On the Sampling Strategy and Absolute Risk Estimation of Nested Case Control Studies

Hongying Li * Ruth Patterson† Loki Natarajan ‡

Abstract

Nested case-control studies are often used in large epidemiology studies as it provides us an efficient means to study associations between a risk factor and a disease outcome at greatly reduced cost. A small subset of controls are sampled from the prospective full cohort for each case. The information for the risk factor (like a biomarker) is only needed to be collected on a much smaller cohort including the selected controls and the cases. However, choice of control sampling strategies can affect study power and possibly even bias risk estimates, and hence needs to be considered carefully. The well-characterized Womens Healthy Eating and Living (WHEL) cohort of over 3000 breast cancer survivors provides us a unique opportunity to investigate the impact of a variety of control sampling strategies on inference in nested case-control studies. In this paper, we conducted an in depth study to evaluate three case-control matching schemes. The hazard ratio estimates for associations between biomarkers of interest and breast cancer death will be estimated based on the nested case control studies and compared with results from the full cohort models, which are deemed to be the gold standard for this analysis. The analytic approach for estimating absolute risk in the nested case-control setting will also be described. Especially, we will describe the methods for the situation of the survival study with left truncation. The results from simulation biomarkers and a real biomarker (e.g. CRP) assayed on the WHEL cohort will be presented.

Key Words: nested case control, matching strategy, sampling variability, absolute risk

1. Introduction

Nested case-control studies provide us a way to study the association of a risk factor with a specific disease with greatly reduced cost. It has been proven to be a valid and unbiased method for estimating the association if the sampling scheme for the controls is appropriate ([Langholz 2009]). When choosing controls for a case in the nested case-control design, there are two things that need consideration. One is the case control ratio. Usually the case control ratio of one to one or two is good. More controls beyond 4 or 5 do not improve the efficiency much ([Ury 1975]). Another is the matching scheme, i.e. if you want to match on covariates or which covariates to match on when choosing controls for each case. For a case who has an event at time t, in a simplest way, a control can be randomly chosen from all subjects who have not dropped out and who are event-free at time t. In this way controls are matched to cases only on follow-up time. It is also common to match controls to cases on other variables that might be related to outcomes. However, if controls are matched to cases on a particular covariate Z, then the log-hazard ratio for this covariate cannot be estimated ([Langholz 2009]). As noted previously choice of matching covariates can affect study power and possibly even bias risk

*Moores UCSD Cancer Center #0901, University of California, La Jolla, CA 92093

†Dept. of Family and Preventive Medicine, Moores UCSD Cancer Center #0901, University of California, La Jolla, CA 92093

‡Dept. of Family and Preventive Medicine, Moores UCSD Cancer Center #0901, University of California, La Jolla, CA 92093

estimates, and hence needs to be considered carefully ([Wacholder et-al. 1992 III]). In this paper we will investigate the sampling variability and bias associated with several control selection schemes for the nested case control studies within a large cohort - WHEL (Women's Healthy Eating and Living Trial)([Pierce et-al. 2007]). We will compare the results from several nested case control designs with the full cohort analysis. Our findings will elucidate methods for design and risk estimation in nested case-control studies of biomarker and cancer outcomes.

We first review the method for estimating risk parameters and absolute risk in full cohort and nested case control cohort in section 2.1 and 2.2. Especially we will show that the method can be easily adapted in the study with left truncation. The case control sampling strategies are described in section 3.3. The results are presented in section 4 and then followed by the discussion in section 5.

2. Theory

2.1 Estimation of Risk Parameters

We first review the method to estimate the risk parameters in a standard Cox proportional hazard model. The foundation is the hazard function that will be used to construct a partial likelihood. For a proportional hazard model, the hazard function at time t for subject i is,

$$h(t; X_i, Z_i) = h_0(t) \exp(\beta X_i + \alpha Z_i) \tag{2.1}$$

where X_i is the exposure of interest (e.g. biomarker CRP) for subject i, Z_i is a vector of potentially confounding covariates that need to be adjusted for, β and α are, respectively, log-hazard ratios associated with the main exposure and covariates, and $h_0(t)$ is an unspecified baseline hazard function, denoting the hazard of an event at time t with all covariate values set equal to 0. The objective is to estimate the log-hazard ratio associated with each covariate as a measure of disease risk. The hazard function is then used to write down a partial likelihood function. Assume there are D distinct event times and there are no ties among them. If subject i experiences an event at time t_i, the probability is expressed as:

$$L_i = \frac{\exp(\beta X_i + \alpha Z_i)}{\sum_{j \in R(t_i)} \exp(\beta X_j + \alpha Z_j)} \tag{2.2}$$

where the summation in the denominator is over all j in the risk set $R(t_i)$. Baseline hazard $h_0(t)$ does not appear as it gets cancelled out in the numerator and denominator. For the standard Cox model, $R(t_i)$ is composed of all subjects that are still under observation and have not experienced the event at time t_i ([Kalfleisch et-al. 2002]). When there is left truncation in the study, the counting time starts from sometime earlier than the actual study entry time. The risk set at time t_i is adjusted and composed of all subjects that already enter the study before time t_i in addition to the requirement that they are still under observation and have not experienced the event at time t_i. The product of the terms L_i over all D individuals who experience the event is the partial likelihood, which is then maximized to obtain estimates of β and α. Hazard ratio estimation can be easily obtained once β and α are estimated.

In nested case control studies, the exposure may be measured only on the cases and selected controls. The estimation of β and α will be based on the partial

likelihood on this subset:

$$L = \prod \frac{\exp(\beta X_i + \alpha Z_i)}{\sum_{j \in \tilde{R}(t_i)} \exp(\beta X_j + \alpha Z_j)} \tag{2.3}$$

Here the product is over all cases. $\tilde{R}(t_i)$ is the risk set at t_i, including the case at t_i and the randomly selected controls for this case. It was showed that under some quite general conditions, the maximum likelihood estimations $\hat{\beta}$ and $\hat{\alpha}$ from equation (2.3) are asymptotically consistent estimators of β and α just like in the standard partial likelihood method ([Borgan et-al. 1995]).

2.2 Estimation of Absolute Risk

Other than the estimation of hazard ratio for the risk factor, absolute risk (i.e., probability of experiencing an event over a specified time-interval) often has clinical importance as well. After $\hat{\beta}$ and $\hat{\alpha}$ are estimated, the baseline cumulative hazard function is estimated using the Breslow estimator:

$$\begin{aligned} \hat{H}_0(t; X_i, Z_i) &= \sum_{t_i \le t} \hat{h}_0(t_i; X_i, Z_i) \\ &= \sum_{t_i \le t} \frac{1}{\sum_{j \in R(t_i)} \exp(\hat{\beta} X_j + \hat{\alpha} Z_j)} \end{aligned} \tag{2.4}$$

Estimates of absolute risk now can be mathematically derived as follows:

$$\hat{A}(t; X_i, Z_i) = 1 - \exp[-\hat{H}_0(t; X_i, Z_i) \exp(\hat{\beta} X_i + \hat{\alpha} Z_i)] \tag{2.5}$$

where $\hat{A}(t; X_i, Z_i)$ is the absolute risk, i.e., the probability of the event occurring before time t; X_i is the primary exposure for subject i(e.g., inflammatory marker CRP level); Z_i is a vector of other covariates([Kalfleisch et-al. 2002]).

[Langholz et-al. 1997] extended the absolute risk estimation method to a nested case control study. The calculation relies on the use of the full cohort information to adjust the estimation of the baseline hazard. The first step of the calculation entails estimating the baseline hazard as

$$h_0(t_i) = \frac{1}{\sum_{j \in \tilde{R}(t_i)} \exp(\beta X_j + \alpha Z_j + \log r_j)}, \tag{2.6}$$

where $\tilde{R}(t_i)$ is the sampled case-control set at time t_i; $r_j = \frac{n_j}{m}$, n_j is the number of subjects in the risk set on the full cohort; and m= size of the case-control set ($m = 2$ for 1 to 1 case control ratio). As before X_j represents biomarker exposure, Z_j is a vector of other covariates, and β and α are, respectively, the associated log-hazard ratios. For stratified sampling, n_j should be the number of subjects in the risk set on the same strata with the case in the full cohort.

Estimates of β and α from fitting the conditional logistic regression model in the case control analysis are substituted into equation 2.6. Having thus obtained the baseline hazard, the absolute risk of disease up to time T is computed the same way as in the full likelihood method (equations 2.4 and 2.5).

3. Materials and Methods

3.1 Study Sample

The WHEL Study was a randomized dietary intervention trial which examined the impact of a high fruit/vegetable/fiber and low-fat diet on breast cancer recurrence and mortality on 3088 breast cancer survivors. It randomized 3088 women within 4 years of breast cancer diagnosis from 1995 to 2000 to either a dietary intervention arm or a control arm. After an average 7.3 years of follow-up, breast cancer recurrence rates in the two study arms were 17% with no difference in recurrence or mortality hazard ratios ([Pierce et-al. 2007]). The WHEL Study provides us a rich database of early breast cancer survivors with demographic, clinical, lifestyle (diet, physical activity), and biomarker measures. It affords a unique opportunity to compare a full cohort versus a nested case-control analysis.

We considered 904 subjects who were overweight/obsese, postmenopausal and had CRP (c-reactive protein, an inflammatory plasma biomarker) measured at the time of randomization from the WHEL study as the full cohort. There were totally 115 breast cancer related death events during follow-up in this subset. We want to estimate the hazard ratio associated with breast cancer related death for CRP, and ultimately estimate the absolute risk using covariates.

3.2 Simulated Biomarkers

Two biomarkers were simulated. Bio1 was simulated to be a strong risk factor for breast cancer related death that was independent of other factors we considered in the multivariate model (e.g. tumor grade, stage). The univariate hazard ratio associated with 1-unit increase of bio1 for breast cancer mortality was 1.11 with 95% confidence interval 1.08 to 1.14 (the adjusted hazard ratio was similar to this). Bio2 was simulated to be a risk factor that had a moderate correlation with tumor stage (spearman correlation=0.5). The univariate hazard ratio was 1.12 with 95% CI (1.08, 1.16). After adjusting for tumor grade, stage, anti-estrogen use, age at diagnosis, time from diagnosis to study entry, the hazard ratio was 1.04 with 95% CI (1.00, 1.10). The simulated biomarkers as well as the log-transformation of baseline CRP will be used to investigate the various case control designs.

3.3 Control Sampling Schemes

Below we describe the various control sampling methods that we will investigate to study the matching variability and estimation bias. The hazard ratio for the biomarker will be estimated from the model that is adjusted for tumor grade, stage, antiestrogen use, age at diagnosis and time from diagnosis to randomization (if they are not matched on) using either full cohort or sampled case control cohorts.

The results from the following control sampling schemes will be compared with the full cohort analysis:

- Design I: simple random sampling where controls are individually matched to cases on follow-up time (controls have a longer follow-up time starting from the diagnosis than their matched cases and also enter the study before the case developed the event);
- Design II: stratified sampling — in addition to the requirement in Design I, controls are also needed to exactly match on the cases' primary tumor stage (stage I, IIA, IIB, IIIA or IIIC).

- Design III: nearest neighborhood matching — controls are exactly matched to cases on primary tumor stage (stage I, IIA, IIB, IIIA or IIIC), closely matched on age at diagnosis, date of diagnosis and date of randomization (by a smallest total distance) in addition to follow-up time.

For the full cohort analysis which will be the gold standard that we want to compare the results from the case-control analysis with, a left truncation survival model will be used as subjects enter the study only some time after they were diagnosed with breast cancer and not yet recurred or died. For Design I and II, we will repeat the case-control matching 1000 times thus we have 1000 such case-control matched data sets. The estimated hazard ratio for a biomarker was calculated for each of the 1000 case-control matched sets. The median along with 2.5% and 97.5% percentiles were reported in section 4 as the estimated hazard ratio and its 95% confidence interval. For Design III, since we matched on age of diagnosis, date of diagnosis and date of randomization by the smallest total distance, there was only one best control for each case so there was no variability associated with the control selection. Stage will not be adjusted in the analysis for Design II and III since it is exactly matched on and can't be estimated.

4. Result

The analysis outlined in section 3.3 was carried out in the subset of 904 subjects from the WHEL Study. Figure 1 showed the boxplots of the estimated hazard ratios from 1000 case control realizations for Design I and II. Design III only had one realization and the results were plotted as a point. The estimated hazard ratios for log(CRP) and simulated biomarkers under various sampling schemes were also summarized in Table 1. It is known that the estimation from the simple random sampling case control design (Design I) is unbiased([Langholz 2009]). But the median estimation seems to be biased upward for bio1, especially under Design I. Also we can see that there are still large variations associated with the control sampling. We might get results that are very different from the estimated value from full cohort in one particular case control matching. The hazard ratio for 1-unit increase in log(CRP) could be as small as 0.86 and as large as 2.46 for Design I from the 1000 realizations, which would lead to contradictory conclusions. The variation associated with repeated control sampling was smaller for Design II comparing with Design I which was reasonable as Design II restricted the controls to those that matched on the case's tumor stage and thus was a special case of Design I. Design II also had a higher proportion of estimations that were within the 95% confidence interval from the full cohort analysis than Design I (Table 1). Design II performs consistently better than Design I for all three markers. Among the three designs, Design III had the smallest variance for the estimated hazard ratio (Table 1). While the HR estimation for simulated biomarkers bio1 and bio2 from Design III seems unbiased, the estimation for log(CRP) is toward null indicating some extent of over matching. This indicates that a good design for one biomarker is not necessarily so for another even in the same cohort.

The estimated 5 year survival probability from diagnosis based on full cohort and nested case control cohorts were estimated as in section 2.2. The results for log(CRP) were shown in Figure 2. Notably, the median estimation from the nested case control cohorts align nicely on the diagonal line for both Design I and II, indicating they are very close to the estimation from the full cohort. The variation associated with random control sampling for Design II is much less than that with

Design I. For Design I, we see that there is a large chance for over estimating the risk especially for high-risk patients indicated by the low 2.5% percentile survival probability from the case control cohorts. For Design III, the absolute risk estimations from the nested case control cohort do not approximate the full cohort estimation well. This finding is consistent with the results reported by [Ganna et-al. 2012] where serious deviations of absolute risk estimation for the fine matched nested case-control design were observed. The results for the simulated biomarkers bio1 and bio2 are similar (plots not shown).

5. Discussion

Case control designs are often used to reduce cost when examining a risk factor (like a biomarker) with a specific disease in large epidemiology studies. Beyond the simple random sampling, there are various novel sampling designs proposed in the literature,like counter-matching, quota-matching ([Langholz2007]). Especially, 'designing a study that is well suited to the particular study situation remains an art' ([Langholz2007]).

CRP is a biomarker for inflammation found in blood plasma. There is some evidence showing that elevated CRP levels at time of breast cancer diagnosis is associated with reduced overall and disease-free survival and with increased risk of death from breast cancer ([Allin et-al.2011]). In this study, CRP is measured for 904 overweight/obsese and postmenopausal breast cancer survivors from the WHEL study, at the time of study entry which is averagely two years later after diagnosis. To estimate the risk, a left truncation model is used since there is a delay for study entry after diagnosis and only patients who are still alive and not recurred can be enrolled. In this group of post menopausal and overweight/obsese patients, after adjusting for known prognosis factors and potential confounders (tumor grade, stage, anti estrogen use, age of diagnosis and time from diagnosis to study entry), the hazard ratio for one unit increase of log(CRP) (log(ng/mL)) for breast cancer related death is 1.19 (p-value=0.07). This is consistent with the literature indicating elevated level of CRP at baseline might be associated with higher risk of breast cancer related death [Allin et-al.2011].

With the complete covariate information for the full cohort, we were able to compare nested case control designs with the full cohort analysis. Three case control designs were examined in this study, namely, simple random sampling (Design I), stratified sampling (Design II) and close matching (Design III). The variability associated with repeated control sampling were investigated by fixing the cases and repeatedly sampling the controls for 1000 times if possible. Design II which randomly selected controls for each case matching on tumor stage effectively reduced the sampling variability comparing with Design I and gave unbiased estimation of the risk parameter. Correspondingly, the absolute risk estimation from Design II also had much smaller variability than Design I. Tumor stage is a known prognostic factor for breast cancer recurrence and death. The analysis using simulated biomarkers showed that matching on stage is a better strategy than simple random sampling when designing the case-control study. Design III is a close matching strategy where there is one best control for each case. While for the simulated biomarkers, the estimation for the risk parameter seemed satisfactory under Design III, the estimated risk for the real biomarker log(CRP) is towards null. This may imply some over-matching. One major drawback with Design III is that the absolute risk estimation is seriously biased. This may be due to that there is just one

best control for each case in Design III so the sampling is determined. The method for how to weight the denominator in equation 2.6 in Design I and II does not apply to Design III well. So overall, Design II seems to be the best choice. This study used the left truncation model for both the full cohort analysis and the nested case control analysis. The method for the standard Cox proportional regression model generalized easily to the left truncation situation with an adjustment of the definition of the risk set at each time point. With a full cohort, simulating a biomarker and then using it to evaluate the various designs is a useful strategy than can give you some insights about how the different case control designs perform in this particular study. Close-matching should be used with caution because of the greater potential of over-matching and the bias for absolute risk estimation.

References

[Borgan et-al. 1995] Borgan, Ø., L. Goldstein, and B. Langholz. (1995). Methods for the analysis of sampled cohort data in the Cox proportional hazards model. *The Annals of Statistics* : 1749-1778.

[Langholz et-al. 1997] Langholz, Bryan, and Ø. Borgan. (1997). Estimation of absolute risk from nested case-control data. *Biometrics* 53, no. 2: 767-774.

[Langholz2007] Langholz, B. (2007). Use of Cohort Information in the Design and Analysis of Case Control Studies. *Scandinavian Journal of Statistics*, 34(1), 120-136.

[Langholz 2009] Langholz, Bryan, and David Richardson. (2009). Are nested case-control studies biased?. *Epidemiology* 20.3: 321-329.

[Wacholder et-al. 1992 II] Wacholder, Sholom, et al. (1992). Selection of controls in case-control studies: II. Types of controls. *American journal of epidemiology* 135.9: 1029-1041.

[Wacholder et-al. 1992 III] Wacholder, S., Silverman, D. T., McLaughlin, J. K., & Mandel, J. S. (1992). Selection of controls in case-control studies: III. Design options. *American Journal of Epidemiology*, 135(9), 1042-1050.

[Ury 1975] Ury HK. (1975). Efficiency of case-control studies with multiple controls per case: continuous or dichotomousdata. *Biometrics*, 31:643-9.

[Pierce et-al. 2007] Pierce, John P., Loki Natarajan, Bette J. Caan, Barbara A. Parker, E. Robert Greenberg, Shirley W. Flatt, Cheryl L. Rock et al. (2007). Influence of a diet very high in vegetables, fruit, and fiber and low in fat on prognosis following treatment for breast cancer: the Women's Healthy Eating and Living (WHEL) randomized trial. *Jama* 298, no. 3: 289-298.

[Kalfleisch et-al. 2002] Kalbfleisch, J. D., & Ross, L. Prentice. (2002). The statistical analysis of failure time data.

[Ganna et-al. 2012] Ganna, A., M. Reilly, et al. (2012). Risk prediction measures for case-cohort and nested case-control designs: an application to cardiovascular disease. *American journal of epidemiology*, 175(7), 715-724.

[Allin et-al.2011] Allin, K. H., Nordestgaard, B. G., Flyger, H., & Bojesen, S. E. (2011). Elevated pre-treatment levels of plasma C-reactive protein are associated with poor prognosis after breast cancer: a cohort study. *Breast Cancer Res*, 13(3), R55.

Table 1: Estimated Hazard Ratio for biomarkers with 95% CI under the full cohort and various case control design schemes

	HR	95% CI[1]	VarRatio[2]	within[3]
logCRP				
Full	1.19	(0.99, 1.43)	1.00	
Design I	1.27	(0.98, 1.70)	3.22	79%
Design II	1.21	(0.99, 1.51)	2.22	92%
Design III	1.02	(0.78, 1.33)	2.00	
bio1				
Full	1.10	(1.07, 1.14)	1.00	
Design I	1.14	(1.08, 1.24)	6.96	41.5%
Design II	1.12	(1.08, 1.19)	3.91	62.2%
Design III	1.12	(1.06, 1.19)	3.65	
bio2				
Full	1.04	(1.00, 1.10)	1.00	
Design I	1.05	(0.98, 1.13)	2.76	80.8%
Design II	1.05	(0.99, 1.11)	2.76	89.4%
Design III	1.04	(0.97, 1.13)	2.48	

[1]: 2.5% and 97.5% quantiles from 1000 realizations are reported for Design I and II.
[2]: the median ratio of estimated variance for log(HR) from 1000 NCCs over that from full cohort for Design I and II.
[3]: proportion of HR estimations from 1000 realizations that are within the 95% CI from full cohort

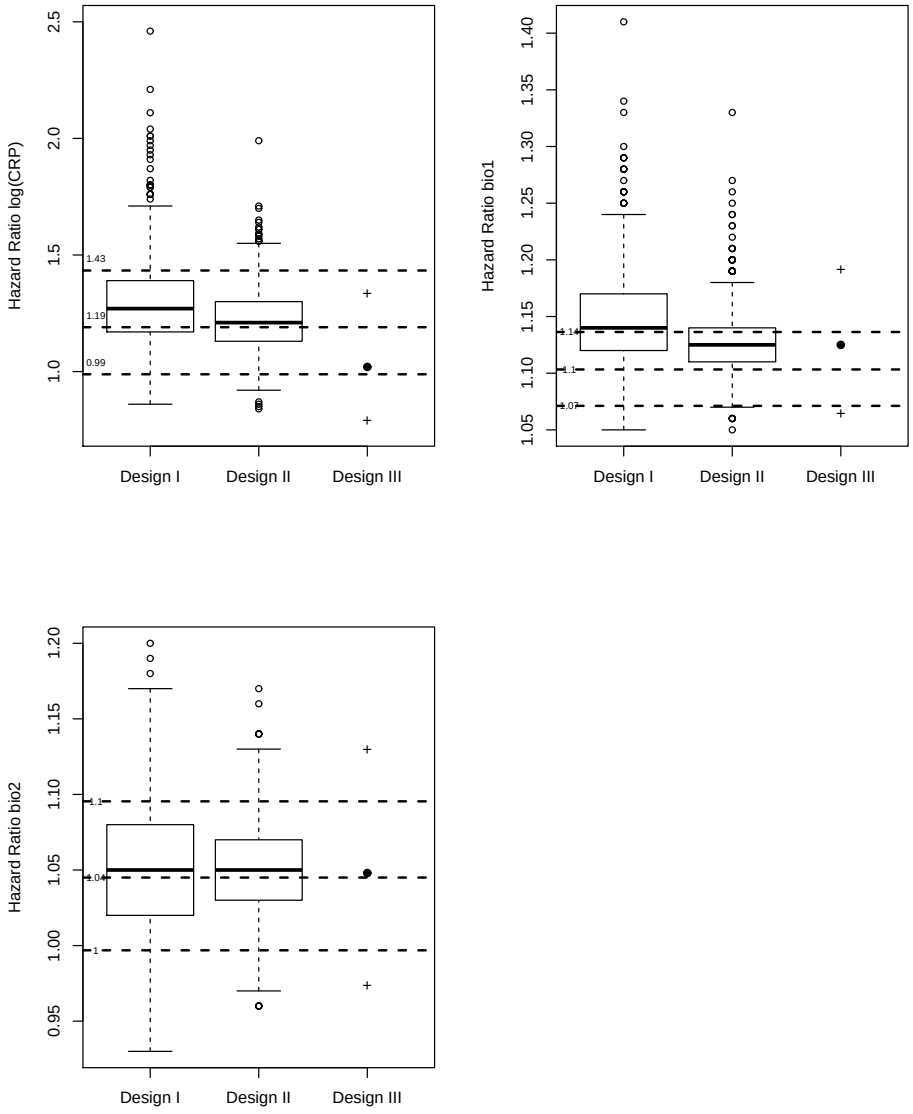

Figure 1: Boxplots for the esitmated hazard ratios from 1000 realizations (Design I, Design II). Design III has only one realization. • denotes the estimation and + denotes the 95% CI from Design III. Dashed horizontal lines indicate the estimations (with 95% CI) from the full cohort.

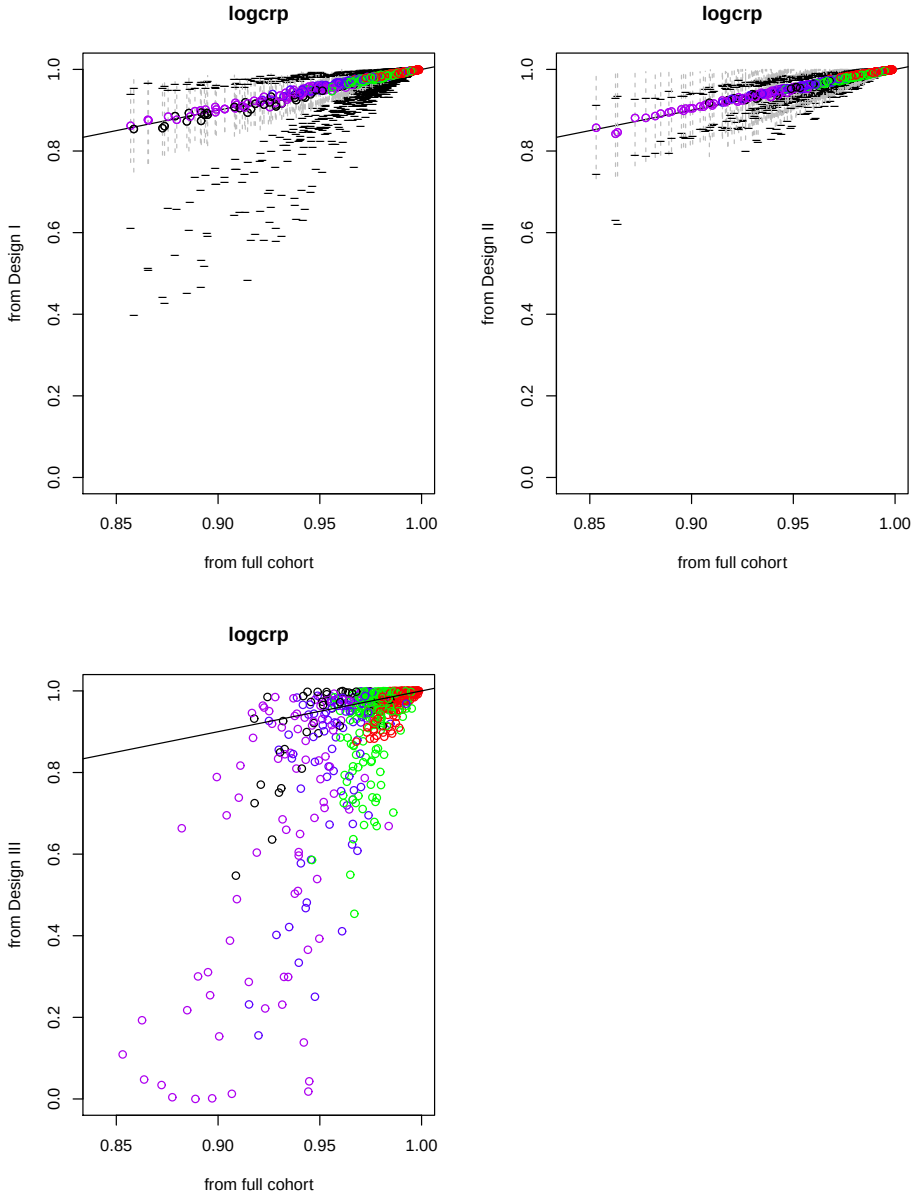

Figure 2: Estimated 5 year from diagnosis survival probability from full cohort analysis (x axis) vs the median estimation from 1000 realizations of nested case control designs (y axis), using log(CRP). The vertical dotted grey lines indicate 95% confidence intervals from full cohort analysis for each subject. The short bars indicate the 2.5% and 97.5% quantiles from the 1000 realizations of nested case control designs. Colors indicate the tumor stage: red, green,blue,purple and black points correspond to stage I, IIA, IIB, IIIA, IIIC respectively.