EVALUATION OF ADJUSTMENTS FOR NONRESPONSE BIAS APPLIED TO PROVIDER NONRESPONSE IN THE NATIONAL IMMUNIZATION SURVEY

Philip J. Smith, CDC; David C. Hoaglin, Abt Associates Inc.; J.N.K. Rao, Carleton University;
Danni Daniels, CDC; and Michael P. Battaglia, Abt Associates Inc.
Philip J. Smith, CDC, 1600 Clifton Road, NE, MS E–62, Atlanta, GA 30333

1 INTRODUCTION

Among the threats to validity of conclusions from a health survey is the bias who can result from differences between people who respond to the survey and people that do not respond. When respondents and nonrespondents have very different characteristics and the proportion of nonrespondents is at least moderately large, unadjusted survey estimates are often severely biased. Thus, it is important to evaluate the effect of nonresponse bias on the survey estimates.

Brick and Kalton[1] indicate that the most common method for adjusting for nonresponse bias in health surveys is by using "adjustment cells." In many national health surveys adjustment cells are defined by demographic and socioeconomic factors. Ezzati and Khare[2] discuss an example of this approach. More recently Smith et al.[3,4] give an example in which adjustment cells are formed using response propensities. Potter et al.[5] apply estimated response propensities directly to sampling weights. This approach is equivalent to choosing as many adjustment cells as there are respondents with complete data.

Regardless of the method used to form cells, a natural question is, "How well did the method work in reducing nonresponse bias?" This paper describes a statistical method for making this evaluation. The approach arises from further research on the statistical methods that the National Immunization Survey (NIS) uses to (i) evaluate the extent of nonresponse bias and (ii) reduce this potential bias.

The design of the NIS has two phases of sampling: a list-assisted random-digit-dialing (RDD) survey of households and a mail survey of vaccination providers of eligible children in sampled households. The RDD phase uses simple random sampling in each of 78 Immunization Action Plan (IAP) areas, which comprise the 50 states and 28 large metropolitan areas, including the District of Columbia. Each IAP area forms a stratum of the sampling design, and within these strata the NIS samples independently. The target population for the NIS is all children living in the 78 IAP areas who are between 19 and 35 months of age at the time of the RDD interview. More complete descriptions of the sample design are given by Smith et al.[3], Ezzati-Rice et al.[6], and Zell et al.[7]

In the RDD phase, the respondent in a sampled household with eligible children is asked to report each child's vaccination history (including the date of each shot, if a "shot card" is available), as well as demographic and socioeconomic information. Parents and guardians are asked for consent to contact the children's vaccination providers to obtain vaccination histories. If verbal consent is obtained, the providers are mailed a vaccination history questionnaire. Data obtained from responding providers include information on whether the children received the recommended number of doses of four vaccines: diphtheria and tetanus toxoids and pertussis vaccine (DTP), poliovirus vaccine (polio), measles-containing vaccine (MCV), and *Haemophilus influenzae* type b vaccine (Hib). A child is said to be "4:3:1 up-to-date" if he/she has received 4 or more doses of DTP, 3 or more doses of polio, and 1 or more doses of MCV. A child who has also received 3 or more doses of Hib is said to be "4:3:1:3 up-to-date." Percentages of children who are up-to-date are referred to in immunization research as vaccination coverage rates.

For a variety of reasons the NIS is unable to obtain any vaccination history from providers for some children who have a completed household interview. In 1998 the NIS completed RDD interviews for 32,511 children. Among these, adequately detailed provider-verified vaccination histories from which children's up-to-date vaccination status could be determined were obtained from 21,827 (67.1%) children. In the remainder of this paper "provider nonresponse" refers to the absence of adequate provider-verified vaccination histories.

Official estimates of vaccination coverage rates in the NIS adjust for provider nonresponse using adjustment cells based on response propensities.[3] More specifically, the NIS estimates each child's response propensity (i.e., the probability that the child has adequate provider response) and uses the quintiles of those response propensities to define five adjustment cells within each stratum.

The present paper explores variations on that method by allowing the data to guide the choice of the number of

cells and the choice of the measure on which the cells are based. Specifically, we illustrate our approach by evaluating how well nonresponse bias is reduced for estimates of 4:3:1 vaccination coverage rates (obtained by using 5 adjustment cells based on propensities). For the comparison alternative estimates use adjustment cells based on response propensities and predictive probabilities of being 4:3:1 up-to-date. To determine the most suitable number of cells, we examine the incremental decrease in bias obtained by increasing the number of cells.

In an econometric application, David et al.[8] describe the use of adjustment cells based only on the use of response propensities. Little[9] describes the merits of using either response propensities alone or predictive probabilities alone. Eltinge and Yansaneh[10] give a method for determining the number of adjustment cells when they are formed using either response propensities or predictive probabilities, but not both. Empirical results from our research indicate that the reduction in bias depends upon whether response propensities or predictive probabilities are chosen to form adjustment cells. None of these authors give recommendations on how to choose between response propensities or predictive probabilities for forming adjustment cells, however. The methods that we describe use both response propensities and predictive probabilities to form adjustment cells and provide a way of evaluating whether to use response propensities, predictive probabilities, or a combination of the two.

2 ADJUSTMENT FOR PROVIDER NONRESPONSE

2.1 Rationale for the Adjustment

Empirical results suggest that children with adequate provider response have characteristics that are believed to be associated with a greater likelihood of being 4:3:1 up-to-date, compared to children with provider nonresponse. Specifically, children with adequate provider response are more likely to live in households that have higher total incomes, to have a white mother, and to live outside a central city of a Metropolitan Statistical Area. These factors are believed to be associated with higher vaccination rates. Also, children with provider nonresponse are less likely to live in the state where they were born and less likely to come from a household whose respondent could locate a written record of the child's vaccination history (i.e., a shot card). Both of these factors indicate a potential lack of continuity of health care, and have been shown to correlate with lower vaccination rates.[11] If no adjustment is made for these differences, estimated vaccination coverage rates may be too high. Then immunization program managers may misjudge the success of their efforts to increase vaccination coverage rates and not realize that improvements in their programs are warranted.

In forming cells to adjust for provider nonresponse in the NIS, the aim is to group together children, within each IAP area, who have similar patterns of vaccination coverage. Because a number of background variables are associated with both provider nonresponse and vaccination coverage, we combine the variables in a model for the child's response propensity (similar to the method in current use) and, separately, in a model for the predictive probability of being 4:3:1 up-to-date. Children who have similar response propensities and similar predictive probabilities of being 4:3:1 up-to-date will also be similar with respect to the background variables that are predictive for these two measures. Thus, within each IAP area, we group children into adjustment cells according to (i) the similarity of their response propensities or (ii) the similarity of their predictive probabilities or (iii) a combination of these two measures.

Within the context of using adjustment cells to reduce nonresponse bias, all of the children within a cell are represented by those who have adequate provider data. For a specific variable, y, the extent of the reduction in bias depends on the degree to which children are comparable with respect to the measures used to construct the adjustment cells, as well as how closely those measures are associated with y.

When adjustment cells are defined by the estimated quantiles of the distributions of estimated response propensities and estimated predictive probabilities of y, nonresponse bias is reduced to an extent that depends on the number of adjustment cells and on how the adjustment cells combine the estimated response propensities and the estimated predictive probabilities.

2.2 Adjustment with a Fixed Number of Cells

We describe our approach in two stages. First, we take the number of adjustment cells as fixed. That is, we fix both the number of quantiles of response propensity (k_1) and the number of quantiles of the predictive probability (k_2). Thus, the focus is simply on forming the cells, and not on the number of cells. Subsections 2.2.1 through 2.2.4 describe how the estimated response propensities and predictive probabilities are obtained, how they are used to form adjustment cells, and the raking that is conducted as a further measure to control bias. Then, Section 2.3 varies both the total number of

cells ($k_1 \times k_2$) and, within that, the values of k_1 and k_2, and uses the incremental reductions in estimated bias to choose the "best" combination of (k_1, k_2).

2.2.1 Estimated Response Propensities

As a first step in forming adjustment cells, a response propensity model was developed using logistic regression. In this model the outcome variable for each child was an indicator variable of whether the child had adequate provider data. The candidates for predictors were categorical variables that have been found to be associated with vaccination status in other CDC research and are listed in Table 1. A forward stepwise procedure selected the actual predictors from these candidates.

At each step of the selection process, the logistic regression examined all possible first-order predictors as candidates for inclusion in the model. Also, after adding a predictor to the model, predictors added previously were considered for removal. Akaike's[12] information criterion (AIC) was used for choosing the optimal set of candidate predictors at each step. Table 2 gives the analysis-of-deviance table[13] for the selection process and lists the predictors selected for the response propensity model. All children with a completed RDD interview received an estimated response propensity.

2.2.2 Estimated Predictive Probabilities

Following the same stepwise process as described above, the predictive probability model used the data from children with adequate provider response in a logistic regression of the indicator of whether the child's provider(s) indicated he/she was 4:3:1 up-to-date. Table 3 gives the analysis-of-deviance table for the selection process and lists the predictors selected for the predictive probability model. All children with a completed RDD interview received an estimated predictive probability.

2.2.3 Forming the Adjustment Cells and Adjusting for Provider Nonresponse

One goal was to have approximately the same unweighted sample count in each adjustment cell. Thus, to produce $k_1 \times k_2$ adjustment cells ($k_1 = 1, 2, \ldots$ and $k_2 = 1, 2, \ldots$), we divided the distribution of estimated response propensities into k_1 equal parts and then, within each of those, divided the conditional distribution of estimated predictive probabilities into k_2 equal parts. To adjust for provider nonresponse, within each adjustment cell, children with adequate provider data were assigned a revised weight by dividing their RDD-phase sampling weight by the cell-specific weighted provider response rate. (From the RDD phase each child with a household interview has a sampling weight, which incorporates an adjustment for unit nonresponse, an adjustment for noncoverage of nontelephone households, and poststratification on race/ethnicity of mother, education of mother, and age of child. Battaglia et al.[14] provide a more detailed description of the adjustments made to the sampling weights.)

2.2.4 Raking to Control Bias and to Maintain the Adjustment

As mentioned in Section 2.1, by dividing the RDD-phase sampling weights of children who have adequate provider data by their adjustment-cell-specific weighted provider response rate, these children more fairly represent all the children in the cell. However, the revised weights may not match the poststratification totals used to construct the RDD-phase sampling weights. Also, the revised weights may not match the RDD-phase sample-weighted totals of variables that are known to be important predictors of being up-to-date. To reduce bias attributable to these differences, we use iterative ratio adjustment[15] to rake the revised weights to match poststratification totals, outcome predictor totals, and variable totals that are required to maintain the face validity of the survey. Equally important, we rake to the adjustment-cell-specific RDD-phase sampling weight totals to maintain the effect of the nonresponse adjustment. The variables used in this process are education of the mother, race of the child, age group of the child, sex of the child, first-born status, MSA status, and the indicator for the adjustment cell.

2.3 A Sequential Method for Choosing the Total Number of Adjustment Cells and (k_1, k_2) Pair

To choose k_1 and k_2, we form cells using both response propensities and predictive probabilities. We proceed in stages (within each IAP area). At stage n, n = 2,..., we seek a "best pair" (k_1, k_2) among all values of k_1 and k_2 such that $k_1 \times k_2 = n$. The search begins by calculating each pair's estimated adjusted rate. Letting (k_1', k_2') denote the best pair at stage n-1 and (k_1, k_2) denote any one of the pairs at stage n, the estimated incremental reduction in bias of $\hat{\bar{Y}}_{k_1,k_2}$ relative to $\hat{\bar{Y}}_{k_1',k_2'}$ is

$$\hat{\Delta}_{k_1,k_2} = \hat{\bar{Y}}_{k_1,k_2} - \hat{\bar{Y}}_{k_1',k_2'}. \quad (1)$$

We obtain its estimated standard error, $\hat{\sigma}(\hat{\Delta}_{k_1,k_2})$, by using a Taylor-series approximation.[16] As discussed in Section 1, estimates of vaccination coverage that are not adjusted for provider nonresponse may be too high. Thus, we are particularly interested in adjusted estimates $\hat{\bar{Y}}_{k_1,k_2}$ that yield negative values of $\hat{\Delta}_{k_1,k_2}$. When n = 2, equation (1) measures the extent to which the positive bias of the unadjusted estimator $\hat{\bar{Y}}_{1,1}$ can be reduced by using only 2 adjustment cells.

To evaluate the statistical significance of the estimated incremental bias reduction for each (k_1, k_2) pair, we compute one-sided p-values

$$p_{k_1,k_2} = \Phi\left(\frac{\hat{\Delta}_{k_1,k_2}}{\hat{\sigma}(\hat{\Delta}_{k_1,k_2})}\right).$$

Here $\Phi(\cdot)$ denotes the standard normal cumulative distribution function. At stage n we determine the (k_1, k_2) pair that yields the most statistically significant reduction in estimated incremental bias (1) obtainable by using n adjustment cells compared to n-1 adjustment cells. This pair corresponds to the smallest value of p_{k_1,k_2}, and we refer to this pair as the "best" (k_1, k_2) pair at stage n. At stage n we let $\hat{\bar{Y}}_n$ denote the estimated coverage rate associated with the best (k_1, k_2) pair. Also, we let $\hat{\Delta}_n$ and p_n denote the corresponding estimated reduction in bias and its p-value.

To choose n, we examine the sequence of p_n, n = 2,..., and define the "best" value of n as the smallest value n' such that $p_l \le 0.05$, $l = 2,...,n'$ and $p_{n'+1} > 0.05$. If no value of n satisfies these relations, no bias reduction is required, and the unadjusted estimate $\hat{\bar{Y}}_{1,1}$ is used to estimate the vaccination coverage rate.

3 ILLUSTRATION OF THE METHOD

Table 4 illustrates the strategy of selecting the best (k_1, k_2) pair for the 4:3:1 series for one IAP area, Oregon. At stage 2 the (1, 2) pair produces a statistically significant reduction in bias, and so does using 3 adjustment cells (the (1, 3) pair) and 4 adjustment cells (the (1, 4) pair). However, the (5, 1) pair does not yield a statistically significant further reduction in bias. In this instance k_1 = 1 for all "best" pairs; that is, the sets of adjustment cells that achieve the most significant reductions in bias use only predictive probabilities. Starting from no adjustment, the actual reductions in nonresponse bias are 0.39 percentage point for (k_1, k_2) = (1, 2), 0.45 percentage point for (1, 3), and 0.34 percentage point for (1, 4). Thus the total reduction in bias from provider nonresponse is 1.18 percentage points.

Among the 78 IAP areas the "best" (k_1, k_2) pairs contain at most 4 adjustment cells, and 53 IAP areas have k_1 = 1 and k_2 > 1. A total of 7 IAP areas have k_1 > 1 and k_2 = 1. For 18 IAP areas the best choice is no adjustment, (1,1). Among the 60 IAP areas with some adjustment, only 6 have a bias reduction greater than 1 percentage point, and the median reduction is 0.51 percentage point. Thus, the estimates of 4:3:1 coverage for most IAP areas require little adjustment to reduce bias attributable to differences between children who have adequate provider data and those who do not have adequate provider data. In this regard, the threat to validity in the NIS posed by provider nonresponse seems small.

In describing the use of stratification to control bias in observational studies, Cochran[17] reported results of an empirical investigation showing that bias is reduced by 90% when as few as 5 adjustment cells are used. This empirical finding has led to the common belief that 5 is a good choice for the number of adjustment cells to reduce bias. Also, Rosenbaum and Rubin[18] have shown the advantages of constructing adjustment cells using estimated response propensities. Thus, for the 50 States and the District of Columbia, we also compared the estimate using 5 cells based on response propensities (k_1 = 5, k_2 = 1) to the estimate based on the best choice of (k_1, k_2). For both of these methods, the estimated total reduction in bias was obtained by comparing the adjusted coverage rate to the unadjusted coverage rate, $\hat{\bar{Y}}_{1,1}$. Results for the 1998 NIS data indicated that the estimated total reductions in the bias for the two methods differed significantly in only 3 states.

4 CONCLUSIONS

As noted by Little, use of response propensities to form adjustment cells may reduce the bias, but not necessarily the variance, of adjusted rates. Also, when adjustment cells are formed using predicted

probabilities, reduced bias and variance are expected. In this regard, the variance of adjusted estimates is expected to be larger when response propensities are used to form adjustment cells than when predictive probabilities are used. As a consequence, one would expect that the statistical power to detect reductions in incremental bias by using more adjustment cells would be smaller when using response propensities (only) than when using predictive probabilities or a "best" combination of both predictive probabilities and response propensities. Empirical results reported by Smith et al.,[3] using data from the 1998 NIS, show that when response propensities (only) were used to form cells, the "best" choice of the number of adjustment cells was 1 for 58 of the 78 IAP areas. In comparison, in this paper we show that, when adjustment cells are formed using a "best" combination of both predictive probabilities and response propensities, the number of adjustment cells is greater than 1 in 60 of the 78 IAP areas.

This potential reduction in statistical power sheds further light on the result that the difference in bias reduction between using 5 adjustment cells based on response propensities and using a "best" combination of both predictive probabilities and response propensities was statistically significant in only 3 states. Because estimates adjusted using response propensities (only) are expected to have higher variance, this may lead to reduced power in making this specific statistical comparison.

As noted previously, many other national health surveys use adjustment cells that are defined by demographic and socioeconomic factors. This approach is equivalent to an implicit model for major survey outcomes. One would anticipate that those implicit models would not be as predictive as explicitly defined models that enable the importance of factors used to form adjustment cells to be evaluated in a formal significance-testing framework. If so, one might expect that cells formed using implicit models would not necessarily reduce bias adequately. Regardless of the method of forming adjustment cells for adjusting for nonresponse bias, it makes sense to evaluate how well bias was reduced. This paper provides a method that tailors adjustment cells to obtain a "best" reduction in bias according to the definition described in Section 2.3. In this regard, it may be used to evaluate the effectiveness of other methods of forming adjustment cells in reducing nonresponse bias.

Acknowledgments

The authors gratefully acknowledge the many helpful comments made by Barry Graubard and Phil Kott during the early stages of this research. Also, the authors thank K.P. Srinath, David Judkins, Monina Klevens, and Lawrence Barker for their constructive comments and encouragement throughout the research.

Table 1. Predictors offered to model selection for response propensity and predictive probability models.

abbreviation	description
shotcard	"shot" card used during RDD interview
all4shot	HH report: up-to-date on 4:3:1:3
full.cpo	HH report: up-to-date on varicella
full.hep	HH report: up-to-date on hepatitis B
racemom	mother's race
educ1	mother's educational status
marital	marital status of the mother
m.agegrp	maternal age group
racekid	race of the child
agegrp	age group of the child
sex	sex of the child
frstbrn	first-born status of the child
childnm	# children < 18 years in the HH
incpov1	poverty status
mobil	mobility status
msa	metropolitan statistical area designation
c5	respondent: mother, father, or other

Table 2. Analysis-of-deviance table from stepwise construction of the logistic regression model for response propensity. The dependent variable is the indicator y: having adequate provider-reported vaccination information.

Step		Df	Deviance	AIC
1				41172
2	+ incpov1	-3	-725.6	40452
3	+ shotcard	-1	-474.8	39979
4	+ c5	-3	-271.5	39714
5	+ mobil	-1	-161.4	39555
6	+ msa	-2	-112.8	39446
7	+ full.hep	-2	-71.0	39379
8	+ racemom	-5	-61.4	39327
9	+ childnm	-2	-23.2	39308
10	+ m.agegrp	-2	-19.9	39292
11	+ all4shot	-2	-13.7	39282
12	+ educ1	-3	-16.1	39272
13	+ full.cpo	-2	-11.5	39265
14	+ sex	-1	-3.2	39264

Table 3. Analysis-of-deviance table from stepwise construction of the logistic regression model for predictive probability. The dependent variable is the indicator y: whether a child is up-to-date on the 4:3:1 vaccination series.

Step		Df	Deviance	AIC
1				20803
2	+ all4shot	-2	-590.9	20216
3	+ childnm	-2	-169.0	20052

4	+ agegrp	-2	-160.0	19896
5	+ educ1	-3	-114.6	19787
6	+ full.cpo	-2	-96.3	19695
7	+ full.hep	-2	-72.7	19626
8	+ shotcard	-1	-40.3	19588
9	+ mobil	-1	-35.9	19554
10	+ racemom	-5	-44.3	19519
11	+ frstbrn	-1	-14.7	19507
12	+ m.agegrp	-2	-20.3	19490
13	+ msa	-2	-9.7	19485
14	+ marital	-2	-7.0	19482

Table 4. Illustration of the strategy of selecting the best (k_1, k_2) pair for estimating vaccination coverage with the 4:3:1 series for the Oregon IAP area.

n	(k_1, k_2)	$\hat{\bar{Y}}_n$	$\hat{\Delta}_n$	p_n
2	1,2	76.20	-0.39	0.04
3	1,3	75.75	-0.45	0.00
4	1,4	75.41	-0.34	0.05
5	5,1	75.48	+0.07	0.55

References

[1] Brick, J.M. and Kalton, G. Handling missing data in survey research. *Statistical Methods in Medical Research*, **5**, 215–238 (1996).

[2] Ezzati, T. and Khare, M. Nonresponse adjustments in a national health survey. *1992 Proceedings of the Section on Survey Research Methods*, American Statistical Association, pp. 339-344 (1992).

[3] Smith, P.J., Rao, J.N.K., Battaglia, M.P., Ezzati-Rice, T.M., Daniels, D., and Khare, M. Compensating for nonresponse bias in the National Immunization Survey using response propensities. NCHS Series 2 Report. National Center for Health Statistics, Hyattsville, MD. To appear.

[4] Smith, P.J., Rao, J.N.K., Battaglia, M.P., Daniels, D., Ezzati-Rice,T.M., Khare,M. Compensating for nonresponse bias in the national immunization survey using response propensities. *2000 Proceedings of the Section on Survey Research Methods*, Alexandria, VA: American Statistical Association, To appear.

[5] Potter, F.J., Iannacchione, V.G., Mosher, W.D., Mason, R.E., and Kavee, J.D. Sample design, sampling weights, imputation, and variance estimation in the 1995 National Survey of Family Growth. Vital and Health Statistics. Series 2, Data Evaluation and methods research; no. 124.

[6] Ezzati-Rice, T.M., Zell, E.R., Battaglia, M.P., Ching, P.L.Y.H., and Wright, R.A. The design of the National Immunization Survey. *1995 Proceedings of the Section on Survey Research Methods*, Alexandria, VA: American Statistical Association, pp. 668-672 (1995).

[7] Zell, E.R., Ezzati-Rice, T.M., Battaglia, M.P., and Wright, R.A. National Immunization Survey: The methodology of a vaccination surveillance system. *Public Health Reports*, **115**, 65-77 (2000).

[8] David, M.H., Little, R.J.A., Samuhel, M., and Triest, R. Nonrandom nonresponse models based on the propensity to respond. *1983 Proceedings of the Business and Economic Statistics Section*, Washington, DC: American Statistical Association, pp. 168–173 (1983).

[9] Little, R.J.A. Survey nonresponse adjustments for estimates of means. *International Statistical Review*, **54**, 139–157 (1986).

[10] Eltinge, J.L. and Yansaneh, I.S. Diagnostics for formation of nonresponse adjustment cells, with an application to income nonresponse in the U.S. Consumer Expenditure Survey. *Survey Methodology*, **23**, 33–40 (1997).

[11] Coronado, V.G., Maes, E.F., Rodewald, L.E., Chu, S., Battaglia, M.P., Hoaglin, D.C., Merced, N.L., Yusuf, H., Cordero, J.F., and Orenstein, W.A. Risk factors for underimmunization among 19-35 month-old children in the United States: National Immunization Survey, July 1996-June 1998. Unpublished manuscript, Centers for Disease Control and Prevention, Atlanta (2000).

[12] Akaike, H. Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory* (eds. B.N. Petrov and F. Csáki). Budapest: Akademia Kiado, pp. 267–281 (1973).

[13] McCullagh, P. and Nelder, J.A. *Generalized Linear Models*, second edition, Chapman and Hall, New York, 1989.

[14] Battaglia, M.P., Malec, D.J., Spencer, B.D., Hoaglin, D.C., and Sedransk, J. Adjusting for noncoverage of nontelephone households in the National Immunization Survey. *1995 Proceedings of the Section on Survey Research Methods*, Alexandria, VA: American Statistical Association, pp. 678–683 (1995).

[15] Deming, W.E. *Statistical Adjustment of Data*. New York: Wiley (1943).

[16] Wolter, K. *Introduction to Variance Estimation*. New York: Springer-Verlag (1985).

[17] Cochran, W.G. The planning of observational studies of human populations. *Journal of the Royal Statistical Society, Series A*, **128**, 234–255 (1965).

[18] Rosenbaum, P.R. and Rubin, D.B. Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association*, **79**, 516–524 (1984).