Childhood Cancer Survivor Study Analysis Concept Proposal

1. Title:
Inverse probability censored weighting (IPCW) to adjust for selection bias and drop out in the context of CCSS analyses

2. Working group and investigators:

Epidemiology/Biostatistics Working Group

Chongzhi Di	cdi@fhcrc.org	206-667-2093
Wendy Leisenring	wleisenr@fhcrc.org	206-667-4374
Kayla Stratton	kstratton@fhcrc.org	206-667-7293
Toana Kawashima	tkawashi@fhcrc.org	206-667-7697
Kiri Ness	kiri.ness@stjude.org	901-595-5157
Yutaka Yasui	yyasui@ualberta.ca	780-492-4220
Greg Armstrong	greg.armstrong@stjude.org	901-495-5892
Ann Mertens	ann.mertens@choa.org	404-785-0691
Aaron McDonald	aaron.mcdonald@stjude.org	
Les Robison	les.robison@stjude.org	901-495-5817

3. Background and rationale

The Childhood Cancer Survivor Study (CCSS) queried participants with a baseline and several follow-up (we focus on Follow-up 2000, 2003 and 2007 here) questionnaires. As in most longitudinal cohort studies, there is non-participation at each follow-up questionnaire. This may be problematic as differential non-participation, by either an exposure of interest or a potential confounder of the association between the exposure and the outcome of interest, has the potential to introduce bias into the estimate of the association between the exposure and the outcome. That is, it is unlikely that the respondent cohort members at a given questionnaire, those who remained in the cohort and who completed the questionnaire, are a representative sample of the original population of interest.

Most analyses in the CCSS have simply omitted individuals who did not complete a particular questionnaire, citing differential non-participation as a potential limitation of the study. Investigators have indicated that differential non-participation may generate biased results, but have not quantified this potential bias. Because the cohort continues to age, and because participation declines gradually over time, it is important to quantify this potential bias so that we have a consistent mechanism to examine bias when conducting analyses and writing manuscripts. This concept proposes an investigation to examine potential bias due to non-participation in the CCSS.

4. Goal

We want to understand which characteristics of survivors are associated with non-participation and to what degree non-participation at follow-up questionnaires are influencing analyses (i.e. associations between risk factors and outcomes). If there is an impact, we want to adjust for this type of selection bias in future analysis, if possible.

5. Analysis framework

We plan to conduct analyses using inverse probability censored weighting (IPCW) methodology. This method can be used for most of the analyses outlined below. When the outcome of interest is from a follow-up questionnaire and the exposures (risk factors) are from baseline (e.g., diabetes at FU 2003 vs. BMI at baseline), we may also explore the augmented IPCW (AIPCW) method. When the data are available, the latter methodology can provide asymptotically more efficient estimates than the former.

Analyses for three follow-ups will be conducted separately, each compared to the baseline population. In the following, we use FU2000 as an example, and methods for FU2003 and FU2007 are the same.

IPCW: Inversely weight regression analyses by the probability of participation (determined based on a logistic regression model for probability of participation given past history covariates and outcomes), effectively inflates the impact of underrepresented subjects, so we can observe associations that would have been observed if all subjects had stayed in the study, assuming the models are correctly specified. Key references for this methodology are:

Robins JM, Rotnitzky A, Zhao LP. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. Journal of the American Statistical Association 1995; 90:106–121.

Robins JM. Marginal structural models versus structural nested models as tools for causal inference. In Statistical Models in Epidemiology: The Environment and Clinical Trials, Halloran ME, Berry D (eds). IMA Volume 116, Springer-Verlag: New York, 1999; 95–134.

Robins JM, Hernan MA, Brumback B. Marginal structural models and causal inference in epidemiology. Epidemiology 2000; 11:550 –560.

Mathematically,the models are represented as follows:

> *Let Y_i be the outcome at FU1, X_i be covariates (FU1 and/or baseline) for the outcome model, Z_i be covariates (baseline) for the missing mechanism model, R_i be the missing indicator (R_i=1 when Y_i is observed and R_i=0 if Y_i is missing).*

Then the probability of missingness at FU1 can be modeled using logistic regression

$$\pi_i = P(R_i = 1 \mid Z_i) = \frac{\exp(\gamma_0 + Z_i^t \gamma)}{1 + \exp(\gamma_0 + Z_i^t \gamma)} \ ,$$

where the outcome is assumed to follow a generalized linear model

$$\mu_i = E(Y_i \mid X_i) = g^{-1}(\beta_0 + X_i^t \beta) \, ,$$

g is a link function, for instance identity for continuous outcome and logit for binary outcome. The IPCW estimation is equivalent to solving the following estimating equations

$$\sum_{i=1}^{n} \frac{R_i}{\pi_i} \frac{\partial \mu_i^{\ t}}{\partial \beta} V_i^{-1} (Y_i - \mu_i) = 0 \, , \textit{where } V_i = \mathrm{var}(Y_i),$$

which uses information from complete cases only, and ignores information from incomplete cases. The standard errors for regression coefficients need to be corrected by the sandwich variance estimators.

In IPCW analysis, assuming the logistic model for R (i.e., factors influence log odds of R linearly) could be questioned as it may or may not approximate the underlying missingness mechanism. Therefore, we will conduct some sensitivity analysis with respect to the logistic model for missingness. We will also consider the possibility of using nonparametric models if needed.

Augmented IPCW (AIPCW): To improve efficiency, the augmented IPCW method (see 4-5) adds an augmentation term to the IPCW estimating equation. This method introduces the augmentation term for non-participants, in contrast to IPCW, which ignores observations from non-participants. However, this method requires that covariates for outcome model X_i are available for both respondents and non-respondents. Thus, Augmented IPCW is applicable in CCSS only if risk factors are observed at baseline.

The mathematical formula for AIPCW estimating equation is given by

$$\sum_{i=1}^{n} \frac{R_i}{\pi_i} \frac{\partial \mu_i^{\ t}}{\partial \beta} V_i^{-1} (Y_i - \mu_i) + \sum_{i=1}^{n} (1 - \frac{R_i}{\pi_i}) h(X_i) = 0$$

where h is a function of X_i. The augmented term incorporates information from incomplete cases, and thus potentially improves efficiency compared to IPCW estimates. The gain in efficiency depends on the choice of the function h. In practice, one can choose some simple functional forms for h (see 4-5).

In addition to possible efficiency gains, another advantage of AIPCW is that it is doubly robust, in the sense that it yields consistent results if either the missingness mechanism or the outcome regression model is correctly specified.

Algorithm for implementation:

1. **Calculate probability weights:** Among subjects who participated at baseline, and who were eligible (i.e alive and eligible) for the relevant FU questionnaire, fit a logistic regression model to predict participation at that questionnaire, using covariates from baseline that are key predictors of participation (**Z – define these as the set we've currently identified in our participation models)** and those that are any baseline versions of the current outcome (**D**) and risk factor of interest (**E**) (if the exact questions are not available, we try to use any similar information available). Results of this modeling process will be summarized to describe factors associated with participation at each questionnaire. Calculate predicted probabilities from this model – call these **P.**

2. **Calculate stabilized versions of probability weights:** A stabilizing calculation involves fitting the same prediction model as above, but only with **E** as the covariate. Calculate predicted probabilities from this model – call these **S.** Then calculate stabilized weights = **SW = S/P**.

3. **Fit weighted logistic regression using 1/P as weights:** Among subjects who responded to FU questionnaire of interest, fit a logistic regression model with **D** as the outcome, and with **E** as the covariate of interest and any other appropriate adjustment covariates for that outcome (would probably do both univariate – w/ **E** as well as multivariate). A sampling weight of **1/P** should be incorporated in this analysis (double check whether SAS automatically takes the inverse of **P** when using weights, or whether you need to give it **1/P**. Use robust variances.

4. **Fit weighted logistic regression using stabilized weights:** Fit the same model as above with stabilized weights, **SW = S/P,** instead of 1/P. Correct standard errors by sandwich estimators.

5. **Fit an unweighted version of the logistic regression for comparison.** Use the same model(s) as above in (3), but without any sampling weights.

Report ORs for E for the three models, 1) weighted w/ P, 2) weighted with SW and 3) unweighted.

Proposed Outcome / Risk factor combinations to look at, along with relevant covariates for probability weight prediction model.

			Covariates for:	
Questionnaire	**Outcome (D)**	**Key Risk Factor (E)**	**Probability weight Model (1,2)**	**Association Models (3, 4, 5, 6)**
FU 2003	Diabetes*	BMI (FU 2003)	BMI(base), Diabetes(base), other RF from participation model (**Z**)	BMI(FU 2003), other RF from participation model (**X**)
FU 2007	Diabetes*	BMI (FU 2007)	BMI(base), Diabetes(base), other RF from participation model (**Z**)	BMI(FU 2007), other RF from participation model (**X**)
FU 2003	Pain (E21)	BSI (scored from FU 2003 – G) – three subscales – depression, somatization, anxiety (dichotomized)	BSI (base - J16 – J37), RF from participation model (**Z**)	BSI (FU 2003), RF from participation model (**X**)
FU 2003	SF-36 Physical function and General health subscales	Age (FU 2003)	Age (Base), Health at base (N15), RF from participation model (**Z**)	Age (FU 2003), RF from participation model (**X**)

***** Use definition used by Yutaka's group

** We will also consider analyses with treatment as a covariate. However, the use of treatment is somewhat problematic since there is a different missing data situation there (missing covariates, in contrast to missing outcomes in other proposed analyses). Yutaka Yasui's group is working on missing treatment issue using a multiple imputation methodology and as that data becomes available, we will evaluate incorporating the multiply imputed data into some treatment related hypotheses.

Table 1: Status at each follow-up

	Baseline	FU 2000	FU 2003	FU 2007
Participant				
Non-Participant				
Dead				
Ineligible				
Total				

Table 2 Distribution of covariates at each follow-up or for MRAF availability

	Baseline eligible		FU 2000 eligible		FU 2003 eligible		FU 2007 eligible	
	Participants	Non-participants	Participants	Non-participants	Participants	Non-participants	Participants	Non-participants
Gender								
Race								
Baseline age								
Diagnosis								
Age of diagnosis								
…								

Table 3: ORs between participation and subject characteristics

	Baseline OR (95% CI)	FU 2000 OR (95% CI)	FU 2003 OR (95% CI)	FU 2007 OR (95% CI)
Gender				
Race				
Baseline age				
Diagnosis				
Age of diagnosis				
…				

Table 4: ORs between outcomes and key risk factors, unadjusted and adjusted

Questionnaire	Outcome	Key risk factor	ORs (95% CIs)		
			Naïve	IPCW	IPCW (stabilized)
FU 2003	Diabetes	BMI			
FU 2003	Diabetes	BMI (base) ?			
FU 2003	Pain	BSI			
FU 2003	Pain	BSI (base) ?			
FU 2003	SF-36	Age			
…					

6. References

1. Robins JM, Rotnitzky A, Zhao LP. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. Journal of the American Statistical Association 1995; 90:106–121.

2. Robins JM. Marginal structural models versus structural nested models as tools for causal inference. In Statistical Models in Epidemiology: The Environment and Clinical Trials, Halloran ME, Berry D (eds). IMA Volume 116, Springer-Verlag: New York, 1999; 95–134.

2. Robins JM, Hernan MA, Brumback B. Marginal structural models and causal inference in epidemiology. Epidemiology 2000; 11:550 –560.

3. Scharfstein DO, Rotnitzky A, Robins JM. (1999). Adjusting for non-ignorable drop-out using semiparametric non-response models. Journal of the American Statistical Association, 94:1096-1120.

4. Rotnitzky A, Robins JM, Scharfstein D. (1999). Semiparametric regression for repeated outcomes with nonignorable nonresponse. Journal of the American Statstical Association, 93(444):1321-1339.