-based variance estimation methods of the sort used in substantive analyses of the Army STARRS data.

Adjusting for item-level non-response

Item-missing data are generally more common in SAQs than interviewer-administered surveys. Army STARRS is no exception to this rule, as indicated by the fact that a meaningful proportion of SAQ respondents failed to complete all SAQ items (Heeringa *et al.*, 2013). In addition, sporadic item-level missing data could be found in a substantial proportion of completed SAQs. A two-step process was used to address this problem. First, SAQs were coded as missing if the data pattern suggested that respondents were giving random responses or if the amount of missing data was so large that imputation was infeasible. Second, item-level missing data were imputed using a three-part process that began with conservative rational imputation for missing items in sections that had selective missing items. For example, in the section on exposure to traumatic experiences, missing values for respondents that endorsed some items but left others blank were recoded as negative responses. The second part of this three-part process involved psychometric scales, where respondents were assigned a total scale score based on partial values using model-based imputation (e.g. estimated true score values on an item response theory [IRT] scale). The third part, finally, involved the use of multiple imputation to assign plausible values to item-missing data based on responses to other questions (Schafer, 1999).

Adjusting for case-level missing data

Recruiting difficult-to-reach cases

One way to deal with case-level missing data is to develop special field procedures aimed at tracking, recruiting, and interviewing hard-to-reach cases. These procedures were not used in the NSS, AAS, or baseline PPDS because of logistical constraints. However, these procedures are being used in the third wave of the PPDS follow-up survey by selecting a probability sub-sample of non-respondents at the end of the standard field period and using special tracing procedures, personalized recruitment procedures, and financial incentives to obtain interview data from as many of these cases as possible. Up-weighting of these cases will be used to adjust for the fact that they are being under-represented in the consolidated analysis dataset. Similar procedures will be used in future planned follow-up surveys of the baseline NSS and AAS samples and further follow-ups of the PPDS sample.

Weighting for case-level non-response

As noted in the Introduction, we were able to adjust for case-level missing data by comparing characteristics of respondents with those of non-respondents. This was done in two ways: by comparing SAQ responses of respondents who did versus did not consent to Army/DoD administrative data linkage; and by comparing profiles of SAQ respondents who consented to linkage with population profiles on the small set of administrative record variables (e.g. age, sex, rank) we were given access to for post-stratification. We developed weights based on both of these comparisons to make weighting adjustments for case-level non-response. Weight 1 (WT1) adjusted for discrepancies in SAQ responses of survey completers with versus without record linkage. Weight 2 (WT2) then adjusted for discrepancies in multivariate profiles of weighted (WT1) survey respondents with administrative record linkage versus the population. Each weight was constructed based on an iterative process of stepwise logistic regression analysis designed to arrive at a stable weighting solution. WT1 was the inverse of the probability of agreement to link administrative data with SAQ data in the sample of SAQ completers based on a prediction equation using SAQ responses as predictors. WT2 was the inverse of the probability of completion of the SAQ based on the comparison of SAQ respondents who agreed to linkage and were weighted (WT1) to represent all SAQ respondents compared to the population based on a prediction equation using administrative record variables as predictors.

Inspection of detailed results for the replicates weighted up to now, which consist of NSS and AAS respondents from Q2–4 2011 and the baseline PPDS, shows that survey respondents who consented to administrative record linkage differ from non-consenters in having experienced more stress in their lifetime and the recent past and in having generally higher self-reported rates of Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition (DSM-IV) mental disorders. However, these differences are not dramatic even though they are statistically significant. This is illustrated in Table 1, which shows that linkage consenters across the three main Army STARRS surveys were somewhat more likely than non-consenters to report having 30-day DSM-IV mental disorders, a history of trauma exposure, and a history of head injuries, but that these differences are quite modest in substantive terms despite being significant from a statistical point of view.

The fact that consenters do not differ dramatically from non-consenters leads to the ratio of high to low weights based on the best-fitting logistic regression equations (i.e. the ratio of $1/p_1$ divided by $1/p_{99}$, where p_1 is the predicted probability of consent for respondents at the first percentile of this probability in the sample and p_{99} is the predicted probability of consent for respondent at the 99th percentile of this probability in the sample) being relatively low: 4.2–8.4 for the NSS, 4.9–9.4 for the AAS, and 1.7 for the PPDS. In addition, the bodies of the weight distributions are fairly symmetrical. These distributional characteristics typically reduce the impact of weights on variances of coefficient estimates (Kish, 1976; Little and Vartivarian, 2005)

Inspection of detailed results of the logistic regression equations used to produce WT2 shows that NSS respondents who provided administrative data linkage consent are somewhat younger than the population of all soldiers eligible for the survey and somewhat more likely than soldiers in the population to be female, non-Hispanic White, never married, and Protestant, but less likely to have no religion, and somewhat more highly educated than all soldiers in the population. NSS respondents with linked administrative data are also somewhat more likely than the population to be in the Regular Army rather than the US Air National Guard (USANG) or US Army Reserve (USAR). Some of these patterns are shown in Table 2, where we see that sample versus population differences are modest in substantive terms even though statistically significant.

Similar patterns of statistically significant but substantially modest sample versus population difference in socio-demographic characteristics were found in the AAS, including the sample being somewhat younger, less female (as opposed to more female in the NSS), more

Int. J. Methods Psychiatr. Res. 22(4): 288–302 (2013). DOI: 10.1002/mpr

Table 1. Selected comparisons on self-administered survey (SAQ) questions of SAQ respondents who consented and provided valid information for administrative data record linkage versus those who did not consent in the three main Army STARRS survey samples[1]

	New Soldier Study (Q2–4 2011)[2]				All Army Study (Q2–4 2011)[2]				Baseline Pre-Post Deployment Study			
	Consented		Did not consent		Consented		Did not consent		Consented		Did not consent	
	Percent	(SE)	Percent	(SE)	Percent	(SE)	Percent	(SE)	Percent	(SE)	Percent	(SE)
Current DSM-IV/CIDI disorders												
30-Day MDE	5.3	(0.3)	4.6	(0.4)	3.3	(0.2)	4.3	(0.4)	4.5	0.2	4.3	0.5
30-Day GAD	4.4	(0.3)	3.6	(0.3)	1.2	(0.1)	1.0	(0.2)	2.4	0.2	2.8	0.4
30-Day Panic	4.2	(0.3)	3.6	(0.3)	3.4	(0.2)	3.4	(0.3)	3.4	0.2	3.2	0.4
30-Day IED	10.7*	(0.4)	8.0	(0.5)	7.5	(0.2)	6.6	(0.4)	8.7	0.3	8.1	0.6
30-Day PTSD	4.5	(0.3)	3.7	(0.3)	4.2	(0.2)	4.7	(0.4)	3.2	0.2	3.0	0.4
Any of the above	17.6	(0.5)	15.5	(0.7)	13.2	(0.3)	13.1	(0.6)	14.3	0.4	13.1	0.8
χ^2_5	18.1*				17.3*				3.4			
History of traumatic life stress												
None	16.6*	(0.5)	24.7	(0.8)	27.7*	(0.4)	31.0	(0.8)	16.3*	0.4	20.3	0.9
Low	28.8	(0.6)	28.3	(0.8)	19.7	(0.4)	18.5	(0.7)	32.9	0.5	33.1	1.1
Intermediate	26.4*	(0.6)	22.7	(0.8)	24.9*	(0.4)	22.3	(0.7)	25.5	0.5	23.6	1.0
High	28.2*	(0.6)	24.3	(0.8)	27.7	(0.4)	28.2	(0.8)	25.3	0.5	22.9	0.9
χ^2_3	86.1*				19.1*				21.8*			
History of head injury												
None	38.9*	(0.7)	48.1	(0.9)	51.8	(0.5)	53.4	(0.9)	36.8*	0.6	43.5	1.1
Low	9.3	(0.4)	8.0	(0.5)	11.5	(0.3)	11.7	(0.6)	10.2	0.4	8.8	0.6
Intermediate	24.7*	(0.6)	19.8	(0.7)	19.5	(0.4)	18.4	(0.7)	24.7	0.5	22.3	0.9
High	27.2*	(0.6)	24.1	(0.8)	17.3	(0.3)	16.5	(0.7)	28.3	0.5	25.4	1.0
χ^2_3	69.6*				4.0				31.5*			
(*n*)	(11,802)		(3256)		(5528)		(2977)		(7425)		(1996)	

*Significant difference between respondents who consented versus did not consent at the 0.05 level, two-sided test.

[1]The NSS target sample consisted of a number of new soldiers in Reception Battalion in designated cohorts equal to the numbers we could accommodate in the group survey administration settings established on the training bases. Survey Research staff worked with Army point-of-contacts (POCs) to select a representative sample of new soldiers from each cohort to fill those quotas. These new soldiers were ordered to attend the Army STARRS consent session but then provided voluntary informed consent for participation in the NSS. The AAS and PPDS target samples were all soldiers in designated units who were ordered to attend the Army STARRS consent session but then provided voluntary informed consent for participation in the study.

[2]The NSS and AAS studies were piloted in Q1 2011 absent the questions about suicidality and the safety plan associated with those questions (which did not receive Institutional Review Board [IRB] approval until after the Q1 replicates were fielded). Full implementation started in Q2 2011, which is why this was the first replicate included in the weighting. Data for 2012 are not reported here because weighting of the 2012 NSS and AAS data are being carried out separately using an updated population post-stratification dataset for that year.

Note: MDE, major depressive episode; GAD, generalized anxiety disorder; Panic, panic disorder; IED, intermittent explosive disorder, PTSD, post-traumatic stress disorder.

Int. J. Methods Psychiatr. Res. 22(4): 288–302 (2013). DOI: 10.1002/mpr

non-Hispanic White, more currently married (as opposed to more "never married" in the NSS), less highly educated (as opposed to more highly educated in the NSS), and less likely to have any religion than soldiers in the population. Some of these patterns are shown in Table 3, where we see that the differences between sample and population are quite modest in substantive terms even though they are statistically significant. Differences between the AAS sample and the population in Army career characteristics are more substantial, though, with a higher proportion of the sample than the population in the lower enlisted ranks (E2–4), having somewhat less time in service, and being more likely to have been deployed exactly once (as opposed either never or more than once). More detailed analyses found that respondents in the sample are more likely than the population to be in the Medical Command and less likely to be in Area Service Component Commands (North/South America, Europe/Central/Africa, Pacific) and to have quite different distributions than the population on Military Occupational Specialties (MOS). These differences are due to differential sampling of units in the first year of the AAS. In the case of the baseline PPDS, finally, differences between sample and population were found to be very modest in all respects other than that the sample was more likely to have deployed two or more times.

The substantial sample versus population differences in the AAS in Command and MOS led to the ratio of consolidated weights (i.e. WT1 × WT2) based on the best-fitting logistic regression equations being a good deal higher (53.3) than for the NSS (14.2) or the PPDS (3.8). However, as with WT1, the consolidated WT1 × WT2 distributions were found to be smooth and fairly symmetric in all three surveys, with no evidence of bimodality toward the extremes. In addition, as respondents with suicidality and mental disorders are over-represented in the samples, respondents with the highest weights tend to be those who do not have these outcomes. This, as shown in the next sub-section, minimizes the adverse effects on sample efficiency that might otherwise occur as a result of weighting. However, it is possible that results will differ in the remaining sample replicates. As a result, all weighting calculations will be repeated in future Army STARRS study replicates once data collection is completed. Consolidated weights will then be created that allow for changes in optimal weighting procedures over the course of the study.

Weighting for under-represented time periods in the ARFORGEN cycle

As noted in an earlier paper in this issue (Kessler *et al.*, 2013), the initial AAS replicates were restricted to the continental United States and only later expanded to include units in other parts of the world. It was not until rather late in the data collection period, furthermore, that we were able to add soldiers who were currently deployed to Afghanistan by interviewing these soldiers when they were passing through Kuwait either leaving for or returning from their mid-tour leave. Other than for those deployed soldiers, the AAS replicates under-represented activated USANG and USAR units in the continental United States due to the fact that soldiers in such units typically activated for only a short time before deployment, spent only a short time in the continental United Stated after returning from deployment prior deactivating, and were reluctant to participate in the AAS during either of these short time periods. For a similar reason, the AAS under-represented units that were scheduled to deploy in the near future as well as units that recently returned from deployment. As we know that the suicide rate is related to these fine-grained time distinctions, the AAS is biased in that it under-represents certain time periods in the unit deployment cycle.

In order to capture such subtleties of a unit's location in the ARFORGEN (Army Forces Generation) cycle we added replicates late in the AAS field period to include USANG and USAR units that (i) were scheduled either to deploy soon after completing the AAS or that (ii) recently returned from Afghanistan and were scheduled to deactivate soon after completing the AAS. In addition, the baseline PPDS sample provided us with information about Brigade Combat Teams that were going to deploy shortly after completing an Army STARRS survey. Importantly, this baseline PPDS survey contained all (and more than) the information in the AAS. In addition, the T2 PPDS survey provided us with comparable information for the same respondents approximately three months after they returned from their deployment. Once the data from all these final surveys are available for analysis, we will combine them with the larger AAS sample to construct a composite portrait of the entire Army with appropriate weights for the cross-classification of Command (i.e. Training and Doctrine Command [TRADOC], Forces Command [FORSCOM], Medical Command [MEDCOM], etc.), Component (i.e. Regular Army, USAR, and USANG), and phase of the ARFORGEN cycle to reproduce the actual distribution of the total Army across the cells of this cross-classification for the time period under study.

Design effects

Conventional methods of estimating significance, which assume a simple random sample, do not take the

Int. J. Methods Psychiatr. Res. 22(4): 288–302 (2013). DOI: 10.1002/mpr

Table 2. Selected comparisons on administrative data record variables of weighted self-administered survey respondents who consented and provided valid information for administrative record linkage versus the population in the three main Army STARRS survey samples[1]

	New Soldier Study (Q2–4 2011)[2]				All Army Study (Q2–4 2011)[2]				Baseline Pre-Post Deployment Study			
	Population		Sample		Population		Sample		Population		Sample	
	Percent	(SE)	Percent	(SE)	Percent	(SE)	Percent	(SE)	Percent	(SE)	Percent	(SE)
Age												
17–20	59.7	(0.1)	60.9	(1.0)	7.2*	(0.0)	8.2	(0.4)	14.3	(0.3)	14.3	(0.4)
21–24	25.2	(0.1)	24.5	(0.7)	22.7*	(0.0)	27.9	(0.6)	34.0	(0.3)	33.7	(0.6)
25–29	9.9	(0.1)	9.9	(0.4)	26.0*	(0.0)	27.5	(0.6)	27.4	(0.3)	27.7	(0.5)
30+	5.1	(0.0)	4.7	(0.2)	44.1*	(0.0)	36.4	(0.6)	24.2	(0.3)	24.2	(0.5)
χ^2_3		9.2*				150.2*				0.3		
Sex												
Male	82.4	(0.1)	81.6	(0.8)	85.8*	(0.0)	87.4	(0.4)	94.5	(0.2)	94.4	(0.3)
Female	17.6	(0.1)	18.4	(0.8)	14.2*	(0.0)	12.6	(0.4)	5.5	(0.2)	5.6	(0.3)
χ^2_1		3.9*				11.7*				0.0		
Race/ethnicity												
Non-Hispanic White	63.7*	(0.1)	65.7	(0.7)	61.2*	(0.0)	66.5	(0.6)	68.3	(0.3)	67.9	(0.6)
Non-Hispanic Black	18.8	(0.1)	19.4	(0.5)	20.5*	(0.0)	16.0	(0.5)	12.6	(0.2)	12.6	(0.4)
Hispanic	11.9*	(0.1)	10.4	(0.4)	11.2	(0.0)	10.6	(0.4)	12.1	(0.2)	12.3	(0.4)
Other	5.6*	(0.1)	4.4	(0.3)	7.1	(0.0)	6.9	(0.3)	7.0	(0.2)	7.2	(0.3)
χ^2_3		51.7*				81.0*				0.4		
Marital status												
Never married	85.7*	(0.1)	87.3	(0.4)	32.0	(0.0)	31.2	(0.6)	44.8	(0.4)	44.3	(0.6)
Previously married	1.3*	(0.0)	0.2	(0.0)	6.4*	(0.0)	5.6	(0.3)	3.7	(0.1)	4.3	(0.3)
Currently married	13.0	(0.1)	12.5	(0.4)	61.5*	(0.0)	63.1	(0.6)	51.5	(0.4)	51.4	(0.6)
χ^2_2		83.1*				9.1*				4.2		
Education												
Less than high school[3]	22.3*	(0.1)	17.9	(0.8)	13.1*	(0.0)	11.2	(0.4)	14.9	(0.3)	14.7	(0.5)
High school	69.2*	(0.1)	72.0	(0.7)	60.0*	(0.0)	68.1	(0.6)	70.6	(0.3)	70.9	(0.6)
Some college	1.9	(0.0)	2.0	(0.2)	3.8	(0.0)	3.9	(0.3)	2.5	(0.1)	3.0	(0.2)
College graduate	6.6*	(0.1)	8.0	(0.6)	23.1*	(0.0)	16.8	(0.5)	12.0	(0.2)	11.4	(0.4)
χ^2_3		131.4*				167.6*				4.5		
Religion												
Christian: Catholic	13.2	(0.1)	12.6	(0.4)	19.0	(0.0)	19.3	(0.5)	18.6	(0.3)	18.6	(0.5)
Christian: Other	52.8*	(0.1)	55.8	(0.6)	52.3	(0.0)	52.1	(0.7)	49.6	(0.4)	49.1	(0.6)
Other religion	1.5	(0.0)	1.7	(0.2)	2.0*	(0.0)	2.5	(0.2)	1.9	(0.1)	1.9	(0.2)

(*Continues*)

Int. J. Methods Psychiatr. Res. 22(4): 288–302 (2013). DOI: 10.1002/mpr

Table 2. (Continued)

	New Soldier Study (Q2–4 2011)[2]				All Army Study (Q2–4 2011)[2]				Baseline Pre-Post Deployment Study			
	Population		Sample		Population		Sample		Population		Sample	
	Percent	(SE)	Percent	(SE)	Percent	(SE)	Percent	(SE)	Percent	(SE)	Percent	(SE)
No religion	29.2*	(0.1)	26.3	(0.6)	18.8*	(0.0)	22.3	(0.6)	21.3	(0.3)	21.8	(0.5)
Unknown	3.4	(0.0)	3.5	(0.3)	8.0*	(0.0)	3.9	(0.3)	8.6	(0.2)	8.6	(0.4)
χ^2_4	53.4*				154.7*				1.0			
Rank												
E1–4					43.8*	(0.0)	53.4	(0.7)	58.9	(0.4)	58.8	(0.6)
E5–9					38.0*	(0.0)	34.0	(0.6)	32.2	(0.3)	32.8	(0.6)
WO or CO					18.2*	(0.0)	12.6	(0.4)	8.9	(0.2)	8.3	(0.3)
χ^2_2					231.5*				2.1			
Time in Army												
0–24 months					16.9*	(0.0)	19.5	(0.5)	31.9	(0.3)	31.5	(0.5)
25–48 months					19.3*	(0.0)	22.6	(0.6)	23.4	(0.3)	23.8	(0.5)
49+ months					63.8*	(0.0)	57.8	(0.7)	44.8	(0.4)	44.7	(0.6)
χ^2_2					85.7*				0.6			
Deployed												
Never					31.7*	(0.0)	29.3	(0.6)	45.0	(0.4)	43.4	(0.6)
Previously 1					31.6*	(0.0)	36.0	(0.6)	31.0	(0.3)	31.1	(0.6)
Previously 2+					36.7*	(0.0)	34.6	(0.6)	24.0	(0.3)	25.5	(0.5)
χ^2_2					44.0*				7.0*			
(*n*)[4]	(212,797)		(11,802)		(3,528,477)		(5528)		(19,182)		(7425)	

*Significant difference between weighted (WT1) SAQ respondents who consented to record linkage and the population.

[1]The samples were weighted (WT1) to adjust for differences on SAQ variables between SAQ respondents who consented and provided linking information versus those that did not, making the weighted sample representative of all SAQ respondents on the SAQ variables. The population data are taken from contemporary administrative data for the populations from which the samples were selected: all soldiers in BCT for the NSS, all non-deployed Regular Army soldiers not in BCT for the AAS, and all soldiers in the three Brigade Combat Teams included in the survey for the PPDS.

[2]The NSS and AAS studies were piloted in Q1 2011 absent the questions about suicidality and the safety plan associated with those questions (which did not receive IRB approval until after the Q1 replicates were fielded). Full implementation started in Q2 2011, which is why this was the first replicate included in the weighting. Data for 2012 are not reported here because weighting of the 2012 NSS and AAS data are being carried out separately using an updated population post-stratification dataset for that year.

[3]Includes alternative education certificate, Army National Guard (ARNG) and General Educational Development (GED).

[4]The population for the NSS is defined as the pooled monthly snapshot of all soldiers in BCT in the time interval April–November 2011. The population for the AAS is defined as the pooled monthly snapshot of all Regular Army soldiers not in BCT and not deployed over the time interval May–December 2011. The population for the baseline PPDS is defined as the pooled monthly snapshot of all soldiers in the three Brigade Combat Teams in the sample in the months before and during baseline data collection.

Table 3. Design effects on selected 30-day outcome variable prevalence estimates due to survey weighting and clustering in the three main Army STARRS survey samples[1]

	New Soldier Study (Q2–4 2011)[2]	All Army Study (Q2–4 2011)[2]	Baseline Pre-Post Deployment Study
Generalized anxiety disorder	1.5	1.0	1.0
Intermittent explosive disorder	1.2	1.6	1.1
Major depressive episode	1.1	1.8	1.0
Panic disorder	1.3	1.3	1.2
Post-traumatic stress disorder	1.2	1.7	1.0
Suicide ideation	1.2	1.5	1.1
Any of the above	1.1	1.9	1.1
(*n*)	(11,802)	(5428)	(7425)

[1]The samples were doubly weighted to adjust for differences on SAQ variables between SAQ respondents who consented and provided linking information for administrative data versus those that did not (WT1) and between the weighted (WT1) sample of SAQ respondents with linked ADS data and the population (WT2).
[2]The NSS and AAS studies were piloted in Q1 2011 absent the questions about suicidality and the safety plan associated with those questions (which did not receive IRB approval until after the Q1 replicates were fielded). Full implementation started in Q2 2011, which is why this was the first replicate included in the weighting. Data for 2012 are not reported here because weighting of the 2012 NSS and AAS data are being carried out separately using an updated population post-stratification dataset for that year.

imprecision introduced by clustering and weighting into account. As a result, special design-based methods of estimating standard errors and significance tests are used in Army STARRS analyses to adjust for the effects of weighting and clustering. The TSL method is the main approach used here (Wolter, 1985), although we also use the more computationally intensive method of JRR (Kish and Frankel, 1974) for applications where a convenient software application using the TSL method is not readily available or for highly non-linear estimation problems in which the linearization of the TSL method might be problematic.

Although the effects of weighting and clustering can be described in a number of ways, a particularly convenient way is to calculate a statistic known as the design effect (DE; Kish, 1965) for a number of variables of interest. The DE is the square of the ratio of the design-based standard error (SE) of a descriptive statistic divided by the simple random sample SE. The DE can be interpreted as the approximate proportional increase in the sample size that would be required to increase the precision of the design-based estimate to the precision of an estimate based on a simple random sample of the same size. DEs due to clustering are usually a good deal larger in estimating means and other first-order statistics than more complex statistics, as the number of respondents having the same characteristics in the same SECU of a single stratum becomes smaller and smaller as the statistics become more complex. This leads to a reduction in the effects of clustering in the estimation of DE. DEs due to weighting are also usually somewhat smaller for multivariate than bivariate descriptive statistics because DEs are due not only to the variance in the weights but also to the strength of the association between the weights and the substantive variables under consideration. Because means typically have higher DEs than other statistics, evaluations of DEs typically focus on the estimation of means. We do the same here.

Seven dichotomous measures of 30-day prevalence of critical outcome variables were included in the evaluation of DEs: suicide ideation and DSM-IV disorder estimates for major depressive episode, generalized anxiety disorder, PTSD, panic disorder, intermittent explosive disorder, and any of the above six outcomes. DEs for these estimates are in the range 1.1–1.5 for the NSS, 1.0–1.9 for the AAS, and 1.0–1.2 for the PPDS. (Table 3) The fact that a number of DEs are 1.0 (i.e.,equal in efficiency to a simple random sample) or only slightly higher than 1.0 can be explained by the same general pattern of the samples with linked administrative data over-representing soldiers with the disorders that are the focus of interest in Army STARRS.

Trimming weights to reduce design effects

As DEs can be sensitive to extreme weights, weight trimming of various sorts is often used to reduce this

sensitivity. We investigated the implications of trimming the final consolidated weight (WT1 × WT2) in each survey. In doing this we took into consideration the fact that even though weight trimming usually reduces the variance of weights, and in this way improves the precision of estimates and the statistical power of tests, it can also lead to bias in estimates if the reduction in variance created due to added efficiency is less than the increase in variance due to bias. It is possible to study this trade-off between bias and efficiency empirically in order to evaluate alternative weight trimming schemes by making use of the equation

$$\mathrm{MSE}_{Yp} = B_{Yp}{}^2 + \mathrm{Var}\left(Y_p\right), \tag{1a}$$

$$= E\left[\left(\hat{B}_{Yp}\right)^2 - \mathrm{Var}\left(\hat{B}_{Yp}\right) + \mathrm{Var}\left(Y_p\right)\right], \tag{1b}$$

where MSE_{Yp} is the mean squared error of the prevalence of outcome variable Y at trimming point p, B_{Yp} is the bias of that prevalence estimate and $\hat{B}_{Yp}$, an unbiased estimate of that bias, $\widehat{\mathrm{Var}}\left(\hat{B}\right)_{Yp}$), is the estimated variance of $\hat{B}_{Yp}$, $\mathrm{Var}(\hat{Y}_p)$ is the estimated variance of estimate $\hat{Y}_p$, and $E[\ ]$ in Equation 1b indicates that the quantity in square brackets is an unbiased estimator of MSE.

Each of the three terms in Equation 1b can be estimated empirically for any value of p, making it possible to calculate MSE across a range of trimming points and select the trimming point that minimizes MSE. The first term, $(\hat{B}_{Yp})^2$, can be estimated directly as $(Y_p - Y_0)^2$, where Y_0 represents the weighted prevalence estimate of Y based on the untrimmed weight. The other two terms in Equation 1b can be estimated using a pseudo-replicate method in which separate estimates for each stratum-SECU are generated for Y_p at each value of p (Zaslavsky *et al.*, 2001). The separate estimates were obtained by sequentially modifying the sample and then generating an estimate based on that modified sample. The modification consisted of removing all cases from one SECU and then weighting the cases in the remaining SECU in the same stratum to have a sum of weights equal to the original sum of weights in that stratum. If we define Y_p as the weighted estimate of Y at trimming point p in the total sample and we define $Y_{p(sn)}$ as the weighted estimate at the same trimming point in the sample that deletes SECU n ($n = 1, 2$) of stratum s ($s = 1$–42), then $\mathrm{Var}(Y_p)$ can be estimated as

$$\widehat{\mathrm{Var}}\left(\hat{Y}_p\right) = \sum_s \left[\left(\hat{Y}_{p(s1)} - Y_p\right)^2 + \left(\hat{Y}_{p(s2)} - \hat{Y}_p\right)^2\right]/2. \tag{2}$$

$\mathrm{Var}(\hat{B}_{Yp})$ was estimated in the same fashion by replacing $\hat{Y}_{p(sn)}$ in Equation 2 with $\hat{B}_{Yp(sn)} = \hat{Y}_{p(sn)} - \hat{Y}_{0(sn)}$ and replacing $\hat{Y}_p$ with $\hat{B}_{Yp(sn)} = \hat{Y}_p - \hat{Y}_0$.

The analysis compared the design-based MSE of 30-day prevalence estimates for the same outcomes as considered in the last sub-section using the consolidated WT1 × WT2 weight and 10 successively more severely trimmed versions of these weights in which between 1% and 10% of cases were trimmed at each tail of the distribution. Trimming consisted of distributing the weights at each of these tails equally across all cases in that tail. MSE_{Y0} was arbitrarily set at 100.0 and all other values were defined in relation to that mean for ease of interpretation. Summary results for illustrative trimming points are presented in Table 4. In the cases of NSS and AAS, while weight trimming reduced MSE for some outcomes (most notably, generalized anxiety disorder in the NSS and major depressive episode in the AAS), it increased MSE for other outcomes, leading us to decide not to trim the consolidated weight for either survey. In the case of PPDS, while the effects of weight trimming were generally positive, they were so modest that we decided not to trim the consolidated weight. As with the weights themselves, it is possible that results regarding the value of weight trimming will differ in the remaining sample replicates. As a result, all weight trimming calculations will be repeated in future Army STARRS study replicates once data collection is completed. Consolidated weight trimming rules will then be created that allow for changes in optimal trimming procedures over the course of the study.

Discussion

As noted in the Introduction, our reading of previous methodological literature led us to expect that Army STARRS survey respondents who agreed to administrative record linkage would have lower rates of self-reported mental disorder than survey respondents who provided identifying information both because those with mental disorders would be less likely to consent to record linkage and because those who did consent would under-report emotional problems. Yet the opposite pattern was found in the data when we examined the predictors of WT1: SAQ respondents who consented to administrative record linkage had significantly higher, not lower, self-reported rates of mental illness than SAQ respondents who did not consent to record linkage.

Why this pattern occurred is unclear. One possibility is that it reflects a positive effect of the message used in respondent recruitment: that Army STARRS is an *independent* research project carried out by academic researchers *outside of the Army* that represents a *unique opportunity* for soldiers to let Army leadership know about issues they are experiencing in the realms of work-related stress and

Int. J. Methods Psychiatr. Res. 22(4): 288–302 (2013). DOI: 10.1002/mpr

Table 4. Effects of weight trimming on the bias-efficiency trade-off for selected outcome variable prevalence estimates in the three main Army STARRS survey samples[1]

	New Soldier Study Q2–4 2011[2]			All Army Study Q2–4 2011[2]			Baseline Pre-Post Deployment Study		
	Trimming point[3]			Trimming point[3]			Trimming point[3]		
	2	5	10	2	5	10	2	5	10
Generalized anxiety disorder	95.9	85.3	72.7	98.7	121.7	192.5	100.2	100.6	98.5
Intermittent explosive disorder	99.8	102.6	82.0	109.1	99.4	101.8	100.1	100.7	100.2
Major depressive episode	105.1	90.9	87.8	116.9	65.5	76.6	98.3	97.5	95.4
Panic disorder	101.4	93.4	94.4	102.9	93.5	101.9	99.2	99.7	97.0
Post-traumatic stress disorder	100.6	111.1	120.2	99.7	138.2	172.4	97.2	95.9	94.9
Suicide ideation	101.2	100.6	148.6	96.1	100.2	148.2	95.6	95.3	92.3
Any 30-day disorder	99.8	89.1	94.2	102.9	102.3	96.0	99.7	99.6	97.4
(*n*)		(11,802)			(5428)			(7425)	

[1]The samples were doubly weighted to adjust for differences on SAQ variables between SAQ respondents who consented and provided linking information for ADS data versus those that did not (WT1) and between the weighted (WT1) sample of SAQ respondents with linked ADS data and the population (WT2).
[2]The NSS and AAS studies were piloted in Q1 2011 absent the questions about suicidality and the safety plan associated with those questions (which did not receive IRB approval until after the Q1 replicates were fielded). Full implementation started in Q2 2011, which is why this was the first replicate included in the weighting. Data for 2012 are not reported here because weighting of the 2012 NSS and AAS data are being carried out separately using an updated population post-stratification dataset for that year.
[3]The trimming point is the proportion of respondents trimmed at each tail of the distribution. Four per cent of respondents (2% at the upper end of the distribution and 2% at the lower end of the distribution) were trimmed in the solution with a trimming point of 2, 10% at a trimming point of 5, and 20% at a trimming point of 10. See the text for a more detailed description of the procedures and rationale for trimming. Results for trimming points at each whole number value in the range 1–10 are available on request.

Int. J. Methods Psychiatr. Res. 22(4): 288–302 (2013). DOI: 10.1002/mpr

emotional problems. This recruitment message went on to say that only a small proportion of all soldiers were invited to participate in the survey, that each respondent's voice consequently speaks for many, and that it is important for those few soldiers who are invited to take advantage of this opportunity to have their voices heard by Army leadership in a fashion that protects confidentiality. This message was presented to all potential Army STARRS survey respondents both in a Study Information Sheet distributed prior to the informed consent session and in the informed consent session. The Army STARRS data collection team worked very closely with local Army Points of Contact to mount a campaign for survey participation while distributing Study Fact Brochures. They also emphasized the high-profile nature of Army STARRS and made it clear that survey results would be used at the highest levels of Army leadership. This recruitment message and the aggressive campaign mounted to disseminate this message might have encouraged both a high response rate and also encouraged soldiers with mental disorders to admit having these disorders, leading to the high reported rates of emotional problems among soldiers who agreed to administrative record linkage.

It is important to put the Army STARRS response rates in perspective by noting that these response rates are a good deal higher than those in a number of other major military surveys, including in surveys that offered complete anonymity to survey respondents (Bray *et al.*, 2006; Ryan *et al.*, 2007). As noted by Heeringa and colleagues in a companion paper in this issue (Heeringa *et al.*, 2013), these high Army STARRS response rates are due to higher proportions of pre-designated respondents in Army STARRS than previous surveys attending the consent sessions coupled with equally or higher proportions of those attending these sessions in Army STARRS than previous surveys agreeing to participate.

The high overall response rates in the Army STARRS surveys had an important implication for WT2, where we compared Army/DoD administrative record data in the population of all soldiers with those in the weighted (WT1) subset of soldiers who both completed the Army STARRS SAQ and provided administrative record linkage. This analysis failed to find evidence of significant differences between the weighted (WT1) sample and the population on a variety of administrative record variables. As a result, while it was important to weight the SAQ data for soldiers who consented to administrative data linkage, this was because failure to do so would have led to *over*-estimation rather than *under*-estimation of mental disorder prevalence in the de-identified survey data.

As we saw in the analysis of DEs, this over-representation of soldiers with mental disorders improved efficiency in estimating prevalence and correlates of these outcomes. Another important finding in this part of the analysis was that the distributions of the consolidated weights were fairly symmetrical and had a relatively narrow range. Taken together, these weight characteristics led to the finding, reported in Table 3, that DEs for the self-reported outcomes of central interest to the initiative are all quite modest.

One important limitation of the earlier analysis is that the weighting adjustments are based on the assumption that self-reports of mental disorders are as valid in the sample of respondents who provided de-identified SAQs as in the sample whose SAQ reports are completely anonymous. This need not be the case. The definitive evaluation of this issue would have required us to carry out an experiment in which a probability sub-sample of soldiers selected to participate in an Army STARRS survey were asked to provide completely anonymous survey data without the option to provide identifying information for administrative record linkage. We did not carry out that experiment. This means that even though prevalence estimates of the disorders assessed in the Army STARRS surveys are higher in the de-identified than anonymous SAQ sub-samples, it might still be the case that prevalence estimates would have been higher yet among respondents whose SAQs are not completely anonymous if they had never been asked to provide identifying information. There is no way to assess this possibility with the data available to us here, but it is a possibility that needs to be kept in mind when interpreting substantive results. To the extent that this bias exists, prevalence estimates of these disorders in the weighted Army STARRS survey data should be considered conservative.

Acknowledgments

On behalf of the Army STARRS Collaborators

Funding/Support

Army STARRS was sponsored by the Department of the Army and funded under cooperative agreement number U01MH087981 with the US Department of Health and Human Services, National Institutes of Health, National Institute of Mental Health (NIH/NIMH). The contents are solely the responsibility of the authors and do not necessarily represent the views of the Department of Health and Human Services, NIMH, the Department of the Army, or the Department of Defense.

Role of the Sponsors

As a cooperative agreement, scientists employed by NIMH (Colpe and Schoenbaum) and Army liaisons/consultants (COL Steven Cersovsky, MD, MPH USAPHC and Kenneth Cox, MD, MPH USAPHC) collaborated to develop the study protocol and data collection instruments, supervise data collection, plan and supervise data analyses, interpret results, and prepare reports. Although a draft of this manuscript was submitted to the Army and NIMH for review and comment prior to submission, this was with the understanding that comments would be no more than advisory.

Additional Contributions

The Army STARRS Team consists of Co-Principal Investigators: Robert J. Ursano, MD (Uniformed Services University of the Health Sciences) and Murray B. Stein, MD, MPH (University of California San Diego and VA San Diego Healthcare System); Site Principal Investigators: Steven Heeringa, PhD (University of Michigan) and Ronald C. Kessler, PhD (Harvard Medical School); NIMH collaborating scientists: Lisa J. Colpe, PhD, MPH and Michael Schoenbaum, PhD; Army liaisons/consultants: COL Steven Cersovsky, MD, MPH (USAPHC) and Kenneth Cox, MD, MPH (USAPHC). Other team members: Pablo A. Aliaga, MA (Uniformed Services University of the Health Sciences); COL David M. Benedek, MD (Uniformed Services University of the Health Sciences); Susan Borja, PhD (National Institute of Mental Health); Gregory G. Brown, PhD (University of California San Diego); Laura Campbell-Sills, PhD (University of California San Diego); Catherine Dempsey, PhD, MPH (Uniformed Services University of the Health Sciences); Richard Frank, PhD (Harvard Medical School); Carol S. Fullerton, PhD (Uniformed Services University of the Health Sciences); Nancy Gebler, MA (University of Michigan); Joel Gelernter, MD (Yale University); Robert K. Gifford, PhD (Uniformed Services University of the Health Sciences); Stephen E. Gilman, ScD (Harvard School of Public Health); Marjan G. Holloway, PhD (Uniformed Services University of the Health Sciences); Paul E. Hurwitz, MPH (Uniformed Services University of the Health Sciences); Sonia Jain, PhD (University of California San Diego); Tzu-Cheg Kao, PhD (Uniformed Services University of the Health Sciences); Karestan C. Koenen, PhD (Columbia University); Lisa Lewandowski-Romps, PhD (University of Michigan); Holly Herberman Mash, PhD (Uniformed Services University of the Health Sciences); James E. McCarroll, PhD, MPH (Uniformed Services University of the Health Sciences); Katie A. McLaughlin, PhD (Harvard Medical School); James A. Naifeh, PhD (Uniformed Services University of the Health Sciences); Matthew K. Nock, PhD (Harvard University); Rema Raman, PhD (University of California San Diego); Nancy A. Sampson, BA (Harvard Medical School); LCDR Patcho Santiago, MD, MPH (Uniformed Services University of the Health Sciences); Michaelle Scanlon, MBA (National Institute of Mental Health); Jordan Smoller, MD, ScD (Harvard Medical School); Nadia Solovieff, PhD (Harvard Medical School); Michael L. Thomas, PhD (University of California San Diego); Christina Wassel, PhD (University of Pittsburgh); and Alan M. Zaslavsky, PhD (Harvard Medical School). The authors would also like to thank John Mann, Maria Oquendo, Barbara Stanley, Kelly Posner, and John Keilp for their contributions to the early stages of Army STARRS development.

Additional Information

A complete list of Army STARRS publications can be found at http://www.ARMYSTARRS.org.

Declaration of interest statement

In the past five years Kessler has been a consultant for Eli Lilly & Company, Glaxo, Inc., Integrated Benefits Institute, Ortho-McNeil Janssen Scientific Affairs, Pfizer Inc., Sanofi-Aventis Groupe, Shire US Inc., and Transcept Pharmaceuticals Inc. and has served on advisory boards for Johnson & Johnson. Kessler has had research support for his epidemiological studies over this time period from Eli Lilly & Company, EPI-Q, GlaxoSmithKline, Ortho-McNeil Janssen Scientific Affairs, Sanofi-Aventis Groupe, Shire US, Inc., and Walgreens Co. Kessler owns a 25% share in DataStat, Inc. Stein has in the last three years been a consultant for Healthcare Management Technologies and had research support for pharmacological imaging studies from Janssen. The remaining authors report no competing interests.

References

Begin G., Boivin M., Bellerose J. (1979) Sensitive data collection through the random response technique: some improvements. *Journal of Psychology*, **101**(1), 53–65.

Bliese P.D., Thomas J.L., McGurk D., McBride S., Castro C.A. (2011) Mental health advisory teams: a proactive examination of mental health during combat deployments. *International Review of Psychiatry*, **23**(2), 127–134, DOI: 10.3109/09540261.2011.558834.

Bray R.M., Hourani L.L., Olmsted K.L.R., Witt M., Brown J.M., Pemberton M.R., Marsden M.E.,

Marriott B., Scheffler S., Vandermaas-Peeler R., Weimer B., Calvin S., Bradshaw M., Close K., Hayden D. (2006) 2005 Department of Defense Survey of Health Related Behaviors Among Active Duty Military Personnel: A Component of the Defense Lifestyle Assessment Program (DLAP), Research Triangle Park, NC, Research Triangle Institute.

Brink T.L. (1995) Sexual behavior and telling the truth on questionnaires. *Psychological Reports*, **76**(1), 218.

Campbell M.J., Waters W.E. (1990) Does anonymity increase response rate in postal questionnaire surveys about sensitive subjects? A randomised trial. *Journal of Epidemiology and Community Health*, **44**(1), 75–76.

Couper M.P., Singer E., Conrad F.G., Groves R.M. (2008) Risk of disclosure, perceptions of risk, and concerns about privacy and confidentiality as factors in survey participation. *Journal of Official Statistics*, **24**(2), 255–275.

Couper M.P., Singer E., Conrad F.G., Groves R.M. (2010) Experimental studies of disclosure risk, disclosure harm, topic sensitivity, and survey participation. *Journal of Official Statistics*, **26**(2), 287–300.

Edwards P.J., Roberts I., Clarke M.J., Diguiseppi C., Wentz R., Kwan I., Cooper R., Felix L. M., Pratap S. (2009) Methods to increase response to postal and electronic questionnaires. *Cochrane Database of Systematic Reviews*, **8**(3), MR000008, DOI: 10.1002/14651858.MR000008.pub4.

Fidler D.S., Kleinknecht R.E. (1977) Random responding versus direct questioning: two data-collection methods for sensitive information. *Psychological Bulletin*, **84**(5), 1045–1049.

Gadermann A.M., Engel C.C., Naifeh J.A., Nock M.K., Petukhova M., Santiago P.N., Wu B., Zaslavsky A.M., Kessler R.C. (2012) Prevalence of DSM-IV major depression among U.S. military personnel: meta-analysis and simulation. *Military Medicine*, **177**(8 Suppl), 47–59.

Heeringa S.G., Colpe L.J., Fullerton C.S., Gebler N., Naifeh J.A., Nock M.K., Sampson N.A., Schoenbaum M., Zaslavsky A.M., Stein M.B., Ursano R.J., Kessler R.C. (2013) Field procedures in the Army Study to Assess Risk and Resilience in Servicemembers (Army STARRS). *International Journal of Methods and Psychiatric Research*, **22**(4), 276–287.

Heeringa S.G., West B.T., Berglund P.A. (2010) Applied Survey Data Analysis. Boca Raton, FL: Taylor and Francis.

Kessler R.C., Colpe L.J., Fullerton C.S., Gebler N., Naifeh J.A., Nock M.K., Sampson N.A., Schoenbaum M., Zaslavsky A.M., Stein M.B., Ursano R.J., Heeringa S.G. (2013) Design of the Army Study to Assess Risk and Resilience in Servicemembers (Army STARRS). *International Journal of Methods in Psychiatric Research*, **22**(4), 267–275.

Kish L. (1965) Survey Sampling. New York: John Wiley and Sons.

Kish L. (1976) Optima and proxima in linear sample designs. *Journal of the Royal Statistical Society, Series A*, **139**(1), 80–95.

Kish L., Frankel M.R. (1974) Inference from complex samples. *Journal of the Royal Statistical Society, Series A*, **36**(1), 1–37.

Little R.J.A., Vartivarian S. (2005) Does weighting for nonresponse increase the variance of survey means? *Survey Methodology*, **31**(2), 161–168.

Ong A.D., Weiss D.J. (2000) The impact of anonymity on responses to "sensitive" questions. *Journal of Applied Social Psychology*, **30**(8), 1691–1708, DOI: 10.1111/j.1559-1816.2000.tb02462.x.

Rogers S.M., Miller H.G., Turner C.F. (1998) Effects of interview mode on bias in survey measurements of drug use: do respondent characteristics make a difference? *Substance Use and Misuse*, **33**(10), 2179–2200, DOI: 10.3109/10826089809069820.

Ryan M.A., Smith T.C., Smith B., Amoroso P., Boyko E.J., Gray G.C., Gackstetter G.D., Riddle J.R., Wells T.S., Gumbs G., Corbeil T.E., Hooper T.I. (2007) Millennium Cohort: enrollment begins a 21-year contribution to understanding the impact of military service. *Journal of Clinical Epidemiology*, **60**(2), 181–191, 10.1016/j.jclinepi.2006.05.009.

Schafer J.L. (1999) Multiple imputation: a primer. *Statistical Methods in Medical Research*, **8**(1), 3–15, DOI: 10.1177/096228029900800102.

Turner C.F., Ku L., Rogers S.M., Lindberg L.D., Pleck J.H., Sonenstein F.L. (1998) Adolescent sexual behavior, drug use, and violence: increased reporting with computer survey technology. *Science*, **280**(5365), 867–873, DOI:10.1126/science.280.5365.867.

Ursano R.J., Heeringa S., Stein M.B., Kessler R.C. (submitted for publication) The Army Study to Assess Risk and Resilience in Servicemembers (STARRS).

Warner C.H., Appenzeller G.N., Grieger T., Belenkiy S., Breitbach J., Parker J., Warner C. M., Hoge C. (2011) Importance of anonymity to encourage honest reporting in mental health screening after combat deployment. *Archives of General Psychiatry*, **68**(10), 1065–1071, DOI: 10.1001/archgenpsychiatry.2011.112.

Warner C.H., Appenzeller G.N., Mullen K., Warner C.M., Grieger T. (2008) Soldier attitudes toward mental health screening and seeking care upon return from combat. *Military Medicine*, **173**(6), 563–569.

Warner C.H., Breitbach J.E., Appenzeller G.N., Yates V., Grieger T., Webster W.G. (2007) Division mental health in the new brigade combat team structure: part II. Redeployment and postdeployment. *Military Medicine*, **172**(9), 912–917.

Werch C.E. (1990) Two procedures to reduce response bias in reports of alcohol consumption. *Journal of Studies on Alcohol*, **51**(4), 327–330.

Wolter K.M. (1985) Introduction to Variance Estimation. New York: Springer-Verlag.

Zaslavsky A.M., Schenker N., Belin T.R. (2001) Downweighting influential clusters in surveys: application to the 1990 Post Enumeration Survey. *Journal of the American Statistical Association*, **96**(455), 858–869, DOI: 10.1198/016214501753208889.

International Journal of Methods in Psychiatric Research
Int. J. Methods Psychiatr. Res. 22(4): 303–321 (2013)
Published online in Wiley Online Library
(wileyonlinelibrary.com) **DOI:** 10.1002/mpr.1398

Clinical reappraisal of the Composite International Diagnostic Interview Screening Scales (CIDI-SC) in the Army Study to Assess Risk and Resilience in Servicemembers (Army STARRS)

RONALD C. KESSLER,[1] PATCHO N. SANTIAGO,[2] LISA J. COLPE,[3] CATHERINE L. DEMPSEY,[2] MICHAEL B. FIRST,[4,5] STEVEN G. HEERINGA,[6] MURRAY B. STEIN,[7,8] CAROL S. FULLERTON,[2] MICHAEL J. GRUBER,[1] JAMES A. NAIFEH,[2] MATTHEW K. NOCK,[9] NANCY A. SAMPSON,[1] MICHAEL SCHOENBAUM,[3] ALAN M. ZASLAVSKY[1] & ROBERT J. URSANO[2]

1 Department of Health Care Policy, Harvard Medical School, Boston, MA, USA
2 Center for the Study of Traumatic Stress, Department of Psychiatry, Uniformed Services University of the Health Sciences, Bethesda, MD, USA
3 National Institute of Mental Health, Bethesda, MD, USA
4 Department of Psychiatry, Columbia University, New York, USA
5 New York State Psychiatric Institute, New York, USA
6 University of Michigan, Institute for Social Research, Ann Arbor, MI, USA
7 Departments of Psychiatry and Family and Preventive Medicine, University of California San Diego, La Jolla, CA, USA
8 VA San Diego Healthcare System, San Diego, CA, USA
9 Department of Psychology, Harvard University, Cambridge, MA, USA

Key words
Composite International Diagnostic Interview (CIDI), CIDI Screening Scales (CIDI-SC), diagnostic concordance, PTSD checklist (PCL), screening scales, validity

Correspondence
Ronald C. Kessler, Department of Health Care Policy, Harvard Medical School, 180 Longwood Avenue, Boston, MA 02115, USA.
Telephone (+1) 617-432-3587 Fax (+1) 617-432-3588
Email: NCS@hcp.med.harvard.edu

Abstract

A clinical reappraisal study was carried out in conjunction with the Army Study to Assess Risk and Resilience in Servicemembers (Army STARRS) All-Army Study (AAS) to evaluate concordance of the Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition (DSM-IV) diagnoses based on the Composite International Diagnostic Interview Screening Scales (CIDI-SC) and post-traumatic stress disorder (PTSD) checklist (PCL) with diagnoses based on independent clinical reappraisal interviews (Structured Clinical Interview for DSM-IV [SCID]). Diagnoses included: lifetime mania/hypomania, panic disorder, and intermittent explosive disorder; six-month adult attention-deficit/hyperactivity disorder; and 30-day major depressive episode, generalized anxiety disorder, PTSD, and substance (alcohol or drug) use disorder (abuse or dependence). The sample ($n = 460$) was weighted for over-sampling CIDI-SC/PCL screened positives. Diagnostic thresholds were set to equalize false positives and false negatives. Good individual-level concordance was found between CIDI-SC/PCL and SCID diagnoses at these thresholds (area under curve

Received 10 July 2013;
accepted 15 July 2013

[AUC] = 0.69–0.79). AUC was considerably higher for continuous than dichotomous screening scale scores (AUC = 0.80–0.90), arguing for substantive analyses using not only dichotomous case designations but also continuous measures of predicted probabilities of clinical diagnoses. *Copyright © 2013 John Wiley & Sons, Ltd.*

Introduction

As described in more detail earlier in this issue (Kessler *et al.*, 2013b) and elsewhere (Ursano *et al.*, submitted for publication), the Army Study to Assess Risk and Resilience in Servicemembers (Army STARRS; http://www.armystarrs.org) is a multi-component epidemiological and neurobiological study of risk and resilience factors for suicidality and its psychopathological correlates in the US Army. The literature on risk and resilience factors for suicidality makes it clear that mental disorders are powerful risk factors (Nock *et al.*, 2008; Nock *et al.*, 2013). As a result, a wide range of mental disorders were assessed in the Army STARRS surveys. However, due to the size and logistical complexities of these surveys, which are described earlier in this issue (Heeringa *et al.*, 2013), it was impossible to administer an in-depth psychiatric diagnostic interview to participants. Instead, mental disorders were assessed with short self-administered screening scales.

A number of screening scales exist to assess such disorders as attention-deficit hyperactivity disorder (ADHD; Kessler *et al.*, 2005a), bipolar disorder (BPD; Hirschfeld *et al.*, 2000), generalized anxiety disorder (GAD; Spitzer *et al.*, 2006), major depressive episode (MDE; Kroenke *et al.*, 2001), and post-traumatic stress disorder (PTSD; Breslau *et al.*, 1999). Although in some cases these scales were developed originally to assess symptom severity among patients in treatment, they subsequently have been adapted for use either as web-based tools for self-diagnosis (Donker *et al.*, 2009; Farvolden *et al.*, 2003) or as brief evaluations of mental disorders in primary care settings or community surveys (Broadhead *et al.*, 1995; Gaynes *et al.*, 2010; Hunter *et al.*, 2005; Kessler *et al.*, 2013a). Clinical reappraisal studies comparing scores on these screening scales with independent clinical diagnoses show that many of these screening scales have good concordance with clinical diagnoses (Kessler and Pennell, in press).

The screening scales that form the core diagnostic assessment in Army STARRS are the World Health Organization (WHO) Composite International Diagnostic Interview Screening Scales (CIDI-SC) (Kessler *et al.*, 2013a). These were selected largely because they are a coordinated set of short scales that cover a wide range of disorders and have good psychometric properties. However, another appeal of the CIDI-SC is that they are embedded in the WHO Composite International Diagnostic Interview (CIDI) (Kessler and Üstün, 2004), the research diagnostic interview used in most large-scale epidemiological surveys of psychiatric disorders throughout the world (Haro *et al.*, 2006). Use of the CIDI-SC in Army STARRS thereby creates a crosswalk to an in-depth diagnostic interview that might be used in more focused follow-up studies of Army STARRS high-risk subsamples. The exception is that we used the PTSD checklist (PCL) (Weathers *et al.*, 1993) to assess PTSD based on the widespread use of this screening scale in previous military studies of PTSD (Barnes *et al.*, 2013; Brown *et al.*, 2012; Jones *et al.*, 2013) coupled with strong evidence for the validity of the PCL in both military and civilian samples (Wilkins *et al.*, 2011).

Although good concordance of the Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition (DSM-IV; American Psychiatric Association, 1994) diagnoses based on the CIDI-SC (Kessler *et al.*, 2005a; Kessler *et al.*, 2007; Kessler *et al.*, 2006a; Kessler *et al.*, 2013a) and PCL (Wilkins *et al.*, 2011) with diagnoses based on independent clinical reappraisal interviews has been reported in a number of studies, this does not guarantee that these screening scales will perform equally well among soldiers in the Army STARRS surveys. As a result, a new clinical reappraisal study (CRS) was carried out in conjunction with the Army STARRS All-Army Study (AAS) (Ursano *et al.*, submitted for publication) to examine the psychometric characteristics of the CIDI-SC and PCL in the context of the field conditions encountered in the Army STARRS surveys. Results of this CRS are presented in the current report.

Methods

The samples

The All-Army Study (AAS)

As described in more detail previously in this issue (Kessler *et al.*, 2013b), the AAS is a cross-sectional survey of active duty Army personnel exclusive of soldiers in basic

Int. J. Methods Psychiatr. Res. 22(4): 303–321 (2013). DOI: 10.1002/mpr

combat training administered in quarterly replicates to a total of nearly 50,000 soldiers during calendar years 2011–2012. Each quarterly AAS replicate consisted of a stratified (by Army Command-location and unit size) probability sample of Army units, excluding units of fewer than 30 soldiers (less than 2% of all Army personnel). All targeted personnel in these units were ordered to attend an informed consent presentation explaining study purposes, confidentiality procedures, and the voluntary nature of participation before requesting written informed consent for a group self-administered questionnaire (SAQ). Respondents were additionally asked for consent to link their Army and Department of Defense administrative records to their SAQ responses and to participate in future longitudinal follow-up data collections. Identifying information (name, birthday, Social Security number for record linkage; telephone number, email, secondary contact information for longitudinal follow-up) was collected from consenting respondents and kept in a separate secure file. These recruitment, consent, and data protection procedures were approved by the Human Subjects Committees of the Uniformed Services University of the Health Sciences for the Henry M. Jackson Foundation (the primary grantee), the Institute for Social Research at the University of Michigan (the organization implementing Army STARRS surveys), and all other collaborating organizations.

The CRS was carried out between March 2012 and November 2012. All quarterly AAS replicates over that time period were based on representative samples of soldiers stationed both in the continental United States and elsewhere in the world other than a combat theater, while the Q2–3 2012 replicates also included probability samples of soldiers stationed in Afghanistan who were surveyed in group-administered sessions while they were passing through Kuwait either leaving for or returning from their mid-tour leave. However, because of logistical issues requiring that the CRS interviews be administered within two weeks of the AAS survey, the CRS was implemented exclusively in the continental United States among Regular (active component) Army AAS respondents providing consent for administrative data linkage and completing the SAQ. Activated Army Reserve and National Guard respondents were excluded from the CRS due to small numbers.

Although, as noted earlier, all unit members in these replicates were ordered to report to the informed consent session, 19.4% of those in the replicates used for the CRS were absent due to conflicting duty assignments. The vast majority of those attending (99.6%) consented to the survey and 98.8% of consenters completed the survey. In addition, 71.4% of completers provided successful record linkage. Most incomplete surveys were due to logistical complications (e.g. units either arriving late to survey sessions or having to leave early), although some respondents needed more than the allotted 90 minutes to complete the survey. The survey completion-successful-linkage *cooperation* rate was 63.9% and the completion-successful-linkage *response* rate was 51.5% based on the American Association of Public Opinion Research COOP1 and RR1 calculation methods (American Association for Public Opinion Research, 2009).

The clinical reappraisal study (CRS) sample

In order to evaluate the concordance of diagnoses based on the CIDI-SC and PCL in the AAS with independent clinical diagnoses, a sample of AAS respondents was selected to participate in clinical follow-up interviews within two weeks of completing the AAS in selected AAS sessions. As soon as the AAS survey was completed in these sessions, each AAS respondent was classified as *threshold*, *subthreshold* or *no* on each of the eight screening scales considered here. A probability subsample of AAS respondents from the session was then invited to participate in a confidential clinical reappraisal interview with the goal of obtaining a total (i.e. over the entire nine-month interview recruitment period) of 30 CRS interviews with respondents selected at random from those classified as threshold cases on each diagnosis, 10 from among those classified as subthreshold on each diagnosis, and 40 respondents selected at random from those classified as meeting neither threshold nor subthreshold criteria for any diagnosis. CRS respondents with each diagnosis were selected *with replacement* (i.e. the same respondent could be selected for more than one diagnosis). The initial sampling fractions varied across disorders due to differences in prevalence among the disorders. These sampling fractions were then modified over sessions in order to achieve a roughly equal distribution of cases within each diagnosis across sessions while meeting the sample quotas. The 460 clinical interviews completed by the end of the CRS is more than the 360 needed (i.e. 30 interviews with threshold CIDI-SC/PCL cases for each of eight disorders plus 10 interviews with CIDI-SC/PCL subthreshold cases for each of these disorders plus 40 respondents screening negative on all eight CIDI-SC/PCL scales) because it was necessary to recruit additional respondents in the later replicates to fill the sample quotas for the least common disorders.

Invitations to participate in the CRS were made through unit points of contact who scheduled two-hour time blocks during which respondents were relieved of

their usual duty assignments in order to report to the Army STARRS office on the installation. Once at the study office, an Army STARRS data collection specialist explained the content and purposes of the CRS and obtained written informed consent to participate. Consenting respondents were then assigned to a private room where they were administered the CRS interview telephonically by one of the CRS clinical interviewers, all of whom were located at the Uniformed Services University of the Health Sciences (USUHS) in Bethesda, Maryland. The CRS clinical supervisor (CLD), also located at USUHS, coordinated with Army STARRS data collection specialists at the local AAS installations to schedule these remote CRS telephone interviewers.

An overview of screening scale content

Screening scales were included in the AAS for eight DSM-IV disorders that have been found in previous general population studies to be significant predictors of suicidality (Nock *et al.*, 2008; Nock *et al.*, 2013; Nock *et al.*, 2009). These include two mood disorders (MDE, mania/hypomania [MHM]), three anxiety disorders (panic disorder with or without agoraphobia, GAD, PTSD), and three externalizing disorders (adult ADHD, intermittent explosive disorder, substance use disorder [SUD]).

Symptom questions in most CIDI-SC ask respondents about the frequency of particular symptoms over the 30 days before interview using the response options *all or almost all of the time*, *most of the time*, *some of the time*, *a little of the time*, and *none of the time*. Each CIDI-SC has an embedded skip logic whereby all respondents are administered one or more entry questions and then either skipped if they fail to endorse these questions or continue to a series of follow-up questions if they endorse the entry question(s). This approach was designed to reduce overall scale administration time and respondent burden while minimizing the number of true positives incorrectly skipped out by the entry questions. Respondents who fail to endorse any of the entry questions are asked a total of 46 questions across all eight scales combined, while respondents who endorse every single question are asked an additional 82 questions.

The CIDI-SC MDE scale begins with four entry questions that ask about *being sad*, *depressed*, *or discouraged*, *having little or no interest or pleasure in things*, and *feeling down on yourself*, *no good*, *or worthless* (Kessler *et al.*, 2013a). Respondents who report that at least one of these symptoms occurred at least *some of the time* in the past 30 days are administered 10 additional questions to assess the inclusion criteria of MDE. The *some of the time* threshold, while low for a DSM-IV diagnosis of MDE (which requires depressive symptoms to last most of the day nearly every day for two weeks or longer), was chosen because we wanted to collect information not only on threshold cases but also on subthreshold manifestions of MDE. A similar attempt to collect information about subthreshold symptoms was made in selecting stem question skip rules for each of the other screening scales.

The CIDI-SC MHM scale focuses on subthreshold hypomania as well as mania and hypomania based on evidence that subthreshold hypomania can be highly impairing (Merikangas *et al.*, 2007). In addition, the questions focus on lifetime rather than 30-day prevalence due to the fact that recent BPD can manifest as either MHM or as MDE. As described in more detail elsewhere (Kessler *et al.*, 2006a; Kessler *et al.*, 2013a), the single MHM entry question begins with a vignette describing a hypomanic episode and then asks respondents if they ever had an episode of this sort at any time in their life. A positive response is followed by four questions about the frequency of core MHM symptoms during *a typical intense episode of this sort*. These symptoms include being *much higher*, *happier*, *or optimistic than usual*; *much more irritable than usual*; *so hyper or wound up that you felt out of control*; *having thoughts race through your mind so fast you could hardly keep track of them*. Respondents who report that at least one of these symptoms occurrs at least *some of the time* during a typical intense episode are then administered six additional questions about the inclusion criteria of MHM and are then asked about episode recency to assess 30-day prevalence of MHM. Lifetime rather than 30-day MHM is evaluated here due to the rarity of 30-day MHM in the AAS sample.

The CIDI-SC panic disorder (PD) scale includes two entry questions about lifetime atacks of *panic*, *anxiety*, *or strong fear that came on very suddenly and made you feel very frightened or uneasy*; and *attacks of heart pounding or chest pain that came on very suddenly and made you feel very frightened or uneasy* (Kessler *et al.*, 2013a). A positive response to either entry question is followed by one additional question on how often these attacks are triggered (i.e. occur in situations where the respondent has a strong fear – like a fear of snakes or heights – or where the respondent is in real danger – like a car accident) versus untriggered (i.e. occur without provocation "out of the blue"). Respondents who report ever having untriggered attacks are then administered 13 additional questions to assess the remaining DSM-IV inclusion criteria of PD. Lifetime rather than 30-day PD is evaluated here due to the rarity of 30-day PD in the AAS sample.

The CIDI-SC GAD scale includes five entry questions about 30-day frequency of being *anxious or nervous*;

Int. J. Methods Psychiatr. Res. 22(4): 303–321 (2013). DOI: 10.1002/mpr

worried about a number of different things; *more anxious or worried than other people in your same situation*; *worried about things most other people don't worry about*; and *having trouble controlling your worry or anxiety* (Kessler *et al.*, 2013a). Respondents who report any of these symptoms at least *some of the time* are administered an additional nine questions to assess the remaining DSM-IV inclusion criteria of GAD along with a final question to assess persistence of symptoms. As a minimum duration of six months is required to meet DSM-IV criteria of GAD, the CIDI-SC assesses duration of symptoms, although the concordance data reported here are for symptoms in the 30-days before interview.

As noted earlier, PTSD is assessed in the AAS with the PCL. The PCL Civilian version (Weathers *et al.*, 1993) was used in Army STARRS because we covered traumatic experiences both in and out of the line of duty. This is a 17-question scale that assesses the 17 DSM-IV Criterion B–D symptoms of PTSD. Although there are no entry questions in the PCL, AAS respondents are first asked 15 questions about traumatic experiences (TEs) that might have happened to them during deployments and 15 additional questions about TEs that might have happened to them at any other time in life. Only respondents who report at least one of these 30 TEs are administered the PCL. The PCL questions ask *how much* respondents were *bothered* in the past 30 days by symptoms associated with any of the TEs they ever experienced. Response categories are *extremely*, *quite a bit*, *moderately*, *a little bit*, and *not at all*.

The CIDI-SC adult ADHD scale includes four entry questions found in previous research to provide an optimal short inclusion screen for ADHD in the adult general population (Kessler *et al.*, 2010a). Respondents who report at least two of these symptoms at least *some of the time* in the past six months then receive an additional eight questions shown in a number of previous studies to detect adult ADHD with good accuracy (Kessler *et al.*, 2007; Kessler *et al.*, 2010a; Kessler *et al.*, 2009).

The CIDI-SC intermittent explosive disorder (IED) scale includes one entry question about lifetime attacks of anger when the respondent *all of a sudden … lost control and either broke or smashed something worth more than a few dollars*, *hit or tried to hurt someone*, *or threatened someone* (Kessler *et al.*, 2006b). A positive response is followed by six additional questions that assess the remaining DSM-IV inclusion criteria of IED. As the assessment of IED followed the same logic as the assessment of PD, lifetime rather than 30-day IED is evaluated here in parallel with the evaluation of PD.

The CIDI-SC assessment of SUD, finally, begins with 12 entry questions about quantity-frequency of alcohol use, illicit drug use, and prescription drug misuse, where the latter is defined as use *either without a doctor's prescription*, *more than prescribed*, *or to get high*, *buzzed*, *or numbed out*. Prescription drug misuse is included in the assessment based on evidence that it is considerably more common than illicit drug use in the Army (Bray *et al.*, 2010). Respondents who report any of these types of substance use are then administered the four CIDI-SC questions about DSM-IV substance abuse in the 30 days before interview and eight additional questions to screen for substance dependence in the 30 days before interview including five from the Severity of Dependence Scale (Gossop *et al.*, 1995) and three additional CIDI-SC questions. SUDs (i.e. either abuse or dependence) are assessed only once for alcohol and/or drugs combined.

Scoring the screening scales

Each screening scale was initially scored continuously by summing values across all items in the scale, assigning respondents who were skipped out after screening questions the lowest possible scores on the remaining items. Receiver operating characteristic (ROC) curve analysis (Margolis *et al.*, 2002) was then used to estimate area under the ROC curve (AUC) for the entire continuous scale and to dichotomize the scale at a point that optimized aggregate concordance between the prevalence estimate based on the Structured Clinical Interview for DSM-IV (SCID) and the prevalence estimate based on the CIDI-SC at the designated threshold. This threshold also makes the number of false positives equal the number of false negatives. It is noteworthy, though, that other criteria exist to select diagnostic thresholds and that decisions about which threshold to choose can vary depending on the criterion used. For example, if we had wanted to use the screening scales in a primary care setting to select patients for more in-depth evaluation, we might have lowered the threshold to the point where the vast majority of SCID cases were detected. Or if we were using the screening scales to select patients for a clinical intervention, we might have raised the threshold to the point where the vast majority of screened positives consisted of SCID cases. If the relative importance of minimizing false positives and minimizing false negatives can be specified based on the considerations of such competing criteria, it is possible to minimize this weighted sum of errors in a formal way (Kraemer, 1992). Based on these considerations, a number of alternative thresholds are examined later.

The clinical reappraisal interview

The clinical reappraisal interview was a modified Research Version, Non-Patient Edition of the Structured Clinical Interview for DSM-IV (SCID-I) (First *et al.*, 2002) focused on the eight syndromes under study with the variations in recall periods noted earlier to match the recall periods used in the screening scales. As noted earlier, these interviews were administered by telephone. Telephone administration is now widely accepted in clinical reappraisal studies based on evidence of comparable validity to in-person administration (Kendler *et al.*, 1992; Rohde *et al.*, 1997; Sobin *et al.*, 1993). A great advantage of telephone administration is that a centralized and closely supervised clinical interview staff can carry out the interviews without the geographic restrictions required for face-to-face clinical assessment. A disadvantage is that people without telephones cannot be included in the assessment. As noted later, though, this difficulty was resolved in the Army STARRS CRS by having pre-designated respondents report to the central Army STARRS research office on their installations, where they were placed in a private room and interviewed remotely by telephone.

A major impediment to making accurate evaluations of concordance between screening scales and clinical diagnoses is the fact that respondents are inconsistent in their reports over time. Indeed, our own previous experience and that of other researchers shows consistently that respondents in community surveys tend to report less and less as they are interviewed more and more due to respondent fatigue (Bromet *et al.*, 1986). Part of this pattern is a tendency for respondents to endorse a smaller number of diagnostic stem questions in follow-up interviews than in initial interviews (Kessler *et al.*, 1998), leading to the biased perception that initial fully-structured assessments overestimate prevalence compared to clinical reappraisal interviews. Consistent with the approach used in a number of other clinical reappraisal studies (Haro *et al.*, 2006; Kessler *et al.*, 2005b; Kessler *et al.*, 1998), we modified the conventional blinded clinical re-interview design in three important ways to address this problem.

First, we unblinded the clinical interviewers to whether respondents endorsed diagnostic stem questions in the CIDI-SC. Importantly, though, we did not unblind clinical interviewers to whether the respondents who endorsed CIDI-SC diagnostic stem questions went on to meet full diagnostic criteria.

Second, we rephrased entry questions in the clinical reappraisal interviews to acknowledge prior endorsement of diagnostic stem questions in the CIDI-SC/PCL in order to minimize the problem of false negative diagnostic stem responses in the SCID. For example, rather than repeating a question about presence-absence of 30-day depressed mood in the SCID to respondents who reported 30-day depressed mood in the CIDI-SC, SCID began the assessment of major depression with a declarative sentence: "In your earlier survey you reported feeling sad or depressed most of the time over the past 30 days. The next questions ask more about those feelings."

Third, in order to guarantee that this partial unblinding did not bias clinical interviewers in the direction of rating all stem-positive respondents as cases, we enriched the clinical reappraisal sample to include a higher proportion of respondents than in the sample who endorsed CIDI-SC/PCL diagnostic stem questions but did *not* meet full CIDI-SC/PCL diagnostic criteria. This third feature of the design actually makes the interviewer task more difficult than it would be in a standard CRS in which there is an over-sample of respondents classified as meeting full diagnostic criteria but not of respondents meeting partial criteria.

Clinical interviewer training and quality control

The SCID were administered by 14 trained clinical interviewers. These included four doctoral-level psychologists, seven MA-level psychologists, and three MSW-level clinical social workers. Half of the interviewers had a decade or more of clinical experience (10–21 years), while the other half had 3–9 years of clinical experience (two with three years of experience and one each with five, six, seven, eight, and nine years of experience). The 32-hour SCID interviewer training program began with a 16-hour centralized group training session taking place over a full weekend that was taught by one of the developers of the SCID (MBF) with the assistance of an experienced SCID supervisor (CLD). Training then continued with biweekly individual and group training sessions with homework assignments totaling 32 hours. The training was carried out at USUHS using a modification of the standard SCID training protocol tailored to the diagnoses assessed by the screening scales. In addition to completing this training, each clinical interviewer was required to pass a proficiency test before they began production interviewing based on trainer and supervisor ratings of three practice interviews using a modified version of the SCID Interviewing Skills Evaluation Form created specifically for this study.

All SCID interviews were audio-recorded with permission of respondents and responses recorded on a hard copy interview. The supervisor reviewed the tape recordings of the first five interviews carried out by each interviewer and a minimum of 10% of all subsequent

interviews carried out by each interviewer. The supervisor also reviewed all hard copy interviews completed by all interviewers and reviewed tape recordings of all interviews in which concerns were raised by the hard copy reviews. The symptom-level hard copy clinical ratings were double-entered into a computerized data file after supervisor review and approval. Each interviewer had a weekly one-on-one feedback meeting with the supervisor and participated in a biweekly group calibration meeting with the supervisor and trainer to prevent rater drift. Diagnoses were made without diagnostic hierarchy rules but with organic exclusions.

Analysis methods

Weighting

The CRS sample was weighted to adjust for over-sampling respondents screened as threshold or subthreshold using a weighting method that adjusted for the fact that sampling was made with replacement. This is important because a number of the statistics used to describe scale characteristics are biased when differential selection of screened positives and negatives is not taken into account.

Analysis of screening scale operating characteristics

As noted earlier in the description of screening scale scoring, a summary continuous screening scale score was created for each diagnosis by summing scores across the screening scale items. ROC curve analysis (Margolis *et al.*, 2002) was then used to estimate AUC for the entire scale. Each continuous screening scale was then dichotomized at a threshold that equalized the (weighted) number of false positives and false negatives, thereby maximizing concordance between prevalence estimates based on the SCID and the screening scales. The McNemar x^2 test was used to evaluate the significance of differences between screening scale and SCID prevalence estimates at this threshold. A range of other thresholds was then selected so that SCID prevalence estimates increased monotonically across screening scale strata but did not differ significantly within strata using the logic of stratum-specific likelihood ratio analysis (Pepe, 2003).

Screening scale operating characteristics were then evaluated for each of these thresholds. Individual-level concordance was evaluated using AUC and Cohen's κ (Cohen, 1960). Although κ is the traditional measure used in psychiatric research, κ is not emphasized here because it varies across populations that differ in prevalence even when sensitivity (SN; the percent of true cases correctly classified) and specificity (SP; the percent of true non-cases correctly classified) are constant (Cook, 1998). AUC, in comparison, is a function of SN and SP, which are considered the fundamental parameters of agreement (Kraemer, 1992). AUC equals (SN + SP)/2 when the screen is dichotomous. AUC scores between 0.5 and 1.0 are often interpreted in parallel with κ as *slight* (AUC = 0.50–0.59; κ = 0.0–0.19), *fair* (AUC = 0.6–0.69; κ = 0.2–0.39), *moderate* (AUC = 0.7–0.79; κ = 0.4–0.59), *substantial* (AUC = 0.8–0.89; κ = 0.6–0.79), and *almost perfect* (AUC = 0.9+; κ = 0.8+) (Landis and Koch, 1977). We also report total classification accuracy (TCA), the proportion of all respondents whose CIDI-SC and SCID classifications are consistent.

In addition, we report disaggregated measures of operating characteristics, including SN and SP, positive predictive value (PPV; the proportion of screened positives confirmed by the SCID), negative predictive value (NPV; the proportion of screened negatives confirmed as non-cases by the SCID), likelihood ratio positive (LR+; [SN/(100 − SP)]), and likelihood ratio negative (LR−; [(100 − SN)/SP)]). LR + and LR − assess *relative* proportions of screened positives versus screened negatives confirmed as cases (LR+) or non-cases (LR−). LR + values greater than or equal to five and LR − values less than or equal to 0.2 are generally considered useful, while LR + values greater than or equal to 10 and LR − values less than or equal to 0.1 are considered sufficient to rule in/out diagnoses (Haynes *et al.*, 2006). Significance tests were based on Taylor series design-based standard errors to adjust for data weighting (Wolter, 1985).

Multiple imputation of predicted probabilities of DSM-IV/SCID diagnoses

As noted earlier in the subsection on scoring the screening scales, each screening scale was originally scored continuously and then dichotomized. However, it is not necessary to dichotomize screening scales to make them useful. This is true even in clinical applications, where simple dichotomous scoring rules can be refined by using polychotomous rules that collapse screening scale scores into strata based on analysis of data in a CRS such that the observed prevalence of the clinical outcome differs significantly across strata but not within strata (Guyatt and Rennie, 2001). Designations of patients into multiple risk strata can be useful for clinical purposes when no sharp distinction between cases and non-cases exists in the screening scale (e.g. borderline hypertension).

An extension of this approach can be used in epidemiological surveys to classify respondents into multiple risk

strata based on screening scale scores and to assign predicted probabilities of clinical diagnoses to respondents in each stratum based on the results of a clinical reappraisal survey. It is also possible to ignore the construction of strata in this approach when a monotonic association exists throughout the scale range between a screening scale and probability of a diagnosis, in which case regression analysis can be used to generate predicted probabilities of clinical diagnoses for each respondent in a large sample based on regression coefficients estimated in a smaller clinical reappraisal subsample. These predicted probabilities can then be used either as continuous variables or as the basis for making dichotomous distinctions using any of several different methods discussed elsewhere (Kessler *et al.*, 2010b; Kessler and Pennell, in press).

The creation of continuous scores of this sort is only useful, though, when significant monotonic associations exist between screening scale scores and probabilities of having the clinical diagnosis. We demonstrate later that such associations exist between screening scale scores and diagnoses based on the SCID in the Army STARRS data by comparing AUC for the continuous versions of the screening scales with AUC based on various dichotomous versions of the scales. Given that these monotonic associations exist, we used the method of multiple imputation (MI) (Rubin, 1987) to assign predicted probabilities of SCID diagnoses based on screening scale scores to all respondents in the Army STARRS surveys. MI is a two-phase method designed to impute missing values of particular variables to respondents who have information on variables strongly related to the variable(s) with missing values in such a way as to maximize the use of all available data in examining multivariate associations.

The first phase of MI develops prediction equations based on any of several different complex search methods (Schafer, 2003; White *et al.*, 2011) to estimate multivariate associations of predictors with the variables to be imputed in the subset of respondents with complete data and to use those equations to generate predicted values (*imputations*) for the missing variables in the remainder of the sample. In order to address the fact that imputed values are less precise than observed values, this first phase uses pseudo-replication (i.e. estimation of a new set of coefficients based on the same model from pseudo-samples selected with replacement from the actual sample of people with complete data) to generate multiple imputations for each missing value. The second phase of MI, in which the multiple imputations are used in substantive analysis, then uses each set of imputed values to carry out the substantive analysis separately and then combines the coefficient values across these replications to adjust standard errors of estimates for the fact that some of the data used in the analyses were imputed rather than observed.

Importantly, the first phase of MI allows the inclusion not only of a screening scale (in this case, the CIDI-SC or PCL) designed to provide a proxy measure for the unmeasured variable of interest (in this case, DSM-IV/SCID diagnoses), but also other variables that might be used in second-phase analyses as predictors or consequences of the imputed variable. This is important because the use of only the CIDI-SC or PCL to impute clinical diagnoses would lead to under-estimation of the associations of predictors and consequences of clinical diagnoses with the components of the clinical diagnoses that are not predicted by the CIDI-SC or PCL scores (Collins *et al.*, 2001). As a result, the multiply-imputed predicted probabilities of DSM-IV/SCID diagnoses in Army STARRS were based on complex multivariate equations that included the complete set of CIDI-SC/PCL scores to impute each clinical diagnosis (to adjust for comorbidities among clinical disorders) along with a wide range of substantive correlates included in the AAS and Army/Department of Defense administrative data systems. We produced 20 imputations for each respondent in Army STARRS, a number at the high end of the number recommended in applying MI (Graham *et al.*, 2007).

Results

Concordance of screening scale scores with DSM-IV/SCID diagnoses

Differences in prevalence estimates based on the dichotomized screening scales and SCID are insignificant for all disorders at optimal screening scale thresholds for estimating prevalence ($x_1^2 = 0.0–0.6$, $p = 0.89–0.43$). (Table 1) This is not surprising, of course, as the thresholds were selected to make CIDI-SC prevalence as similar as possible to SCID prevalence. But this is no guarantee of good concordance at the individual level. Individual-level diagnostic concordance at these thresholds is *moderate* for seven diagnoses (AUC = 0.70–0.79) and *fair* for the other diagnosis (ADHD; AUC = 0.69). Total classification accuracy is in the range 86.0–95.9%. The screening scale estimate of 30-day prevalence of any of the seven disorders assessed for 30-day prevalence (the exception being MHM, which was only assessed over the entire lifetime), like most of the individual disorders, has moderate concordance with the estimate based on the SCID (AUC = 0.78).

Int. J. Methods Psychiatr. Res. 22(4): 303–321 (2013). DOI: 10.1002/mpr

Table 1. Aggregate (McNemar x^2) and individual-level (AUC, κ, TCA) consistency of DSM-IV diagnoses based on the CIDI screening scales (CIDI-SC) at their optimal (to estimate prevalence) thresholds and on blinded SCID clinical reappraisal interviews ($n = 460$)[a]

	Aggregate concordance[b]					Individual-level concordance[c]		
	Prevalence estimates							
	CIDI-SC		SCID		McNemar			
	Percent	(SE)	Percent	(SE)	x_1^2	AUC	κ	TCA
I. Mood disorders								
Major depressive episode	6.8	(1.0)	6.7	(1.0)	0.0	0.78	0.55	94.3
Mania/hypomania	4.9	(1.0)	5.2	(0.9)	0.1	0.70	0.42	94.4
II. Anxiety disorders								
Panic disorder	5.0	(0.9)	5.1	(0.7)	0.0	0.78	0.57	95.9
Generalized anxiety disorder	6.6	(0.9)	6.8	(1.0)	0.0	0.70	0.41	92.6
Post-traumatic stress disorder	6.7	(1.0)	6.4	(0.8)	0.1	0.75	0.49	93.7
III. Externalizing disorders								
Adult attention-deficit/hyperactivity disorder	8.2	(1.1)	7.1	(1.1)	0.6	0.69	0.35	90.8
Intermittent explosive disorder	20.8	(2.3)	20.4	(2.0)	0.1	0.79	0.57	86.0
Substance use disorder	4.9	(0.4)	5.4	(0.8)	0.1	0.73	0.47	94.8
IV. Any disorder[d]	18.9	(1.6)	20.3	(1.9)	0.6	0.78	0.58	86.6

[a]Analyses are based on weighted data to adjust for the over-sampling of respondents screening positive on the CIDI-SC scales.
[b]The CIDI-SC prevalence estimates are set at the thresholds designed to maximize concordance with prevalence estimates based on the blinded SCID clinical reappraisal interviews. The McNemar x^2 tests evaluate concordance of these two prevalence estimates.
[c]AUC = area under the receiver operating characteristic curve; κ = Cohen's κ; TCA = total classification accuracy. See the text for definitions of these statistics, all three of which provide information about the overall individual-level concordance between diagnoses based on the CIDI-SC and the blinded SCID clinical reappraisal interviews.
[d]Any of the seven disorders other than mania/hypomania, as mania/hypomania were assessed only over the entire lifetime.

Operating characteristics of the tests

The proportions of SCID cases detected (SN) at the optimal screening scale diagnostic thresholds for estimating SCID prevalence are in the range 42.8–66.8% and the proportions of screening scale cases confirmed by the SCID (PPV) at these thresholds are in the range 37.1–65.3% (68.3% for any 30-day disorder) (Table 2). The proportions of SCID non-cases classified correctly (SP) are 90.9–97.9% and the proportions of screening scale non-cases confirmed as non-cases by the SCID (NPV) are 91.5–97.8%. Lower SN and PPV than SP and NPV are expected for thresholds designed to estimate prevalence without bias when only a minority of respondents has a disorder. LR + is generally considered more informative than SN in such cases (Haynes *et al.*, 2006). LR + is in the *definitive* range (i.e. greater than 10.0) at these thresholds for six of the eight disorders and in the *informative* range (i.e. greater than 5.0) for the others (7.3 for IED; 7.8 for ADHD) and for any 30-day disorder (8.5), indicating that screened positives at these thresholds are much more likely than screened negatives to be confirmed as cases in the clinical reappraisal interviews. LR − values, in comparison, are in a range that would not be considered useful in screening out true non-cases (0.4–0.6).

The implications of modifying diagnostic thresholds

The proportions of screened positives confirmed as SCID cases (PPV) could be increased by raising the screening scale diagnostic thresholds beyond the optimal for estimating prevalence. However, this increase in PPV would be obtained at the expense of decreasing SN and creating downwardly biased (conservative) prevalence estimates. The value of making such a change in threshold while still attempting to approximate clinical prevalence can be

Table 2. CIDI screening scale (CIDI-SC) operating characteristics at optimal thresholds for estimating DSM-IV/SCID prevalence ($n = 460$)[a]

	Positive operating characteristics[b]					Negative operating characteristics[c]				
	SN	(SE)	PPV	(SE)	LR+	SP	(SE)	NPV	(SE)	LR−
I. Mood disorders										
Major depressive episode	58.8	(9.0)	57.5	(6.6)	19.0	96.9	(0.8)	97.0	(0.9)	0.4
Mania/hypomania	43.5	(8.9)	45.8	(6.3)	15.5	97.2	(0.6)	96.9	(0.8)	0.6
II. Anxiety disorders										
Panic disorder	58.5	(11.2)	59.5	(7.7)	27.9	97.9	(0.4)	97.8	(0.6)	0.4
Generalized anxiety disorder	43.9	(4.7)	45.6	(7.4)	11.5	96.2	(0.8)	95.9	(0.7)	0.6
Post-traumatic stress disorder	53.7	(6.9)	50.8	(7.0)	15.3	96.5	(0.8)	96.8	(0.6)	0.5
III. Externalizing disorders										
Adult attention-deficit/hyperactivity disorder	42.8	(7.8)	37.1	(7.0)	7.8	94.5	(1.1)	95.6	(1.0)	0.6
Intermittent explosive disorder	66.8	(6.2)	65.3	(5.2)	7.3	90.9	(1.6)	91.5	(1.6)	0.4
Substance use disorder	47.6	(8.5)	51.6	(8.0)	19.0	97.5	(0.5)	97.0	(0.8)	0.5
IV. Any disorder[d]	63.7	(4.3)	68.3	(4.0)	8.5	92.5	(1.2)	90.9	(1.6)	0.4

[a]Analyses are based on weighted data to adjust for the over-sampling of respondents screening positive on the CIDI-SC scales.
[b]SN = sensitivity (the percent of SCID cases detected by the CIDI-SC); PPV = positive predictive value (the percent of CIDI-SC cases confirmed by the SCID); LR + = likelihood ratio positive (the relative proportions of SCID cases among CIDI-SC cases versus non-cases).
[c]SP = specificity (the percent of SCID non-cases classified as non-cases by the CIDI-SC); NPV = negative predictive value (the percent of CIDI-SC non-cases confirmed as non-cases by the SCID); LR − = likelihood ratio negative (the relative proportions of SCID non-cases among CIDI-SC cases versus non-cases).
[d]Any of the seven disorders other than mania/hypomania, as mania/hypomania were assessed only over the entire lifetime.

evaluated by examining relative changes in PPV versus SN associated with modest increases in screening scale thresholds around the optimal thresholds for estimating SCID prevalence. When we make these small increases in threshold we see that the increases in PPV are much less than the decreases in SN for four disorders (MDE, GAD, ADHD, SUD) (proportional screening scales decreases of 20%, 7%, 25%, and 31%, respectively; proportional PPV increases of 2%, 0%, 18%, and 4%, respectively) (Table 3). In addition, PPV actually *decreases* slightly for the other four disorders due to respondents with CIDI-SC scores just above the optimal thresholds for estimating SCID prevalence of these disorders having high SCID prevalence. These results argue against small changes to increase the screening scale thresholds in the service of making diagnoses more conservative while still maintaining estimates that approximate the SCID prevalence estimates.

We also examined the implications of making small changes in the thresholds in the other direction to increase the proportions of clinical cases screening positive by lowering the screening scale thresholds. Such changes increase SN by definition. This is desirable for purposes of guaranteeing comprehensive detection in treatment samples when PPV does not decrease more than SN increases. However, such anticonservative changes can lead to upward bias in prevalence estimates as well as to reductions in LR + when the proportional increases in SN are lower than the proportional decreases in SP. An analysis of these changes associated with modest decreases in screening scale thresholds shows that LR + consistently decreases when modest changes are made to decrease thresholds (Table 3). These results argue against making the screening scale thresholds less conservative while still maintaining estimates that approximate SCID prevalence.

Selecting alternative optimization rules in selecting screening scale diagnostic thresholds

As noted earlier in the section on analysis methods, the most useful thresholds for screening scales differ depending on the uses to which the screening scales are put. As Army STARRS is an epidemiological study rather than a clinical study, we place a premium on accurate estimation of SCID prevalence. But in a clinical study, where screening scales

Int. J. Methods Psychiatr. Res. 22(4): 303–321 (2013). DOI: 10.1002/mpr

Table 3. Variation in CIDI screening scale (CIDI-SC) operating characteristics when diagnostic thresholds are changed from the optimal for estimating prevalence to either more conservative or more anticonservative thresholds ($n = 460$)[a]

	CIDI-SC prevalence estimate[b]		Positive operating characteristics[c]					Negative operating characteristics[d]				
	Percent	(SE)	SN	(SE)	PPV	(SE)	LR+	SP	(SE)	NPV	(SE)	LR−
Major depressive episode												
Conservative	6.0	(0.8)	49.1	(8.8)	54.9	(8.5)	16.9	97.1	(0.7)	96.4	(1.0)	0.5
Optimal	6.8	(1.0)	58.8	(9.0)	57.5	(6.6)	19.0	96.9	(0.8)	97.0	(0.9)	0.4
Anticonservative	7.5	(1.2)	62.5	(9.2)	55.6	(6.3)	17.4	96.4	(0.9)	97.3	(0.9)	0.4
Mania/hypomania												
Conservative	2.7	(0.5)	20.9	(5.9)	39.5	(8.3)	12.3	98.3	(0.4)	95.8	(1.0)	0.8
Optimal	4.9	(1.0)	43.5	(8.9)	45.8	(6.3)	15.5	97.2	(0.6)	96.9	(0.8)	0.6
Anticonservative	11.6	(1.4)	72.6	(9.1)	32.3	(6.6)	8.7	91.7	(1.4)	98.4	(0.6)	0.3
Panic disorder												
Conservative	3.4	(0.7)	37.1	(9.9)	54.8	(8.9)	23.2	98.4	(0.4)	96.7	(0.7)	0.6
Optimal	5.0	(0.9)	58.5	(11.2)	59.5	(7.7)	27.9	97.9	(0.4)	97.8	(0.6)	0.4
Anticonservative	6.4	(0.9)	71.4	(10.4)	57.1	(6.9)	24.6	97.1	(0.5)	98.4	(0.5)	0.3
Generalized anxiety disorder												
Conservative	6.0	(0.8)	40.8	(4.6)	46.7	(7.7)	10.9	96.6	(0.8)	95.7	(0.7)	0.5
Optimal	6.6	(0.9)	43.9	(4.7)	45.6	(7.4)	11.5	96.2	(0.8)	95.9	(0.7)	0.6
Anticonservative	7.1	(1.0)	44.2	(4.8)	42.6	(7.8)	10.0	95.6	(1.0)	95.9	(0.7)	0.6
Post-traumatic stress disorder												
Conservative	6.2	(1.0)	46.0	(7.0)	47.5	(7.2)	13.1	96.5	(0.9)	96.3	(0.7)	0.6
Optimal	6.7	(1.0)	53.7	(6.9)	50.8	(7.0)	15.3	96.5	(0.8)	96.8	(0.6)	0.5
Anticonservative	7.7	(1.1)	56.8	(7.2)	47.2	(6.3)	13.2	95.7	(0.9)	97.0	(0.6)	0.4
Adult attention-deficit/hyperactivity disorder												
Conservative	6.8	(1.1)	31.8	(6.2)	33.4	(6.8)	6.5	95.1	(1.1)	94.8	(1.0)	0.7
Optimal	8.2	(1.1)	42.8	(7.8)	37.1	(7.0)	7.8	94.5	(1.1)	95.6	(1.0)	0.6
Anticonservative	8.8	(1.1)	44.1	(7.9)	35.5	(6.5)	7.2	93.9	(1.1)	95.7	(1.0)	0.5
Intermittent explosive disorder												
Conservative	16.8	(1.4)	47.3	(4.5)	57.2	(5.5)	5.2	90.9	(1.6)	87.1	(2.0)	0.6
Optimal	20.8	(2.3)	66.8	(6.2)	65.3	(5.2)	7.3	90.9	(1.6)	91.5	(1.6)	0.4
Anticonservative	26.7	(3.3)	73.5	(7.4)	56.0	(5.7)	5.0	85.3	(2.7)	92.6	(1.9)	0.9
Substance use disorder												
Conservative	4.1	(0.5)	38.1	(8.3)	49.5	(8.4)	17.3	97.8	(0.4)	96.5	(0.8)	0.6
Optimal	4.9	(0.4)	47.6	(8.5)	51.6	(8.0)	19.0	97.5	(0.5)	97.0	(0.8)	0.5
Anticonservative	6.7	(0.6)	57.3	(9.1)	45.6	(5.6)	14.7	96.1	(0.5)	97.5	(0.8)	0.4

[a]Analyses are based on weighted data to adjust for the over-sampling of respondents screening positive on the CIDI-SC scales.
[b]The CIDI-SC prevalence estimates are varied by changing the threshold to values both above (conservative) and below (anticonservative) the thresholds designed to maximize concordance with prevalence estimates based on the blinded SCID clinical reappraisal interviews.
[c]SN = sensitivity (the percent of SCID cases detected by the CIDI-SC); PPV = positive predictive value (the percent of CIDI-SC cases confirmed by the SCID); LR + = likelihood ratio positive (the relative proportions of SCID cases among CIDI-SC cases versus non-cases).
[d]SP = specificity (the percent of SCID non-cases classified as non-cases by the CIDI-SC); NPV = negative predictive value (the percent of CIDI-SC non-cases confirmed as non-cases by the SCID); LR − = likelihood ratio negative (the relative proportions of SCID non-cases among CIDI-SC cases versus non-cases).

might be used for case-finding to select people for additional assessment and treatment, it might make more sense to lower the threshold to capture as large a proportion of clinical cases as feasible within the constraints of the cost-benefit ratio of screening and treatment. To investigate the implications of using such a rule in setting

screening scale thresholds, we compared scale operating characteristics when the threshold was selected to detect 80% of DSM-IV/SCID cases (i.e. SN = 80.0%).

This change leads to a lowering of screening scale thresholds for all disorders because SN is consistently lower than 80% at the optimal threshold for estimating SCID prevalence. And this, in turn, leads to substantial increases in screening scale prevalence (2.5–7.0 times the prevalence estimates based on the optimal threshold for estimating SCID prevalence) for all disorders other than PD and IED (where CIDI-SC prevalence estimates increase to 1.2–1.3 times the optimal for estimating SCID prevalence) and to correspondingly large reductions in PPV (Table 4). While PPV at the optimal threshold for estimating SCID prevalence averages 51.6% (i.e. 51.6% of screened positives are true clinical cases, with a range 37.1–65.3%), average PPV drops to 30.0% (range: 11.9–57.1%) when thresholds are selected so that SN exceed 80%. This

Table 4. Variation in CIDI screening scale (CIDI-SC) operating characteristics when diagnostic thresholds are changed from (i) the optimal for estimating prevalence to (ii) having high SN (i.e. detecting at least 80% of DSM-IV/SCID cases) ($n = 460$)[a]

	CIDI-SC prevalence estimate[b]		Positive operating characteristics[c]					Negative operating characteristics[d]				
	Percent	(SE)	SN	(SE)	PPV	(SE)	LR+	SP	(SE)	NPV	(SE)	LR−
Major depressive episode												
Optimal for prevalence	6.8	(1.0)	58.8	(9.0)	57.5	(6.6)	19.0	96.9	(0.8)	97.0	(0.9)	0.4
High SN	25.2	(2.1)	80.2	(11.8)	21.3	(3.1)	3.8	78.7	(2.1)	98.2	(1.2)	0.3
Mania/hypomania												
Optimal for prevalence	4.9	(1.0)	43.5	(8.9)	45.8	(6.3)	15.5	97.2	(0.6)	96.9	(0.8)	0.6
High SN	20.5	(2.2)	82.7	(8.4)	20.8	(4.1)	4.8	82.9	(2.1)	98.9	(0.6)	0.2
Panic disorder												
Optimal for prevalence	5.0	(0.9)	58.5	(11.2)	59.5	(7.7)	27.9	97.9	(0.4)	97.8	(0.6)	0.4
High SN[e]	6.4	(0.9)	71.4	(10.4)	57.1	(6.9)	24.6	97.1	(0.5)	98.4	(0.5)	0.3
Generalized anxiety disorder												
Optimal for prevalence	6.6	(0.9)	43.9	(4.7)	45.6	(7.4)	11.5	96.2	(0.8)	95.9	(0.7)	0.6
High SN	19.1	(2.1)	80.6	(5.2)	28.9	(5.3)	5.5	85.4	(2.3)	98.4	(0.5)	0.2
Post-traumatic stress disorder												
Optimal for prevalence	6.7	(1.0)	53.7	(6.9)	50.8	(7.0)	15.3	96.5	(0.8)	96.8	(0.6)	0.5
High SN	43.5	(4.0)	81.2	(6.6)	11.9	(2.0)	2.0	59.0	(4.2)	97.9	(0.8)	0.3
Adult attention-deficit/hyperactivity disorder												
Optimal for prevalence	8.2	(1.1)	42.8	(7.8)	37.1	(7.0)	7.8	94.5	(1.1)	95.6	(1.0)	0.6
High SN	40.3	(2.9)	84.3	(7.5)	14.9	(2.5)	2.3	63.1	(3.1)	98.1	(1.0)	0.2
Intermittent explosive disorder												
Optimal for prevalence	20.8	(2.3)	66.8	(6.2)	65.3	(5.2)	7.3	90.9	(1.6)	91.5	(1.6)	0.4
High SN[e]	26.7	(3.3)	73.5	(7.4)	56.0	(5.7)	5.0	85.3	(2.7)	92.6	(1.9)	0.3
Substance use disorder												
Optimal for prevalence	4.9	(0.4)	47.6	(8.5)	51.6	(8.0)	19.0	97.5	(0.5)	97.0	(0.8)	0.5
High SN[e]	12.4	(1.6)	66.8	(9.,4)	28.8	(5.4)	7.1	90.6	(1.7)	98.0	(0.8)	0.4

[a]Analyses are based on weighted data to adjust for the over-sampling of respondents screening positive on the CIDI-SC scales.
[b]The CIDI-SC prevalence estimates are varied by changing the threshold to have a minimum SN of 80.0% based on the blinded SCID clinical reappraisal interviews.
[c]SN = sensitivity (the percent of SCID cases detected by the CIDI-SC); PPV = positive predictive value (the percent of CIDI-SC cases confirmed by the SCID); LR+ = likelihood ratio positive (the relative proportions of SCID cases among CIDI-SC cases versus non-cases).
[d]SP = specificity (the percent of SCID non-cases classified as non-cases by the CIDI-SC); NPV = negative predictive value (the percent of CIDI-SC non-cases confirmed as non-cases by the SCID); LR − = likelihood ratio negative (the relative proportions of SCID non-cases among CIDI-SC cases versus non-cases).
[e]As none of the CIDI-SC thresholds for this disorder had SN as high as 80%, the threshold with the highest SN is reported.

Int. J. Methods Psychiatr. Res. 22(4): 303–321 (2013). DOI: 10.1002/mpr

means that it would require an average of about three SCID interviews to detect each clinical case among the screened positives at the lower threshold compared to roughly two at the higher threshold. Clinical intervention cost-effectiveness calculations would be needed to determine whether this additional expense of case-finding could be justified based on the human costs (i.e. quality of life, morbidity, mortality) of an untreated case, the costs of treatment, and the likely effectiveness of treatment in reducing human costs. From the perspective of epidemiological research, lowering the thresholds below the optimal for estimating prevalence might still be desirable even though such an anticonservative change introduces upward bias in prevalence estimates, as it is possible that lowering thresholds will lead to greater proportional increases in SN than in (100 − SP), in which case LR + will increase. However, LR + decreases consistently when the screening scale thresholds are lowered, arguing against making these thresholds less conservative for purposes of epidemiological analysis of the Army STARRS data.

Another goal of screening might be to select screening scale thresholds to have a minimum proportion of screened positives confirmed in clinical interviews (i.e. high PPV). For example, minimum PPV might be set at 50% to guarantee that the majority of screened positives are true clinical cases or at 80% to guarantee that the vast majority of screened positives are true clinical cases. However, this will lead to a reduction in SN that might make the true cases detected unrepresentative of all true cases. If minimum PPV is set at 50%, the thresholds selected to maximize estimation of SCID prevalence meet the PPV criterion in five of eight cases, the exceptions being MHM (PPV = 45.8%), GAD (PPV = 45.6%), and ADHD (PPV = 37.1%). In the case of MHM, the threshold can be raised to increase PPV to 71.1%, but this leads to a dramatic reduction in estimated prevalence (from 4.9% to 0.7%) and in SN (from 43.5% to 9.7%) (Table 5). While more than two-thirds of the small fraction of respondents defined as positive for MHM in the CIDI-SC are SCID cases, the exclusion of the vast majority of SCID cases of MHM from this small fraction (100 − SN = 90.3% of SCID cases not detected) means that the proportion of SCID cases among the screened negatives is nearly as high as the proportion among screened negatives (LR − = 0.9), arguing against making the screening scale thresholds this conservative for purposes of epidemiological analysis of the Army STARRS data.

In the case of GAD, raising the CIDI-SC threshold to make PPV exceed 50% leads to halving both estimated prevalence (from 6.6% to 3.2%) and SN (from 43.9% to 23.6%) in the service of only a relatively modest increase in PPV (from 45.6% to 50.2%) compared to when the threshold is set to maximize estimation of SCID prevalence. It is difficult to argue for a threshold that decreases SN so dramatically for such a modest increase in PPV. The situation is similar but less dramatic for ADHD, where a change in the CIDI-SC threshold that increased PPV by roughly 50% (from 37.1% to 55.9%) decreased estimated prevalence by 70% (from 8.2% to 2.4%) and SN by 55% (from 42.8% to 19.3%). Selecting thresholds to have even higher PPV (a minimum of 80%) for disorders where screening scale PPV is greater than 50% at the optimal threshold for estimating SCID prevalence consistently has the same negative effects in that the proportional increases in PPV (in the range 47–59%) are much less than the proportional decreases in prevalence (84–91%), resulting in extremely low levels of SN (7.3–13.4%). These results argue against using such restrictive thresholds for purposes of epidemiological analysis of the Army STARRS data.

Continuous versus dichotomous diagnostic classification

As noted earlier in the section on analysis methods, we calculated ROC curves for the entire screening scale distributions (Figure 1). AUC was calculated for each of these curves and compared to the AUC of the dichotomous version of the same screening scale. AUC was found to be substantially higher for the continuous than dichotomous scoring rule for each of the eight screening scales (Range: 0.80–0.90 continuous versus 0.69–0.79 dichotomous; inter-quartile range: 0.85–0.87 continuous versus 0.70–0.78 dichotomous) (Table 6). This suggests that meaningful variation in SCID prevalence exists at other places on the screening scale ranges than the optimal diagnostic threshold for estimating SCID prevalence. The important implication of this finding for our purposes is that continuous screening scale scores defining predicted probabilities of clinical diagnoses might be more useful than dichotomous diagnostic classifications based on the screening scales for purposes of epidemiological analysis. We consequently calculated both continuous (predicted probability of having a DSM-IV/SCID diagnosis) and dichotomous versions of each screening scale for use in analysis of the Army STARRS data. The continuous versions were produced using the MI method. Importantly, not only the screening scale scores but also a wide range of other significant correlates of the DSM-IV/SCID diagnoses were used in the first-phase MI analysis in order to minimize bias in subsequent substantive analyses that

Table 5. Variation in CIDI screening scale (CIDI-SC) operating characteristics when diagnostic thresholds are changed from (i) the optimal for estimating prevalence to (ii) having high PPV (i.e. at least 80% of screened positives having a DSM-IV/SCID diagnosis) ($n = 460$)[a]

	CIDI-SC prevalence estimate[b]		Positive operating characteristics[c]					Negative operating characteristics[d]				
	Percent	(SE)	SN	(SE)	PPV	(SE)	LR+	SP	(SE)	NPV	(SE)	LR–
Major depressive episode												
Optimal for prevalence	6.8	(1.0)	58.8	(9.0)	57.5	(6.6)	19.0	96.9	(0.8)	97.0	(0.9)	0.4
High PPV	0.6	(0.3)	7.3	(4.1)	84.7	(11.5)	73.0	99.9	(0.1)	93.8	(1.0)	0.9
Mania/hypomania												
Optimal for prevalence	4.9	(1.0)	43.5	(8.9)	45.8	(6.3)	15.5	97.2	(0.6)	96.9	(0.8)	0.6
High PPV[e]	0.7	(0.2)	9.7	(3.0)	71.1	(12.5)	48.5	99.8	(0.1)	95.3	(1.0)	0.9
Generalized anxiety disorder												
Optimal for prevalence	6.6	(0.9)	43.9	(4.7)	45.6	(7.4)	11.5	96.2	(0.8)	95.9	(0.7)	0.6
PPV GT 50%	3.2	(0.6)	23.6	(3.9)	50.2	(8.5)	13.9	98.3	(0.5)	94.6	(0.9)	0.8
High PPV[e]	1.0	(0.3)	10.4	(3.5)	69.1	(12.2)	34.7	99.7	(0.2)	93.8	(0.9)	0.9
Post-traumatic stress disorder												
Optimal for prevalence	6.7	(1.0)	53.7	(6.9)	50.8	(7.0)	15.3	96.5	(0.8)	96.8	(0.6)	0.5
High PPV[e]	1.1	(0.3)	13.4	(3.3)	79.5	(10.5)	67.0	79.5	(10.5)	99.8	(0.1)	0.9
Adult attention-deficit/hyperactivity disorder												
Optimal for prevalence	8.2	(1.1)	42.8	(7.8)	37.1	(7.0)	7.8	94.5	(1.1)	95.6	(1.0)	0.6
PPV GT 50%	2.4	(0.5)	19.3	(3.8)	55.9	(12.0)	16.1	98.8	(0.4)	94.1	(1.0)	0.8
High PPV[e]	2.0	(0.4)	17.8	(3.7)	63.1	(12.7)	22.3	99.2	(0.3)	94.1	(0.9)	0.8
Substance use disorder												
Optimal for prevalence	4.9	(0.4)	47.6	(8.5)	51.6	(8.0)	19.0	97.5	(0.5)	97.0	(0.8)	0.5
High PPV	0.6	(0.2)	8.8	(3.6)	81.7	(13.3)	88.0	99.9	(0.1)	95.1	(0.8)	0.9

[a]Analyses are based on weighted data to adjust for the over-sampling of respondents screening positive on the CIDI-SC scales.
[b]The CIDI-SC prevalence estimates are varied by changing the threshold to have a minimum PPV of 80.0% based on the blinded SCID clinical reappraisal interviews. Results are not reported for PD or IED because optimal thresholds for predicting SCID prevalence of these disorders also had the highest values of PPV.
[c]SN = sensitivity (the percent of SCID cases detected by the CIDI-SC); PPV = positive predictive value (the percent of CIDI-SC cases confirmed by the SCID); LR + = likelihood ratio positive (the relative proportions of SCID cases among CIDI-SC cases versus non-cases).
[d]SP = specificity (the percent of SCID non-cases classified as non-cases by the CIDI-SC); NPV = negative predictive value (the percent of CIDI-SC non-cases confirmed as non-cases by the SCID); LR − = likelihood ratio negative (the relative proportions of SCID non-cases among CIDI-SC cases versus non-cases).
[e]As none of the CIDI-SC thresholds for this disorder had PPV as high as 80%, the threshold with the highest PPV is reported.

will use these variables as correlates of predicted probabilities of DSM-IV/SCID disorders.

Discussion

Previous research has shown that CIDI-SC operating characteristics are equivalent to or better than those of alternative screening scales in samples of the general population (Kessler *et al.*, 2005a; Kessler *et al.*, 2006a; Kessler *et al.*, 2013a) and that the PCL has very good concordance with clinical diagnoses of PTSD in samples of both the military and the general population (Wilkins *et al.*, 2011). We nonetheless carried out an independent CRS of these screening scales in Army STARRS due to the fact that the operating characteristics of the same screening scale can differ substantially across surveys depending on such fundamental survey conditions as auspices, level of confidentiality (e.g. complete anonymity versus de-identification), mode of data collection, and situational factors, such as constraint on the amount of time available to complete the survey (Kessler and Pennell, in press).

It is not surprising in light of the challenging survey conditions in Army STARRS – including group-administration in settings with suboptimal physical facilities (e.g. sitting on

Int. J. Methods Psychiatr. Res. 22(4): 303–321 (2013). DOI: 10.1002/mpr

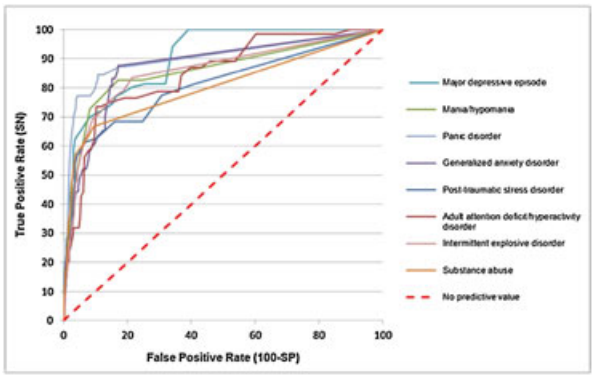

Figure 1. ROC curves for the associations between continuous screening scales and DSM-IV/SCID diagnoses ($n = 460$) ROC = receiver operating characteristic; SN = sensitivity; SP = specificity; DSM-IV = Diagnostic and Statistical Manuel of Mental Disorders, Fourth Edition; SCID = Structured Clinical Interview for DSM-IV.

Table 6. Comparison of area under the receiver operating characteristic curve (AUC) based on the dichotomous versions of the CIDI-SC scales as the optimal thresholds for estimating DSM-IV/SCID prevalence and based on the continuous versions of the CIDI-SC scales ($n = 460$)[a]

	Area under the curve (AUC)	
	Dichotomous	Continuous
I. Mood disorders		
Major depressive episode	0.78	0.90
Mania/hypomania	0.70	0.86
II. Anxiety disorders		
Panic disorder	0.78	0.90
Generalized anxiety disorder	0.70	0.87
Post-traumatic stress disorder	0.75	0.81
III. Externalizing disorders		
Adult attention deficit/ hyperactivity disorder	0.69	0.85
Intermittent explosive disorder	0.79	0.86
Substance use disorder	0.73	0.80

[a]Analyses are based on weighted data to adjust for the over-sampling of respondents screening positive on the CIDI-SC scales.

folding chairs in full field gear in temporary data collection locations) – that we found that the CIDI-SC and PCL AUCs are somewhat lower than in previous psychometric studies of these scales. Individual-level concordance of diagnoses based on the CIDI-SC and PCL with diagnoses based on independent SCID clinical reappraisal interviews in the AAS is for most part *moderate* (AUC = 0.70–0.79; $\kappa = 0.4$–0.6), whereas most previous evaluations found concordance of the CIDI-SC and the PCL with SCID diagnoses to be *substantial* (AUC = 0.80–0.89; $\kappa = 0.6$–0.8). However, the administrative conditions of the screening scales in most previous studies that carried out clinical reappraisals were much better than in Army STARRS, including self-administration in primary care waiting rooms (Kessler *et al.*, 2013a), face-to-face interviewer administration in household surveys (Kessler *et al.*, 2006a), and interviewer administration over the telephone with health plan subscribers (Kessler *et al.*, 2005a).

Perhaps the more striking result in light of the challenging Army STARRS field conditions is that the positive CIDI-SC/PCL operating characteristics for dichotomous versions of the scales designed to optimize aggregate concordance with SCID prevalence estimates are generally quite good. LR + values for six of the eight disorders are in the range 11.5–27.9, all of which are well above the 10.0 value generally considered sufficient to rule in diagnoses (Haynes *et al.*, 2006), while the 7.3–7.8 LR + values for the other two diagnoses and the 8.5 LR + value for any 30-day disorder are well above the 5.0 value considered useful in ruling in diagnoses. However, these good LR + values are accompanied by LR − values generally considered not to be useful in screening out true negatives (0.4–0.6); that is, to contain proportions of true negative that are not strikingly different from the proportions found among screened positives.

As discussed in more detail elsewhere (Kessler *et al.*, 2013a), the definitions of screened positives and screened negatives could be purified for clinical purposes by selecting thresholds at the tails of the distributions that have operating characteristics deemed useful for clinical purposes. For example, an upper threshold of a screening scale could be selected to have a minimum PPV of 0.5 in order to make sure that at least 50% of screened positives are SCID cases. As we saw, though, this desirable feature of that threshold would generally mean that a substantial proportion of SCID cases are missed. Alternatively, the upper threshold of a screening scale could be set at a minimum SN of 0.80 to make sure that the vast majority of SCID cases are picked up by the screen, but this desirable feature of that threshold would mean that only a small proportion of screened positives have SCID diagnoses. In a similar way, a lower threshold of a screening could be purified by requiring NPV to be, say, at least $1 - p/5$, where p = SCID prevalence of the disorder, thereby guaranteeing that the proportion of SCID cases among patients screening negative is no more than 20% as high

as the prevalence of the disorder in the sample, but this desirable feature of that threshold might mean that a substantial proportion of true non-cases are excluded from this ruled-out group.

It is also possible to select multiple thresholds at upper and lower tails both to maximize the positives (i.e. definitive screen-ins and/or screen-outs) and minimize the negatives (i.e. minimizing the numbers of false positives and/or false negatives) and leave one or more intermediate strata that define those with high-but-not-definitively-high scores, low-but-not-definitively-low scores, and uninformative intermediate scores. We noted earlier that such polychotomous scoring rules are fairly common in screening scales developed for clnical practice (Guyatt and Rennie, 2001). Indeed, CIDI-SC polychotomous thresholds have been developed for exactly this reason to facilitate the use of these scales in primary care screening (Kessler *et al.*, 2013a).

However, a more useful approach for purposes of epidemiological analysis of the screening scales considered here is likely to be retention of the entire screening scale range given that AUCs of continuous versions of the screening scales are higher than AUCs of dichotomized versions of the scales at their unbiased thresholds. Based on this observation, we are using MI to assign predicted probabilities of DSM-IV/SCID diagnoses to all Army STARRS respondents who completed the screening scales. We are addressing the uncertainty of inference from prediction equations using imputed rather than observed values by estimating 20 MI estimates of the predicted probability of having each clinical diagnosis for each respondent. The practical use of this approach is illustrated in a more detailed methodological exposition published previously in this journal (Kessler and Üstün, 2004) as well as in a number of subsequent substantive reports that used this approach to estimate the prevalence and correlates of several different DSM-IV/SCID disorders in other psychiatric epidemiological studies (Fayyad *et al.*, 2007; Huang *et al.*, 2009; Kessler *et al.*, 2005a). However, second-phase of MI analysis can be computationally intensive even after the first-phase multiple imputations, as each model has to be estimated 20 separate times rather than once and the coefficients in these 20 replicates then need to be combined to calculate adjusted standard errors. As a result, we also plan to work with dichotomously-scored screening scale measures at the optimal diagnostic thresholds and to investigate the extent to which substantive results differ depending on whether this dichotomous approach is used instead of MI. Dichotomous screening scale scoring will be used in cases where results are relatively insensitive to the more refined estimates using MI.

Acknowledgements

On behalf of the Army STARRS Collaborators

Funding/Support: Army STARRS was sponsored by the Department of the Army and funded under cooperative agreement number U01MH087981 with the US Department of Health and Human Services, National Institutes of Health, National Institute of Mental Health (NIH/NIMH). The contents are solely the responsibility of the authors and do not necessarily represent the views of the Department of Health and Human Services, NIMH, the Department of the Army, or the Department of Defense.

Role of the Sponsors: As a cooperative agreement, scientists employed by NIMH (Colpe and Schoenbaum) and Army liaisons/consultants (COL Steven Cersovsky, MD, MPH USAPHC and Kenneth Cox, MD, MPH USAPHC) collaborated to develop the study protocol and data collection instruments, supervise data collection, plan and supervise data analyses, interpret results, and prepare reports. Although a draft of this manuscript was submitted to the Army and NIMH for review and comment prior to submission, this was with the understanding that comments would be no more than advisory.

Additional Contributions: The Army STARRS Team consists of Co-Principal Investigators: Robert J. Ursano, MD (Uniformed Services University of the Health Sciences) and Murray B. Stein, MD, MPH (University of California San Diego and VA San Diego Healthcare System); Site Principal Investigators: Steven Heeringa, PhD (University of Michigan) and Ronald C. Kessler, PhD (Harvard Medical School); NIMH collaborating scientists: Lisa J. Colpe, PhD, MPH and Michael Schoenbaum, PhD; Army liaisons/consultants: COL Steven Cersovsky, MD, MPH (USAPHC) and Kenneth Cox, MD, MPH (USAPHC). Other team members: Pablo A. Aliaga, MA (Uniformed Services University of the Health Sciences); COL David M. Benedek, MD (Uniformed Services University of the Health Sciences); Susan Borja, PhD (National Institute of Mental Health); Gregory G. Brown, PhD (University of California San Diego); Laura Campbell-Sills, PhD (University of California San Diego); Catherine L. Dempsey, PhD, MPH (Uniformed Services University of the Health Sciences); Richard Frank, PhD (Harvard Medical School); Carol S. Fullerton, PhD (Uniformed Services University of the Health Sciences); Nancy Gebler, MA (University of Michigan); Joel Gelernter, MD (Yale University); Robert K. Gifford, PhD (Uniformed Services University of the Health Sciences); Stephen E. Gilman, ScD (Harvard School of Public Health); Marjan G. Holloway, PhD (Uniformed Services University of the Health Sciences); Paul E. Hurwitz, MPH (Uniformed Services University of the Health Sciences); Sonia Jain, PhD (University of California San Diego); Tzu-Cheg Kao, PhD (Uniformed Services University of the Health Sciences);

Int. J. Methods Psychiatr. Res. 22(4): 303–321 (2013). DOI: 10.1002/mpr

Karestan C. Koenen, PhD (Columbia University); Lisa Lewandowski-Romps, PhD (University of Michigan); Holly Herberman Mash, PhD (Uniformed Services University of the Health Sciences); James E. McCarroll, PhD, MPH (Uniformed Services University of the Health Sciences); Katie A. McLaughlin, PhD (Harvard Medical School); James A. Naifeh, PhD (Uniformed Services University of the Health Sciences); Matthew K. Nock, PhD (Harvard University); Rema Raman, PhD (University of California San Diego); Nancy A. Sampson, BA (Harvard Medical School); LCDR Patcho N. Santiago, MD, MPH (Uniformed Services University of the Health Sciences); Michaelle Scanlon, MBA (National Institute of Mental Health); Jordan Smoller, MD, ScD (Harvard Medical School); Nadia Solovieff, PhD (Harvard Medical School); Michael L. Thomas, PhD (University of California San Diego); Christina Wassel, PhD (University of Pittsburgh); and Alan M. Zaslavsky, PhD (Harvard Medical School).

Declaration of interest statement

In the past five years Kessler has been a consultant for Eli Lilly & Company, Glaxo, Inc., Integrated Benefits Institute, Ortho-McNeil Janssen Scientific Affairs, Pfizer Inc., Sanofi-Aventis Groupe, Shire US Inc., and Transcept Pharmaceuticals Inc. and has served on advisory boards for Johnson & Johnson. Kessler has had research support for his epidemiological studies over this time period from Eli Lilly & Company, EPI-Q, GlaxoSmithKline, Ortho-McNeil Janssen Scientific Affairs, Sanofi-Aventis Groupe, Shire US, Inc., and Walgreens Co. Kessler owns a 25% share in DataStat, Inc. First received consultation fees from the Henry M. Jackson Foundation for the Advancement of Military Medicine, the sponsor of the study. Stein has in the last three years been a consultant for Healthcare Management Technologies and had research support for pharmacological imaging studies from Janssen. The remaining authors report nothing to disclose.

References

American Association for Public Opinion Research. (2009) *Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys*, Deerfield, IL, American Association for Public Opinion Research.

American Psychiatric Association. (1994) *Diagnostic and Statistical Manual of Mental Disorders (DSM-IV)*, Fourth Edition, Washington, DC, American Psychiatric Association.

Barnes J.B., Nickerson A., Adler A.B., Litz B.T. (2013) Perceived military organizational support and peacekeeper distress: A longitudinal investigation. *Psychological Services*, **10**(2), 177–185, DOI: 10.1037/a0032607

Bray R.M., Pemberton M.R., Lane M.E., Hourani L.L., Mattiko M.J., Babeu L.A. (2010) Substance use and mental health trends among U.S. military active duty personnel: key findings from the 2008 DoD Health Behavior Survey. *Military Medicine*, **175**(6), 390–399.

Breslau N., Peterson E.L., Kessler R.C., Schultz L.R. (1999) Short screening scale for DSM-IV posttraumatic stress disorder. *American Journal of Psychiatry*, **156**(6), 908–911.

Broadhead W.E., Leon A.C., Weissman M.M., Barrett J.E., Blacklow R.S., Gilbert T.T., Keller M.B., Olfson M., Higgins E.S. (1995) Development and validation of the SDDS-PC screen for multiple mental disorders in primary care. *Archives of Family Medicine*, **4**(3), 211–219.

Bromet E.J., Dunn L.O., Connell M.M., Dew M.A., Schulberg H.C. (1986) Long-term reliability of diagnosing lifetime major depression in a community sample. *Archives of General Psychiatry*, **43**(5), 435–440.

Brown J.M., Williams J., Bray R.M., Hourani L. (2012) Postdeployment alcohol use, aggression, and post-traumatic stress disorder. *Military Medicine*, **177**(10), 1184–1190.

Cohen J. (1960) A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, **20**(1), 37–46, DOI: 10.1177/001316446002000104

Collins L.M., Schafer J.L., Kam C.M. (2001) A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods*, **6**(4), 330–351.

Cook R.J. (1998) Kappa and its dependence on marginal rates. In Armitage P., Colton T. (eds) *The Encyclopedia of Biostatistics*, p. 2166–2168, New York, John Wiley & Sons.

Donker T., van Straten A., Marks I., Cuijpers P. (2009) A brief Web-based screening questionnaire for common mental disorders: development and validation. *Journal of Medical Internet Research*, **11**(3), e19, DOI: 10.2196/jmir.1134

Farvolden P., McBride C., Bagby R.M., Ravitz P. (2003) A Web-based screening instrument for depression and anxiety disorders in primary care. *Journal of Medical Internet Research*, **5**(3), e23, DOI: 10.2196/jmir.5.3.e23

Fayyad J., De Graaf R., Kessler R., Alonso J., Angermeyer M., Demyttenaere K., De Girolamo G., Haro J.M., Karam E.G., Lara C., Lepine J.P., Ormel J., Posada-Villa J., Zaslavsky A.M., Jin R. (2007) Cross-national prevalence and correlates of adult attention-deficit hyperactivity disorder. *British Journal of Psychiatry*, **190**, 402–409, DOI: 10.1192/bjp.bp.106.034389

First M.B., Spitzer R.L., Gibbon M., Williams J.B.W. (2002) *Structured Clinical Interview for DSM-IV Axis I Disorders, Research Version, Non-patient Edition (SCID-I/NP)*, New York, Biometrics Research, New York State Psychiatric Institute.

Gaynes B.N., DeVeaugh-Geiss J., Weir S., Gu H., MacPherson C., Schulberg H.C., Culpepper L., Rubinow D.R. (2010) Feasibility and diagnostic validity of the M-3 checklist: a brief, self-rated screen for depressive, bipolar, anxiety, and post-traumatic stress disorders in primary care. *Annals of Family Medicine*, **8**(2), 160–169, DOI: 10.1370/afm.1092

Gossop M., Darke S., Griffiths P., Hando J., Powis B., Hall W., Strang J. (1995) The Severity of Dependence Scale (SDS): psychometric properties of the SDS in English and Australian samples of heroin, cocaine and amphetamine users. *Addiction*, **90**(5), 607–614.

Graham J.W., Olchowski A.E., Gilreath T.D. (2007) How many imputations are really

needed? Some practical clarifications of multiple imputation theory. *Prevention Science*, **8**(3), 206–213, DOI: 10.1007/s11121-007-0070-9

Guyatt G., Rennie D. (2001) *User's Guide to the Medical Literature: A Manual for Evidence-based Clinical Practice*, Chicago, IL, AMA Press.

Haro J.M., Arbabzadeh-Bouchez S., Brugha T.S., de Girolamo G., Guyer M.E., Jin R., Lepine J.P., Mazzi F., Reneses B., Vilagut G., Sampson N. A., Kessler R.C. (2006) Concordance of the Composite International Diagnostic Interview Version 3.0 (CIDI 3.0) with standardized clinical assessments in the WHO World Mental Health surveys. *International Journal of Methods in Psychiatric Research*, **15**(4), 167–180, DOI: 10.1002/mpr.196

Haynes R.B., Sackett D.L., Guyatt G.H., Tugwell P. (2006) *Clinical Epidemiology: How to Do Clinical Practice Research*, Third Edition, Philadelphia, PA, Lippincott Williams & Wilkins.

Heeringa S.G., Colpe L.J., Fullerton C.S., Gebler N., Naifeh J.A., Nock M.K., Sampson N.A., Schoenbaum M., Zaslavsky A.M., Stein M.B., Ursano R.J., Kessler R.C. (2013) Field procedures in the Army Study to Assess Risk and Resilience in Servicemembers (Army STARRS). *International Journal of Methods and Psychiatric Research*, **22**(4), 276–287.

Hirschfeld R.M., Williams J.B., Spitzer R.L., Calabrese J.R., Flynn L., Keck P.E., Jr., Lewis L., McElroy S.L., Post R.M., Rapport D.J., Russell J.M., Sachs G.S., Zajecka J. (2000) Development and validation of a screening instrument for bipolar spectrum disorder: the Mood Disorder Questionnaire. *American Journal of Psychiatry*, **157**(11), 1873–1875, DOI: 10.1176/appi.ajp.157.11.1873

Huang Y., Kotov R., de Girolamo G., Preti A., Angermeyer M., Benjet C., Demyttenaere K., de Graaf R., Gureje O., Karam A.N., Lee S., Lepine J.P., Matschinger H., Posada-Villa J., Suliman S., Vilagut G., Kessler R.C. (2009) DSM-IV personality disorders in the WHO World Mental Health Surveys. *British Journal of Psychiatry*, **195**(1), 46–53, DOI: 10.1192/bjp.bp.108.058552

Hunter E.E., Penick E.C., Powell B.J., Othmer E., Nickel E.J., Desouza C. (2005) Development of scales to screen for eight common psychiatric disorders. *Journal of Nervous and Mental Disease*, **193**(2), 131–135, DOI: 10.1097/01.nmd.0000152786.61048.a1

Jones M., Sundin J., Goodwin L., Hull L., Fear N.T., Wessely S., Rona R.J. (2013) What explains post-traumatic stress disorder (PTSD) in UK service personnel: deployment or something else? *Psychological Medicine*, **43**(8), 1703–1712, DOI: 10.1017/S0033291712002619

Kendler K.S., Neale M.C., Kessler R.C., Heath A.C., Eaves L.J. (1992) A population-based twin study of major depression in women. *The impact of varying definitions of illness. Archives of General Psychiatry*, **49**(4), 257–266.

Kessler R.C., Adler L., Ames M., Demler O., Faraone S., Hiripi E., Howes M.J., Jin R., Secnik K., Spencer T., Ustun T.B., Walters E. E. (2005a) The World Health Organization Adult ADHD Self-Report Scale (ASRS): a short screening scale for use in the general population. *Psychological Medicine*, **35**(2), 245–256, DOI: 10.1017/S0033291704002892

Kessler R.C., Adler L.A., Gruber M.J., Sarawate C. A., Spencer T., Van Brunt D.L. (2007) Validity of the World Health Organization Adult ADHD Self-Report Scale (ASRS) Screener in a representative sample of health plan members. *International Journal of Methods and Psychiatric Research*, **16**(2), 52–65, DOI: 10.1002/mpr.208

Kessler R.C., Akiskal H.S., Angst J., Guyer M., Hirschfeld R.M., Merikangas K.R., Stang P.E. (2006a) Validity of the assessment of bipolar spectrum disorders in the WHO CIDI 3.0. *Journal of Affective Disorders*, **96**(3), 259–269, DOI: 10.1016/j.jad.2006.08.018

Kessler R.C., Berglund P., Demler O., Jin R., Merikangas K.R., Walters E.E. (2005b) Lifetime prevalence and age-of-onset distributions of DSM-IV disorders in the National Comorbidity Survey Replication. *Archives of General Psychiatry*, **62**(6), 593–602, DOI: 10.1001/archpsyc.62.6.593

Kessler R.C., Calabrese J.R., Farley P.A., Gruber M. J., Jewell M.A., Katon W., Keck P.E., Nierenberg A.A., Sampson N.A., Shear M.K., Shillington A.C., Stein M.B., Thase M.E., Wittchen H.U. (2013a) Composite International Diagnostic Interview screening scales for DSM-IV anxiety and mood disorders. *Psychological Medicine*, **43**(8), 1625–1637, DOI: 10.1017/S0033291712002334

Kessler R.C., Coccaro E.F., Fava M., Jaeger S., Jin R., Walters E. (2006b) The prevalence and correlates of DSM-IV intermittent explosive disorder in the National Comorbidity Survey Replication. *Archives of General Psychiatry*, **63**(6), 669–678, DOI: 10.1001/archpsyc.63.6.669

Kessler R.C., Colpe L.J., Fullerton C.S., Gebler N., Naifeh J.A., Nock M.K., Sampson N.A., Schoenbaum M., Zaslavsky A.M., Stein M.B., Ursano R.J., Heeringa S.G. (2013b) Design of the Army Study to Assess Risk and Resilience in Servicemembers (Army STARRS). *International Journal of Methods and Psychiatric Research*, **22**(4), 267–275.

Kessler R.C., Green J.G., Adler L.A., Barkley R.A., Chatterji S., Faraone S.V., Finkelman M., Greenhill L.L., Gruber M.J., Jewell M., Russo L.J., Sampson N.A., Van Brunt D.L. (2010a) Structure and diagnosis of adult attention-deficit/hyperactivity disorder: analysis of expanded symptom criteria from the Adult ADHD Clinical Diagnostic Scale. *Archives of General Psychiatry*, **67**(11), 1168–1178, DOI: 10.1001/archgenpsychiatry.2010.146

Kessler R.C., Green J.G., Gruber M.J., Sampson N. A., Bromet E., Cuitan M., Furukawa T.A., Gureje O., Hinkov H., Hu C.Y., Lara C., Lee S., Mneimneh Z., Myer L., Oakley-Browne M., Posada-Villa J., Sagar R., Viana M.C., Zaslavsky A.M. (2010b) Screening for serious mental illness in the general population with the K6 screening scale: results from the WHO World Mental Health (WMH) survey initiative. *International Journal of Methods and Psychiatric Research*, **19**(Suppl 1), 4–22, DOI: 10.1002/mpr.310

Kessler R.C., Lane M., Stang P.E., Van Brunt D.L. (2009) The prevalence and workplace costs of adult attention deficit hyperactivity disorder in a large manufacturing firm. *Psychological Medicine*, **39**(1), 137–147, DOI: 10.1017/S0033291708003309

Kessler R.C., Pennell B.-E. (in press) Developing and selecting mental health measures. In T.P. Johnson (ed.) *Handbook of Health Survey Methods*, New York, John Wiley & Sons.

Kessler R.C., Üstün T.B. (2004) The World Mental Health (WMH) Survey Initiative Version of the World Health Organization (WHO) Composite International Diagnostic Interview (CIDI). *International Journal of Methods and Psychiatric Research*, **13**(2), 93–121, DOI: 10.1002/mpr.168

Kessler R.C., Wittchen H.-U., Abelson J.M., McGonagle K.A., Schwarz N., Kendler K.S., Knäuper B., Zhao S. (1998) Methodological studies of the Composite International Diagnostic Interview (CIDI) in the US National Comorbidity Survey. *International Journal of Methods in Psychiatric Research*, **7**(1), 33–55.

Kraemer H.C. (1992) *Evaluating Medical Tests: Objective and Quantitative Guidelines*, Newbury Park, CA, Sage Publications.

Kroenke K., Spitzer R.L., Williams J.B. (2001) The PHQ-9: validity of a brief depression severity measure. *Journal of General Internal Medicine*, **16**(9), 606–613, DOI: 10.1046/j.1525-1497.2001.016009606.x

Landis J.R., Koch G.G. (1977) The measurement of observer agreement for categorical data. *Biometrics*, **33**(1), 159–174, DOI: 10.2307/2529310

Margolis D.J., Bilker W., Boston R., Localio R., Berlin J.A. (2002) Statistical characteristics of area under the receiver operating characteristic curve for a simple prognostic model using traditional and bootstrapped approaches. *Journal of Clinical Epidemiology*, **55**(5), 518–524, DOI: S0895435601005121 [pii]

Merikangas K.R., Akiskal H.S., Angst J., Greenberg P.E., Hirschfeld R.M., Petukhova M., Kessler R.C. (2007) Lifetime and 12-month prevalence of bipolar spectrum disorder in the National Comorbidity Survey replication. *Archives of General Psychiatry*, **64**(5), 543–552, DOI: 10.1001/archpsyc.64.5.543

Nock M.K., Borges G., Bromet E.J., Cha C.B., Kessler R.C., Lee S. (2008) Suicide and suicidal behavior. *Epidemiologic Reviews*, **30**(1), 133–154, DOI: 10.1093/epirev/mxn002

Nock M.K., Deming C.A., Fullerton C.S., Gilman S.E., Goldenberg M., Kessler R.C., McCarroll J.E., McLaughlin K.A., Peterson C., Schoenbaum M., Stanley B., Ursano R.J. (2013) Suicide among Soldiers: a review of psychological risk and protective factors. *Psychiatry*, **76**(2), 97–125, DOI: 10.1521/psyc.2013.76.2.97

Nock M.K., Hwang I., Sampson N., Kessler R.C., Angermeyer M., Beautrais A., Borges G., Bromet E., Bruffaerts R., de Girolamo G., de Graaf R., Florescu S., Gureje O., Haro J.M., Hu C., Huang Y., Karam E.G., Kawakami N., Kovess V., Levinson D., Posada-Villa J., Sagar R., Tomov T., Viana M.C., Williams D.R. (2009) Cross-national analysis of the associations among mental disorders and suicidal behavior: findings from the WHO World Mental Health Surveys. *PLoS Medicine*, **6**(8), e1000123, DOI: 10.1371/journal.pmed.1000123

Pepe M.S. (2003) *Statistical Analysis of Medical Tests for Classification and Prediction*, New York, Oxford University Press.

Rohde P., Lewinsohn P.M., Seeley J.R. (1997) Comparability of telephone and face-to-face interviews in assessing axis I and II disorders. *American Journal of Psychiatry*, **154**(11), 1593–1598.

Rubin D.B. (1987) *Multiple Imputation for Nonresponse in Surveys*, New York, John Wiley & Sons.

Schafer J.L. (2003) Multiple imputation in multivariate problems when the imputation and analysis models differ. *Statistica Neerlandica*, **57**(1), 19–35.

Sobin C., Weissman M.M., Goldstein R.B., Adams P., Wickramaratne P., Warner V., Lish J.D. (1993) Diagnostic interviewing for family studies: comparing telephone and face-to-face methods for the diagnosis of lifetime psychiatric disorders. *Psychiatric Genetics*, **3**(4), 227–233.

Spitzer R.L., Kroenke K., Williams J.B., Lowe B. (2006) A brief measure for assessing generalized anxiety disorder: the GAD-7. *Archives of Internal Medicine*, **166**(10), 1092–1097, DOI: 10.1001/archinte.166.10.1092

Ursano R.J., Heeringa S., Stein M.B., Kessler R.C. (submitted for publication) The Army Study to Assess Risk and Resilience in Servicemembers (STARRS).

Weathers F., Litz B., Herman D., Huska J., Keane T. (1993) The PTSD checklist (PCL): reliability, validity, and diagnostic utility. *Annual meeting of the International Society for Traumatic Stress Studies*, San Antonio, TX.

White I.R., Royston P., Wood A.M. (2011) Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in Medicine*, **30**(4), 377–399, DOI: 10.1002/sim.4067

Wilkins K.C., Lang A.J., Norman S.B. (2011) Synthesis of the psychometric properties of the PTSD checklist (PCL) military, civilian, and specific versions. *Depression and Anxiety*, **28**(7), 596–606, DOI: 10.1002/da.20837

Wolter K.M. (1985) *Introduction to Variance Estimation*, New York, Springer-Verlag.