AD_____

Award Number: DAMD17-03-1-0273

TITLE: Genes Involved in Oxidation and Prostate Cancer Progression

PRINCIPAL INVESTIGATOR: Elizabeth A. Platz, ScD, MPH

CONTRACTING ORGANIZATION: Johns Hopkins University
Bloomberg School of Public Health
Baltimore, MD 21205

REPORT DATE: January 2007

TYPE OF REPORT: Final Addendum

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
01-01-2007	Final Addendum	1 Apr 2003 – 14 Dec 2006

4. TITLE AND SUBTITLE

Genes Involved in Oxidation and Prostate Cancer Progression

5a. CONTRACT NUMBER

5b. GRANT NUMBER
DAMD17-03-1-0273

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Elizabeth A. Platz, ScD, MPH

Email: eplatz@jhsph.edu

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Johns Hopkins University
Bloomberg School of Public Health
Baltimore, MD 21205

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT
We are evaluating whether polymorphisms in genes involved in the genesis of oxidative species, the detoxification of oxidative species, or the repair of oxidative DNA damage influence the risk of prostate cancer progression in men with clinically organ-confined prostate cancer. We requested a no-cost extension through 01/15/2008 to complete the genotyping component of this work. During the past year, we completed locating all but 4 of the 742 (number of unique samples, 524 progressor-control pairs) archived prostate samples (unaffected lymph nodes removed at prostatectomy), which was the unanticipated rate limiting step for this project. We tested the amount of tissue needed and the methods of DNA extraction from the paraffin-embedded tissue that would produce a quantity and quality of DNA that was adequate for amplification by PCR. For each subject, we confirmed that the nodes did not contain cancer and we took 10 cores per block. DNA extraction has been completed for a third of the samples by Bioserve. We tested the ability of the Mass Array system to give accurate genotyping calls for these paraffin-embedded samples. The remaining steps are genotyping and statistical analysis. We generated a manuscript entitled "A Simulation Study of Control Sampling: Methods for Nested Case-Control Studies of Candidate Genes and Prostate Cancer Progression" that compares methods of control sampling for the type of progression study we are conducting to support that the approach we used yields the least biased effect estimates.

15. SUBJECT TERMS
Prostate cancer, oxidation, genes, polymorphisms, risk

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			USAMRMC
U	U	U	UU	37	19b. TELEPHONE NUMBER *(include area code)*

Table of Contents

INTRODUCTION

We are evaluating whether polymorphisms in genes involved in the genesis of oxidative species, the detoxification of oxidative species, or the repair of oxidative DNA damage influence the risk of prostate cancer progression in men with clinically organ-confined prostate cancer who were treated with radical prostatectomy. We hypothesize that men with an inherently greater burden of oxidative stress or inability to repair DNA damage caused by oxidative stress is associated with a higher risk of men. We recently requested a no-cost extension through 01/15/2008, and thus this progress report is an annual progress report rather than final progress report.

During the past year, we completed locating all but 4 of the 742 (number of unique samples, 524 progressor-control pairs) archived prostate samples (unaffected lymph nodes removed at prostatectomy), which was the unanticipated rate limiting step for this project. We tested the amount of tissue needed and the methods of DNA extraction from the paraffin-embedded tissue that would produce a quantity and quality of DNA that was adequate for amplification by PCR. For each subject, we confirmed that the nodes did not contain cancer and we took 10 cores per block. DNA extraction has been completed for a third of the samples by Bioserve. We tested the ability of the Mass Array system to give accurate genotyping calls for these paraffin-embedded samples. The remaining steps are genotyping and statistical analysis. We generated a draft manuscript entitled “A Simulation Study of Control Sampling: Methods for Nested Case-Control Studies of Candidate Genes and Prostate Cancer Progression” that compares methods of control sampling for the type of progression study we are conducting to support that the approach we used yields the least biased effect estimates.

We are pleased to report that the nested case-control set for prostate cancer progression generated under this DOD funding has been viewed as valuable resource by the Hopkins basic, translational, and epidemiologic prostate cancer investigators and the set is now being proposed for wide use at Hopkins. In addition to the Career Development award on genes involved in inflammation received by Dr. Platz and the NCI R01 on genes involved in metastasis received by Dr. Isaacs (Platz co-I), the set is now proposed as the basis for a genome-wide scan by Drs. Isaacs, and Trock as a component of the prostate SPORE (Platz co-I). In addition, Drs. Angelo De Marzo and George Netto are currently assembling tissue microarrays for these cases and controls for expression studies. Dr. Netto (De Marzo, Platz, co-Is) have submitted a proposal to evaluate the presence of the fusion of the androgen-controlled TMPRSS2 gene (21q22.3) with one of 3 members of the ETS family of transcription factors ERG (21q22.2), ETV1 (7p21.2), or ETV4 (17q21) as a predictor of prostate cancer progression.

BODY

The aims of this proposal were:

1) Using expression data from cDNA microarrays coupled with published information on the functionality of sequence changes, we plan to identify 5 single nucleotide polymorphisms (SNP) in each of 25 genes encoding enzymes involved in production of ROS, detoxification of ROS, and repair of oxidative DNA damage.

2) To test whether these SNPs are independently and in combination associated with risk of prostate cancer progression.

We had proposed that these aims be accomplished by the following tasks. After each task, progress is described.

Task 1. *Select 25 polymorphic genes involved in production of ROS, detoxification of ROS, and repair of oxidative damage, Months 1-2*

a. Review cDNA expression data for prostate tumors generated in laboratory of Dr. Isaacs to identify genes involved in oxidation that are expressed above the 80th percentile or below the 20th percentile compared to normal tissue.

Task 2. *Select 200 cases (progressors) and 200 matched controls (nonprogressors) Months 3-5*

a. Link the Hopkins Pathology Tissue Core database to electronic hospital records to identify prostate cancer patients treated with radical prostatectomy and who experienced biochemical failure.
b. From the total set of eligible patients, select 200 men who had biochemical failure and 200 men who still had undetectable PSA at the date of the case's failure, same follow-up time, and who are similar on demographic and tumor characteristics.

Tasks 1 and 2 were completed in the prior funding year. As we indicated for Task 2 in the prior funding year, we consulted with a biostatistician and two other statistical epidemiologists to confirm our approach to control sampling and to ensure that planned analytical approach could handle the data structure that would be imposed by the method chosen. When describing the approach to statistical geneticists we encountered a difference of opinion in the optimal approach. We chose to use incidence density sampling of controls. In this method, a man's person-time at risk is sampled and thus, a man may be sampled more than once represent different person-years at risk and a man who goes on to recur may be sampled as a control prior to failure and then also be counted as a case. The statistical geneticists and urologists suggested that we sampling controls from among the men who did not progress by the end of the follow-up period and who were still under follow-up. Their thought was that genes dictate progression and thus need a pure group of men who were never to progress. We disagreed and suspected that the latter approach would generate a distorted allele frequency from that in the population that gave rise to the cases. To resolve the controversy, we conducted a simulation study in which we sampled controls using 3 methods: incidence density sampling with replacement (meaning a man's person-time experience may be sampled more than once), sampling without replacement, and sampling from the end of follow-up from among the men who were still under follow-up. We compared the estimates of the association of genes and prostate cancer progression from these simulated nested case-control studies to what would be observed if the entire cohort had been studied (gold standard). As we hypothesized, incidence density sampling with replacement was the least biased approach and sampling from the end of the interval was the most biased approach. Sampling without replacement performed only slightly more poorly that sampling with replacement. A manuscript has been prepared and will be submitted for publication: Wang MS, Shugart YY, Zarfas K, Cole SR and Platz EA. "A Simulation

Study of Control Sampling: Methods for Nested Case-Control Studies of Candidate Genes and Prostate Cancer Progression". This work forms one aim of MS Wang's doctoral dissertation; he is pursuing a degree in genetic epidemiology at the Johns Hopkins Bloomberg School of Public Health. Dr. Shugart is a statistical geneticist and Dr. Cole is a statistical epidemiologist who collaborated with us.

Task 3. Genotyping, Months 6-12

a. Pull samples for the 400 patients from Hopkins Pathology Tissue Core archive and review for normal regions.
b. Extract genomic DNA in laboratory of Dr. Isaacs.
c. Ship samples to laboratory of Dr. Xu and perform high throughput genotyping.

In the prior funding period, we decided that an abundant source of DNA would be paraffin-bedded lymph nodes, which are routinely removed at prostatectomy. Also recall that we leveraged additional funds to increase the sample size to 524 progressors and 524 controls. Also, note that the number of unique men is 742, the difference being due to the method of control sampling (incidence density sampling, as described above). During the past year, we completed locating all but 4 of the 742 archived lymph node samples, which was the unanticipated rate limiting step for this project. Many of the case's blocks had been checked out the Hopkins pathology archive and not returned or had been returned but misplaced. Our pathology colleague Dr. De Marzo had his laboratory technician, Ms. Helen Fedor, track down these samples. We are grateful to them for their efforts on our behalf.

We tested the amount of tissue needed and the methods of DNA extraction from the paraffin-embedded tissue that would produce a quantity and quality of DNA that was adequate for amplification by PCR. 10 cores proved to be adequate. For each subject, we confirmed that the nodes did not contain cancer and we took 10 cores per block. DNA extraction has been completed for a third of the samples by Bioserve. We had originally planned to perform DNA extraction in the laboratory of Dr. Isaacs (co-I), but because of the increased sample size, we decided to use a company with high through-put technology and experience with extraction from paraffin-embedded samples.

Genotyping will be done in the laboratory of our collaborator Dr. Jianfeng Xu at Wake Forest. His group uses Sequenom's (San Diego, CA) high-throughput MassARRAY system. We tested the ability of the Mass Array system to give accurate genotyping calls for a small number of these paraffin-embedded samples. We conclude that this approach will be successful.

Task 4. Data management and interim analysis, Months 13-18

Not done.

Task 5. Final analyses and report/manuscript preparation, Months 19-24

Not done.

KEY RESEARCH ACCOMPLISHMENTS

- Showed that our method of control sampling is the least biased based on a simulation study.
- Generated a resource for other studies on genetic variation and gene expression in the etiology of prostate cancer progression
- Accomplishments of Dr. Platz related to this New Investigator Award
    - Since 2004, she heads the cancer epidemiology, prevention and control training program for pre- and post-docs and in 2006 the T32 grant supporting the training program was refunded by the National Cancer Institute
    - In 2006, she was appointed as a Staff Investigator in the Cancer Prevention and Control Program at the Sidney Kimmel Comprehensive Cancer Center.
    - With Dr. James Herman and other Sidney Kimmel Comprehensive Cancer Center Investigators, in 2006 she submitted a DISCOVER Center Grant proposal (P50) to the National Institutes of Environmental Health Sciences. Dr. Platz is proposed as the deputy director and the leader of the project entitled “Prostate Cancer: Environment, Methylation, and Inflammation”.
    - Dr. Platz has become well-known for her research on genes and prostate cancer and has been asked to collaborate with groups outside of Hopkins:

Michaud DS, Daugherty SE, Berndt SI, Platz EA, Yeager M, Crawford ED, Hsing A, Huang WY, Hayes RB. Genetic polymorphisms of interleukin-1B (IL-1B), IL-6, IL-8, and IL-10 and risk of prostate cancer. Cancer Res. 2006;66:4525-30.

Daugherty SE, Hayes RB, Yeager M, Andriole GL, Chatterjee N, Huang WY, Isaacs WB, Platz EA. RNASEL Arg462Gln polymorphism and prostate cancer in PLCO. Prostate (in press).

REPORTABLE OUTCOMES

- Generated a manuscript for publication” A Simulation Study of Control Sampling: Methods for Nested Case-Control Studies of Candidate Genes and Prostate Cancer Progression”, as described above.

CONCLUSIONS

- In genetic epidemiology studies of prostate cancer progression, incidence density sampling with replacement is the least biased approach to control sampling.

REFERENCES

- None to date

APPENDIX

- Draft manuscript: Wang et al. A Simulation Study of Control Sampling: Methods for Nested Case-Control Studies of Candidate Genes and Prostate Cancer Progression.

Comparing Different Control Sampling Methods in Nested Case-Control Studies Applied to Searching for Genes and Prostate Cancer Progression – A Simulation Study

Ming-Hsi Wang[1], Katherine Zarfas[1], Yin Yao Shugart[1], Stephen Cole[1], Elizabeth A. Platz[123]

Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland[1], The James Buchanan Brady Urological Institute, Department of Urology, Johns Hopkins University School of Medicine, Baltimore, Maryland[2], The Sidney Kimmel Comprehensive Cancer Center at Johns Hopkins, Baltimore, Maryland[3]

Corresponding author: Elizabeth A. Platz (e-mail: eplatz@jhsph.edu)
Johns Hopkins Bloomberg School of Public Health
Department of Epidemiology, Room E6138
615 North Wolfe Street
Baltimore, MD 21205
TEL: 410-614-9674
FAX: 410-614-2632

Email addresses:
MHW: mwang2@jhsph.edu
KZ: kzarfas@jhsph.edu
SC: scole@jhsph.edu
YYS: yyao@jhsph.edu

Abstract:

Background: Up to one third of early stage prostate cancer patients who underwent radical prostatectomy experienced disease progression. Different controls sampling methods applied in nested case-control studies, which searched for biomarkers that provide prognostic information and help understand disease progression mechanisms, have been adopted by researchers in different studies. ***Aims***: Through a simulation, to compare the validity of the estimates of relative risk from three different controls sampling schemes. ***Methods***: Hypothetical biomarker effects and disease progression probabilities were constructed by a stratified proportional hazard model. Three different controls sampling schemes, scheme 1 “incidence density sampling with replacement”, scheme 2 “incidence density sampling without replacement”, and scheme 3 “pure control”, were used and compared for their estimates of the true effect size. ***Results***: Scheme 1 and 3 show unbiased and biased estimates of relative risk, respectively. Scheme 2 did not show evidence of bias because of few subjects have been selected as controls more than once. ***Conclusions***: Nested case-control study, for searching biomarkers contributing to prostate cancer progression, within a defined cohort could efficiently provide valid estimate of relative risk with adequate sampling method like scheme 1.

Introduction:

Prostate cancer (PCa) has been a major health problem in US and Western Europe for decades. According to the American Cancer Society (ACS) estimation based on data from the US Surveillance, Epidemiology and End Results (SEER) Program (1), PCa is the most commonly diagnosed cancer in US men. In recent years, men who were diagnosed with PCa at early stage have higher than 99% 5-yr survival rate (2). Compared with past decades when survival was poorer, this high survival rate may be due to early detection with prostate surface antigen (PSA) when the cancer may be treated by surgery, radical retro-pubic prostatectomy (RRP). However, up to 33% of men who are treated with RRP experience biochemical failure, PSA re-elevation from non-detectable, in men with early or localized PCa (i.e. clinically organ-confined disease) (3;4). In the Johns Hopkins series, 32% and more than 50% of men, who underwent RRP, developed clinically evident metastasis at 5 and 10 years follow-up time, respectively, since PSA re-elevation. Identification of relevant molecular/ genetic markers for their prognostic significance with reference to cancer recurrence could provide complementary information to that gained from traditional histo-pathological features or clinical parameters, e.g. Partin table (5), and also help understand disease progression mechanisms.

Two primary types of non-experimental studies in epidemiology could be used to evaluate markers of progression - cohort study and case-control study. A prospective cohort study has the advantage of providing correct temporal sequence between exposure and disease, however, the time and cost involved in collecting exposure and covariate information may be substantial. In biomarker studies including genetic variants, an efficient design is needed to reduce the cost and still preserve the correct temporal sequence. A commonly used approach is to conduct a case control study nested within a pre-existing cohort (6-8). Through the adequate control sampling method, like incidence density sampling with replacement, valid estimates of true effect size could be obtained in a more efficient way, i.e. offering impressive reductions in costs and efforts of data collection and with relatively minor loss in statistical efficiency (9). This approach has been defined as in which controls are selected from all persons at risk, excluding the index case itself, at the time of case occurrence, and the selected control at one certain time could still be eligible to be selected as control, if he still hasn't had event, for another case who occurred at later time or become case in the later follow-up time. In addition, this method has been shown as one choice for obtaining unbiased results (10;11). To understand the role of genetic variants can play on the PCa progression, two different types of cohorts could be applied to conduct a nested case-control study. One is following

up men who are diagnosed with clinically organ confined PCa and underwent RRP. The other is embedded in a large cohort of healthy men, in whom some of them developed clinically organ confined PCa later.

In clinic-based studies of biomarkers of PCa progression, different control sampling methods have been chosen by investigators based on an intuitive sense; for example, they have selected subjects who never progressed by the end of study follow-up as controls for incident cases who progressed during the period of follow-up(12-15). However, restricting controls to be individuals who never progressed would lead to biased estimates of the relative risk (RR) (10). In this paper, we conduct a simulation study, based on the scenario of long term follow up of a hypothetical group of men, who were diagnosed with clinically organ-confined PCa and underwent treatment with RRP, to contrast the estimates of RR of progression using different control sampling schemes. We compare the results from three different schemes, including incidence density sampling with or without replacement and an often used variant scheme in which controls are only selected from those who survived and still free of progression to the end of follow up, to document their unbiased or biased estimates of the parameters in related to the RR from an optimally conducted cohort study (gold standard).

Methods:

Background of simulated cohorts:

According to a large case series report from Johns Hopkins Hospital (16), 2404 men with a mean follow-up of 6.3 +/- 4.2 years after RRP for clinically localized PCa, 412 men (17%) have recurred. The overall actuarial 5-, 10-, 15-year recurrence-free survival rates were 84%, 74%, and 66%, respectively. Based on these data, we simulated a cohort of men diagnosed with clinically organ-confined PCa, treated with RRP, and were closely post-operative followed for up to 15 years for progression. Progression was defined as PSA re-elevated after becoming undetectable post-operatively. Among the clinical parameters of PCa patients, pathological stage and pathological Gleason score (GS) are the two most important predictors for the disease progression (16-18). Patients with advanced stage, or higher GS, PCa would progress more rapidly than those with less advanced disease. Therefore, it would be difficult to find enough number of non-progressed subjects in the stratum of advanced stage and high GS PCa while doing stratified analysis. This situation is called a sparse-data problem. Without matching, many strata in a stratified analysis would have very few subjects or even none. Through matching, each case would have one or more matched controls for comparison in the subsequent stratified analysis. We grouped patients according to the combinations of

pathologic stage and pathological Gleason score (GS) which could provide good stratification of progression-free probability. For simplicity, we simulated three strata of patients each with different 5-, 10-, and 15-year progression-free survival rates (PFSR), respectively; stratum 1, GS less than 7 with or without extra-prostatic extension (EPE), with 96%, 90%, and 80% at 5-, 10- and 15-year PFSR, respectively; stratum 2, GS large than or equal to 7 with EPE, with 82%, 69% and 59% at 5-, 10- and 15-year PFSR, respectively; and stratum 3, seminal vesicle (SV) or lymph node involvement, with 68%, 51% and 43% at 5-, 10- and 15-year PFSR, respectively (16). Under these simulated parameters (Table 1), we were able to sample for each incident case a control matched exactly on the pathological stage and GS under each of the three different control sampling schemes, described below.

Description of Models:

Based on the stratified proportional hazards assumption, we simulated the cohort composed of three different strata, which were characterized as the above. The survival time of the ith person in the jth strata of a cohort is assumed to follow its own hazard function, $\lambda_{ij}(t_{ij})$ which could be expressed as

$$\lambda_{ij}(t_{ij}) = \lambda_{0j}(t_{ij}) \exp(X_{ij}\beta)$$

where $\lambda_{0j}(t_{ij})$ is the baseline hazard for the jth strata, X is the explanatory variable, here it

could be a biomarker of risk for progression including genetic variant, for the ith person in the jth strata, and β is the unknown regression parameter. We assumed that the effect of X on the hazard is constant within each stratum. X_{ij} is obtained from independent uniform (0,1) distribution. In this study, we will use a common regression coefficient for the effect of the biomarker over all stage and grade strata in each model. However, we modeled different baseline hazards or each stage and grade stratum.

We assessed the suitability of applying a Weibull parametric model of progression to the real data by drawing the log-cumulative hazard plot and the results from the three groups showed pretty close to a straight line which means Weibull distribution is an adequate parametric model to simulate the survival time (19). According to the suitability check for Weibull model using survival rate from the real data, we assumed the survival time for the ith person in the jth stratum T_{ij} following a Weibull distribution ($(\gamma/\alpha)^* (\exp(X_{ij}\beta) * (t_{ij}/\alpha)^{(\gamma-1)})$), where$\gamma$is the shape parameter and α is the scale parameter . These two parameters for each jth stratum could be determined according to the real survival data of each stratum described above.

Men who were diagnosed with clinically organ-confined PCa, with the average age around sixty (16), who were closely followed post-RRP follow-up, and could also die from other diseases, like heart attack, stroke etc. Therefore, we simulated a competitive

death probability in the simulated cohort using data from U.S. white men in 1988-1992, which also approximately followed a Weibull distribution (Table1) (20).

We performed 2000 replicates for each scheme using the Cox proportional hazards regression model which was implemented in PROC PHREG procedure in SAS (SAS Institute Cary, NC). In each replicate, the total sample size was fixed at 2000, with 1400 in stratum 1, 400 in stratum 2, and 200 in stratum 3. A total of nine models, the combinations of three different x1 covariate effect size (RR=1.0, 1.5, 2.0, which are hypothetical effect size for genetic or biomarker studies) and three different x1 exposure prevalence (10%, 30%, 50%), were fitted. In addition, we assumed that the each simulated cohort was observed for 15 years. Failed subjects were defined as those who had had PSA failure occurred before the end of follow-up and also before the time of dying from other competitive risks. All the other subjects were coded as censored.

Cases/ Controls selection:

All observed failed subjects (“progressors”) were selected to be “cases” from the cohort in 15 years follow-up. Three different control sampling schemes were adopted. Scheme 1 (see figure 1a) is called “incidence density sampling with replacement”, which is the gold standard method of control sampling. A given case at the time of PSA failure is matched on factors, pathological stage and Gleason score (GS), with one control

randomly selected from all those at risk at that time (i.e., risk set sampling) (21-23). A case developed at time *t* may serve as a control for a prior case, and a single control may be selected for more than one case, though these situations arise rarely if there are few cases and many non-cases. This approach is exactly analogous to the risk sets in a proportional hazards model. Scheme 2 (see figure 1b) is called "incidence density sampling without replacement", in which, the difference between this from scheme 1 is that once a subject has been selected as control at time *t*, then that subject could not be eligible again to be selected as control at later time, even that subject still doesn't have event occurred yet. This approach could provide some practical convenience while doing genetic or biomarker studies and sample is precious. Scheme 3 (see figure 1c) is called "pure control", in which, a given case at the time of PSA failure is matched on pathological stage and GS, with one control randomly selected from all those who are at risk at that time and also never observed to develop PSA failure or dying from other competitive risk factor through the whole period of observation. This approach has been adopted in some studies (12-15).

Statistical Analysis:

To estimate the β regression coefficient and its standard error for the simulated full cohort and each of three case-control sampling schemes, we applied the

Newton-Raphson procedure to estimate β and the maximization of the Breslow likelihood function is easily accomplished on a computer program, SAS package version 9.1 (SAS institute, Cary, NC). The estimated significance levels (rejection probabilities), which was defined as the fraction of replicates for which [βhat / standard error (βhat)] exceeds 1.96 underα=0.05, for testingβ=0 were also presented in the Tables 2-4. In models with RR=1.0, this estimated significance level is the observed type I error. In models with RR=1.5 or 2.0, the estimated significance level is the observed power.

Results:

In Table 2-4, we give the average values of the parameter estimates, natural logarithm of RR for full cohort or the odds ratio (OR) for nested case-control, over the 2000 replicates of simulation under different exposure of interest prevalence in the cohort population. Convergence was achieved for all replicates. The standard deviation of the sampling distribution of the parameter estimator is also shown. For comparison purposes, the left column of the tables lists the simulation summary statistics for the full cohort partial likelihood estimation procedure. The estimated significance levels (rejection probabilities) for testing null hypothesisβ=0 are within sampling error of their nominal values with the exception of the scheme 3, at the exposure prevalence 30%, which exceeds the 5% significance level, i.e. exceeding its nominal value by two standard

errors.

There was no evidence of bias in the estimates of β for scheme 1, however biased results observed in scheme 3, when $\beta=0$ (i.e. RR=1.0) across the three different prevalence of exposure. For scheme 2, there was no evident bias observed here because there was only a few controls have been selected for more than once. From Table 2-4, no matter how prevalent the exposure in the cohort population is, scheme 3 tended to overestimate the β at the scenarios of $\beta=0.405$ (RR=1.5) and 0.693 (RR=2.0). It also showed a higher type I error proportion for $\beta=0$ when compared with sampling schemes 1 and 2. The bias of the estimated effect size, away from the true value, became bigger with the increase in the true effect size (β). However, the extent of the bias did not change with the prevalence of the causal exposure.

The only difference between scheme 1 and 2 is the former allows the prior selected control to be eligible for subsequent control if the subject is still at risk, but the latter doesn't. From Table 2-4, both the estimates of β from schemes 1 and 2 are not appreciably different from the truth and also not different from each other. However, the sample standard error of the estimates of β from scheme 2 is slightly smaller than that from scheme 1 by 6-8%. The power to reject null hypothesis $\beta=0$ is similar at both scheme 1 and 2 across different exposure prevalences.

Discussion:

As is well known in the epidemiologic theory, it is obvious that different case-control sampling schemes could lead to unbiased or biased estimates of RR. From this simulation study, we shows that unbiased estimate could be achieved by selecting controls randomly from all those at each case failure time *t*, i.e. scheme 1 “incidence density sampling with replacement”. However, bias could be induced by selecting “pure control”, i.e. excluding subsequent cases from the control group. This bias could be substantial while applying to clinical trials or epidemiological studies with disease of appreciable incidence, like the PSA failure rate among men who had received RRP.

For practical reasons, one might intend to use “biased” control sampling scheme, i.e. scheme 3, to conduct a case-control study. The reason for adopting this approach could be intuitive; that is, if we have already known one patient, who was diagnosed as clinically organ-confined PCa and had undergone RRP, and was observed with PSA failure during the follow-up, then this patient should not be eligible to be selected as “control”. Therefore, researchers might only select “control” from those who remained non-progressed to the end of follow-up. This is exactly the idea of control sampling scheme 3 and, from our study, the result shows the evidence of biased estimates of RR. The reason of this approach leading to biased estimates is that the exposure

distribution among controls is not the same as it is in the source population which derives the cases. In scheme 3, the controls group has to remain non-progressed till the end of 15 years follow-up. If the genetic variant X1 is a risk factor for progression, the distribution of genetic variant in the controls pool of scheme 3 must be very different from it is in the source population of cases. Furthermore, the probability of carrying this genetic risk factor in controls would be much lower than it is in the source population and the calculated OR must be overestimated, because the controls pool is “too healthier” to compare with.

Why the scheme 1 is the optimal way to estimate the RR? To answer this question we should first acknowledge how the RR was calculated in a cohort. Following Prentice and Breslow et al. (24), here we define the cause-specific hazard rate in the unexposed group for cause x1, which could be a genetic risk factor, as

$$\lambda_0(t) = \lim_{\triangle t \to 0} \text{pr\{ fail from x1 in } [t, t+\triangle t] \mid \text{at risk at } t \} / \triangle t$$

Then we make a very crucial “proportional hazard” assumption that the cause-specific hazard rate in the exposed group is

$$\lambda(t) = \lambda_0(t) \exp(x1*\beta)$$

This assumption basically assumes that the ratio of the hazard rates for exposed and unexposed groups remains constant, i.e. $\exp(x1*\beta)$ which is exactly the RR, over time.

The maximum likelihood estimates of β could be obtained using Newton-Raphson procedure to maximize (25). The calculated RR would be unbiasedly approximated by the Odds Ratio (OR) from scheme 1 "incidence density sampling with replacement". Through the linear algebraic calculation, the OR derived from scheme 1 will be

$$OR = \frac{[\ pr(x1=1 \mid \text{case at } t) / pr\ (x=0 \mid \text{case at } t)\ ]}{[\ pr(x1=1 \mid \text{control at } t) / pr\ (x=0 \mid \text{control at } t)\ ]} = RR$$

i.e. the OR from scheme 1 is unbiased estimate of RR which is derived from full cohort, if and only if control was selected randomly from all patients at risk at time *t* when the case occurred, which is also called "risk-set sampling" (10). In addition, even more importantly, a person selected as a control at time *t* remains in the study population at risk after selection should remain eligible to be selected as control again at later time *t'*, where $t' > t$.

However, another possible choice of a control sampling scheme, i.e. scheme 2 "incidence density sampling without replacement", could be adopted and executed by some investigators. They would select controls randomly from the risk set at time *t* when cases occur, and those subjects selected as controls prior to time *t* would be excluded from the controls pool at time *t*. In a prospective cohort study, the disease incidence rate might be defined as,

Incidence Rate = (No. of disease onsets) / ($\Sigma_{persons}$ time spent in population)

Each person who develops disease, here is “progression”, would contribute not only to the numerator of the disease incidence rate but also to the person-time experience tallied in the denominator of the rate, until the time of disease onset. Therefore, each case should have been eligible to be a control before the time of disease onset. Each control should be eligible to become a case as of the time of selection as a control (10;11;26). This explains why the scheme 3 is not appropriate. Furthermore, a person selected as a control remains in the study population at risk after selection should remain eligible to be selected once again as a control later. Thus, the same person may appear in the control group two or more times. This explains why the scheme 2 is not an appropriate method too. The reason for no difference on the estimates of β between schemes 1 and 2 shown in Table 2-4 is there is only few selected controls in scheme 1 served more than once. To prove this conceptually difference, we did another simulation analysis by modifying the model parameters to let the disease incidence, here is progression incidence, be higher and the result did show biased estimate of the RR (data not shown).

. To study the genetic variants which might cause post-operative PSA failure in men who were diagnosed with clinically organ-confined PCa and received RRP, we already known that pathological stage and GS were both very significantly outcome

predictors (16-18), which means higher stage or GS has higher probabilities having PSA failure. However, it is difficult to find sufficient number of unique controls, who have the same stage and GS as cases, at the same risk sets as cases occurred. This indicates that applying scheme 2 to select controls in this kind of situation would not be feasible; oppositely, scheme 1 could easily overcome this problem.

Case-control studies might be best understood by defining a source population, which represent a hypothetical study population in which a cohort study might have been conducted. The major task in a cohort study is to identify the exposed and unexposed denominator experience in person-time units and then to identify the number of cases occurring in each person-time category. In a case-control study, cases are identified and their exposure status is determined just as in a cohort study, but denominators from which rates could be calculated are not measured. Instead, a control group of subjects, which is sampled from the time experience of the source population, is used to determine the relative size of the exposed and unexposed person-time denominators within the source population, i.e. RR. To achieve the goal of obtaining unbiased estimate of RR, the cardinal requirement of control selection is that the controls must be sampled independently of their exposure status. In our simulation study, scheme 1 exactly followed this rule and obtained the valid estimates of RR. Scheme 3 is against the rule

and its result shows biased estimates from the truth because of selected controls less likely to have casually hazardous exposure. Scheme 2 is also against the principle of controls selection but did not show bias here, Table 2-4, because there is only few subjects selected as controls more than once. The bias would be more visible when disease incidence increases. In conclusion, nested case-control study, for searching genetic variants contributing to PCa progression, within a defined cohort could efficiently provide valid estimate of RR with adequate sampling method, like scheme 1 in this simulation study.

Reference List

(1) American Cancer Society. Cancer Factors and Figures 2006. Atlanta American Cancer Society 2006.

(2) Ries LAG, Harkins D, Krapcho M etc. SEER Cancer Statistics Review, 1975-2003. National Cancer Institute, Bethesda, MD; 2006.

(3) Partin A, Pound C, Clemens J, Epstein JI, Walsh PC. Serum PSA after anatomical radical prostatectomy: The Johns Hopkins experience after 10 years. Uro Clin North Am 1993;20:713.

(4) Trapasso J, DeKernion J, Smith R, Dorey F. The incidence and significance of detectable levels of serum prostate specific antigen after radical prostatectomy. J Urol 1994;152:1821.

(5) Partin AW, Kattan MW, Subong EN, Walsh PC, Wojno KJ, Oesterling JE, et al.

Combination of prostate-specific antigen, clinical stage and Gleason score to predict pathological stage of localized prostate cancer. A multi-institutional update. JAMA 1997;277:1445-51.

(6) Arana A, Varas C, Gonzalez-Perez A, Gutierrez L, Bjerrum L, Gracia-Rodriguez LA. Hormone therapy and cerebrovascular events: a population-based nested case-control study. Menopause: The Journal of The North American Menopause Society 2006;13(5):730-6.

(7) Haiman CA, Hankinson SE, Spiegelman D, Colditz G, Willett WC, Speizer FE, et al. The relationship between a polymorphism in CYP17 with plasma hormone levels and breast cancer. Cancer Research 1999;59:1015-20.

(8) Pedersen L, Norgaard M, Skriver MV, Olsen J, Sorensen HT. Prenatal exposure to loratadine in children with hypospadias: a nested case-control study within the Danish national birth cohort. Am J Ther 2006;13:320-4.

(9) Ernster VL. Nested Case-Control Studies. Preventive Medicine 1994;23:587-90.

(10) Lubin JH, Gail MH. Biased selection of controls for case-control analyses of cohort studies. Biometrics 1984;40:63-75.

(11) Robbin JM, Gail MH. More on "Biased selection of controls for case-control analyses for cohort studies". Biometrics 1986;42:293-9.

(12) Murata M, Watanabe M, Yamanaka M, Kubota Y, Shiraishi T. Genetic polymorphisms in cytochrome P450(CYP)1A1, CYP1A2, CYP2E1, glutathione S-transferase (GST)M1 and GSTT1 and susceptibility to prostate cancer in the Japanese population. Cancer Letters 2001;165:171-7.

(13) Hsing AW, Chen C, Chokkalingam AP, Reichardt JKV. Polymorphic Markers in the SRD5A2 Gene and Prostate Cancer Risk: A Population-based Case-control Study. Cancer Epidemiology Biomarkers & Prevention 2001 Oct;10:1077-82.

(14) Hein DW, Leff MA, Ishibe N, Caporaso NE. Association of Prostate Cancer With Rapid N-acetyltransferase 1 in Combination With Slow N-acetyltransferase 2 Acetylator Genotypes in a Pilot Case-Control Study. Environmental and Molecular Mutagenesis 2002;40:161-7.

(15) Rebbeck TR, Walker AH, Zeigler-Johnson C, Malkowicz SB. Association of HPC2/ELAC2 Genotypes and Prostate Cancer. Am J Hum Genet 2000;67:1014-9.

(16) Han M, Partin AW, Pound CR, Epstein JI, Walsh PC. Long-term biochemical disease-free and cancer-specific survival following anatomic radical retropubic prostatectomy:the 15-year Johns Hopkins Experience. Urologic Clinics of North America 2001 Aug 1;28(3):555-65.

(17) Epstein JI, Pizov G, Wash PC. Correlation of pathologic findings with progression after radical retropubic prostatectomy. Cancer 1993;71:3582-93.

(18) Epstein JI, Partin AW, Sauvageot J, Walsh PC. Prediction of Progression Following Radical Prostatectomy: A Multivariate Analysis of 721 Men with Long-term Follow-up. Am J Surg Pathol 1996;20:286-92.

(19) Collett D. The Weibull model for survival data. Modelling survival data in medical research.London: Chapman & Hall; 1994. p. 107-47.

(20) Pickle LW, Mungiole M, Jones GK, White AA. Atlas of United States Mortality. Maryland USA: CDC NCHS; 1996.

(21) Wacholder S, McLaughlin JK, Silverman DT, Mandel JS. Selection of controls in case-control studies, I: principles. Am J Epidemiol 1992;135:433-7.

(22) Wacholder S, Silverman DT, McLaughlin JK, Mandel JS. Selection of controls in case-control studies, III: design options. Am J Epidemiol 1992;135:1042-50.

(23) Wacholder S, Silverman DT, McLaughlin JK, Mandel JS. Selection of controls in case-control studies, II: types of controls. Am J Epidemiol 1992;135:1019-28.

(24) Prentice RL, Breslow NE. Retrospective studies and failure-time models. Biometrika 1978;65:153-8.

(25) Prentice RL. On the design of synthetic case-control studies. Biometrics 1986;42:301-10.

(26) Greenland S, Thomas DC. On the need for the rare disease assumption in case-control studies. Am J Epidemiol 1982;116:547-53.

Table 1. Parameters used to simulate cohorts, of men who were diagnosed with clinically organ-confined prostate cancer and underwent RRP, which composed of three different strata for risk of progression and followed up for 15 years.

		Progression-free probability			Progression hazard rate function					Competitive deaths rate function				
	Sample size	5-yr	10-yr	15-yr	5-yr	10-yr	15-yr	γ*	α*	5-yr cumulative deaths	10-yr cumulative deaths	15-yr cumulative deaths	γ*	α*
Stratum 1	1400	96%	90%	80%	0.007	0.010	0.013	1.5	60	55	89	156	8.74	95.59
Stratum 2	400	82%	69%	59%	0.023	0.021	0.019	0.85	52	9	24	44	8.74	95.59
Stratum 3	200	68%	51%	43%	0.032	0.027	0.023	0.75	40	5	8	13	8.74	95.59

***** Under Weibull distribution assumption ($(\gamma/\alpha)^* (\exp(X_{ij}\beta) * (t_{ij}/\alpha)^{(\gamma-1)})$), where γ is the shape parameter and α is the scale parameter

Table 2. Simulation statistics, based on 2000 replicates, of various controls sampling methods of estimating β, the natural logarithm of relative risk, with exposure of interest prevalence Pe=10%.

	Full Cohort (N=2000)	Scheme1 / (Bias)	Scheme 2 / (Bias)	Scheme 3 / (Bias)
Relative Risk = 1.0 (β=0)				
1> Number of cases or case-control pairs	339.0 +/- 16.5	339.0 +/- 16.5	339.0 +/- 16.5	339.0 +/- 16.5
2> Sample mean β(hat)	-0.014	0.0001 / (0.0001)	-0.002 / (0.002)	-0.011 / (0.011)
3> Sample standard error	0.185	0.260	0.244	0.260
4> Rejection rate for null hypothesis (Ho:β=0); here, it is Type I error	4.85%	4.50%	3.05%	4.75%
Relative Risk = 1.5 (β=0.405)				
1> Number of cases or case-control pairs	353.3 +/- 16.4	353.3 +/- 16.4	353.3 +/- 16.4	353.2 +/- 16.3
2> Sample mean β(hat)	0.403	0.419 / (0.014)	0.412 / (0.007)	0.491 / (0.086)
3> Sample standard error	0.154	0.238	0.223	0.250
4> Rejection rate for null hypothesis (Ho:β=0); here, it is Power	72.35%	40.25%	38.30%	49.85%
Relative Risk = 2.0 (β=0.693)				

1> Number of cases or case-control pairs	365.5 +/- 15.9	365.5 +/- 15.9	365.5 +/- 15.9	365.5 +/- 15.9
2> Sample mean β(hat)	0.694	0.708 / (0.015)	0.699 / (0.006)	0.852 / (0.159)
3> Sample standard error	0.138	0.235	0.214	0.248
4> Rejection rate for null	99.80%	87.85%	88.65%	95.35%

hypothesis (Ho:β=0); here, it is

Power

*** Bias: the absolute value of the difference between the sample mean of parameter β (hat) and the true β.**

Table 3. Simulation statistics, based on 2000 replicates, of various controls sampling methods of estimating β, the natural logarithm of relative risk, with exposure of interest prevalence Pe=30%.

	Full Cohort (N=2000)	**Scheme1 / (Bias)**	**Scheme 2 / (Bias)**	**Scheme 3 / (Bias)**
Relative Risk = 1.0 (β=0)				
1> Number of cases or case-control pairs	339.5 +/- 16.4	339.5 +/- 16.4	339.5 +/- 16.4	339.5 +/- 16.4
2> Sample mean β(hat)	-0.001	0.004 / (0.004)	0.002 / (0.002)	0.002 / (0.002)
3> Sample standard error	0.118	0.169	0.156	0.172
4> Rejection rate for null	5.35%	4.90%	3.45%	6.00%

hypothesis (Ho:β=0); here, it is

Type I error

Relative Risk = 1.5 (β=0.405)

1> Number of cases or case-control pairs	381.6 +/- 16.4	381.6 +/- 16.4	381.6 +/- 16.4	381.4 +/- 16.3
2> Sample mean β(hat)	0.403	0.406 / (0.001)	0.402 / (0.003)	0.480 / (0.075)
3> Sample standard error	0.104	0.157	0.144	0.159
4> Rejection rate for null hypothesis (Ho:β=0); here, it is Power	96.75%	73.55%	74.9%	86.55%
Relative Risk = 2.0 (β=0.693)				
1> Number of cases or case-control pairs	417.8 +/- 17.2	417.8 +/- 17.2	417.8 +/- 17.2	417.2 +/- 16.5
2> Sample mean β(hat)	0.692	0.694 / (0.001)	0.683 / (0.01)	0.846 / (0.153)
3> Sample standard error	0.101	0.157	0.143	0.160
4> Rejection rate for null hypothesis (Ho:β=0); here, it is Power	100%	99.80%	99.90%	100%

*** Bias: the absolute value of the difference between the sample mean of parameter β (hat) and the true β.**

Table 4. Simulation statistics, based on 2000 replicates, of various controls sampling methods of estimating β, the natural logarithm of relative risk, with exposure of interest prevalence Pe=50%.

	Full Cohort (N=2000)	**Scheme1 / (Bias)**	**Scheme 2 / (Bias)**	**Scheme 3 / (Bias)**
Relative Risk = 1.0 (β=0)				

1> Number of cases or case-control pairs	339.8 +/- 16.3	339.8 +/- 16.3	339.8 +/- 16.3	339.8 +/- 16.3
2> Sample mean β(hat)	0.003	0.001 / (0.001)	0.002 / (0.002)	0.002 / (0.002)
3> Sample standard error	0.109	0.157	0.145	0.155
4> Rejection rate for null hypothesis (Ho:β=0); here, it is Type I error	5.35%	4.75%	3.15%	4.90%
Relative Risk = 1.5 (β=0.405)				
1> Number of cases or case-control pairs	408.7 +/- 17.1	408.7 +/- 17.1	408.7 +/- 17.1	407.5 +/- 16.6
2> Sample mean β(hat)	0.405	0.407 / (0.002)	0.401 / (0.004)	0.482 / (0.077)
3> Sample standard error	0.101	0.145	0.131	0.145
4> Rejection rate for null hypothesis (Ho:β=0); here, it is Power	97.50%	82.70%	83.55%	94.10%
Relative Risk = 2.0 (β=0.693)				
1> Number of cases or case-control pairs	469.2 +/- 18.2	469.2 +/- 18.2	469.2 +/- 18.2	460.2 +/- 17.2
2> Sample mean β(hat)	0.694	0.696 / (0.003)	0.683 / (0.01)	0.849 / (0.156)
3> Sample standard error	0.097	0.137	0.125	0.144
4> Rejection rate for null	100%	99.95%	100%	100%

hypothesis (Ho:β=0); here, it is

Power

*** Bias: the absolute value of the difference between the sample mean of parameter β (hat) and the true β.**

Figure 1 a. Diagram of control selection scheme 1, "incidence density sampling with replacement", the gold standard.

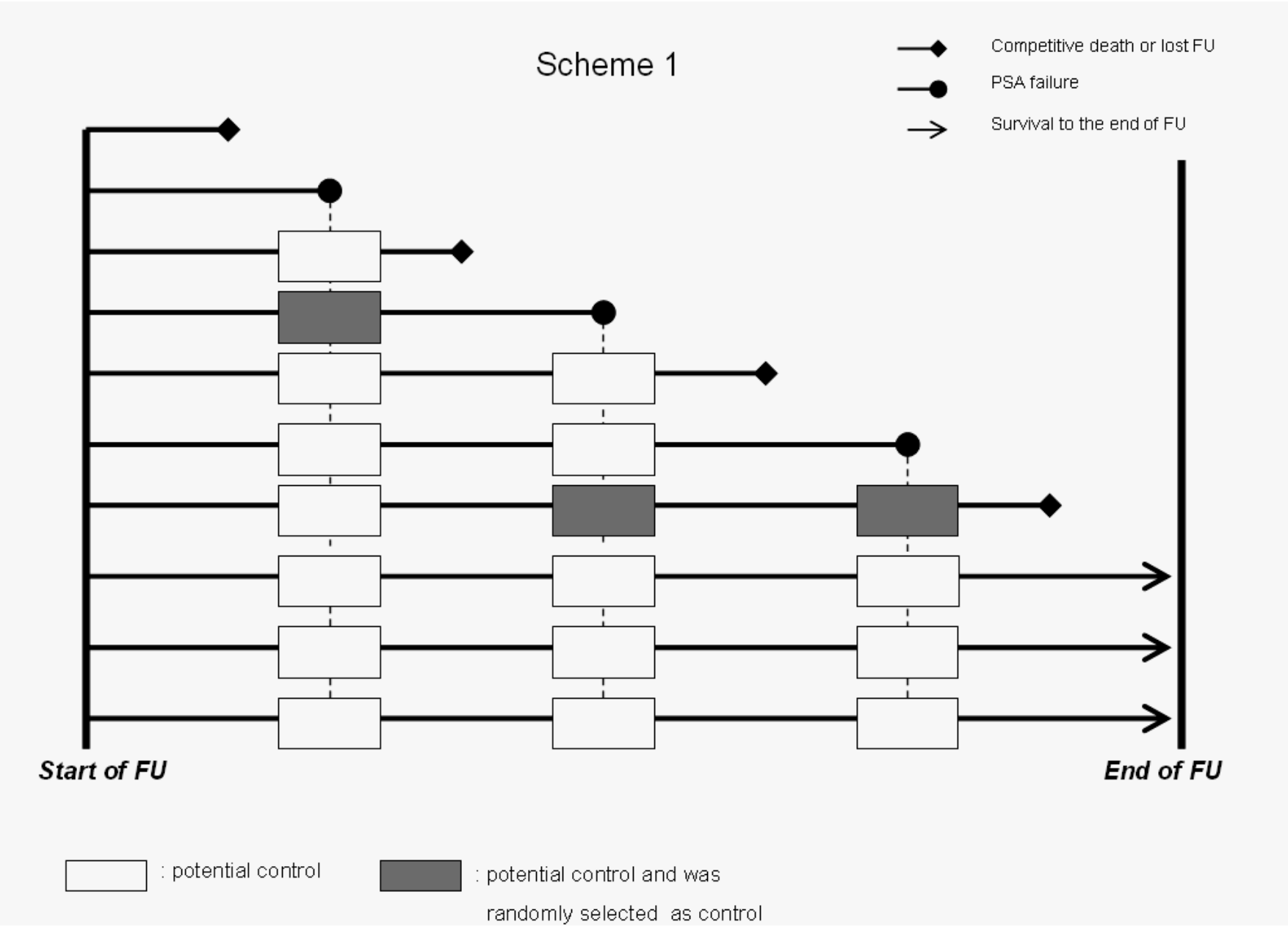


Figure 1 b. Diagram of control selection scheme 2, "incidence density sampling without replacement", a biased method.

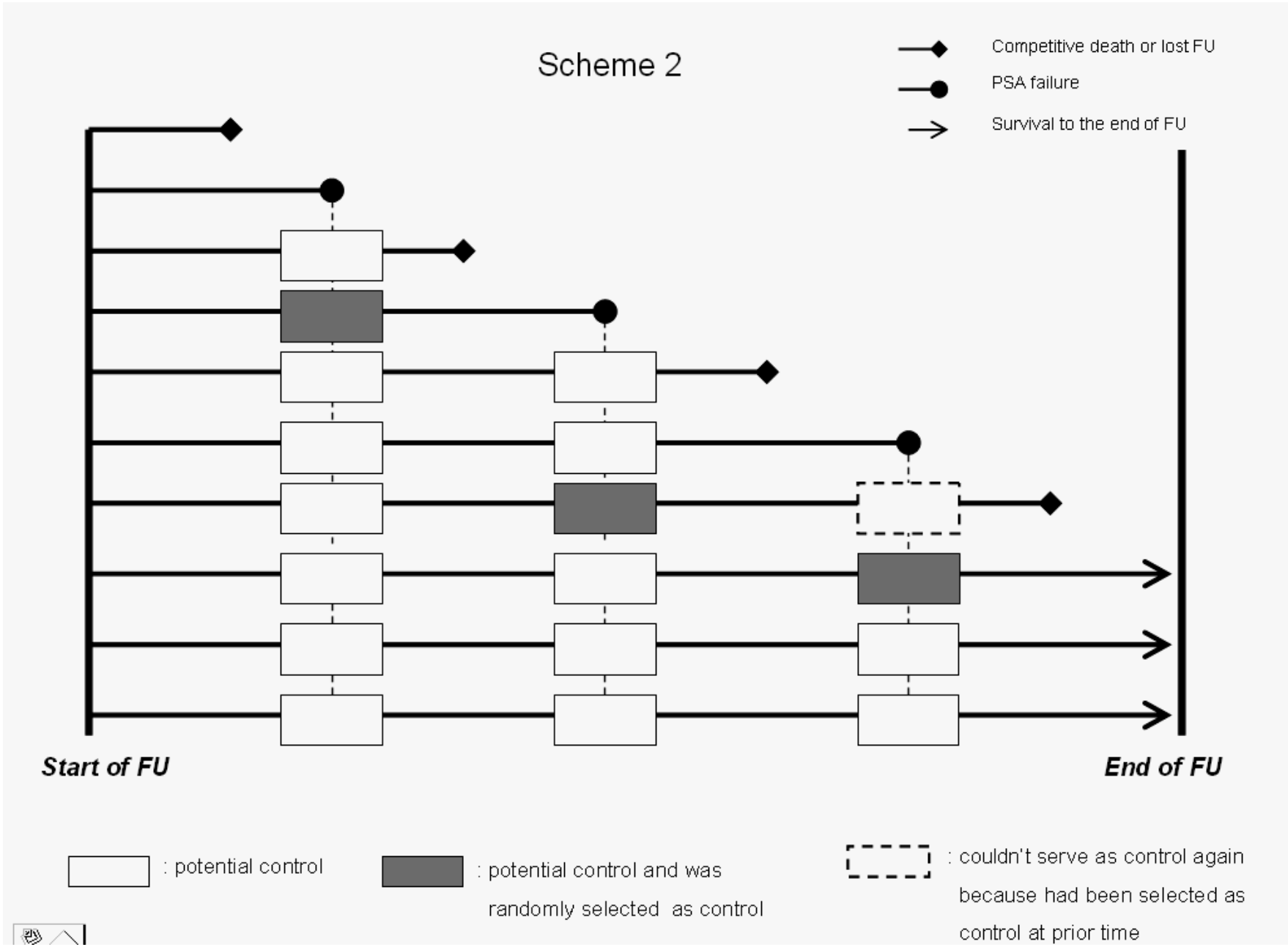
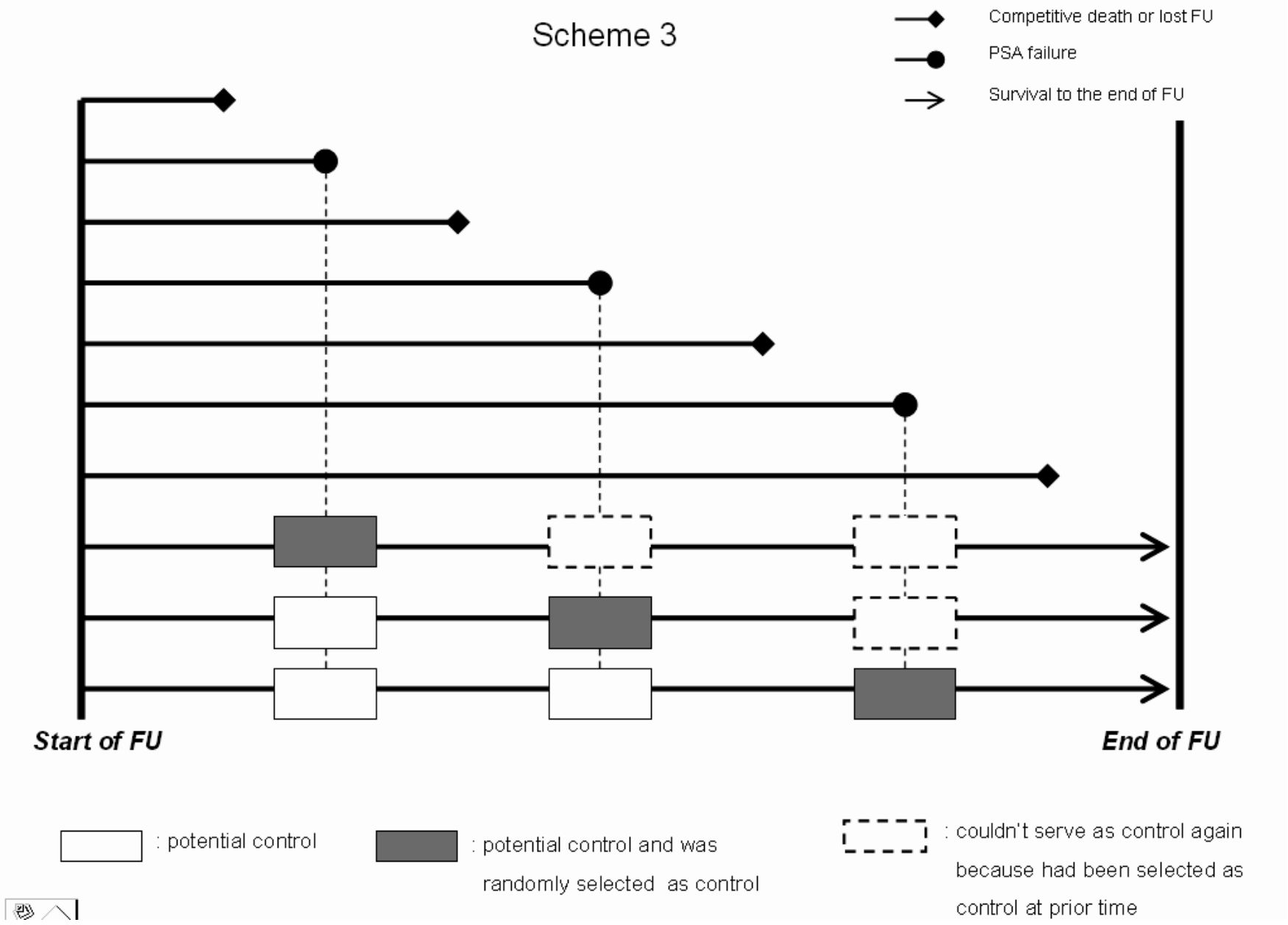


Figure 1 c. Diagram of control selection scheme 2, "pure control sampling without replacement", a biased method.


Scheme 3
Competitive death or lost FU
PSA failure
Survival to the end of FU
Start of FU
End of FU
: potential control
: potential control and was randomly selected as control
: couldn't serve as control again because had been selected as control at prior time