Biostatistics (2010), **11**, 3, *pp.* 413–418
doi:10.1093/biostatistics/kxq004
Advance Access publication on February 19, 2010

Transparency and reproducibility in data analysis: the Prostate Cancer Prevention Trial

STUART G. BAKER*

Biometry Research Group, Division of Cancer Prevention, National Cancer Institute, Executive Plaza North 3131, 6130 Executive Boulevard, Mail Stop Code 7354, Bethesda, MD 20892-7354, USA
sb16i@nih.gov

AMY K. DARKE

Statistical Center, Fred Hutchinson Cancer Research Center, Seattle, WA 98109, USA

PAUL PINSKY

Early Detection Research Group, Division of Cancer Prevention, National Cancer Institute, Bethesda, MD 20892, USA

HOWARD L. PARNES

Prostate and Urologic Cancer Research Group, Division of Cancer Prevention, National Cancer Institute, Bethesda, MD 20892, USA

BARNETT S. KRAMER

Office of Disease Prevention, National Institutes of Health, Bethesda, MD 20892, USA

SUMMARY

With the analysis of complex, messy data sets, the statistics community has recently focused attention on “reproducible research,” namely research that can be readily replicated by others. One standard that has been proposed is the availability of data sets and computer code. However, in some situations, raw data cannot be disseminated for reasons of confidentiality or because the data are so messy as to make dissemination impractical. For one such situation, we propose 2 steps for reproducible research: (i) presentation of a table of data and (ii) presentation of a formula to estimate key quantities from the table of data. We illustrate this strategy in the analysis of data from the Prostate Cancer Prevention Trial, which investigated the effect of the drug finasteride versus placebo on the period prevalence of prostate cancer. With such an important result at stake, a transparent analysis was important.

Keywords: Categorical data; Maximum likelihood; Missing data; Multinomial–Poisson transformation; Propensity-to-be-missing score; Randomized trials.

[*] To whom correspondence should be addressed.

Published by Oxford University Press 2010.

1. INTRODUCTION

The Prostate Cancer Prevention Trial (PCPT) investigated the effect of finasteride versus placebo on the period prevalence of prostate cancer determined by biopsy (Thompson *and others*, 2003). The PCPT randomized 18 882 men to receive either placebo or finasteride over 7 years. At annual intervals, men in both arms were scheduled to receive tests for prostate-specific antigen (PSA) and also a digital rectal examination (DRE). If PSA was elevated or DRE indicated an abnormal finding, a biopsy was recommended. A biopsy was also offered at the end of the study to all living participants who had not been previously diagnosed with prostate cancer. A nonrandom sample of participants with prostate cancer on biopsy received surgery. The original endpoint was prostate cancer on biopsy. However, due to problems with interpreting this endpoint, it was later decided that a more relevant endpoint was prostate cancer determined on surgery and classified as either low grade (less severe) or high grade (more severe) based on a Gleason score (GS). The analysis is complicated because endpoint data were missing in 2 stages, biopsy and surgery following biopsy, both of which depended on whether or not biopsy was recommended and the grade on biopsy.

Three recent papers proposed different ways to estimate the effect of finasteride on high-grade prostate cancer ascertained by surgery in the presence of missing data (Pinsky *and others*, 2008; Redman *and others*, 2008; Shepherd *and others*, 2008). However, none of these papers was conducive to reproducibility. A reproducible and transparent analysis is important because the data are very complex and the results could affect a large number of men who currently receive finasteride for benign conditions.

One standard for reproducibility is the availability of data sets and computer code (Peng, 2009). However, due to confidentiality considerations and the messiness of the data, it was not feasible to disseminate the raw data. Instead, we chose another approach to reproducible research, namely presenting tables of counts and reporting simple closed-form maximum likelihood estimates. Please also see the supplementary material available at *Biostatistics* online for additional clinical background and important documentation for reproducible research on definitions, classification of subjects, derivations, and calculations.

2. DATA TABLES

We define

$X =$ randomization group with $x = 0$ if placebo and $x = 1$ if finasteride,

$A =$ indicator of biopsy recommendation with $a = 1$ if positive recommendation and 0 otherwise,

$Y =$ the prostate cancer outcome if everyone were biopsied, with $y = 0$ if no cancer, $y = 1$ if low-grade cancer, and $y = 2$ if high-grade cancer,

$D =$ “definitive” (for this study) prostate cancer outcome with $d = 0$ if no cancer on biopsy, 1 if low-grade prostate cancer on surgery, and 2 if high-grade prostate cancer on surgery,

$M_Y =$ 0 if missing the biopsy and 1 otherwise,

$M_D =$ 0 if missing the surgery and 1 otherwise.

The data involve 3 types of counts:

$m_{xa} =$ the number of persons in randomization group x with biopsy recommendation a, missing a biopsy outcome and missing a surgery outcome,

$n_{xay} =$ the number of persons in randomization group x with biopsy recommendation a, biopsy outcome y, and missing surgery outcome,

$k_{xayd} =$ the number of persons in randomization group x with biopsy recommendation a, biopsy outcome y, and surgery outcome d.

Because persons who had surgery must have had prostate cancer on biopsy and hence had to be diagnosed with prostate cancer at surgery, k_{xayd} is only defined for $y \neq 0$ and $d \neq 0$. Investigators were interested in 2 possible definitions for high-grade prostate cancer, either a GS greater than or equal to 7 or a GS greater than or equal to 8 (Tables 1 and 2).

Table 1. *Data from PCPT using GS greater than or equal to 7 as the definition of high-grade prostate cancer*

Persons without a biopsy and without surgery (m_{xa})		
Randomization group (x)	Biopsy recommendation (a)	
Placebo	No	3955
	Yes	215
Finasteride	No	4169
	Yes	214

Persons with biopsy but no surgery (n_{xay})				
Randomization	Biopsy	Biopsy outcome (y)		
group (x)	recommendation (a)	No cancer	GS $<$ 7	GS $\geqslant$ 7
Placebo	No	3675	417	78
	Yes	479	249	104
Finasteride	No	3791	230	75
	Yes	458	145	132

Persons with biopsy and surgery (k_{xayd})				
Randomization	Biopsy	Surgery outcome (d)		
group (x)	recommendation (a)	Biopsy outcome (y)	GS $<$ 7	GS $\geqslant$ 7
Placebo	No	GS $<$ 7	83	19
		GS $\geqslant$ 7	8	13
	Yes	GS $<$ 7	83	33
		GS $\geqslant$ 7	9	46
Finasteride	No	GS $<$ 7	50	6
		GS $\geqslant$ 7	7	14
	Yes	GS $<$ 7	44	23
		GS $\geqslant$ 7	10	55

3. A TRANSPARENT ANALYSIS

We analyzed the data in the tables using a transparent approach involving closed-form maximum likelihood estimates. The parameter of interest is the probability of "definitive" prostate cancer outcome d conditional on randomization group x,

$$\beta_{d|x} = \mathrm{pr}(D = d|x). \tag{3.1}$$

Other parameters are the probability of cancer outcome y on biopsy conditional on randomization group x and definitive prostate cancer outcome d,

$$\lambda_{y|xd} = \mathrm{pr}(Y = y|x, d), \tag{3.2}$$

and the probability of biopsy recommendation a conditional on randomization group x, biopsy outcome y, and definitive prostate cancer outcome d,

$$\phi_{a|xyd} = \mathrm{pr}(A = a|x, y, d). \tag{3.3}$$

We make 2 reasonable assumptions.

ASSUMPTION 3.1 The probability of missing a biopsy outcome depends on randomization group x and biopsy recommendation a, but not biopsy outcome y, namely $\pi_{xa} = \mathrm{pr}(M_Y = 0|\ x, a) = \mathrm{pr}(M_Y = 0|\ x, a, y)$.

Table 2. *Data from PCPT using GS greater than or equal to* 8 *as the definition of high-grade prostate cancer*

Persons without a biopsy and without surgery (m_{xa})				
Randomization group (x)	Biopsy recommendation (a)			
Placebo	No	3955		
	Yes	215		
Finasteride	No	4169		
	Yes	214		
Persons with biopsy but no surgery (n_{xay})				
Randomization	Biopsy	Biopsy outcome (y)		
group (x)	recommendation (a)	No cancer	GS $<$ 8	GS $\geqslant$ 8
Placebo	No	3675	488	7
	Yes	479	316	37
Finasteride	No	3791	291	14
	Yes	458	231	46
Persons with biopsy and surgery (k_{xayd})				
Randomization	Biopsy	Surgery outcome (d)		
group(x)	recommendation (a)	Biopsy outcome (y)	GS $<$ 8	GS $\geqslant$ 8
Placebo	No	GS $<$ 8	120	0
		GS $\geqslant$ 8	3	0
	Yes	GS $<$ 8	152	4
		GS $\geqslant$ 8	5	10
Finasteride	No	GS $<$ 8	70	1
		GS $\geqslant$ 8	5	1
	Yes	GS $<$ 8	101	6
		GS $\geqslant$ 8	12	12

ASSUMPTION 3.2 The probability of missing a surgery outcome following a biopsy depends on randomization group x, biopsy recommendation a, and outcome of biopsy y, but not the outcome of the surgery d, namely $\gamma_{xay} = \text{pr}(M_D = 0|\, x, a, y, M_Y = 1) = \text{pr}(M_D = 0|\, x, a, y, d, M_Y = 1)$.

Because the model is saturated, we computed maximum likelihood estimates by setting observed cell counts equal to their expected values. Let "+" in a subscript denote summation. The maximum likelihood estimate of the relative risk (RR) for finasteride versus placebo group of high-grade prostate cancer ($d = 2$) is

$$\text{RR} = \widehat{\beta}_{2|1}/\widehat{\beta}_{2|0}, \text{ where}$$

$$\widehat{\beta}_{2|x} = \left(\sum_{a=0}^{1} \sum_{y=1}^{2} \upsilon_{xay} \right) / N_x,$$

$$N_x = k_{x+++} + n_{x++} + m_{x+} = \text{ number in group } x,$$

$$\upsilon_{xay} = \frac{k_{xay2}}{(1 - \widehat{\pi}_{xa})(1 - \hat{\gamma}_{xay})} = \text{ imputed number with } x, a, y, \text{ who have } d = 2,$$

Table 3. *Estimated RRs for high-grade prostate cancer for finasteride versus placebo (95% confidence intervals)*

	Definition of high-grade prostate cancer as GS	
	7 or above	8 or above
Our method unadjusted[†]	0.83 (0.65, 1.05)	1.53 (0.85, 2.75)
Adjusted[‡]	0.82 (0.64, 1.06)	1.40 (0.71, 2.76)
Redman *and others* (2008)	0.73 (0.56,0.96)	1.25[§]
Pinsky *and others* (2008)	0.84 (0.58, 1.06)	1.39 (0.79, 2.50)

[†] Based on data from Tables 1 and 2.

[‡] Based on family history, race, age. See supplementary material available at *Biostatistics* online.

[§] No confidence interval reported, said to be imprecise.

$$\widehat{\pi}_{xa} = \frac{m_{xa}}{k_{xa++} + n_{xa+} + m_{xa}} = \text{ fraction with } x \text{ and } a \text{ who are missing biopsy,}$$
$$\hat{\gamma}_{xay} = \frac{n_{xay}}{k_{xay+} + n_{xay}} = \text{ fraction biopsied with } x, a, y \text{ who are missing surgery.} \tag{3.4}$$

The estimated variance of RR was computed using the multinomial–Poisson transformation (Baker, 1994). We also computed estimates after adjusting for baseline variables using propensity-to-be-missing scores (Baker *and others*, 2006).

4. Results and conclusions

Based on the results in Table 3, we draw 2 clinically important conclusions. First, finasteride likely lowers the risk of high-grade prostate cancer defined as a GS of 7 or greater, with a point estimate indicating a lower risk under finasteride and an upper bound of the 95% confidence interval indicating a very slight increase in risk (hence borderline statistical significance). Second, finasteride possibly increases the risk of high-grade prostate cancer defined as a GS of 8 or greater, with a point estimate indicating higher risk under finasteride but a lower bound of the 95% confidence interval indicating a substantial decrease in risk (hence statistically not significant). The main caveats, shared by all analyses of the PCPT data, are unmeasured confounders, misclassification of data, and lack of a definitive prostate cancer mortality endpoint. Although the results are qualitatively similar to previous results, it does not diminish the clinical importance of this analysis. Because our basic analysis is transparent and reproducible, it should carry extra weight with the clinical community.

Supplementary material

Supplementary material is available at http://biostatistics.oxfordjournals.org.

Acknowledgments

We thank Ian Thompson, the principal investigator for PCPT, for facilitating our use of the data, Cathy Tangen and Phyllis Goodman for assistance in processing the data, and Mary Redman for helpful comments. *Conflict of Interest:* None declared.

FUNDING

National Cancer Institute, Division of Cancer Prevention (5U10CA037429-25); Southwest Oncology Group CCOP Research Base, PCPT/PCPT LTFU, and SELECT Programs; Division of Cancer Prevention in the National Cancer Institute and the National Institutes of Health.

REFERENCES

BAKER, S. G. (1994). The multinomial-Poisson transformation. *The Statistician* **43**, 495–504.

BAKER, S. G., FITZMAURICE, G., FREEDMAN, L. S. AND KRAMER, B. S. (2006). Simple adjustments for randomized trials with nonrandomly missing or censored outcomes arising from informative covariates. *Biostatistics* **7**, 29–40.

PINSKY, P., PARNES, H. AND FORD, L. (2008). Estimating rates of true high-grade disease in the Prostate Cancer Prevention Trial. *Cancer Prevention Research* **1**, 182–186.

PENG, R. D. (2009). Reproducible research and Biostatistics. *Biostatistics* **10**, 405–408.

REDMAN, M. W., TANGEN, C. M., GOODMAN, P. J., LUCIA, S., COLTMAN, JR, C. A. AND THOMPSON, I. M. (2008). Finasteride does not increase the risk of high-grade prostate cancer: a bias-adjusted modeling approach. *Cancer Prevention Research* **1**, 174–181.

SHEPHERD, B. E., REDMAN, M. W. AND ANKERST, D. P. (2008). Does finasteride affect the severity of prostate cancer? A causal sensitivity analysis. *Journal of the American Statistical Association* **103**, 1392–1404.

THOMPSON, I. M., GOODMAN, P. J., TANGEN, C. M., LUCIA, M. S. MILLER, G. J., FORD, L. G., LIEBER, M. M., Cespedes, R. D., Atkins, J. N., Lippman, S. M. *and others* (2003). The influence of finasteride on the development of prostate cancer. *The New England Journal of Medicine* **349**, 215–224.

[*Received August 13, 2009; revised November 12, 2009; accepted for publication January 11, 2010*]