The first ‘nested case-control’ study
{and the first conditional logistic regression}

James Hanley
Department of Epidemiology, Biostatistics and Occupational Health
McGill University, Montréal, Québec, Canada

McGill Biostatistics Seminar Series
Nov 27, 2024

Biometrika (2024), **0**, 0, *pp.* 1–8
https://doi.org/10.1093/biomet/asae038
Advance Access publication 8 August 2024

Studies in the history of probability and statistics, LI: the first conditional logistic regression

BY J. A. HANLEY

Department of Epidemiology, Biostatistics and Occupational Health, McGill University, 2001 McGill College Avenue, Montréal, Québec H3A 1G1, Canada
james.hanley@mcgill.ca

SUMMARY

Statisticians and epidemiologists generally cite the publications of Prentice & Breslow (1978) and Breslow et al. (1978) as the first description and use of conditional logistic regression, while economists cite the book chapter by Nobel laureate McFadden (McFadden, 1973). We describe the until-now-unrecognized use of, and way of fitting, this model in 1934 by Lionel Penrose and Ronald Fisher.

Some key words: Birth order; Down's syndrome; Estimating equation; Family-based selection; Maternal age; Peer review; Relative odds; Standard error.

OUTLINE

CONDITIONAL LOGISTIC REGRESSION &
THE NESTED CASE-CONTROL STUDY DESIGN:

- Modern Non-Epi Example of Conditional logistic Regression (for Orientation)
  &
  Annotated Epi. Examples of Nested CC studies [ONLINE]
- 1970s: **The** Etiologic Study Comes of Age
- 1934: Penrose (& Fisher) – Overlooked Until Now [ONLINE]

Why I am telling this story ...

Mix of epidemiology | statistics | computing

An opportunity to reflect on 90 years of

- growth in statistical methods & computing
- understanding of **the** etiologic study
- role of McGill epidemiologists and biostatisticians

Modern, non-Epi, Example (for Orientation)

1 Applying discrete choice models to predict Academy Award winners

J. R. Statist. Soc. A (2008)
171, *Part* 2, *pp.* 375–394

Iain Pardoe

His Oscar Predictions http://iainpardoe.com/oscars/

University of Oregon, Eugene, USA

and Dean K. Simonton

University of California at Davis, USA

[Received September 2005. Revised June 2007]

Summary. Every year since 1928, the Academy of Motion Picture Arts and Sciences has recognized outstanding achievement in film with their prestigious Academy Award, or Oscar. Before the winners in various categories are announced, there is intense media and public interest in predicting who will come away from the awards ceremony with an Oscar statuette. There are no end of theories about which nominees are most likely to win, yet despite this there continue to be major surprises when the winners are announced. The paper frames the question of predicting the four major awards—picture, director, actor in a leading role and actress in a leading role—as a discrete choice problem. It is then possible to predict the winners in these four categories with a reasonable degree of success. The analysis also reveals which past results might be considered truly surprising—nominees with low estimated probability of winning who have overcome nominees who were strongly favoured to win.

Keywords: Bayesian; Conditional logit; Films; Forecasting; Mixed logit; Motion pictures; Movies; Multinomial logit

Reference to (Nobel Laureate) McFadden, D. (1974) Conditional logit analysis of qualitative choice behavior.

Table 4: Explanatory variables for Best Actress in a Leading Role

1. Indicator for Best Picture Oscar nomination [1939–2004]. Only 25 actresses have won the Best Actress in a Leading Role Oscar for a movie that did not receive a Best Picture nomination (most recently, Charlize Theron for Monster in 2003).

2. Natural logarithm of the number of previous Best Actress in a Leading Role Oscar wins [1938–2004]. 24 percent of Best Actress Oscar nominees with no previous lead actress wins have won the Oscar, whereas 13 percent of Best Actress Oscar nominees with one or more previous lead actress wins have won. This variable has been log-transformed because it is highly skewed.

3. Indicator for winning a Golden Globe for Best Actress in a Leading Role (Drama) [1944–2004]. Of the 62 Best Actress Oscar winners from 1943 to 2004, 31 had won a Golden Globe for Best Actress (Drama) a few weeks earlier.

4. Indicator for winning a Golden Globe for Best Actress in a Leading Role (Musical or Comedy) [1952–2004]. Of the 55 Best Actress Oscar winners from 1950 to 2004, 11 had won a Golden Globe for Best Actress (Musical or Comedy) a few weeks earlier.

5. Indicator for winning a Screen Actor's Guild award [1996–2004]. Of the 11 Best Actress Oscar winners since 1994, eight had already won a SAG award.

Chance 2005, vol 18(4), 32-39

Why not regular (unconditional) logistic regression?

- Data are organized by competition & year ('set')
- There's a winner in each competition [~~indep. Bernoulli r.v.s~~]
- Some elements of profile did not exist in earlier years

Data, (relative & scaled-to-sum-to-1, modelled) Win Probabilities, LogLikelihood Contributions

		Profile				Rel. Prob	Prob. Win	Winner?	LogLik
Year	Nominee	X_1	X_2	...	X_K	$e^{\mathbf{X}\beta}$	(P)	(Y)	$(Y \log P)$
2024	Nominee$_1$	✓	✓	✓	✓	ω_1	$\omega_1 / \sum \omega$	0	-
2024	Nominee$_2$	✓	✓	✓	✓	ω_2	$\omega_2 / \sum \omega$	0	-
2024	Nominee$_3$	✓	✓	✓	✓	ω_3	$\omega_3 / \sum \omega$	0	-
2024	Nominee$_4$	✓	✓	✓	✓	ω_4	$\omega_4 / \sum \omega$	1	$\log P_4$
2024	Nominee$_5$	✓	✓	✓	✓	ω_5	$\omega_5 / \sum \omega$	0	-
						$\sum \omega$	1		
2023	Nominee$_1$	✓	✓	✓	✓			0	-
2023	Nominee$_2$	✓	✓	✓	✓	etc	etc	1	$\log P_2$
2023	Nominee$_3$	✓	✓	✓	✓			0	-
				etc					
1938	Nominee$_1$	✓	✓		✓			0	-
1938	Nominee $_2$	✓	✓		✓	etc	etc	0	-
1938	Nominee$_3$	✓	✓		✓			1	$\log P_3$
1938	Nominee$_4$	✓	✓		✓			0	-

DATA in Black

$\sum$ *LogLik*

FITTING in Red

$\Downarrow$

$\hat{\beta} = argmax_{\beta} \sum LogLik$

$\Downarrow$

Predictions for 2025: compute $e^{X\hat{\beta}}$ for each 2025 nominee, and rescale to P's

Modern Epidemiological Examples

ONLINE SLIDES & LYRICS

2. (Transient) Exposures and Risk of Acute Events [self-matched]

- Association between Cellular-Telephone Calls and Motor Vehicle Collisions [*NEJM 1997*]
- A Case-Crossover Study of Sleep and Work Hours and the Risk of Road Traffic Accidents [*Sleep 2010*]
- Association between high ambient temperature and acute work-related injury: a case-crossover analysis [*Scand J Work, Env & Health 2017*]
- Effects of cold temperature and snowfall on stroke mortality: A case crossover analysis [*Environment International 2019*]
- Snowfall, Temperature, and the Risk of Death From Myocardial Infarction: A Case-Crossover Study [*AJE 2020*]
- Ambient heat and risks of emergency department visits among adults in the United States: time stratified case crossover study [*BMJ 2021*]

Nature 1953

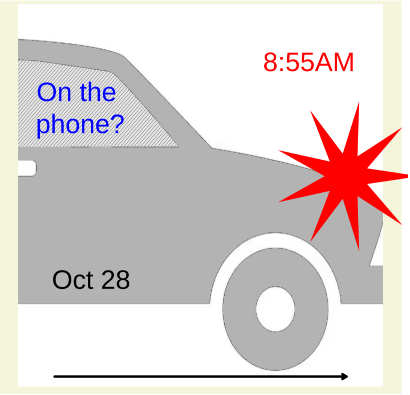
- Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid
- A Structure for Deoxyribose Nucleic Acid: **an X-ray diffraction study**

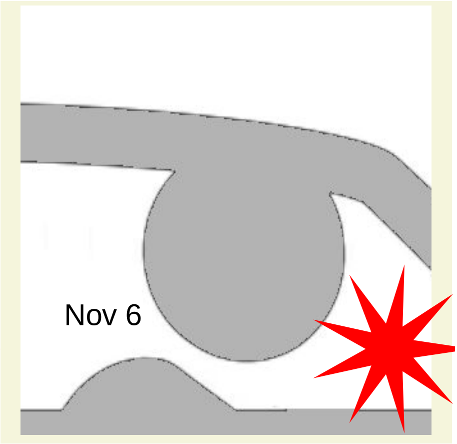
Cellular-Telephone Calls and Motor Vehicle Collisions, July 1994 - Aug 1995

≈ 6,000 Drivers who came to the North York Collision Reporting Centre, Toronto during peak hours (10AM-6PM Monday to Friday) after having been in a collision with substantial property damage (but no injury)

- ≈ 1000 (1/6!) owned a cell phone; some 699 agreed to have billing records examined.

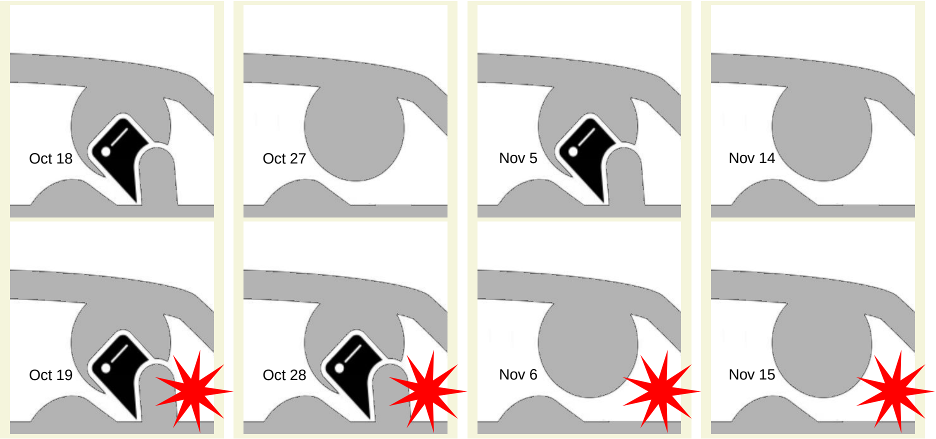
- focus (here): use of cell phone in 10 minutes before collision

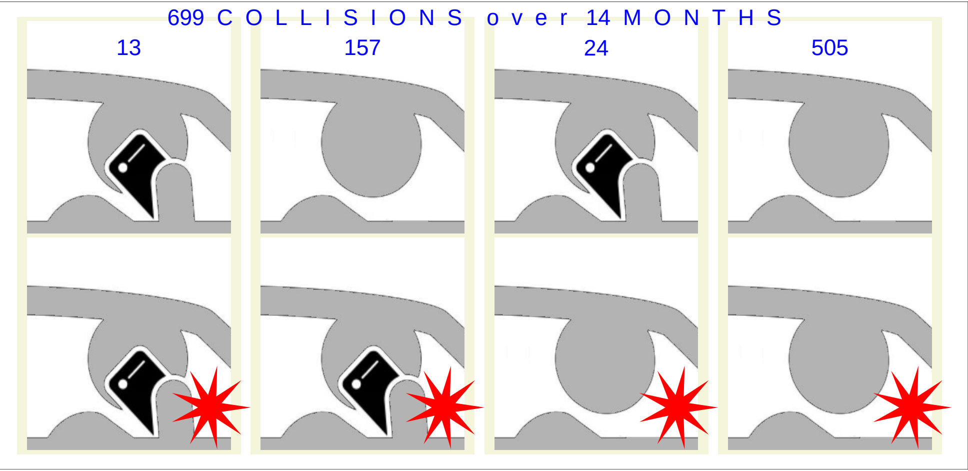




170 collisions
(24% of 699)

529 collisions


Oct 18
Oct 27
Nov 5
Nov 14
Oct 19
Oct 28
Nov 6
Nov 15


699 C O L L I S I O N S o v e r 14 M O N T H S
13
157
24
505

From “Ambient heat and risks of emergency department visits...”

To estimate the association between county specific daily maximum temperature centile and all cause and cause specific Emergency Department visits for May to September 2010-19, we used a **time stratified case crossover design**.

In this study design, **participants serve as their own control**; inference is based on the comparison of daily ambient temperatures on the case day versus daily ambient temperatures on control days***.

[...] **case day** was defined as the admission **date of each visit**; **control days** were selected at **same year and month** as case day to control for seasonal and long term time trends. They were the **other days in the same month and day of week as the case day**.

This [self- and county-matched] design has the advantage of controlling for potential confounding by all known and unknown individual and county level covariates that do not vary day to day; including, for example, age, sex, race, socioeconomic status, and population density, and behavior risk factors, such as smoking.

*** JH: Isn’t this an ‘un-modern’ way of viewing etiological studies? Why not study and **compare visit rates** at various temperatures?

Mini-example with 10 events (tornadoes)

Column header: Day of Week & Month when tornado occurred (first one: 3rd Thursday in May)
*Number in **bold***: **Temperature** (°C) on day it occurred (13.5°C the day the first one occurred)
Other numbers in same column: (°C) for other 3 or 4 days in same month, and same day of week.

Thu May (1)	Sun Jun (2)	Sat Jun (3)	Tue Jul (4)	Wed Jul (5*)	Mon Jul (6)	Sun Aug (7)	Tue Aug (8)	Thu Sep (9)	Fri Sep (10)
15.0	25.0	28.0	26.5	26.0	26.5	29.0	20.5	**24.0**	20.0
23.0	18.5	20.0	24.0	24.0	27.5	**26.0**	23.5	19.5	**26.5**
13.5	**30.0**	29.0	**28.5**	**30.0**	26.5	20.0	23.5	15.0	16.0
20.0	21.5	25.5	26.0	28.0	21.5	29.5	**29.0**	20.0	22.5
20.5		**22.5**	22.5		**32.0**				

- We start by identifying each tornado **instance** (‘case’)
- **Then** assemble the full ’set’ of possible (candidate) days on which it could have occured. (Usual to **match** on the day of the week when, (e.g., in traffic fatalities) risks, and the triggers being studied, vary by day of week.)
- The variate(s) in probability model can be multi-dimensional (e.g., Temperature, Humidity) and lagged (can use history)
- **Given** that it happened on one of those candidate days, why did it happen **on the day it did**? We find the parameter value(s) that maximize(s) overall logLik.
- Each set has **same structure as in Oscar dataset**, but (for feasibility and economy reasons) is assembled **after** the fact.

3. Unintended Effects of Medications

- Prescription of antidepressants and the risk of road traffic crash in the elderly: a case-crossover study [Br J Clin Pharmacol 2013]

- Concurrent Use of Benzodiazepines and Antidepressants and the Risk of Motor Vehicle Accident in Older Drivers: A Nested Case-Control Study [Neurology and Therapy 2015]

- Testosterone treatment and risk of venous thromboembolism: population based case-control study [BMJ 2016]

- Menopausal Hormone Therapy Formulation and Breast Cancer Risk [Obstetrics & Gynecology 2022]

From “Menopausal Hormone Therapy Formulation and Breast Cancer Risk...”

Once an instance of a new diagnosis of breast cancer was identified within the Clinical Practice Research Datalink, we **matched** the woman with 10 others [forming a **‘riskset’ of size 11**]

The 10 were **randomly selected** from the list of women who

- were (still) registered within the Datalink on the date of the diagnosis
- had no history of breast cancer
- were **born within 1 year** of the woman
- had been **registered for the same duration (**$\pm$ **1 year)**

We estimated the odds ratio (OR) of breast cancer associated with any menopausal hormone therapy exposure, then to the different estrogens and progestins using **conditional multivariate logistic regression, adjusted for the baseline covariates***
.

SAME STRUCTURE ! ; Confounder-control: combination of **matching** & **modelling**

*including obesity, smoking status (ever or never), alcohol consumption (heavy drinker, social drinker or abstainer), and medical history of endometrial cancer, hysterectomy, oophorectomy, oral contraceptive use and family history of breast cancer

Another nested case control study from same base, and same research group

We used risk-set sampling to select appropriate controls.

Each AD case was matched to up to 40 AD-free controls randomly selected from the risk set defined by the case (those still being followed and event-free at the date of the AD event).

Given the use of risk-set sampling, the ORs derived from our nested case-control analysis calculated via conditional logistic regression could be interpreted as unbiased estimators of the hazard ratios derived from the underlying cohort analysis calculated via Cox regression with minimal loss in precision.

1970s: THE ETIOLOGIC STUDY comes of age

The sophisticated use and understanding of case-control studies is the most outstanding methodologic development of modern epidemiology

(Rothman 1986,p. 62, quoted by Breslow 1996)

Timeline

'Case−Control' Studies		Statistical Developments
		Textbook: The Design and Analysis of Cohort Studies
	1980	Textbook: The Analysis of Case−Control Studies
		Conditional logistic regression for matched c−c studies
		Cox's likelihood for reduced risksets (nested c−c study)
Abortion & secondary infertility, 2 indiv.controls:case		Incidence Density (NO 'rare disease' assumption)
Asbestos & lung cancer; 4:1 'nested' c−c study		
		'Synthetic Retrospective Study': reduces effort/computations
		Regression models and life tables (risksets)
Maternal stilbestrol therapy & tumor appearance in daughters		
Age at first birth & breast cancer risk	1970	Textbook: Analysis of Binary Data
Abortion & ectopic pregnancy 4 indiv. matched controls:case		(ML) Estimation of Rel. Risk from Individually Matched Series
		Individual matching with multiple controls, 0/1 exposure
		Estimation of prob[event] as fn. of several indep. variables
US Surgeon General's Report on Smoking & Health		
		Risk of Coronary Heart Disease as logistic function
	1960	
Psych. Aspects of Rheumatoid Arthritis (nearest sib control)		Tabular methods for confounder control
Tonsillectomy and Risk of Poliomyelitis (family controls)		Regression models for binary data (logit)
ABO Blood Groups in Duodenal Ulcer−A Study of Sibships		Blood Gp. & peptic ulcer; Summary Incidence Ratio, NO Odds
Relation b/w ABO Blood Groups and cancer, peptic ulcer		
		Exposure OR = disease OR (if 'rare disease')
Smoking & lung cancer	1950	
Epidemiology of (CNS) Congenital Malformations		
	1940	
Tobacco smoking & lung cancer		
		Conditioning (2x2 table) => Odds Ratio (OR) parameter
Epidemiology of Down's Syndrome		
Pipe smoking & oral cancer	1930	
Why did THOSE five persons perish when bridge collapsed?		
Reproductive history & breast cancer		
	1920	

Table 1. *Calculation of combined estimate of incidence ratio of peptic ulcer in groups O and A*

'Numerator' Series　　'Denominator' Series

City	Peptic ulcer		~~Control~~		$x=\frac{hK}{Hk}$	$y=\log_e x$	$w=\frac{1}{\boxed{\frac{1}{h}+\frac{1}{k}+\frac{1}{H}+\frac{1}{K}}}$	$wy^2=\chi^2$
	Group O (h)	Group A (k)	Group O (H)	Group A (K)				
London	911	579	4578	4219	1·4500	0·3716	304·9	42·11
Manchester	361	246	4532	3775	1·2224	0·2008	136·6	5·50
Newcastle	396	219	6598	5261	1·4418	0·3659	134·5	18·01
					$\Sigma wy=189{\cdot}94$		576·0	65·62

Today:
1/a + 1/b + 1/c + 1/d

$Y=\Sigma wy/\Sigma w=0{\cdot}3289.$
$Y^2\Sigma w=62{\cdot}63.$
S.D. of $Y=(\Sigma w)^{-\frac{1}{2}}=0{\cdot}0417.$
95 % fiducial limits of $Y=0{\cdot}2472–0{\cdot}4106.$
$X=\text{antilog } Y=1{\cdot}39.$
95 % fiducial limits of $X=1{\cdot}28–1{\cdot}51.$

χ^2 analysis

	D.F.	
Y	1	62·63
Heterogeneity	2	2·99
Total	3	65·62

[Human Genetics **1955** :] **Directly "work with (i.e. contrast) incidence rates. The data usually do not permit calculation of absolute rates, nor are they needed.** What is wanted and readily obtained is an estimate of the ratio of one **rate** to another: the incidence in the [population time in the index category] will be $\frac{h}{H \times \textit{some constant}}$, and that in the reference category will be $\frac{k}{K \times \textit{the SAME constant}}$. An estimate of the **[incidence] ratio** will be $\frac{hK}{Hk}$, and it may readily be shown that this is the ML estimate."
Note use of lower/upper case for **(entirely separate)** numerator & denominator series.

Evolution of Epidemiologic Ideas

1987

Annotated Readings on Concepts and Methods

Sander Greenland, Editor

Table of Contents

Illustrative computations for chi square and for summary measures of relative risk (R) relating to the association of epidermoid and undifferentiated pulmonary carcinoma

in women with smoking history

Group		Epidermoid-undifferentiated pulmonary carcinoma			Controls			Cases and controls		
		1 + Pack cigarettes daily	Nonsmokers	Total	1 + Pack cigarettes daily	Nonsmokers	Total	1 + Pack cigarettes daily	Nonsmokers	Total
		A (1)	B (2)	N_1 (3)	C (4)	D (5)	N_2 (6)	M_1 (7)	M_2 (8)	T (9)
Housewives	under age 45	0	2	2	0	7	7	0	9	9
	45–54	2	5	7	1	24	25	3	29	32
	55–64	3	6	9	0	49	49	3	55	58
	65 and over	0	11	11	0	42	42	0	53	53
White-collar workers	under age 45	3	0	3	2	6	8	5	6	11
	45–54	2	2	4	2	18	20	4	20	24
	55–64	2	4	6	2	23	25	4	27	31
	65 and over	0	6	6	1	11	12	1	17	18
Other occupations	under age 45	1	0	1	3	10	13	4	10	14
	45–54	4	1	5	1	12	13	5	13	18
	55–64	0	6	6	1	19	20	1	25	26
	65 and over	1	3	4	0	15	15	1	18	19
Total		18	46	64	13	236	249	31	282	313

Cases Controls

1959

Derivative computations

$\frac{AD}{T}$ $\frac{(1)(5)}{(9)}$ (10)	$\frac{BC}{T}$ $\frac{(2)(4)}{(9)}$ (11)	E(A) $\frac{(3)(7)}{(9)}$ (12)	E(D) $\frac{(6)(8)}{(9)}$ (13)	V(A) $\frac{(12)(13)}{(9)-1.0}$ (14)	$\frac{N_1C}{N_2}$ $\frac{(3)(4)}{(6)}$ (15)	$\frac{N_1D}{N_2}$ $\frac{(3)(5)}{(6)}$ (16)	$\frac{N_2A}{N_1}$ $\frac{(1)(6)}{(3)}$ (17)	$\frac{N_2B}{N_1}$ $\frac{(2)(6)}{(3)}$ (18)
0	0	0	7.000	0	0	2.000	0	7.000
1.500	0.156	0.656	22.656	0.480	0.280	6.720	7.143	17.857
2.534	0	0.466	46.466	0.380	0	9.000	16.333	32.667
0	0	0	42.000	0	0	11.000	0	42.000
1.636	0	1.364	4.364	0.595	0.750	2.250	8.000	0
1.500	0.167	0.667	16.667	0.483	0.400	3.600	10.000	10.000
1.484	0.258	0.774	21.774	0.562	0.480	5.520	8.333	16.667
0	0.333	0.333	11.333	0.222	0.500	5.500	0	12.000
0.714	0	0.286	9.286	0.204	0.231	.769	13.000	0
2.667	0.056	1.389	9.389	0.767	0.385	4.615	10.400	2.600
0	0.231	0.231	19.231	0.178	0.300	5.700	0	20.000
0.790	0	0.211	14.211	0.166	0	4.000	3.750	11.250
12.825	1.201	6.375	224.375	4.036	3.325	60.675	76.960	172.040

Chi-square: $X^2 = (|\text{discrepancy}| - 0.5)^2/\Sigma V(A) = (|Y| - 0.5)^2/\Sigma(14) = 30.66$

Relative risk: $R = \Sigma(AD/T)/\Sigma(BC/T) = \Sigma(10)/\Sigma(11) = \boxed{10.68}$

[crude relative risk, $r = \Sigma A \Sigma D/\Sigma B \Sigma C = \Sigma(1)\Sigma(5)/\Sigma(2)\Sigma(4) = \boxed{7.10}$

3. THE MANTEL–HAENSZEL ERA

Epidemiologists who have done case-control studies during the past 20 years . . . have stood on the shoulders of giants. And, lest we epidemiologists lose sight of one major root of our discipline, we should remember that all of these men are, or were, statisticians (Cole 1979, p. 15).

The statisticians to whom Cole refers are Cornfield and Dorn and their colleagues Mantel and Haenszel, who in 1959 published their landmark paper in the *Journal of the National Cancer Institute*. This paper clarified the relationship between case-control (or retrospective) and cohort (forward or prospective) studies with the observation that "*a primary goal is to reach the same conclusions in a retrospective study as would have been obtained from a forward study, if one had been done*" (Mantel and Haenszel 1959, p. 733). Anticipating the development of the nested case-control study (see Sec. 5), Mantel and Haenszel suggested that one might adopt the case-control approach even to the sampling of subjects already ascertained in a cohort study, to collect additional data items. Clearly, the only *conceptual* difference between cohort and case-control studies was that the latter involved sampling from the cohort rather than complete enumeration of it.

Breslow, 1996
Fisher Lecture,
Statistics in Epidemiology: The Case-Control Study

$\widehat{\textbf{PT}}$ in lieu of **PT** Denominators

Rates* in Exposed (E) vs. Unexposed ($\overline{E}$) Population-Time (PT)

$* \ \frac{\textit{Numbers of Cases } (C)}{\textit{Amount of PT}}$

$$\frac{C_E}{PT_E} \ / \ \frac{C_{\overline{E}}}{PT_{\overline{E}}}$$ Entire base

$$\frac{C_E}{\widehat{PT}_E} \ / \ \frac{C_{\overline{E}}}{\widehat{PT}_{\overline{E}}}$$ ‘Fair sample of base’

2 ‘**Series**’:

Case *or ‘Numerator’*:

$\rightarrow C_E \ \& \ C_{\overline{E}}$

~~***Control***~~ ‘Denominator’ :

$\rightarrow \widehat{PT}_E : \widehat{PT}_{\overline{E}}$

JH Eur. J. Epi 2018

1969

INDIVIDUAL MATCHING WITH MULTIPLE CONTROLS IN THE CASE OF ALL-OR-NONE ~~RESPONSES~~ EXPOSURES

OLLI S. MIETTINEN

Departments of Epidemiology and Biostatistics, Harvard School of Public Health, and Cardiology Division, Department of Medicine, Children's Hospital Medical Center, Boston, Massachusetts, U. S. A.

SUMMARY

The one-to-one individual matching principle of the matched pairs design is generalized to R-to-one individual matching in the case of all-or-none responses and fixed sample size procedures. A test is given; its asymptotic power function is derived; the selection of the matching ratio (R) is considered in relation to the unit costs in the two comparison groups; and finally, procedures for sample size determination are described.

1. INTRODUCTION

Matching is a common feature in the design of nonexperimental studies concerned with the evaluation of causal propositions (such as hypotheses on disease etiology). Its main purpose typically is the attainment of validity for the inferences, but it has implications for design efficiency as well.

As nonexperimental studies with matched comparison series are frequently quite expensive, it is important to understand the properties of matching designs so as to be able to make the best use of them. The matched pairs design in the case of all-or-none responses and fixed sample size has recently been studied rather extensively (Worcester [1964], Billewicz [1964, 1965], Miettinen [1966, 1968a, b], Bennett [1967], Chase [1968]). The present paper deals with the extension of this design to the case where the number of control subjects obtained for each propositus is not necessarily one but some general number R. We will use the term 'R-to-one individual matching design.' This generalization and an intelligent choice of R are important whenever several control subjects can be obtained at a unit cost substantially lower than that of the propositi.

Previous history of induced abortion in propositi with ectopic pregnancy and matched controls. Trichopoulos *et al.*

Index number	History of induced abortion				
	"Case" Propositus	Control number			
		1	2	3	4
1	−	−	−	−	−
2	+	−	+	−	−
3	+	−	−	−	−
4	−	−	−	−	−
5	−	+	−	−	−
6	+	−	−	−	−
7	+	−	−	−	−
8	−	−	−	−	−
9	+	+	−	−	+
10	+	−	+	−	−
11	+	−	+	+	−
12	−	−	−	−	−
13	+	+	+	+	+
14	+	−	−	+	−
15	+	−	−	+	−
16	+	+	−	−	−
17	−	−	−	−	−
18	+	+	−	−	+

1970

ESTIMATION OF RELATIVE RISK FROM INDIVIDUALLY MATCHED SERIES

OLLI S. MIETTINEN

Departments of Epidemiology and Biostatistics, Harvard School of Public Health, and the Cardiology Division of the Department of Medicine, Children's Hospital Medical Center, Boston, Mass., U.S.A.

SUMMARY

Point and interval estimation of relative risk is investigated for the purpose of case-control studies of disease etiology with individual matching of cases and controls. It is assumed that the disease is rare and that the relative risk bears no relation to the matching factors. The resulting maximum likelihood estimate is expressed in a closed form up to the case of two-to-one matching, while with 3 or more controls for each case a simple iterative procedure of obtaining the estimate is presented. Results for exact and approximate interval estimation are also derived.

1. INTRODUCTION

Ever since its introduction in a classical paper by Cornfield [1951], relative risk has been the focal point of the statistical analysis of data from case-control (retrospective) studies of disease etiology, and considerable attention has been given to problems encountered in its estimation (Cornfield [1951; 1956], Woolf [1955], Haldane [1956], Cox [1958], Mantel and Haenszel [1959], Cornfield and Haenszel [1960], Berger [1961], Gart [1962a, b], Goodman [1963; 1964], etc). Results have thus become available for setting confidence limits for the relative risk on the basis of studies involving independent series of cases and controls, for testing homogeneity of several relative risks, and

D. R. Cox

Regression Models and Life-Tables

JRSS B 1972

Summary

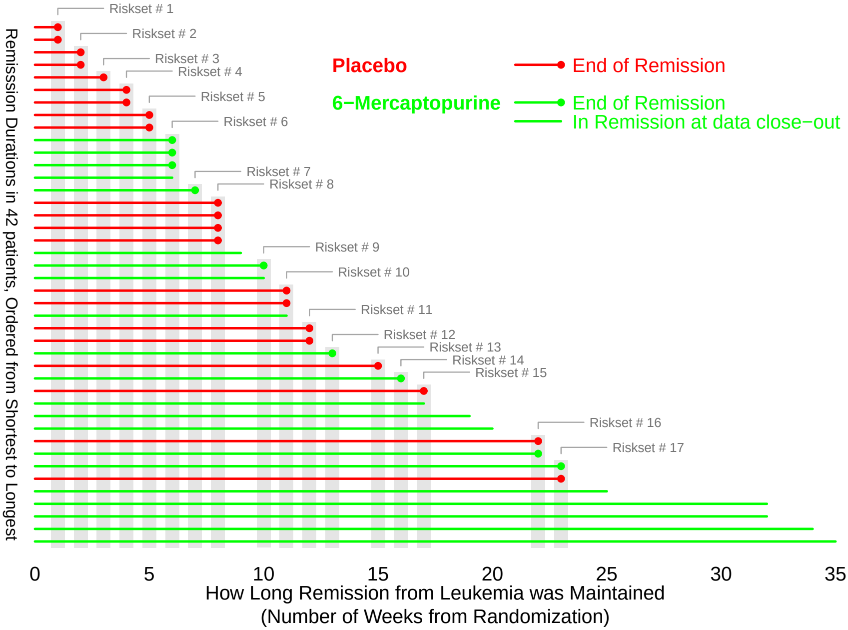
The analysis of censored failure times is considered. It is assumed that on each individual are available values of one or more explanatory variables. The hazard function (age-specific failure rate) is taken to be a function of the explanatory variables and unknown regression coefficients multiplied by an arbitrary and unknown function of time. A conditional likelihood is obtained, leading to inferences about the unknown regression coefficients. Some generalizations are outlined.

3. Regression Models

Suppose now that on each individual one or more further measurements are available, say on variables $z_1, \ldots, z_p$. We deal first with the notationally simpler case when the failure-times are continuously distributed and the possibility of ties can be ignored. For the *j*th individual let the values of $\mathbf{z}$ be $\mathbf{z}_j = (z_{1j}, \ldots, z_{pj})$. The z's may be functions of time. The main problem considered in this paper is that of assessing the relation between the distribution of failure time and $\mathbf{z}$. This will be done in terms of a model in which the hazard is

$$\lambda(t; \mathbf{z}) = \exp(\mathbf{z}\boldsymbol{\beta})\, \lambda_0(t), \tag{9}$$

where $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown parameters and $\lambda_0(t)$ is an unknown function giving the hazard function for the standard set of conditions $\mathbf{z} = \mathbf{0}$. In fact $(\mathbf{z}\boldsymbol{\beta})$ can be replaced by any known function $h(\mathbf{z}, \boldsymbol{\beta})$, but this extra generality is not needed at this stage. The following examples illustrate just a few possibilities.



Data from Freireich and Gehan (1963), used by Gehan(1965) and Cox(1972)

For the particular failure at time $t_{(i)}$, conditionally on the risk set $\mathscr{R}(t_{(i)})$, the probability that the failure is on the individual as observed is

$$\exp\{\mathbf{z}_{(i)}\boldsymbol{\beta}\}\Big/\sum_{l\in\mathscr{R}(t_{(i)})}\exp\{\mathbf{z}_{(l)}\boldsymbol{\beta}\}. \tag{12}$$

Each failure contributes a factor of this nature and hence the required conditional log likelihood is

$$L(\boldsymbol{\beta})=\sum_{i=1}^{k}\mathbf{z}_{(i)}\boldsymbol{\beta}-\sum_{i=1}^{k}\log\left[\sum_{l\in\mathscr{R}(t_{(i)})}\exp\{\mathbf{z}_{(l)}\boldsymbol{\beta}\}\right]. \tag{13}$$

Direct calculation from (13) gives for $\xi,\eta=1,\ldots,p$

$$U_{\xi}(\boldsymbol{\beta})=\frac{\partial L(\boldsymbol{\beta})}{\partial\beta_{\xi}}=\sum_{i=1}^{k}\{z_{(\xi i)}-A_{(\xi i)}(\boldsymbol{\beta})\}, \tag{14}$$

where

$$A_{(\xi i)}(\boldsymbol{\beta})=\frac{\sum z_{\xi l}\exp(\mathbf{z}_{l}\boldsymbol{\beta})}{\sum\exp(\mathbf{z}_{l}\boldsymbol{\beta})}, \tag{15}$$

the sum being over $l\in\mathscr{R}(t_{(i)})$. That is, $A_{(\xi i)}(\boldsymbol{\beta})$ is the average of z_{ξ} over the finite population $\mathscr{R}(t_{(i)})$, using an "exponentially weighted" form of sampling. Similarly

$$\mathscr{I}_{\xi\eta}(\boldsymbol{\beta})=-\frac{\partial^{2}L(\boldsymbol{\beta})}{\partial\beta_{\xi}\,\partial\beta_{\eta}}=\sum_{i=1}^{k}C_{(\xi\eta i)}(\boldsymbol{\beta}), \tag{16}$$

where

$$C_{(\xi\eta i)}(\boldsymbol{\beta})=\{\sum z_{\xi l}z_{\eta l}\exp(\mathbf{z}_{l}\boldsymbol{\beta})/\sum\exp(\mathbf{z}_{l}\boldsymbol{\beta})\}-A_{(\xi i)}(\boldsymbol{\beta})\,A_{(\eta i)}(\boldsymbol{\beta}) \tag{17}$$

is the covariance of z_{ξ} and z_{η} in this form of weighted sampling.

SYNTHETIC RETROSPECTIVE STUDIES AND RELATED TOPICS

1973

NATHAN MANTEL

Biometry Branch, National Cancer Institute, Bethesda, Maryland 20014, U.S.A.

SUMMARY

Prospective and retrospective approaches for estimating the influence of several variables on the occurrence of disease are discussed. The assumptions under which these approaches would tend to yield the same estimates as would be given by an ideal but unattainable experimental design approach are stated. It is then brought out that in a large prospective study in which comparatively few cases of disease have occurred, computational problems can be so burdensome as to preclude a comprehensive and imaginative analysis of the data. The prospective study can be converted into a synthetic retrospective study by selecting a random sample of the cases and a random sample of the noncases, the sampling proportion being small for noncases, but essentially unity for cases. It is demonstrated that such sampling will tend to leave the dependence of the log odds on the variables unaffected except for an additive constant.

THE SYNTHETIC RETROSPECTIVE STUDY—SAMPLING FROM A PROSPECTIVE STUDY

My present purpose is to propose, discuss, and validate use of retrospective approach procedures in a prospective study situation. A particular prospective-study situation which I encountered gave rise to only 165 cases of a particular condition in a cohort of about 4,000 individuals.[1] Preliminary analyses were undertaken using a limited number of the variables on which data had been collected. But even with simple maximum likelihood analyses of the form used involving only one or two of the study variables, the computer time required was somewhat prolonged. This could then preclude making analyses as comprehensive and as extensive as we should have liked.

As it turned out, the key cause for prolonged computer time was the large number of observations involved. Computation was simple and rapid once the necessary totals were obtained for all 4,000 individuals. But time was consumed for entering all the information and for computing at each iterative stage certain quantities appropriate for each individual. The number of iterative cycles for convergence could be reduced by a device for obtaining suitable entering approximations (see below), but even this would not resolve our problem.

A possible remedy envisaged was to convert the study, in principle, to a retrospective one. Suppose we included in the analysis a random proportion, π_1, of our cases and another random proportion, π_2, of the negatives. If we chose π_1 as 1 and π_2 as 0.15, we would have all the cases and 3.5 negatives per case. By the reasoning that $n_1 n_2/(n_1 + n_2)$ measures the relative information in a comparison of two averages based on sample sizes of n_1 and n_2 respectively, we might expect by analogy, which would of course not be exact in the present case, that this approach would result in only a moderate loss of information. (The practicing statistician is generally aware of this kind of thing. There is little to be gained by letting the size of the control group, n_2, become arbitrarily large if the size of the experimental group, n_1, must remain fixed.) But the reduction in computer time would permit much more effective analyses. Ostensibly we would be meeting the additional conditions assumed for validity of the retrospective study approach; that is the retained individuals would be a random sample of the cases and disease-free individuals arising in the prospective study.

Suppose the randomness conditions are met. Still it seems that we are selecting our retained individuals on the basis of their response variable,

---

[1] The actual number of individuals was substantially less than 4,000. An initial cohort of about 1,350 men was studied to evaluate the short-term prognostic value of various factors in coronary heart disease. Men remaining free of disease for two years could be reentered into the analysis for the next two years using their new X_i values.

$\boxed{\text{Info} = 1/\text{var} = 1/(1/n_1 + 1/n_2) = n_1 n_2 / (n_1 + n_2)}$

1971

Mortality in the Chrysotile Asbestos Mines and Mills of Quebec

J. Corbett McDonald, MD; Alison D. McDonald, MD; Graham W. Gibbs, MSc; Jack Siemiatycki; and Charles E. Rossiter, MA, Montreal

Of 11,788 persons born between 1891 and 1920 employed in the Quebec asbestos mining industry, 88.4% were traced. Of these 2,457 (23.6%) had died. Exposure indexes for each worker were calculated from job dust levels and duration of employment. The overall mortality was lower than expected for the population of Quebec but in the highest dust category, comprising 5% of the cohort, the age-standardized rate was 20% higher than in the other groups. Respiratory, cardiovascular, and malignant disease in equal proportions accounted for the excess. There were 101 deaths from respiratory cancer including three from malignant mesothelioma, an estimated excess of about 15 deaths. The difference in rates for respiratory cancer between those maximally and minimally exposed was fivefold and, though perhaps exaggerated, was apparently determined by accumulated dust exposure and duration of employment.

Arch Environ Health—Vol 22, June 1971

Submitted for publication Aug 17, 1970; accepted Nov 10.

From the Department of Epidemiology and Health, McGill University, Montreal. Mr. Rossiter is presently with the Medical Research Council Pneumoconiosis Unit, Penarth, South Wales.

Reprint requests to 3775 University St, Montreal 112 (Dr. J. C. McDonald).

THE REMARKABLE qualities of the asbestos group of fibrous minerals have been recognized since antiquity, but mining and milling on an industrial scale began only at the end of the 19th century. In the Eastern Townships region of Quebec, deposits of chrysotile asbestos in serpentine rock were noted in the 1847 Canadian Geological Survey. The first mine was opened at Thetford in 1878, and within 30 years the region was producing most of the world's asbestos. The proportion fell as Russian, South African, and Italian mines came into operation, but Quebec still produces about 40% of the world's supply, now estimated at about 4 million tons a year.[1]

There are two main mining areas, one at Thetford Mines and neighboring towns of Black Lake and Broughton, and the other at Asbestos. The Thetford area was developed by many different companies, but with amalgamation the number has now been reduced to six. At Asbestos, the mining has been carried out since 1882 by one large company which also operates a small factory in the town for the manufacture of mixed asbestos products. There is a small mine owned by another company a few miles away.

Arch Environ Health/Vol 28, Feb 1974

Scientific Communications

The results of studies of respiratory symptoms and function, roentgenographic changes, and mortality* in relation to dust exposure in the Quebec chrysotile industry, which has employed some 28,000 workers, are brought together and their implications for control examined.

* in 9,692 men who had worked for one month or more, and who were born 1891 to 1920.

Submitted for publication July 2, 1973; accepted July 30.

From the Department of Epidemiology and Health, McGill University, Montreal. Mr. Rossiter is presently with the Medical Research Council Pneumoconiosis Unit, Penarth, South Wales.

Reprint requests to Department of Epidemiology and Health, McGill University, 3775 University St, Montreal 112, Quebec, Canada (Dr. McDonald).

The Health of Chrysotile Asbestos Mine and Mill Workers of Quebec

J. Corbett McDonald, MD; Margaret R. Becklake, MD; Graham W. Gibbs, PhD; Alison D. McDonald, MD; Charles E. Rossiter, MA, Montreal

Meantime, another method of analysis (G. Eyssen, MSc, and F. D. K. Liddell, MA, unpublished data) has been employed that, we believe, eliminates certain of these problems, though it does not make full use of the data available and provides only estimates of relative rather than absolute risk. For this analysis, the dust exposures of the 134 men included in the 1969 analysis of respiratory cancer mortality were compared with a sample of men, four for each case, selected at random among persons living at the time of the death of the respiratory cancer case and born in the same year.

McDonald et al. Arch Env. Health 1974

Table 3.—Age-Corrected Death Rates Per 1,000 Men Born 1891-1920: Deaths to December 1969

Cause	Dust Exposure Level, mpcf-yr					
	<10	10	100	200	400	800+
All causes	365	355	354	313	323	395
Respiratory cancers*	10.3	13.1	13.4	15.5	21.4	32.1
Abdominal cancers	18.0	13.6	18.7	11.6	26.3	28.7
Pneumoconiosis	1.6	1.5	0.8	4.9	4.9	23.6

Includes malignant pleural mesothelioma.

Table 4.—Relative Risk of Death From Respiratory Cancer in Men by Dust Group, Estimated From Retrospective Analysis

Dust Exposure Level, mpcf-yr	No. of Cases	No. of Controls	Relative Risk
< 10	32	186	1.0
10	41	188	1.3
100	13	39	1.9
200	14	61	1.3
400	15	40	2.2
800+	19	22	5.0
Total	134	536	

The distribution of cases and controls by dust exposure category is presented in Table 4. It can be seen that the pattern of relative risk obtained by this approach is quite similar to that for respiratory cancer mortality shown in Table 3.

J. R. Statist. Soc. A,
(1977), **140**, *Part* 4, pp. 469–491

Methods of Cohort Analysis: Appraisal by Application to Asbestos Mining

By F. D. K. Liddell, J. C. McDonald† and D. C. Thomas

McGill University

[Read before the Royal Statistical Society on Wednesday, June 22nd, 1977, the President, Miss Stella V. Cunliffe, in the Chair]

Summary

Longitudinal studies of occupational mortality have usually been analysed *a priori*: the cohort is subdivided in terms of potential stimuli and comparisons made between sub-cohorts in their patterns of mortality. The alternative *a posteriori* argument compares the dead with the living, searching for differences in the potential stimuli. We selected the following methods for appraisal: (a) comparative composite cohort analysis (Case and Lea, 1955), against external and internal standards; (b) the use of a fixed number of controls for each death (following Miettinen, 1969); and (c) that of Cox (1972) based on regression models. Method (a) argues *a priori*, the others *a posteriori*. These three methods have been applied to a large cohort study of mortality in the Quebec chrysotile asbestos-producing industry, focusing on lung cancer. The methods agreed in demonstrating a clear direct relationship, which may well be linear, between excess lung cancer mortality and total dust exposure. Method (a), with an external standard, is useful for placing the cohort in demographic context. In method (b), only three or four controls should suffice for each case, leading to possibilities of improved quality of data. Similar advantages might be achieved for method (c) through some sampling of the living, but it would remain more complex; while it facilitates the study of interactions and, without sampling, can provide absolute risks, it was very expensive.

6. Applications of *A Posteriori* Reasoning to Lung Cancer Mortality in the Quebec Cohort

6.1. *Case and Fixed Multiple Controls*

The Miettinen approach was evaluated by considering five controls for each of the 215 lung cancer deaths.

The selection of controls was strictly at random from among men born in the same year as the case and known to have survived at least into the year following that in which the case died.

That there were only 1,290 men in this study made it possible to re-examine all smoking history questionnaires which failed validity checks or otherwise aroused doubts. As a result, codes were changed for 122 men (nearly 10 per cent of 1,290), although altering the classification for only 39 men (3 per cent). However, the opportunity was taken to reclassify those who had given up smoking, according to the report, into those who had been *ex-smokers* for at least seven years when the case died, and *recent smokers*.

Liddell, McDonald, Thomas JRSS A 1977

MIETTINEN, O. S. (1969). Individual matching with multiple controls in the case of all-or-none response. *Biometrics*, **25**, 339–355.
NELDER, J. A. (1975). GLIM (Generalized Linear Interactive Modelling Program). *Appl. Statist.*, **24**, 259–261.
OLDHAM, P. D. and ROSSITER, C. E. (1965). Mortality in coal worker's pneumoconiosis related to lung function: A prospective study. *Brit. J. Industr. Med.*, **22**, 92–100.
THOMAS, D. C. (1976). Analysis of longitudinal studies with interval-censored response times. Ph.D. Thesis, McGill University.
—— (1977). Applications of methods of response–time analysis to long-term epidemiological studies. (In preparation.)
URY, H. K. (1975). Efficiency of case-control studies with multiple controls per case: continuous or dichotomous data. *Biometrics*, **31**, 643–649.

ADDENDUM

By D. C. THOMAS

As most of the computing cost for the Cox method was in the calculations for the living, doing them on only a sample of each risk set can yield considerable savings. In the study of Section 6.1, the five controls are a random sample of the risk set for their case. The obvious way to reconstruct Cox's likelihood is to divide the contributions of the controls by their sampling fractions. However, an alternative generalization results from ignoring the sampling fraction, to obtain:

$$L^* = \prod_{i=1}^{n} [\exp(\beta z_{i0}) / \sum_{k=0}^{K_i} \exp(\beta z_{ik})],$$

where subscript i indexes the case/control sets and 0 and k represent cases and controls respectively. This is the conditional likelihood that the particular subjects $i0$ are the cases, given that one of each set of K_i+1 subjects is a case.

Repeating the analyses of Section 6.2 using L^* led to similar results, neither method producing systematically larger χ^2 statistics. The cost was about one twentieth that of the full cohort analyses.

Statistics in Epidemiology: The Case-Control Study, N. E. Breslow (1996). JASA

5. MATCHING AND NESTING

I was introduced to the case-control design in 1972 during collaborative work at IARC on a study of esophageal cancer among Singapore Chinese (DeJong, Breslow, Hong, Sridharan, and Shanmugaratnam 1974). This was a typical hospital-based interview study, with two control groups, that focused on ethnicity, diet, alcohol, and tobacco as possible risk factors. Of particular interest were questions relating to the temperature at which various beverages were consumed. We were well aware that differential "recall bias" in the interview responses of cases and controls was a strong possibility for this item. D. R. Cox's (1970) text covering logistic regression had recently appeared. Having had some previous experience with this methodology in a clinical setting (Breslow and McCann 1971), I jumped at the chance to apply the technique to the case-control study. In retrospect, the enthusiasm seems rather naive, because we simply ignored the apparent problems posed by the outcome-dependent sampling.

One aspect of our analysis that did bother me was its failure to account for the pair matching of controls to cases on age, gender, hospital ward (for one of the control groups), and time of diagnosis. Such matching was widely used to select "comparable" controls, but there was little appreciation among epidemiologists for the complexities that it introduced for rigorous statistical analysis. Special procedures for matched case-control designs with binary exposures were available (Miettinen 1970; Pike and Morrow 1970), but a general treatment was lacking. The problem occupied my attention on my return to Seattle in 1974 and, with the help of colleagues and students, we developed a solution based on stratified logistic regression (Breslow, Day, Halvorsen, Prentice, and Sabai 1978).

Suppose that the population at risk is so finely stratified that each case occupies a single stratum, and that the matched controls are drawn from the same stratum as the case. With S denoting the stratum, the population model is

$$\Pr(D=1|S=j,\mathbf{X}=\mathbf{x}) = \frac{\exp(\alpha_j+\mathbf{x}\boldsymbol{\beta})}{1+\exp(\alpha_j+\mathbf{x}\boldsymbol{\beta})}.$$

This involves a separate parameter for each matched set and allows inclusion of possible interactions between exposures and matching variables among the explanatory variables $\mathbf{x}$. Following Fisherian principles, the stratum parameters α_j are eliminated by conditioning on an appropriate ancillary statistic, in this case the unordered set of exposures for the case and controls in each stratum. Thus the conditional likelihood that the exposures $\mathbf{x}_{j0}$ are those of the case and $(\mathbf{x}_{j1},\ldots,\mathbf{x}_{jM})$ are those of the M controls in stratum j, as observed, given the set of $M+1$ exposures, is proportional to [...]

Writing in the usual fashion, the marginal probabilities drop out, and we are left with

$$\prod_{j=1}^{J} \frac{\exp(\mathbf{x}_{j0}\boldsymbol{\beta})}{\exp(\mathbf{x}_{j0}\boldsymbol{\beta})+\sum_{m=1}^{M}\exp(\mathbf{x}_{jm}\boldsymbol{\beta})} \tag{10}$$

for inference about $\boldsymbol{\beta}$. Note that the terms $\exp(\mathbf{x}_{jm}\boldsymbol{\beta})$ are the relative risks for each subject relative to someone with a standard ($\mathbf{X}=\mathbf{0}$) set of exposures. These arguments are easily generalized to situations with a variable number of controls per case, and even to matched sets with an arbitrary number of cases and controls (Breslow et al. 1978). The conditional likelihood (10) also arises from the stratified logistic regression model for a cohort study, by conditioning on the number of cases that occur in each stratum. This further strengthens the notion that one is estimating the same parameters in cohort studies and case-control studies.

INTERNATIONAL AGENCY FOR RESEARCH ON CANCER

STATISTICAL METHODS IN CANCER RESEARCH

VOLUME 1 - The analysis of case-control studies

BY

1980

N.E. BRESLOW & N.E. DAY

CONTENTS

APPENDIX III:

MATCHED DATA FROM THE LOS ANGELES STUDY OF ENDOMETRIAL CANCER USED FOR ILLUSTRATION IN CHAPTERS 5 AND 7

CASE OR CONTROL	AGE	GALL BLADDER DISEASE	HYPER TENSION	OBESITY	ESTROGEN (ANY) USE	CONJUGATED DOSE (SEE CODE)	ESTROGEN DURATION (MONTHS)	NON ESTROGEN DRUG
CASE	74	NO	NO	YES	YES	3	96+	YES
CONTROL	75	NO	NO	UNK	NO	0	0	NO
CONTROL	74	NO	NO	UNK	NO	0	0	NO
CONTROL	74	NO	NO	UNK	NO	0	0	NO
CONTROL	75	NO	NO	YES	YES	1	48	YES
CASE	67	NO	NO	NO	YES	3	96+	YES
CONTROL	67	NO	NO	NO	YES	3	5	NO
CONTROL	67	NO	YES	YES	NO	0	0	YES
CONTROL	67	NO	NO	NO	YES	2	53	NO
CONTROL	68	NO	NO	NO	YES	2	45	YES

- etc

MAIN PROGRAM

(READS DATA, SETS UP RISK VARIABLES, CALLS SUBROUTINE)

```
C      MASTER MAIN
       DIMENSION NR(63),IVAR(6),W(6),EZB(6),
      * Z(6,5,63),B(6),SCORE(6),COV(21),COVI(21)
C
C      DIMENSION IVAR(NM),NCA(NS),NCT(NS),IP(NMAX2),W(NM),WW(NM),
C     *Z(NM,NRMAX,NS),B(NM),SCORE(NM),COV(NM1),COVI(NM1)
C      SEE SUBROUTINE FOR DEFINITIONS
C
       DATA NR/63*5/,B/6*0.0/,IVAR/1,2,3,4,5,6/,NP/6/
       DATA NM,NRMAX,NIT,EPS/6,5,10,0.0001/
       NM1=NM*(NM+1)/2
       I-0
       READ(1,100) GALL,OB,EST
 1000  CONTINUE
 100   FORMAT(10X,F5.0,5X,F5.0,15X,F5.0)
C
C    I IS THE ORDER NUMBER OF THE SET
C
       I=I+1
       K=0
       DO 6 KK-1,5
       K=K+1
       Z(1,K,I)=2-GALL
       IF (OB.EQ.9)  OB=2
       Z(2,K,I)=2-OB
       Z(3,K,I)=2-EST
       Z(4,K,I)=(2-GALL)*(2-EST)
       Z(5,K,I)=(2-OB)*(2-EST)
       Z(6,K,I)=(2-GALL)*(2-OB)
       READ(1,100,END=2000) GALL,OB,EST
 6     CONTINUE
       GOTO 1000
 2000  CONTINUE
       NS = I
       DO 2 I=1,NP
       B(I)=0.0
 2     CALL MATCH(NS,NR,NRMAX,NM,I,NIT,B,Z,SCORE,COVI,COV,
      * EZB,IVAR,NM1,EPS,W)
       STOP
       END
```


SUBPROGRAM MATCH

```
C      REFERENCES:
C      N. E. BRESLOW,N.E. DAY,K. T. HALVORSEN, R. L. PRENTICE,C. SABAI :
C        ESTIMATION OF MULTIPLE RELATIVE RISK FUNCTIONS IN MATCHED
C        CASE-CONTROL STUDIES.  AMERICAN JOURNAL OF EPIDEMIOLOGY
C        VOL108,NO4, P 299-307 , 1978
C
C        THIS SUBROUTINE COMPUTES A LINEAR LOGISTIC REGRESSION ANALYSIS FOR
C        MATCHED SETS OF 1  CASE A VARIABLE NO. OF CONTROLS PER CASE
C        THE VARIABLES APPEARING IN THE CALL STATEMENT ARE DEFINED AS FOLLO
C        NS  NUMBER OF MATCHED SETS
C        NR  VECTOR OF NO. OF CONTROLS IN EACH SET + 1
C        NRMAX  (MAX NO. OF CONTROLS PER CASE)+1
C        NM  MAXIMUM NUMBER OF VARIABLES TO BE ANALYZED
C        NP  NUMBER OF VARIABLES ANALYZED IN THIS RUN
C        NIT   MAXIMUM NUMBER OF ITERATIONS OF THE NEWTON-RAPHSON TYPE
C        B  PARAMETER VECTOR OF LENGTH NM
C        Z  NM BY NRMAX BY NS MATRIX CONTAINING COVARIATES
C        SCORE FIRST DERIVATIVE OF THE LN-LIKELIHOOD OF LENGTH NM
C        COVI INFORMATION MATRIX(2ND DERIVATIVE OF LN-LIKELIHOOD)
C        COV INVERSE INFORMATION MATRIX(ESTIMATED COVARIANCE MATRIX)
C        EZB WORKING VECTOR OF LENGTH NRMAX
C        IVAR VECTOR OF VARIABLES USED IN THIS RUN (DIMENSION NM)
C        NM1 = NM*(NM+1)/2 DIMENSION OF COVI AND COV
C        EPS   CHANGE IN LIKELIHOOD BELOW WHICH ITERATION STOPS
C        W  WORKING VECTOR OF LENGTH NM
C
C        NOTE(Z,J,K,I) IS THE VALUE OF THE JTH COVARIATE FOR THE KTH
C        MEMBER IN THE ITH SET.IT IS ASSUMED THAT THE FIRST MEMBER IS THE
C        CASE AND THAT THE REMAINING NR(I)-1 MEMBERS ARE CONTROLS.
C
C        NOTE(Z MUST BE DIMENSIONED TO HAVE NM ROWS,NRMAX COLUMNS
C        AND AT LEAST NS SLICES IN THE MAIN PROGRAM,
C        COVI AND COV ARE ARRAYS OF LENGTH NM*(NM+1)/2 SINCE THEY USE
C        THE SYMMETRIC STORAGE MODE.
C
         SUBROUTINE MATCH(NS,NR,NRMAX,NM,NP,NIT,B,Z,SCORE,COVI,COV,
        * EZB,IVAR,NM1,EPS,W)
         DIMENSION Z(NM,NRMAX,NS),B(NM),EZB(NRMAX),SCORE(NM),COV(NM1),
        * COVI(NM1),IVAR(NM),NR(NS),W(NM)
         REAL LOGLIK
         DATA TEST/1.0/,ISUB/1/
         WRITE(6,100)
100      FORMAT(///' LOGISTIC REGRESSION ANALYSIS FOR MATCHED SETS',/)
         WRITE(6,101)NS
101      FORMAT(1H ,'NUMBER OF MATCHED SETS',I4)
         WRITE(6,102)((I,NR(I)),I=1,NS)
102      FORMAT(1H ,'SET NUMBER AND NUMBER OF MEMBERS',
        *' IN EACH SET,(INCLUDING CASE)',/,50(1X,10(I4,I3),/))
         WRITE(6,103)NP,(IVAR(J),J=1,NP)
```



```
C  UPON CONVERGENCE OF MAXIMUM ITERATIONS WRITE OUT RESULTS
      GO TO 1
9     WRITE(6,108)(B(J),J=1,NP)
108   FORMAT(' ESTIMATED PARAMETER VECTOR',(/1X,10F12.6))
      WRITE(6,109)(SCORE(J),J=1,NP)
109   FORMAT(' FIRST DERIVATIVE LOG-LIKELIHOOD',(/1X,10F12.6))
      WRITE(6,110)
110   FORMAT(' INFORMATION MATRIX')
      DO 10 J=1,NP
      K=J*(J-1)/2+1
      JJ=J*(J+1)/2
10    WRITE(6,111) (COVI(I),I=K,JJ)
111   FORMAT(1X,10F12.6)
C  INVERT INFORMATION MATRIX AND WRITE OUT
C     CALL SYMINV(COVI,NP,COV,W,NULLTY,IFAULT,NM1)
      DO 600 I=1,NM1
      COV(I)=COVI(I)
600   CONTINUE
      CALL SINV(COV,NP,EPS,IER)
      IF (IER.NE.0) WRITE (6,501)IER
      WRITE(6,112)
112   FORMAT(' ESTIMATED COVARIANCE MATRIX')
      DO 11 J=1,NP
      K=J*(J-1)/2+1
      JJ=J*(J+1)/2
      SCORE(J)=B(J)/SQRT(COV(JJ))
11    WRITE(6,111)(COV(I),I=K,JJ)
      WRITE(6,113)(SCORE(J),J=1,NP)
113   FORMAT(' STANDARDIZED REGRESSION COEFFICIENTS',(/1X,100F12.6))
      RETURN
      END
```

MODEL WITH SINGLE VARIABLE: GALL

```
LOGISTIC REGRESSION ANALYSIS IN STRATA

NUMBER OF STRATA  63
STRATUM NUMBER AND NUMBERS OF CASES AND CONTROLS
  1  1  4   2  1  4   3  1  4   4  1  4   5  1  4   6  1  4   7  1  4   8  1  4
  9  1  4  10  1  4  11  1  4  12  1  4  13  1  4  14  1  4  15  1  4  16  1  4
 17  1  4  18  1  4  19  1  4  20  1  4  21  1  4  22  1  4  23  1  4  24  1  4
 25  1  4  26  1  4  27  1  4  28  1  4  29  1  4  30  1  4  31  1  4  32  1  4
 33  1  4  34  1  4  35  1  4  36  1  4  37  1  4  38  1  4  39  1  4  40  1  4
 41  1  4  42  1  4  43  1  4  44  1  4  45  1  4  46  1  4  47  1  4  48  1  4
 49  1  4  50  1  4  51  1  4  52  1  4  53  1  4  54  1  4  55  1  4  56  1  4
 57  1  4  58  1  4  59  1  4  60  1  4  61  1  4  62  1  4  63  1  4
NUMBER OF VARIABLES IN THIS ANALYSIS  1
THESE VARIABLES ARE  1
MAXIMUM NUMBER OF ITERATIONS  10
ITER LOG-LIKELIHOOD  SCORE    PARAMETER ESTIMATES
  1     -101.3945   13.829         0.0
  2      -95.6567    0.512         1.5714
  3      -95.4042    0.000         1.3011
  4      -95.4041    0.000         1.3061
  5      -95.4041    0.000         1.3061
ESTIMATED PARAMETER VECTOR
   1.306143
FIRST DERIVATIVE LOG-LIKELIHOOD
   0.000002
INFORMATION MATRIX
   7.232529
ESTIMATED COVARIANCE MATRIX
   0.138264
STNDIZED REGRESSION COEFFICIENTS
   3.512656
```

MODEL WITH 2 VARIABLES: GALL + OB

```
NUMBER OF VARIABLES IN THIS ANALYSIS  2
THESE VARIABLES ARE  1  2
MAXIMUM NUMBER OF ITERATIONS  10
ITER LOG-LIKELIHOOD  SCORE    PARAMETER ESTIMATES
 1     -95.4041      5.031    1.3061    0.0
 2     -92.8559      0.016    1.2673    0.7044
 3     -92.8481      0.000    1.3085    0.7254
 4     -92.8483      0.000    1.3086    0.7255
 5     -92.8483      0.000    1.3086    0.7255
ESTIMATED PARAMETER VECTOR
   1.308584    0.725525
FIRST DERIVATIVE LOG-LIKELIHOOD
  -0.000006    0.000004
INFORMATION MATRIX
   7.115862
  -0.413175    9.369920
ESTIMATED COVARIANCE MATRIX
   0.140892
   0.006213    0.106998
STNDIZED REGRESSION COEFFICIENTS
   3.486253    2.218013
```

MODEL WITH 3 VARIABLES: GALL + OB + EST

(Model 4, Table 7.7)

```
NUMBER OF VARIABLES IN THIS ANALYSIS  3
THESE VARIABLES ARE  1  2  3
MAXIMUM NUMBER OF ITERATIONS  10
ITER  LOG-LIKELIHOOD  SCORE    PARAMETER ESTIMATES
 1       -92.8483    26.837   1.3086    0.7255    0.0
 2       -78.4745     1.213   1.0256    0.4851    1.6166
 3       -77.8259     0.017   1.2543    0.5085    1.9860
 4       -77.8176     0.000   1.2746    0.5113    2.0394
 5       -77.8177     0.000   1.2748    0.5113    2.0403
 6       -77.8178     0.000   1.2748    0.5113    2.0403
ESTIMATED PARAMETER VECTOR
   1.274838    0.511341    2.040295
FIRST DERIVATIVE LOG-LIKELIHOOD
  -0.000002   -0.000016   -0.000012
INFORMATION MATRIX
   5.996716
  -0.295262    7.917883
  -0.581353    0.554115    5.222651
ESTIMATED COVARIANCE MATRIX
   0.168775
   0.005016    0.127390
   0.018255   -0.012958    0.194880
STNDIZED REGRESSION COEFFICIENTS
   3.103141    1.432656    4.621776
```

MODEL WITH 4 VARIABLES: GALL + OB + EST + GALL×EST

(Model 8, Table 7.7)

```
NUMBER OF VARIABLES IN THIS ANALYSIS  4
THESE VARIABLES ARE NUMBERS  1  2  3  4
MAXIMUM NUMBER OF ITERATIONS  10
ITER  LOG-LIKELIHOOD  SCORE    PARAMETER ESTIMATES
 1       -77.8178     4.392   1.2748    0.5113    2.0403     0.0
 2       -75.9028     0.221   3.0330    0.4859    2.5096    -2.2027
 3       -75.7904     0.000   2.8467    0.4901    2.6172    -2.0003
 4       -75.7906     0.000   2.8446    0.4901    2.6206    -1.9974
 5       -75.7907     0.000   2.8446    0.4901    2.6206    -1.9974
 6       -75.7904     0.000   2.8446    0.4901    2.6206    -1.9975
 7       -75.7906     0.000   2.8446    0.4901    2.6206    -1.9974
 8       -75.7906     0.000   2.8446    0.4901    2.6206    -1.9974
ESTIMATED PARAMETER VECTOR
   2.844558    0.490128    2.620618   -1.997445
FIRST DERIVATIVE LOG-LIKELIHOOD
   0.000001   -0.000002    0.000000    0.000002
INFORMATION MATRIX
   6.692931
  -0.287092    7.667618
  -1.389091    0.432643    4.504441
   4.758771   -0.133909    0.584442    4.981022
ESTIMATED COVARIANCE MATRIX
   0.774393
  -0.003830    0.131275
   0.340365   -0.014950    0.376375
  -0.779880    0.008943   -0.369741    0.989468
TNDIZED REGRESSION COEFFICIENTS
   3.232466    1.352753    4.271628   -2.008047
```

WORLD HEALTH ORGANIZATION

INTERNATIONAL AGENCY FOR RESEARCH ON CANCER

STATISTICAL METHODS IN CANCER RESEARCH

VOLUME II – THE DESIGN AND ANALYSIS OF COHORT STUDIES

BY

N.E. BRESLOW & N.E. DAY

TECHNICAL EDITOR FOR IARC

E. HESELTINE

IARC Scientific Publications No. 82

INTERNATIONAL AGENCY FOR RESEARCH ON CANCER
LYON

1987

CONTENTS

LIST OF PARTICIPANTS AT IARC WORKSHOP
25–27 May 1983

Professor E. Bjelke
Institute of Hygiene and Social Medicine
University of Bergen
5016 Haukeland Sykehus, Norway

Professor N.E. Breslow
Department of Biostatistics, SC-32
University of Washington
Seattle, WA 98195, USA

Dr T. Hirayama
Chief, Epidemiology Division
National Cancer Center Research Institute
Tokyo, Japan

Dr B. Langholz
German Cancer Research Center
Im Neuenheimer Feld 280
6900 Heidelberg 1, Federal Republic of Germany

Dr O. Møller Jensen
Director, Danish Cancer Registry
2100 Copenhagen Ø, Denmark

Professor J. Peto
Division of Epidemiology
Institute of Cancer Research
Sutton, Surrey, UK

Dr P.G. Smith
Department of Medical Statistics
London School of Hygiene and Tropical Medicine
London WC1E 7HT, UK

Dr D.C. Thomas
Department of Family and Preventive Medicine
University of Southern California
Los Angeles, CA 90033, USA

Dr A. Whittemore
Department of Family, Community and Preventive Medicine
Stanford University School of Medicine
Stanford, CA 94305, USA

IARC participants:

Dr N.E. Day
Dr J. Estève
Dr J. Wahrendorf
Dr A.M. Walker

THE (singular) ETIOLOGIC STUDY JH, Eur J Epi 2018

ORIGINALLY

A mere 'case-control' study, which involves a group of cases of the illness in question and a comparable control group without the illness; and these groups are compared with respect to the histories of the etiologic factor under study.

Many older 'case-control' studies did not have an explicit study base.

~~Compare Cases vs. 'Controls~~
Like Woolf, we should Compare Rates in Exposed vs. Unexposed Population-Time (OSM: "We are Students of Rates")

~~Exposure Odds and their Ratio~~
~~"The baseline risk of a crash is low (< 1%) during an average day, making an odds ratio a good estimate of relative risk." [2016]~~
In 1955, Woolf did not need, or mention, the term Odds or Odds Ratio.

MODERN CONCEPT

Constructed on a defined aggregate of study population-time, constituting the base of the study. Its elements are:

(1) the suitably documented case series, constituted by the entirety of the cases (as defined) occurring in the study base;

(2) the similarly documented base series (denominator series), derived as a fair sample of the study base; and

(3) the data on these two series (of person-moments) translated into the corresponding value for the confounder-conditional rate-ratio of the occurrence of the illness in the study base, and into its associated inferential statistic(s).

The result is an incidence-density ratio, free of any 'rare-disease assumption'.

A MULTIVARIATE STATISTICAL APPROACH TO THE PROBLEM OF INFECTIONS DURING THE EARLY MONTHS OF PREGNANCY AND THEIR RELATIONSHIP TO ABORTION, STILLBIRTH, CONGENITAL MALFORMATIONS, AND NEONATAL AND INFANT MORTALITY.

Author

John A. Stewart

A thesis submitted to the Faculty of Graduate Studies and Research in partial fulfillment of the requirements for the degree of Master of Sciences

Department of Epidemiology and Health

McGill University

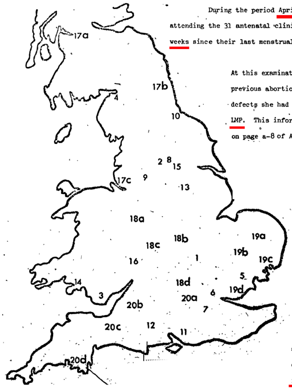
Montreal

March, 1974

ABSTRACT

For a study of the effects of infection during pregnancy on outcome, information about 11,815 pregnancies in 1959-60 had been collected in 31 clinics throughout Great Britain, and a stratified sample of 2,172 sera from these pregnant women had been tested for anti-bodies to some 20 infections. Starting in 1971, the author has analysed these data. After standardizing for age, parity, etc., it was found that, in most clinics, mothers of liveborn infants had lower titers to some infections than had mothers of deadborn infants. No marked differences in titers were found between mothers of normal infants and mothers of liveborn infants with major congenital defect(s). A reason for the weak associations may be that the only assessment of infections utilised had been from a single serum sample; thus, the infection may have been subclinical or have occurred before conception. Another possible reason lies in the low expectations of abnormal outcomes due to specific infections.

Figure (1) Participating Clinics (see Table 1 for identities)



During the period April 1959 through December 1960 all women attending the 31 antenatal clinics between the twelfth and sixteenth weeks since their last menstrual period (LMP) were admitted to the study.

At this examination each woman provided information of her age; how many previous abortions, stillbirths, livebirths, and children with congenital defects she had had; and the number and date of any infections since her LMP. This information was recorded on the Antenatal Record Card shown on page a-8 of Appendix A as M/S(1). (See page a-4 of Appendix A for the

At this same examination a blood sample of 20 ml. was taken and tested for Rh and ABO blood groups and hemoglobin content. The results were recorded on M/S(3) (see page a-11 of Appendix A). The serum from the remaining blood was inactivated by heating to 56°C for thirty minutes and then stored at -20°C by each center.

A report on the outcome of each pregnancy (abortion, stillbirth, or livebirth) was sought by the antenatal clinic, the staff of the clinic also noted if any congenital defects were present. These data were recorded on the Follow-up Record Card; see M/S(2) in Appendix A on pages

Shortly after the first birthday of each livebirth, the clinic staff recorded whether the child was alive or dead and examined living children to discover if there were any congenital defects not previously identified. This information was added to the Follow-up Record Card.

3) Serological sample

In the latter part of 1962 (after the last child born had passed his first birthday), certain sera were selected for serological examination. First, the sera were included from all mothers who had had a spontaneous abortion or stillbirth (463), or whose liveborn infant had died in the first year (228), or whose infant despite survival for at least a year had a major congenital defect (438). These totaled 1,129,

and the available resources allowed the inclusion of a like number of sera from a sample of all the other mothers. A one in ten systematic sample, selecting each mother whose serial number ended in the digit 5, provided 1,043 sera. These 2,172 selected sera were sent to a central laboratory to be tested. The 20 types of antibodies for which all sera in this sample were tested are shown in Table 2. The results of these tests were recorded on M/S(3) shown in figure (A3) of Appendix A.

figure (2) The outcomes at birth and at the assessment at one year for all pregnancies which ıtered the study.

In parentheses are the 2,172 outcomes of pregnancy included in the serological sample.

ENTRY
11,815 (2,172)

Outcome at Birth: LIVEBORN		Losses	Infant Deaths	Outcome at Assessment at one Year: Major Defect	Minor Defect	Normal
Total	10,691 (1,709)	823 (97)	228 (228)	420 (420)	1,232 (132)	7,988 (832)
Normal	10,062 (1,366)	789 (77)	151 (151)	203 (203)	931 (103)	7,988 (832)
Minor Defect	324 (38)	16 (2)	1 (1)	6 (6)	301 (29)	
Major Defect	305 (305)	18 (18)	76 (76)	211 (211)		

DEADBORN	
Total	463 (463)
Stillbirths	217 (217)
Abortions	246 (246)

Losses
661 (0)

A MULTIVARIATE STATISTICAL APPROACH TO THE PROBLEM OF INFECTIONS DURING THE EARLY MONTHS OF PREGNANCY AND THEIR RELATIONSHIP TO ABORTION, STILLBIRTH, CONGENITAL MALFORMATIONS, AND NEONATAL AND INFANT MORTALITY.

John A Stewart, MSc thesis, McGill University, 1974

The Lancet · Saturday 18 October 1980

LOW SERUM-VITAMIN-A AND SUBSEQUENT RISK OF CANCER

Preliminary Results of a Prospective Study

NICHOLAS WALD MARIANNE IDLE
JILLIAN BOREHAM

I.C.R.F. Cancer Epidemiology and Clinical Trials Unit, Radcliffe Infirmary, Oxford OX2 6HE

ALAN BAILEY

B.U.P.A. Medical Research, 300 Gray's Inn Road, London WC1X 8DU

Summary In a prospective study of about 16 000 men, serum samples were collected and stored. Vitamin-A (retinol) levels were later measured in the stored samples from the 86 men who were subsequently notified as having developed cancer and in the stored samples from 172 controls who did not develop cancer. Low retinol levels were associated with an increased risk of cancer. The association was independent of age, smoking habits, and serum-cholesterol level and was greatest for men who developed lung cancer (mean retinol level 187 i.u./dl compared with 229 i.u./dl for the controls, $p<0 \cdot 005$). The risk of cancer at any site for men with retinol levels in the lowest quintile was 2·2 times greater than the risk for men with levels in the highest quintile ($p<0 \cdot 025$). These results suggest that measures taken to increase serum-retinol levels in man may lead to a reduction in cancer risk.

Introduction

VITAMIN-A deficiency causes cell dedifferentiation and

Retrospective studies of serum-vitamin-A, measured as retinol, have demonstrated lower levels in patients with cancer than in controls without cancer, although the low levels may have been a result of the cancer rather than a precursor.[10,11]

To investigate whether vitamin A was related to future incidence of cancer, and lung cancer in particular, we made a prospective study of serum-retinol in men attending a medical screening centre in London.

Methods

The study population consisted of about 16 000 men aged 35–64 years who attended the B.U.P.A. medical centre in London for a comprehensive health-screening examination between March, 1975, and December, 1978. The men were asked about their medical history and their smoking habits. A physical examination was carried out, together with lung-function tests, an electrocardiogram, a chest X-ray, and a series of blood tests. The blood was also used to provide serum, which was stored at −40°C. If, on the basis of any information collected, there were any grounds for suspecting the presence of cancer this was noted. The N.H.S. records of the men were flagged and, through the assistance of the Office of Population Censuses and Surveys, notification was received in the event of cancer or death. By the end of 1979, 86 men were identified who had developed cancer (subjects). 172 control men who were alive and without cancer were selected from the remainder of the study population. Controls were chosen to be of similar age (within 5 years) and similar smoking habits as the subjects to take account of any indirect association between retinol and cancer which might have arisen if age and smoking were related to retinol as well as to cancer. Controls were also matched with subjects for the date blood was taken (within 4 months).

American Journal of Epidemiology

[1] Department of Occupational Health, Faculty of Medicine, McGill University, Montreal, Quebec, Canada.

Cancer Risks Associated with Occupational Exposure to Magnetic Fields among Electric Utility Workers in Ontario and Quebec, Canada, and France: 1970–1989

G. Thériault,[1] M. Goldberg,[2] A. B. Miller,[3] B. Armstrong,[1] P. Guénel,[2] J. Deadman,[1] E. Imbernon,[4] T. To,[3] A. Chevalier,[4] D. Cyr,[1] and C. Wall[3]

To determine whether occupational exposure to magnetic fields of 50–60 Hz was associated with cancer among electric utility workers, the authors used a case-control design nested within three cohorts of workers at electric utilities: Électricité de France–Gaz de France, 170,000 men; Ontario Hydro, 31,543 men; and Hydro-Québec, 21,749 men. During the observation period, 1970–1989, 4,151 new cases of cancer occurred. Each participant's cumulative exposure to magnetic fields was estimated based on measurements of current exposure of 2,066 workers performing tasks similar to those in the cohorts using personal dosimetry. Estimates were also made of past exposure based on knowledge of current loading, work practices, and usage.

6,106

Controls were other employees from the same utility matched to the case on year of birth. For each case, a "risk set" was generated comprising all study participants who were born in the same year, were alive, and were members of the cohort at the date of diagnosis of the case. Controls were selected at random from these "risk sets." A man selected as a control could become a case later on in the study.

For cancers defined a priori as of special interest, the case-control ratio was 1:4; for the other cancers, the ratio was 1:1.

To respect the matched design of the study and allow for adjustment for possible confounding factors, we estimated odds ratios and their 95 percent confidence intervals by conditional logistic regression (16) using the EPICURE program

Hi Jean-François (Boivin) and Samy

I have modified the title (and focus) of the talk, so it deals more with the 1970s than the 1930s!

And even though the focus is the case-control study, I will also add in another first (I think): Jean-François' strategy when his RA was waiting, with time on her hands, to extract radiotherapy details for the cases of secondary cancer, and the controls he went ahead and anticipated.. and sampled from the cohort so she could have work to do and not let the budget run out!!

I am always impressed that we don't need 'special names' for designs.. just good smart common sense. And I'm not sure that the names we have given them are the best ones we could have come up with. Court-Brown and Doll (1957), and Smith and Doll (1962) didn't gave their sampling design a name either.

Jean-François : I will be looking at Pubmed today to find a nice excerpt and example from your **case-cohort** work ... but happy as well if you want to give me any of the backstory (besides what I say ?remember above)

Jim

Dear Jim:

Very nice to hear from you.

The first time I came across this peculiar design was while I was a student at Harvard under the mentorship of George Hutchison. I had decided to read all his publications, and I came across his cohort study of radiotherapy and leukemia (JNCI 1968; attached). In this paper, Hutchison refer to a sample of 10% of the entire cohort to be used to estimate expected numbers of cancers. I could not make sense of this design as I thought that the 10% sampling applied to both the numerator and the denominator – what was the point then ? Hutchison seemed surprised by my question, and he explained that the 10% sampling applied only to the denominator. He did not make a big deal of this approach, and it certainly did not occur to him that his design should receive a special name. When I returned to that design later and then named it (Wacholder, Boivin AJE 1987), Hutchison's interest was keener.

I used that (case-cohort) design in a Cancer paper (1992; attached). I had some difficulty publishing it because the peer-reviewers lectured me about the appropriate procedures for the selection of controls in case-control studies. You will see in the Methods section of my paper that I had to respond to such comments.

I hope to attend your seminar next week. It is nice to see that you are still pursuing your historical research.

Best regards.

Jean-François Boivin, md, ScD , Médecin-conseil
Institut national d'excellence en santé et en services sociaux (INESSS) Québec

1934: ? first conditional logistic regression

ONLINE

https://jhanley.biostat.mcgill.ca/Penrose/

Down syndrome

87 languages

Article Talk | Read View source View history Tools

From Wikipedia, the free encyclopedia

Down syndrome or **Down's syndrome**, also known as **trisomy 21**, is a genetic disorder caused by the presence of all or part of a third copy of chromosome 21.[3] It is usually associated with developmental delays, mild to moderate intellectual disability, and characteristic physical features.[1][12] There are three types of Down syndrome, all with the same features: Trisomy 21, the most common type; Mosaic Down syndrome, and Translocation Down syndrome.[13][14]

The parents of the affected individual are usually genetically normal.[15] The probability increases from less than 0.1% in 20-year-old mothers to 3% in those of age 45.[4] The extra chromosome is provided at conception as the egg and sperm combine.[16] A very small percentage of 1-2% gets the additional chromosome in the embryo stage and it only impacts some of the cells in the body; this is known as Mosaic Down syndrome.[17][18] Usually, babies get 23 chromosomes from each parent for a total of 46, whereas in Down syndrome, a third 21st chromosome is attached.[18] It is believed to occur by chance, with no known behavioral activity or environmental factor that changes the probability.[2] Down syndrome can be identified during pregnancy by prenatal screening, followed by diagnostic testing, or after birth by direct observation and genetic testing.[6] Since the introduction of screening, Down syndrome pregnancies are often aborted (rates varying from 50 to 85% depending on maternal age, gestational age, and maternal race/ethnicity).[19][20][21]

There is no cure for Down syndrome.[22] Education and proper care have been shown to provide good quality of life.[7] Some children with Down syndrome are educated in typical school classes, while others require more specialized education.[8] Some individuals with Down syndrome graduate from high school, and a few attend post-secondary education.[23] In adulthood, about 20% in the United States do paid work in some capacity,[24] with many requiring a sheltered work environment.[8] Support in financial and legal matters is often needed.[10] Life expectancy is around 50 to 60 years in the developed world, with proper health care.[9][10] Regular screening for health issues common in Down syndrome is recommended throughout the person's life.[9]

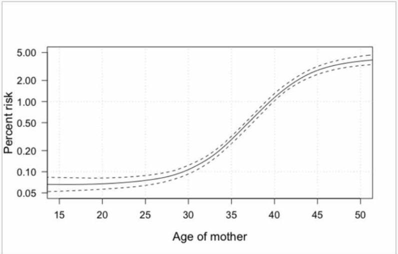
Down syndrome	
Other names	Down's syndrome, Down's, trisomy 21

An eight-year-old boy displaying characteristic facial features of Down syndrome

Specialty	Medical genetics, pediatrics

Epidemiology

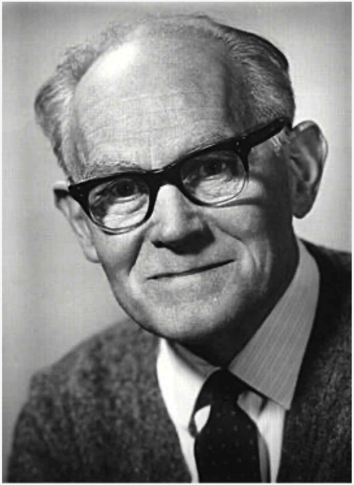
Down syndrome is the most common chromosomal abnormality in humans.[9] Globally, as of 2010, Down syndrome occurs in about 1 per 1,000 births[1] and results in about 17,000 deaths.[132] More children are born with Down syndrome in countries where abortion is not allowed and in countries where pregnancy more commonly occurs at a later age.[1] About 1.4 per 1,000 live births in the United States[133] and 1.1 per 1,000 live births in Norway are affected.[9] In the 1950s, in the United States, it occurred in 2 per 1,000 live births with the decrease since then due to prenatal screening and abortions.[92] The number of pregnancies with Down syndrome is more than two times greater with many spontaneously aborting.[10] It is the cause of 8% of all congenital disorders.[1]



The risk of having a Down syndrome pregnancy in relation to a mother's age[4]

Maternal age affects the chances of having a pregnancy with Down syndrome.[4] At age 20, the chance is 1 in 1,441; at age 30, it is 1 in 959; at age 40, it is 1 in 84; and at age 50 it is 1 in 44.[4] Although the probability increases with maternal age, 70% of children with Down syndrome are born to women 35 years of age and younger, because younger people have more children.[4] The father's older age is also a risk factor in women older than 35, but not in women younger than 35, and may partly explain the increase in risk as women age.[134]

Lionel Penrose



Born	Lionel Sharples Penrose 11 June 1898[1] London, UK[3]
Died	12 May 1972 (aged 73) London, UK
Alma mater	St John's College, Cambridge University of Vienna King's College London
Known for	Penrose triangle Penrose method Penrose stairs[4] Penrose's Law[5][6] Penrose square root law Penrose–Banzhaf index
Spouse	Margaret Leathes (m. 1928)
Children	Oliver Penrose Roger Penrose Jonathan Penrose Shirley Hodgson
Awards	Fellow of the Royal Society[1] Lasker Award[2] James Spence Medal 1964.
Scientific career	
Fields	Pediatrics, Psychiatry, Genetics
Institutions	University of Cambridge University College London

J Genetics 1933

THE RELATIVE EFFECTS OF PATERNAL AND MATERNAL AGE IN DOWN'S SYNDROME

BY L.S. PENROSE, M.D.*

150 families, each containing one or more Down's syndrome children.

.

After accounting for the high correlation in the parents' ages,
he concluded that the father's age is 'not a significant factor,'
while the **mother's age** 'is to be regarded as **very important**.'

PROCEEDINGS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

The Relative Aetiological Importance of Birth Order and Maternal Age in [redacted]

L. S. Penrose

Proc. R. Soc. Lond. B 1934 **115**, doi: 10.1098/rspb.1934.0051, published 1 August 1934

First submission received by the Royal Society on November 25, 1933.

217 families
(210 had 1 affected child, 7 had 2: $\rightarrow$ 224 'Cases')

Table I—Scatter Diagram showing Relationship of Maternal Age to Birth Rank

(Suffixes in bold type indicate [redacted])

Age ↓ Birth Order →

Age	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	Total	(N)	([redacted])
17	1	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1	1	—
18	2	1	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	3	3	—
19	8_3	—	1	—	—	—	—	—	—	—	—	—	—	—	—	—	—	9	6	3
20	11_1	3	—	1	—	—	—	—	—	—	—	—	—	—	—	—	—	15	14	1
21	12_3	7	1	—	—	—	—	—	—	—	—	—	—	—	—	—	—	20	17	3
22	7_2	9_2	3	1	—	—	—	—	—	—	—	—	—	—	—	—	—	20	16	4
23	13_2	11_2	5	1	—	—	—	—	—	—	—	—	—	—	—	—	—	30	26	4
24	16_1	10	13	4	—	—	—	—	—	—	—	—	—	—	—	—	—	43	42	1
25	15	13	6	2	1	2	—	—	—	—	—	—	—	—	—	—	—	39	39	—
26	9_2	13	10_1	9	3	—	—	—	—	—	—	—	—	—	—	—	—	44	41	3
27	5_1	13_3	9_1	8_1	10	—	1	—	—	—	—	—	—	—	—	—	—	46	40	6
28	11_1	9_1	5	3	6	4_1	2	1	—	—	—	—	—	—	—	—	—	41	38	3
29	6_1	7_1	9	10	4	5	1	1	—	—	—	—	—	—	—	—	—	43	41	2
30	8	10_1	11_1	6	6	5	5	—	2	—	—	—	—	—	—	—	—	53	51	2
31	5_1	4	9_1	5_1	9	8_1	2	2	1	1	—	—	—	—	—	—	—	46	42	4
32	5_1	13_2	4_1	7	8	7_1	6	2	2_1	—	—	—	—	—	—	—	—	54	48	6
33	5	5_3	7_2	8	7_1	8	6	2	2	1	—	—	—	—	—	—	—	51	45	6
34	3	9_2	8_4	5_2	9	2	6	6	1_1	1	1	—	—	—	—	—	—	51	42	9
35	3_1	2_1	8_1	7_1	8_2	10_3	10_2	8_2	3	2	—	1	—	—	—	—	—	62	49	13
36	2_1	3_3	4	5_2	5	6	1	6_1	3	4	3_1	—	—	—	—	—	—	42	34	8
37	1	2	4_3	8_4	4_1	4_1	3_2	7_3	5_1	3	1	—	1	—	—	—	—	43	28	15
38	2_2	4	1	4_2	5	5_2	11_6	1	4_1	3_1	2	2	1	1	—	—	—	46	32	14
39	2_1	1_1	4_2	7_3	5_1	3_1	5_2	2	3_1	4_2	1	3	1	—	—	—	—	41	27	14
40	1_1	3_1	4_1	2_1	2_1	6_3	9_4	2_2	4_2	—	2	2	1_1	1_1	—	—	—	39	21	18
41	—	1	5_2	3_1	—	—	3_2	4_2	2_1	6_3	4_2	4_2	1_1	1	—	—	—	34	18	16
42	—	3_3	5_5	3_3	2_1	2_2	2_1	5_3	1	2_1	2_1	2	4_2	1	1_1	—	—	35	12	23
43	—	1_1	1_1	2_1	2_2	3_2	—	4_2	2_1	1	—	1	1	1	3_2	1	—	23	11	12
44	—	—	—	—	4_3	4_4	2	3_2	2_1	—	—	—	2_1	2_1	—	1	—	20	8	12
45	1_1	—	—	—	—	2_1	1_1	1	4_1	1_1	1_1	1	1	1	—	—	1_1	15	8	7
46	—	—	2_2	1_1	1_1	1_1	—	1	1_1	1_1	—	—	2	2_1	1	—	1	14	6	8
47	—	—	—	—	—	—	—	2_2	1_1	1	—	—	1_1	—	1_1	—	1_1	7	1	6
48	—	—	—	—	—	—	—	—	—	—	—	—	—	1_1	—	—	—	1	—	1
Total	154	157	139	112	101	87	76	60	43	31	17	16	16	11	6	2	3	1031	—	—
(N)	128	130	111	89	88	64	56	41	30	22	12	14	10	7	2	2	1	—	$\boxed{807}$	—
([redacted])........	26	27	28	23	13	23	20	19	13	9	5	2	6	4	4	—	2	—	—	$\boxed{224}$

The birth order was also recorded with particular care: miscarriages and stillbirths were deemed to affect the ordinal number of subsequent births, but they have been excluded from the data as presented here. It is very uncertain whether they represent offspring affected or not by Down's syndrome and I wish to include in the data only those individuals in the 217 sibships of whom it could be said with certainly that they were either Down's syndrome or not.

Peer Review: 1933

8th December 1933.

Dr. Penrose,
Royal Eastern Counties Institution,
Essex Hall,
Colchester.

Dear Dr. Penrose,

I have had your paper on age and birth order among [redacted] sent to me as referee by the Royal Society, and as I have a good deal to say, I am writing directly to you, instead of letting it trickle through anonymously as extracts from the referee's report. Either way I am afraid you will find it a confounded nuisance.

The whole difficulty turns on the point made in section three, but that section makes it far from clear. You do not mention the essential point, that choosing families only containing [redacted], the proportion of [redacted] must be highest in the smallest families, which generally contain early, but not late children by birth rank.

Now it seems to me that your family data are much too important for you to be satisfied with an unconvincing statistical analysis.

I mean, that no one reading your paper critically will feel sure that a more exact treatment would not have yielded a different result.

I may add that I entirely expect your actual conclusions to be the right ones, but that is no sufficient reason why they should not be adequately established.

The only convincing test for a theory, is a direct comparison between what has been observed, and what must be expected on that theory.

The appropriate theory here is, that the probability of a Down's syndrome child depends on age, in some manner unknown prior to the data, but not, given the age, on the birth rank.

As I think you already see the only relevant facts available to test this theory consist of the distribution of Down's syndrome children within families of given constitution in respect of (a) number of children recorded, (b) birth rank of these children, (c) maternal ages, and (d) number of Down's syndrome children. Families wholly Down's syndrome, like families wholly normal will give no information.

You're full publication of the data is excellent but in table one I think you ought to give the number of Down syndrome children at each age and birth, rank, either as a suffix or in brackets, following the total number of cases.

Revised Manuscript

“To avoid these sources of ambiguity the data have been subjected to analysis by an entirely different method which was suggested by Professor R. A. Fisher. By use of this new process we are able, after a single complex reconstruction, [i.e. a conditional logistic model]

to compare the observed number of Down’s syndrome cases in any given birth rank with the number which is to be expected on the hypothesis that the probability of a Down’s syndrome child depends upon maternal age (in some manner unknown prior to the data) but not, given age, upon birth rank."

He didn’t fit a model with both age and birth order; he fitted one based just on age, and then (effectively) grouped the residuals by birth order.

I will focus on this age-only model, where he categorized age as 7 age-bins, each 5 years wide. So his model had 6 free age-effect parameters.

Table 1. *Trial ω values, and age-specific fitted ('calculated') frequencies of Down's syndrome (DS) children*

Data	Maternal age group	15–19	20–24	25–29	30–34	35–39	40–44	45–49
	Observed no. of normal children	10	114	199	228	170	67	15
Fitting	Observed no. of DS children	3	13	14	27	64	81	22
Trial no.								
1	ω values	83[4]	31[2]	19[**1**]	33[2]	104[5]	321[15]	407[20]
	Calculated no. of DS children **(fitted)**	3.69	16.87	19.50	29.11	59.48	76.99	18.35
⋮								
7	ω values	22[4]	10[2]	6[**1**]	19[3]	88[15]	296[50]	558[90]
	Calculated no. of DS children	2.98	12.87	13.82	26.58	64.14	81.46	22.17
....	....							
`clogit`	Scaled ω values	[3.47]	[1.59]	[**1**]	[2.96]	[13.19]	[43.62]	[81.53]

Source: Page 440 of Penrose (1934a).

Since the ω values are relative odds, values in brackets have been scaled so that the lowest risk age group (25–29) serves as the reference category, with a scaled odds of **1**:1. See Penrose (1934b) for how he chose the ω's for each trial. The scaled ω values fitted in five iterations by the `clogit` function in the `R survival` package (R Development Core Team, 2024) yielded calculated frequencies that were, in absolute terms, within 10^{-9} of the observed ones.

Fisher's Model

In a cryptic passage that puzzled me for years, and that I explain in the next section, Penrose then presents a model, of an as-yet-to-be-specified functional form, for a specific pair of maternal ages. He adopted Fisher's symbol x to denote a relative odds. Here, I have replaced it by the Greek letter ω, and replaced his letter S for the sum by today's $\sum$.

> Let us suppose that there are a number of families containing only two children born at the maternal ages of 32 and 42, respectively, and that one child in each family has Down's syndrome. Call p_{32} and p_{42} the [age-specific] probabilities that a Down's syndrome child is born at these maternal ages. The frequencies of families which have the Down's syndrome child at age 32 to those which have the Down's syndrome child at 42 will be in the ratio $p_{32}/(1-p_{32}) : p_{42}/(1-p_{42})$, or, say, $\omega_{32} : \omega_{42}$ where ω is proportional to [the odds] $p/(1-p)$. In any such family the expectation that the child born at 32 is a, or in this case, the, child with Down's syndrome is $\omega_{32}/(\omega_{32}+\omega_{42})$.

He then explains that, in general, this means that

> for families containing only one Down's syndrome child, the expectation that a child whose (relative odds) was ω, is the affected one is $\omega/\sum\omega'$, where $\sum\omega'$ is the sum of the values of ω for the maternal ages of [each of the] children in the family.

With the children in the family regarded as a set, this expectation has the same structure as the conditional probability defined in § 2; thus, the likelihood contribution also has the same structure. Fisher had not yet specified a functional form for the age specific p. **(7 such families)**

The more complex expressions for the expectations involving families with more than one Down's syndrome child were left for the technical paper, and will be addressed in the next section of the

POPULATION MODEL

MODEL INDUCED BY OUTCOME-BASED-SAMPLING

(For the observed data)

Fisher's Criterion for the best-fitting ω values

Before he addressed the form of the ω function, Penrose stated the operational criterion of fit, which in his review, Fisher had simply set out, without justification:

> *"the best-fitting ω values will be those where the number of Down's syndrome children observed at any given maternal age tallies with (equals) the sum of the expectations attributed to each child at that maternal age."*

In each age bin, Observed Number = Fitted Number

Cox1972 & JH 2024: the ω's that satisfy this estimating equation are Maximum Likelihood estimates.

Families with 2 affected children ['tied' observations in 'survival' data]

The calculations of expectations (and thus the likelihood contribution from each such family/'set') are admirably laid out in the (separate) technical paper

A METHOD OF SEPARATING THE RELATIVE AETIOLOGICAL EFFECTS OF BIRTH ORDER AND MATERNAL AGE, WITH SPECIAL REFERENCE TO [redacted]

Annals of Eugenics
Vol 6, Issue 1 Oct 1934
pp 108-131

By L. S. PENROSE, M.D.

From the Research Department, Royal Eastern Counties' Institution

§ I

[redacted] is a not very uncommon developmental abnormality which tends to affect children who are born at the end of a family, and the incidence of the condition increases as maternal age increases. There are, quite possibly, other human diseases or characters which occur frequently either at the beginning or at the end of sibships. Wright (1) observed that coat colour in guinea-pigs varied in association with the age of the dam, and a similar effect was observed in a certain type of polydactyly. Wright was able to show, by use of the method of partial correlation, that the number of pregnancies of the dam had no effect upon the incidence of these characters. In a paper recently published in the *Proceedings of the Royal Society* (2) an analysis of human data concerning [redacted] was undertaken. Two alternative methods were used in this analysis. The first method corresponded to Wright's technique of partial correlation, but it was complicated by the necessity for reconstructing the data in order to allow for the varying sizes of human families and the mode of their selection. The first method had several disadvantages, which were avoided by using the second method suggested by Prof. R. A. Fisher. It was not, however, possible to deal with the second method in full in the paper just referred to, and, in particular, it was not possible to give an account of how the sampling errors of the expectations were obtained. The purpose of this paper is to describe Prof. Fisher's method in detail, so that it may be possible for a future investigator to repeat the process on fresh data concerning mongolism or any other condition in which maternal age or birth order is suspected of being aetiologically significant.

The data on which the calculations which follow are based are given as an appendix to this paper in the same form as that given in the paper referred to above. The data consist of 217 sibships containing at least one [redacted] each. In order that the results of family history investigation may be suitable for the application of the analytical method described here it is necessary for sibships to be recorded giving the order of birth of each individual child. Affected and normal children must be clearly distinguished and, when it is impossible to know whether offspring are affected or normal, they must be excluded: thus miscarriages and still-births will not appear as individuals in the data, but they will affect the birth

and repeated in modern notation in the Bka 2024 piece.

$p\,q' : p'\,q$

or, $\frac{1}{4}\,\frac{p}{q} : \frac{1}{2}\,\frac{p'}{q'}$

or say, $x : x'$

where ~~it~~ x is proportional to p/q .

For families containing one [redacted] but more than one normal, the expectations are clearly $\frac{x}{S(x)}$ $\frac{x}{S(x)}, \frac{x'}{S(x)}, \frac{x''}{S(x)}$

~~When~~ where $S(x)$ is the sum of the value x for the different maternal ages in the family.

B For families containing two [redacted]s the expectations of [redacted]s at each place will be

$$\frac{x\,S'(x)}{SS\,(x\,x')}$$ adding up to two..

When $S'(x)$ is the sum of the other values, and $SS\,(x\,x')$ stands for the sum of all the products two at a time.

Given the series of x values, therefore, for all ages, the expectations of each recorded child being a [redacted] can be set down, and the number in each birth rank compared with what is actually observed. The assigned x values will be correct when these observed numbers at each age tally with those expected. The correct procedure is therefore to start with a trial series of x values, increasing with age, based on the proportions of mongols observed at those ages. On your theory of the simple recessive, the values of q should not fall below three-quarters, so that the trial values

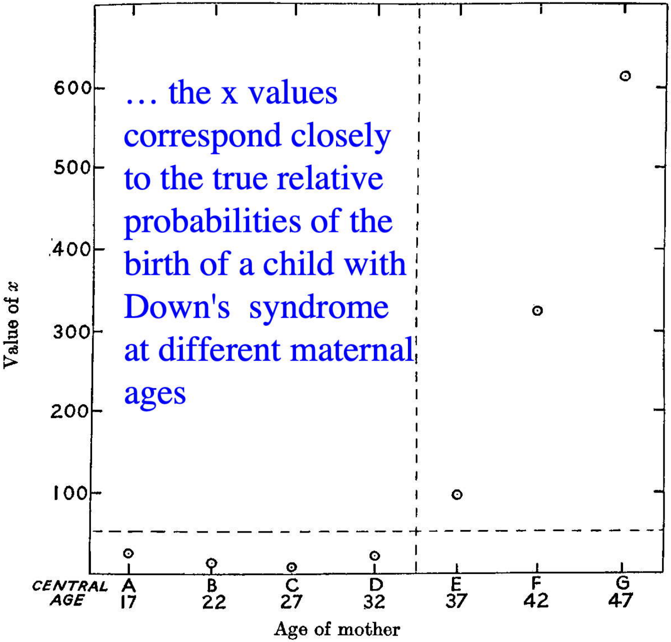


Fig. 1. Final estimate of values of x for maternal age groups

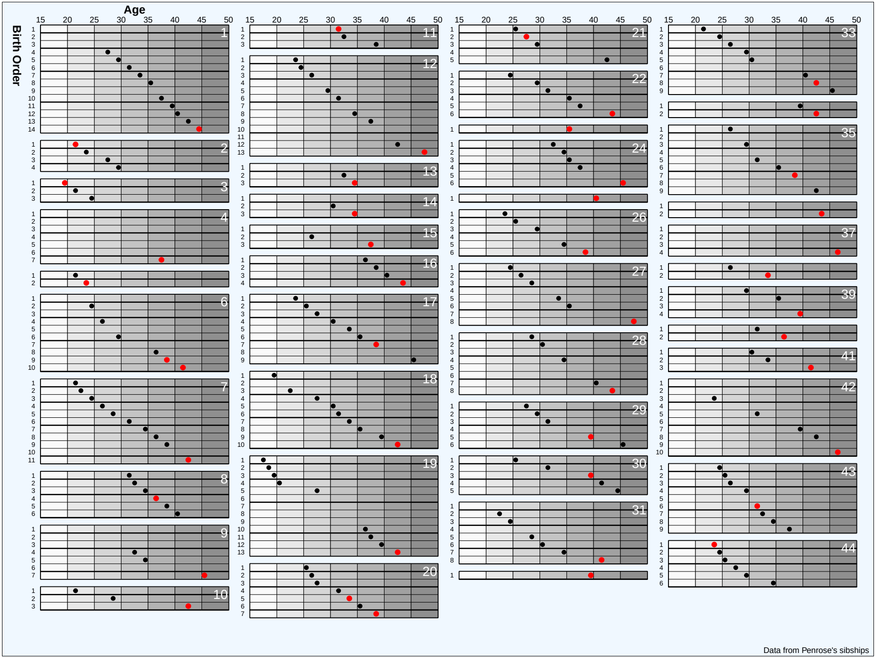
Maternal Age.

Serial number	Sex	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48
1	f											4		5		6		7		8		10		11	12		13		**14**				
2	m					**1**		2				3		4																			
3	m			**1**		2			3																								
4	m																					**7**											
5	m					1		**2**																									
6	m, f								2		4			6							8		**9**			**10**							
7	m					1	2		3		4		5			6			7		8		9				**11**						
8	m															1	2		3		**4**		5		6								
9	m																4		5											**7**			
10	m					1							2														**3**						
11	f															**1**	2						3										
12	f							1	2		3			5		6			8			9					12					**13**	
13	f																2		**3**														
14	m														2				**3**														
15	f										2											**3**											
16	f																				1		2		3			**4**					
17	m							1		2		3			4			5		6			**7**							9			
18	f			1			3					4			5	6		7		8				9			**10**						
19	m	1	2	3	4							5									10	11		12			**13**						
20	m, f									1	2	3				4		**5**		6			**7**										
21	f									1		**2**		3													5						
22	m								1					2		3				4		5						**6**					
23	m																			**1**													
24	m																1		2	3		4								**6**			
25	f																								**1**								
26	f							1		2				3					5				**6**										
27	f								1		2		3					5		6												**8**	
28	f												1		2				4						7			**8**					
29	m											1		2		3								**5**						6			
30	m									1						2								**3**		4			5				

Maternal Age—(*continued*)

Serial number	Sex	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48
196	m						1			3			7				**9**																
197	m					1		2	3		4		5		6							**7**											
198	m, f															**3**			4				5						**6**				
199	f								1	2							4	5							**7**								
200	f																					1			2			**3**					
201	m														1		2		3			**4**											
202	f												1						2								**3**						
203	m						**1**																										
204	m													1								**3**											
205	f												1			3	4				5				6							**8**	
206	m			**1**																													
207	f																													**1**			
208	f															1				4		6	7				**8**						
209	f																		1						**2**								
210	m									1		2									3			**4**									
211	m														9	10						13						14				**15**	
212	f												1		**2**	3																	
213	m											4			5		6			**7**													
214	m																		2		3						**6**						
215	m				1		2			3										8		9		10		11		12			13		**14**
216	f															5	6	7			10						14	**15**					
217	m, m																		2												{**3**, **3**		
		1	3	6	14	17	16	26	42	39	41	40	38	41	51	42	48	45	42	49	34	28	32	27	21	18	12	11	8	8	6	1	—
		—	—	**3**	**1**	**3**	**4**	**4**	**1**	—	**3**	**6**	**3**	**2**	**2**	**4**	**6**	**6**	**9**	**13**	**8**	**15**	**14**	**14**	**18**	**16**	**23**	**12**	**12**	**7**	**8**	**6**	**1**

JH has assembled these data into a ‘long’ .csv file that is available on his website `https://jhanley.biostat.mcgill.ca/Penrose/`


Age
Birth Order
Data from Penrose's sibships

RECORDS OF SPECIAL CONVOCATIONS DURING THE CONGRESS

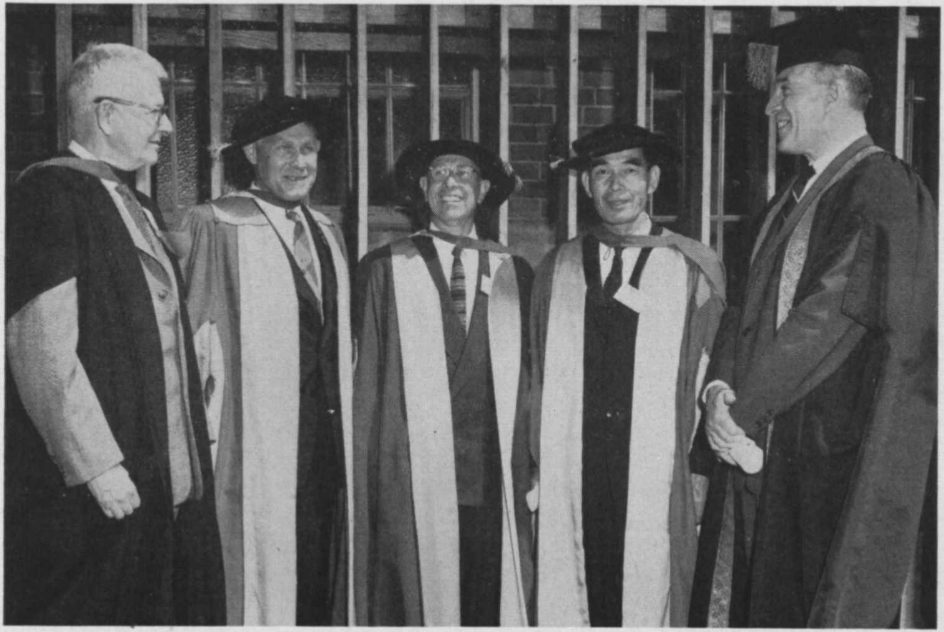
10th International Congress of Genetics, McGill University, Aug 20-27, 1958

A Special Convocation of McGill University was held in the Percival Molson Memorial Stadium on Wednesday, August 20. Degrees of Doctor of Science, *honoris causa*, were conferred upon Professor Hitoshi Kihara, Professor Lionel S. Penrose, and Professor Curt Stern. The remarks of the Principal and Vice-Chancellor, Dr. F. Cyril James, are included below together with the citation of the recipients prepared by Dr. Lloyd G. Stevenson, Dean of the Faculty of Medicine, and the response of Professor Curt Stern for the recipients.

Remarks of the Principal and Vice-Chancellor

It is my privilege this morning, from this Convocation platform, to offer to each of you a warm welcome to Montreal, and especially to McGill University. I hope that the arrangements made for your comfort by the Committee headed by my colleague Professor J. W. Boyes will make your stay pleasant, and that the scientific discussions during this X International Congress of Genetics will be intellectually rewarding. When you meet again, at some other place, for the eleventh Congress, I hope that the tenth will be a pleasant memory to evoke nostalgic talk during the intervals between more serious discussion.

Genetics is a comparatively new science, but we at McGill were early indoctrinated by the enthusiasm and skill of a master. Although Leonard Huskins—whom the Genetics Society of Canada commemorates in its annual Memorial Lecture—came to McGill as Associate Professor of Botany in 1930, his enthusiasm



Special Convocation, McGill University: *left to right*—Professor Sewall Wright, Professor L. S. Penrose, Professor Curt Stern, Professor H. Kihara, Professor F. Cyril James (Principal and Vice-Chancellor)

WRAP-UP

Take-home messages

- Our methods are older than we think
- And are born of necessity
- The 'practising statistician' (collaboration with researchers)
- BMJ: $\rightarrow$ The ETIOLOGIC STUDY
- (RCT) Study bases can also be sampled for the PROGNOSTIC STUDY

  JH & OSM (2009) Fitting Smooth-in-Time Prognostic Risk Functions via Logistic Regression

  `casebase` (2024) package by Bhatnagar/Turgeon/Islam/Saarela/Hanley

Thanks to:

- Penrose (UCL) & Fisher (U Adelaide) Online Archives
- Andrea Benedetti and her `bios624` class, Fall 2014.
- All of my colleagues since 1973
- My (trained-at-Rothamsted under Yates) University College Cork statistics professor Tadgh Carey who said to our small (1966-1969) class

  You are still too young but 'one day' I will let you see the statistics journals where Fisher and Pearson were so nasty to each other.