The Bio in Biostatistics

James A. Hanley

Dept. of Epidemiology, Biostatistics & Occupational Health
Faculty of Medicine, McGill University

SSC Impact Award Address
Annual Meeting of Statistical Society of Canada
Montréal, Québec
2018-06-05

Things They Don't Teach You in Graduate School[1]

James A Hanley

McGill University

Abstract

Much of what statisticians teach and use in practice is learnt 'on the job.' I recount here some of my early statistical experiences, and the lessons we might learn from them. They are aimed at those of you starting out in the profession today, and at the teachers who train you. I stress the importance of communication.

Key Words

communication; communication; communication.

James A. Hanley (http://www.biostat.mcgill.ca/hanley) is Professor, Department of Epidemiology, Biostatistics and Occupational Health, Montreal, Quebec, H3A 1A2, Canada. (email: James.Hanley@McGill.CA). This work was partially supported by the Natural Sciences and Engineering Research Council of Canada, and Le Fonds Québécois de la recherche sur la nature et les technologies. The author thanks Marvin Zelen for helpful comments.

[1] An expansion on some after-dinner remarks made at the Conference of Applied Statisticians of Ireland, held in Killarney, May 17-19, 2006. The article is dedicated to two former colleagues – and superb communicators – Fred Mosteller and Steve Lagakos, who are no longer with us.

2012.07.19

[Reprinted from RADIOLOGY, Vol. 143, No. 1, Pages 29–36, April, 1982.]

The Meaning and Use of the Area under a Receiver Operating Characteristic (ROC) Curve[1]

James A. Hanley, Ph.D.
Barbara J. McNeil, M.D., Ph.D.

TABLE I: Rating of 109 CT Images

True Disease Status	Rating					Total
	Definitely Normal (1)	Probably Normal (2)	Question-able (3)	Probably Abnormal (4)	Definitely Abnormal (5)	
Normal	33	6	6	11	2	$n_N = 58$
Abnormal	3	2	2	11	33	$n_A = 51$
Totals	36	8	8	22	35	109

JOURNAL OF MATHEMATICAL PSYCHOLOGY: 6, 487–496 (1969)

Maximum-Likelihood Estimation of Parameters of Signal-Detection Theory and Determination of Confidence Intervals—Rating-Method Data[1]

DONALD D. DORFMAN[2]

San Diego State College, San Diego, California 92115

AND

EDWARD ALF, JR.

U.S. Naval Personnel Research Activity, San Diego, California 92133

Procedures have been developed for obtaining maximum-likelihood estimates of the parameters of the Thurstonian model for the method of successive intervals. The signal-detection model for rating-method data is a special case of the Thurstonian model with fixed boundaries, in that there are two stimuli rather than an unspecified set. The present paper presents the solution to the two-stimulus case, and in addition, provides procedures for obtaining the variance–covariance matrix and confidence intervals. The expected values of the second partial derivatives are presented in analytic form to ensure accurate computation of the variance–covariance matrix. An application of these methods was employed on some data collected by others.

Dorfman and Alf (1968) recently developed procedures for obtaining maximum-likelihood estimates of the parameters of signal-detection theory from data of yes–no ROC curves. Ogilvie and Creelman (1968) recently developed maximum-likelihood estimates and confidence intervals for the parameters of signal-detection theory from rating-method data, by using the logistic distribution rather than the normal distribution to make the mathematics more tractable. They estimated d' by means of an empirical relation which they obtained between d' and an analogous parameter in the logistic model. This relation was found through numerical experiments on a high-speed computer. Unfortunately, a stable empirical relation could not be found between the sigma ratio of signal-detection theory and the analogous parameter of the logistic model. Consequently, a procedure assuming underlying normal distributions would be preferred. Schönemann and Tucker (1967) developed maximum likelihood

from Encyclopedia of Biostatistics, 2005

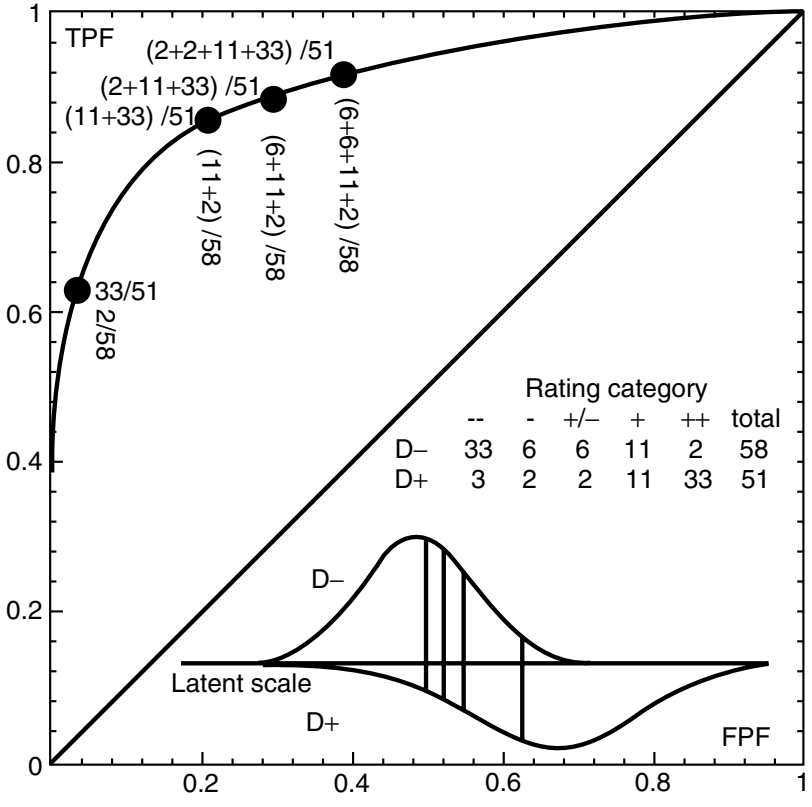
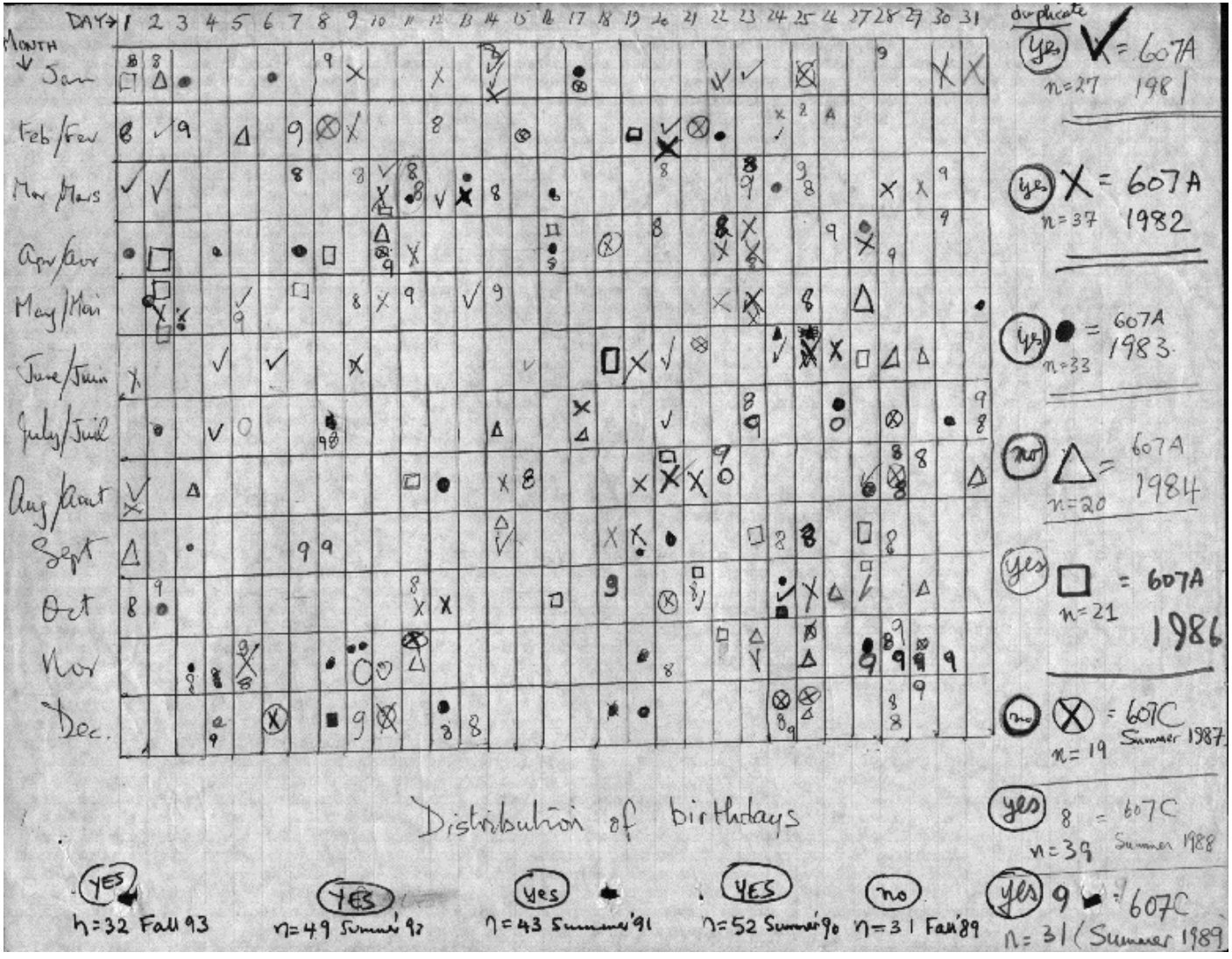


Figure 1 Example of empirical ROC points and smooth curve fitted to them. The empirical points are calculated from successively more liberal definitions of test positivity applied to the 2×5 table (inset) of disease status (D+ or D−) and rating category (−− to ++). The smooth ROC curve is derived from the fitted binormal model (inset, lower right, with parameters $a = 1.657$ and $b = 0.713$ on a continuous latent scale) by using all possible scale values for test positivity. The fitted parameters a and b, together with the four estimated cutpoints, produce fitted frequencies of $\{32.9, 6.4, 5.9, 10.7, 2.1\}$ and $\{3.2, 1.5, 2.1, 11.2, 32.9\}$ for the D− and D+ rows of the 2×5 table. Note that a monotonic transformation of the latent axis may produce overlapping distributions with nonbinormal shapes, but will yield the same multinomial distributions and the same fitted ROC curve

No.s of students in intro. course ’81-’93 (n=7 in ’80) and Distribution of Birthdays



JOURNAL OF MATHEMATICAL PSYCHOLOGY 12, 387–415 (1975)

The Area above the Ordinal Dominance Graph and the Area below the Receiver Operating Characteristic Graph

DONALD BAMBER

Psychology Service, Veterans Administration Hospital, St. Cloud, Minnesota 56301

Receiver operating characteristic graphs are shown to be a variant form of ordinal dominance graphs. The area above the latter graph and the area below the former graph are useful measures of both the size or importance of a difference between two populations and/or the accuracy of discrimination performance. The usual estimator for this area is closely related to the Mann–Whitney U statistic. Statistical literature on this area estimator is reviewed. For large sample sizes, the area estimator is approximately normally distributed. Formulas for the variance and the maximum variance of the area estimator are given. Several different methods of constructing confidence intervals for the area measure are presented and the strengths and weaknesses of each of these methods are discussed. Finally, the Appendix presents the derivation of a new mathematical result, the maximum variance of the area estimator over convex ordinal dominance graphs.

[Reprinted from RADIOLOGY, Vol. 143, No. 1, Pages 29–36, April, 1982.]

James A. Hanley, Ph.D.
Barbara J. McNeil, M.D., Ph.D.

The Meaning and Use of the Area under a Receiver Operating Characteristic (ROC) Curve[1]

A representation and interpretation of the area under a receiver operating characteristic (ROC) curve obtained by the "rating" method, or by mathematical predictions based on patient characteristics, is presented. It is shown that in such a setting the area represents the probability that a randomly chosen diseased subject is (correctly) rated or ranked with greater suspicion than a randomly chosen nondiseased subject. Moreover, this probability of a correct ranking is the same quantity that is estimated by the already well-studied nonparametric Wilcoxon statistic. These two relationships are exploited to (*a*) provide rapid closed-form expressions for the approximate magnitude of the sampling variability, *i.e.*, standard error that one uses to accompany the area under a smoothed ROC curve, (*b*) guide in determining the size of the sample required to provide a sufficiently reliable estimate of this area, and (*c*) determine how large sample sizes should be to ensure that one can statistically detect differences in the accuracy of diagnostic techniques.

TABLE I: Rating of 109 CT Images

True Disease Status	Rating					Total
	Definitely Normal (1)	Probably Normal (2)	Question-able (3)	Probably Abnormal (4)	Definitely Abnormal (5)	
Normal	33	6	6	11	2	$n_N = 58$
Abnormal	3	2	2	11	33	$n_A = 51$
Totals	36	8	8	22	35	109

TABLE II: Computation of W and Its Standard Error

Row	Contents	Column (Rating) x = 1	x = 2	x = 3	x = 4	x = 5	Total	Remarks
1	Number of normals rated x	33	6	6	11	2	$58 = n_N$	Obtained from TABLE I
2	Number of abnormals rated $>x$	48	46	44	33	0		Obtained from 3 by successive subtractions from $n_A = 51$
3	Number of abnormals rated x	3	2	2	11	33	$51 = n_A$	Obtained from TABLE I
4	Number of normals rated $<x$	0	33	39	45	56		Obtained from 1 by successive additions to 0
5	$(1) \times (2) + \frac{1}{2} \times (1) \times (3)$	$1{,}633\frac{1}{2}$	282	270	$423\frac{1}{2}$	33	2,642	W = Total (5) $\div (n_N \cdot n_A) = 0.893$
6	$(3) \times [(4)^2 + (4) \times (1) + \frac{1}{3} \times (1)^2]$	1,089	2,598	3,534	$28{,}163\frac{2}{3}$	107,228	$142{,}612\frac{2}{3}$	Q_2 = Total (6) $\div (n_A \cdot n_N^2) = 0.8313$
7	$(1) \times [(2)^2 + (2) \times (3) + \frac{1}{3} \times (3)^2]$	80,883	13,256	12,152	$16{,}415\frac{2}{3}$	726	$123{,}432\frac{2}{3}$	Q_1 = Total (7) $\div (n_N \cdot n_A^2) = 0.8182$

$W = \hat{\theta}$ = total (5) $\div (n_N \cdot n_A) = 2{,}642 \div (58 \cdot 51) = 0.893 = 89.3\%$

$$SE(\hat{\theta}) = \sqrt{\frac{\hat{\theta}(1-\hat{\theta}) + (n_A - 1)(Q_1 - \hat{\theta}^2) + (n_N - 1)(Q_2 - \hat{\theta}^2)}{n_A \cdot n_N}} = \sqrt{\frac{0.099551 + 0.037764 + 1.926686}{51 \cdot 59}} = 0.032 = 3.2\%$$

VisiCalc

From Wikipedia, the free encyclopedia

VisiCalc (for "visible calculator")[1] was the first spreadsheet computer program for personal computers, originally released for the Apple II by VisiCorp. It is often considered the application that turned the microcomputer from a hobby for computer enthusiasts into a serious business tool, prompting IBM to introduce the IBM PC two years later.[2] VisiCalc is considered the Apple II's killer app. It sold over 700,000 copies in six years, and as many as 1 million copies over its history.

Initially developed in a 6502 assembler running on the Multics time sharing system,[3][4][5] VisiCalc was ported to numerous platforms, both 8-bit and some of the early 16-bit systems. In order to do this, the company developed porting platforms that produced bug compatible versions. The company took the same approach when the IBM PC was launched, producing a product that was essentially identical to the original 8-bit Apple II version. Sales were initially brisk, with about 300,000 copies sold.

VisiCalc used the A1 notation in formulas.[6]

When Lotus 1-2-3 was launched in 1983, taking full advantage of the expanded memory and screen of the PC, VisiCalc sales practically ended overnight. Sales imploded so rapidly that the company was soon insolvent. Lotus Development purchased the company in 1985, and immediately ended sales of VisiCalc and the company's other products.

Contents [hide]

VisiCalc

An example VisiCalc spreadsheet on an Apple II

Developer(s)	Software Arts
Initial release	1979; 38 years ago
Stable release	VisiCalc Advanced Version / 1983; 34 years ago
Operating system	Apple II, Apple SOS, CP/M, Atari 8-bit family, Commodore PET, TRSDOS, Sony SMC-70, DOS, HP series 80
Type	Spreadsheet
License	Commercial proprietary software
Website	danbricklin.com/visicalc.htm

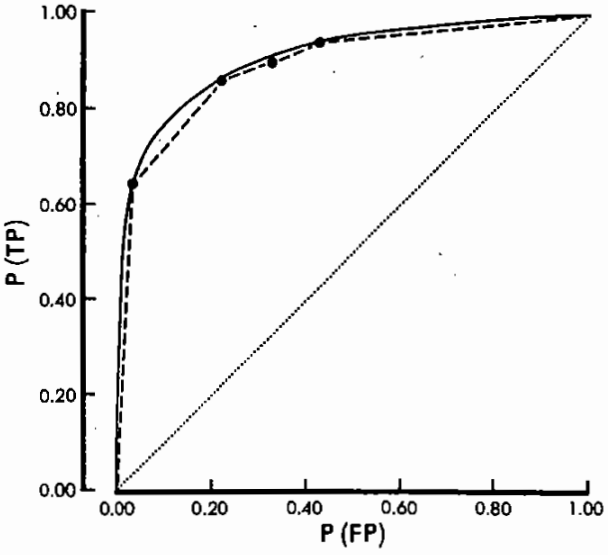
Med Decis Making
Vol. 5, No. 2, 1985

A Visicalc Program for Estimating the Area Under a Receiver Operating Characteristic (ROC) Curve

Robert M. Centor, M.D.

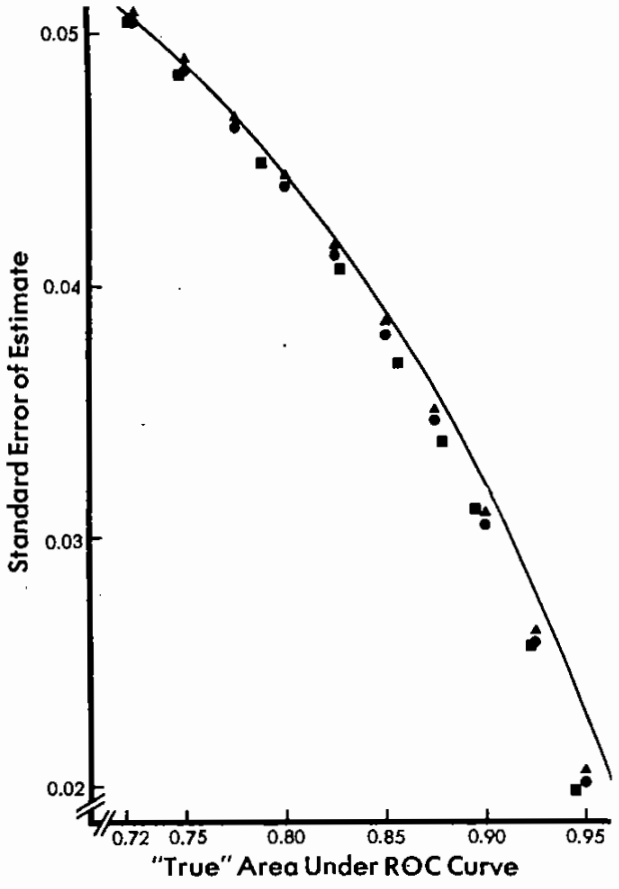
The area under the ROC curve interests us as a method for analyzing discrimination or detectability. One can assess a diagnostic test or probability assessor with respect to its degree of discrimination. The area under the ROC curve gives us the probability of correctly identifying abnormal from normal in a forced-choice, two-alternative problem. Previous methods used for calculating the area involved maximum likelihood estimation on a mainframe or minicomputer. This paper demonstrates an adaptation of a recently published nonparametric method for estimating the area. This adaptation takes advantage of electronic spreadsheet software and may be used on most (if not all) microcomputers. The paper develops the construction of the program needed for the necessary calculations. (Med Decis Making 5:139–148, 1985)

Figure 1



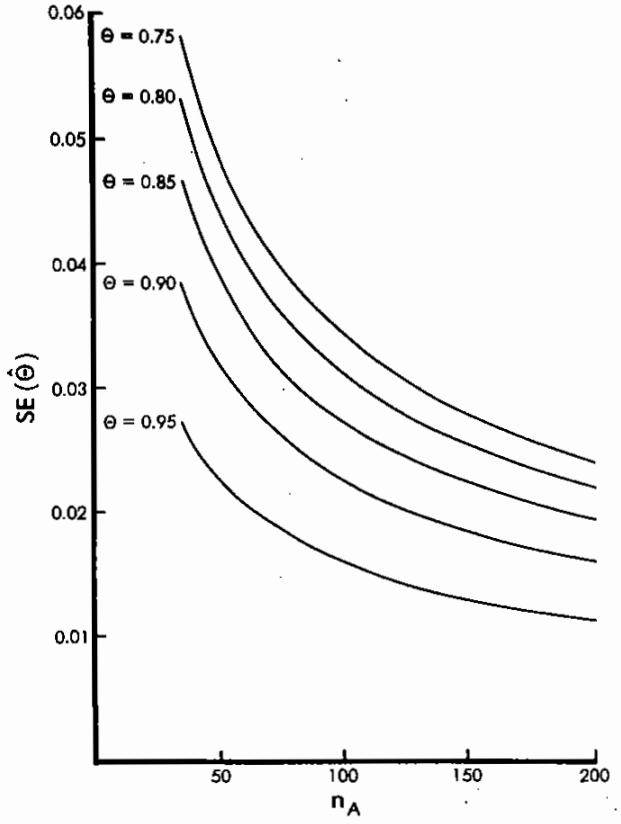
ROC curve for data in TABLE I. Dashed line = empirical curve; solid line = smoothed (Gaussian-based) curve; dotted diagonal line = no discrimination.

Figure 2



Anticipated standard error for area under ROC curve generated from different underlying distributions. Circles = Gaussian with variance ratio 1:1.5; triangles = Gaussian with variance ratio 1:0.5; squares = gamma with various degrees of freedom; solid line = negative exponential.

Figure 3



Standard error (SE) for estimated area under ROC curve ($\hat{\theta}$) in relation to sample size (n_A = number of abnormal cases) and true area under ROC curve (θ). Calculations assume an equal number (n_N) of normal cases.

n_N, Formula 1 contains three other parameters—θ, Q_1, and Q_2. While one can use anticipated values of the true area θ, the quantities Q_1 and Q_2 are complex functions of the underlying distributions for x_A and x_N. Fortunately, for any specified pair of distributions Formula 1 is almost entirely determined by θ, and only very slightly

the fraction $\hat{\theta}$ of m pairs of images

TABLE III: Number of Normal and Abnormal Subjects Required to Provide a Probability of 80%, 90%, or 95% of Detecting Various Differences between the Areas θ_1 and θ_2 under Two ROC Curves (Using a One-Sided Test of Significance with $p = 0.05$)

	θ_2									
θ_1	0.750	0.775	0.800	0.825	0.850	0.875	0.900	0.925	0.950	0.975
0.700	652	286	158	100	68	49	37	28	22	18
	897	392	216	135	92	66	49	38	29	23
	1131	493	271	169	115	82	61	46	36	29
0.725		610	267	148	93	63	45	34	26	20
		839	366	201	126	85	61	45	34	27
		1057	459	252	157	106	75	55	42	33
0.750			565	246	136	85	58	41	31	23
			776	337	185	115	77	55	41	31
			976	423	231	143	96	68	50	38
0.775				516	224	123	77	52	37	27
				707	306	167	104	69	49	36
				889	383	209	129	86	60	44
0.800					463	201	110	68	46	33
					634	273	149	92	61	43
					797	342	185	113	75	53
0.825						408	176	96	59	40
						557	239	129	79	52
						699	298	160	97	64
0.850							350	150	81	50
							477	203	108	66
							597	252	134	81
0.875								290	123	66
								393	165	87
								491	205	107
0.900								960	228	96
								1314	308	127
								1648	383	156
0.925									710	165
									966	220
									1209	272
0.950										457
										615
										765

80% probability = top number; 90% probability = middle number; 95% probability = bottom number.

[Reprinted from RADIOLOGY, Vol. 148, No. 3, Pages 839–843, September, 1983.]

James A. Hanley, Ph.D.
Barbara J. McNeil, M.D., Ph.D.

A Method of Comparing the Areas under Receiver Operating Characteristic Curves Derived from the Same Cases[1]

BIOMETRICS 44, 837–845
September 1988

Comparing the Areas Under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach

Elizabeth R. DeLong

Quintiles, Inc., 1829 East Franklin Street,
Chapel Hill, North Carolina 27514, U.S.A.

David M. DeLong

SAS Institute, Cary, North Carolina 27511, U.S.A.

and

Daniel L. Clarke-Pearson

Division of Oncology, Department of OBGYN, Duke University Medical Center,
Durham, North Carolina 27710, U.S.A.

SUMMARY

Methods of evaluating and comparing the performance of diagnostic tests are of increasing importance as new tests are developed and marketed. When a test is based on an observed variable that lies on a continuous or graded scale, an assessment of the overall value of the test can be made through the use of a receiver operating characteristic (ROC) curve. The curve is constructed by varying the cutpoint used to determine which values of the observed variable will be considered abnormal and then plotting the resulting sensitivities against the corresponding false positive rates. When two or more empirical curves are constructed based on tests performed on the same individuals, statistical analysis on differences between curves must take into account the correlated nature of the data. This paper presents a nonparametric approach to the analysis of areas under correlated ROC curves, by using the theory on generalized *U*-statistics to generate an estimated covariance matrix.

1. Introduction

Methods of evaluating and comparing the performance of diagnostic tests or indices are of increasing importance as new tests or indices are developed or measured. When a test is based on an observed variable that lies on a continuous or graded scale, an assessment of the overall value of the test can be made through the use of a receiver operating characteristic (ROC) curve (Hanley and McNeil, 1982; Metz, 1978). The underlying population curve is theoretically given by varying the cutpoint used to determine the values of the observed variable to be considered abnormal and then plotting the resulting sensitivities against the corresponding false positive rates. If a test could perfectly discriminate, it would have a value above which the entire abnormal population would fall and below which all normal values would fall (or vice versa). The curve would then pass through the point (0, 1) on the unit grid. The closer an ROC curve comes to this ideal point, the better its discriminating ability. A test with no discriminating ability will produce a curve that follows the diagonal of the grid.

For statistical analysis, a recommended index of accuracy associated with an ROC curve is the area under the curve (Swets and Pickett, 1982). The area under the population ROC

Key words: Jackknifing; Mann–Whitney test; Receiver operating characteristic (ROC) curve; Structural components; *U*-statistics.

Statistics in Radiology

Harold L. Kundel, MD, *Editor*

Sampling Variability of Nonparametric Estimates of the Areas under Receiver Operating Characteristic Curves: An Update

James A. Hanley, PhD[1,2], Karim O. Hajian-Tilaki, PhD[1]

Rationale and Objectives. Several methods have been proposed for calculating the variances and covariances of nonparametric estimates of the area under receiver operating characteristic curves (AUC). The authors provide an explanation of the relationships between them and illustrate the factors that determine sampling variability.

Methods. The authors investigated the algebraic links between two methods, that of "placements" and that of "pseudovalues" based on jackknifing. They also performed a numerical investigation of the comparative performance of the two methods.

Results. The "placement" method has a simple structure that illustrates the determinants of the sampling variability and does not require specialized software. The authors show that the pseudovalues used in the jackknife method are directly linked to the placement values.

Conclusion. Because of the close link, borne out in a numeric investigation of the sampling variation, and because of the ease of computation, the choice between the two methods can be based on users' preferences. For indexes other than the AUC, however, the use of pseudovalues holds greater promise.

Key Words. Nonparametric ROC analysis; area under the curve, DeLong method; jackknife pseudovalues

From the [1]Department of Epidemiology and Biostatistics, McGill University, Montreal, Canada; and [2]Division of Clinical Epidemiology, Royal Victoria Hospital, Montreal, Canada.

Supported by funds from the Natural Sciences and Engineering Council of Canada and the Fonds de la recherche en santé du Québec.

Address reprint requests to J. A. Hanley, PhD, Department of Epidemiology and Biostatistics, McGill University, 1020 Pine Ave W, Montreal, Quebec, Canada H3A 1A2.

Received March 25, 1996, and accepted for publication after revision September 18, 1996.

Acad Radiol 1997;4:49–58

The area under the receiver operating characteristic (ROC) curve (AUC) is commonly used as a measure of the accuracy of a diagnostic test. It can be estimated parametrically or nonparametrically [1–6]. Although this statistic has a helpful interpretation, the assessment of its sampling variability—especially in the nonparametric case—is less intuitive. At least four formulas or approaches have been proposed for calculating the variance of a nonparametric AUC estimate, two of which are extendable to the covariance between estimates from two curves.

The first of these four approaches was initially suggested by Bamber [7], who noted the connection between the AUC and the parameter estimated with the Wilcoxon statistic. Hanley and McNeil [6] used this link to give a

TABLE 2: DeLong Method: Calculation of Placements (and of the AUC and Its Variance) from Rating Data for Six Diseased and Nine Nondiseased Subjects

Ratings for $n = 9$ Nondiseased Subjects*	Ratings for $m = 6$ Diseased Subjects*						Placement
	$Y_1 = 1$	$Y_2 = 5$	$Y_3 = 1$	$Y_4 = 2$	$Y_5 = 2$	$Y_6 = 5$	V_X
$X_1 = 2$	0.0	1	0.0	0.5	0.5	1	0.50
$X_2 = 1$	0.5	1	0.5	1	1	1	0.83
$X_3 = 1$	0.5	1	0.5	1	1	1	0.83
$X_4 = 1$	0.5	1	0.5	1	1	1	0.83
$X_5 = 2$	0.0	1	0.0	0.5	0.5	1	0.50
$X_6 = 1$	0.5	1	0.5	1	1	1	0.83
$X_7 = 1$	0.5	1	0.5	1	1	1	0.83
$X_8 = 1$	0.5	1	0.5	1	1	1	0.83
$X_9 = 1$	0.5	1	0.5	1	1	1	0.83
Placement V_Y	0.39	1	0.39	0.89	0.89	1	. . .

Note.—Data indicate the placement of each Y with respect to each X, with 1 indicating the "correct" ordering, 0 an "incorrect" ordering, and 0.5 if Y and X are equal. The data in the right column and bottom row of the Table, obtained as the averages of the corresponding rows/columns, are the placements or pseudoaccuracies corresponding to each X and each Y. Calculations in this and later tables were performed with spreadsheet precision, but numbers were rounded for presentation. Data were obtained with the first of the two field strengths in Table 1. AUC = average of V_X's = average of V_Y's = 0.76. $\text{Var}(V_X)$ = 0.0216; $\text{Var}(V_Y)$ = 0.0848. Var(AUC) = 0.0216/9 + 0.0848/6 = 0.0165. SE(AUC) = $\sqrt{0.0165}$ = 0.13.

*Rating scale ranged from 1 (definitely negative) to 5 (definitely positive).

Up to now, the reader may ask why one would bother to calculate these six individual V_Y's and nine V_X's, since one can simply calculate the AUC directly from the average of the $6 \times 9 = 54$ comparisons of each Y with each X. The answer is that the variations of these six and nine V's can be used directly to estimate the variance of the AUC estimate.

Variance of the AUC Estimate

In the method used by DeLong et al [9], the variance of the AUC estimate is calculated as the sum of two contributions, one relating to the number and variability of the V_X's, the other to the number and variability of the V_Y's, as follows:

$$\text{Var[AUC]} = \frac{\text{Variance of } V_x\text{'s}}{n\text{: number of } V_x\text{'s}} + \frac{\text{Variance of } V_y\text{'s}}{m\text{: number of } V_x\text{'s}}. \quad (1)$$

Those interested in the equivalence of this equation and the formula given in Hanley and McNeil's first article [6] can consult the textbook by Hettmansperger [14]. DeLong et al [9] omitted the third variance component, $\text{AUC}(1 - \text{AUC})/mn$, since it is negligible when n and m are large.

$$\text{Var[AUC]} = \frac{0.0216}{9} + \frac{0.0848}{6} = 0.0165,$$

so that the standard error (SE) is

$$\text{SE[AUC]} = \sqrt{0.0165} = 0.13$$

The structure of Equation (1) reveals one additional insight into the sampling variability (and its control) that does not appear to have been commented on previously. This insight comes from the nature of the component variances (0.0216 and 0.0848 in our example). These are estimates of the variance of the true-positive fraction (TPF) and false-positive fraction (FPF) points on the smooth ROC curve underlying the data. One can imagine the smooth ROC curve as a very large number (say 1,000 or 10,000) of TPF points corresponding to 100 or 10,000 equally spaced FPF points. If the ROC curve were the 45° diagonal line, these TPF points would be uniform on the (0,1) scale, and their variance would be 1/12 or 0.0833. The closer the curve is to the top left corner, the more concentrated and closer to the 1 than the 0 end of the (0,1) scale the TPF points will be and the smaller will be their variance. The V_Y's

1994

Transfer of Technology From Statistical Journals to the Biomedical Literature

Past Trends and Future Predictions

Douglas G. Altman, Steven N. Goodman, MD, PhD

Objective.—To investigate the speed of the transfer of new statistical methods into the medical literature and, on the basis of current data, to predict what methods medical journal editors should expect to see in the next decade.

Design.—Influential statistical articles were identified and the time pattern of citations in the medical literature was ascertained. In addition, longitudinal studies of the statistical content of articles in medical journals were reviewed.

Main Outcome Measures.—Cumulative number of citations in medical journals of each article in the years after publication.

Results.—Annual citations show some evidence of decreasing lag times between the introduction of new statistical methods and their appearance in medical journals. Newer technical innovations still typically take 4 to 6 years before they achieve 25 citations in the medical literature. Few methodological advances of the 1980s seem yet to have been widely cited in medical journals. Longitudinal studies indicate a large increase in the use of more complex statistical methods.

Conclusions.—Time trends suggest that technology diffusion has speeded up during the last 30 years, although there is still a lag of several years before medical citations begin to accrue. Journals should expect to see more articles using increasingly sophisticated methods. Medical journals may need to modify reviewing procedures to deal with articles using these complex new methods.

(*JAMA*. 1994;272:129-132)

using computer searches of the SciSearch database (Institute of Scientific Information, Philadelphia, Pa). These searches were carried out in July and August 1993, by which time citations for 1992 should have been virtually complete. We did not search for articles that had incorrect citations of the articles of interest. It is our impression that the rate of incorrect citations of these articles was about 10% (excluding errors in titles). Some minor inconsistency between the two methods of searching may have arisen through problems in identifying what constitutes a medical journal. For comparison, similar citation analyses were performed for two heavily cited expository statistical articles published in medical journals.[21,22]

We also sought evidence from longitudinal studies of the statistical content of articles in medical journals to examine changes in the methods used over

Table 1.—Statistical Articles Included in This Study

Source, y	Topic
Methodological articles	
Cornfield,[3] 1951	Odds ratio
Cochran,[4] 1954	χ^2 Trend test
Woolf,[5] 1955	Combining 2×2 tables
Kaplan and Meier,[6] 1958	Survival curve
Mantel and Haenszel,[7] 1958	Stratified 2×2 table
Cohen,[8] 1960	κ Statistic
Mantel,[9] 1963	Survival analysis
Box and Cox,[10] 1964	Transformations
Mantel,[11] 1966	Survival analysis
Elston and Stewart,[12] 1971	Heredity
Peto and Peto,[13] 1972	Log rank test
Cox,[14] 1972	Proportional hazards regression
Dempster et al,[15] 1977	EM algorithm
Efron,[16] 1979	Bootstrap
Hanley and McNeil,[17] 1982	Receiver operating characteristic curve
Geman and Geman,[18] 1984	Gibbs sampling
Breiman et al,[19] 1984	Classification and regression trees
Zeger and Liang,[20] 1986	Longitudinal data
Expository articles	
Peto et al,[21] 1977	Log rank test
Bland and Altman,[22] 1986	Method comparison

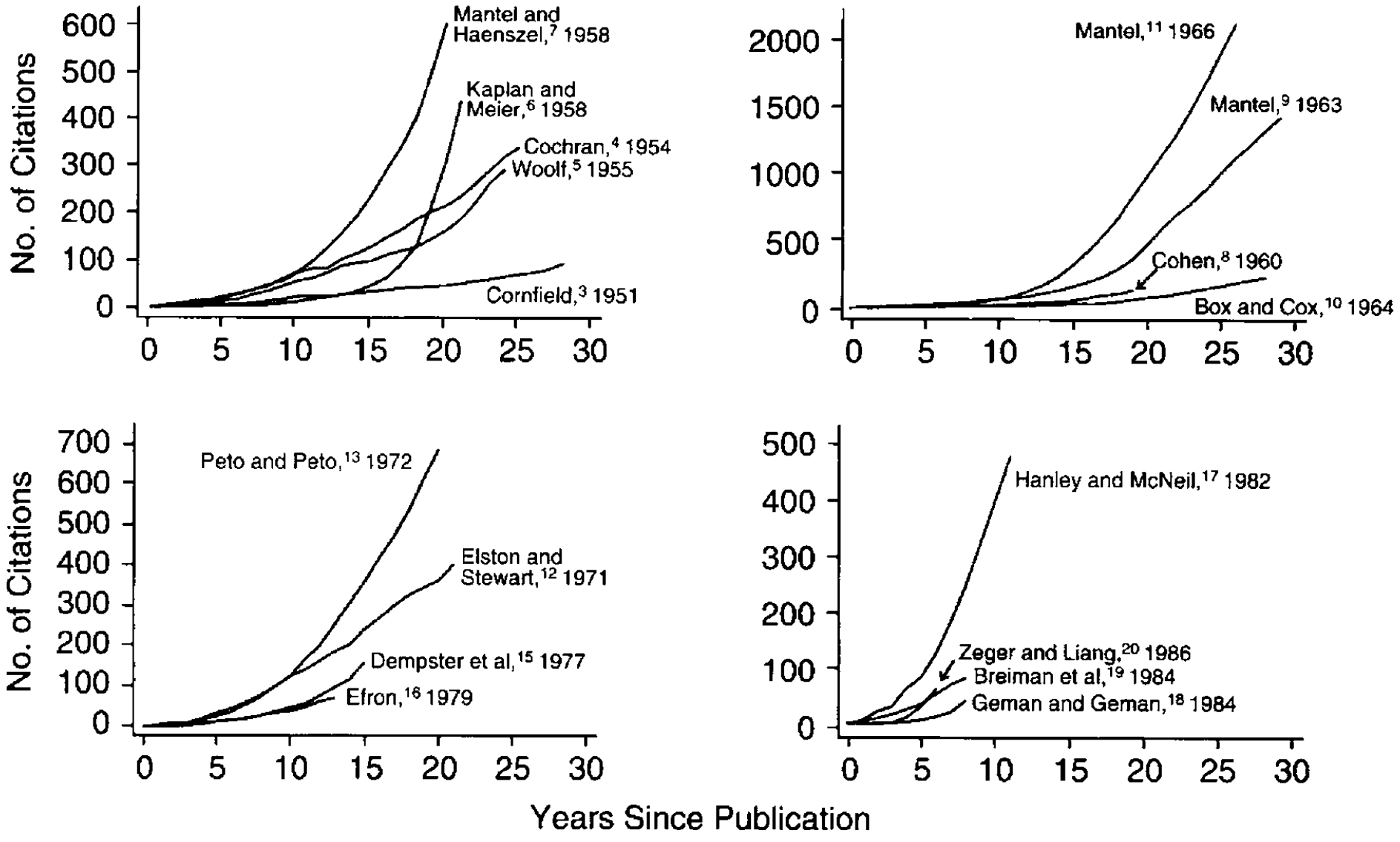


Fig 1.—Cumulative citations in medical journals for selected articles published in 1950 through 1959 (top left), 1960 through 1969 (top right), 1970 through 1979 (bottom left), and 1980 through 1989 (bottom right).

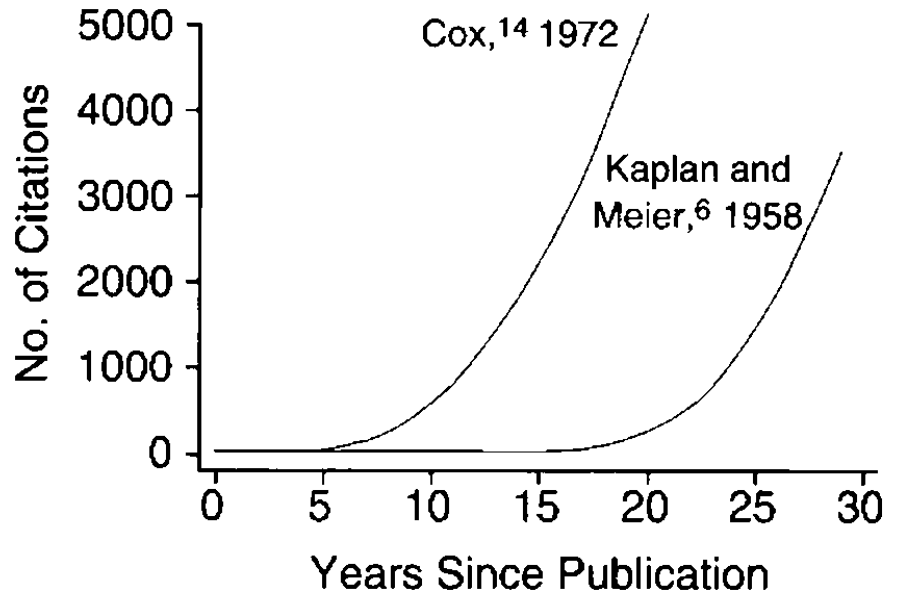


Fig 2.—Cumulative citations in medical journals for two heavily cited articles on survival analysis methods.

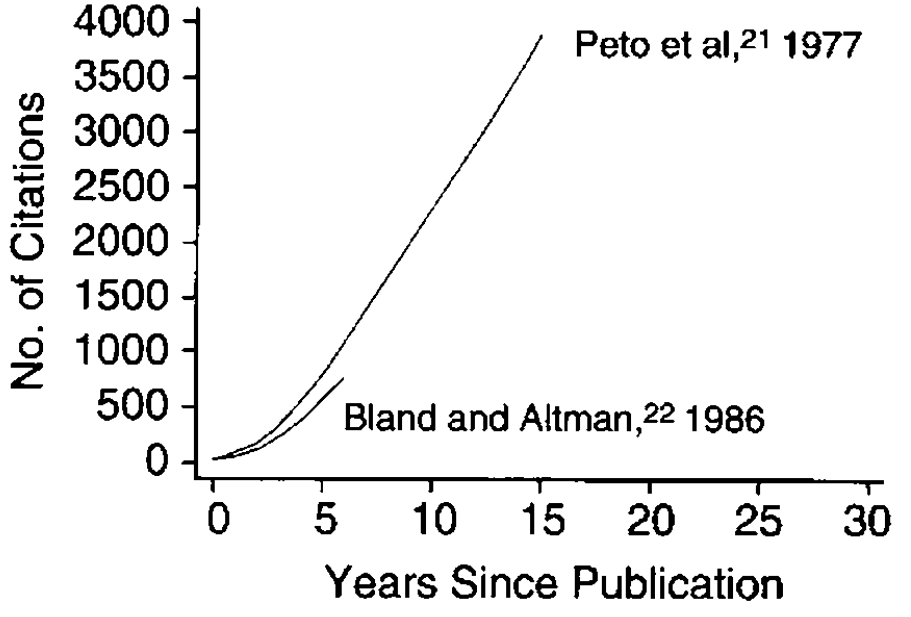
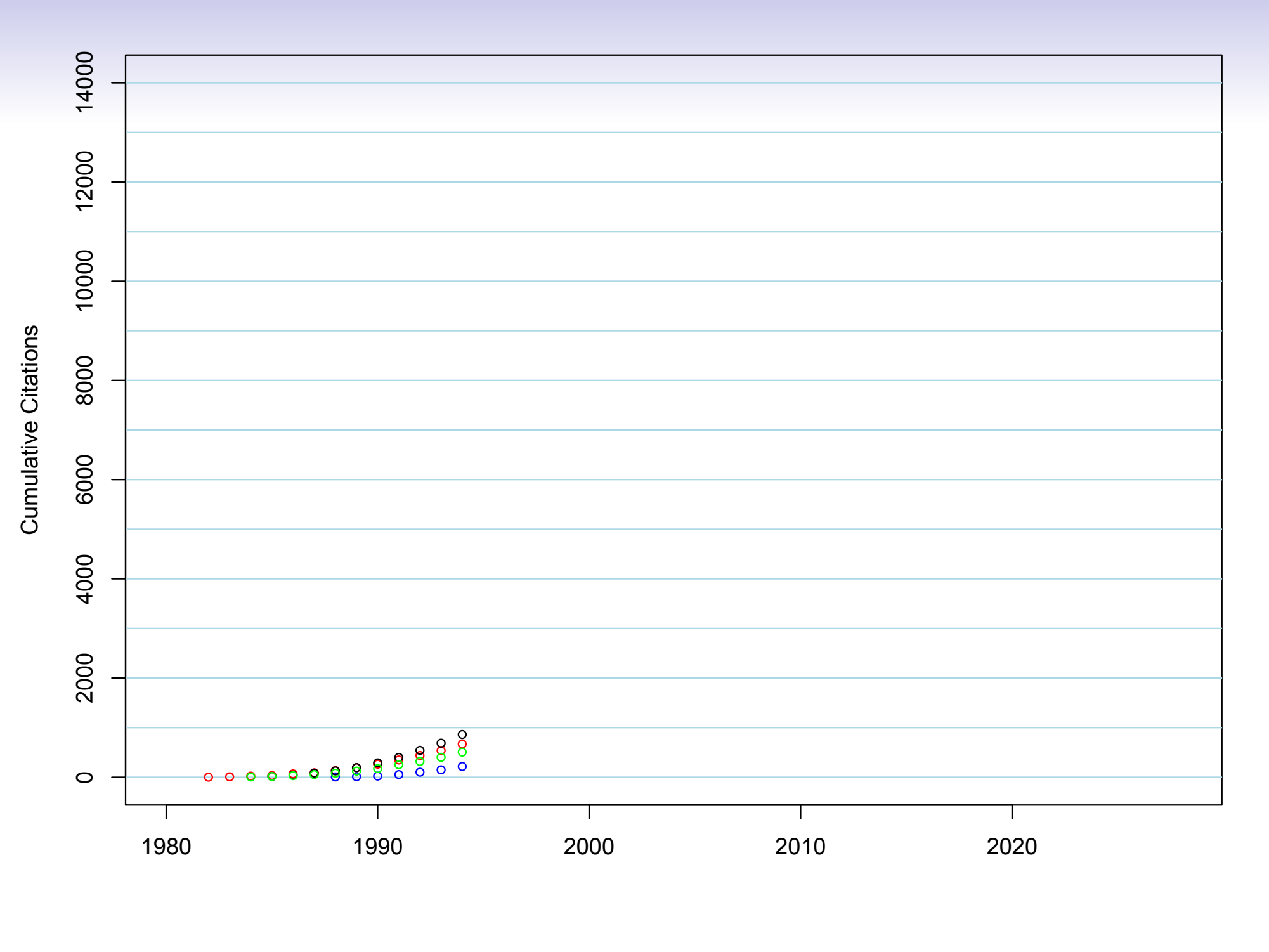
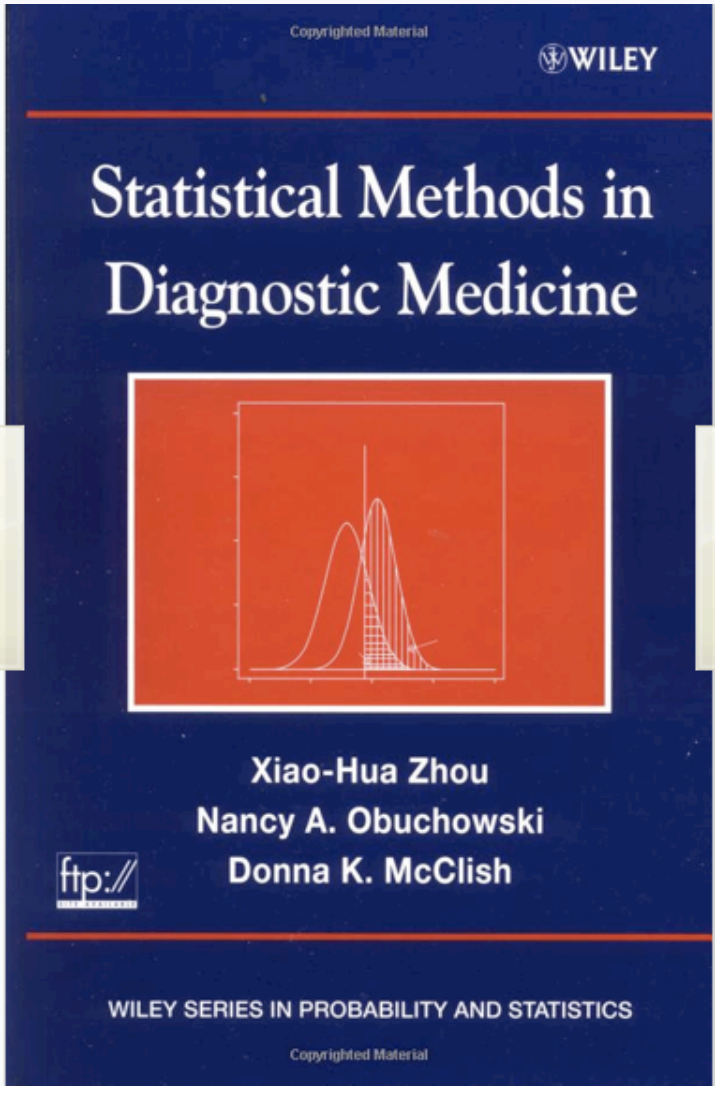
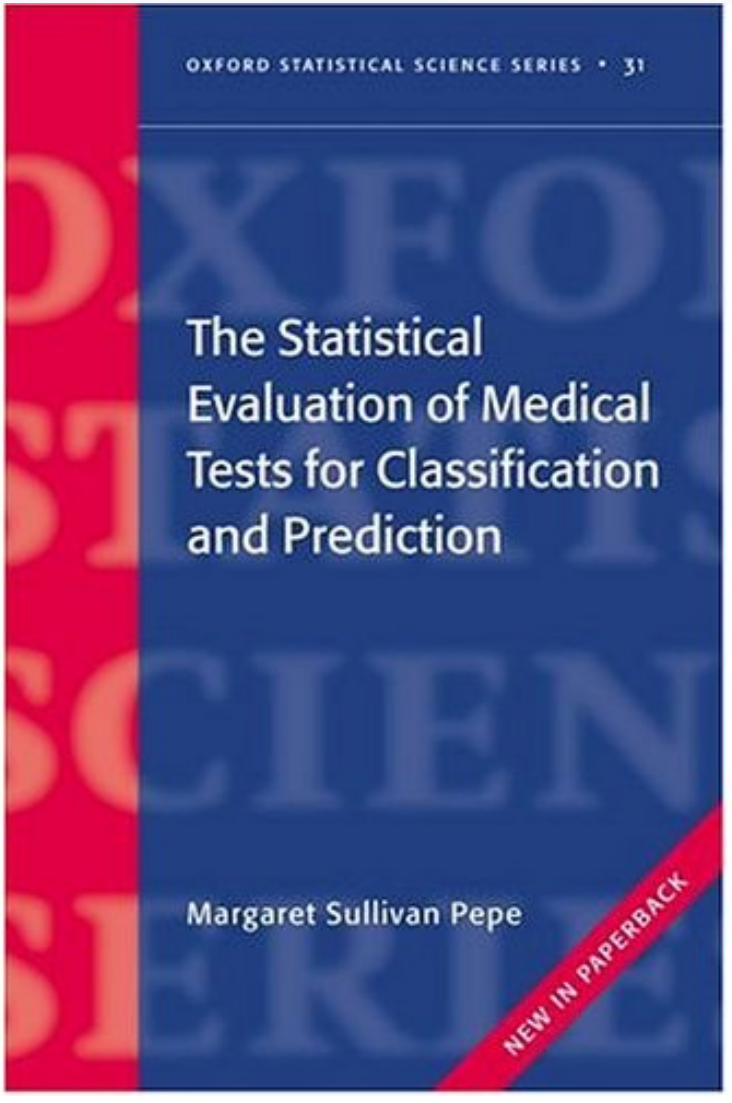


Fig 3.—Cumulative citations in medical journals for two expository articles.

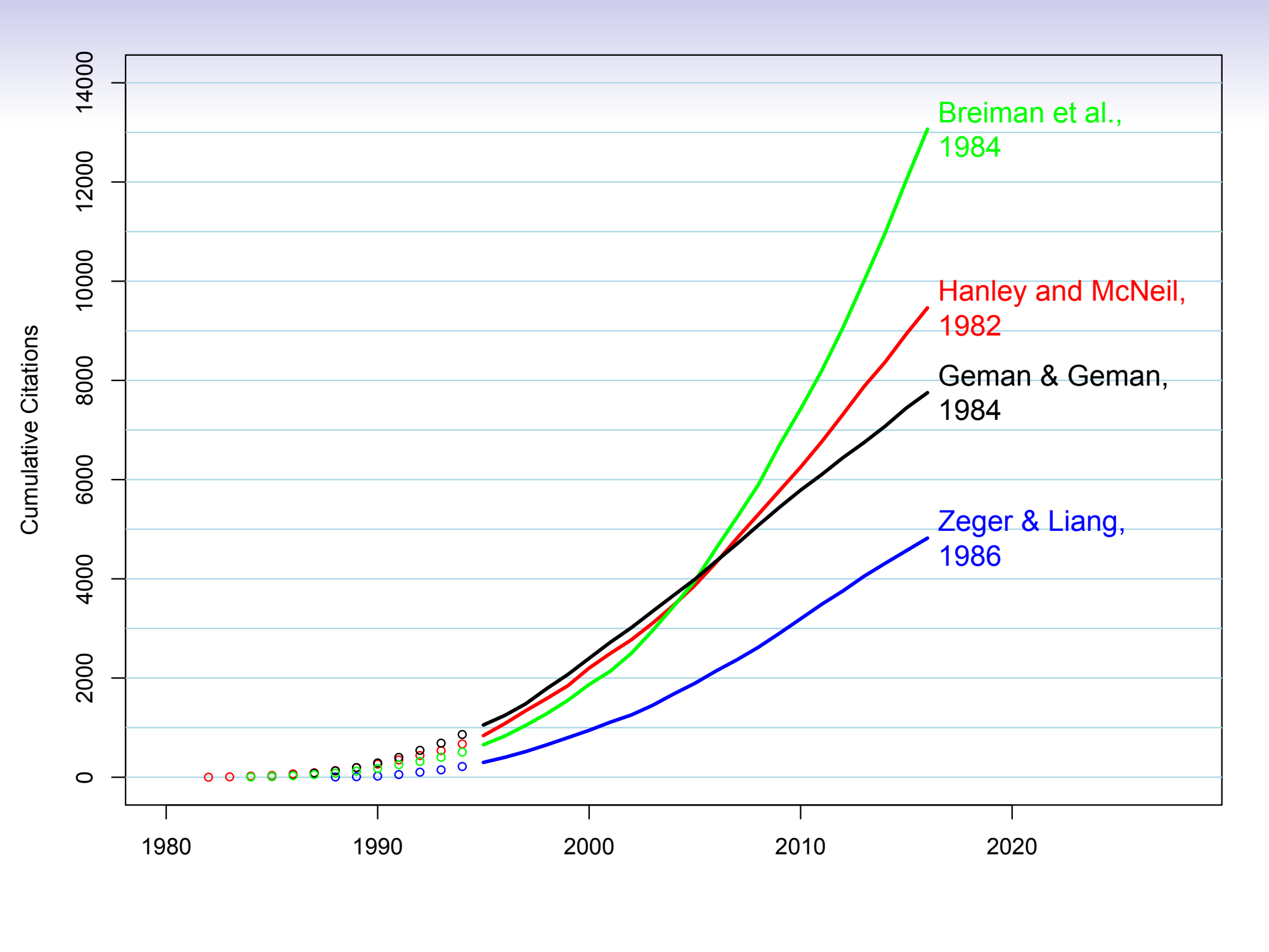

Cumulative Citations
0
2000
4000
6000
8000
10000
12000
14000
1980
1990
2000
2010
2020

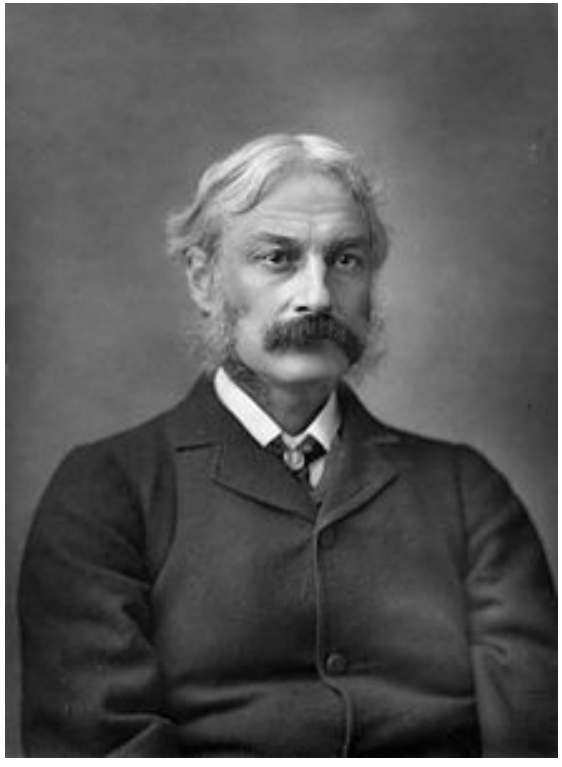

Copyrighted Material
WILEY
Statistical Methods in Diagnostic Medicine
Xiao-Hua Zhou
Nancy A. Obuchowski
Donna K. McClish
ftp://
WILEY SERIES IN PROBABILITY AND STATISTICS
Copyrighted Material

2002


OXFORD STATISTICAL SCIENCE SERIES • 31
The Statistical Evaluation of Medical Tests for Classification and Prediction
Margaret Sullivan Pepe
NEW IN PAPERBACK

2003


Cumulative Citations
0
2000
4000
6000
8000
10000
12000
14000
1980
1990
2000
2010
2020
Breiman et al., 1984
Hanley and McNeil, 1982
Geman & Geman, 1984
Zeger & Liang, 1986



Politicians use statistics in the same way that a drunk uses lamp-posts:

for support rather than illumination.

Andrew Lang (1844 – 1912), Scottish poet, novelist, and literary critic, and contributor to anthropology.

Extra-mural consultations

1. 2007. CANWEST vs Government of Canada:
   – Direct to Consumer Advertising of Medications

2. 2013. World Anti-Doping Agency (WADA):
   – Detection limits for Human Growth Hormone tests

3. 1994-5. Québec Health Ministry:
   – Should it pay for PSA screening for Prostate Cancer?

How does direct-to-consumer advertising (DTCA) affect prescribing? A survey in primary care environments with and without legal DTCA

Barbara Mintzes, Morris L. Barer, Richard L. Kravitz, Ken Bassett, Joel Lexchin, Arminée Kazanjian, Robert G. Evans, Richard Pan, Stephen A. Marion

ß See related articles pages 421 and 425

Abstract

Background: Direct-to-consumer advertising (DTCA) of prescription drugs has increased rapidly in the United States during the last decade, yet little is known about its effects on prescribing decisions in primary care. We compared prescribing decisions in a US setting with legal DTCA and a Canadian setting where DTCA of prescription drugs is illegal, but some cross-border exposure occurs.

Methods: We recruited primary care physicians working in Sacramento, California, and Vancouver, British Columbia, and their group practice partners to participate in the study. On preselected days, patients aged 18 years or more completed a questionnaire before seeing their physician. We asked these patients' physicians to complete a brief questionnaire immediately following the selected patient visit. By pairing individual patient and physician responses, we determined how many patients had been exposed to some form of DTCA, the frequency of patients' requests for prescriptions for advertised medicines and the frequency of prescriptions that were stimulated by the patients' requests. We measured physicians' confidence in treatment choice for each new prescription by asking them whether they would prescribe this drug to a patient with the same condition.

Results: Seventy-eight physicians (Sacramento $n = 38$, Vancouver $n = 40$) and 1431 adult patients (Sacramento $n = 683$, Vancouver $n = 748$), or 61% of patients who consulted participating physicians on pre-set days, participated in the survey. Exposure to DTCA was higher in Sacramento, although 87.4% of Vancouver patients had seen prescription drug advertisements. Of the Sacramento patients, 7.2% requested advertised drugs as opposed to 3.3% in Vancouver (odds ratio [OR] 2.2, 95% confidence interval [CI] 1.2–4.1). Patients with higher self-reported exposure to advertising, conditions that were potential to be only "possible" or "unlikely" choices for other similar patients, as compared with 12.4% of new prescriptions not requested by patients ($p < 0.001$).

Interpretation: Our results suggest that more advertising leads to more requests for advertised medicines, and more prescriptions. If DTCA opens a conversation between patients and physicians, that conversation is highly likely to end with a prescription, often despite physician ambivalence about treatment choice.

CMAJ 2003;169(5):405-12

From 1996 to 2000, spending on direct-to-consumer advertising (DTCA) of prescription drugs in the United States more than tripled,[1] reaching US$2.7 billion in 2001.[2] The United States and New Zealand are the only industrialized countries that allow such advertising, although restrictive legislation in the European Union[3] and Canada[4] has recently been under review. Canada allows advertising of over-the-counter (OTC) drugs but prohibits DTCA of prescription medicines, although a 1978 exemption, which was intended to allow price comparisons, permits advertising of product name, price and quantity.[4] Nevertheless, Canadians see advertisements in US magazines and on US cable television, as well as an increasing volume of domestically generated DTCA of questionable legality.[5] Proponents of DTCA argue that advertisements empower patients, whereas critics counter that they encourage wasteful prescribing.[6] Empirical research is needed to assess the effects of DTCA on prescribing decisions, the patient–physician relationship and, ultimately, health outcomes.

We surveyed primary care patients and their physicians in Sacramento, California, and Vancouver, British Columbia.

Court File No. 05-CV-303001PD2

ONTARIO
SUPERIOR COURT OF JUSTICE

BETWEEN:

CANWEST MEDIAWORKS INC.

Applicant

and

ATTORNEY GENERAL OF CANADA

Respondent

AFFIDAVIT OF

I, in the Province of Ontario, make oath and say as follows:

Personal qualificiations

Form 4D – AFFIDAVIT
Rules of Civil Procedure, (Rule 4.06)

Court File No.: 05-CV-303001PD2

ONTARIO
SUPERIOR COURT OF JUSTICE

BETWEEN:

CANWEST MEDIAWORKS INC.

Applicant

and

ATTORNEY GENERAL OF CANADA

Respondent

AFFIDAVIT OF JAMES HANLEY

I, **JAMES HANLEY**, of the City of Montreal, in the Province of Quebec, solemnly AFFIRM:

‘10 events per variable’ rule for logistic regression

- ‘One common criterion for the validity of such statistical models: **a minimum of at least 10 outcome events per model parameter**.

- The model has a sample size of 74 events; collectively at least **12 main parameters** being estimated from these **74 events** [in the 1400 patients studied].

- The sample size is therefore clearly too small to support an analysis of this complexity with any reliability. ’

‘10 events per variable’ rule for [logistic] regression

- ‘Work by others has shown that conclusions from such models fitted with insufficient sample size can be substantially in error with respect to the magnitude, precision, statistical significance, and even the direction of the associations indicated in the results.

- These concerns are particularly pertinent when the factors included in the model may themselves be related to one another. ’

J. CLINICAL EPIDEMIOLOGY 1996

A Simulation Study of the Number of Events per Variable in Logistic Regression Analysis

Peter Peduzzi,[1,4,*] *John Concato*,[2,3] *Elizabeth Kemper*,[1,4] *Theodore R. Holford*,[4] *and Alvan R. Feinstein*[2,3,4]

[1]COOPERATIVE STUDIES PROGRAM COORDINATING CENTER AND THE [2]MEDICAL SERVICE, VETERANS AFFAIRS MEDICAL CENTER, WEST HAVEN CONNECTICUT 06516; AND THE DEPARTMENTS OF [3]MEDICINE (CLINICAL EPIDEMIOLOGY UNIT) AND [4]EPIDEMIOLOGY AND PUBLIC HEALTH, YALE UNIVERSITY SCHOOL OF MEDICINE, NEW HAVEN, CONNECTICUT 06510

ABSTRACT. We performed a Monte Carlo study to evaluate the effect of the number of events per variable (EPV) analyzed in logistic regression analysis. The simulations were based on data from a cardiac trial of 673 patients in which 252 deaths occurred and seven variables were cogent predictors of mortality; the number of events per predictive variable was (252/7=) 36 for the full sample. For the simulations, at values of EPV = 2, 5, 10, 15, 20, and 25, we randomly generated 500 samples of the 673 patients, chosen with replacement, according to a logistic model derived from the full sample. Simulation results for the regression coefficients for each variable in each group of 500 samples were compared for bias, precision, and significance testing against the results of the model fitted to the original sample.

For EPV values of 10 or greater, no major problems occurred. For EPV values less than 10, however, the regression coefficients were biased in both positive and negative directions; the large sample variance estimates from the logistic model both overestimated and underestimated the sample variance of the regression coefficients; the 90% confidence limits about the estimated values did not have proper coverage; the Wald statistic was conservative under the null hypothesis; and paradoxical associations (significance in the wrong direction) were increased. Although other factors (such as the total number of events, or sample size) may influence the validity of the logistic model, our findings indicate that low EPV can lead to major problems. J CLIN EPIDEMIOL 49;12:1373–1379, 1996.

KEY WORDS. Monte Carlo, bias, precision, significance testing

American Journal of Epidemiology
Vol. 165, No. 6
DOI: 10.1093/aje/kwk052
Advance Access publication December 20, 2006

Original Contribution

Relaxing the Rule of Ten Events per Variable in Logistic and Cox Regression

Eric Vittinghoff and Charles E. McCulloch

From the Department of Epidemiology and Biostatistics, University of California, San Francisco, CA.

Received for publication March 15, 2006; accepted for publication August 15, 2006.

The rule of thumb that logistic and Cox models should be used with a minimum of 10 outcome events per predictor variable (EPV), based on two simulation studies, may be too conservative. The authors conducted a large simulation study of other influences on confidence interval coverage, type I error, relative bias, and other model performance measures. They found a range of circumstances in which coverage and bias were within acceptable levels despite less than 10 EPV, as well as other factors that were as influential as or more influential than EPV. They conclude that this rule can be relaxed, in particular for sensitivity analyses undertaken to demonstrate adequate control of confounding.

bias (epidemiology); coverage probability; event history analysis; model adequacy; type I error; variable selection

Abbreviation: EPV; events per predictor variable.

FAILURE TO DISTINGUISH

SAME EQUATION

$$E[Y|X_1,\ X_2,\ \ldots,\ X_p] = \beta_0 + \beta_1 X_1 + \beta_2 X_2,\ \cdots + \beta_p\ X_p$$

DIFFERENT OBJECTIVES (TARGETS)

- β_1
- $\sum \beta_j X_j$, for various $\{X_1,\ X_2,\ \ldots, X_p\}$ 'profiles'
- $\{\beta_1,\ \beta_2,\ \ldots, \beta_p\}$

ELSEVIER

CrossMark

Journal of Clinical Epidemiology 68 (2015) 627–636

Journal of Clinical Epidemiology

The number of subjects per variable required in linear regression analyses

Peter C. Austin[a,b,c,*], Ewout W. Steyerberg[d]

[a]*Institute for Clinical Evaluative Sciences, G1 06, 2075 Bayview Avenue, Toronto, Ontario, Canada M4N 3M5*

[b]*Institute of Health Policy, Management and Evaluation, University of Toronto, 155 College Street, Suite 425 Toronto, ON M5T 3M6, Canada*

[c]*Schulich Heart Research Program, Sunnybrook Research Institute, 2075 Bayview Avenue, Toronto, ON M4N 3M5, Canada*

[d]*Department of Public Health, Erasmus MC—University Medical Center Rotterdam, 's-Gravendijkwal 230 3015 CE, Rotterdam, The Netherlands*

Accepted 24 December 2014; Published online 22 January 2015

Abstract

Objectives: To determine the number of independent variables that can be included in a linear regression model.

Study Design and Setting: We used a series of Monte Carlo simulations to examine the impact of the number of subjects per variable (SPV) on the accuracy of estimated regression coefficients and standard errors, on the empirical coverage of estimated confidence intervals, and on the accuracy of the estimated R^2 of the fitted model.

Results: A minimum of approximately two SPV tended to result in estimation of regression coefficients with relative bias of less than 10%. Furthermore, with this minimum number of SPV, the standard errors of the regression coefficients were accurately estimated and estimated confidence intervals had approximately the advertised coverage rates. A much higher number of SPV were necessary to minimize bias in estimating the model R^2, although adjusted R^2 estimates behaved well. The bias in estimating the model R^2 statistic was inversely proportional to the magnitude of the proportion of variation explained by the population regression model.

Conclusion: Linear regression models require only two SPV for adequate estimation of regression coefficients, standard errors, and confidence intervals.

Keywords: Regression; Linear regression; Bias; Monte Carlo simulations; Explained variation; Statistical methods

ELSEVIER

CrossMark

Journal of Clinical Epidemiology 79 (2016) 112–119

Journal of Clinical Epidemiology

Simple and multiple linear regression: sample size considerations

James A. Hanley*

Department of Epidemiology, Biostatistics and Occupational Health, McGill University, 1020 Pine Avenue West, Montreal, Quebec H3A 1A2, Canada

Accepted 6 May 2016; Published online 5 July 2016

Abstract

Objective: The suggested ''two subjects per variable'' (2SPV) rule of thumb in the Austin and Steyerberg article is a chance to bring out some long-established and quite intuitive sample size considerations for both simple and multiple linear regression.

Study Design and Setting: This article distinguishes two of the major uses of regression models that imply very different sample size considerations, neither served well by the 2SPV rule. The first is etiological research, which contrasts mean Y levels at differing ''exposure'' (X) values and thus tends to focus on a single regression coefficient, possibly adjusted for confounders. The second research genre guides clinical practice. It addresses Y levels for individuals with different covariate patterns or ''profiles.'' It focuses on the profile-specific (mean) Y levels themselves, estimating them via linear compounds of regression coefficients and covariates.

Results and Conclusion: By drawing on long-established closed-form variance formulae that lie beneath the standard errors in multiple regression, and by rearranging them for heuristic purposes, one arrives at quite intuitive sample size considerations for both research genres.

Keywords: Precision; Power; Prediction; Confounding; Degrees of freedom

World Anti-Doping Agency (WADA):

Detection limits for Human Growth Hormone tests

Clinical Chemistry 50:10
1762–1768 (2004)

REFERENCE VALUES

C-Reactive Protein and Features of the Metabolic Syndrome in a Population-Based Sample of Children and Adolescents

Marie Lambert,[1*] Edgard E. Delvin,[2] Gilles Paradis,[4] Jennifer O'Loughlin,[4] James A. Hanley,[4] and Emile Levy[3]

Background: **C-Reactive protein (CRP) is a risk marker for type 2 diabetes and cardiovascular diseases. In youth, limited data are available on the distribution of high-sensitivity CRP as well as on its association with components of the metabolic syndrome.**
Methods: **In 1999, we conducted a school-based survey of a representative sample of youths 9, 13, and 16 years of age in the province of Quebec, Canada. Standardized clinical measurements and fasting plasma lipid, glucose, insulin, and CRP concentrations were available for 2224 individuals.**
pressure was no longer statistically significant after adjustment for BMI.
Conclusions: **The metabolic correlates of excess weight, including a state of low-grade systemic inflammation, are detectable early in life. Their health impact in adults remains to be fully examined.**

Measurement of the concentration of C-reactive protein (CRP),[5] an acute-phase reactant, has been used for decades in the diagnosis and monitoring of active infections

Table 2. Percentile values for plasma CRP concentration by age and sex.

Sex	Age, years	Exclusion[a]	n	CRP concentration by percentiles (95% CI), mg/L		
				50th	75th	95th
Boys	9	No	340	<0.2 (<0.2 to 0.20)	0.47 (0.37–0.76)	3.13 (2.25–4.32)
		Yes	221	<0.2 (<0.2 to 0.20)	0.47 (0.36–0.68)	2.73 (2.09–4.22)
	13	No	365	0.21 (<0.2 to 0.25)	0.71 (0.56–1.0)	4.24 (2.80–5.71)
		Yes	192	<0.2 (<0.2 to 0.23)	0.66 (0.54–1.0)	4.44 (2.96–5.74)
	16	No	372	0.30 (0.25–0.37)	1.09 (0.82–1.33)	5.06 (3.77–10.7)
		Yes	125	0.31 (0.27–0.38)	0.88 (0.71–1.08)	3.28 (2.29–5.04)
Girls	9	No	366	0.31 (0.23–0.37)	1.06 (0.73–1.66)	5.65 (4.04–10.1)
		Yes	236	0.28 (0.22–0.32)	0.88 (0.63–1.58)	5.02 (3.81–6.36)
	13	No	349	<0.2 (<0.2 to 0.21)	0.54 (0.40–0.71)	2.72 (2.00–4.02)
		Yes	142	<0.2 (<0.2 to 0.22)	0.59 (0.45–0.75)	2.43 (2.03–3.94)
	16	No	432	0.56 (0.42–0.73)	1.90 (1.44–2.17)	6.28 (5.11–7.85)
		Yes	85	0.38 (0.34–0.42)	1.63 (0.88–2.16)	5.29 (4.32–6.33)

[a] Excludes current smokers and individuals who took antibiotics or medications for pain/fever, cold/allergies, or respiratory problems in the 2 weeks before blood sampling.

Growth Hormone & IGF Research 24 (2014) 205–215

Contents lists available at ScienceDirect

Growth Hormone & IGF Research

journal homepage: www.elsevier.com/locate/ghir

hGH isoform differential immunoassays applied to blood samples from athletes: Decision limits for anti-doping testing

James A. Hanley [a,b,*], Olli Saarela [a], David A. Stephens [b], Jean-Christophe Thalabard [c,d]

[a] Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montreal, Canada
[b] Department of Mathematics and Statistics, McGill University, Montreal, Canada
[c] Paris Descartes University, MAP5, UMR CNRS 8145, Paris, France
[d] Endocrine Gynaecology Unit, Hôpital Cochin, Paris, France

ARTICLE INFO

Article history:
Received 17 May 2014
Accepted 2 June 2014
Available online 11 June 2014

Keywords:
Quantile
Regression
Decision limits
Isoforms
Human Growth Hormone
Doping

ABSTRACT

Objective: To detect hGH doping in sport, the World Anti-Doping Agency (WADA)-accredited laboratories use the ratio of the concentrations of recombinant hGH ('rec') versus other 'natural' pituitary-derived isoforms of hGH ('pit'), measured with two different kits developed specifically to detect the administration of exogenous hGH. The current joint compliance decision limits (DLs) for ratios derived from these kits, designed so that they would both be exceeded in fewer than 1 in 10,000 samples from non-doping athletes, are based on data accrued in anti-doping labs up to March 2010, and later confirmed with data up to February–March 2011. In April 2013, WADA asked the authors to analyze the now much larger set of ratios collected in routine hGH testing of athletes, and to document in the peer-reviewed literature a statistical procedure for establishing DLs, so that it be re-applied as more data become available.

Design: We examined the variation in the rec/pit ratios obtained for 21,943 screened blood (serum) samples submitted to the WADA accredited laboratories over the period 2009–2013. To fit the relevant sex- and kit-specific

'LMS' ($\lambda,\ \mu,\ \sigma$) method

(developed for growth charts by Tim Cole)

JRSS A, 1988

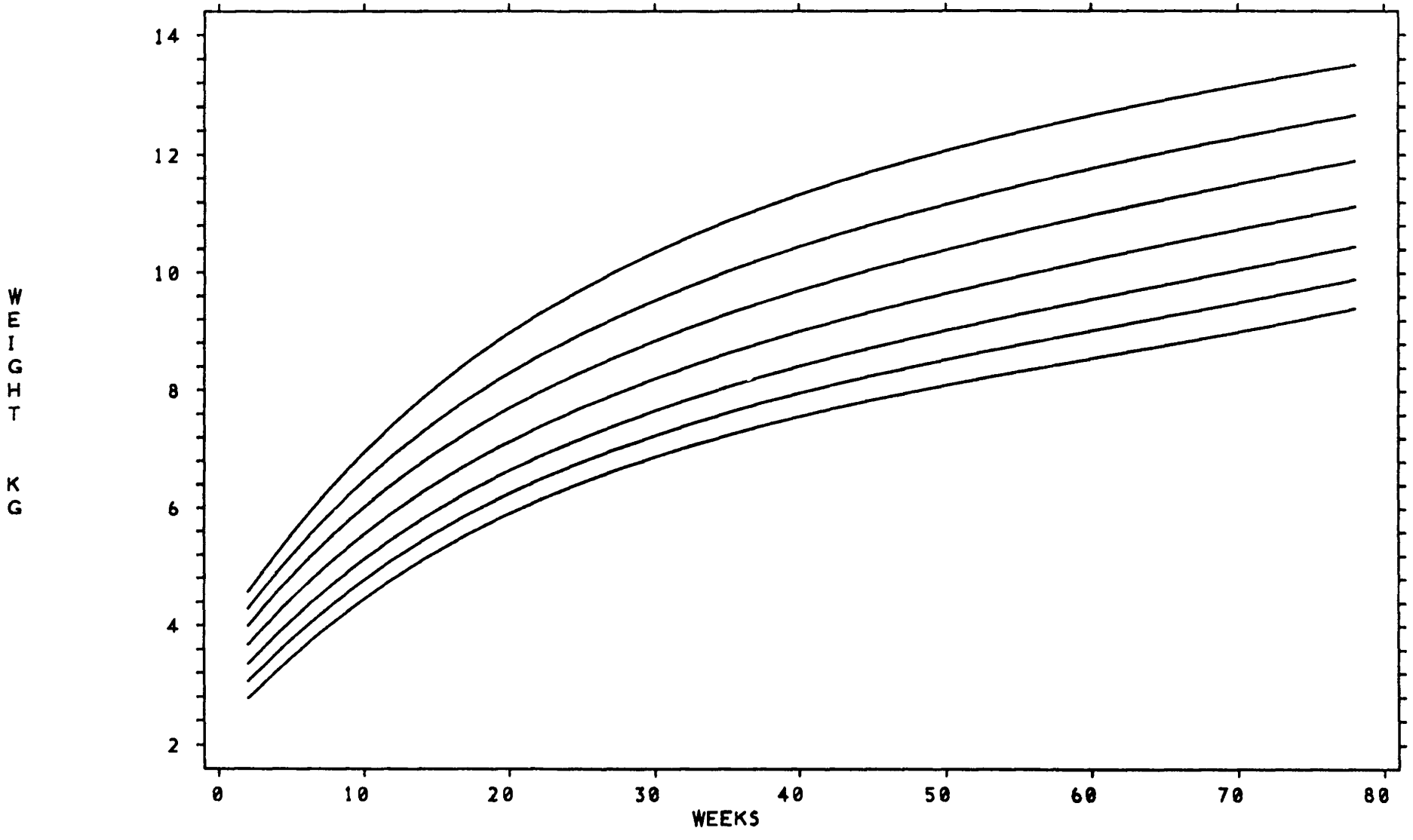
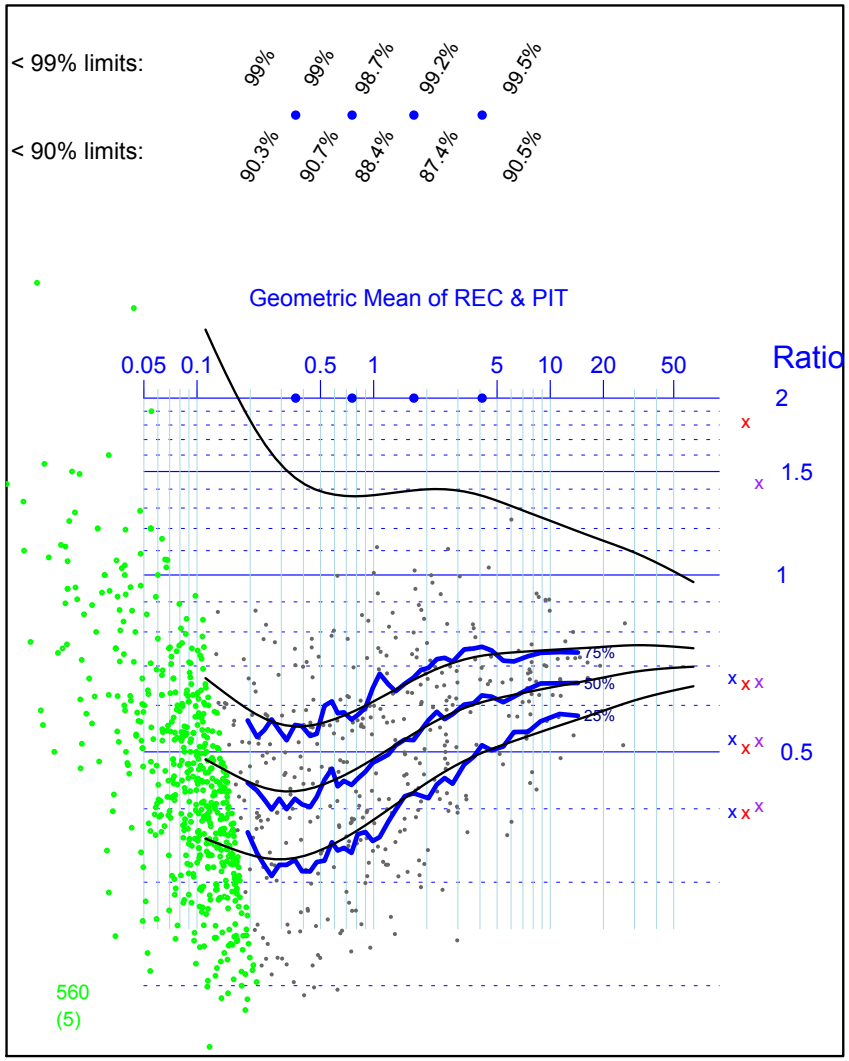


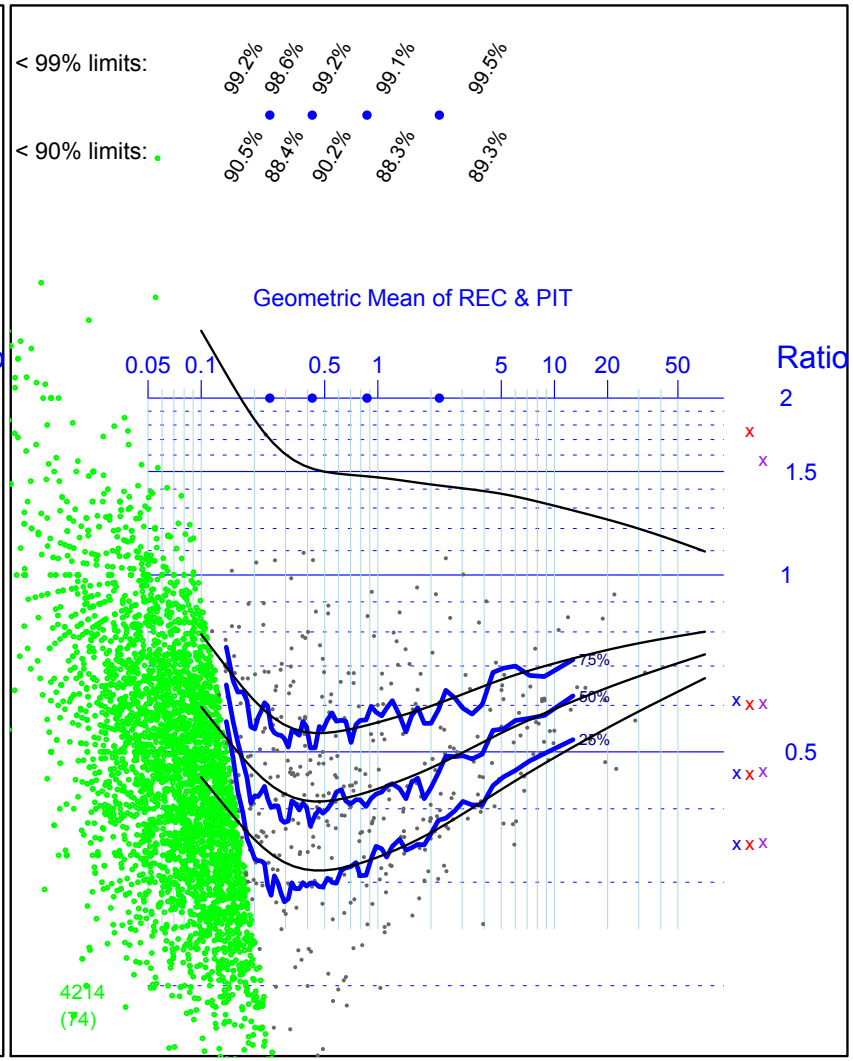
Fig. 6. Cambridge infant growth study weight: fitted centile curves for boys as derived from equation (7): $L(t)$ and $S(t)$ are shown in Figs 4 and 5, while $M(t)$ is shown here as the 50th centile

LMS : ‘X’ = Geom. Mean of REC & PIT

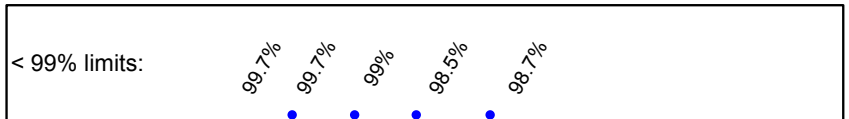
Kit 1 Females (N = 4543 ; 3983 with REC >= 0.1 & PIT >= 0.05)



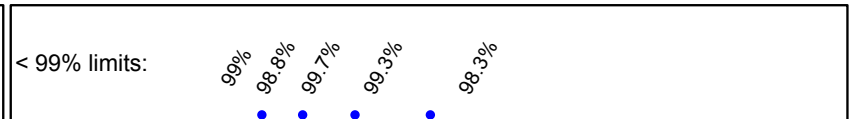
Kit 1 Males (N = 10144 ; 5930 with REC >= 0.1 & PIT >= 0.05)



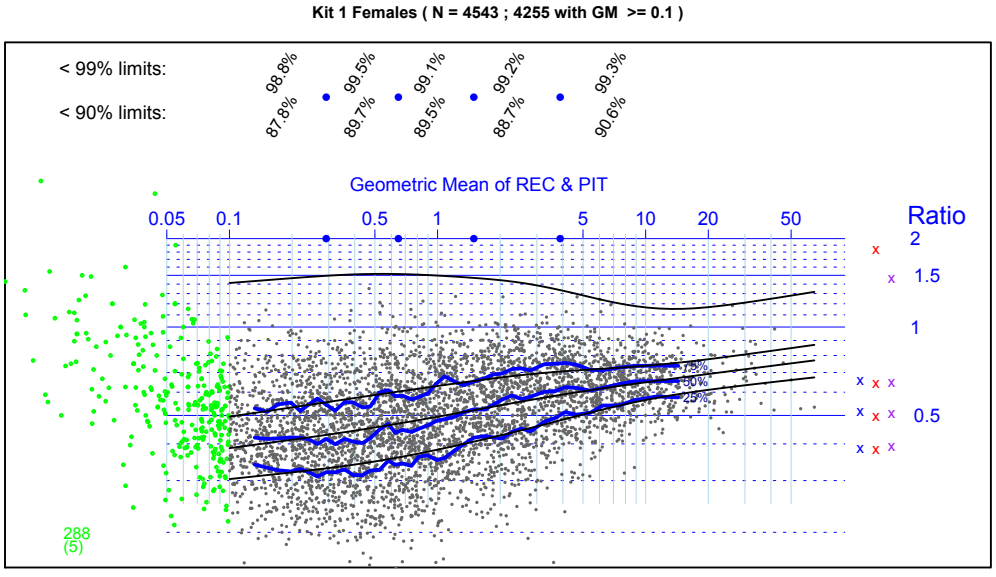
Kit 2 Females (N = 2120 ; 1974 with REC >= 0.1 & PIT >= 0.05)

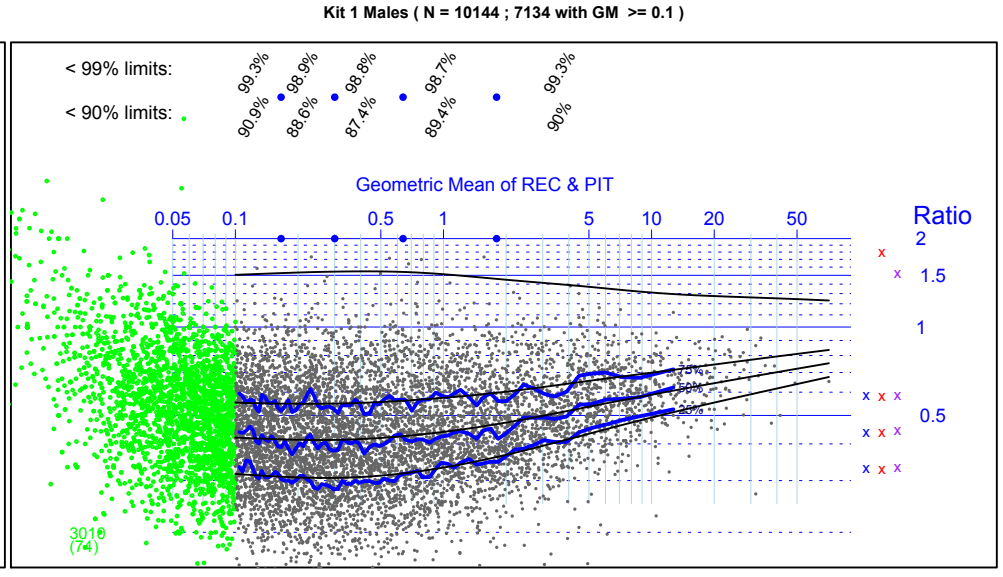


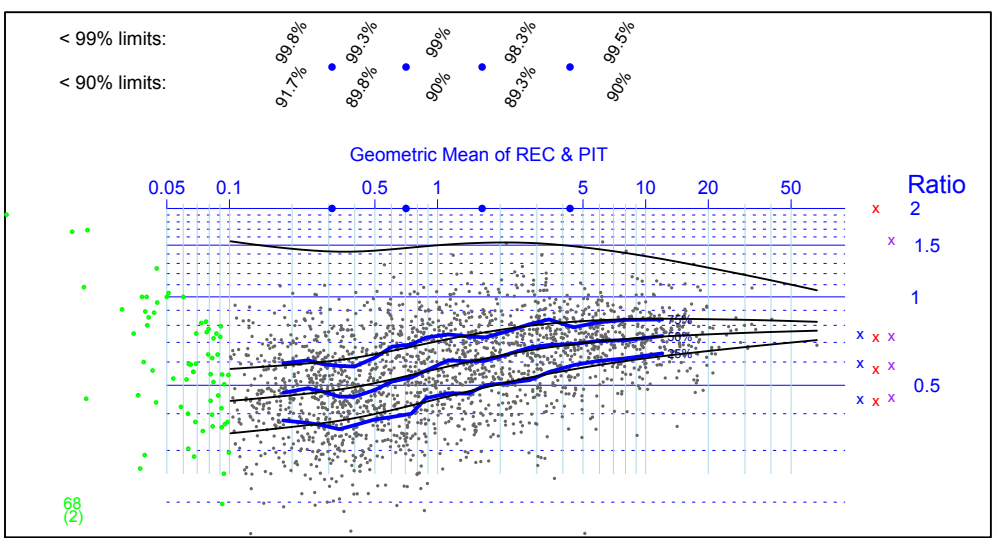
Kit 2 Males (N = 5083 ; 3024 with REC >= 0.1 & PIT >= 0.05)

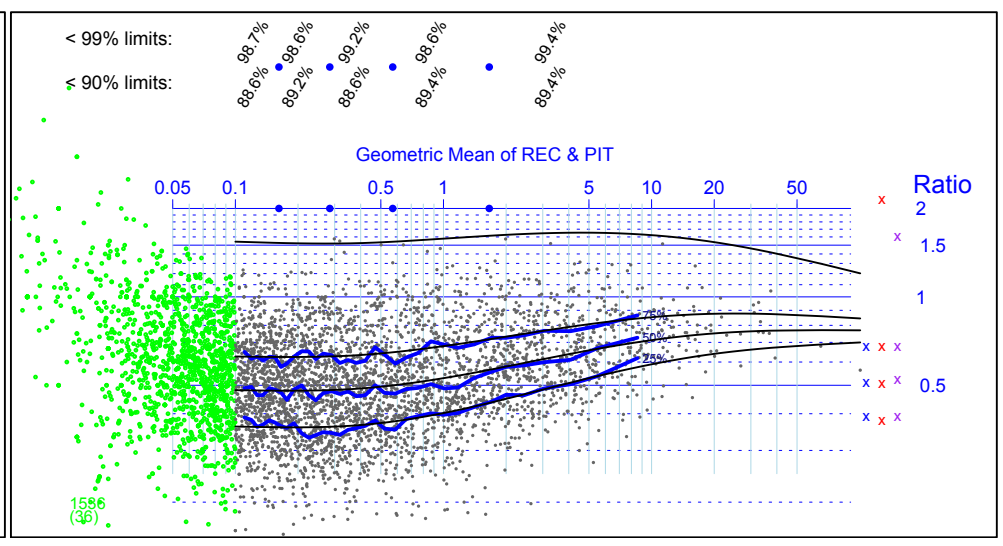


Harmonize: so ‘Low’ cutoff is based on ‘X’





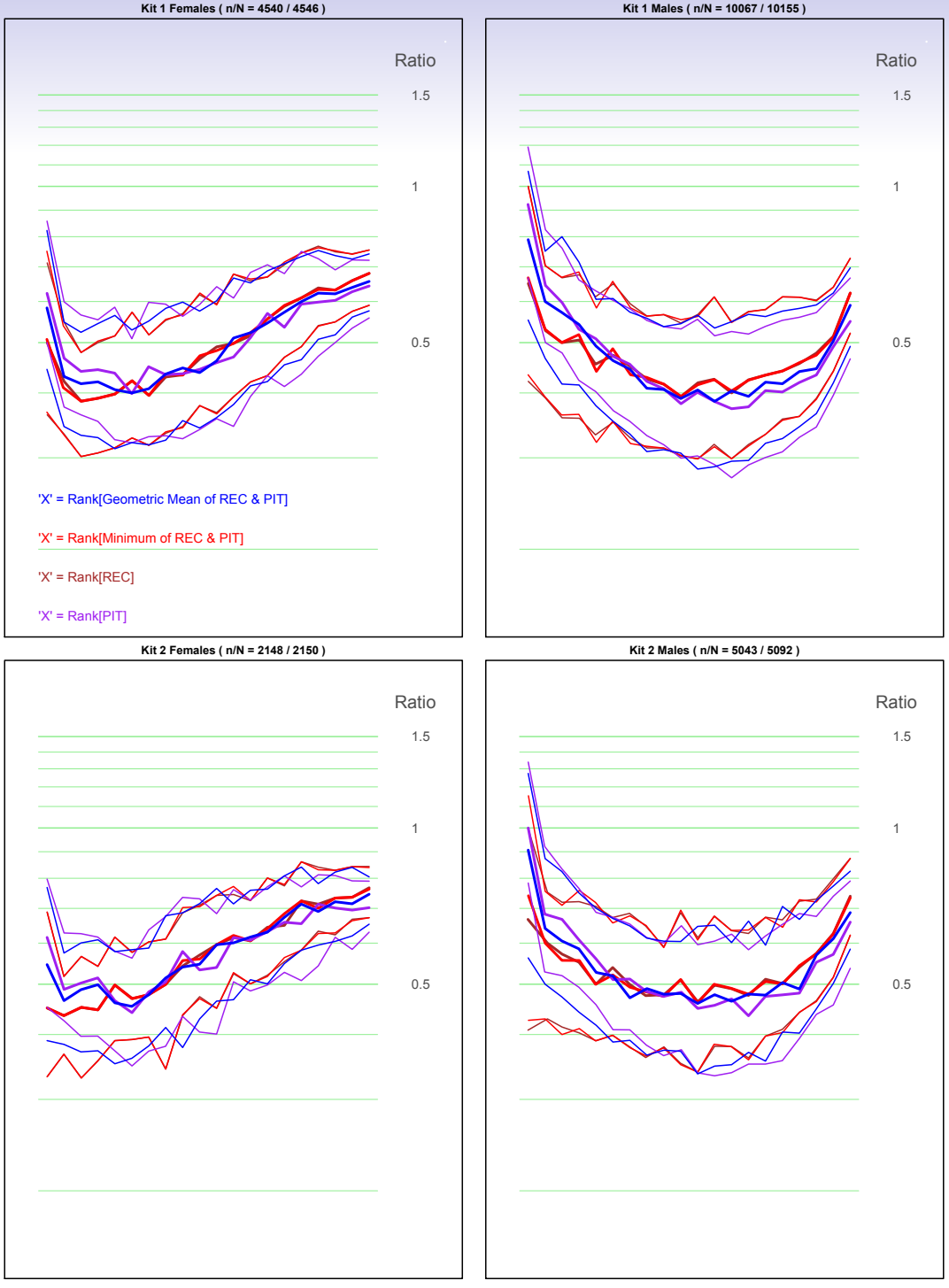


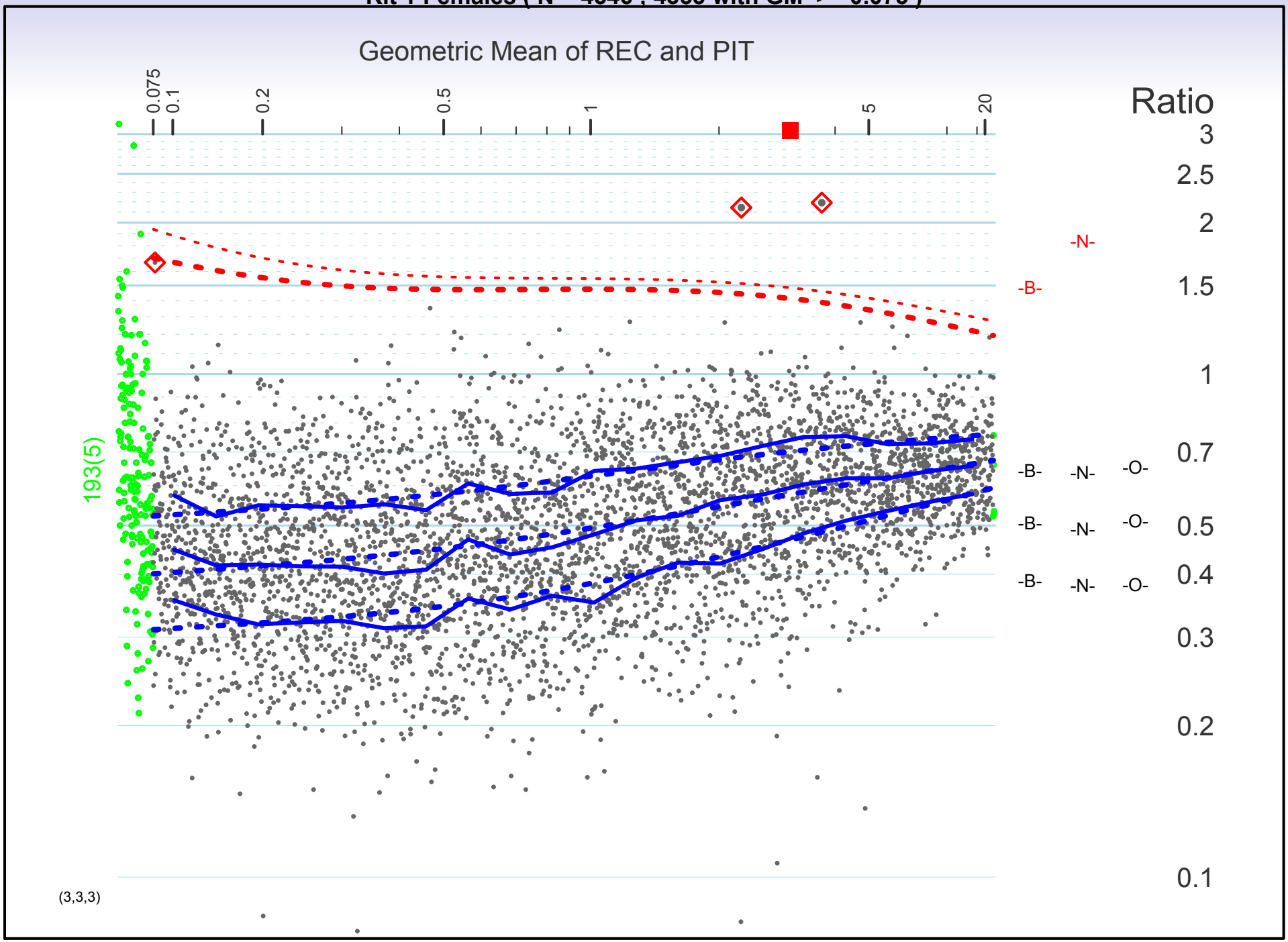


Choice of ‘X’ and its representation

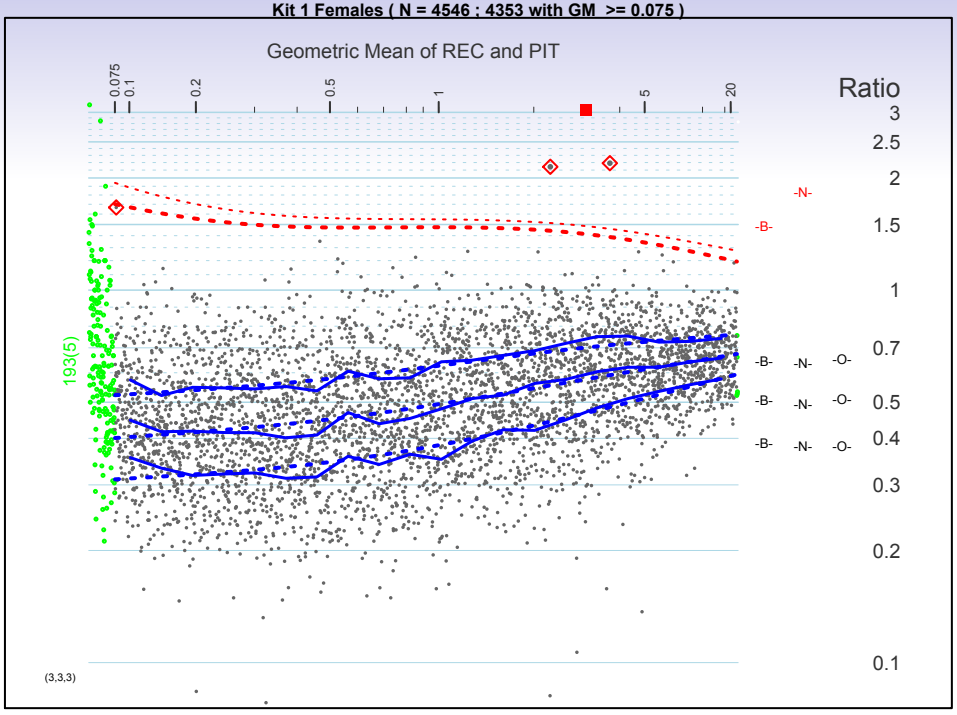
X = Rank(GM of REC & PIT)

Evenly distributed along X-axis

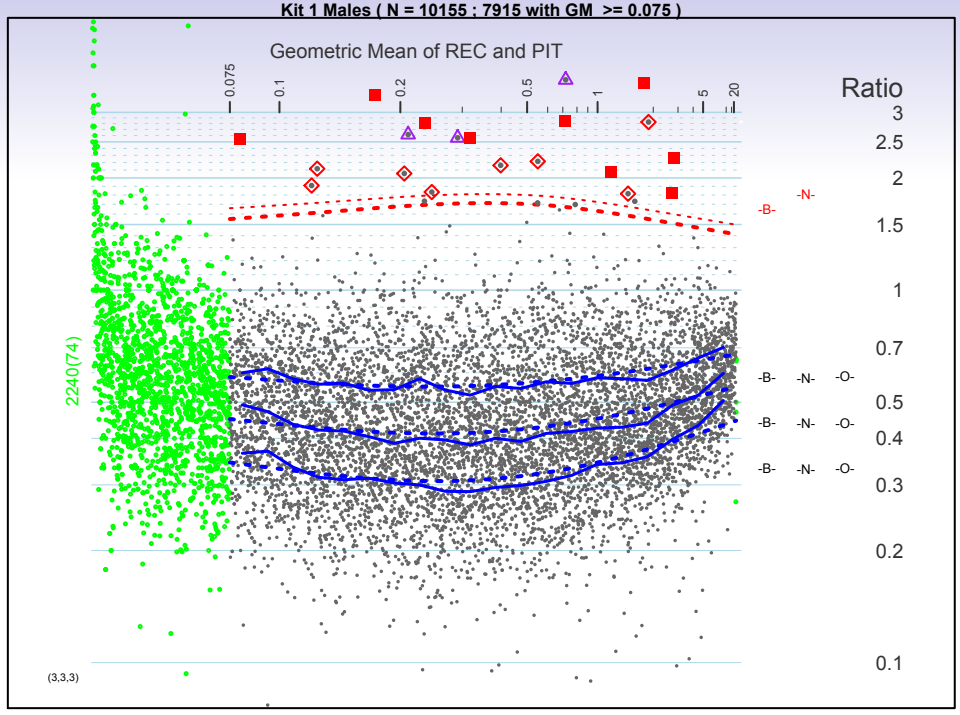



Kit 1 Females (N = 4546 ; 4353 with GM >= 0.075)
Geometric Mean of REC and PIT
0.075
0.1
0.2
0.5
1
5
20
Ratio
3
2.5
2
1.5
1
0.7
0.5
0.4
0.3
0.2
0.1
-N-
-B-
-B- -N- -O-
-B- -N- -O-
-B- -N- -O-
193(5)
(3,3,3)

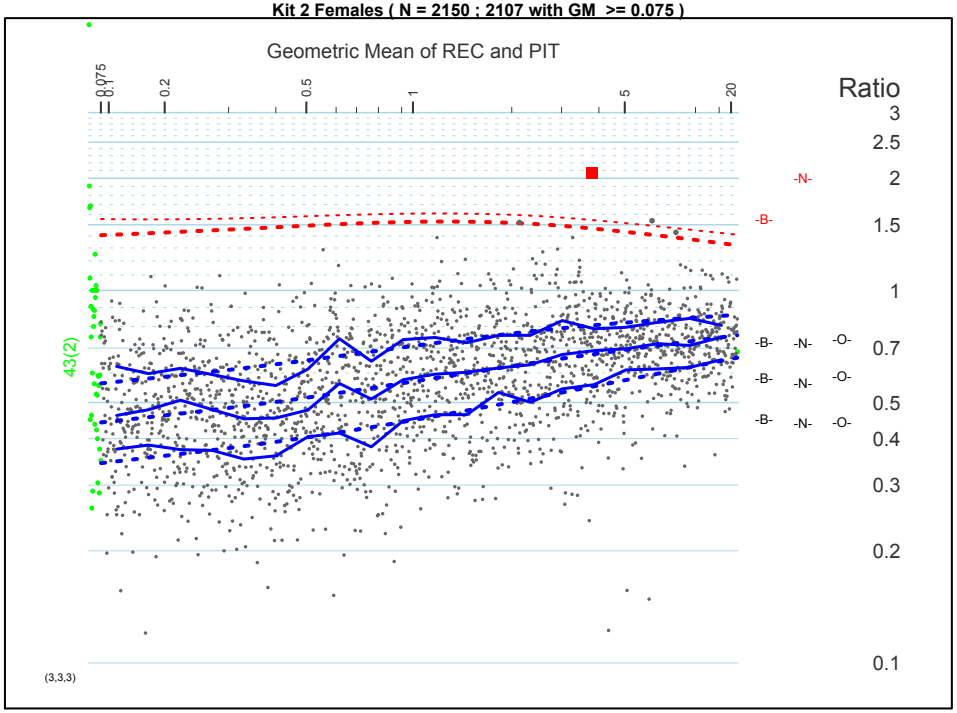
Kit 1 Females (N = 4546 ; 4353 with GM >= 0.075)



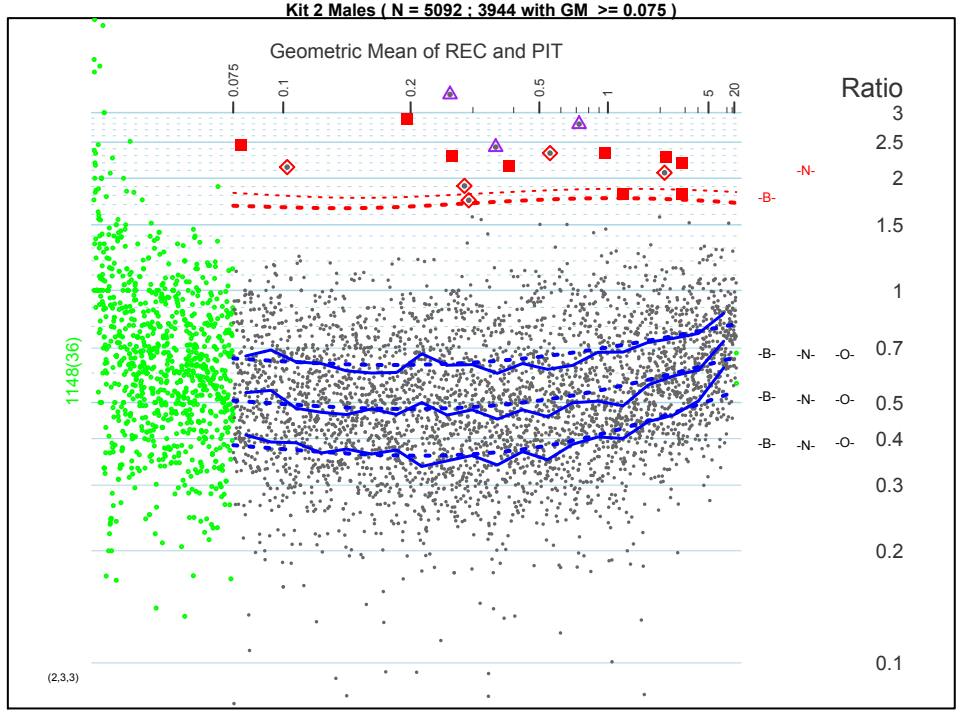
Kit 1 Males (N = 10155 ; 7915 with GM >= 0.075)



Kit 2 Females (N = 2150 ; 2107 with GM >= 0.075)

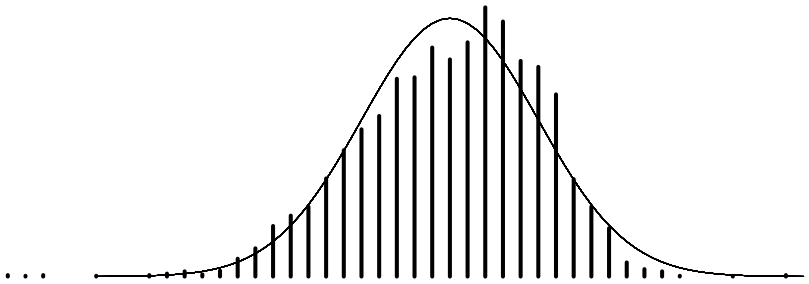


Kit 2 Males (N = 5092 ; 3944 with GM >= 0.075)

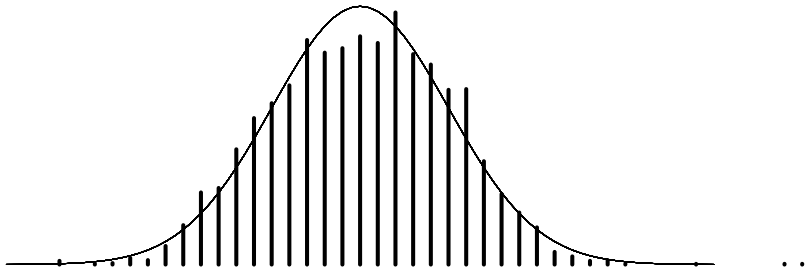


Kit 1 Females (N = 4546 ; 4353 with GM >= 0.075)

Distribution of z-scores
(a) actual, with ref. to a constant mean/variance but otherwise untransformed
(b) actual, with ref. to a constant mean/variance and a single Box-Cox power

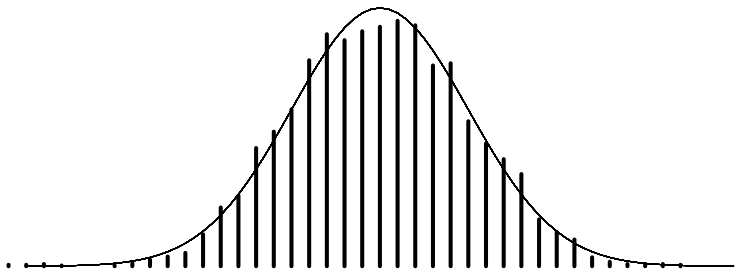


m1, m2, m3, excess kurtosis: 0, 1, -0.41, 0.28

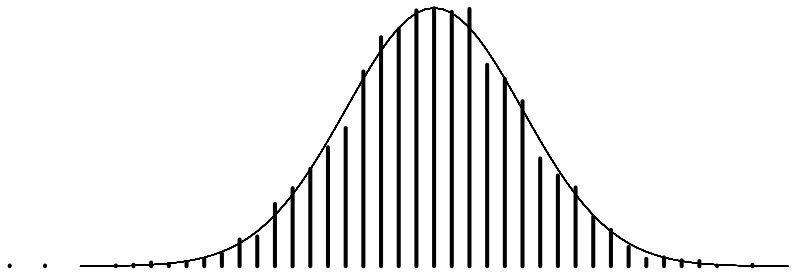


0, 1, -0.01, -0.19

(a) actual, with reference to LMS model
(b) selected randomly from a N(0,1) distribution



m1, m2, m3, excess kurtosis: -0.01, 1.01, 0.04, 0.31



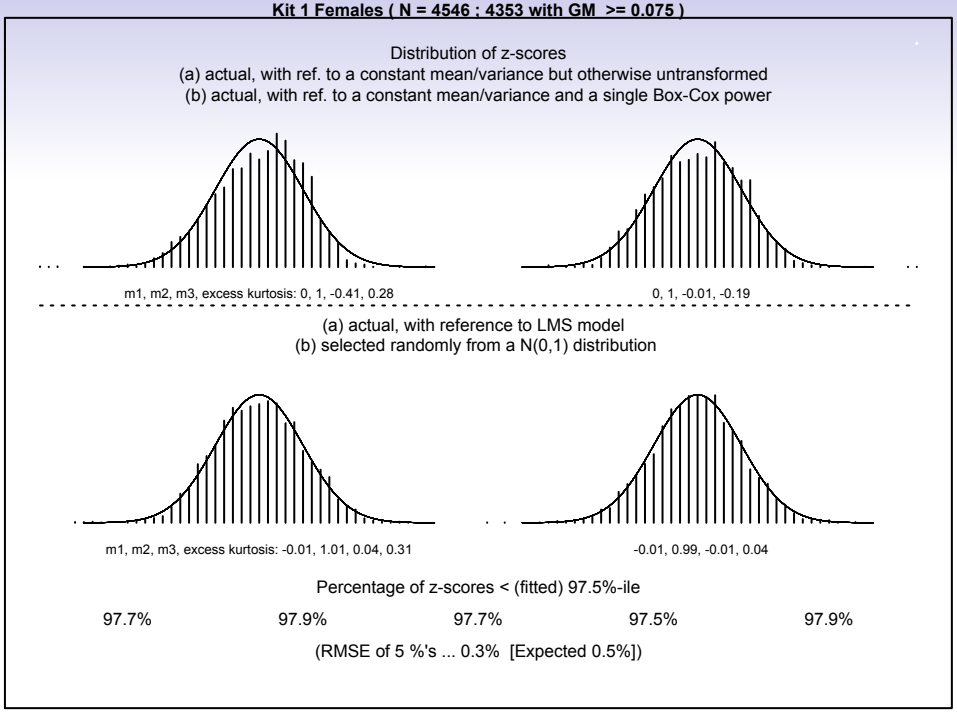
-0.01, 0.99, -0.01, 0.04

Percentage of z-scores < (fitted) 97.5%-ile

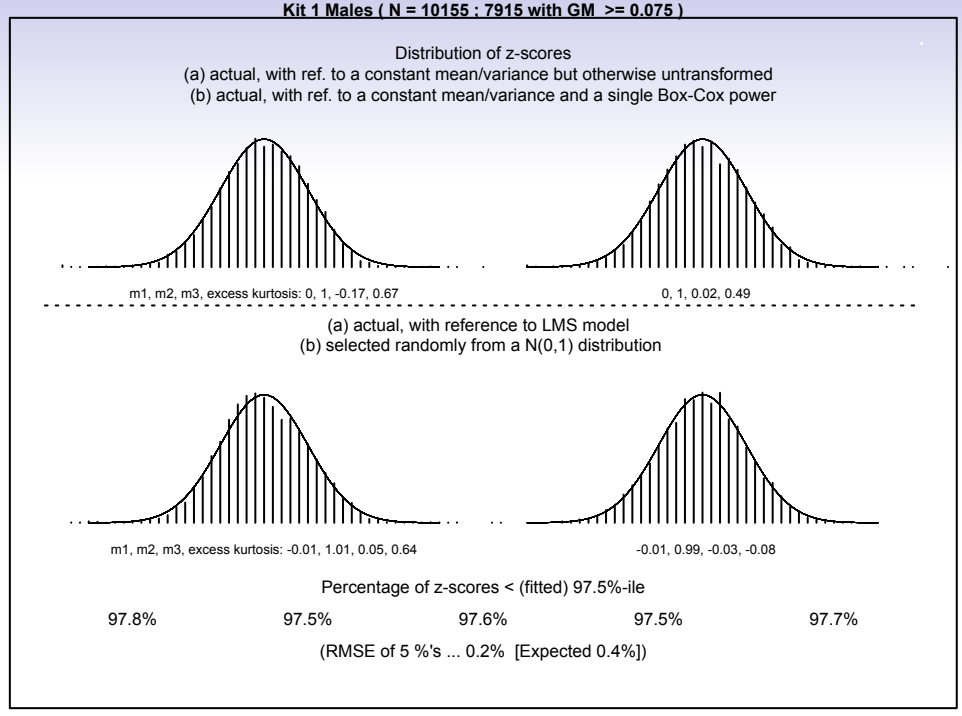
97.7% 97.9% 97.7% 97.5% 97.9%

(RMSE of 5 %'s ... 0.3% [Expected 0.5%])

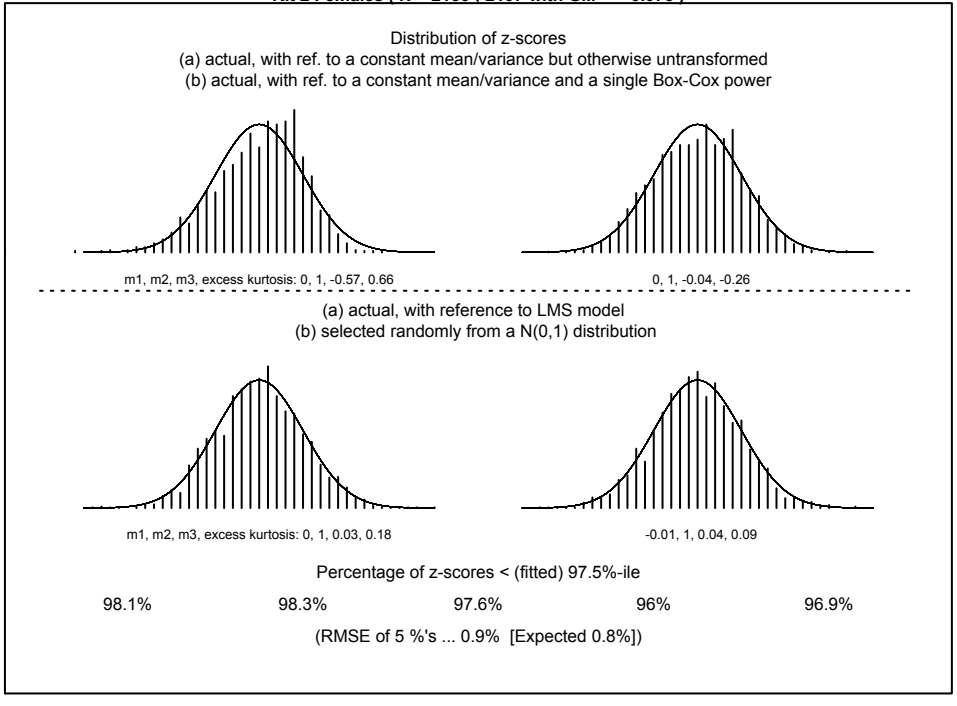
Kit 1 Females (N = 4546 ; 4353 with GM >= 0.075)



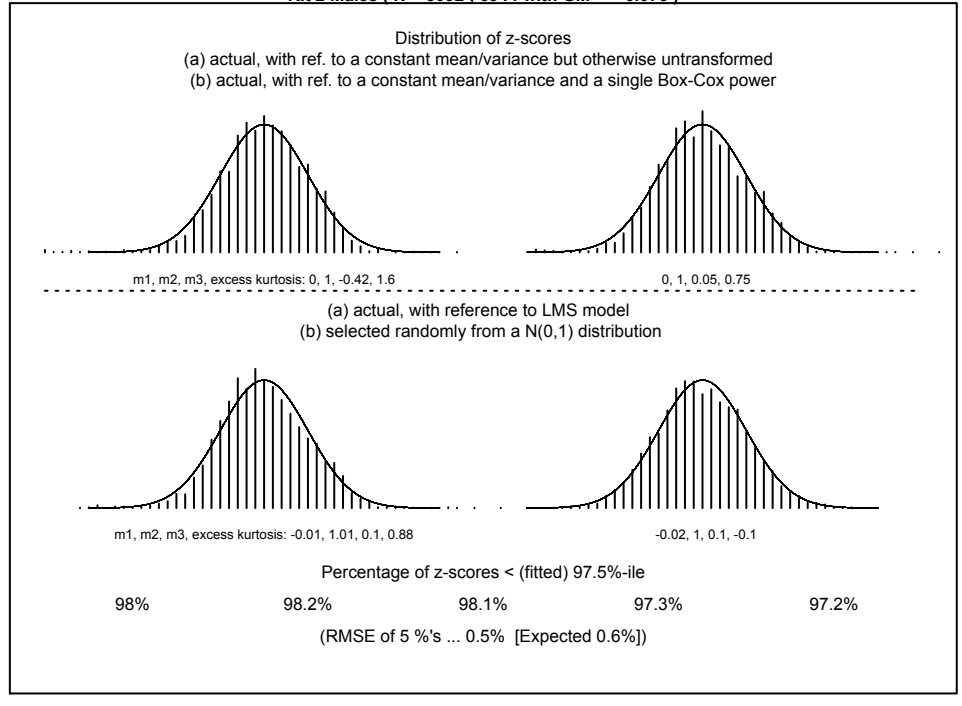
Kit 1 Males (N = 10155 ; 7915 with GM >= 0.075)



Kit 2 Females (N = 2150 ; 2107 with GM >= 0.075)



Kit 2 Males (N = 5092 ; 3944 with GM >= 0.075)



Analysis of the data from human Growth Hormone (hGH) Isoforms Differential Immunoassays in sportspersons, with the objective of setting test compliance decision limits to detect doping with hGH.

Report prepared for the

World Anti-Doping Agency ["WADA"],
800 Square Victoria, Suite 1700, Montreal, QC, Canada, H4Z 1B7,

by

James A. Hanley[1], Olli Saarela[1],
David A. Stephens[2],

[1] Department of Epidemiology, Biostatistics and Occupational Health
[1020 Pine Ave. West, H3A 1A2]

[2] Department of Mathematics and Statistics
[805 Sherbrooke Street West, H3A 0B9]

McGill University, Montreal, Canada

August 11, 2013

(text amended Aug 26 to make clearer references to Figures)

Determining the Decision Limits for the hGH Isoform Differential Immunoassays

Jean- Christophe Thalabard

03/11/2013

Contents

List of Tables

Growth Hormone & IGF Research 24 (2014) 205–215

Contents lists available at ScienceDirect

Growth Hormone & IGF Research

journal homepage: www.elsevier.com/locate/ghir

hGH isoform differential immunoassays applied to blood samples from athletes: Decision limits for anti-doping testing

James A. Hanley [a,b,*], Olli Saarela [a], David A. Stephens [b], Jean-Christophe Thalabard [c,d]

[a] Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montreal, Canada
[b] Department of Mathematics and Statistics, McGill University, Montreal, Canada
[c] Paris Descartes University, MAP5, UMR CNRS 8145, Paris, France
[d] Endocrine Gynaecology Unit, Hôpital Cochin, Paris, France

ARTICLE INFO

Article history:
Received 17 May 2014
Accepted 2 June 2014
Available online 11 June 2014

Keywords:
Quantile
Regression
Decision limits
Isoforms
Human Growth Hormone
Doping

ABSTRACT

Objective: To detect hGH doping in sport, the World Anti-Doping Agency (WADA)-accredited laboratories use the ratio of the concentrations of recombinant hGH ('rec') versus other 'natural' pituitary-derived isoforms of hGH ('pit'), measured with two different kits developed specifically to detect the administration of exogenous hGH. The current joint compliance decision limits (DLs) for ratios derived from these kits, designed so that they would both be exceeded in fewer than 1 in 10,000 samples from non-doping athletes, are based on data accrued in anti-doping labs up to March 2010, and later confirmed with data up to February–March 2011. In April 2013, WADA asked the authors to analyze the now much larger set of ratios collected in routine hGH testing of athletes, and to document in the peer-reviewed literature a statistical procedure for establishing DLs, so that it be re-applied as more data become available.

Design: We examined the variation in the rec/pit ratios obtained for 21,943 screened blood (serum) samples submitted to the WADA accredited laboratories over the period 2009–2013. To fit the relevant sex- and kit-specific

Measuring the Reductions Produced by Cancer Screening:

My campaign for NON-Proportional Hazards

Biometrika 1969

HARVARDgazette

Campus & Community > Obituaries

HSPH's Marvin Zelen dies at 87

Was considered a 'tremendous force' in biostatistics

November 19, 2014 | Editor's Pick

Photo by Shaina Andelman

Harvard Professor Marvin Zelen was noted for developing the statistical methods and study designs that are used in clinical cancer trials, in which experimental drugs are tested for toxicity, effectiveness, and proper dosage.

HSPH Communications

Professor Marvin Zelen of the Department of Biostatistics at the Harvard T.H. Chan School of Public Health (HSPH) died on Nov. 15 after a battle with cancer. He was 87.

On the theory of screening for chroni

By M. ZELEN

State University of New York at Buffalo

AND M. FEINLEIB

National Institutes of Health

Biometrika 1997

Planning clinical trials to evaluate early detection p

By PING HU AND MARVIN ZELEN

Division of Biostatistics, Dana Farber Cancer Institute, 44 Binney Str
Massachusetts 02115, U.S.A.

e-mail: phu@jimmy.harvard.edu zelen@jimmy.harvard.e

Biometrics 2008

Mortality Modeling of Early Detection H

Sandra J. Lee* and Marvin Zelen

Harvard School of Public Health and the Dana-Farber Cancer
Boston, Massachusetts 02115, U.S.A.

* *email:* lee.sandra@jimmy.harvard.edu

Members' Corner

Order of Military Merit

Constitution

Print:

Back to Search

Order of Canada

Maurice McGregor, O.C., C.Q., M.D., F.R.C.P.C.

Montréal, Quebec [Canada]

Officer of the Order of Canada

Awarded on November 16, 2010; Invested on May 27, 2011

The impact of Maurice McGregor's contributions on medicine has spanned cardiology, education and health policy. An exceptional teacher and role model, he served as head of cardiology and head of medicine at the Royal Victoria Hospital, and as dean of medicine at McGill University and at the University of the Witwatersrand, in South Africa. He is recognized for having pioneered and championed the field of health technology assessment in Quebec and throughout Canada, and for having served as founding president of the Conseil d'évaluation des technologies de la santé du Québec. Now professor emeritus, he is chair of the Technology Assessment Unit of the McGill University Health Centre.

Evidence

Études

From the Departments of *Medicine and of †Epidemiology and Biostatistics, McGill

Screening for prostate cancer: estimating the magnitude of overdetection

Canadian Medical Association Journal, 1998

Maurice McGregor,*‡ MD; James A. Hanley,*†§ PhD; Jean-François Boivin,†¶ MD, DSc; Richard George McLean, MB, BS**

Abstract

Olli S MIETTINEN

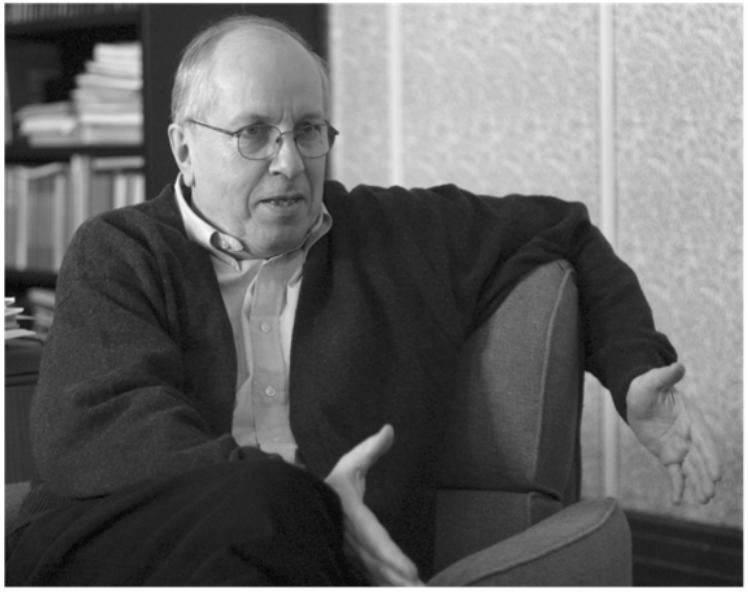
Claudia I Henschke, Mark W Pasmantier, James P Smith, Daniel M Libby, David F Yankelevitz

Year	Screened cohort		Control cohort		Rate ratio (95% CI)
	Actual number	Moving average	Actual number	Moving average	
1	0		0		
2	4	1·3	5	2·0	0·7
3	0	3·3	1	2·7	1·2
4	6	4·0	2	2·7	1·5
5	6	5·3	5	4·0	1·3
6	4	5·7	5	5·7	1·0
7	7	5·0	7	7·3	0·7 (0·36–1·31)
8	4	4·3	10	8·3	0·5 (0·27–1·00)*
9	2	2·7	8	6·3	0·4 (0·19–0·94)*
10	2	3·3	1	7·0	0·5 (0·23–0·99)*
11	6†		12†		

*Based on years 8–11, rate ratio point estimate is 14/31=0·45 (95% CI 0·24–0·84). †Some of these deaths (from 1987) probably belong to year 10 or even to year 9.

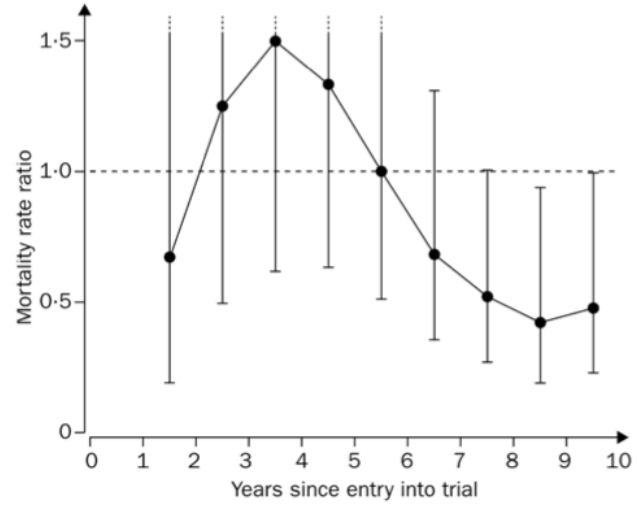
Table 1: **Number of breast-cancer deaths by year after entry into Malmö study for women 55–69 years of age at entry**

have occurred only after a suitable delay.



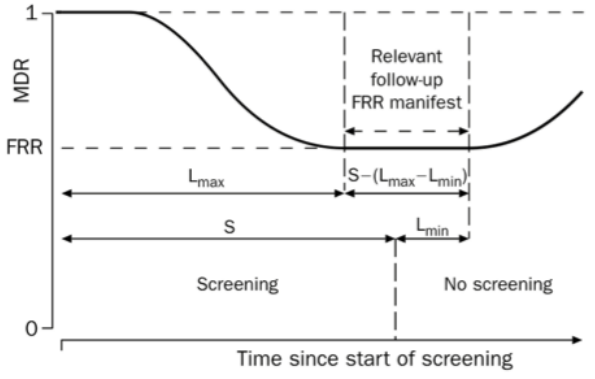
Mammographic screening: no reliable supporting evidence?

The Lancet, 2002



Breast-cancer mortality ratio for women at least 55 years of age in the Malmö study

Shown are point estimates and 95% CI, based on the deaths in the year at issue together with those in the preceding and following years.



Follow-up experience in a randomised controlled trial comparing screening for cancer with no screening in respect to cause-specific mortality: interrelations of parameters

At any given point in the follow-up there is a particular mortality density, MD, among the screened and the not screened; for an interval of t to t+dt, with dC cases expected in it, $MD_t=dC/Pdt$, where P is the size of the population. Contrasting the screened with the not screened, there is the corresponding mortality-density ratio, MDR. This ratio is depicted as a function of time since entry into the trial. The early excess mortality among the screened is not shown, since focus is on the intended result of reduced fatality rate, FR, quantified in terms of fatality-rate ratio, FRR. MDR coincides with FRR in a particular interval of follow-up time if the duration of screening, S, exceeds the difference between the maximum,

A single Hazard Ratio is Appropriate if Reduction is VIRTUALLY IMMEDIATE & ...

- SUSTAINED
  - **Adult circumcision** quickly reduces the risk of **getting HIV** by about 50%; reduced rate is lifelong.
  - **Polio, HPV, ...** Once there is full immunity, **vaccine** protection lasts for decades.

or if we ...

- STOP COUNTING AS SOON AS PROTECTION STOPS
  - **Blood thinners**
  - **beta blockers**

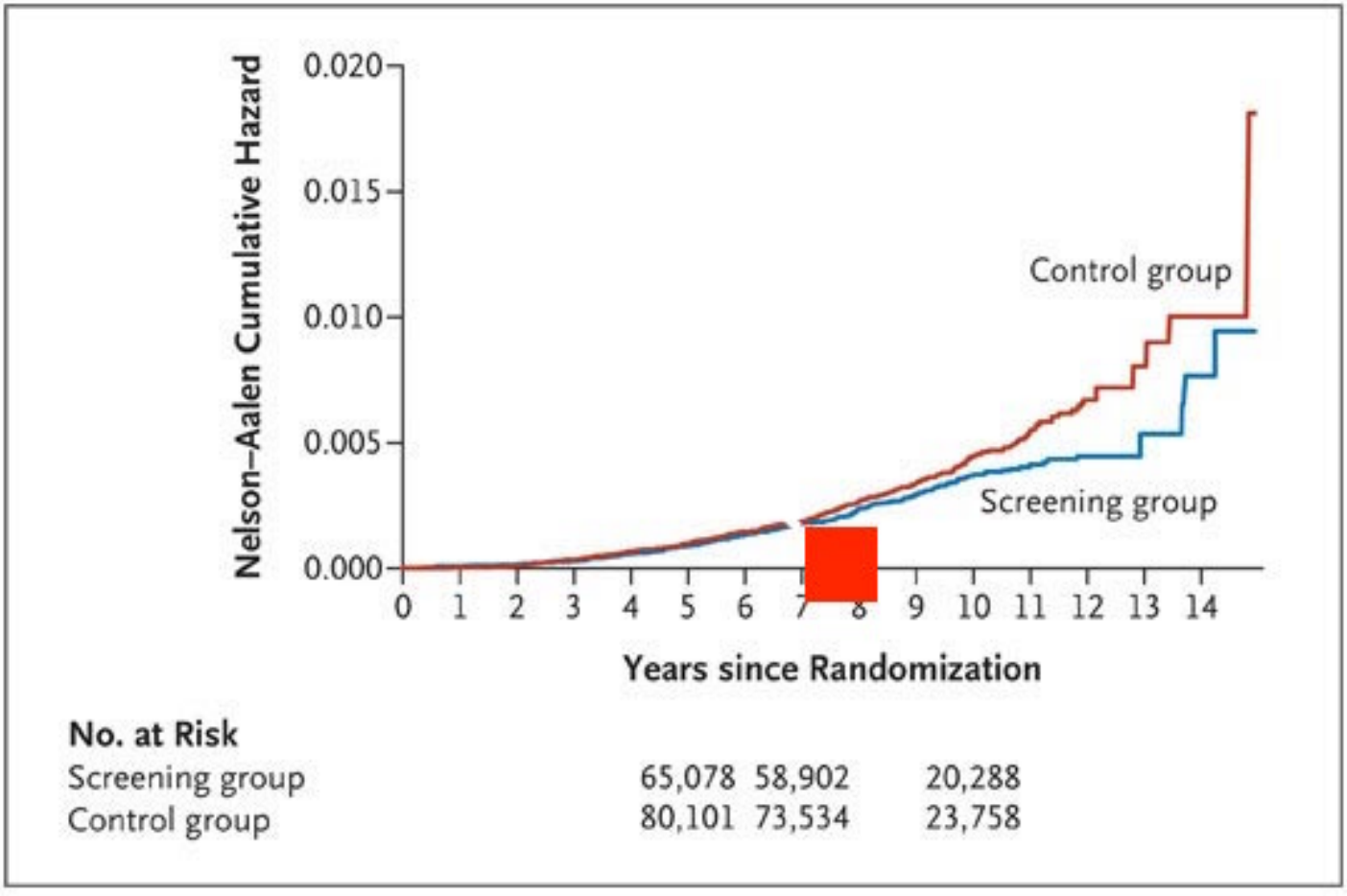
Reduction is CONSIDERABLY DELAYED following ...

PROSTATE CANCER SCREENING

Screening & Prostate-Ca Mortality in Randomized European Study '92-'08 ("ERSPC" nejm2009.04)

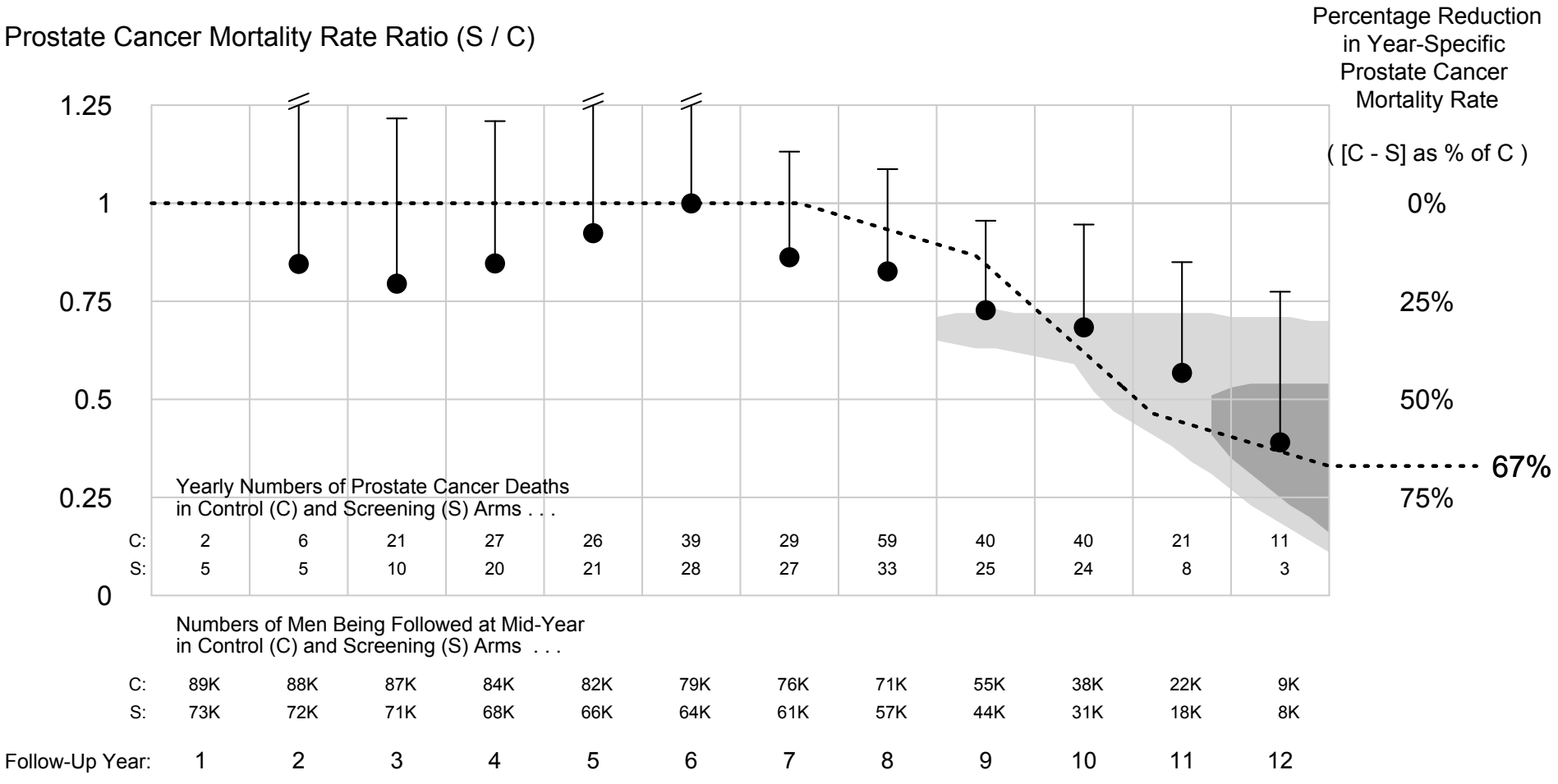
8.8 years mean F.U., 214 & 326 deaths: **HAZARD RATIO: 0.80**

"PSA-based screening reduced rate of [pr. ca.] death **by 20%**. "



RE-ANALYSIS OF ERSPC DATA
using
year-specific prostate cancer mortality ratios

Year-specific mortality ratios



Hanley, *J Medical Screening*, 2010.

Time-Distribution of Deaths was ‘FRONT-LOADED’

very staggered entry;

many more man-years & deaths at front than at back end

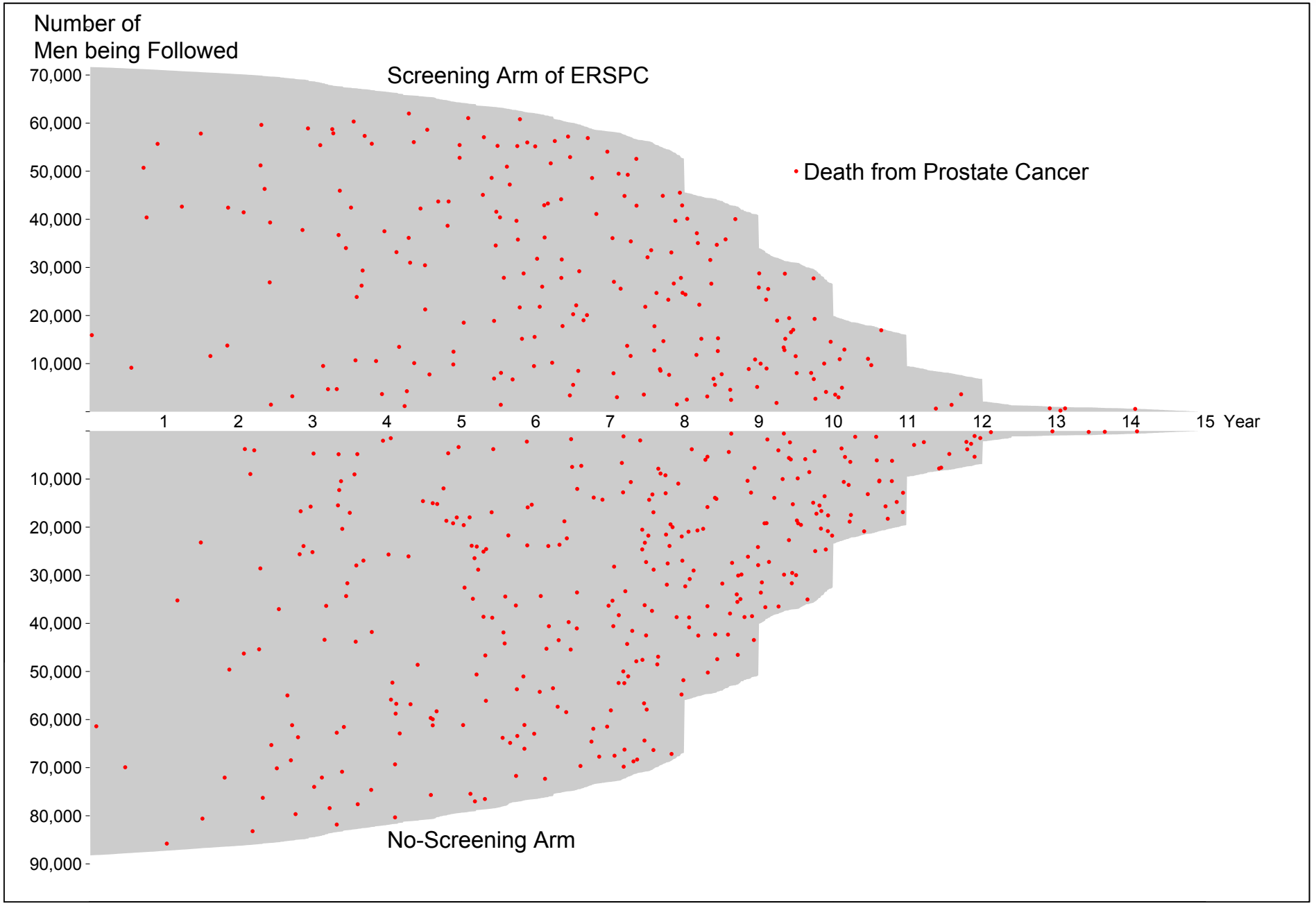
Yearly Numbers of Prostate Cancer Deaths in Control (C) and Screening (S) Arms . . .

C:	2	6	21	27	26	39	29	59	40	40	21	11
S:	5	5	10	20	21	28	27	33	25	24	8	3

Numbers of Men Being Followed at Mid-Year in Control (C) and Screening (S) Arms . . .

C:	89K	88K	87K	84K	82K	79K	76K	71K	55K	38K	22K	9K
S:	73K	72K	71K	68K	66K	64K	61K	57K	44K	31K	18K	8K
Year:	1	2	3	4	5	6	7	8	9	10	11	12

‘POPULATION-TIME’ Plot



COLON CANCER

FOBT screening for colon cancer – Minnesota Trial 1976-2008

Long-Term Mortality after Screening for Colorectal Cancer

Aasma Shaukat, M.D., M.P.H., Steven J. Mongin, M.S., Mindy S. Geisser, M.S., Frank A. Lederle, M.D., John H. Bond, M.D., Jack S. Mandel, Ph.D., M.P.H., and Timothy R. Church, Ph.D.

ABSTRACT

BACKGROUND

In randomized trials, fecal occult-blood testing reduces mortality from colorectal cancer. However, the duration of the benefit is unknown, as are the effects specific to age and sex.

METHODS

In the Minnesota Colon Cancer Control Study, 46,551 participants, 50 to 80 years of age, were randomly assigned to usual care (control) or to annual or biennial screening with fecal occult-blood testing. Screening was performed from 1976 through 1982 and from 1986 through 1992. We used the National Death Index to obtain updated information on the vital status of participants and to determine causes of death through 2008.

From the Divisions of Gastroenterology (A.S., J.H.B.) and Internal Medicine (F.A.L.), Minneapolis Veterans Affairs Health Care System, and the Department of Medicine, School of Medicine (A.S., F.A.L., J.H.B.), and the Division of Environmental Health Sciences, School of Public Health (S.J.M., M.S.G., T.R.C.), University of Minnesota — both in Minneapolis; and Exponent, Menlo Park, CA (J.S.M.). Address reprint requests to Dr. Shaukat at 1 Veterans Dr., 111-D, Minneapolis, MN 55417.

N Engl J Med 2013;369:1106-14.
DOI: 10.1056/NEJMoa1300720

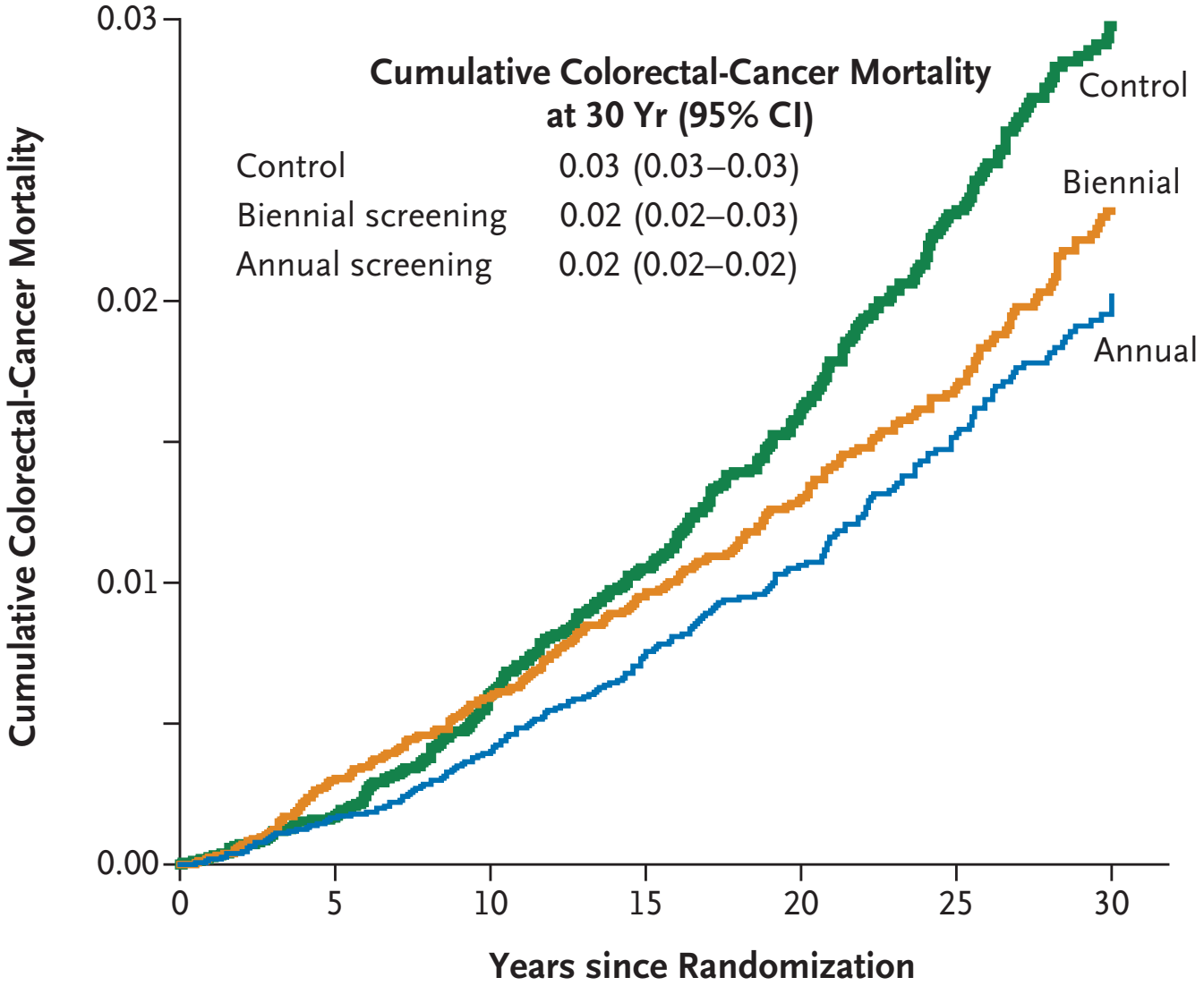
FOBT screening for colon cancer – Minnesota Trial 1976-2008

RESULTS

Through 30 years of follow-up, 33,020 participants (70.9%) died. A total of 732 deaths were attributed to colorectal cancer: 200 of the 11,072 deaths (1.8%) in the annual-screening group, 237 of the 11,004 deaths (2.2%) in the biennial-screening group, and 295 of the 10,944 deaths (2.7%) in the control group. Screening reduced colorectal-cancer mortality (relative risk with annual screening, 0.68; 32% onfidence interval [CI], 0.56 to 0.82; relative risk with biennial screening, 0.78; 22% , 0.65 to 0.93) through 30 years of follow-up. No reduction was observed in all-cause mortality (relative risk with annual screening, 1.00; 95% CI, 0.99 to 1.01; relative risk with biennial screening, 0.99; 95% CI, 0.98 to 1.01). The reduction in colorectal-cancer mortality was larger for men than for women in the biennial-screening group (P=0.04 for interaction).

CONCLUSIONS

The effect of screening with fecal occult-blood testing on colorectal-cancer mortality persists after 30 years but does not influence all-cause mortality. The sustained reduction in colorectal-cancer mortality supports the effect of polypectomy. (Funded by the Veterans Affairs Merit Review Award Program and others.)



No. at Risk

Control	14,497	13,103	11,320	9157	6741	4450
Biennial screening	14,635	13,243	11,445	9323	6802	4583
Annual screening	14,658	13,294	11,437	9219	6802	4498

Figure 1. *Rep. Alexander Pirnie, R-NY, draws the first capsule in the lottery drawing held on Dec. 1, 1969. The capsule contained the date. Sept. 14.*

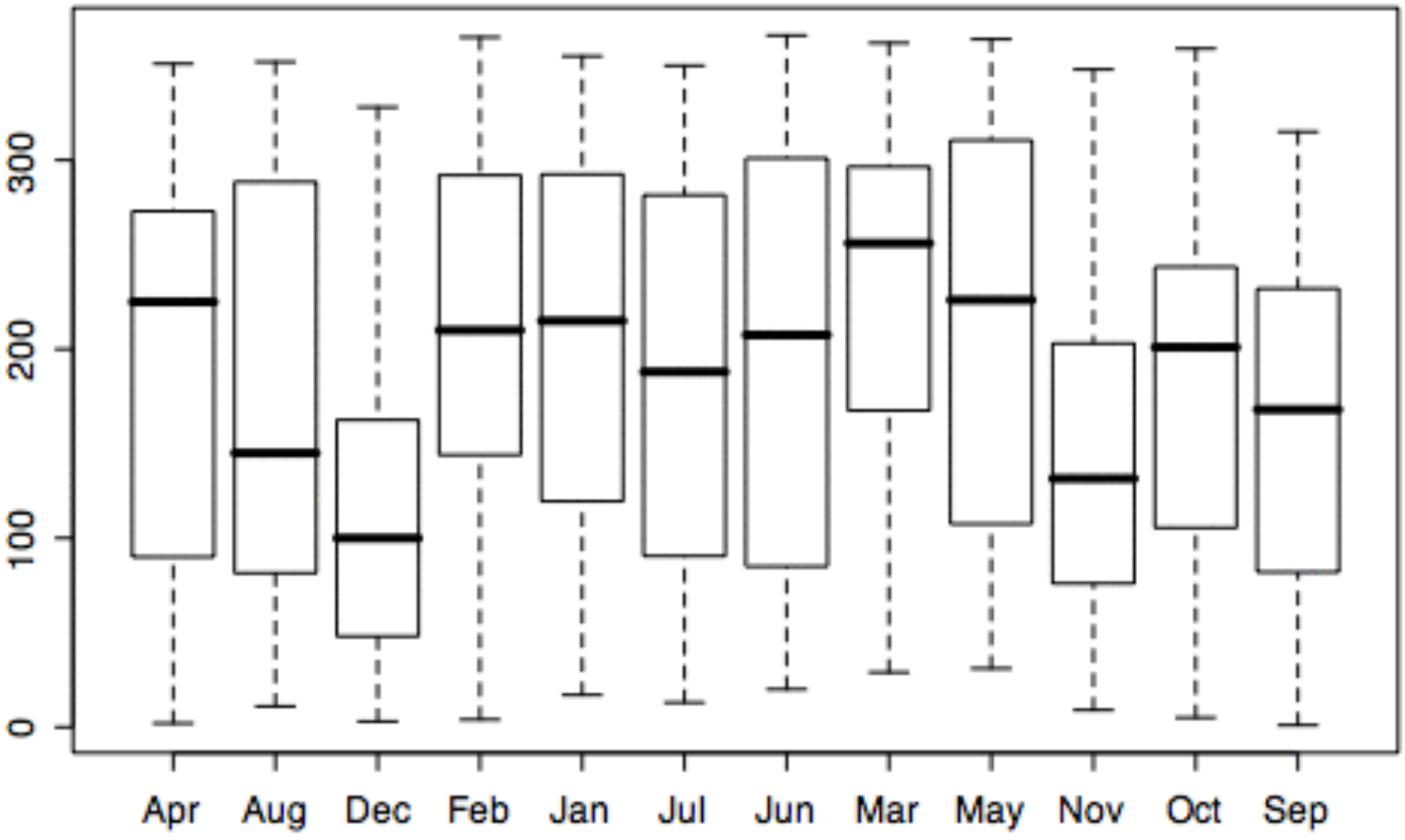


Figure 4. Side-by-side boxplots of draft numbers for each month.

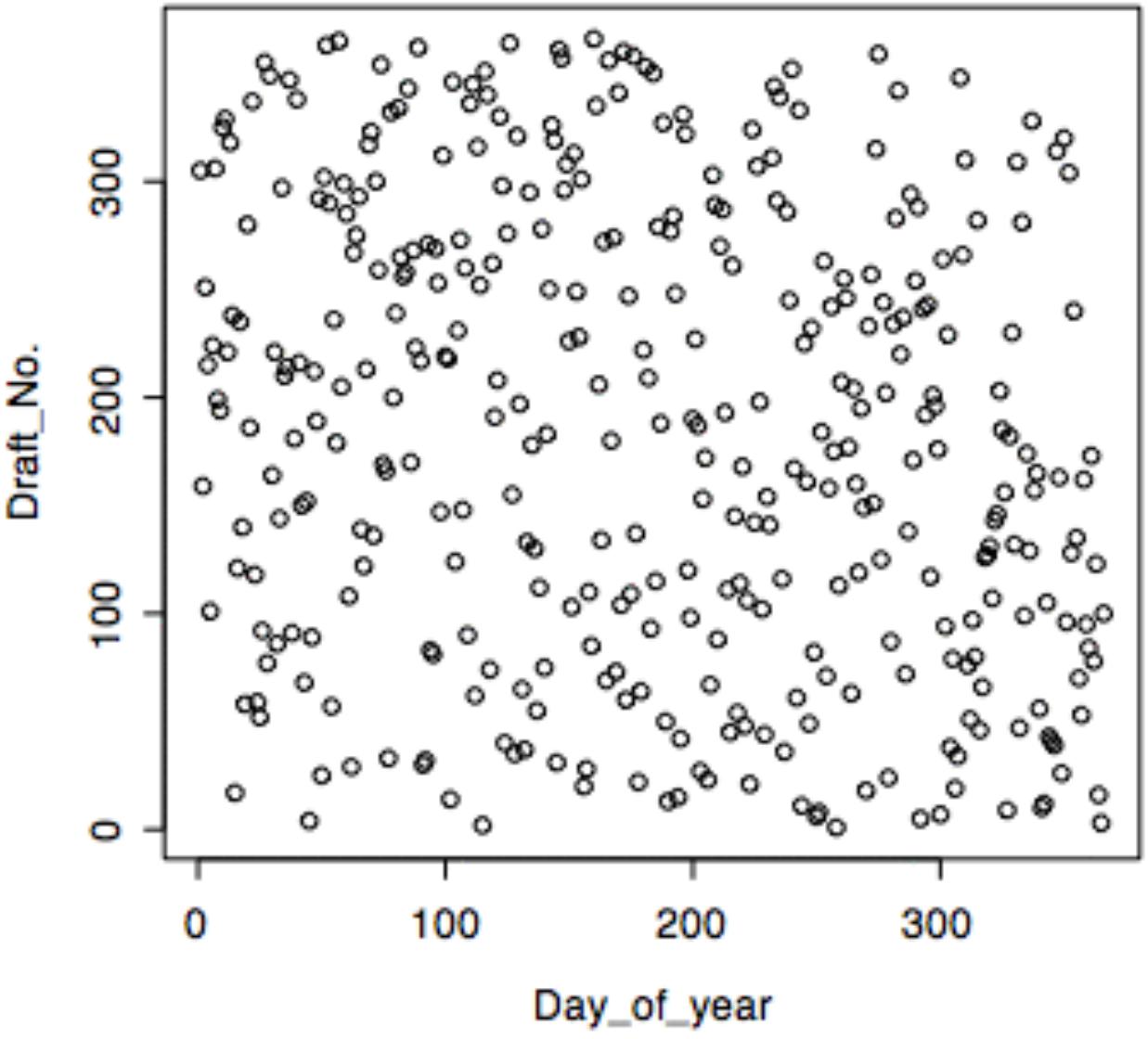


Figure 2. A scatterplot of ***Draft_No.*** *versus* ***Day_of_year****.*

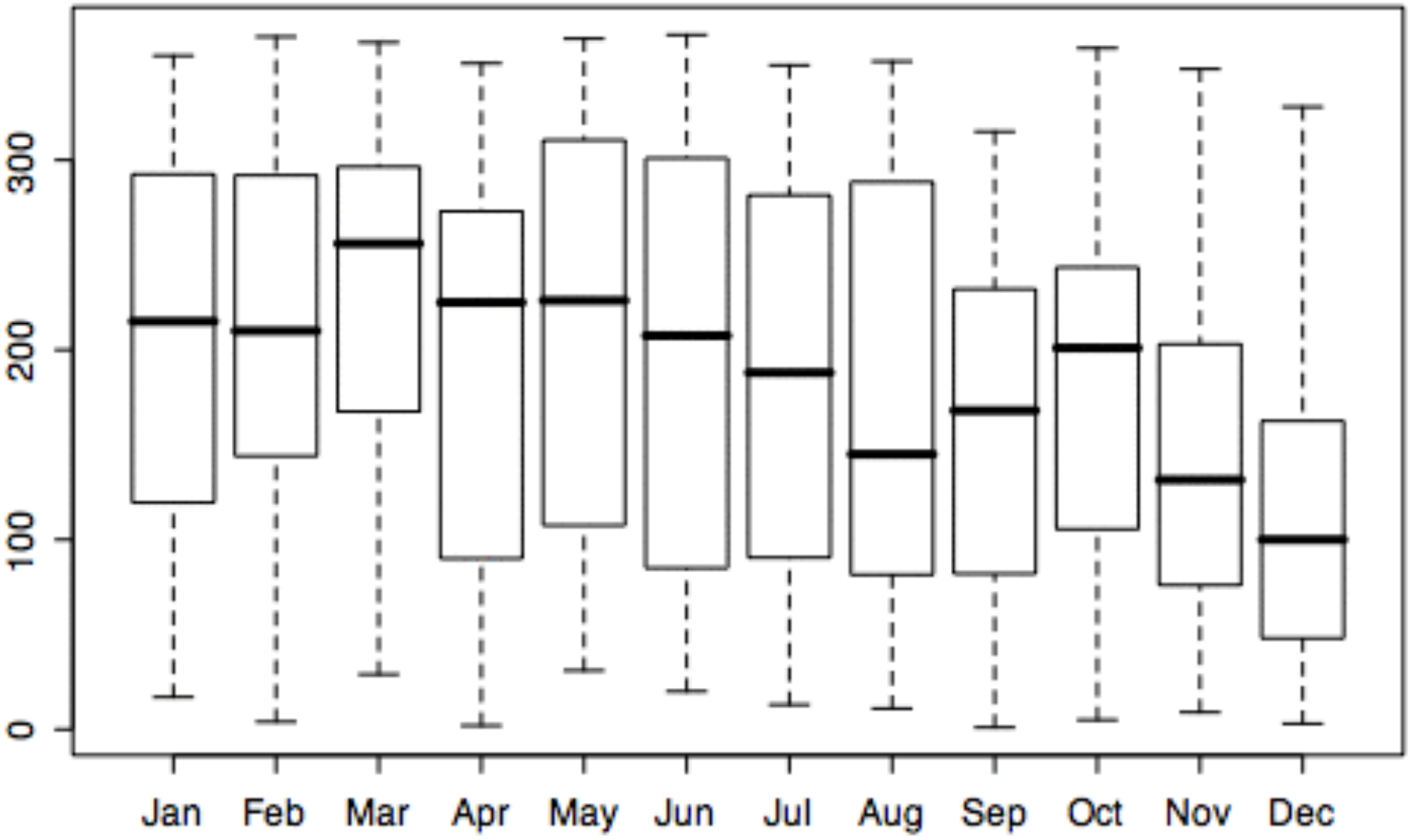
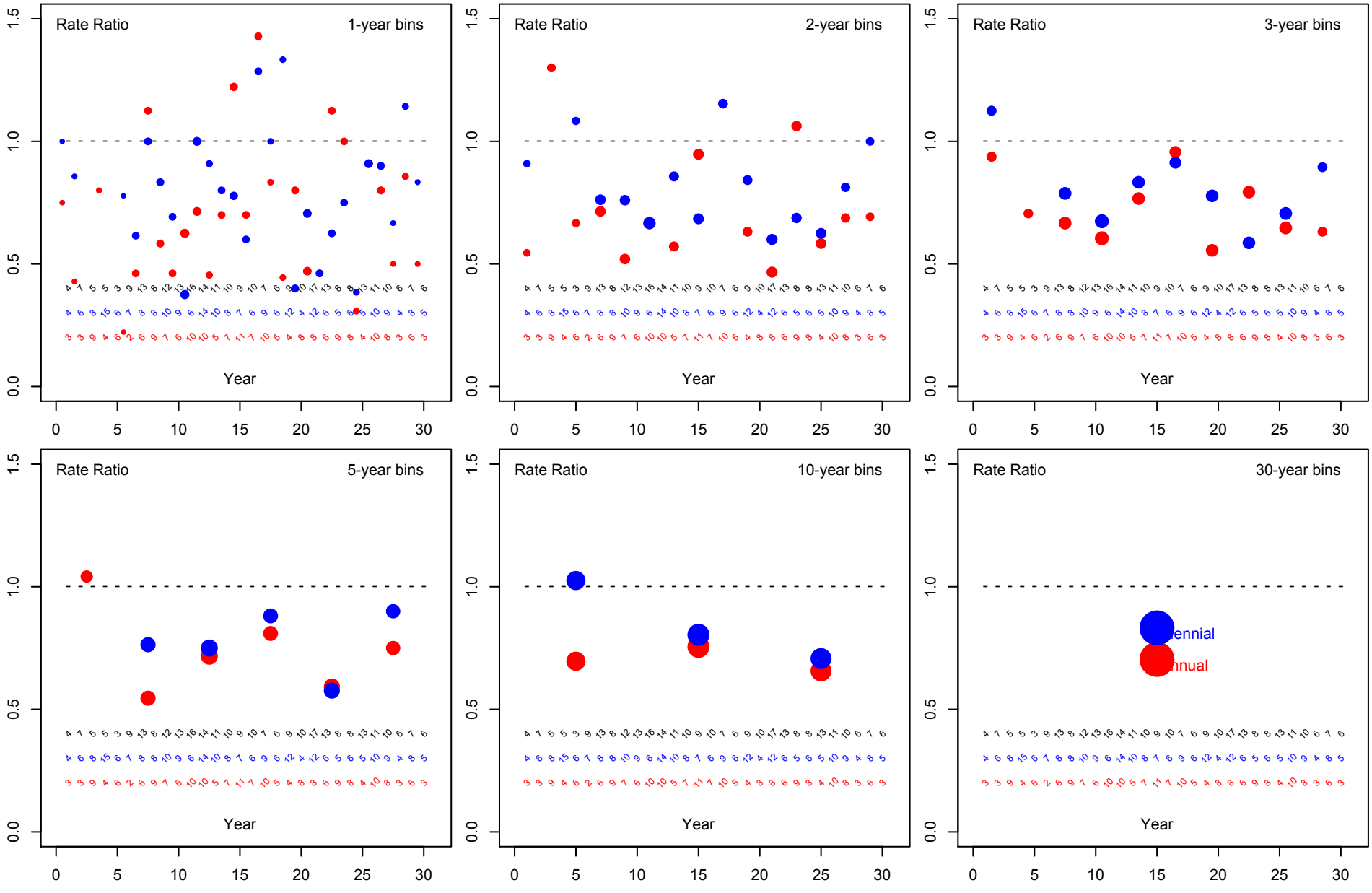
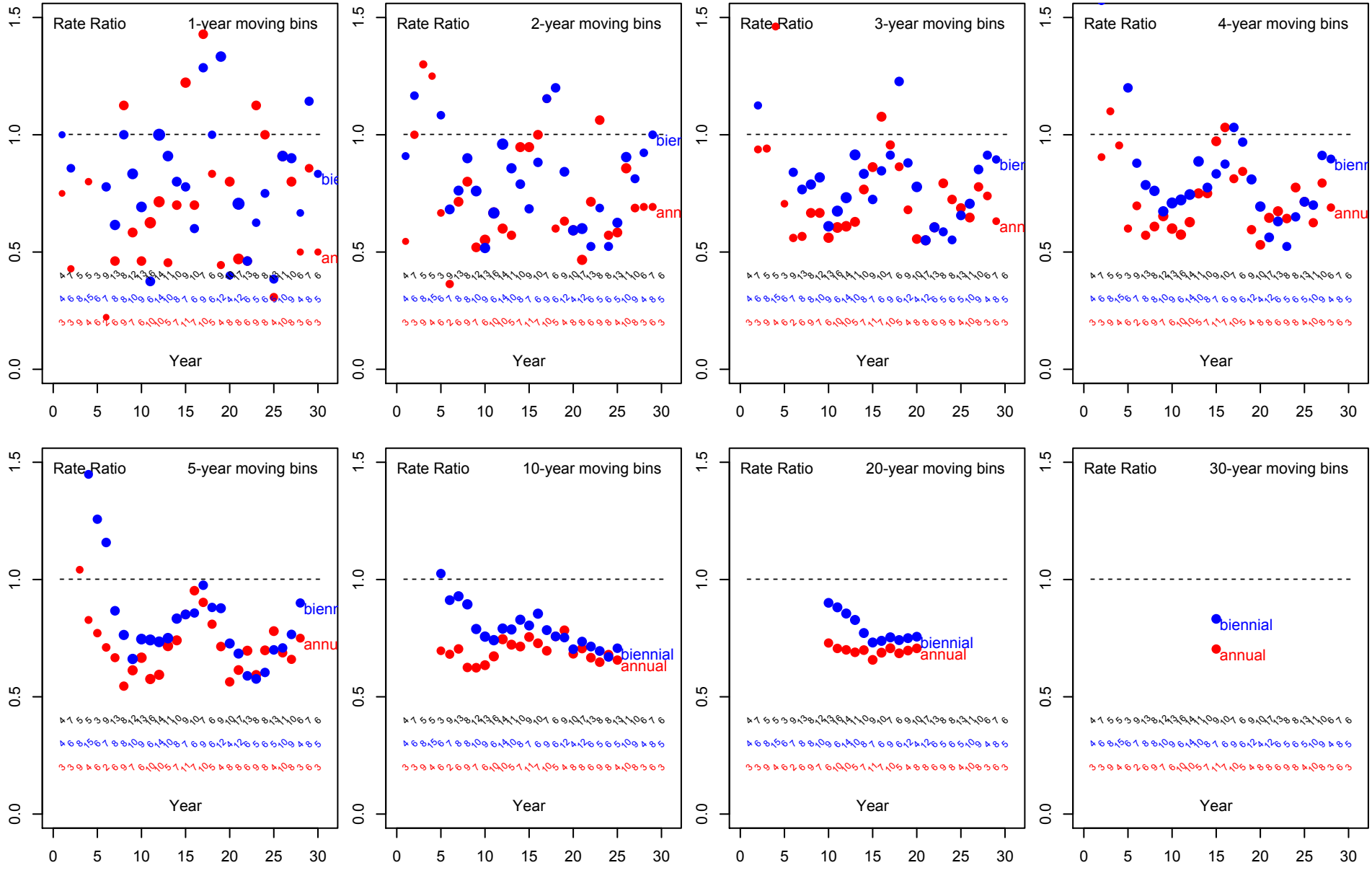


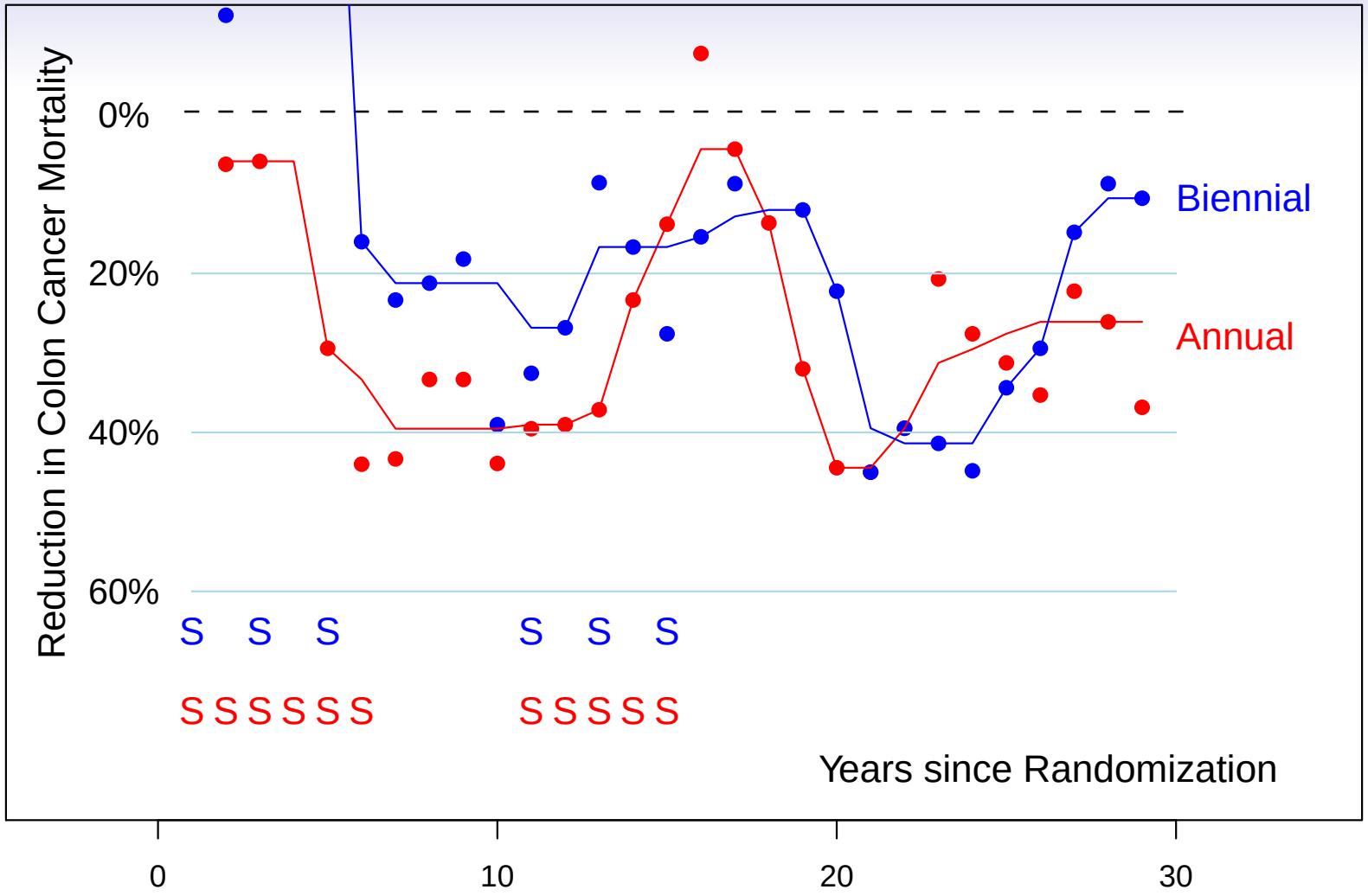
Figure 6. Side-by-side boxplots of draft numbers sorted by month.

Time-**split** versus time-**lumped** Rate Ratios



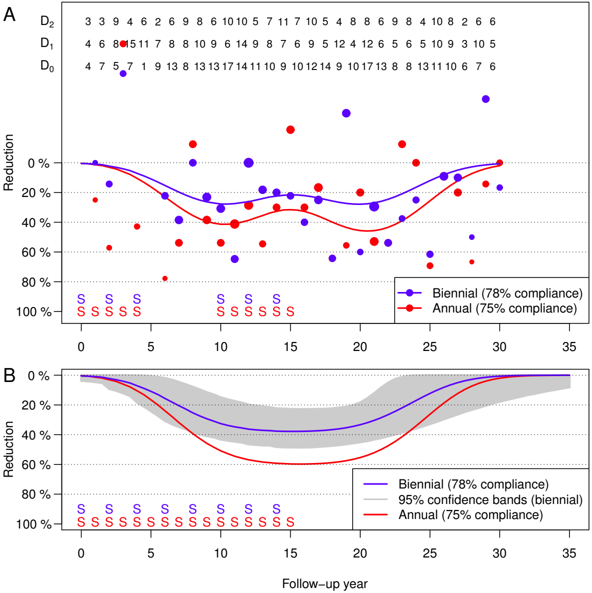
Time-**split** versus time-**lumped** Rate Ratios





Yearly reductions in colon cancer mortality in two screening arms. Each dot is based on number of deaths in a three year moving window; smooth curves were fitted though them. Because the hiatus was in calendar-time rather than follow-up time, and entries were staggered, the timing of the screens (each denoted by an ‘S’) is only approximate.

Liu Model: A Fitted to Data; B Projected i.e., no interruption. 6 & 11 Rounds

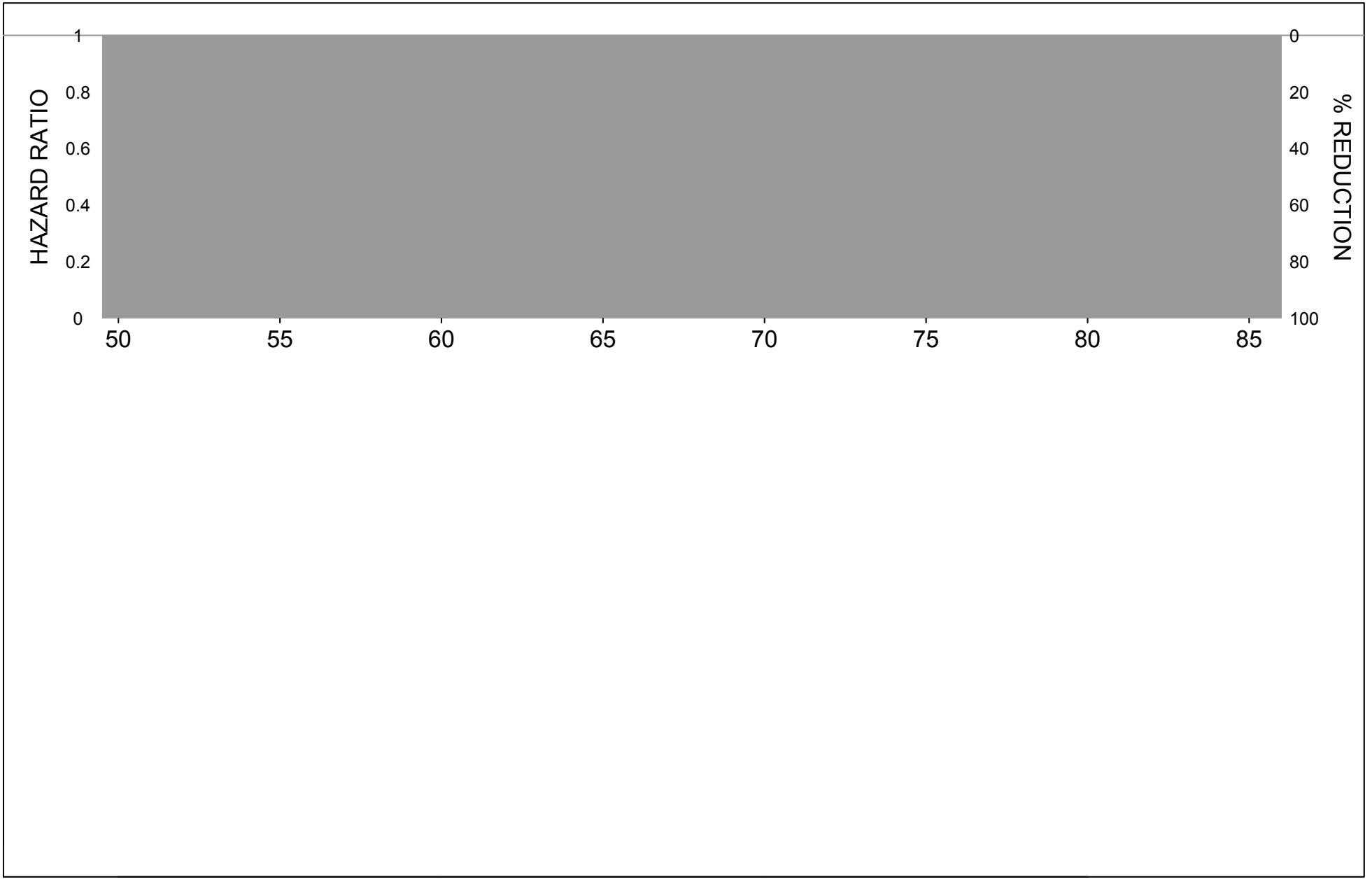


STATISTICAL MODEL

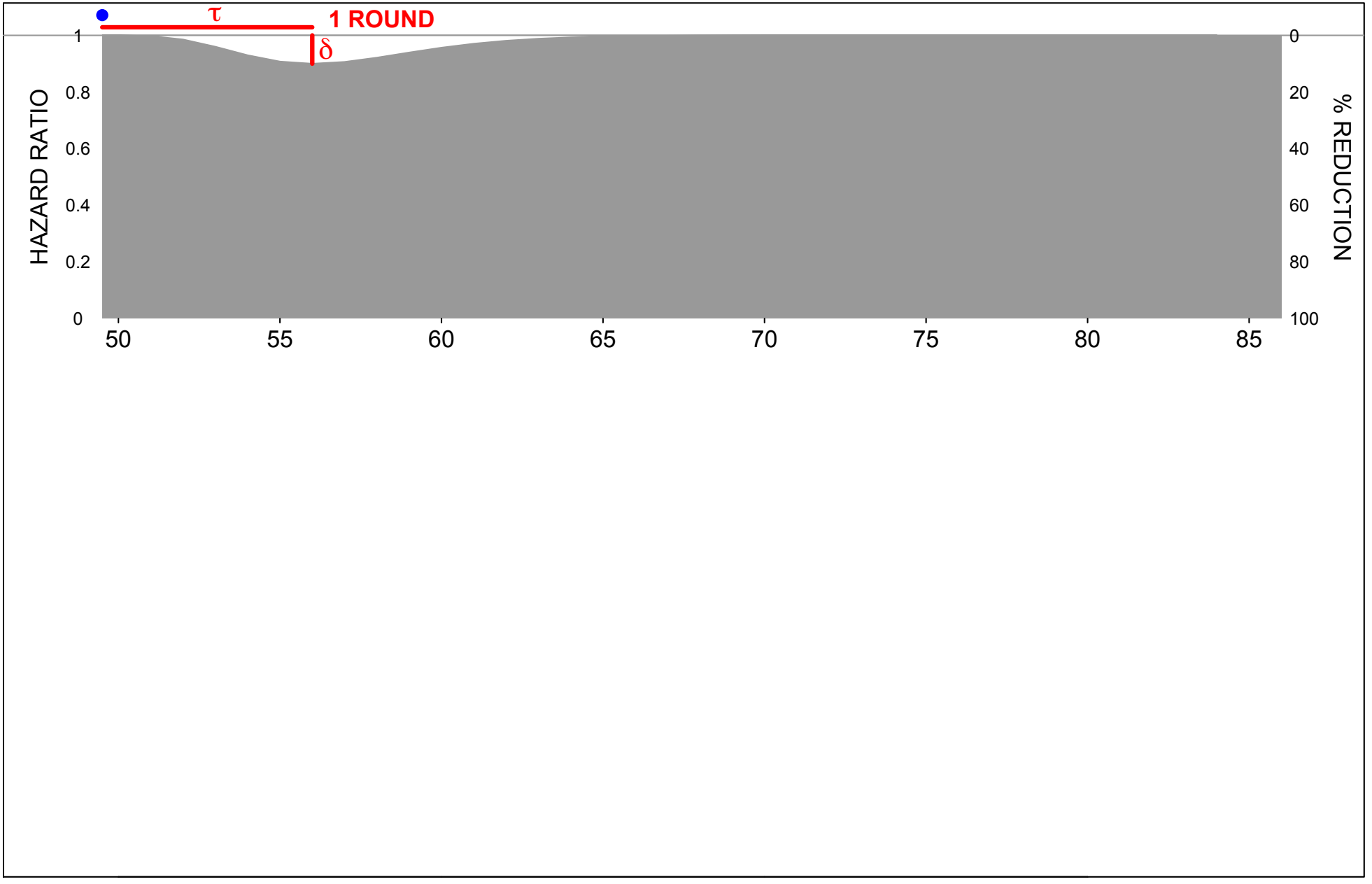
leading to

BATHTUB-SHAPED HAZARD-RATIO FUNCTION

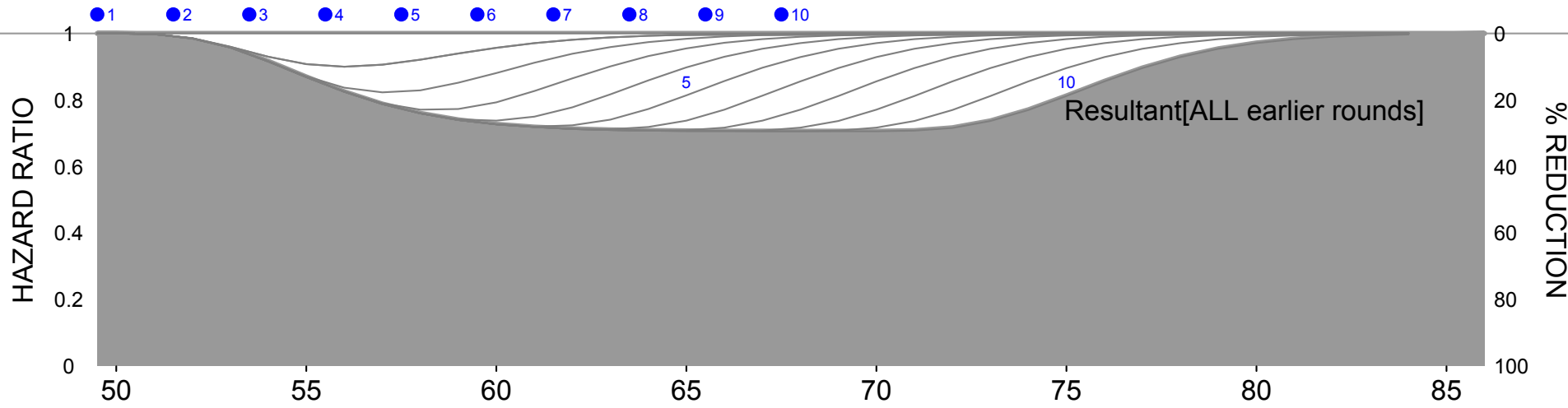
Time-pattern of mortality deficits (HRs) if NO screening: age 50 onwards



MODEL for time-pattern of mortality deficits (HRs) if • round



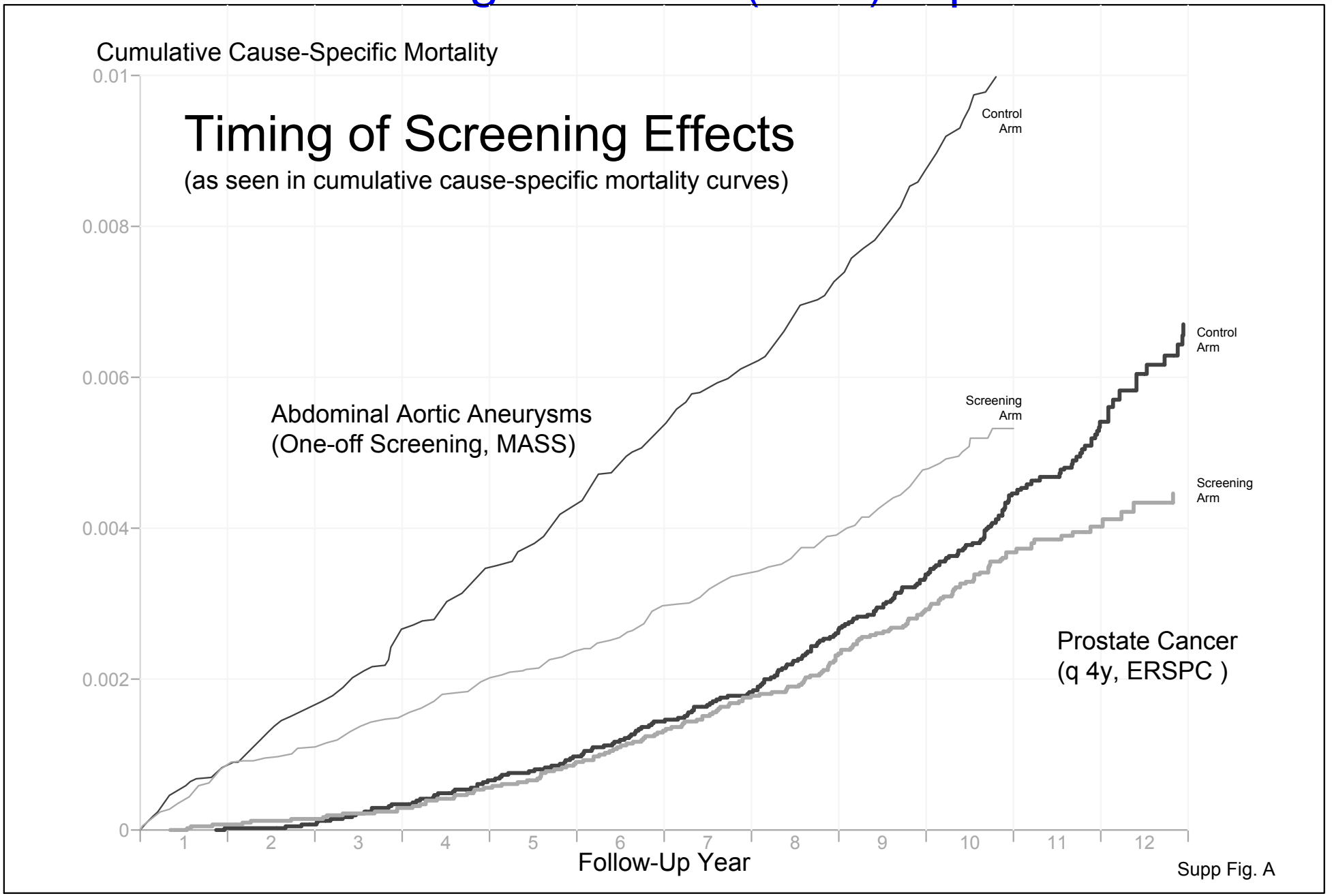
MODEL for 1-D HR pattern if ● ● ● ● ● ● ● ● ● ● rounds



Zhihui (Amy) LIU **PhD Thesis** 2104

Zhihui (Amy) LIU, James A. Hanley , Olli Saarela and Nandini Dendukuri
A Conditional Approach to Measure Mortality Reductions Due to Cancer Screening
International Statistical Review (2015), 0, 0, 1?18 doi:10.1111/insr.12088

Loneliness of Long-Distance (non-)Experimentalist



Supp Fig. A

OUR PROGRAM / MY HOME PAGE / FUNDING

http://www.biostat.mcgill.ca/hanley

or Google "James Hanley McGill"

Economic and Social Research Institute (Ireland)
1969

Natural Sciences and Engineering Research Council
1984-2012

Canadian Institutes of Health Research
2011-2018