Published in final edited form as:

Stat Biosci. 2013 November 1; 5(2): . doi:10.1007/s12561-013-9095-8.

An overview of Ross Prentice's contributions to statistical science

D. Y. Lin

Department of Biostatistics University of North Carolina at Chapel Hill

Abstract

Ross Prentice's work has had the most profound impact on the theory and practice of statistics. His research interests range from survival analysis, longitudinal data analysis, epidemiologic designs and analysis, to genomic studies. His contributions are so broad and so deep that it would be impossible to provide a comprehensive review in any limited amount of space. In this commentary, I will attempt to give a brief tour of some of his statistical work, focusing on ten of my favorite papers of his. I will describe the main ideas in those papers and their influence on the directions of statistical research and on the designs and analysis of medical studies. I will mention a few stories along the way.

Keywords

Case-cohort design; case-control studies; competing risks; linear rank tests; logistic regression; longitudinal data; marginal likelihood; measurement error; multivariate failure time data; quadratic exponential model; surrogate endpoints; survival analysis; Women's Health Initiative

Ross Prentice is a towering figure in statistics. He has made pathbreaking contributions to many areas of statistics, especially survival analysis and epidemiological methods. He was the Senior Vice President and Director of the Division of Public Health Sciences at the Fred Hutchinson Cancer Research Center for more than 20 years. He was the principal investigator of the Clinical Coordinating Center of the Women's Health Initiative (WHI). I am not aware of anyone else who has excelled so well in both the theory and practice of statistics as well as in high-level administration. Despite his deep involvements in biomedical applications and his enormous administrative duties, Ross has remained focused throughout his career on the development of the core statistical discipline.

Figure 1 is a Web of Knowledge's Citation Report for Ross Prentice in February of 2013. Several interesting observations can be made from this report. Clearly, Ross became more productive as he got older: his publications peaked in the last few years. Not surprisingly, the citations of his papers grew exponentially. Ross had a h -index of 66. A scientist has index h if h of his/her papers have at least h citations each. A h -index of 66 is extraordinary for a statistician. Not many other statisticians get close to that. There are two major reasons that Ross' work has been so highly cited. First, many of his statistical papers provided fundamentally new and practically useful methods, so they have been frequently cited in both statistical and applied journals. Second, Ross has played a leadership role in the WHI and other important epidemiologic studies, and there are many high-impact publications out of those studies.

Figure 2 shows the ten most cited papers written by Ross. Half of those papers, including the two most cited ones, reported findings from the WHI. The top two WHI papers contain the results of the WHI Hormone Therapy clinical trials, which have had a profound impact on public health in the US. The other five papers provided new statistical concepts and methods that have immediate applications in clinical and epidemiologic studies.

In this commentary, I describe my top 10 statistical papers of Ross Prentice. I picked the papers not only on the basis of citations, but also on the basis of intellectual content. I also try to cover as many areas of Ross' work as possible. I describe below the 10 papers and related work, telling a few stories along the way.

10. Kalbfleisch, J. D. and Prentice, R. L. (1972). Discussion of the paper by D. R. Cox. *J. R. Statist. Soc., B*, **34**, 215-216.

Sir David Cox's 1972 *JRSS(B)* discussion paper on proportional hazards regression is the cornerstone of modern survival analysis. Nowadays, the Royal Statistical Society posts the preprints of discussion papers online, and everyone can contribute to the discussion in person or in writing. Back in the early 1970's, however, there was no internet or email, and copying machine was in its infancy. In order to discuss a paper, one had to get a hard copy somehow. Marvin Zelen was able to get hold of Cox's paper because of his early work on exponential regression with Polly Feigl (Feigl and Zelen, 1965). Marvin took the paper with him on a ski trip in the winter of 1971. When he was reading the paper, Jack (Kalbfleisch) and Ross, who were also on the ski trip, asked him what he was reading. Marvin was kind enough to share the paper with the two of them, who ended up being key discussants of the paper.

The proportional hazards model (Cox, 1972) specifies that the hazard function of failure time T conditional on a p -vector of covariates Z takes the form

$$\lambda(t|Z) = e^{\beta^T Z} \lambda_0(t),$$

where β is a vector of unknown regression parameters, and $\lambda_0(\cdot)$ is an arbitrary baseline hazard function. For a random sample of size n , the data consist of $(\tilde{T}_i, \Delta_i, Z_i)$ ($i=1, \dots, n$), where $\tilde{T}_i = \min(T_i, C_i)$, $\Delta_i = I(T_i \leq C_i)$, C_i is the censoring time for the i th subject, and $I(\cdot)$ is the indicator function. To eliminate the infinite-dimensional nuisance parameter $\lambda_0(\cdot)$ in the estimation of β , Cox (1972) proposed to use the "conditional" likelihood

$$\prod_{i=1}^n \left(\frac{e^{\beta^T Z_i}}{\sum_{j \in \mathcal{R}_i} e^{\beta^T Z_j}} \right)^{\Delta_i}, \quad (1)$$

where $\mathcal{R}_i = \{j: \tilde{T}_j \geq \tilde{T}_i\}$.

During the discussion of Cox's paper, several researchers questioned the interpretation of (1) as a conditional likelihood. Kalbfleisch and Prentice (1972) pointed out that it is rather the marginal likelihood of the ranks. Breslow (1972) suggested an alternative estimation approach which yields Cox's estimator for β as well as a simple estimator for the cumulative baseline hazard function. In response to the discussion of his paper, Cox (1975) later introduced the concept of partial likelihood and showed that (1) is in fact a partial likelihood. The concept of partial likelihood, although ingenious, has little usage outside of

the Cox regression model. The derivation of (1) as the marginal likelihood of the ranks plays a useful role in the next paper.

9. Prentice, R. L. (1978). Linear rank tests with right censored data. *Biometrika*, **65**, 167-179.

Consider the linear regression model

$$Y = \beta^T Z + \epsilon, \quad (2)$$

where ϵ is an error variable with an unknown distribution. When $Y = \log T$, (2) is called the accelerated failure time model to reflect the multiplicative effects of Z on T .

Prentice (1978) developed linear rank statistics for testing with censored data. The statistics were derived as score statistics based on the marginal probability of a generalized rank vector. The statistics take the quadratic form

$$Q(\beta) = U(\beta)^T V(\beta)^{-1} U(\beta),$$

where U is the score function, and V is the observed Fisher information matrix. Under the null hypothesis that $\beta = \beta_0$, the statistic $Q(\beta_0)$ has an asymptotic χ^2 distribution with p degrees of freedom.

Prentice (1978) suggested to obtain rank estimators of β by minimizing $Q(\beta)$ or solving the equation $U(\beta) = 0$. Because it is a rank statistic, $U(\beta)$ is a discrete function of β . The discrete nature of $U(\beta)$ makes the estimation very challenging both theoretically and numerically. The asymptotic properties of rank estimators were successfully established in the early 1990's (Tsiatis, 1990; Lai and Ying, 1991). The asymptotic variances of rank estimators involve the density of T which is difficult to estimate well with censored data.

In 1989, L. J. Wei, Zhiliang Ying and I attempted to bypass the estimation of the asymptotic variances when making inferences about individual components or subsets of β . Suppose that β comprises of β_1 and β_2 and it desired to test the hypothesis that $\beta_1 = \beta_1^0$. L. J. suggested to minimize $Q(\beta)$ over β_2 while setting $\beta_1 = \beta_1^0$. Before we submitted our paper (Wei, Ying and Lin, 2000) to *Biometrika*, I took a close look at Prentice's (1978) paper. I was surprised to find that our test was mentioned in the second paragraph of Section 4.3 of that paper (in two sentences). There was no formal theoretical justification, but the idea was there!

8. Prentice, R. L. (1986). A case-cohort design for epidemiologic cohort studies and disease prevention trials. *Biometrika*, **73**, 1-11.

The score function corresponding to the partial likelihood given in (1) is

$$U(\beta) = \sum_{i=1}^n \Delta_i \left\{ Z_i - \bar{Z}(\beta; \tilde{T}_i) \right\},$$

where

$$\overline{Z}(\beta; t) = \frac{\sum_{j \in \mathcal{R}_i} e^{\beta^T Z_j} Z_j}{\sum_{j \in \mathcal{R}_i} e^{\beta^T Z_j}}.$$

Note that the censored observations contribute only to the calculation of $\overline{Z}$. Because $\overline{Z}(\beta, t)$ is a weighted average of Z at time t , it can be estimated with 500 subjects nearly as well as with 10,000 subjects. This fact motivated Ross to propose the case-cohort design, under which expensive covariates are measured on all cases (i.e., uncensored subjects) and on a subset of subjects randomly selected from the entire cohort (Prentice, 1986). This design provides a cost-effective strategy to conduct large cohort studies with rare events and expensive covariate measurements and is particularly useful in large-scale biomarker studies and genetic association studies.

7. Prentice, R. L. (1992). Covariate measurement errors and parameter estimation in a failure time regression model. *Biometrika*, **69**, 331-342.

Ross has a long-standing interest in measurement errors because of his research in nutrition and dietary factors. He pioneered the study of measurement errors in proportional hazards regression and has continued to work in that area. In the 1992 paper, he derived the conditional hazard function of the failure time given the mismeasured covariate. The corresponding partial likelihood involves the baseline hazard function. Ross provided useful inference procedures for testing zero regression parameters and for estimation in important special cases, such as rare events. This paper is one of the earliest applications of the regression calibration method.

6. Prentice, R. L., Kalbfleisch, J. D., Peterson, Jr. A. V., Flournoy, N., Farewell, V. T. and Breslow, N. E. (1978). The analysis of failure times in the presence of competing risks. *Biometrics*, **34**, 541-554.

Competing risks arise when there is more than one distinct cause or type of failure. The traditional formulation of this problem is in terms of latent failure times for each failure type. That approach makes strong assumptions and lacks physical interpretations. Prentice et al. (1978) advocated the use of cause-specific hazard functions, which they showed are the basic estimable quantities. They argued that the estimation of failure rates under the removal of certain causes is not a well-posed problem unless a mechanism for cause removal is specified. This paper has become a standard reference on competing risks.

5. Prentice, R. L. (1989). Surrogate endpoints in clinical trials: definition and operational criteria. *Statistics in Medicine*, **8**, 431-440.

Motivated by the desire for rapid discoveries of effective treatments for HIV/AIDS, there was a tremendous interest in using surrogate endpoints in clinical trials in the late 1980's. Replacing a rare or late-occurring clinical endpoint (e.g., time to development of AIDS or death) with a frequent or short-term outcome variable (e.g., CD4 counts or viral loads) can lead to substantial reduction in sample size and trial duration. There was a misconception that an outcome variable that is highly correlated with the clinical endpoint would be a good surrogate endpoint. Prentice (1989) showed that for the treatment comparison based on a surrogate response variable to have an unequivocal implication for the corresponding true endpoint treatment comparison, a test of the null hypothesis of no treatment difference in the surrogate endpoint should also be a valid test of the corresponding null hypothesis based on the true endpoint. This definition, now commonly known as the Prentice criterion,

essentially requires the surrogate variable to capture any relationship between the treatment and the true endpoint. The Prentice criterion is unlikely to be fully satisfied in practice and provides a cautionary note about the use of surrogate endpoints. This paper is Ross' most cited statistical paper. It was actually a discussion of three papers presented at the 1987 International Biometric Society ENAR Spring Meeting.

4. Prentice, R. L. and Zhao, L. P. (1991). Estimating equations for parameters in means and covariances of multivariate discrete and continuous responses. *Biometrics*, **47**, 825-539.

There was a tremendous interest in regression analysis of multivariate responses in the 1980's, and that interest remains strong today. A major approach is the generalized estimating equations (GEE), which formulates the marginal distributions of multivariate responses with generalized linear models and estimates regression parameters through quasi-likelihood score equations (Liang and Zeger, 1986). The approach is focused on mean parameters. In many applications involving multivariate responses, investigators are interested in the dependence of covariances on covariates. In a series of papers, including Prentice and Zhao (1991), Ross used a class of quadratic exponential models to develop joint estimating equations for mean and covariance parameters. This approach is known as GEE2.

3. Prentice, R. L. and Hsu, L. (1997). Regression on hazard ratios and cross ratios in multivariate failure time analysis. *Biometrika* **84**, 349-363.

Multivariate failure time data arise when a study subject may experience multiple events or when there exists natural or artificial clustering such that the failure times within the same cluster are correlated. Along with the interest in (non-censored) multivariate responses, there has been a strong interest in multivariate failure times. Regression analysis of multivariate failure time data has been one of Ross' major research interests in recent years. It is highly challenging to develop efficient estimators for semiparametric regression models with censored data. In a similar spirit to GEE2, Prentice and Hsu (1997) developed joint estimating equations for hazard ratio and correlation parameters.

2. Prentice, R. L. and Cai, J. (1992). Covariance and survivor function estimation using censored multivariate failure time data. *Biometrika* **79**, 495-512.

Ross is very passionate about the estimation of the multivariate survival function, which is a notoriously hard problem in survival analysis. While vacationing in Hawaii, Ross conceived the idea behind the Prentice-Cai (1992) estimator. The estimator is based on a clever representation of the bivariate survival function in terms of the marginal survival functions and a covariance function between counting process martingales. Ross was not satisfied with the estimator because it may assign negative mass, so he has continued to work on this problem.

Ross and I participated in a workshop on *Survival and Event History Analysis* organized by Richard Cook and Jerry Lawless in May, 2011. When I ran into Ross at the workshop, I asked him what he was up to these days. He said that he was still working on multivariate survival function estimation. He pulled a manuscript out of his suitcase. It was a paper he had been working on with a graduate student entitled "Nonparametric maximum likelihood estimation of multivariate survivor function". It was a very rough draft with Ross' pencil marks all over it. He said that the approach was promising although he had not figured out how to implement it numerically or prove the asymptotic theory yet. He speculated that it would be possible to derive efficient estimators for the regression analysis of multivariate failure time data once he solves the multivariate survival function estimation problem.

1. Prentice, R. L. and Pyke, R. (1979). Logistic disease incidence models and case-control studies. *Biometrika*, **66**, 403-411.

The case-control design is a primary tool for the study of factors related to disease incidence. Under this design, diseased cases and disease-free controls are identified and “followed back” to ascertain their exposure levels. Prentice and Pyke (1979) showed that the odds ratio estimators and their asymptotic covariance matrix may be obtained by applying the usual logistic regression model to the case-control study as if the data had been collected in a prospective study. This remarkable discovery allows epidemiologists to use standard logistic regression as implemented in commercial software packages to analyze their case-control data. Although the result is very simple and elegant, the derivation was very tedious. The derivation can be greatly simplified by using the profile likelihood theory.

I asked Ross to provide some background information on this famous paper. The following is his reply:

Ron (Pyke) asked to spend some time at the Hutch during the summer of 1978. I think he was looking for some diversion from teaching responsibilities. He was a delightful visitor. When he arrived, I had worked out the type of NPMLE argument that appears in the paper. I asked Ron for help with the associated asymptotic distribution theory. He thought about it for a while and formulated it as a k -sample problem from which the asymptotics flowed pretty readily, but the variance estimator form took some effort. I worked that through in a brute force manner, while Ron worked on a more elegant approach. For some years, he planned to lead a follow-up paper to include this nicer approach, but for some reason it didn't happen. Ron referred to this particular paper as RP^2 , given our initials. It was a brief collaboration, but one of my most enjoyable — with another mathematically oriented statistician from Ontario, Canada with initials of RP.

I asked Ross whether the “more elegant approach” Ron worked on is related to profile likelihood. Ross did not think it relied on profile likelihood directly. We will never find out Ron's approach since he has passed away.

The papers I selected obviously do not encompass all of Ross' statistical work. I did not attempt to describe all the papers Ross has written on any topic, let alone provide a comprehensive review of relevant work by others. This commentary is meant to be informal and fun.

Ross' contributions to statistics go well beyond his research papers. The biostatistics program he has built at the Fred Hutchinson Cancer Research Center is the envy of the world. Many of his former students have become extraordinary leaders in our field. Ross' achievements have greatly advanced the field of statistics and inspired generations of statisticians. It was my honor and pleasure to prepare this brief commentary.

Acknowledgments

This paper is a written account of my presentation at the Prentice Symposium. A more entertaining version is available at my website: <http://dlin.web.unc.edu/>. This work was supported by the National Institutes of Health awards R37GM047845 and R01CA082659. I thank Ross for clarifying a few details during this writing. I also thank Charles Kooperberg for reading the paper and providing helpful comments.

References

- Breslow NE. Cox DR. J. R. Statist. Soc. B. 1972; 34:216–217. Discussion of the paper.
Cox DR. Regression models and life-tables (with discussion). J. R. Statist. Soc. B. 1972; 34:187–220.

- Cox DR. Partial likelihood. *Biometrika*. 1975; 62:269–276.
- Feigel P, Zelen M. Estimation of exponential survival probabilities with concomitant information. *Biometrics*. 1965; 21:826–833. [PubMed: 5857231]
- Kalbfleisch JD, Prentice RL. Cox DR. J. R. Statist. Soc. B. 1972; 34:215–216. Discussion of the paper.
- Lai TL, Ying Z. Rank regression methods for left-truncated and right-censored data. *Ann. Statist.* 1991; 19:531–556.
- Liang K-Y, Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika*. 1986; 73:13–22.
- Prentice RL. Linear rank tests with right censored data. *Biometrika*. 1978; 65:167–179.
- Prentice RL. A case-cohort design for epidemiologic cohort studies and disease prevention trials. *Biometrika*. 1986; 73:1–11.
- Prentice RL. Surrogate endpoints in clinical trials: definition and operational criteria. *Statistics in Medicine*. 1989; 8:431–440. [PubMed: 2727467]
- Prentice RL. Covariate measurement errors and parameter estimation in a failure time regression model. *Biometrika*. 1992; 69:331–342.
- Prentice RL, Cai J. Covariance and survivor function estimation using censored multivariate failure time data. *Biometrika*. 1992; 79:495–512.
- Prentice RL, Hsu L. Regression on hazard ratios and cross ratios in multivariate failure time analysis. *Biometrika*. 1997; 84:349–363.
- Prentice RL, Kalbfleisch JD, Peterson AV Jr, Flournoy N, Farewell VT, Breslow NE. The analysis of failure times in the presence of competing risks. *Biometrics*. 1978; 34:541–554. [PubMed: 373811]
- Prentice RL, Pyke R. Logistic disease incidence models and case-control studies. *Biometrika*. 1979; 66:403–411.
- Prentice RL, Zhao LP. Estimating equations for parameters in means and covariances of multivariate discrete and continuous responses. *Biometrics*. 1991; 47:825–539. [PubMed: 1742441]
- Tsiatis AA. Estimating regression parameters using linear rank tests for censored data. *Ann. Statist.* 1990; 18:354–372.
- Wei LJ, Ying Z, Lin DY. Linear regression analysis of censored survival data based on rank tests. *Biometrika*. 1990; 77:845–851.

Figure 1.
Ross Prentice's publications and citations according to Web of Knowledge's Citation Report.

Citation report, ranked by times cited from highest to lowest		2009	2010	2011	2012	2013	Total	Average Citations per Year
		2162	2050	2054	1940	221	27053	659.10
1.	Risks and benefits of estrogen plus progestin in healthy postmenopausal women - Principal results from the Women's Health Initiative randomized controlled trial Authors: Rossouw, JE; Anderson, GL; Prentice, RL; et al. Journal of the American Medical Association Volume: 288 Issue: 3 Pages: 321-333 Published: JUL 17 2002	643	591	564	492	54	7315	609.67
2.	Effects of conjugated, equine estrogen in postmenopausal women with hysterectomy - The Women's Health Initiative randomized controlled trial Authors: Anderson, GL; Limacher, M; Assaf, AR; et al. Group Authors: WHI Steering Comm Journal of the American Medical Association Volume: 291 Issue: 14 Pages: 1701-1712 Published: APR 14 2004	209	189	197	154	19	2073	207.40
3.	Surrogate endpoints in clinical trials - definition and operational criteria Author: Prentice, RL Statistics in Medicine Volume: 8 Issue: 4 Pages: 431-440 DOI: 10.1002/sim.4780080407 Published: APR 1989	77	61	69	55	3	767	30.68
4.	Design of the Women's Health Initiative clinical trial and observational study Authors: Anderson, GL; Cummings, S; Freedman, LR; et al. Group Authors: Women's Health Initiative Study Group Controlled Clinical Trials Volume: 19 Issue: 1 Pages: 61-109 Published: FEB 1998	48	56	502	65	5	648	40.50
5.	Calcium plus vitamin D supplementation and the risk of fractures Authors: Jackson, RD; LaCroix, AZ; Gass, M; et al. Group Authors: Women's Health Initiative Investigators New England Journal of Medicine Volume: 354 Issue: 7 Pages: 659-663 DOI: 10.1056/NEJMo055218 Published: FEB 16 2006	77	98	75	77	8	600	75.33
6.	Postmenopausal hormone therapy and risk of cardiovascular disease by age and years since menopause Authors: Rossouw, JE; Prentice, RL; Manson, JoAnn E; et al. Journal of the American Medical Association Volume: 297 Issue: 13 Pages: 1465-1477 DOI:10.1001/jama.297.13.1465 Published: APR 4 2007	133	102	98	91	12	593	84.71
7.	Analysis of failure times in presence of competing risks Authors: Prentice, RL; Kalbfleisch, JD; Peterson, RV; et al. Biometrika Volume: 34 Issue: 4 Pages: 541-554 DOI: 10.2307/2330374 Published: 1978	35	43	39	39	9	548	15.22
8.	A case-cohort design for epidemiologic cohort studies and disease prevention trials Author: Prentice, RL Biometrika Volume: 73 Issue: 1 Pages: 1-11 Published: APR 1986	30	31	27	38	5	531	19.00
9.	Regression-analysis of grouped survival data with application to breast cancer data Authors: Prentice, RL; Gloeckler, LA Biometrics Volume: 34 Issue: 1 Pages: 57-67 DOI: 10.2307/2529588 Published: 1978	35	33	21	31	3	511	17.19
10.	Logistic disease incidence models and case-control studies Authors: Prentice, RL; Pyke, R Biometrika Volume: 66 Issue: 3 Pages: 403-411 DOI: 10.2307/2335158 Published: 1979	34	39	30	41	3	500	14.29

Figure 2.
Ross Prentice’s most cited papers according to Web of Knowledge’s Citation Report.