Web Appendix: Truth or Consequences: The intertemporal consistency of adolescent self-report on the Youth Risk Behavior Survey

Janet E. Rosenbaum

Revision: February 12, 2009

SUPPLEMENTARY DATA

Data analysis: error due to inconsistency

Estimating risk behavior prevalence using imperfect survey questions of unknown accuracy is analogous to the epidemiological problem of estimating disease prevalence using imperfect diagnostic instruments of unknown accuracy. We use a Bayesian method designed for this epidemiological problem.

The data from two medical binary tests for a disease have three degrees of freedom, but are described by seven parameters: population prevalence, sensitivity and specificity for each test, the probability of true positives on both tests, and the probability of true negatives on both tests. The problem is underdetermined, that is, any combination of sensitivity, specificity, and prevalence may be consistent with the observed data. The below data show the results of two tests for *Strongyloides* infection from a study to estimate the prevalence of *Strongyloides* in Cambodian refugees to Canada in 1982–83 for which only two imperfect tests are available: a stool test and a serology test. The two tests estimated dramatically different prevalences — 25% and 77%, respectively — due to the properties of the tests. As pointed out by original analysis of data (1) these different prevalences do not even take into account the possibility of sampling error, false positives, and false negatives. The below data for *Strongyloides* infection are equally consistent with many sets of parameters: 100% of the population could have the disease, but one test only finds the disease 77% of the time and the other test only finds it 25%; similarly, 0% of the population could have the disease, but the tests have false positive rates of 77% and 25%. Prior knowledge leads us to believe that neither is the case, but the data itself does not exclude these possibilities. We use priors in a Bayesian model to exclude such improbable answers.

		Stool test T_1		
		+	-	Total
Serology T_2	+	38	87	125
	-	2	35	37
	Total	40	122	162

Frequentist methods for two tests in one population assume that tests are independent conditional on disease status and that two of the test parameters are known with precision, and compute the remaining three parameters and estimate variance (2). Fixing parameters underestimates uncertainty, and the values which are fixed are often chosen arbitrarily.

Bayesian methods use a prior distribution on the seven parameters to compensate for lack of identifiability and view all seven parameters as uncertain. Frequentist methods which assign an exact value to some parameters yield answers equivalent to Bayesian method which uses a prior which is a point mass at the assigned value (1). The Bayesian methods are more realistic because it's unlikely to know any parameters so precisely, and they also avoid parametric assumptions in creating credible intervals around estimated parameters (1). Bayesian methods to estimate prevalence of a disease with two non-gold standard tests are described in (1, 3, 4), and in greater detail below. The prior compensates for lack of identifiability, so the prior influences posterior results, even with more data (4).

Parameters and notation

Two tests with binary results yield a two-by-two contingency table with cells n_{ij} where n_{ij} is the number of respondents giving answer i at wave 1 and answer j at time 2. Divide up each cell n_{ij} into the latent data: truly positive (denoted a_{ij}) and the truly negative (denoted b_{ij}).

Let C_1 and C_2 be specificity of tests 1 and 2, respectively, S_1 and S_2 sensitivity, and π be the prevalence of the behavior in the population. Let $p_{ij|k}$ be the probability that someone with disease state k will give responses (i, j). The tests are not independent. Due to inter-test agreement, $p_{ii|k} > p_{i.|k}p_{.i|k}$ for test result i and true disease state k. We include two parameters to describe correlation between answers $p_{00|0}, p_{11|1}$ which must satisfy $S_1S_2 < p_{11|1} < \min(S_1, S_2)$ and $C_1C_2 < p_{00|0} < \min(C_1, C_2)$ (3, 4).

Priors: original approach

Rare behaviors, with 1999 YRBS prevalence less than 15 %, are given a uniform distribution on the interval (0,0.5). All other risk behaviors are given a uniform distribution on the unit interval (0.0,1.0). The 29 rare behaviors are: lifetime forced sex, cocaine, inhalant, heroin, methamphetamine, steroid, injected illegal drugs, and pregnancy; past year injured in fight, threatened with weapon at school, physically hurt by boyfriend/girlfriend, attempted suicide, injured in suicide attempt; past month drove after drinking, carried gun, carried weapon at school, felt too unsafe for school, bought cigarettes, bought cigarettes and carded, smoke at school, smokeless tobacco, smokeless tobacco at school, no usual cigarette brand, alcohol at school, marijuana at school, cocaine, inhalants, diet pills for weight control, vomit for weight control.

We tried several priors for sensitivity and specificity: uniform, Beta(8,1), Beta(16,2), Beta(32,4), and choose the most diffuse prior which minimizes the number of prevalence estimates which differ substantially from past survey data. In the end, we use Beta(16,2) which means that we assume that sensitivity and specificity are probably at least 50 %, with a near-zero probability that these parameters are less than 50 %.

Parameter	Prior	Conditions
Prevalence, rare behavior	$\pi \sim \mathrm{U}(0, 0.5)$	
Prevalence, non-rare behavior	$\pi \sim \mathrm{U}(0, 1)$	
Sensitivity	$S_i \sim \mathrm{Beta}(\alpha_{\mathrm{i}}, \beta_{\mathrm{i}})$	$(\alpha_i, \beta_i) \in \{(1,1),(8,1),(16,2)\}$
Specificity	$C_i \sim \mathrm{Beta}(\alpha_{\mathrm{i}}, \beta_{\mathrm{i}})$	C_i such that $C_i + S_i > 1$
Pr[true positive agreement]	$p_{11\|1} \sim \mathrm{U}(S_1 S_2, \min(S_1, S_2))$	
Pr[true negative agreement]	$p_{00\|0} \sim \mathrm{U}(C_1 C_2, \min(C_1, C_2))$	

Priors

Sensitivity and specificity are given beta priors.We allow for potential retest effects by not requiring sensitivity or specificity to be identical in the first and second surveys since a respondent may view a question differently the second time than the first time.

The survey questions are assumed to convey some information, so $S + C > 1$ for each wave of the survey. Sensitivity and specificity are constant over the population due to lack of covariates.

The correlation between answers $p_{00|0}, p_{11|1}$ have uniform distribution on the intervals to which they are constrained due to inter-test agreement. $p_{11|1} \sim U(S_1 S_2, \min(S_1, S_2)), p_{00|0} \sim U(C_1 C_2, \min(C_1, C_2))$.

Gibbs sampler

After drawing the seven parameters from the above priors, the probabilities of response patterns conditional on true behavior status $p_{ij|k}$ are computed from the values of the six test properties. Bayes's rule uses these probabilities and the prevalence estimate to compute $p_{1|ij}$ the probability that the respondent truly engaged in the behavior conditional on response pattern. The number of people who truly engaged in the behavior and gave survey response pattern ij is drawn from a binomial distribution $a_{ij} \sim \mathrm{Bin}(\mathrm{n_{ij}}, \mathrm{p_{1|ij}})$. The posterior is a Beta distribution with parameters updated by the simulated latent data a_{ij}. Due to the large amount of data, the estimates converge quickly, but we use 5000 iterations and discard the first 1000 as burn-in. The programming was checked using the original *Strongyloides* data to ensure the same answers were reached.

The Gibbs sampler gives accurate results, but if the model is insufficiently specific the estimates will diverge. Due to lack of identifiability, some estimates of rare events differ substantially from past survey results. For each variable, we do 100 Gibbs samplers to be able to detect divergent prevalence estimates, and use the most vague prior that also avoids divergent prevalence estimate, that is, estimates which differ more than 50 percentage-points from the prevalence found in past surveys.

Gibbs sampler procedure

1. Draw initial values for the seven parameters from the prior distributions.

Parameter	Prior	
Prevalence	$\pi \sim \mathrm{Beta}(1,1)$	truncated at 0.5 if rare
Sensitivity	$S_i \sim \mathrm{Beta}(\alpha_{\mathrm{i}}, \beta_{\mathrm{i}})$	
Specificity	$C_i \sim \mathrm{Beta}(\alpha_{\mathrm{i}}, \beta_{\mathrm{i}})$	require $C_i + S_i > 1$
Prob both true positive	$p_{11\|1} \sim \mathrm{U}(S_1 S_2, \min(S_1, S_2))$	
Prob both true negative	$p_{00\|0} \sim \mathrm{U}(C_1 C_2, \min(C_1, C_2))$	

α, β are chosen to avoid divergent estimates.

If tests are independent $p_{11|1} = P[T_1^+ T_2^+ | D^+] = P(T_1^+|D^+)P(T_2^+|D^+) = S_1 S_2)$, and if the tests are totally correlated/dependent, then $p_{11|1} = P[T_1^+ T_2^+ | D^+] = P(T_1^+|D^+)P(T_2^+|T_1^+) = S_1 \times 1 = S_1$ or vice versa, so we constrain $p_{11|1}$ to the interval between these two estimates, and similarly for $p_{00|0}$.

2. Compute the probabilities of each response pattern conditional on true behavior status and the seven parameter values.

Parameter	Function
$p_{10\|1}$	$S_1 - p_{11\|1}$
$p_{01\|1}$	$S_2 - p_{11\|1}$
$p_{00\|1}$	$1 - S_1 - S_2 + p_{11\|1}$
$p_{10\|0}$	$C_2 - p_{00\|0}$
$p_{01\|0}$	$C_1 - p_{00\|0}$
$p_{11\|0}$	$C_1 - C_2 + p_{00\|0}$

3. Compute the probability of truly having engaged in the behavior, based on response pattern.

$p_{1|ij} = \frac{\pi p_{ij|1}}{\pi p_{ij|1} + (1-\pi) p_{ij|0}}$

4. Draw latent data: the number of respondents with each response pattern who are truly positive:

$a_{ij} \sim \mathrm{Bin}(\mathrm{n_{ij}}, \mathrm{p(1|ij)})$

$b_{ij} \sim n_{ij} - a_{ij}$

5. Calculate the likelihood given the latent data.

$p_{11|1}^{a_{11}} p_{10|1}^{a_{10}} p_{01|1}^{a_{01}} p_{00|1}^{a_{00}} p_{11|0}^{n_{11}-a_{11}} p_{10|0}^{n_{10}-a_{10}} p_{01|0}^{n_{01}-a_{01}} p_{00|0}^{n_{00}-a_{00}}$

6. The posterior probabilities are

Parameter	Prior	
Prevalence	$\pi \sim \mathrm{Beta}(1+\sum \mathrm{a}_{\mathrm{ij}}, 1+\sum \mathrm{b}_{\mathrm{ij}})$	
Sensitivity 1	$S_1 \sim \mathrm{Beta}(\alpha+\mathrm{a}_{11}+\mathrm{a}_{10}, \beta+\mathrm{a}_{01}+\mathrm{a}_{00})$	
Specificity 1	$C_1 \sim \mathrm{Beta}(\alpha+\mathrm{b}_{01}+\mathrm{b}_{00}, \beta+\mathrm{b}_{11}+\mathrm{b}_{10})$	require $C_1 + S_1 > 1$
Sensitivity 2	$S_2 \sim \mathrm{Beta}(\alpha+\mathrm{a}_{11}+\mathrm{a}_{01}, \beta+\mathrm{a}_{10}+\mathrm{a}_{00})$	
Specificity 2	$C_2 \sim \mathrm{Beta}(\alpha+\mathrm{b}_{10}+\mathrm{b}_{00}, \beta+\mathrm{b}_{11}+\mathrm{b}_{01})$	require $C_2 + S_2 > 1$
Prob both true positive	$p_{11\|1} \sim \mathrm{U}(S_1 S_2, \min(S_1, S_2))$	
Prob both true negative	$p_{00\|0} \sim \mathrm{U}(C_1 C_2, \min(C_1, C_2))$	

Additional analysis

The question about doctor visits in the past year had relative retraction of 21.1 % and relative initiation of 20.0 %; dentist visits in the past year had 14.8 % retraction and 10.6 % initiation.

Within the category of violence, the most consistent question was about forced sex; the five questions about perpetrating violence were the next most consistent questions; and four questions about being victim of violence were the least consistent.

The question about visiting the doctor when not sick in the past year has lower TCC than the question about visiting the dentist in the past year (0.72 versus 0.85).

Within the category of depressive symptoms in the past year, the highest TCC was considered suicide (0.94), attempted suicide (0.94), planned suicide (0.90), injured in suicide attempt to need medical attention (0.87), and felt sad or hopeless for at least two weeks (0.80).

Within the category of sex, the highest TCC question was whether the respondent had ever had sexual intercourse (TCC=0.99), followed by sex in the past 3 months (0.91), 4 or more lifetime sex partners (0.82), ever pregnant or made another pregnant (0.81), and had sex before age 13 (0.66).

Adolescents admit perpetrating violence more consistently than being the victim of violence, except forced sex. Forced sex has substantially higher TCC than other victim questions. Victims of forced sex may report more consistently due to saliency compared with other acts of violence. Adolescents may be reluctant to admit being victims of other crimes, so only report it sometimes, or may regard the crime as less significant than the perpetrators. Victim questions may be more ambiguous or subjective, such as the low TCC question asking whether respondents have stayed home from school in the past year because it felt too unsafe, or whether they were threatened with a weapon at school (5).

The questions about doctor and dentist appointments may have substantially different TCCs (0.72 vs. 0.85) because the doctor question has an extra qualification, "visited the doctor in the past 12 months when you were not sick" as opposed to "visited the dentist in the past 12 months." The qualification may decrease the salience of the answer since it requires evaluation and memory whether the respondent was sick when they visited the doctor the last time. Respondents may not have understood the intent of the question was to measure wellness visits as opposed to whether they happened to be sick at their last wellness visit.

The inconsistency for pregnancy is presumably being driven by boys because inconsistency is much higher in this two week period for both sexes compared with the one year inconsistency for girls only in the one year period between Add Health waves 1 and 2 (relative retraction is 45% compared with 19% in Add Health.) The high inconsistency for pregnancy may indicate that partners' pregnancies are not as memorable to boys, possibly due to uncertainty about paternity. Pregnancy questions were less consistent than traffic safety and substance use. Unfortunately reporting the results for this item together for boys and girls mixes two separate issues, and it's impossible to look at inconsistency in boys and girls separately without access to the full data.

Learn about HIV at school may have lowest TCC because virtually all adolescents have learned about HIV, and they may not reliably recall whether school was one of the places that they learned about HIV.

Consistency appears so strongly related to topic that no relationship between inconsistency and any other factor is evident across all topics. Within topics, however, factors such as time-frame seem to affect inconsistency. Within single risk behaviors, inconsistency is associated with more recent time-frames and more specific questions. Similarly, although we expect a relationship between prevalence and the estimated error due to inconsistency, with more rare behaviors having larger deviations, no such relationship is evident across all questions. After isolating the substance use questions, we see the expected inverse relationship between prevalence and error.

Questionnaire

The YRBS questionnaire from 1999 is available at

http://web.archive.org/web/19991128160242/www.cdc.gov/nccdphp/dash/yrbs/survey99.htm

References used in supplement

[1] Black M, Craig B. Estimating disease prevalence in the absence of a gold standard. *Statistics in Medicine.* 2002;21:2653–2669.

[2] Joseph L, Gyorkos T, Coupal L. Bayesian estimation of disease prevalence and the parameters of diagnostic tests in the absence of a gold standard. *American Journal of Epidemiology.* 1995; 141(3):263–272.

[3] Hui S, Zhou X. Evaluation of diagnostic tests without gold standards. *Statistical methods in medical research.* 1998;7:354–370.

[4] Denkdukuri N, Joseph L. Bayesian approaches to modeling the conditional dependence between multiple diagnostic tests. *Biometrics.* 2001;57:158–167.

[5] Turner CF. Why do surveys disagree? Some preliminary hypotheses and some disagreeable examples. In: Turner CF and Martin E, eds. Surveying subjective phenomena, vol. 2. New York, NY: Russell Sage Foundation; 1984; 159–214.