Supplementary Online Content

Godoy Garraza L, Walrath C, Goldston DB, Reid H, McKeon R. Effect of the Garrett Lee Smith Memorial Suicide Prevention Program on suicide attempts among youths. *JAMA Psychiatry*. Published online October 14, 2015. doi:10.1001/jamapsychiatry.2015.1933.

eTable 1. Average Value Across Intervention and Control Counties of Selected Covariates Before and After Matching by Subclass, 2004-2006

eTable 2. Estimated Average Effect of GLS Implementation (5% extreme weights truncated)

eTable 3. Estimated Average Effect of GLS on 16-23 Suicide Attempts Rate (per thousand) the Year Following Implementation, by Subclass

eTable 4. Estimated Average Effect of GLS Implementation by age subgroup

eAppendix. Methods

eReferences

eTable 1: Average Value Across Intervention and Control Counties of Selected Covariates Before and After Matching by Subclass, 2004-2006

	% of Respondents									
	Subclass 1		Subclass 2		Subclass 3		Subclass 4		Subclass 5	
Covariates	Control[a] (*n* = 24 000)[b]	Intervention (*n* = 5000)[b]	Control[a] (*n* = 22 000)[b]	Intervention (*n* = 7000)[b]	Control[a] (*n* = 19 000)[b]	Intervention (*n* = 16 000)[b]	Control[a] (*n* = 13 000)[b]	Intervention (*n* = 11 000)[b]	Control[a] (*n* = 6000)[b]	Intervention (*n* = 18 000)[b]
Female	51.9	52.0	53.2	50.2	52.1	50.0	55.3	53.2	54.4	53.7
Age Group, y										
12-17	12.1	13.1	12.2	11.4	10.6	11.4	10.3	9.6	11.6	10.4
18-25	13.8	13.1	14.3	14.3	14.4	16.4	17.2	16.9	18.9	17.9
26+	74.1	73.8	73.4	74.3	74.9	72.2	72.4	73.5	69.5	71.7
Race/Ethnicity										
White, non-H.	82.5	83.6	84.4	82.8	82.2	79.7	77.1	82.5	72.0	75.6
African American, non-H.	9.0	8.0	7.3	10.4	10.3	11.1	9.7	8.2	11.7	5.7
American Indian/AN, non-H.	0.5	2.4	1.1	0.5	0.7	0.6	1.1	1.6	6.9	5.2
Nat. Haw. and Oth. Pac. Isl., non-H.	0.1	0.2	0.2	0.1	0.1	0.1	0.1	0.1	0.1	0.4
Asian, non-H.	1.1	0.7	1.0	0.7	1.3	2.0	2.3	1.8	2.2	2.7
Multiple races, non-H.	0.9	1.1	1.2	0.7	1.3	0.8	1.6	1.1	1.6	2.1
Hispanic	6.0	4.0	4.9	4.8	4.1	5.5	8.0	4.7	5.5	8.2
Education (age ≥18 y)										
<High school	18.8	21.1	17.9	21.7	18.0	18.1	18.3	15.0	16.9	14.7
High school graduate	39.7	39.6	38.8	39.5	33.8	33.1	34.9	35.7	29.9	33.3
Some college	23.8	23.0	24.6	20.9	25.0	25.8	25.0	25.4	32.1	25.1
College graduate	17.7	16.3	18.7	17.9	23.2	23.0	21.8	23.9	21.1	27.0
In school (age ≤23 y)	72.4	76.7	72.3	66.2	75.1	76.8	76.9	74.7	81.4	71.8
Married (age >14 y)	56.3	52.8	54.5	55.6	53.4	53.8	53.1	51.2	48.4	53.3
Total family Income, $										

<20 000	21.2	21.0	22.6	21.3	24.8	23.1	24.9	22.6	27.0	22.9
20 000-49 999	37.9	41.0	37.5	40.8	37.5	38.5	35.3	36.0	35.7	38.3
50 000-74 999	19.7	18.1	18.5	17.9	16.0	17.1	18.8	18.2	14.5	19.0
$75 000	21.1	20.0	21.4	20.0	21.7	21.3	21.0	23.2	22.8	19.9
Employment status										
Employed full-time	51.2	51.0	49.5	52.6	48.6	51.4	48.5	50.7	48.5	51.2
Employed part-time	14.4	11.5	14.0	16.7	12.9	12.8	15.3	15.0	18.4	15.7
Unemployed	3.8	4.5	3.9	3.5	4.1	4.6	3.4	3.3	3.6	3.0
Other (incl. not in labor force)	30.6	33.0	32.6	27.2	34.4	31.3	32.9	31.0	29.5	30.1
Has health insurance	85.4	83.3	85.8	85.4	85.7	84.1	85.6	87.1	85.9	87.6
In rural segment	48.2	52.5	42.9	37.4	29.3	31.7	24.0	29.0	23.3	18.5
Had lifetime major depressive episode	15.3	15.0	14.6	14.6	15.7	17.1	13.8	15.6	15.0	17.0
Had past year major depressive episode	9.1	9.0	8.2	6.8	8.2	10.4	6.7	7.7	9.0	8.5

[a] Weighted average across the 5 subclasses of counties, where the weights are given by the proportion of counties in each subclass among the intervention counties.
[b] Unweighted number of respondents in intervention and control counties rounded to the nearest thousand due to NSDUH disclosure restrictions.

eTable 2: Estimated Average Effect of GLS Implementation (5% extreme weights truncated)

	Estimate	Std. Error	*Pr*(>|*t*|)
Suicide attempts rate 16-23 (per thousand)			
GLS previous year	-3.16	1.54	.045
GLS 2 years ago or more	-0.72	1.80	.690
Suicide attempts rate 24 plus (per thousand)			
GLS previous year	2.53	2.91	.389
GLS 2 years ago or more	-2.16	2.80	.445

eTable 3: Estimated Average Effect of GLS on 16-23 Suicide Attempts Rate (per thousand) the Year Following Implementation, by Subclass

	Estimate	Std. Error	*Pr*(>|*t*|)
Untruncated weights			
Subclass 1	-2.34	2.61	.374
Subclass 2	-8.19	2.33	.001
Subclass 3	-6.99	2.36	.005
Subclass 4	-1.26	2.18	.565
Subclass 5	-3.10	8.88	.729
5% extreme weights truncated			
Subclass 1	2.55	3.34	.447
Subclass 2	-7.55	2.43	.003
Subclass 3	-6.37	2.44	.012
Subclass 4	-1.07	2.43	.660
Subclass 5	-1.58	7.57	.836

eTable 4. Estimated Average Effect of GLS Implementation by age subgroup [a]

	Estimate	Std. Error	*Pr*(>|*t*|)
Untruncated weights			
Suicide attempts rate 16–19 (per thousand)			
GLS previous year	-4.46	2.14	.042
GLS 2 years ago or more	-2.70	2.98	.369
Suicide attempts rate 20–23 (per thousand)			
GLS previous year	-5.68	2.46	.025
GLS 2 years ago or more	3.09	3.63	.399
5% extreme weights truncated			
Suicide attempts rate 16–19 (per thousand)			
GLS previous year	-1.80	2.31	.440
GLS 2 years ago or more	-2.33	3.19	.470
Suicide attempts rate 20–23 (per thousand)			
GLS previous year	-5.30	2.33	.027
GLS 2 years ago or more	3.64	3.37	.285

[a] Weighted average across the 5 subclasses of counties, where the weights are given by the proportion of counties in each subclass among the intervention counties.

eAppendix

Methods

Sample Selection and Propensity Score

This analysis used the same sample of counties and years as well as procedures to increase comparability to the analysis of suicide mortality by Walrath et al.[1] This section summarizes these procedures.

Sample Selection

All U.S. counties with a population of at least 3,000 youths between the ages of 10 and 24 were considered for inclusion in the sample.[*] Of these, 479 were exposed to GLS suicide prevention efforts at some point between 2006 and 2009 as signaled by the implementation of at least one GLS gatekeeper training (intervention counties). In the remaining 1,616 counties, there were no GLS trainings (our proxy for GLS program activities) implemented during that period. This group of counties constituted a pool of potential control counties from which a sample of 1,161 counties that shared key pre-intervention characteristics with the intervention counties was selected using propensity score matching techniques. It was not possible to find adequate matches for 13 of the 479 intervention counties, and these were therefore excluded from the analysis. In some of the 466 intervention counties, training implementation occurred in more than 1 year; as a result, the sample contained a total of 776 county-years in which at least one GLS training was implemented.

Addressing Comparability

[*] Smaller counties were not considered for inclusion since the large variability of youth suicide mortality among them makes it extremely difficult to detect any systematic difference. A total of 1,047 out of 3,142 counties did not reach the 3,000-youth threshold.

Three sequential steps were taken to address comparability between intervention and control counties and county-years before the analysis of outcome information (and, in fact, before accessing NSDUH data on suicide attempts). The three steps relied on the estimation of the *propensity score*, the probability of being in the intervention sample as opposed to the control sample as a function of the observed covariates.[2,3] Specifically, propensity score models were used sequentially to (1) select the sample of control counties (trimming step), (2) create five homogeneous subgroups across the intervention and control samples (subclassification step), and, finally, (3) develop inverse probability of exposure weights within each subgroup (weighting step). In the first two steps, the propensity models were used to predict the implementation of at least one GLS training at any time between 2006 and 2009 as a function of historical county characteristics, ie, average values of the covariates between 2000 and 2006. The third step incorporated time-varying covariates in addition to time-fixed covariates in order to predict training implementation in each year.

The trimming and subclassification steps had the common goal of making intervention and control counties as similar as possible regarding relatively permanent historical characteristics. An additional benefit of subclassification was that variation in the effect of training implementation by subclass could be explored. Furthermore, the propensity model for the weighting step was fitted separately within each subclass, making it more likely that the overall model for the weights was correctly specified. The goal of the weighting step was to account for time-varying covariates and, in particular, recent changes in suicide rates before the implementation and the history of exposure to the GLS program in previous years.[4, 5, 6]

Covariates Used in the Propensity Score

Covariates considered for the propensity score included county's total population, age-group composition, race-ethnic composition (percentage Hispanic and non-Hispanic whites, African American, American Indian or Alaskan Native, Asian and other race), gender composition, median household income, poverty rate, unemployment rate, and percentage of rural population. In addition, pre-intervention rates of suicide mortality were included as covariates. Relatively permanent characteristics of the counties were assessed by averaging each covariate between 2000 and 2006 (time-fixed covariates). Recent change in these characteristics was assessed through the value in each of the previous 4 years as well as the moving average throughout that 4-year period (time-varying covariates). The source of demographic information was the U.S. Census Bureau of Statistics' Intercensal Estimates.[7] Income, poverty, and unemployment rates were based on small area estimations by the U.S. Census Bureau of Statistics[8] and the Bureau of Labor Statistics.[9] Pre-intervention mortality information was sourced from the Compressed Mortality File[10]

Propensity Score Model Specifications

Each of the three steps described above (trimming, subclassification and weighting) requires a separate specification of a propensity score model. We used the same set of procedures to reach the final specification in all cases. Starting with a list of candidate covariates to include (all time-fixed covariates in the case of the trimming and subclassification steps, and all the time-fixed and time-varying covariates in the weighting step), two sequential stepwise regression procedures were implemented. The first procedure was used to determine which subset of covariates was to be included. The

second procedure was used to decide whether two-way interaction or second-order polynomials among the retained covariates had to be added. Higher order interactions or polynomial terms were not considered. The time-varying propensity model included indicators of the history of previous implementations and year indicators, before the other covariates were considered for inclusion.

Trimming Step

For each intervention county, we selected the closest in terms of propensity from among the counties in the control pool that were no more than 0.2 standard deviations away. In the few cases where there were no counties that close, we dropped the intervention county from the sample. We used *with replacement* selection, which prioritizes quality over quantity of matches.[11] On the other hand, in situations where multiple neighbors were equally close, which frequently was the case for smaller counties, all of them were included. This strategy increased the number of matches with no cost in terms of quality.

Subclassification

To determine the subclasses, we divided the sample into two subclasses using the median propensity score among the intervention counties. We computed a *t* statistic for the difference in mean propensity score between intervention and control samples in each subclass. If the *t* statistic was larger than 1.96, we split the sample in that subclass again, using the median propensity score among the intervention counties. This strategy resulted in 5 groups containing approximately 0.25, 0.25, 0.25, 0.125 and 0.125 of the intervention counties. The approach overcomes some of the criticism of the more popular approach that relies on the quintile of the distribution of the estimated propensity score.[12]

Weighting

Stabilized weights were used in all cases. In order to increase efficiency, the estimated probability of implementation at each time point conditional on the history of previous implementations is incorporated into the numerator.[4]

Software

All procedures were implemented with "R," and, in particular, the "Matching" and "ipw" packages.[13-15]

NSDUH Complex Design

Unlike the mortality analysis for which only county and year aggregated data were available, NSDUH contains data on individual respondents within each county and year. This section further describes a few additional technical details regarding the treatment of the NSDUH sample design for the analysis.

Sample Design Weights

Since county-years were defined as the unit of analysis of interest, a possible approach would have been to first compute the county-year rates using the sample design weights and subsequently run the main analysis on the aggregated data. In the present analysis, however, we directly analyzed respondent-level data. While our unit of analysis remained the county, we used these *microdata* in the regression (as opposed to computing the county rates first, in a 2-step procedure), incorporating the NSDUH's sample design weights. We also included linear terms for individual demographic characteristics, both to gain precision and to obtain *adjusted* rates. The longitudinal aspect of the analysis and the fact that the NSDUH used a complex sample design were 2 possible sources of lack of independence (ie, clustering) across observations. As in the previous study, we relied

on generalized estimating equation (GEE) with robust inference, assuming only independence between states but allowing for any pattern of lack of independence within a state. Sensitivity of the results to extreme inverse probability of exposure weights was assessed by refitting the regression after truncating 1% and 5% of the weights at each extreme of the distribution.

We previously rescaled sample design weights to add up to 1 within each county and year. Point estimates of the quantities of interest were identical in both approaches. The selected approach incorporated the additional layer of uncertainty associated with the fact that county-year rates were based only on a sample and not on the universe of cases.

Other Designs Features

The use of cluster sampling in NSDUH resulted in a loss of precision when compared with simple random sampling. We accounted for this in the analysis through the use of GEE[1] and standard errors that assumed independence only between states but were robust to any pattern of correlation within states, including clustering resulting from the sample design.[16]

While we did not directly take into account NSDUH stratification in the analysis, we included basic individual-level demographic characteristics as covariates in the analysis to partially account for this fact.

Missing Data

As noted in the Methods section, suicide attempt rates were not available for a fraction of counties and years in the original mortality sample. No special treatment for these missing data was undertaken. This amounted to assuming that the availability of the rate for a certain county and year was not related with the potential outcomes of the

county following the implementation (or absence of implementation) of the GLS. The documented balance between intervention and control samples somewhat validates this assumption (see, in particular, Table 1 in the body of the manuscript).

Software

All procedures were implemented using "R," and, in particular, the "survey" packages.[13,17]

eReferences

1. Walrath CM, Godoy Garraza L, et al. The impact of the Garrett Lee Smith (GLS) Suicide Prevention Program on suicide mortality. Submitted for publication. 2014.

2. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika.* 1983;70(1):41–55.

3. Rosenbaum PR, Rubin DB. Reducing bias in observational studies using subclassification on the propensity score. *J Am Stat Assoc.* 1984;79(387):516–524.

4. Cole SR, Hernán MA. Constructing inverse probability weights for marginal structural models. *Am J Epidemiol.* 2008;168(6):656–664.

5. Hernán MA, Brumback BA, Robins JM. Estimating the causal effect of zidovudine on CD4 count with a marginal structural model for repeated measures. *Stat Med.* 2002;21(12):1689–1709.

6. Robins JM, Hernán MA, Brumback B. Marginal structural models and causal inference in epidemiology. *Epidemiology.* 2000;11(5):550–560.

7. 2000-2010 County Totals Intercensal Population Estimates.In: Population Estimates: County Intercensal Estimates (2000-2010) [database online]. Washington, DC: U.S. Department of Commerce, U.S. Census Bureau. http://www.census.gov/popest/data/intercensal/county/county2010.html. Published September 2011. Accessed February 12 2013.

8. State and county data files. In: Model-based Small Area Income & Poverty Estimates (SAIPE) for School Districts, Counties, and States [database online]. Washington, DC: U.S. Department of Commerce, U.S. Census Bureau. http://www.census.gov/did/www/saipe/index.html. Published November 2012. Accessed August 19 2013.

9. Labor Force Data by County. In: Local Area Unemployment Statistics [database online]. Washington, DC: Bureau of Labor Statistics. http://www.bls.gov/lau/#data. Published April 2013. Accessed August 19 2013.

10. CDC WONDER: Compressed Mortality File 1999-2010 Series 20 No. 2P. In: About Compressed Mortality, 1999-2011 [database online]. Atlanta, GA: Centers for Disease Control and Prevention, National Center for Health Statistics. http://wonder.cdc.gov/cmf-icd10.html. Published 2013. Accessed November 27 2013.

11. Dehejia RH, Wahba S. Propensity score matching methods for nonexperimental causal studies. *Review of Economics and Statistics.* 2002;84(1):151–161.

12. Lunceford JK, Davidian M. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Stat Med.* 2004;23(19):2937–2960.

13. R Development Core Team. *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing; 2013.

14. Sekhon JS. Multivariate and propensity score matching software with automated balance optimization: the matching package for R. *Journal of Statistical Software.* 2011;42(7).

15. van der Wal WM, Geskus RB. ipw: an R package for inverse probability Weighting. *Journal of Statistical Software.* 2011;43(13).

16. Zeger SL, Liang K. Longitudinal data analysis for discrete and continuous outcomes. *Biometrics.* 1986;42(121);30.

17. Lumley T. Analysis of complex survey samples. *Journal of Statistical Software.* 2004;9(8).