

# How well can post-traumatic stress disorder be predicted from pre-trauma risk factors? An exploratory study in the WHO World Mental Health Surveys

RONALD C. KESSLER<sup>1</sup>, SHERRI ROSE<sup>1</sup>, KARESTAN C. KOENEN<sup>2</sup>, ELIE G. KARAM<sup>3</sup>, PAUL E. STANG<sup>4</sup>, DAN J. STEIN<sup>5</sup>, STEVEN G. HEERINGA<sup>6</sup>, ERIC D. HILL<sup>1</sup>, ISRAEL LIBERZON<sup>7</sup>, KATIE A. MCLAUGHLIN<sup>8</sup>, SAMUEL A. MCLEAN<sup>9</sup>, BETH E. PENNELL<sup>6</sup>, MARIA PETUKHOVA<sup>1</sup>, ANTHONY J. ROSELLINI<sup>1</sup>, AYELET M. RUSCIO<sup>10</sup>, VICTORIA SHAHLY<sup>1</sup>, ARIEH Y. SHALEV<sup>11</sup>, DERRICK SILOVE<sup>12</sup>, ALAN M. ZASLAVSKY<sup>1</sup>, MATTHIAS C. ANGERMEYER<sup>13</sup>, EVELYN J. BROMET<sup>14</sup>, JOSÉ MIGUEL CALDAS DE ALMEIDA<sup>15</sup>, GIOVANNI DE GIROLAMO<sup>16</sup>, PETER DE JONGE<sup>17</sup>, KOEN DEMYTTENAERE<sup>18</sup>, SILVIA E. FLORESCU<sup>19</sup>, OYE GUREJE<sup>20</sup>, JOSEP MARIA HARO<sup>21</sup>, HRISTO HINKOV<sup>22</sup>, NORITO KAWAKAMI<sup>23</sup>, VIVIANE KOVÉSS-MASFETY<sup>24</sup>, SING LEE<sup>25</sup>, MARIA ELENA MEDINA-MORA<sup>26</sup>, SAMUEL D. MURPHY<sup>27</sup>, FERNANDO NAVARRO-MATEU<sup>28</sup>, MARINA PIAZZA<sup>29</sup>, JOSE POSADA-VILLA<sup>30</sup>, KATE SCOTT<sup>31</sup>, YOLANDA TORRES<sup>32</sup>, MARIA CARMEN VIANA<sup>33</sup>

<sup>1</sup>Department of Health Care Policy, Harvard Medical School, 180 Longwood Ave., Boston, MA 02115, USA; <sup>2</sup>Mailman School of Public Health, Columbia University, New York, NY, USA; <sup>3</sup>Balamand University Medical School and Institute for Development, Research, Advocacy and Applied Care (IDRAAC), Beirut, Lebanon; <sup>4</sup>Janssen Research & Development, Titusville, NJ, USA; <sup>5</sup>University of Cape Town, Cape Town, South Africa; <sup>6</sup>Institute for Social Research, University of Michigan, Ann Arbor, MI, USA; <sup>7</sup>Department of Psychology, University of Michigan, Ann Arbor, MI, USA; <sup>8</sup>University of Washington, Seattle, WA, USA; <sup>9</sup>University of North Carolina at Chapel Hill, Chapel Hill, NC, USA; <sup>10</sup>University of Pennsylvania, Philadelphia, PA, USA; <sup>11</sup>Hadassah University Hospital, Jerusalem, Israel; <sup>12</sup>University of New South Wales and Liverpool Hospital, Sydney, Australia; <sup>13</sup>Center for Public Mental Health, Gösingam Wagram, Austria; <sup>14</sup>State University of New York at Stony Brook, Stony Brook, NY, USA; <sup>15</sup>Universidade Nova de Lisboa, Lisbon, Portugal; <sup>16</sup>IRCCS Centro San Giovanni di Dio Fatebenefratelli, Brescia, Italy; <sup>17</sup>University of Groningen, Groningen, The Netherlands; <sup>18</sup>University Hospital Gasthuisberg, Leuven, Belgium; <sup>19</sup>National School of Public Health Management and Professional Development, Bucharest, Romania; <sup>20</sup>University of Ibadan, Ibadan, Nigeria; <sup>21</sup>Parc Sanitari Sant Joan de Déu, Sant Boi de Llobregat, Spain; <sup>22</sup>National Center for Public Health Protection, Sofia, Bulgaria; <sup>23</sup>University of Tokyo, Tokyo, Japan; <sup>24</sup>Université Paris Descartes and EHESP School for Public Health, Paris, France; <sup>25</sup>Chinese University of Hong Kong, Hong Kong SAR, China; <sup>26</sup>Instituto Nacional de Psiquiatria Ramon de La Fuente Muñiz, Tlalpan, Mexico City, Mexico; <sup>27</sup>University of Ulster, Londonderry, Northern Ireland, UK; <sup>28</sup>Servicio Murciano de Salud and CIBER de Epidemiología y Salud Pública (CIBERESP), El Palmar, Spain; <sup>29</sup>Universidad Peruana Cayetano Heredia, Lima, Peru; <sup>30</sup>Universidad Colegio Mayor de Cundinamarca, Bogota, Colombia; <sup>31</sup>University of Otago, Dunedin, New Zealand; <sup>32</sup>University Center of Excellence on Mental Health Research, Medellín, Colombia; <sup>33</sup>Federal University of Espirito Santo (UFES), Vitoria, Brazil

*Post-traumatic stress disorder (PTSD) should be one of the most preventable mental disorders, since many people exposed to traumatic experiences (TEs) could be targeted in first response settings in the immediate aftermath of exposure for preventive intervention. However, these interventions are costly and the proportion of TE-exposed people who develop PTSD is small. To be cost-effective, risk prediction rules are needed to target high-risk people in the immediate aftermath of a TE. Although a number of studies have been carried out to examine prospective predictors of PTSD among people recently exposed to TEs, most were either small or focused on a narrow sample, making it unclear how well PTSD can be predicted in the total population of people exposed to TEs. The current report investigates this issue in a large sample based on the World Health Organization (WHO)'s World Mental Health Surveys. Retrospective reports were obtained on the predictors of PTSD associated with 47,466 TE exposures in representative community surveys carried out in 24 countries. Machine learning methods (random forests, penalized regression, super learner) were used to develop a model predicting PTSD from information about TE type, socio-demographics, and prior histories of cumulative TE exposure and DSM-IV disorders. DSM-IV PTSD prevalence was 4.0% across the 47,466 TE exposures. 95.6% of these PTSD cases were associated with the 10.0% of exposures (i.e., 4,747) classified by machine learning algorithm as having highest predicted PTSD risk. The 47,466 exposures were divided into 20 ventiles (20 groups of equal size) ranked by predicted PTSD risk. PTSD occurred after 56.3% of the TEs in the highest-risk ventile, 20.0% of the TEs in the second highest ventile, and 0.0-1.3% of the TEs in the 18 remaining ventiles. These patterns of differential risk were quite stable across demographic-geographic sub-samples. These results demonstrate that a sensitive risk algorithm can be created using data collected in the immediate aftermath of TE exposure to target people at highest risk of PTSD. However, validation of the algorithm is needed in prospective samples, and additional work is warranted to refine the algorithm both in terms of determining a minimum required predictor set and developing a practical administration and scoring protocol that can be used in routine clinical practice.*

**Key words:** Post-traumatic stress disorder, predictive modeling, machine learning, penalized regression, random forests, ridge regression

(*World Psychiatry* 2014;13:265–274)

Post-traumatic stress disorder (PTSD) is a commonly occurring and seriously impairing disorder (1). Many people exposed to the traumatic experiences (TEs) that lead to PTSD come to the attention of the criminal justice or health care system short-

ly after exposure and could be targeted through these systems for early preventive interventions. In recognition of this fact, an increasing amount of research has been carried out to develop and evaluate early preventive interventions for PTSD.

While the interventions developed for delivery in the first few hours after TE exposure have so far proven ineffective (2), cognitive-behavioral (3) and prolonged exposure (4) therapies delivered within a few weeks after TE exposure have been shown to be moderately effective in preventing chronic PTSD. In addition, ongoing research suggests that a wider range of potentially cost-effective preventive interventions might become available in the future (5).

Importantly, though, these preventive interventions for PTSD are labor-intensive, making them infeasible to offer cost-effectively to all people exposed to TEs (1). Prediction rules that successfully target people at highest PTSD risk shortly after TE exposure could improve intervention cost-effectiveness.

Meta-analyses (6-8) and reviews (9-11) of studies that searched for these predictors point to six especially promising predictor classes: type-severity of TE (highest PTSD risk associated with physical or sexual assault) (7,12); socio-demographics (e.g., female gender and young age) (6,8,9); cumulative prior TE exposure (including exposure to childhood family trauma) (6,7,10); prior mental disorders (especially anxiety, mood, and conduct disorders) (10,11); acute emotional and biological responses (6,7,11,13); and proximal social factors occurring in the days and weeks *after* TE exposure (e.g., low social support, heightened life stress) (6,7).

This literature offers no guidance on how to combine information about these predictors into an optimal PTSD risk algorithm. Machine learning methods have been used to develop similar algorithms in other areas of medicine (14,15). However, most studies using information obtained shortly after TE exposure to predict PTSD are based on samples too small (typically N=100-300) to apply these methods. This limitation could be overcome if future prospective studies were either much larger or used much more consistent measures (to allow individual-level data pooling for secondary analysis) than studies carried out up to now.

Prior to that time, a preliminary PTSD risk algorithm could be developed from the first four classes of predictors enumerated above (i.e., socio-demographics, type of focal TE, prior TE exposure, prior psychopathology), based on analysis of existing cross-sectional community epidemiological studies. The latter studies tend to be quite large, which means that machine learning methods could be applied. Although limited by being cross-sectional and relying on retrospective reports to examine associations of putative predictors with subsequent PTSD, these preliminary prediction algorithms could be validated in small prospective studies (that are themselves too small for algorithm development).

The current report presents the results of developing a preliminary PTSD risk algorithm from cross-sectional data in the World Health Organization (WHO)'s World Mental Health (WMH) Surveys ([www.hcp.med.harvard.edu/WMH](http://www.hcp.med.harvard.edu/WMH)), a series of community epidemiological surveys in 24 countries that included retrospective assessments of PTSD associated with 47,466 lifetime TE exposures. The large and geographically

dispersed sample, coupled with the great variety of TEs and predictors assessed, make this database attractive for developing a preliminary PTSD risk algorithm. If the algorithm appears to perform well, it could subsequently be validated in smaller prospective studies and used as a starting point for data collection in future prospective studies.

## METHODS

### Samples

The WMH surveys were conducted in thirteen countries classified by the World Bank (16) as high income (Australia, Belgium, France, Germany, Israel, Italy, Japan, Spain, The Netherlands, New Zealand, Northern Ireland, Portugal, United States), seven upper-middle income (São Paulo in Brazil, Bulgaria, Lebanon, Mexico, Romania, South Africa, Ukraine), and four lower-middle income (Colombia, Nigeria, Beijing and Shanghai in the People's Republic of China, Peru). Most surveys were based on national household samples, the exceptions being surveys of all urbanized areas in Colombia and Mexico, specific metropolitan areas in Brazil, China and Spain, a series of cities in Japan, and two regions in Nigeria. Response rates were in the range 45.9-97.2% and averaged 70.4%. More detailed sample descriptions are presented elsewhere (17).

Interviews were administered face-to-face in two parts after obtaining informed consent using procedures approved by local institutional review boards. Part I, administered to all respondents (N=126,096), assessed core DSM-IV mental disorders. Part II, administered to all Part I respondents with any lifetime Part I disorder plus a probability sub-sample of other Part I respondents (N=69,272), assessed additional disorders, including PTSD, and correlates. Part II respondents were weighted by the inverse of their probability of selection from Part I. More details about WMH sample designs and weighting are presented elsewhere (17). The 42,634 Part II respondents who reported lifetime TEs included a sub-sample of 13,610 subjects who were exposed only once to only a single TE and an additional sub-sample of 29,024 subjects who reported multiple TE exposures.

PTSD was assessed for each of the 13,610 exposures in the first sub-sample. The 29,024 respondents with multiple TEs were asked to select a "worst" TE using a two-part question sequence. The first of the two-part sequence asked: "Let me review. You had (two/three/quite a few) different traumatic experiences. After an experience like this, people sometimes have problems like upsetting memories or dreams, feeling emotionally distant or depressed, trouble sleeping or concentrating, and feeling jumpy or easily startled. Did you have any of these reactions after (either/any) of these experiences?"

The 9,791 respondents answering "yes" were then asked the second question in the two-part series: "Of the experiences

you reported, which one caused you the most problems like that?”. PTSD was assessed for each exposure reported in response to this question. However, as these “worst” TEs cannot be taken to describe all TEs these respondents experienced, we also assessed PTSD for one exposure selected at random for a probability sub-sample of respondents with multiple exposures (N=4,832). The observational record for each “worst” exposure was assigned a weight of 1, while that for each randomly selected exposure was assigned a weight of  $1/tp$  ( $t$ =number of TEs reported by the respondent other than the worst TE;  $p$ =probability of case selection), in order to make the total sample of 47,466 exposures assessed representative of all lifetime TE exposures of all respondents.

## PTSD diagnosis

Mental disorders were assessed with the Composite International Diagnostic Interview (CIDI, 18), a fully-structured lay-administered interview yielding DSM-IV diagnoses. A clinical reappraisal study carried out in several WMH Surveys (19), assessing the CIDI concordance for DSM-IV PTSD with the Structured Clinical Interview for DSM-IV (SCID) (20) used as the gold standard, found an area under the curve (AUC) of 0.69, a sensitivity of 38.3, and a specificity of 99.1. The resulting likelihood ratio positive (LR+) of 42 is well above the threshold of 10 typically used to consider screening scale diagnoses definitive. Consistent with the high LR+, positive predictive value was 86.1%, suggesting that the vast majority of CIDI cases would be judged to have PTSD in independent clinical evaluations.

## Predictors of PTSD

### *Socio-demographics*

Socio-demographics included gender along with age, education, and marital status at focal TE exposure.

### *Traumatic experiences*

WMH Surveys assessed 29 TE types, including 27 specific types from a list, one open-ended question about TEs not included in the list, and a final yes-no question about any other lifetime TE that respondents did not wish to describe concretely (referred to as a “private event”). Respondents were probed separately about number of lifetime occurrences and age at first occurrence of each TE type reported.

Exploratory factor analysis found that the vast majority of TE types loaded on one of five broad factors (Table 1) referred to below as “exposure to organized violence”, “participation in organized violence”, “interpersonal violence”, “sexual-relationship violence”, and “other life-threatening TEs”. Predictors of PTSD included a separate dummy variable for each focal TE type in addition to 29 dummy variables for prior life-

time exposure to the same types. Temporal clustering among TEs was captured by creating counts of prior lifetime exposure to TEs in each factor and of other TEs in each factor in the same year as exposure to the focal TE.

### *Prior mental disorders*

The CIDI assessed seven lifetime DSM-IV internalizing disorders in addition to PTSD (separation anxiety disorder, specific phobia, social phobia, agoraphobia and/or panic disorder, generalized anxiety disorder, major depressive disorder and/or dysthymia, bipolar disorder I-II) and six lifetime externalizing disorders (attention-deficit/hyperactivity disorder (ADHD), intermittent explosive disorder, oppositional-defiant disorder (ODD), conduct disorder (CD), alcohol abuse with or without dependence, drug abuse with or without dependence).

Age of onset of each disorder was assessed using special probing techniques shown experimentally to improve recall accuracy (21). DSM-IV organic exclusion rules and diagnostic hierarchy rules were used other than for ODD (defined with or without CD) and substance abuse (defined with or without dependence). As detailed elsewhere (19), generally good concordance was found between diagnoses based on the CIDI and blinded clinical diagnoses based on the SCID (20).

## Analysis methods

Conventional multiple regression (with all predictors in the model) (22) and four machine learning algorithms were used to predict PTSD. The machine learning algorithms included random forests (23) and three elastic net penalized logistic regressions (24) designed to address two problems in conventional multiple regression: that coefficients are unstable when high correlations exist among predictors, which is the case for the predictors considered here, leading to low replication of predictions in independent samples (25); and that conventional regression assumes additivity, whereas the predictors considered here might have non-additive effects (7,8,10).

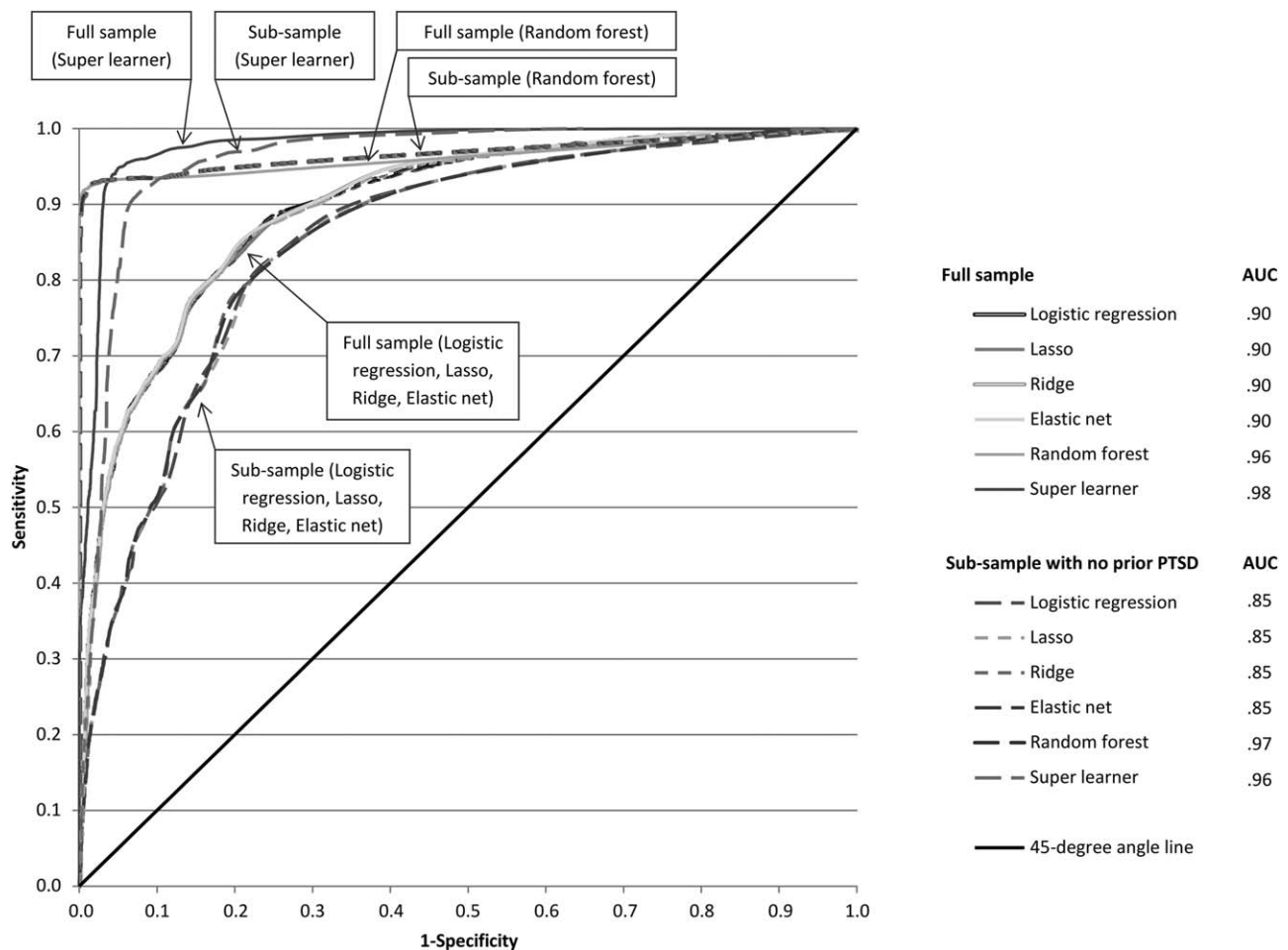
Random forests is an ensemble machine learning method that generates many regression trees to detect interactions, each based on a separate bootstrapped pseudo-sample to protect against over-fitting, and assigns individual-level predicted probabilities of outcomes based on modal values across replicates (23). The algorithm was implemented in the R-package randomForest (26). The R-package r-part (27) was also used to examine the distribution of higher-order interactions underlying the data.

Elastic net penalized regression is an approach that trades off bias to decrease standard errors of estimates, reducing instability caused by high correlations among predictors using a mixing parameter penalty (MPP) that varies in the range 0-1. The three penalties we used included: the lasso penalty (MPP=1.0), which favors sparse models that force coefficients for all but one predictor in each strongly

**Table 1** Distribution and conditional risk of DSM-IV/CIDI PTSD associated with exposure to the 29 types of traumatic experience (TE) assessed in the WMH Surveys (N=47,566)

	Proportion of all TE exposures % (SE)	Conditional risk of PTSD % (SE)	Proportion of all PTSD % (SE)
<b>Exposed to organized violence</b>			
Civilian in war zone	1.4 (0.1)	1.3 (0.5)	0.5 (0.2)
Civilian in region of terror	1.0 (0.1)	1.6 (0.6)	0.4 (0.1)
Relief worker in war zone	0.3 (0.1)	0.8 (0.7)	0.1 (0.1)
Refugee	0.7 (0.1)	4.5 (2.0)	0.8 (0.4)
Kidnapped	0.4 (0.1)	11.0 (3.0)	1.0 (0.3)
Any	3.9 (0.2)	2.9 (0.5)	2.8 (0.5)
<b>Participated in organized violence</b>			
Combat experience	1.0 (0.1)	3.6 (0.8)	0.9 (0.2)
Witnessed death/serious injury or discovered dead body	16.2 (0.5)	1.3 (0.3)	5.3 (1.0)
Saw atrocities	2.7 (0.3)	5.4 (4.1)	3.7 (2.8)
Accidentally caused death/serious injury	0.7 (0.1)	2.8 (1.0)	0.5 (0.2)
Purposefully caused death/serious injury	0.7 (0.1)	4.0 (3.1)	0.7 (0.5)
Any	21.3 (0.6)	2.1 (0.6)	11.2 (3.1)
<b>Interpersonal violence</b>			
Witnessed physical fights at home as a child	2.4 (0.1)	3.9 (0.7)	2.3 (0.4)
Childhood physical abuse	2.7 (0.1)	5.0 (1.0)	3.4 (0.7)
Beaten by someone else (not spouse/partner)	3.3 (0.2)	2.5 (0.6)	2.1 (0.5)
Mugged or threatened with weapon	8.2 (0.3)	1.8 (0.4)	3.8 (0.8)
Any	16.6 (0.4)	2.8 (0.3)	11.5 (1.3)
<b>Sexual-relationship violence</b>			
Beaten by spouse/partner	1.4 (0.1)	11.7 (1.3)	4.1 (0.5)
Raped	1.8 (0.1)	19.0 (2.2)	8.4 (1.0)
Sexually assaulted	3.2 (0.2)	10.5 (1.5)	8.4 (1.2)
Stalked	2.9 (0.2)	7.6 (2.0)	5.4 (1.4)
Other event	1.4 (0.1)	9.1 (1.0)	3.1 (0.4)
“Private event” (see text)	1.5 (0.1)	9.2 (1.1)	3.5 (0.4)
Any	12.1 (0.3)	10.9 (0.8)	32.9 (2.1)
<b>Other life-threatening TEs</b>			
Life-threatening illness	5.1 (0.2)	2.0 (0.3)	2.5 (0.4)
Life-threatening motor vehicle accident	6.2 (0.2)	2.6 (0.4)	4.1 (0.7)
Other life-threatening accident	3.0 (0.2)	4.9 (2.3)	3.7 (1.8)
Natural disaster	3.9 (0.4)	0.3 (0.1)	0.3 (0.1)
Toxic chemical exposure	3.5 (0.3)	0.1 (0.0)	0.1 (0.0)
Other man-made disaster	1.9 (0.2)	2.9 (1.3)	1.4 (0.7)
Any	23.7 (0.6)	2.0 (0.4)	12.0 (2.1)
<b>Network traumatic experiences</b>			
Unexpected death of loved one	16.8 (0.4)	5.4 (0.5)	22.6 (1.9)
Life-threatening illness of child	3.3 (0.1)	4.8 (0.6)	4.0 (0.5)
Other traumatic experience of loved one	2.4 (0.2)	5.1 (1.3)	3.1 (0.8)
Any	22.5 (0.4)	5.3 (0.4)	29.7 (2.0)
<b>Total</b>	100.0	4.0 (0.2)	100.0

CIDI – Composite International Diagnostic Interview, PTSD – post-traumatic stress disorder, WMH – World Mental Health



**Figure 1** Receiver operating characteristic (ROC) curves for predicted probability of DSM-IV/CIDI PTSD after TE exposure based on the different algorithms in the total sample (N=47,466) and the sub-sample of exposures occurring to respondents with no history of prior PTSD (N=45,556). CIDI – Composite International Diagnostic Interview, TE – traumatic experience, PTSD – post-traumatic stress disorder, AUC – area under the curve

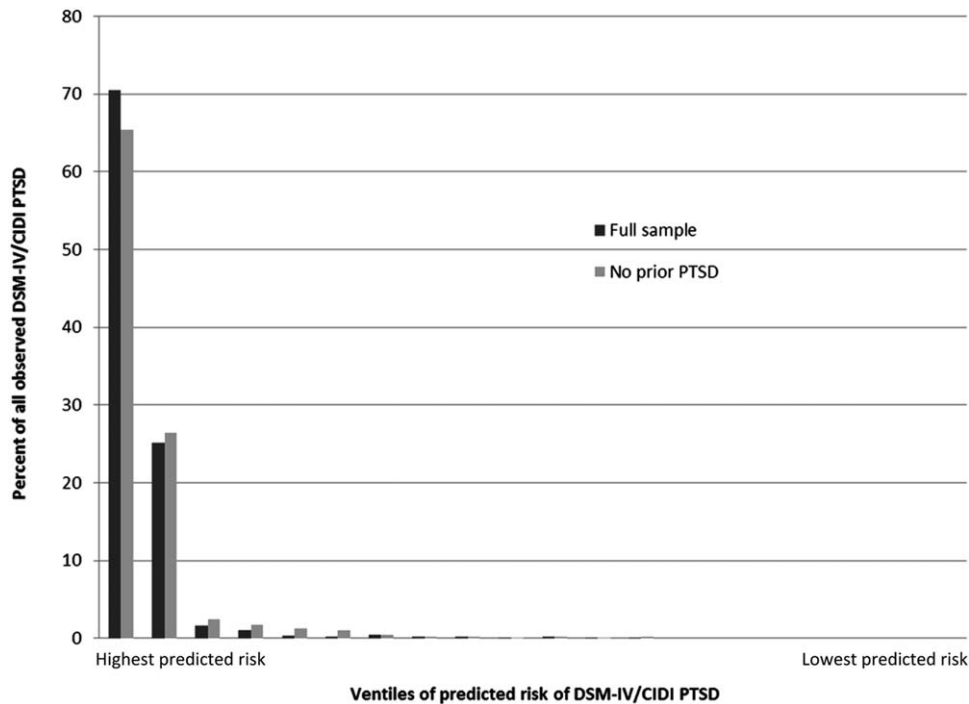
correlated set to zero; the ridge penalty (MPP=0), which uses proportional coefficient shrinkage to retain all predictors; and an intermediate elastic net (MPP=0.5), which combines both approaches. Internal cross-validation was used to select the coefficient in front of the penalty. The algorithms were implemented in the R-package glmnet (24).

Finally, we used an ensembling method known as super learner (28,29) to generate an optimally weighted composite prediction algorithm averaged across the five individual algorithms using internal cross-validation implemented in the R-package Super Learner (30).

It is important to note that the internal cross-validation used in the penalized regressions improves on a more conventional approach, that fits a model in a discovery sample and then tests the model fit in a hold-out sample, in two ways. First, the internal cross-validation used the 10-fold cross-validation technique, which divides the sample into 10 equal-sized sub-samples and estimates a model for each of a large number of fixed coefficients in front of the penalty 10 times, in each of which one of the sub-samples is held out

and then the coefficients are applied to the hold-out sample. Model fit was then estimated across the 10 hold-out sub-samples to evaluate model fit for the fixed value of the coefficient in front of the penalty. The value of that coefficient was then selected to maximize cross-validated model fit. Second, MPP itself was varied, which leads to variation in the number of predictors in the model. Super learner applied a separate 10-fold cross-validation to this entire set of procedures to assign differential weights to the models with different MPP values as well as to the other algorithms.

Individual-level predicted PTSD probabilities based on the separate algorithms and super learner were created, receiver operating characteristic (ROC) curves generated, and AUC calculated to evaluate prediction accuracy. Super learner predicted probabilities were then discretized into ventiles (20 groups of equal size ordered by percentiles) and cross-classified with observed PTSD. As prior PTSD was a dominant predictor in all algorithms, analysis was replicated for the 45,556 TE exposures that occurred to respondents without a history of prior PTSD. All analyses were based on



**Figure 2** Concentration of risk for DSM-IV/CIDI PTSD. CIDI – Composite International Diagnostic Interview, PTSD – post-traumatic stress disorder

weighted data to adjust for individual differences in probabilities of TE selection.

## RESULTS

### Distribution and associations of TEs with PTSD

Weighted DSM-IV/CIDI PTSD prevalence was 4.0% in the total sample, and ranged across TEs between 0.1-0.3% (natural and man-made disasters) and 19.0% (rape) ( $\chi^2=639.4$ ,  $df=28$ ,  $p<0.001$ ) (Table 1).

The three TEs accounting for the highest proportions of PTSD cases included unexpected death of a loved one, rape, and other sexual assault. Unexpected death of a loved one was the most commonly reported TE (accounting for 16.8% of all TE exposures) and accounted for a somewhat higher proportion of PTSD (22.6%) than of all TE exposures, due to a conditional risk of PTSD slightly higher than average (5.4%). The other two TEs with highest proportions of PTSD cases (8.4% each) were rape and sexual assault. Rape and sexual assault were both much less common than unexpected death of a loved one (rape accounting for 1.8% of all TE exposures and sexual assault for 3.2%), but had much higher conditional PTSD risks (19.0 and 10.5%, respectively).

These high conditional PTSD risks associated with rape and sexual assault were part of a broader pattern of highest PTSD risk being associated with TEs involving interpersonal violence (kidnapping, beaten by spouse/partner, rape,

sexual assault), which together accounted for 6.8% of TE exposures and 21.9% of PTSD.

### Concentration of risk

ROC curves show that super learner substantially outperformed the individual algorithms other than random forests (AUC=0.96 vs. 0.90 in the total sample; 0.97 vs. 0.85 in the sub-sample with no prior PTSD) (Figure 1).

Inspection of observed PTSD distributions across ventiles of predicted risk based on super learner shows that 95.6% of observed PTSD occurred after the 10% of exposures having highest predicted risk (Figure 2). Conditional PTSD risk was 56.3% in the highest ventile, 20.0% in the second highest ventile, and 0.0-1.3% in the remaining 18 ventiles.

In the sub-sample with no history of prior PTSD, 91.9% of observed PTSD occurred after the 10% of exposures having highest predicted risk. Conditional PTSD risk was 32.2% for the highest ventile, 13.0% for the second highest ventile, and 0.0-1.2% in the remaining 18 ventiles.

### Stability of results

Results were found to be stable across sub-samples defined by individual-level characteristics (sex, age, income) and country-level characteristics (economic development, recent history of war or sectarian violence) (Table 2). Between

**Table 2** Concentration of observed DSM-IV/CIDI PTSD in the 10% of exposures having highest predicted risk based on the super learner algorithm across sub-samples

	Proportion of all PTSD associated with the 10% of exposures having highest predicted risk		Conditional observed PTSD risk in the 10% of exposures having highest predicted risk vs. other exposures			
	Total sample % (SE)	Respondents with no prior PTSD % (SE)	Total sample		Respondents with no prior PTSD	
			Top 10% % (SE)	Other 90% % (SE)	Top 10% % (SE)	Other 90% % (SE)
<b>Gender</b>						
Male	97.6 (0.6)	96.4 (1.1)	43.4 (3.8)	0.06 (0.01)	27.7 (4.2)	0.05 (0.01)
Female	94.3 (1.0)	92.6 (1.4)	35.9 (1.4)	0.40 (0.07)	21.4 (1.1)	0.34 (0.07)
<b>Age at TE exposure</b>						
Less than 25	95.2 (0.9)	93.3 (1.4)	41.5 (2.4)	0.22 (0.04)	23.0 (2.3)	0.18 (0.03)
25 or older	95.8 (1.2)	94.6 (1.8)	33.9 (2.0)	0.18 (0.05)	23.3 (1.6)	0.16 (0.06)
<b>Education*</b>						
Low/low-average	94.7 (0.8)	92.6 (1.2)	37.5 (1.8)	0.24 (0.04)	21.0 (1.4)	0.20 (0.03)
High-average/high	96.9 (1.4)	96.1 (2.1)	39.5 (3.1)	0.13 (0.06)	28.2 (2.9)	0.11 (0.06)
<b>Country World Bank income level</b>						
High	95.5 (0.6)	94.3 (0.8)	34.3 (1.5)	0.26 (0.03)	20.5 (1.2)	0.20 (0.03)
All others	95.3 (2.0)	92.6 (3.2)	52.6 (3.6)	0.14 (0.06)	33.2 (4.2)	0.13 (0.06)
<b>Country involved in war or sectarian violence**</b>						
Yes	95.9 (1.7)	94.4 (2.5)	44.4 (5.2)	0.13 (0.05)	28.8 (5.2)	0.13 (0.05)
No	95.3 (0.8)	93.6 (1.2)	36.6 (1.3)	0.23 (0.04)	21.6 (1.0)	0.19 (0.04)

PTSD – post-traumatic stress disorder, TE – traumatic experience

\*Educational level relative to others in the same country; \*\*countries classified “yes” include Colombia, Israel, Lebanon, Nigeria, Northern Ireland, and South Africa

94.3% and 97.6% of observed PTSD in each sub-sample was associated with the 10% of TEs having highest predicted risk (92.6-96.4% in the sub-sample with no prior PTSD). PTSD prevalence in these high-risk sub-samples was 33.9-52.6% (20.5-33.2% in the sub-sample with no prior PTSD).

### Components of risk

Although it is hazardous to interpret individual machine learning model coefficients, since the algorithms maximize overall prediction accuracy at the expense of individual coefficient accuracy, an understanding of important predictors is nonetheless useful. We used a two-part approach to achieve this understanding. We first examined multivariate predictor profiles in 100 bootstrapped r-part trees to determine which interactions were important. These profiles all involved history of prior PTSD interacting either with history of prior unexpected death of a loved one and/or with history of prior sexual trauma. We then included dummy varia-

bles for those multivariate profiles along with variables for the main effects of individual predictors in lasso regressions.

Socio-demographic differences in PTSD risk were restricted to elevated odds ratios among women (1.5-1.6) and the previously married (1.5). Nine TE types also had elevated odds ratios, all but one of which (the exception being exposure to a life-threatening accident, with odds ratios of 1.4-1.8) involved interpersonal violence: rape (3.2-3.5), kidnap (1.8-3.4), childhood physical abuse (1.5-1.8), witnessing atrocities (1.4), and four other (than rape) TE types in the relationship-sexual violence factor (1.5-1.8). Three TE types had meaningfully reduced odds ratios: witnessing death/injury, toxic chemical spill, and natural disasters (0.4-0.7) (Table 3).

Five summary measures of collateral TEs occurring in the same year as the focal TE had meaningful odds ratios. Four of these five (the exception being unexpected death of a loved one) involved violence: two or more TEs in the organized violence factor, three or more TEs in the interpersonal violence factor, and any as well as two or more TEs in

**Table 3** Lasso penalized logistic regression coefficients (odds ratios) to predict onset of DSM-IV/CIDI PTSD after exposure to a traumatic experience (TE)

	Total sample OR	Sub-sample without prior PTSD OR	Total sample with interaction OR
<b>Focal TE</b>			
Organized violence			
Witnessed atrocities	1.4	1.4	1.4
Witnessed death/injury or discovered dead bodies	0.6	0.7	0.6
Kidnapped	3.0	1.8	3.4
Interpersonal violence			
Childhood physical abuse	1.5		
Sexual-relationship violence			
Beaten by spouse-partner	1.8	1.5	1.7
Raped	3.2	3.5	3.5
Sexually assaulted			1.5
“Private TE” (see text)	1.8	1.5	1.8
Some other TE	1.6		1.5
Other			
Toxic chemical exposure	0.5	0.6	0.5
Natural disaster	0.4		0.5
Other life-threatening accident	1.8	1.4	1.6
<b>Collateral TEs</b>			
Multiple (2+) participants in organized violence	2.5	3.8	2.7
High (3+) exposure to interpersonal violence	6.8	11.5	9.3
Any exposure to sexual-relationship violence	1.6	1.7	1.6
Multiple (2+) exposures to sexual-relationship violence	4.0	3.2	3.5
Unexpected death of loved one	2.1	2.1	2.2
<b>Socio-demographics</b>			
Female		1.6	1.5
Previously married	1.5		1.5
<b>Lifetime prior TEs</b>			
Witnessed atrocities	1.6	1.8	1.6
Raped		2.3	1.6
Sexually assaulted			1.4
Unexpected death of loved one			1.6
<b>Lifetime prior DSM-IV/CIDI disorders</b>			
Separation anxiety disorder	2.0	1.7	1.9
Specific phobia			1.5
Attention-deficit/hyperactivity disorder	1.6		1.7
Generalized anxiety disorder	2.2	2.5	2.2
PTSD	27.2		
<b>Interactions of lifetime prior PTSD with lifetime prior</b>			
Sexual violence and unexpected death of loved one			5.4
Sexual violence but no unexpected death of loved one			12.9
Unexpected death of loved one but no sexual violence			5.7
Neither sexual violence nor unexpected death of loved one			134.7

CIDI – Composite International Diagnostic Interview, PTSD – post-traumatic stress disorder, OR – odds ratio



the relationship-sexual violence factor. The collateral TEs involving single exposures (sexual or death) had odds ratios of 1.6-2.2, while those involving two-three or more exposures had odds ratios of 2.5-11.5. Four of the 29 prior lifetime TE types had meaningful odds ratios, all of them elevated: witnessing atrocities (1.6-1.8), being raped (1.6-2.3) or sexually assaulted (1.4), and experiencing the unexpected death of a loved one (1.6) (Table 3).

Five of the 14 prior lifetime DSM-IV/CIDI disorders had meaningful odds ratios: ADHD (1.6-1.7), separation anxiety disorder (1.7-2.0), specific phobia (1.5), generalized anxiety disorder (2.2-2.5), and PTSD (27.2) (Table 3). The high odds ratio for prior PTSD was due to the 3.5% of exposures occurring to respondents with prior PTSD accounting for 40.5% of all episodes of PTSD. Disaggregation into underlying multivariate profiles showed that this strong effect of prior PTSD was limited to people marked as vulnerable by virtue of having past PTSD associated with TEs generally not associated with high PTSD risk.

## DISCUSSION

Although caution is needed in interpreting these results, since the WMH Survey data were based on retrospective reports and fully structured lay-administered diagnostic interviews, it is nonetheless striking that we were able to produce a prediction algorithm in which the vast majority of PTSD cases were associated with the 10% of TE exposures having highest predicted risk. By far the most powerful predictor in the algorithm was history of prior PTSD, but a number of other prior lifetime mental disorders were also significant predictors, along with a number of measures of prior lifetime trauma exposure as well as socio-demographic characteristics of respondents and information about characteristics of the focal TE.

Limitations introduced by the retrospective nature of the data could have led to upward bias in odds ratios if respondents defined as having a history of PTSD were more likely than others to recall prior lifetime TE exposures and/or mental disorders. Importantly, evidence has been presented in the literature that this type of bias does, in fact, exist in retrospective reports among people with PTSD (31-33). In addition, the concentration-of-risk estimates could have been upwardly biased compared to those that would be found among people who sought help in the immediate aftermath of TE exposure in criminal justice or health care settings, to the extent that this help-seeking predicted subsequent PTSD.

Despite these limitations, the WMH Survey results are important in suggesting that a PTSD risk algorithm based on machine learning methods might help improve targeting and subsequent cost-effectiveness of preventive interventions for PTSD by pinpointing the small proportion of TE-exposed people having high PTSD risk that account for

most subsequent PTSD. Our study is much larger than all other previous studies attempting to predict onset of PTSD from information about trauma types and pre-trauma predictors. We showed that a composite risk score can be constructed from such data that classifies the vast majority of people who go on to develop PTSD into a high-risk segment of the population.

External validation of the risk algorithms in prospective samples is, however, needed. A number of such prospective studies exist. All these studies are much smaller than the WMH Survey database and not all assessed the full range of predictors considered in our analysis. Nonetheless, the strong results found here suggest that it would be valuable to carry out replications in these prospective studies. We are currently involved in several collaborative replication analyses of this sort and are eager to work with others to evaluate the extent to which our algorithm fits in independent prospective samples. In addition, we are interested in collaborating with other groups to apply the methods used here to determine the predictive accuracy of algorithms based on data involving an expanded set of predictors, including biomarkers. If the results of this ongoing work are encouraging, subsequent prospective studies should be designed so that they include the full range of predictors found to be important. This would advance the agenda of creating broadly useful PTSD risk algorithms (and subsequent algorithms to predict other psychopathological responses to TE exposure) to target preventive interventions across a wide range of settings.

It is important to note that different predictors will almost certainly be found to be important in different populations (e.g., military personnel, first responders in disaster situations, civilians in less developed countries), in association with different TEs (e.g., sectarian violence in war zones or regions of terror, large-scale natural disasters) and in different screening settings (e.g., temporary emergency clinics established at the site of natural or man-made disasters, medical clinics in war zones, trauma units, emergency departments) (6,7,11,13). This means that an expansion of the current line of work will ultimately lead either to the development of a family of risk prediction algorithms or to a consolidated master algorithm that allows for complex interactions across populations, TEs, and screening settings.

It will also be important, in developing such algorithms, to be sensitive to variation in the costs of collecting different types of data (e.g., self-reports versus biomarkers), even data ostensibly assessing the same underlying constructs (e.g., self-reports of impulsivity versus neurocognitive tests of impulsivity), as well as the burdens associated with administering detailed risk factors assessments (both in terms of time burden and the psychological burden of asking people detailed questions of a sensitive nature in the immediate aftermath of TE exposure). Thoughtful analysis will be needed of the cost-benefit trade-offs associated with including-excluding expensive and burdensome elements of the assessments depending on strength of predictions.

## Acknowledgements

The research reported here is supported by U.S. Public Health Service grant 1 R01MH101227-01A1 and is carried out in conjunction with the WHO's WMH Survey initiative, which is supported by the National Institute of Mental Health (R01 MH070884 and R01 MH093612-01), the John D. and Catherine T. MacArthur Foundation, the Pfizer Foundation, the U.S. Public Health Service (R13-MH066849, R01-MH069864, and R01 DA016558), the Fogarty International Center (FIRCA R03-TW006481), the Pan American Health Organization, Eli Lilly and Company, Ortho-McNeil Pharmaceutical, GlaxoSmithKline, and Bristol-Myers Squibb. The authors thank the staff of the WMH Data Collection and Data Analysis Coordination Centres for assistance with instrumentation, fieldwork, and consultation on data analysis. None of the funders had any role in the design, analysis, interpretation of results, or preparation of this paper. A complete list of all within-country and cross-national WMH Survey publications can be found at <http://www.hcp.med.harvard.edu/WMH/>. The views and opinions expressed in this report are those of the authors and should not be construed to represent the views of the sponsoring organizations, agencies, or governments.

## References

1. Kessler R. Posttraumatic stress disorder: the burden to the individual and to society. *J Clin Psychiatry* 2000;61(Suppl. 5):4-12.
2. Forneris CA, Gartlehner G, Brownley KA et al. Interventions to prevent post-traumatic stress disorder: a systematic review. *Am J Prev Med* 2013;44:635-50.
3. Kliem S, Kroger C. Prevention of chronic PTSD with early cognitive behavioral therapy. A meta-analysis using mixed-effects modeling. *Behav Res Ther* 2013;51:753-61.
4. Shalev AY, Ankri Y, Israeli-Shalev Y et al. Prevention of posttraumatic stress disorder by early treatment: results from the Jerusalem Trauma Outreach And Prevention study. *Arch Gen Psychiatry* 2012;69:166-76.
5. Ostrowski SA, Delahanty DL. Prospects for the pharmacological prevention of post-traumatic stress in vulnerable individuals. *CNS Drugs* 2014;28:195-203.
6. Brewin CR, Andrews B, Valentine JD. Meta-analysis of risk factors for posttraumatic stress disorder in trauma-exposed adults. *J Consult Clin Psychol* 2000;68:748-66.
7. Ozer EJ, Best SR, Lipsey TL et al. Predictors of posttraumatic stress disorder and symptoms in adults: a meta-analysis. *Psychol Bull* 2003;129:52-73.
8. Tolin DF, Foa EB. Sex differences in trauma and posttraumatic stress disorder: a quantitative review of 25 years of research. *Psychol Bull* 2006;132:959-92.
9. Breslau N. Epidemiologic studies of trauma, posttraumatic stress disorder, and other psychiatric disorders. *Can J Psychiatry* 2002;47:923-9.
10. DiGangi JA, Gomez D, Mendoza L et al. Pretrauma risk factors for posttraumatic stress disorder: a systematic review of the literature. *Clin Psychol Rev* 2013;33:728-44.
11. Heron-Delaney M, Kenardy J, Charlton E et al. A systematic review of predictors of posttraumatic stress disorder (PTSD) for adult road traffic crash survivors. *Injury* 2013;44:1413-22.
12. Roberts AL, Gilman SE, Breslau J et al. Race/ethnic differences in exposure to traumatic events, development of post-traumatic stress disorder, and treatment-seeking for post-traumatic stress disorder in the United States. *Psychol Med* 2011;41:71-83.
13. van Zuiden M, Kavelaars A, Geuze E et al. Predicting PTSD: pre-existing vulnerabilities in glucocorticoid-signaling and implications for preventive interventions. *Brain Behav Immun* 2013;30:12-21.
14. Baca-Garcia E, Perez-Rodriguez MM, Basurte-Villamor I et al. Using data mining to explore complex clinical decisions: a study of hospitalization after a suicide attempt. *J Clin Psychiatry* 2006;67:1124-32.
15. Connor JP, Symons M, Feeney GF et al. The application of machine learning techniques as an adjunct to clinical decision making in alcohol dependence treatment. *Subst Use Misuse* 2007;42:2193-206.
16. World Bank. Data: countries and economies. <http://data.worldbank.org/country>.
17. Heeringa SG, Wells EJ, Hubbard F et al. Sample designs and sampling procedures. In: Kessler RC, Üstün TB (eds). *The WHO World Mental Health Surveys: global perspectives on the epidemiology of mental disorders*. New York: Cambridge University Press, 2008:14-32.
18. Kessler RC, Üstün TB. The World Mental Health (WMH) Survey Initiative Version of the World Health Organization (WHO) Composite International Diagnostic Interview (CIDI). *Int J Methods Psychiatr Res* 2004;13:93-121.
19. Haro JM, Arbabzadeh-Bouchez S, Brugha TS et al. Concordance of the Composite International Diagnostic Interview Version 3.0 (CIDI 3.0) with standardized clinical assessments in the WHO World Mental Health surveys. *Int J Methods Psychiatr Res* 2006;15:167-80.
20. First MB, Spitzer RL, Gibbon M et al. *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, 2002.
21. Knäuper B, Cannell CF, Schwarz N et al. Improving accuracy of major depression age-of-onset reports in the US National Comorbidity Survey. *Int J Methods Psychiatr Res* 1999;8:39-48.
22. Hosmer DW, Lemeshow S. *Applied logistic regression*, 2nd ed. New York: Wiley, 2001.
23. Breiman L. Random forests. *Mach Learn* 2001;45:5-32.
24. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. *J Stat Softw* 2010;33:1-22.
25. Berk RA. *Statistical learning from a regression perspective*. New York: Springer, 2008.
26. Liaw A, Wiener M. Classification and regression by randomForest. *R News* 2002;2:18-22.
27. Thernau T, Atkinson B, Ripley B. Rpart: recursive partitioning. *R Package* 4.1-0., 2012.
28. van der Laan MJ, Polley EC, Hubbard AE. Super learner. *Stat Appl Genet Mol Biol* 2007;6:Article 25.
29. van der Laan MJ, Rose S. *Targeted learning: causal inference for observational and experimental data*. New York: Springer, 2011.
30. Polley EC, van der Laan MJ. SuperLearner: Super Learner prediction, Package Version 2.0-4. Vienna: R Foundation for Statistical Computing, 2011.
31. Harvey AG, Bryant RA. Memory for acute stress disorder symptoms: a two-year prospective study. *J Nerv Ment Dis* 2000;188:602-7.
32. Roemer L, Litz BT, Orsillo SM et al. Increases in retrospective accounts of war-zone exposure over time: the role of PTSD symptom severity. *J Trauma Stress* 1998;11:597-605.
33. Zoellner LA, Foa EB, Brigidi BD et al. Are trauma victims susceptible to "false memories"? *J Abnorm Psychol* 2000;109:517-24.

DOI 10.1002/wps.20150