Paper SAS1844-2018

SAS/STAT® 14.3: Roundup: Modern Methods for the Modern Statistician

Maura Stokes and Statistical Staff
SAS Institute Inc.

Abstract

The latest release of SAS/STAT® software has something for everyone. The new CAUSALMED procedure performs causal mediation analysis for observational data, enabling you to obtain unbiased estimates of the direct causal effect. You can now fit compartment models for pharmacokinetic analysis with the NLMIXED and MCMC procedures. In addition, variance estimation by the bootstrap method is available in the survey data analysis procedures, and the PHREG procedure provides cause-specific proportional hazards analysis for competing-risk data. Several other procedures have been enhanced as well. Learn about the latest methods available in SAS/STAT software that can modernize your statistical practice.

Overview

Fall of 2017 brought additional updates to SAS/STAT software. Key updates were made in our current emphasis areas of causal inference, pharmaceutical statistics, survey data analysis, and survival analysis.

Causal Mediation Analysis

All data analysts know not to confuse association with causation. So how do you determine whether a “treatment” contributes to a particular outcome when you aren’t dealing with a randomized experiment or clinical trial? Causal analysis methods were developed to assess causation for observational data. You can pursue the goals of a randomized clinical trial through the use of propensity scores to provide covariate balance with the PSMATCH procedure. You can also directly estimate causal effects for a binary outcome with the CAUSALTRT procedure. in addition, you can now perform causal mediation analysis in SAS.

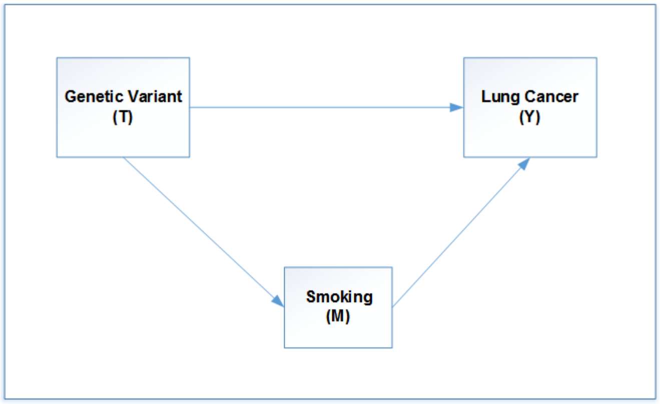
In causal mediation analysis, there are four main variables of interest:

- outcome variable Y
- treatment variable T, which is hypothesized to have direct and indirect causal effects on the outcome variable Y
- mediator variable M, which is hypothesized to be causally affected by the treatment variable T and which itself has a direct effect on the outcome variable Y
- set of pretreatment or background covariates C, which confound the observed relationships among Y, T, and M

The relationships among the first three variables represent the primary causal mediation effects of interest. The set of covariates represent background or pretreatment characteristics that confound the observed relationships among Y, T, and M.

Figure 1 represents the first three main variables in an illustrative causal diagram (see Pearl 2009 for a general theory of causal diagrams).

Figure 1 Mediation Model



In the terminology of causal diagrams, there are two causal pathways of effects from the treatment variable T:

- direct pathway: $T \rightarrow Y$
- mediated or indirect pathway: $T \rightarrow M \rightarrow Y$

Causal mediation analysis quantifies and estimates the total, direct, and indirect (or mediated) effects. It allows causal interpretations of the effects under the assumptions of the counterfactual framework (Robins and Greenland 1992; Pearl 2001).

In practice, one of the biggest issues in causal analysis is the presence of confounding covariates. They complicate the observed relationships among Y, T, and M by introducing extraneous associations that are not due to the direct and indirect pathways in the causal diagram for mediation analysis.

The new CAUSALMED procedure performs causal mediation analysis. It fits generalized linear models that have binary, negative binomial, Poisson, or normal distributions for the outcome and binary or normal distributions for the mediator. PROC CAUSALMED implements the regression approach of VanderWeele (2014), which relies on correct specification of covariate effects in the outcome and mediator models for covariate effect adjustment. The CAUSALMED procedure enables you to specify such covariate effects. You can include interaction effects between the treatment and mediator variables and among the covariates.

PROC CAUSALMED computes the four main causal mediation effects. For continuous outcomes, these effects are computed on the original continuous scale. For binary outcomes, these effects are computed on the odds ratio scale and excess relative risk scale. For definitions of these and other effects, see Valeri and VanderWeele (2013) and VanderWeele (2014).

The following statements demonstrate how you specify an analysis with PROC CAUSALMED:

```
proc causalmed data=Cognitive;
   model    CogPerform  = Encourage Motivation;
   mediator Motivation  = Encourage;
   covar FamSize SocStatus;
run;
```

For more information, see Yung, Lamm, and Zhang (2018).

Pharmacokinetic Analysis

Pharmacokinetics (PK) is a branch of medical research that models the movement of a drug through the body. PK models are nonlinear models that are widely used in the biopharmaceutical industry to predict pharmacokinetic changes in a body system. You can use them to model the drug concentrations by using drug dosage, time elapsed, and other covariates. The concentrations can be addressed with a set of ordinary differential equations (ODE), whose solutions can be further used to model the nonlinear relationship of drug concentration and time. A large class of PK models depend on the compartment scheme, where the drug is kinetically homogeneous in that “compartment”.

Common models are one-, two-, and three compartment models. ODE solutions are known for intravenous (bolus and infusion) and extravascular types of drug administration for these models.

Figure 2 and Figure 3 illustrate one- and two- compartment models for extravascular drug administration.

Figure 2 One-Compartment Model with Absorption Phase

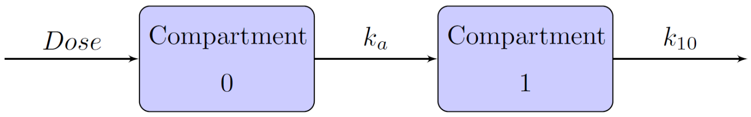
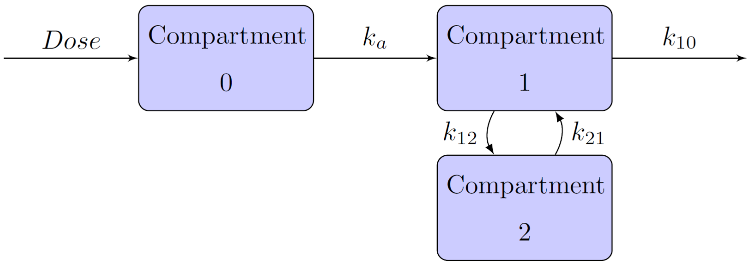


Figure 3 Two-Compartment Model with Absorption Phase



One-, two-, and three-compartment models all have a central compartment and can have one or more peripheral compartments that are linked only to the central compartment but not to each other. After a drug is administered to an administration site, it is distributed to the central compartment and then to peripheral compartments. The rates at which the drug moves from the central compartment to and from the peripheral compartments are characterized by transfer rate constants. The rates at which the drug is eliminated from the central or peripheral compartments are characterized by elimination rate constants.

Structural PK models describe the relationship between concentrations, and time and covariate PK models describe the relationship between the concentrations and other variables. In practice, the final structural model is usually combined with the covariate model.

The NLMIXED procedure now supports the CMPTMODEL statement, which enables you to fit one-, two-, and three compartment models with bolus, infusion, and oral dose adminstration, for a total of nine types of models. It uses an internal ODE solver so that you don't have to provide solutions yourself. The CMPTMODEL statement also supports PK models for multiple dosages and those that have various parameterizations. You can combine the results of the PK model with a PD model.

The MCMC procedure also provides the CMPTMODEL statement, thus offering a Bayesian approach to these models.

For more information, see Kurada and Chen (2018).

Bootstrap Computation for Variances in Survey Data Analysis

The bootstrap method of variance estimation is now available for the SURVEYFREQ, SURVEYIMPUTE, SURVEYMEANS, SURVEYLOGISTIC, SURVEYPHREG, and SURVEYREG procedures. It's an effective method for smooth functions of population means and for some unsmooth functions such as quartiles. It's particular useful for finding confidence intervals directly. However, it can require more computations than the BRR or jackknife method (Lohr 2010).

The bootstrap method can be used for stratified sample designs and for designs that have no stratification. If your design is stratified, the bootstrap method requires at least two PSUs in each stratum. You can provide bootstrap replicate weights for the analysis by using a REPWEIGHTS statement, or the procedure can construct bootstrap replicate weights for the analysis. Note that the naive bootstrap variance estimator is not consistent for complex survey data, which require specific methods similar to that of Rao, Wu, and Yue (1992).

If you do not provide the replicate weights, the procedures construct bootstrap replicate weights by using with-replacement random sampling of PSUs within strata. You can specify the number of bootstrap replicates; the default is 250. You can improve the estimation precision by increasing the number of replicates, but the computation time also increases. In each replicate sample, the original sampling weights of the selected units are adjusted to reflect the full sample. These adjusted weights are the bootstrap replicate weights.

Other updates to the survey procedures in SAS/STAT 14.3 include:

- The new HAZARDRATIO statement in the SURVEYPHREG procedure enables you to request hazard ratios for any variable in the model at customized settings.
- The SURVEYFREQ procedure provides additional agreement statistics.
- The SURVEYIMPUTE procedure computes bootstrap replicate weights. If METHOD=FEFI or METHOD=FHDI is specified, then the bootstrap weights are further adjusted for imputation.
- With the new OUTORDER=RANDOM option in the PROC SURVEYSELECT statement, you can randomly order the selected observations in the output data set.

See *SAS/STAT User's Guide* for further details.

Cause-Specific Proportional Hazards Analysis of Competing-Risks Data

Competing risks arise in studies where individuals are subject to a number of potential failure events and the occurrence of one event might impede the occurrence of other events. For example, after a bone marrow transplant, a patient might experience a relapse, or the patient might die while in remission. For relapse, death is a competing risk because the event of relapse can no longer occur after the patient dies.

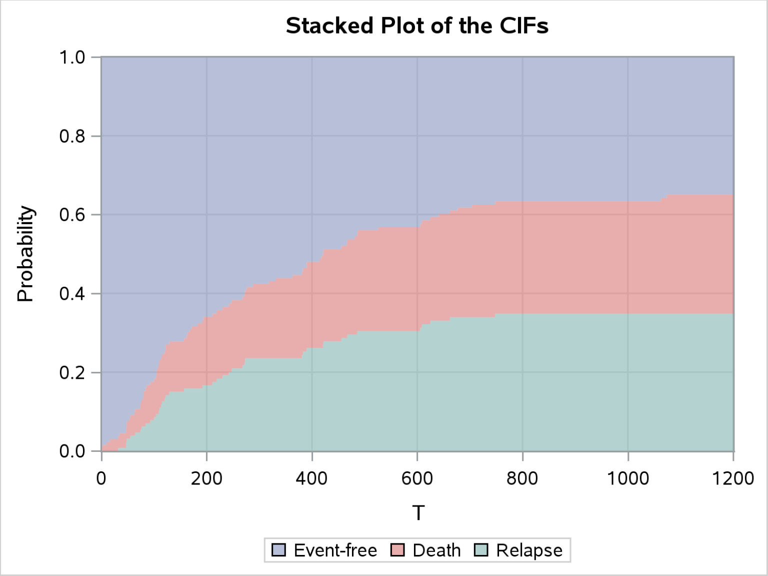
The concepts of a survival function and a hazard function, which form the basis for standard survival analysis, are inadequate for studying competing risks because once a subject experiences an event other than the event of interest, information about the latter can no longer be ascertained reliably. Instead, the analysis of competing risks is based on the analogous concepts of a cumulative incidence function (CIF) and a cause-specific hazard (CSH) function. The CIF, which is defined as the probability subdistribution function of failure from a specific cause, characterizes the occurrence of a cause-specific outcome over time. The CSH function measures the instantaneous rate of failing from a specific cause in the presence of other causes.

The model of Fine and Gray (1999) extends the Cox model to the CIF setting and is often referred to as the proportional subdistribution hazards model. This model was implemented in the PHREG procedure in SAS/STAT 13.1, and you can request it with the EVENTCODE(FG)= option in the MODEL statement; see So, Lin, and Johnston (2015). This approach models the CIF based on the subhazard function.

In SAS/STAT 14.3, the PHREG procedure provides an alternative approach, which performs cause-specific proportional hazards analysis. You can request this approach with the EVENTCODE(COX)= option in the MODEL statemen, and it provides cause-specific analysis of competing-risks data by applying the Cox model to the cause-specific hazard for each event type separately. Estimates of the cumulative incidence function are based on a synthesis of all failure causes and can be obtained with the BASELINE statement.

Figure 4 displays the stacked plot of the CIFs from one such analysis. It effectively shows the progression of failures from different event types over time. The event-free area is the predicted overall survival function, which is the complement of the sum of all the predicted CIFs.

Figure 4 Stacked CIF Plot



For more information, see Guo and So (2018).

Bootstrapped Estimates in the TTEST Procedure

The new BOOTSTRAP statement in the TTEST procedure provides bootstrap standard error, bias estimates, and confidence limits for means and standard deviations in one-sample, paired, and two-sample designs. These estimates can be useful when the usual assumptions such as normality are likely to be violated. The following bootstrap confidence intervals are available:

- bias-corrected percentile intervals
- bootstrap t intervals, which use a traditional standard error estimate and quantiles of the bootstrap distribution of the t statistic
- percentile-based confidence intervals that include a narrowness bias adjustment
- uncorrected percentile-based confidence intervals
- t-based confidence intervals that use the bootstrap standard error estimate
- normal-based confidence intervals that use the bootstrap standard error estimate

In addition to displaying the standard errors of the within-class means and the pooled standard error of the mean difference, the TTEST procedure now also displays the unpooled (Satterthwaite) standard error of the mean class difference. The following statements illustrate how you would request this analysis:

```
data scores;
  input Gender $ Score @@;
datalines;
f 75  f 76  f 80  f 77  f 80  f 77  f 73
m 82  m 80  m 85  m 85  m 78  m 87  m 82
;
proc ttest ci=equal umpu;
   class Gender;
   var Score;
   bootstrap / seed=837;
run;
```

For more detail, see *SAS/STAT User's Guide*.

GAMPL Procedure Supports Tweedie Distribution

The Tweedie distribution is useful for modeling response variables that are continuous for positive values and take the value 0 with a positive probability. For example, in the insurance industry the Tweedie distribution is often assumed for claims, which are positive for customers who have filed claims and 0 for other customers. You can use the Tweedie distribution to model these responses without transformations. Since the Tweedie distribution belongs to the exponential family, you can fit a generalized linear model.

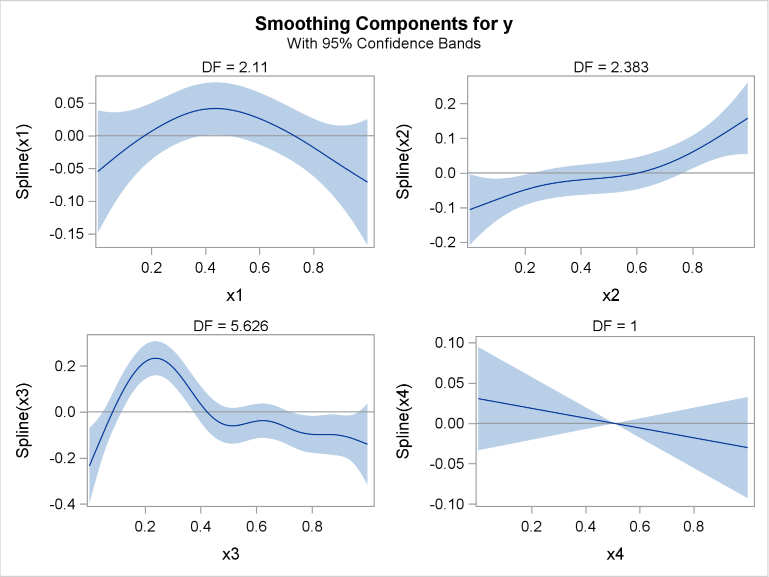
The GENMOD procedure fits Tweedie regression models that incorporate linear effects for the covariates. However, sometimes the assumption of linearity is too restrictive, and you might want to fit a nonparametric regression model with spline effects. This distribution is now available with the GAMPL procedure so that you can explore a nonlinear dependency structure.

Say you fit a model to a response assuming the Tweedie distribution in the GENMOD procedure but were unhappy with the results and wanted to investigate possible nonlinearity. The following PROC GAMPL statements fit a model with splines for four continuous covariates:

```
proc gampl data=one seed=1234 plots;
  model y=spline(x1) spline(x2) spline(x3) spline(x4)/dist=tweedie;
run;
```

The smoothing components in Figure 5 shows that X1–X3 are reasonably smooth and close to true functions, but the plot for X4 displays Bayesian confidence bands that cover the horizontal line at 0. This variable does not contribute to the model and can be removed. This is reinforced with Wald tests for the smoothing components (not shown).

Figure 5 Smoothing Components



See *SAS/STAT User's Guide* for more information.

Common Risk Difference in 2 by 2 Tables

The difference of proportions (risk difference) is a measure of association that has a more natural interpretation than the odds ratio. Besides providing the risk difference for 2 by 2 tables, the FREQ procedure now enables you to request the common risk (stratified) difference for sets of 2 by 2 tables, where the risk difference is the difference between the row 1 proportion and the row 2 proportion, as well as tests. The default is to provide Mantel-Haenszel and summary score estimates of the common risk difference, along with their confidence limits. Another option is minimum risk estimates, confidence limits, and tests. They can improve precision and reduce bias (compared to other weighting strategies) as well as minimize the power loss that can occur when underlying assumptions are not met. The stratified Newcombe estimates and confidence limits are also available.

For more information, see the "Details" section of the FREQ documentation.

The following data come from a study investigating the impact of an experimental treatment on the survival of patients with sepsis (Dmitrienko et al. 2005).

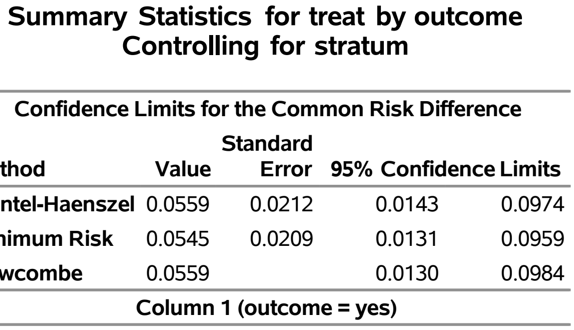
```
data sepsis;
   input stratum treat $ outcome $ count @@;
datalines;
1 exp yes 185 1 exp no 33
1 placebo yes 189 1 placebo no 26
2 exp yes 169 2 exp no 49
2 placebo yes 165 2 placebo no 57
3 exp yes 156 3 exp no 48
3 placebo yes 104 3 placebo no 58
4 exp yes 130 4 exp no 80
4 placebo yes 123 4 placebo no 118
;
```

The COMMONRISKDIFF option requests various common risk differences, confidence limits, and tests.

```
proc freq order=data;
   weight count;
   tables stratum*treat*outcome /
   commonriskdiff(test cl=newcombe cl=mh cl=score cl=minrisk) cmh nocol nopct;
run;
```

Figure 6 displays the common risk differences and confidence limits. These estimates of the average proportion difference are very close.

Figure 6 Common Risk Difference



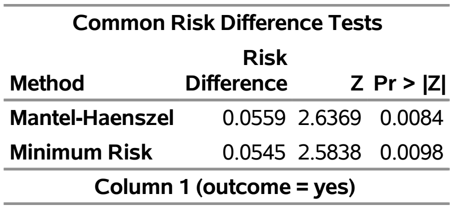
Summary Statistics for treat by outcome
Controlling for stratum

Confidence Limits for the Common Risk Difference				
Method	Value	Standard Error	95% Confidence Limits	
Mantel-Haenszel	0.0559	0.0212	0.0143	0.0974
Minimum Risk	0.0545	0.0209	0.0131	0.0959
Newcombe	0.0559		0.0130	0.0984

Column 1 (outcome = yes)

Figure 7 displays the available tests of the null hypothesis, which confirm what is seen in the confidence limits.

Figure 7 Common Risk Difference Tests

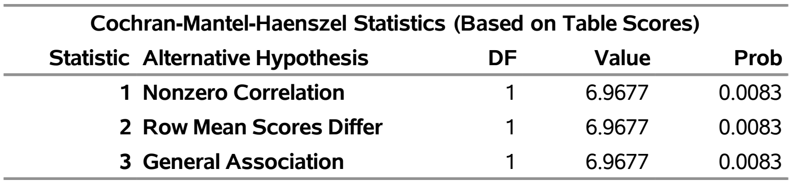


Common Risk Difference Tests			
Method	Risk Difference	Z	Pr > |Z|
Mantel-Haenszel	0.0559	2.6369	0.0084
Minimum Risk	0.0545	2.5838	0.0098

Column 1 (outcome = yes)

The Mantel-Haenszel difference test is asymptotically equivalent to the CMH test, and their p-values are nearly the same.

Figure 8 Mantel Haenszel Tests



Cochran-Mantel-Haenszel Statistics (Based on Table Scores)

Statistic	Alternative Hypothesis	DF	Value	Prob
1	Nonzero Correlation	1	6.9677	0.0083
2	Row Mean Scores Differ	1	6.9677	0.0083
3	General Association	1	6.9677	0.0083

Additional Highlights

Numerous other enhancements are available in this release. The following are some highlights:

- The BCHOICE procedure allows varying weights among choice sets for allocation choice experiments to indicate how many times the allocated percentages of a choice set should be counted.
- The IRT procedure now supports the nominal response model for item analysis.
- The QUANTREG and QUANTSELECT procedures now support fast quantile process regression.
- The MIXED and GLIMMIX procedures now include the TYPE=SP(LEAR) covariance structure for modeling a linear exponent autoregressive covariance, as proposed by Simpson et al. (2010). This can be viewed as an extension of TYPE=ARMA(1,1) to unequally spaced repeated measurements, in the same way that TYPE=SP(EXP) and TYPE=SP(POW) extend TYPE=AR(1).
- The LACKFIT option in the MODEL statement for PROC LOGISTIC now performs the Hosmer-Lemeshow test for polytomous response models and provides more suboptions for controlling the test. Also, the FIRTH option in the MODEL statement performs penalized-likelihood optimization for binary response models for all link functions.
- The HPMIXED procedure now includes the TYPE=TOEP(1) covariance structure, which specifies a Toeplitz structure with one band. This can be useful for specifying the same variance component for several effects.

For more detail, see *SAS/STAT User’s Guide*.

What’s Changed

In general, SAS/STAT does not change its defaults or behaviors. We try to ensure that customer programs run as written. Of course, this is especially important for production jobs. We generally add newer methods with new options, but occasionally, when good statistical practice dictates, we feel obliged to change a default method. Also, we might alter the output as we determine better ways to display results. The “What’s Changed” section is a feature of each What’s New chapter in the documentation and lists any changes by procedure. You should always review this section when you begin working with a new SAS/STAT release.

Conclusions

The latest release of SAS/STAT software provides causal mediation analysis for observational data, easy access to the compartment models of pharmacokinetic analysis, variance estimation by the bootstrap method for survey data analysis, and cause-specific proportional hazards analysis for competing-risk data. Common risk differences and tests are now available for table analysis, and many other enhancements are also included in this release.

Acknowledgements

Thanks to Bob Rodriguez for careful review and input and thanks for Ed Huddleston for careful editing.

Contact Information

Your comments and questions are valued and encouraged. Contact the author:

Maura Stokes
SAS Institute Inc.
SAS Campus Drive
Cary, NC 27513
maura.stokes@sas.com

Version 2.0

REFERENCES

Dmitrienko, A., Molenberghs, G., Chuang-Stein, C., and Offen, W. (2005). *Analysis of Clinical Trials Using SAS: A Practical Guide*. Cary, NC: SAS Institute Inc.

Guo, C., and So, Y. (2018). "Cause-Specific Analysis of Competing Risks Using the PHREG Procedure." In *Proceedings of the SAS Global Forum 2018 Conference*. Cary, NC: SAS Institute Inc.

Kurada, R. R., and Chen, F. (2018). "Fitting Compartment Models Using PROC NLMIXED." In *Proceedings of the SAS Global Forum 2018 Conference*. Cary, NC: SAS Institute Inc.

Lohr, S. L. (2010). *Sampling: Design and Analysis*. 2nd ed. Boston: Brooks/Cole.

Pearl, J. (2001). "Direct and Indirect Effects." In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, edited by J. Breese, and D. Koller, 411–420. San Francisco: Morgan Kaufmann.

Pearl, J. (2009). *Causality: Models, Reasoning, and Inference*. 2nd ed. Cambridge: Cambridge University Press.

Rao, J. N. K., Wu, C. F. J., and Yue, K. (1992). "Some Recent Work on Resampling Methods for Complex Surveys." *Survey Methodology* 18:209–217.

Robins, J. M., and Greenland, S. (1992). "Identifiability and Exchangeability for Direct and Indirect Effects." *Epidemiology* 3:143–155.

So, Y., Lin, G., and Johnston, G. (2015). "Using the PHREG Procedure to Analyze Competing-Risks Data." In *Proceedings of the SAS Global Forum 2015 Conference*. Cary, NC: SAS Institute Inc. http://support.sas.com/resources/papers/proceedings15/SAS1855-2015.pdf.

Valeri, L., and VanderWeele, T. J. (2013). "Mediation Analysis Allowing for Exposure-Mediator Interactions and Causal Interpretation: Theoretical Assumptions and Implementation with SAS and SPSS Macros." *Psychological Methods* 18:137–150.

VanderWeele, T. J. (2014). "A Unification of Mediation and Interaction: A 4-Way Decomposition." *Epidemiology* 25:749–761.

Yung, Y.-F., Lamm, M., and Zhang, W. (2018). "Causal Mediation Analysis with the CAUSALMED Procedure." In *Proceedings of the SAS Global Forum 2018 Conference*. Cary, NC: SAS Institute Inc.