Published in final edited form as:

J Am Stat Assoc. 2010 December ; 105(492): 1323–1332. doi:10.1198/jasa.2010.ap09185.

Using a short screening scale for small-area estimation of mental illness prevalence for schools

Fan Li and

Department of Statistical Science, Duke University

Alan M. Zaslavsky

Department of Health Care Policy, Harvard Medical School

Fan Li: fli@stat.duke.edu; Alan M. Zaslavsky: zaslavsk@hcp.med.harvard.edu

Abstract

We use data collected in the National Comorbidity Survey - Adolescent (NCS-A) to develop a methodology to estimate the small-area prevalence of serious emotional distress (SED) in schools in the United States, exploiting the clustering of the main NCS-A sample by school. The NCS-A instrument includes both a short screening scale, the K6, and extensive diagnostic assessments of the individual disorders and associated impairment that determine the diagnosis of SED. We fitted a Bayesian bivariate multilevel regression model with correlated effects for the probability of SED and a modified K6 score at the individual and school levels. Our results provide evidence for the existence of variation in the prevalence of SED across schools and geographical regions. Although the concordance between the modified K6 scale and SED is only modest for individuals, the school-level random effects for the two measures are strongly correlated. Under this model we obtain a prediction equation for the rate of SED based on the mean K6 score and covariates. This finding supports the feasibility of using short screening scales like the K6 as an alternative to more comprehensive lay assessments in estimating school-level rates of SED. These methods may be applicable to other studies aiming at small-area estimation for geographical units.

Keywords

Bayesian; bivariate; hierarchical models; National Comorbidity Survey; prediction; serious emotional distress; small-area estimation; survey

1. Introduction

In the United States, mental disorders collectively account for more than 15 percent of the overall burden of disease from all causes, slightly more than the burden associated with all forms of cancer (Murray and Lopez 1996). An important reason for the significance of mental disorders is that they have a dramatically earlier age-of-onset distribution than other commonly occurring and seriously impairing chronic conditions. The importance of early onset is leading health policymakers and epidemiologists to focus on the prevalence and correlates of these disorders among children and adolescents to facilitate intervention for prevention and early treatment. If the prevalence of mental illness varies significantly across schools and geographical regions, providing reliable small area (e.g., state, county, or school) estimates of prevalence will have crucial value in mental health treatment planning

and resource allocation. Since clinical diagnosis of mental disorders is very expensive, it is worthwhile to study the feasibility of predicting the small-area prevalences of mental disorders from various short screening scales.

The National Comorbidity Survey - Adolescent (NCS-A) is the first nationally representative general population survey in the U.S. to evaluate the mental health of adolescents using a fully structured research diagnostic interview (Kessler et al. 2009a; Kessler et al. 2009b). The NCS-A instrument includes both extensive diagnostic assessments of the individual disorders and associated mental distress that determine the diagnosis of the serious emotional distress (SED), and a short (6-item) screening scale, the K6 (Kessler and Mroczek 1994). The NCS-A provides a unique opportunity for developing a methodology for bivariate small-area estimation of SED prevalence from the K6, which is highly desirable when small-area data include only short screening scales but not a full diagnostic interview. Such data arise from a number of other national surveys of adults, such as the CDC's Behavioral Risk Factor Surveillance System (BRFSS) and the National Health Interview Survey (NHIS), and could be collected by schools in adolescent health screening surveys.

"Small-area estimation" (SAE) refers to any method that produces estimates for domains for which sample sizes are inadequate to produce direct estimates of adequate precision. Comprehensive reviews of SAE models can be found in Ghosh and Rao (1994), Rao (2003), and Jiang and Lahiri (2006). Bayesian SAE methods are reviewed in Ghosh and Natarajan (1999) and Nandram (2000). In SAE, the choice of models depends on the nature of the data available, particularly the level of detail at which the target and auxiliary variables (covariates) are measured, and the theoretically and empirically justified forms of relationships among the variables. Univariate SAE models have been well developed (e.g., Fay and Herriot 1979; Fuller and Harter 1987; Ghosh et al. 1998; among others). Multivariate models have been also widely studied and applied (e.g., DeSouza 1992; Datta et al. 1996; Datta et al. 1998; Raghunathan et al. 2007).

In this report, we propose a methodology for bivariate SAE and apply it to the NCS-A to provide general prediction rules for the prevalence of SED from the K6 short screening scale in the NCS-A and similar mental health surveys. Most of the existing multivariate SAE methods deal with the situation where all outcomes are observed in each domain. In our study, however, only one outcome (the K6) is observed outside a relatively small calibration sample. Instead, using a model estimated from the calibration sample, we predict the small area quantities of another (missing) outcome (SED) from the observed one. In Section 2 we introduce statistical models for multivariate SAE for both continuous and binary outcomes. We analyze the NCS-A data and obtain prediction equations for SED based on K6 scores and covariates in Section 3. In Section 4 we evaluate the results through simulations. Section 5 concludes with a discussion.

2. Models and estimation

2.1. Models

We first describe models for bivariate hierarchical data with a two-level structure, where individual (i, j) ($j = 1, \dots, J_i$) belongs to cluster i ($i = 1, \dots, I$). Each individual has two continuous outcomes Y_{ijm} ($m = 1, 2$), with corresponding vectors of covariates (both individual-level and cluster-level), $\mathbf{X}_{ijm}$. The continuous outcome Y_{ij2} might be observed, or might be latent but dichotomized to generate an observed binary outcome Y_{ij3} . Such binary measurements occur commonly in epidemiology and other fields, arising either from dichotomizing a measured continuous scale or from an algorithm or observation procedure that is not based on observable scale, such as a diagnostic algorithm requiring a particular combination of criteria to be satisfied or based on clinical judgement.

For two continuous outcomes (Y_1, Y_2) , we assume the two-level bivariate random effects model

$$Y_{ijm} = \mathbf{X}_{ijm} \boldsymbol{\beta}_m + v_{im} + e_{ijm}, \quad \text{for } m=1, 2 \quad (1)$$

with random effects $(v_{i1}, v_{i2})' \sim N(0, \boldsymbol{\Sigma}_v)$, $(e_{ij1}, e_{ij2})' \sim N(0, \boldsymbol{\Sigma}_e)$, and

$$\boldsymbol{\Sigma}_e = \begin{pmatrix} \sigma_{e1}^2 & \rho_e \sigma_{e1} \sigma_{e2} \\ \rho_e \sigma_{e1} \sigma_{e2} & \sigma_{e2}^2 \end{pmatrix}, \quad \boldsymbol{\Sigma}_v = \begin{pmatrix} \sigma_{v1}^2 & \rho_v \sigma_{v1} \sigma_{v2} \\ \rho_v \sigma_{v1} \sigma_{v2} & \sigma_{v2}^2 \end{pmatrix}.$$

Parameters ρ_e, ρ_v are the correlations between the error terms of the two outcomes, which might be related to different sets of covariates. Model (1) is a natural extension of the univariate nested error regression model of Fuller and Harter (1987), applied by Battese, Harter, and Fuller (1988) to estimate county crop areas.

For a continuous outcome Y_1 and a binary outcome Y_3 , in addition to model (1), we assume that

$$Y_{ij3} = I(Y_{ij2} > 0), \quad (2)$$

and fix $\sigma_{e2}^2 = 1$. In this case Y_2 is merely a latent variable and not an observable outcome as in model (1). The above formulation is equivalent to a generalized linear mixed model (Breslow and Clayton 1993) for Y_{ij3} with probit link,

$$Y_{ij3} \sim \text{Bernoulli}(\pi_{ij3}), \text{ with } \Phi^{-1}(\pi_{ij3}) = \mathbf{X}_{ij2} \boldsymbol{\beta}_2 + v_{i2}.$$

Other link functions, such as the logistic link, can be assumed. We choose the probit link mainly for its computational simplicity since the corresponding latent variable model is a hierarchical normal linear model like (1).

In our analysis of the NCS-A, the continuous measure Y_1 is a modified K6 score (described in Sec. 3.1). We consider two alternative versions of the diagnostic outcome (described further in Sec. 3.1). In one, the continuous outcome Y_2 is the probit transformation of the probability of SED predicted by a set of diagnostic measurements in a probit model derived in a separate analysis. The probit transformation of probability of SED, i.e., the linear predictor of SED in the probit model, fits model (1) better than the untransformed probability, because its distribution is less skewed. In this case the small-area inferences of interest involve means of the transformed value $\Phi(Y_2)$, which are predicted rate of SED. A second analysis dichotomizes the predicted probability of SED to obtain a binary outcome Y_3 .

The two-level models can be easily extended to incorporate higher multilevel structure, e.g., a third level for the sampling strata in the NCS-A. Model (1), for instance, can be extended to a three-level random effects model by adding a third-level random effect u ,

$$Y_{ijkm} = \mathbf{X}_{ijkm} \boldsymbol{\beta}_m + u_{im} + v_{ijm} + e_{ijkm}, \quad \text{for } m=1, 2 \quad (3)$$

where $(u_{i1}, u_{i2})' \sim N(0, \boldsymbol{\Sigma}_u)$ and

$$\boldsymbol{\Sigma}_u = \begin{pmatrix} \sigma_{u1}^2 & \rho_u \sigma_{u1} \sigma_{u2} \\ \rho_u \sigma_{u1} \sigma_{u2} & \sigma_{u2}^2 \end{pmatrix}.$$

2.2. Model fitting

Estimation approaches for generalized linear mixed models like (1) and (2) include empirical Bayes (EB) (Morris 1983), empirical best linear unbiased prediction (EBLUP) (Harville and Jeske 1992) and hierarchical Bayes (HB) (Gelfand and Smith 1990; Albert and Chib 1993). We have selected the latter approach, which enables us to separate the model-fitting and prediction steps and account for parameter uncertainty in the latter through simulation using draws from the former.

In the HB approach, a prior distribution is specified for each model parameter, $\theta = (\boldsymbol{\beta}, \boldsymbol{\Sigma}_e, \boldsymbol{\Sigma}_v)$, and draws from the posterior distributions of these parameters are used to estimate posterior distributions, summarized by posterior means and variances. For fixed effects $\boldsymbol{\beta}$, we assume independent uninformative (improper) uniform priors: $\boldsymbol{\beta} \propto 1$. For the covariance matrices $\boldsymbol{\Sigma}$ at both individual and cluster level, we assume the conventional inverse Wishart prior: $\boldsymbol{\Sigma}^{-1} \sim \text{Wishart}(b_0, b_0 \boldsymbol{\Sigma}_0^{-1})$, where b_0 is the prior sample size and $\boldsymbol{\Sigma}_0$ is the prior covariance matrix. Although the Wishart prior is informative, we can express prior ignorance by setting $b_0 = 3$ and $\boldsymbol{\Sigma}_0$ to a diagonal matrix, which yields a prior under which the marginal distributions of the correlations are uniform on $[-1, 1]$ (Barnard et al. 2000). We set diagonals to the maximum likelihood estimates (MLE) of the corresponding variances, thus setting the MLE of $\boldsymbol{\Sigma}$ at the center of a weakly informative prior. Sensitivity analyses in Section 4.1 show that this specification yields a robust and conservative estimate of the correlation. Markov chain Monte Carlo (MCMC) is then used to sample from the joint

posterior distribution of the parameters. Computation of bivariate models with outcomes of different types as in (2) is greatly simplified by the probit link, which is available in the software we used, MLwiN 2.02 (Browne 2005). More details appear in Section 3.2.

2.3. Small-area estimation

2.3.1. Overview—Our objective is to calibrate an easily available measure for the prediction of cluster means of a related but usually missing measure. We estimate parameters of a model relating measure Y_1 to either Y_2 or Y_3 from a “calibration sample” of clusters in which we have both measures on samples of individuals. We then use the estimated model to predict small-area means involving Y_2 or its dichotomized version, e.g., $\sum_j \Phi(Y_{ij2})/J_i$ or $\sum_j Y_{ij3}/J_i$, in a distinct population of clusters in which we have data on Y_1 and auxiliary variables X_1, X_2 for a “survey sample” of individuals of each cluster and auxiliary variables for the “target population” of individuals in the cluster, but not Y_2 or Y_3 . Thus, our application differs from most others in which hierarchical models are used for small-area estimation in that even within the survey sample in which Y_1 is observed, prediction is still required for the quantity of interest Y_2 or Y_3 . Specifically, in our application we have data on both the short K6 scale and an extensive diagnostic interview to assess SED in the NCS-A. Using these data to estimate model parameters θ , we want to predict the school-level average SED prevalence, for schools in a distinct new sample where only the K6 is collected. We substitute estimates of θ and the auxiliary information Y_1, X_1, X_2 into formulas that are derived in this section to obtain these small-area estimates. Uncertainty about the model parameters is taken into account by combining the posterior draws of θ .

We consider two cases differing in the assumed relationship between the survey sample for which Y_1 is measured and the target population for which Y_2 or Y_3 must be predicted. In the first case, the sample is distinct from the target population, but within the same cluster. The goal is to predict Y_2 or Y_3 for the unsampled units. One reason for the practical importance of this case is that if the sample constitutes only a small part of the target population (as when a sample of 200 students is screened out of a school with 1500 enrolled), this “out-of-sample” prediction will be a good approximation to the analysis that correctly takes into account that some cases are “in sample” and some are not, yielding an upper bound on variance, analogous to ignoring the finite population correction in ordinary analysis of surveys. In the second case, the target population is equal to or a subset of the survey sample in the cluster; Y_1 is measured for every individual and we want to predict the other, missing measurement Y_2 for the same or a subset of individuals in the survey sample (“in-sample” prediction). In terms of the analytical models (1) and (2), in the out-of-sample prediction, one first estimates v_1 from outcome Y_1 in the sample, from which v_2 can be predicted. The individual errors e_2 of the target population, distinct from the sample, are independent of the predicted v_{i2} (Figure 1(a)). This independence no longer holds for in-sample prediction where correlation between Y_1 and Y_2 at both the individual and cluster levels must be considered (Figure 1(b)). Information on e_1 can improve the estimation of e_2 and thus of Y_2 .

2.3.2. In-sample small-area prediction—We first consider in-sample small-area prediction. Let J_{i1}, J_{i2} be the sizes of the survey sample and the target population in the cluster, respectively, which are not necessarily equal. Let $Y_{i,1} = \{Y_{ij1} : j = 1, \dots, J_{i1}\}, X_{i,m} =$

$\{\mathbf{X}_{ijm} : j = 1, \dots, J_{im}\}$, $\bar{Y}_{im} = \sum_j Y_{ijm}/J_{im}$, $\bar{\mathbf{X}}_{im} = \sum_j \mathbf{X}_{ijm}/J_{im}$, and the observed data $\mathbf{D}_i = \{Y_{i1}, X_{i1}, X_{i2}\}$. In-sample small-area prediction involves the conditional distribution of the sum of random effects, $t_{ijm} = v_{im} + e_{ijm}$, given data and parameters, for which the sufficient statistics are Y_{ij1} and the mean of other Y_1 's in the same cluster, $\bar{Y}_{i(-j)1} = \sum_{j' \neq j} Y_{ij'1}/(J_{i1} - 1)$. Under model (1), we can show (Appendix) that $f(t_{ij2}|\mathbf{D}_i, \theta) = N(\tilde{\mu}_{tij}, \tilde{\sigma}_{ti}^2)$, where

$$\tilde{\mu}_{tij} = c_{ti}(Y_{ij1} - \mathbf{X}_{ij1}\beta_1) + d_{ti}(\bar{Y}_{i(-j)1} - \bar{\mathbf{X}}_{i(-j)1}\beta_1), \quad (4)$$

$$\tilde{\sigma}_{ti}^2 = \sigma_{v2}^2 + \sigma_{e2}^2 - \frac{(\rho_v \sigma_{v1} \sigma_{v2} + \rho_e \sigma_{e1} \sigma_{e2})^2 + (J_{i1} - 1)(\rho_v^2 \sigma_{v2}^2 + \rho_e^2 \sigma_{e2}^2) \sigma_{v1}^2}{J_{i1} \sigma_{v1}^2 + \sigma_{e1}^2}. \quad (5)$$

with $c_{ti} = \{\rho_v \sigma_{v1} \sigma_{v2} + \rho_e \sigma_{e1} \sigma_{e2} + (J_{i1} - 1) \rho_e \sigma_{v1}^2 \sigma_{e2} / \sigma_{e1}\} / (J_{i1} \sigma_{v1}^2 + \sigma_{e1}^2)$ and $d_{ti} = (J_{i1} \rho_v \sigma_{v1} \sigma_{v2} + \rho_e \sigma_{e1} \sigma_{e2}) / (J_{i1} \sigma_{v1}^2 + \sigma_{e1}^2) - c_{ti}$.

It is useful to write $t_{ijm} = s_{im} + w_{ijm}$, where $s_{im} = v_{im} + \sum_j e_{ijm}/J_{im} = v_{im} + \bar{e}_{im}$ is the second-level mean random effect and $w_{ijm} = e_{ijm} - \bar{e}_{im}$ is the first-level residual random effect. Then s and w are conditionally independent, $w_{ijm} \perp s_{im} | (\mathbf{D}_i, \theta)$, and we can show (Appendix) that $f(s_{i2}|\mathbf{D}_i, \theta) = f(s_{i2}|\bar{Y}_{i1} - \bar{\mathbf{X}}_{i1}\beta_1, \theta) = N(\tilde{\mu}_{si}, \tilde{\sigma}_{si}^2)$, where

$$\tilde{\mu}_{si} = c_{si}(\bar{Y}_{i1} - \bar{\mathbf{X}}_{i1}\beta_1), \quad \tilde{\sigma}_{si}^2 = \sigma_{v2}^2 + \frac{\sigma_{e2}^2}{J_{i2}} - \frac{(\rho_v \sigma_{v1} \sigma_{v2} + \rho_e \sigma_{e1} \sigma_{e2} / \sqrt{J_{i1} J_{i2}})^2}{\sigma_{v1}^2 + \sigma_{e1}^2 / J_{i1}}, \quad (6)$$

with $c_{si} = c_{ti} + d_{ti} = (\rho_v \sigma_{v1} \sigma_{v2} + \rho_e \sigma_{e1} \sigma_{e2} / \sqrt{J_{i1} J_{i2}}) / (\sigma_{v1}^2 + \sigma_{e1}^2 / J_{i1})$, which we call the “prediction coefficient”. The conditional mean μ_{si} can be interpreted as the regression of s_2 on s_1 estimated by residuals, with shrinkage for estimation of s_1 . Then $\sum_j \mu_{tij} / J_{i1} = \mu_{si}$ and $f(w_{ij2}|\mathbf{D}_i, \theta) = N(\tilde{\mu}_{wij}, \tilde{\sigma}_{wi}^2)$, where $\mu_{wij} = \mu_{tij} - \mu_{si}$ and $\tilde{\sigma}_{wi}^2 = \tilde{\sigma}_{ti}^2 + \tilde{\sigma}_{si}^2 - 2\sigma_{v2}^2 - 2\sigma_{e2}^2 / J_{i2}$.

The individual-level small-area prediction for the continuous outcome Y_{ij2} is

$\hat{Y}_{ij2}^{\text{in}} = E_\theta E_t(Y_{ij2}|\mathbf{D}_i, \theta) = E_\theta(\mathbf{X}_{ij2}\beta_2 + \tilde{\mu}_{tij})$. The predicted second-level mean, $\bar{Y}_{i2}$, is:

$$\hat{\bar{Y}}_{i2}^{\text{in}} = E_\theta E_s(\bar{Y}_{i2}|\mathbf{D}_i, \theta) = E_\theta E_s(\bar{\mathbf{X}}_{i2}\beta_2 + s_{i2}|\mathbf{D}_i, \theta) = E_\theta(\bar{\mathbf{X}}_{i2}\beta_2 + \tilde{\mu}_{si}). \quad (7)$$

with predictive variance the sum of two terms:

$$\text{Var}(\bar{Y}_{i2}^{\text{in}}|\mathbf{D}_i) = E_\theta \text{Var}_s(\bar{Y}_{i2}^{\text{in}}|\mathbf{D}_i, \theta) + \text{Var}_\theta E_s(\bar{Y}_{i2}^{\text{in}}|\mathbf{D}_i, \theta). \quad (8)$$

Expectations with respect to θ can be estimated by simulation or approximated by substituting the posterior means of the parameters. The second term also can be estimated by simulation, calculating $E_s(\bar{Y}_{i2}^{\text{in}}|\mathbf{D}_i, \theta)$ for several random draws of θ and then the empirical variance. In our application, this term contributed little to the total variance, but we suggest simulations to check the magnitude of this term in other applications.

When the main interest is in the transformed outcome $\Phi(Y_{ij2})$, the formula can be simplified by the expression: $\Phi(Y_{ij2}) = \Pr(Y_{ij2} + z > 0)$, where $z \sim N(0, 1)$. Specifically, the small-area prediction of the individual-level $\Phi(Y_{ij2})$ can be derived as follows

$$E_{\theta} E_t [\Phi(Y_{ij2}) | \mathbf{D}_i, \theta] = E_{\theta} E_t [\Pr(\mathbf{X}_{ij2}\beta_2 + t_{ij2} + z > 0 | \mathbf{D}_i, \theta)] = E_{\theta} \Phi \left(\frac{\mathbf{X}_{ij2}\beta_2 + \tilde{\mu}_{tij}}{\sqrt{1 + \tilde{\sigma}_{ti}^2}} \right). \quad (9)$$

The last equation follows the standardization of the conditional normal distribution $t_{ij2} + z$. The second-level mean, $P_{i2} = \Sigma \Phi(Y_{ij2}) = J_{i2}$, can be obtained by averaging the individual estimates in the cluster, or be approximated by integrating over the normal approximation to the population distribution of the individual-level covariates $\mathbf{X}_{ij2}$ and μ_{tij} in cluster i , i.e., assuming $\mathbf{X}_{ij2}\beta_2 \sim N(\bar{\mathbf{X}}_{i2}\beta_2, \sigma_{xi}^2)$ and $\tilde{\mu}_{tij} \sim N(\tilde{\mu}_{si}, \tilde{\sigma}_{\mu i}^2)$, from which

$$\hat{P}_{i2}^{\text{in}} = E_{\mathbf{X}, \mu_t} [\hat{\Phi}^{\text{in}}(Y_{ij2})] \approx E_{\theta} \Phi \left(\frac{\bar{\mathbf{X}}_{i2}\beta_2 + \tilde{\mu}_{si}}{\sqrt{1 + \tilde{\sigma}_{ti}^2 + \sigma_{xi}^2 + \tilde{\sigma}_{\mu i}^2}} \right), \quad (10)$$

where $\sigma_{xi}^2, \tilde{\sigma}_{\mu i}^2$ can be estimated by their empirical counterpart. Although the sample distribution of $\mathbf{X}_{ij2}\beta_2$ is often not exactly normal, simulations suggest that approximation (10) is reasonably accurate even for a non-normal linear predictor as long as $\sigma_{xi}^2 < 0.5$.

The predictive variance of $\hat{P}_{i2}^{\text{in}}$ can be approximated by a first order Taylor expansion. Due to the curvature of Φ in the tails, such an approximation, however, is accurate only for small values of $\tilde{\sigma}_{si}^2 (< 0.1)$. A second order Taylor expansion might improve the estimation, but involves an overly complex formula that lacks intuitive interpretation. Instead, a simple alternative based on simulation gives satisfactory results in applications. To obtain predictive intervals, we randomly draw one $\tilde{s}$ from $f(s_{i2} | \mathbf{D}_i, \theta)$ and $\tilde{w}$'s from $f(w_{ij2} | \mathbf{D}_i, \theta)$ for each cluster. The individual linear prediction is then $\hat{Y}_{ij2}^{\text{in}} = \mathbf{X}_{ij2}\beta_2 + \tilde{s} + \tilde{w}$, from which the simulated transformed outcomes $\Phi(Y_{ij2})$ or the second-level means can be calculated. Repeating this N (e.g. 1000) times, the posterior variance and predictive intervals are given by the empirical variance, and (2.5%, 97.5%) quantiles of the N draws.

Prediction of the binary outcome Y_{ij3} from Y_{ij1} and the parameters θ of model (2) must take into account the nonlinearity of the probit link function. Using the same trick as in equation (9), we have

$$\hat{P}_{ij3}^{\text{in}} = E_{\theta} E_t (Y_{ij3} | \mathbf{D}_i, \theta) = E_{\theta} [\Pr(\mathbf{X}_{ij2}\beta_2 + t_{ij2} > 0 | \mathbf{D}_i, \theta)] = E_{\theta} \Phi \left(\frac{\mathbf{X}_{ij2}\beta_2 + \tilde{\mu}_{tij}}{\tilde{\sigma}_{ti}} \right), \quad (11)$$

with $e_{ij2} \sim N(0, 1)$. Analogous to the transformed continuous outcome, the second-level mean, $\bar{Y}_{i3} = \Sigma_j Y_{ij3} / J_{i2}$, can be approximated by either randomly drawing J_{i2} outcomes with Bernoulli probabilities $\hat{P}_{ij3}^{\text{in}}$ and averaging them, or by integrating over the normal

approximation to the population distribution of the individual-level covariates $\mathbf{X}_{ij2}$ and μ_{ij} in cluster i :

$$\hat{Y}_{i3}^{\text{in}} = E_X(\hat{P}_{ij3}^{\text{in}}) \approx E_\theta \Phi \left(\frac{\bar{\mathbf{X}}_{i2} \beta_2 + \tilde{\mu}_{si}}{\sqrt{\tilde{\sigma}_{ti}^2 + \sigma_{xi}^2 + \tilde{\sigma}_{\mu i}^2}} \right). \quad (12)$$

The predictive interval of $\bar{Y}_{i3}$ can be obtained using the same simulation procedure as in the continuous case.

A measure of the accuracy of the small-area estimation procedure is the “reliability”, defined as the prior-to-posterior proportional reduction in variance of s_{i2} , $\zeta = 1 - \text{Var}(s_{i2} | \mathbf{D}_i, \theta) / \text{Var}(s_{i2} | \theta)$, where $\zeta = 1$ means perfect prediction and $\zeta = 0$ means that the small-area data for Y_1 is completely uninformative. From (6),

$$\zeta_i^{\text{in}} = \frac{[\rho_v + \rho_e \sigma_{e1} \sigma_{e2} / (J_{i1} \sigma_{v1} \sigma_{v2})]^2}{[1 + \sigma_{e1}^2 / (J_{i1} \sigma_{v1}^2)] [1 + \sigma_{e2}^2 / (J_{i2} \sigma_{v2}^2)]}. \quad (13)$$

High correlations of both first-level and second-level random effects between the two outcomes, large observed sample size and large target population size all contribute to high reliability, approaching ρ_v^2 as $J_{i1}, J_{i2} \rightarrow \infty$. When the ratios of variances of second-level to first-level random effects for the two outcomes are similar, $\sigma_{v2}^2 / \sigma_{e2}^2 \approx \sigma_{v1}^2 / \sigma_{e1}^2$, larger ratios lead to greater reliability. Otherwise, the relation between the ratios and the reliability is a complex function of $\rho_e, \rho_v, J_{i1}, J_{i2}$.

Formulas for bivariate prediction and reliability for three-level models can be derived by analogous arguments.

An important special case of the in-sample prediction is for individuals ($J_{i2} = 1$), as when responses to the screening scale are used to identify individuals with high predicted Y_2 for further evaluation. The mean score in the cluster affects the prediction for an individual through the second term of (4), reflecting the information they contain about the random effect v_{i2} . Hence two individuals with identical scores Y_{ij1} might have differing predictions when the mean levels in their clusters are taken into account.

2.3.3. Out-of-sample small-area prediction—Formulae for out-of-sample small-area prediction can be readily derived from those for in-sample prediction. In this case, random effects e_{ij1} and $e_{ij'2}$ are independent for (i, j) in the survey sample and (i, j') in the target population, since the two are disjoint. This independence can be modeled by setting $\rho_e = 0$ in (1). The key to the out-of-sample analysis is the conditional distribution

$v_{i2} | (v_{i1}, \theta) \sim N(\tilde{\mu}_{vi}, \tilde{\sigma}_{vi}^2)$, where

$$\tilde{\mu}_{vi} = c_{vi} (\bar{Y}_{i1} - \bar{\mathbf{X}}_{i1} \beta_1), \quad \tilde{\sigma}_{vi}^2 = \sigma_{v2}^2 - \frac{\rho_v^2 \sigma_{v1}^2 \sigma_{v2}^2}{\sigma_{v1}^2 + \sigma_{e1}^2 / J_{i1}}, \quad (14)$$

and $c_{vi} = \rho_v \sigma_{v1} \sigma_{v2} / (\sigma_{v1}^2 + \sigma_{e1}^2 / J_{i1})$. When $\rho_e = 0$, we have $c_{vi} = c_{si}$ and $\tilde{\sigma}_{ti}^2 = \tilde{\sigma}_{vi}^2 + \sigma_{e2}^2$. In fact, the discussion in the previous section applies directly to out-of-sample prediction, where one replaces the components of (s, w) by the corresponding ones of (v, e) under $\rho_e = 0$.

Specifically, replacing $\tilde{\mu}_{tij}$ and $\tilde{\mu}_{si}$ by $\tilde{\mu}_{vi}$, $\tilde{\sigma}_{wi}^2$ by σ_{e2}^2 , and $\tilde{\sigma}_{\mu i}^2$ by 0 in formulas (7)–(12), we obtain the corresponding out-of-sample small-area prediction formulas. Reliability is obtained by setting $\rho_e = 0$ in (13) and is always less than that for the in-sample prediction as long as $\rho_v, \rho_e > 0$.

3. Application to NCS-A data

3.1. Data

Between August 2001 and August 2002, the NCS-A administered face-to-face diagnostic interviews to 10,148 adolescents in the age range 13 to 18 years. The survey consists of two distinct samples: 9,244 students at a sample of schools and 904 adolescents in households that participated in a preceding survey of adults. Since our objective is small-area estimation for schools, we only use the data from the school sample. A stratified sample of two primary sampling units (PSUs) per geographical stratum was drawn, and then a probability-proportional-to-size sample of schools in each sampled PSU; non-cooperating sampled schools were replaced with similar substitutes. Finally, a sample of students in the eligible age range, stratified by age, was drawn from each participating school. We removed data from schools with fewer than 10 students, giving a sample of 9,022 students, drawn from 282 schools grouped into 42 geographically-defined sampling strata, which constitutes a three-level hierarchical structure (Kessler et al. 2009a; Kessler et al. 2009b).

The diagnostic instrument used in the NCS-A is a modification of the World Health Organization's Composite International Diagnostic Interview (CIDI; World Health Organization, 1997) appropriate for adolescents (CIDI-A) (Merikangas et al. 2008). The definition of SED includes serious impairment, defined as a score on the Clinical Global Assessment of Functioning (CGAF) below 50. Because the CGAF requires administration by a clinical interviewer rather than the lay interviewers of the NCS-A, it was not included in the CIDI-A instrument but was administered in a smaller validation sample ($n = 347$). We then fitted a probit model for the probability of SED as a function of variables available in the CIDI-A instrument, including diagnoses, disorder-specific impairment items, and personality scales. Details of this model are described elsewhere (Green et al. 2009).

The NCS-A instrument also included a short screening scale, the K6 (Kessler and Mroczek 1994), consisting of six short questions on mental health during the past 30 days: "how often did you feel (1) nervous; (2) hopeless; (3) restless or fidgety; (4) so depressed that nothing could cheer you up; (5) everything was an effort; (6) worthless?" Each item uses a frequency response scale coded from 0 (never) to 4 (all the time). The standard K6 score is the sum of the six items. In this study we used a modified K6 score that supplements the standard six-item K6 with five additional items that improve its sensitivity (at the individual level) for detection of externalizing disorders that are often missed by the standard K6 (Green et al. 2009); we henceforth refer to this modified scale simply as the K6. We explored unequally-weighted scores and quadratic functions of this score and found that they did not improve on

the summed scale as a predictor of SED. This summary of mental status is easy to obtain, and corresponds to outcome Y_1 in our analysis. The outcomes Y_2 and Y_3 correspond to measures of SED based on the full CIDI-A instrument. Because the CIDI-A instrument only generates a predicted probability of SED rather than a strict determination, as described above, we illustrate two approaches to modeling SED. In one we define Y_2 as the linear predictor in a probit model for probability of SED, and thus $\Phi(Y_2)$ is the continuous SED score, as mentioned earlier. In the second, we defined a binary variable Y_3 by dichotomizing the predicted probabilities at a cutpoint chosen to make SED prevalence match that obtained by clinical assessments in the clinical validation sample, a common (but not necessarily efficient) procedure in psychiatry epidemiology.

The NCS-A collects a large set of covariates at both the individual and school levels. From preliminary analysis, we selected a set of covariates that are most predictive of the outcomes. The selected individual-level predictors (coded as dummy variables) include age (13–18 years), sex, race (white, black, Hispanic, other) and school entrance age (6, 7, > 7 years); the school-level predictors include school size (< 50 or ≥ 50 teachers) and the public/private school indicator. No interaction term was found to be a significant predictor of SED. Other covariates tested were student nationality and religion and the school's urbanicity, average student age, race proportion, and an "education climate" variable summarizing neighborhood socioeconomic characteristics. These variables were excluded because they were not statistically significant predictors or are hard to obtain in routine school-based surveys.

3.2. Results of model-fitting

Using the hierarchical Bayes approach, we fitted the two-level bivariate models (1) and (2) to the NCS-A school sample, ignoring the third-level clustering by sampling strata. We used MLwiN 2.02 to run 5 chains of hybrid adaptive Gibbs sampling (to estimate variance components) and Metropolis-Hastings sampling (to estimate regression coefficients). Each chain has different starting values and all have 10,000 burn-in and 10,000 iterations. Convergence is checked by the Gelman-Rubin statistic (Gelman and Rubin 1992). Table 1 shows the posterior means of coefficients from the models with a continuous outcome (columns 1, 2) and with a dichotomized outcome (columns 3, 4). As expected, the coefficients for the K6 score, $\hat{\beta}_1$, are very similar across models. The signs of the coefficients for the K6 and the two SED outcomes are the same. The coefficients for the binary SED are roughly 1.5 to 2 times those for continuous SED for most covariates. This concordance is plausible since the K6, the linear SED score and the dichotomized SED all measure mental status, with higher score indicating worse status. We also fitted the corresponding three-level bivariate models with the sampling strata as the third level. The coefficient estimates are similar to those of two-level models and therefore are not shown.

Estimates of variance components from the two-level models (Table 2) suggest that although concordance between the K6 and SED scores is only modest for individuals ($\rho_e = 0.544$ and 0.453 for models (1) and (2), respectively), the school-level random effects for the two measures are strongly correlated ($\rho_v = 0.845$ and 0.754). The lower correlation in the binary model is due to the information loss in dichotomization. In three-level models,

stratum-level variance only accounts for a quarter of the total cluster (second- and third-) level variance in Y_2 . The estimated individual-level and school-level correlations are similar to those from two-level models. These findings support the feasibility of using short screening scales like the K6 as an alternative to more comprehensive lay assessments in estimating school-level rates of SED among adolescents.

3.3. Predictive models and predictions

In a school-wide screening, the target population is exactly the survey sample, and in the NCS-A the same covariates are used for all outcomes. Therefore, we simplify the presentation by letting $J_{i1} = J_{i2} = J_i$, $X_1 = X_2 = X$. From results in Section 2.3.2, the key component in the in-sample small-area prediction of cluster mean is $X_i\bar{\beta}_2 + \mu_{si} = X_i(\bar{\beta}_2 - c_{si}\beta_1) + c_{si}\bar{Y}_{i1}$. The prediction coefficients c_{si} depend on the cluster sample size J_i . Table 3 shows coefficients in the continuous model with four representative sample sizes: $J_i = 1$ (individual prediction), $J_i = 30$ (average school sample in the NCS-A), $J_i = 200$ (a realistic school size for screening) and $J_i = \infty$. In both models, c_{si} increases to $\rho_v\sigma_{v2}/\sigma_{v1}$ as J_i increases. With our data, prediction coefficients with $J_i = 200$ are close to those with $J_i = \infty$.

The dependence of the prediction coefficients on sample size explains why we performed a bivariate analysis instead of directly regressing Y_2 on Y_1 . The coefficient in our bivariate model increases with J_i , as reliability of estimation improves. The coefficient of $\bar{Y}_{i1}$ in a simple univariate regression model is the same regardless of J_i . In fact, the sample size of a school screening, potentially hundreds of adolescents, is much larger than the NCS-A sample in any school, so the correct coefficients c_{si} can only be obtained by extrapolation under our bivariate model, not by direct empirical regression modeling in any available dataset.

As discussed in Section 2.3.2, small-area estimation of the school-level random effects is also useful in screening of individuals. Consider two large schools, one with moderately low ($v_{i1} = -\sigma_{v1}$) and one with moderately high ($v_{i1} = \sigma_{v1}$) mean K6 scores. This implies a difference of $2\rho_v\sigma_{v2} = 0.32$ on the probit scale in their estimates of v_2 . Then two adolescents with the same moderately high K6 score and characteristics X might have predicted probabilities of SED of 6.7% and 12.3% respectively at the two schools. Conversely, if referral to followup evaluation were determined by predicted probability of SED after administering the K6 screener, we would set the K6 cutoff lower in a school with a higher mean K6 score.

3.4. Reliability

The reliability of small-area estimates defined in (13) for a given cluster sample size J_i (Table 4) is a measure of variance reduction. For out-of-sample prediction from a school sample of size 200, small-area data on the K6 and covariates reduces variance in predicting the linear and dichotomized SED scores by 60% and 54%, respectively. Reliability is consistently lower from the discrete than the continuous model due to the information loss in dichotomization. The reliability of in-sample prediction is larger than that of out-of-sample case, both converging to the upper bound ρ_v^2 .

4. Model Validation

4.1. Sensitivity to prior specification

We evaluated the validity of models with continuous and binary outcomes, here presenting results for the former, our preferred model. We first examine the sensitivity of parameter estimates to the prior specification for the covariance matrix $\sum_v^{-1} \sim \text{Wishart}(b_{0v}, b_{0v} \sum_{0v}^{-1})$, focusing in correlation. With only 282 schools, the prior distribution could substantially affect estimates of the second-level correlation. In our main analysis, we set the prior sample size $b_{0v} = 3$ and the scale matrix $\sum_{0v}$ to a diagonal matrix with MLEs of variance components on the diagonal. Setting $b_{0v} < 3$, $b_{0v} = 3$, $b_{0v} > 3$ yields marginal prior distributions of the correlations that are convex, uniform on $[-1, 1]$ and concave, respectively (Barnard et al. 2000). We fitted the same model with the $b_{0v} = 2, 3, 4$ and $(\sigma_{0v1}^2, \sigma_{0v2}^2)$ proportional to the MLE in the bivariate continuous model (1). We compare the corresponding posterior means of the correlation ρ_v in Table 5, as well as that with

$\sum_{0v} = \sum_v^{\text{mle}}$. With $\sum_{0v}$ fixed, ρ_v decreases as b_{0v} increases; and with b_{0v} fixed, ρ_v decreases as prior variances increase, as expected, because larger prior sample size b_{0v} means stronger prior belief in independence. Although centering the prior for variances at larger values of $(\sigma_{0v1}^2, \sigma_{0v2}^2)$ leads to larger posterior means of variances, estimates of covariance are similar across different priors. We also fitted the above settings with different prior correlations ($\rho_{0v} = 0, 0.5, 0.8, \rho^{\text{mle}}$). As expected, the independence prior ($\rho_{0v} = 0$) induces slightly smaller posterior estimates of correlation than with dependent priors, but all posterior means were between 0.77 and 0.91. Thus, even though the scale of the prior (but not the prior for correlation) is data-dependent, correlation estimates are robust over a wide range of weakly informative priors; findings were similar under the discrete model (2). Comparable analyses for Σ_e revealed minimal sensitivity to a similar range of priors, due to the much larger sample size at the individual level.

4.2. Posterior predictive checks

To check model fit, we obtained 1000 posterior draws of parameters (β, Σ) from the fitted model, and simulated models (1) and (2) with the same covariates to generate new outcomes Y . In effect, we simulated new subjects in new schools with the same sample sizes and individual and school covariates as in the NCS-A data. We compare summary statistics of the observed data to their simulated expectations in Table 6, and report posterior predictive p -values, defined here as the posterior predictive probability that the simulated statistic is smaller than its observed value (Gelman et al. 1996; also see Rao 2003 ch. 10.2.6). The posterior predictive distributions of the summary statistics cover the observed values, with p -values close to 0.5. Simulated correlations are slightly smaller than the observed values, not surprising given our conservative prior.

We next predicted SED prevalence from K6 in the schools in the NCS-A as if only K6 had been measured, and compared the predictions to the observed values by checking how often the observed prevalence for each school is covered by the 95% “in-sample” predictive interval. Because the observed SED school prevalence is highly skewed, we used the

simulation-based procedure suggested in Section 2.3.2 with $N = 1000$ to obtain the 95% intervals. The mean squared error (MSE) is 1.54×10^{-3} and the mean absolute error (MAE) is 0.023. The observed SED prevalence of 254 out of 282 (90.7%) schools falls into predictive intervals under the continuous model, with similar coverage in schools with smaller and larger samples, suggesting that these intervals are fairly well calibrated.

The predicted and observed distributions of school means are well matched except in the upper tail (observed prevalence > 0.10). The average observed school prevalence is 0.06 in the NCS-A and about 10% of the schools have prevalence > 0.10 , the average of which is 0.152. The predictive expectation of the mean prevalence at the schools with the highest 10% of simulated prevalences is only 0.113, implying some lack of predictive fit at the high end. Nonetheless, it might be more important to identify the schools with the highest prevalences as targets for intervention than to predict their exact prevalences. In the NCS-A data, 21 out of the 29 (72.4%) schools among the top 10% by predicted prevalence are also among the top 10% by observed prevalence, showing success in identifying outlier schools despite underestimation of the highest prevalences.

As another check on model adequacy, we compared direct and model-based estimates for aggregates of schools. In the NCS-A, the data are collected from 42 geographical strata, some of which have small sample sizes and large sampling variances. We collapse the strata into 14 larger geographical strata (details available upon request). The direct estimates are weighted averages of the observed school means of $\Phi(Y_2)$ within each strata and the weights are proportional to school sizes. In 10 out of 14 strata, the errors of the bivariate predictions fall within the 95% confidence intervals of the direct estimates. The same strata were identified as having the two highest observed and predicted prevalences, but those prevalences were underestimated; these results parallel those for the highest prevalence schools.

4.3. Comparison with alternative methods

To demonstrate the value of the various information sources synthesized in our bivariate model, we compared it to four alternative prediction approaches:

- A. A simple regression-synthetic model (Rao 2003, ch. 4.2): We first regress individual-level SED score on covariates without K6, and then calculate individual predictions and thus predict school means. This approach represents the standard in previous small-area estimation of mental health (e.g., Hudson 2010).
- B. A similar regression-synthetic model including K6 as a predictor.
- C. A univariate random-effects logistic model without K6 as a predictor.
- D. The same as (C) but including K6 as a predictor.

Both (C) and (D) implement the two-level univariate model

$$Y_{ij2} = Y_{ij1}\alpha + \mathbf{X}_{ij}\beta + v_i + e_{ij}, \quad v_i \sim N(0, \sigma_v^2), e_{ij} \sim N(0, \sigma_e^2) \quad (15)$$

(D) includes Y_1 but treats it like other covariates (Rao 2003, ch. 10). We fitted NCS-A data to model (15) assuming flat priors for σ_v^2, σ_e^2 , and obtained predictive draws of SED

prevalence by drawing parameters $(\alpha, \beta, \sigma_v, \sigma_e)$, simulating random effects v_i . The outcome $\hat{Y}_{ij2} = Y_{ij1}\alpha + X_{ij}\beta + v_i + \epsilon_{ij}$, using 1000 draws to estimate the posterior mean and 95% predictive interval for prevalence at each school.

Predictions from these models are compared with our bivariate model in Table 7. The models with K6 clearly outperform their counterparts without K6, showing the importance of small-area screening data. The random-effects models with K6 improve on the regression-synthetic models, demonstrating the value of the “shrinkage” of estimates from the those models, whether univariate or bivariate. The bivariate model has smaller MSE and MAE than the univariate model.

Sample sizes per school in the NCS-A are small, ranging from 10 to 45. An actual school-based screening exercise would obtain data from most of the enrolled students, typically hundreds of students. To compare the performance of the proposed bivariate method and the alternative univariate model under more realistic conditions, we simulated a dataset with a larger school sample size using parameters estimated from the NCS-A data. We also simulate schools with small sample sizes for comparison. We generated 800 schools, consisting of 4 sets of 200 schools with sample sizes $J = 1, 25, 200, 400$, respectively. We gave each simulated school the covariate distribution of a randomly chosen actual school (sample size larger than 30) by sampling $J \mathbf{X}$'s with replacement from that school. We drew a value of (β, Σ) from the posterior parameter distribution given the NCS-A data and substituted it and the simulated $\mathbf{X}$ into model (1) to generate the outcomes, SED and K6. To represent parameter uncertainty, we drew (β, Σ) (and outcomes) repeatedly from the same posterior distribution. All later inference is based on the sample combined across draws. We then applied the bivariate and univariate models to the simulated data to get the prediction of the SED prevalence in each school and the associated 95% predictive intervals, repeating the entire process 10 times to get an average prediction and interval coverage. The simulated school SED prevalence ranges from 0.014 to 0.143 with mean 0.053. The MSE's and MAE's are displayed in Table 8. As expected, the prediction errors and the length of the predictive intervals decreased significantly with increasing sample size. With the smallest sample sizes, the univariate model performs slightly better than the bivariate, but the bivariate model dominates with larger samples. With $J = 400$, the MSE of the bivariate model is 21% less than that of the univariate model (D). This matches our theoretical observation that the prediction coefficients of univariate model are the same regardless of sample size and thus this model fitted to the NCS-A does not predict as well as the bivariate model in more realistic large samples.

5. Discussion

Motivated by the NCS-A, we developed a methodology to estimate the small-area prevalence of serious emotional disturbance from a short screening scale, the K6, in schools in the United States. We modeled the joint distribution of SED and K6 scores in the NCS-A sample by a bivariate random effects model, from which we derived small-area estimation procedures. Under this model, predictive means and intervals for both continuous and binary outcomes can be calculated for schools at which the short screening scale, but not the full diagnostic interview, is administered to students. These bivariate methods can be useful for a

wide range of applications where a desired outcome is difficult or expensive to measure, but an alternative measure is more readily available.

A simpler approach, used in many previous small-area estimation applications, is to treat the cluster mean of the inexpensive source as a cluster-level covariate and fit a univariate random-effects model. This approach may not be adequate, however, when sample sizes per cluster are small enough so estimates of cluster means for the inexpensive source have substantial sampling variability relative to between-cluster variation. In that case, the regression coefficient will vary depending on sample size, and the bivariate model is required to determine model coefficients that would be used in applications with sample sizes very different from those of the original survey used for estimation of model parameters.

Two cases that differ in the relationship between the screening sample and the target population were explored in detail. Out-of-sample prediction assumes that the sample and target population are distinct. It yields conservative estimates of precision when there is a small overlap. In-sample prediction takes into account both the cluster- and individual-level relationships between the K6 and SED, and should be used when there is substantial overlap. Two important applications of this analysis are prediction for individuals, and small-area estimation when almost the entire population of the cluster is assessed as would be the case in a school screening survey. Estimates from our data suggest that even individual-level predictions of the probability of SED can be substantially improved by taking into account estimated school-level prevalence.

Dichotomization of the predicted probability of SED leads to information loss, leading to lower reliability and larger predictive error which is also reflected in worse convergence of the MCMC sampler. Even though dichotomization is commonplace in mental health diagnosis, we recommend continuous measurement of the probability of diagnosis whenever possible.

The NCS-A application has significant practical importance (Li et al. 2010). We are now developing a simple computerized tool for administration of the survey and calculation of predicted prevalence and individual probabilities. Coefficients for various sample sizes are simple functions of J_i that can be programmed into the tool.

Our models assumed normality of the random effects at both levels. Moderate deviations from normality should have little affect on model performance. Diagnostics applied to the NCS-A data suggested some lack of fit (longer than predicted tails) at the high end of the distribution. Our method approximately preserves the rank of prevalences in the upper tail, but underestimate their magnitudes. Thus the possible lack of normality might be more of an issue for estimating resources needed by schools than for determining which schools have the greatest needs. Alternative models might accommodate non-normality of the level-2 random effects using an explicit asymmetrical long-tailed bivariate distribution, by transformation of bivariate normal variables, or by a normal mixture or Dirichlet process approximation. Each approach would require new efficient computational tools since no analytical forms would be available for bivariate analysis.

Acknowledgments

We thank David Banks, Hal Stern, the associate editor, and two reviewers for constructive comments that helped to improve the manuscript significantly. We also thank Ronald C. Kessler and Jennifer Green for insights into the NCS-A data and school assessment of emotional distress, and Nancy Sampson and Michael Gruber for preparation of data. This research was funded by grant R01-MH66627 from the National Institute for Mental Health (NIMH) and by Duke University. The NCS-A is supported by NIMH (U01-MH60220) with supplemental support from the National Institute on Drug Abuse (NIDA), the Substance Abuse and Mental Health Services Administration (SAMHSA), the Robert Wood Johnson Foundation (RWJF; grant 044780), and the John W. Alden Trust. The views and opinions expressed in this report are those of the authors and should not be construed to represent the views of any of the sponsoring organizations, agencies, or U.S. Government. The NCS-A is carried out in conjunction with the World Health Organization World Mental Health (WMH) Survey Initiative. We thank the staff of the WMH Data Collection and Data Analysis Coordination Centres for assistance with instrumentation, fieldwork, and consultation on data analysis. The WMH Data Coordination Centres have received support from NIMH (R01-MH070884, R13-MH066849, R01-MH069864, R01-MH077883), NIDA (R01-DA016558), the Fogarty International Center of the National Institutes of Health (FIRCA R03-TW006481), the John D. and Catherine T. MacArthur Foundation, the Pfizer Foundation, and the Pan American Health Organization. The WMH Data Coordination Centres have also received unrestricted educational grants from Astra Zeneca, BristolMyersSquibb, Eli Lilly and Company, GlaxoSmithKline, Ortho-McNeil, Pfizer, Sanofi-Aventis, and Wyeth.

References

- Albert J, Chib S. Bayesian Analysis of Binary and Polychotomous Response Data. *Journal of the American Statistical Association*. 1993; 88:669–679.
- Barnard J, McCulloch R, Meng XL. Modeling Covariance Matrices in Terms of Standard Deviations and Correlations, with Application to Shrinkage. *Statistica Sinica*. 2000; 10:1281–1311.
- Battese GE, Harter RM, Fuller WA. An Error Component Model for Prediction of County Crop Areas Using Survey and Satellite Data. *Journal of the American Statistical Association*. 1988; 83:28–36.
- Breslow NE, Clayton DG. Approximate Inference in Generalized Linear Mixed Models. *Journal of the American Statistical Association*. 1993; 88:9–25.
- Browne, WJ. MCMC Estimation in MLwiN (Version 2.0). Bristol: University of Bristol, Centre for Multilevel Modelling; 2005.
- Datta G, Day B, Maiti T. Multivariate Bayesian Small Area Estimation: Application to Survey and Satellite Data. *Sankhya, Series A*. 1998; 60:1–19.
- Datta, G.; Ghosh, M.; Nangia, N.; Natarajan, K. Estimation of Median Income of Four-Person families: A Bayesian Approach. In: Berry, W.; Chaloner, K.; Geweke, J., editors. *Bayesian Analysis in Statistics and Econometrics*. Wiley; New York, NY: 1996. p. 129-140.
- DeSouza CM. An Appropriate Bivariate Bayesian Method for Analysing Small Frequencies. *Biometrics*. 1992; 48:1113–1130.
- Fay R, Herriot R. Estimation of Income from Small Places: An Application of James-Stein Procedures to Census Data. *Journal of the American Statistical Association*. 1979; 74:269–277.
- Fuller, WA.; Harter, R. The Multivariate Components of Variance Model for Small Area Estimation. In: Platek, R.; Rao, JNK.; Sarndal, C.; Singh, MP., editors. *Small Area Statistics*. Wiley; New York, NY: 1987. p. 103-123.
- Gelfand AE, Smith AFM. Sample-Based Approaches to Calculate Marginal Densities. *Journal of the American Statistical Association*. 1990; 85:872–885.
- Gelman A, Meng X-L, Stern H. Posterior predictive assessment of model fitness via realized discrepancies (with discussion). *Statistica Sinica*. 1996; 6:733–807.
- Gelman A, Rubin D. Inference from Iterative Simulation Using Multiple Sequences. *Statistical Science*. 1992; 7:457–472.
- Ghosh, M.; Natarajan, K. Small Area Estimation: A Bayesian Perspective. In: Ghosh, S.; Dekker, M., editors. *Multivariate Analysis, Design of Experiments and Survey Sampling*. John Wiley and Sons; New York, NY: 1999. p. 69-92.
- Ghosh M, Natarajan K, Stroud TWF, Carlin BP. Generalized Linear Models for Small Area Estimation. *Journal of the American Statistical Association*. 1998; 93:273–282.
- Ghosh M, Rao JNK. Small Area Estimation: An Appraisal. *Statistical Science*. 1994; 9(1):55–76.

- Green, JG.; Gruber, MJ.; Kessler, RC.; Sampson, NA.; Zaslavsky, AM. Technical report. Harvard University, Department of Health Care Policy; 2009. Improving the K6 Short Scale to Predict Serious Emotional Distress in Adolescents in the US.
- Harville DA, Jeske DR. Mean Squared Error of Estimation or Prediction Under a General Linear Model. *Journal of the American Statistical Association*. 1992; 87:724–731.
- Hudson C. Validation of a model for estimating state and local prevalence of serious mental illness. *International Journal of Methods in Psychiatric Research*. 2010; 18(4) In press.
- Jiang J, Lahiri P. Mixed model prediction and small area estimation. *Test*. 2006; 15:1–96.
- Kessler, R.; Mroczek, D. Final Versions of Our Non-specific Psychological Distress Scale. University of Michigan, Survey Research Centre of the Institute for Social Research; 1994.
- Kessler RC, Avenevoli S, Costello EJ, Green JG, Gruber MJ, Heeringa S, Merikangas KR, Pennell B-E, Sampson NA, Zaslavsky AM. Design and Field Procedures in the US National Comorbidity Survey Replication Adolescent Supplement (NCS-A). *International Journal of Methods of Psychiatric Research*. 2009a; 18:69–83.
- Kessler RC, Avenevoli S, Costello EJ, Green JG, Gruber MJ, Heeringa S, Merikangas KR, Pennell BE, Sampson NA, Zaslavsky AM. The National Comorbidity Survey Adolescent Supplement (NCS-A): II. Overview and Design. *Journal of the American Academy of Child and Adolescent Psychiatry*. 2009b; 48:380–385. [PubMed: 19242381]
- Li, F.; Green, J.; Kessler, R.; Zaslavsky, A. Technical report. Harvard University, Department of Health Care Policy; 2010. Estimating prevalence of serious emotional disturbance in schools using a brief screening scale.
- Merikangas KR, Avenevoli S, Costello EJ, Koretz D, Kessler RC. The National Comorbidity Survey Adolescent Supplement (NCS-A): I. Background and Measures. *Journal of the American Academy of Child and Adolescent Psychiatry*. 2008; 48:367–369. [PubMed: 19242382]
- Morris CA. Parametric Empirical Bayes Inferences: Theory and Applications. *Journal of the American Statistical Association*. 1983; 78:47–54.
- Murray, CJL.; Lopez, AD. The Global Burden of Disease: A Comprehensive Assessment of Mortality and Disability from Diseases, Injuries and Risk Factors in 1990 and Projected to 2020. Cambridge, MA: Harvard University Press; 1996.
- Nandram, B. Bayesian Generalized Linear Models for Inference about Small Areas. In: Dey, D.; Ghosh, SK.; Mallick, BK., editors. *Generalized Linear models — A Bayesian Perspective*. Boca Raton: CRC Press; 2000. p. 89-109.
- Raghunathan TE, Xie D, Schenker N, Parsons VL, Davis WW, Feuer EJ, Dodd KW. Combining Information from Two Surveys to Estimate County-level Prevalence Rates of Cancer Risk Factors and Screening. *Journal of the American Statistical Association*. 2007; 102:474–486.
- Rao, JNK. *Small Area Estimation*. John Wiley and Sons, Inc; Hoboken, NJ: 2003.

Appendix

To get $f(t_{ij2}|D_i, \theta)$, we first obtain the joint distribution of $(v_{i1} + e_{ij1}, v_{i1} + \bar{v}_{i(-j)1}, v_{i2} + e_{ij2})$:

$$\begin{pmatrix} v_{i1} + e_{ij1} \\ v_{i1} + \bar{v}_{i(-j)1} \\ v_{i2} + e_{ij2} \end{pmatrix} = N \left(\mathbf{0}, \begin{pmatrix} \sigma_{v1}^2 + \sigma_{e1}^2 & \sigma_{v1}^2 & \rho_v \sigma_{v1} \sigma_{v2} + \rho_e \sigma_{e1} \sigma_{e2} \\ \sigma_{v1}^2 & \sigma_{v1}^2 + \frac{\sigma_{e1}^2}{J_{i1} - 1} & \rho_v \sigma_{v1} \sigma_{v2} \\ \rho_v \sigma_{v1} \sigma_{v2} + \rho_e \sigma_{e1} \sigma_{e2} & \rho_v \sigma_{v1} \sigma_{v2} & \sigma_{v2}^2 + \sigma_{e2}^2 \end{pmatrix} \right).$$

By the standard results of multivariate normal distribution, the conditional distribution of $t_{ij2} = v_{i2} + e_{ij2}$ given the observed data on the first two variables, i.e., $Y_{ij1} - X_{ij1}\beta_1$ and $\bar{Y}_{i(-j)1} - X_{i(-j)1}\beta_1$, is normal with mean and variance given in equations (4), (5). Similarly for s_i , we first write down the joint distribution of (s_{i1}, s_{i2}) :

$$\begin{pmatrix} s_{i1} \\ s_{i2} \end{pmatrix} = N \left(\mathbf{0}, \begin{pmatrix} \sigma_{v1}^2 + \sigma_{e1}^2 / J_{i1} & \rho_v \sigma_{v1} \sigma_{v2} + \rho_e \sigma_{e1} \sigma_{e2} / \sqrt{J_{i1} J_{i2}} \\ \rho_v \sigma_{v1} \sigma_{v2} + \rho_e \sigma_{e1} \sigma_{e2} / \sqrt{J_{i1} J_{i2}} & \sigma_{v2}^2 + \sigma_{e2}^2 / J_{i2} \end{pmatrix} \right),$$

and then easily obtain the conditional distribution of $s_{i2} | (\mathbf{D}_i, \theta)$ as in (6).

Figure 1.
Dependence among outcomes and random effects

Table 1

Estimates of coefficients of two-level models with posterior standard deviations.

Covariate	K6 (Y_1) vs. linear SED (Y_2)		K6 (Y_1) vs. binary SED (Y_3)	
	β_1 (s.d.)	β_2 (s.d.)	β_1 (s.d.)	β_2 (s.d.)
intercept	-2.022 (.037)	-2.381 (.048)	-2.021 (.037)	-2.195 (.078)
sex (male)	0.058 (.013)	-0.108 (.017)	0.060 (.013)	-0.088 (.045)
age 14-year	0.051 (.022)	0.080 (.028)	0.055 (.022)	0.153 (.077)
age 15-year	0.070 (.024)	0.146 (.032)	0.074 (.024)	0.243 (.069)
age 16-year	0.059 (.025)	0.147 (.032)	0.065 (.024)	0.298 (.069)
age 17-year	0.077 (.026)	0.218 (.033)	0.086 (.025)	0.412 (.072)
age 18-year	0.042 (.034)	0.196 (.043)	0.051 (.033)	0.380 (.096)
black	0.060 (.021)	0.019 (.027)	0.060 (.021)	0.099 (.062)
Hispanic	0.110 (.020)	0.100 (.026)	0.109 (.021)	0.122 (.059)
other race	0.149 (.029)	0.090 (.037)	0.151 (.030)	0.137 (.094)
start school at 7	0.058 (.015)	0.062 (.019)	0.059 (.015)	0.143 (.048)
start school > 7	0.219 (.029)	0.192 (.037)	0.222 (.028)	0.208 (.089)
school size	0.071 (.022)	0.097 (.029)	0.068 (.023)	0.046 (.056)
public school	0.130 (.030)	0.151 (.040)	0.122 (.030)	0.227 (.068)

Table 2

Estimated variance components of two-level models.

	Y_1 vs. Y_2		Y_1 vs. Y_3	
	ν	e	ν	e
σ_1^2	0.019	0.371	0.019	0.371
σ_2^2	0.037	0.597	0.089	1
ρ	0.845	0.544	0.754	0.453

Table 3

In-sample prediction coefficients of models (1) estimated from the NCS-A.

$\beta_2 - c_{si}\beta_1$	$J_i = 1$	$J_i = 30$	$J_i = 200$	$J_i = \infty$
intercept	-0.937	-0.405	-0.118	-0.036
sex(male)	-0.149	-0.165	-0.173	-0.175
age 14-year	0.043	0.030	0.022	0.020
age 15-year	0.096	0.078	0.068	0.065
age 16-year	0.105	0.089	0.081	0.078
age 17-year	0.163	0.143	0.132	0.129
age 18-year	0.167	0.156	0.150	0.148
black	-0.024	-0.040	-0.049	-0.051
Hispanic	0.021	-0.008	-0.023	-0.028
other race	-0.016	-0.056	-0.077	-0.083
start school at 7	0.021	0.006	0.002	0.004
start school > 7	0.036	-0.022	-0.053	-0.061
school size	0.046	0.027	0.017	0.014
public school	0.058	0.024	0.006	0.000
c_{si}	0.714	0.977	1.119	1.161

Table 4

Reliability of small-area estimates (defined in eq. (13)) in NCS-A data.

J_i	Y_1 vs. Y_2		Y_1 vs. Y_3	
	ζ^{out}	ζ^{in}	ζ^{out}	ζ^{in}
1	0.00	0.31	0.22	0.00
30	0.28	0.54	0.42	0.25
200	0.60	0.67	0.54	0.49
∞	0.72	0.72	0.57	0.57

Table 5

Posterior means of level-2 correlation ρ_v under different Wishart priors in model (1).

Σ_{0v}	$b_{0v} = 2$	$b_{0v} = 3$	$b_{0v} = 4$
$\hat{\Sigma}^{\text{mle}}$	0.90	0.91	0.91
$\text{diag}(\hat{\sigma}_{11}^{\text{mle}}, \hat{\sigma}_{22}^{\text{mle}})$	0.86	0.85	0.84
$(1/3) \cdot \text{diag}(\hat{\sigma}_{11}^{\text{mle}}, \hat{\sigma}_{22}^{\text{mle}})$	0.90	0.89	0.88
$3 \cdot \text{diag}(\hat{\sigma}_{11}^{\text{mle}}, \hat{\sigma}_{22}^{\text{mle}})$	0.81	0.79	0.77

Table 6

Summary statistics from observed and simulated data. Posterior predictive p -values are shown in parenthesis.

summary statistics	K6		$\Phi^{-1}(\text{SED})$	
	obs	sim	obs	sim
Individual mean	-1.72	-1.73(0.41)	-2.08	-2.07(0.37)
Individual s.d.	0.63	0.63(0.47)	0.81	0.81(0.46)
Mean of school means	-1.72	-1.72(0.50)	-2.08	-2.07(0.58)
S.d. of school means	0.20	0.21(0.58)	0.27	0.26(0.88)
Individual correlation with K6	-	-	0.56	0.53(0.57)
School correlation with K6	-	-	0.79	0.60(0.85)

Table 7

Comparison of errors of prediction of school-level SED prevalences in NCS-A from different SAE models.

Model	MSE ($\times 10^3$)	MAE ($\times 10^2$)
(A) Synthetic without K6	2.35	2.94
(B) Synthetic with K6	1.90	2.55
(C) Univariate without K6	2.44	2.95
(D) Univariate with K6	1.67	2.38
Bivariate	1.54	2.30

MSE = mean squared error, MAE = mean absolute error.

Table 8

Comparison of prediction performance between univariate and bivariate models under different sample size

Sample size	MSE ($\times 10^4$)		MAE ($\times 10^3$)	
	Bivariate	Univariate	Bivariate	Univariate
1	50.61	47.73	39.64	39.72
25	4.16	4.13	15.33	15.41
200	1.34	1.48	9.12	9.38
400	0.94	1.13	7.39	8.01