B. Burt Gerstman

Dep. of Health Science, San Jose State University, San Jose, California

Comments regarding "*On prognosis*" by William Farr (1838), with reconstruction of his longitudinal analysis of smallpox recovery and death rates

The editor has kindly asked me to comment from a statistical perspective on the historical article *"On prognosis"* by William Farr (1838). To place this article in context, I note that it was published about a year *after* the Registrar-General's office was established in 1837, and a year *before* Farr's appointment to his first official post as Compiler of Abstracts in this office, Farr later assumed the position of Superintendent of the Statistical Department in the Registrar-General's office, a position he held until his retirement in 1880 (Humphreys 1885).

The article *On prognosis* is notable in many ways. It reveals Farr's early insights into population-based epidemiologic principles, a fundamental understanding of longitudinal analysis, an interest in clinical research, and a commitment to explaining the course of health and disease in terms of "simple laws susceptible to calculation". It is no wonder that Susser and Adelstein (1975) referred to Farr as "a founder, even the founder of epidemiology in its modern form".

Population-based epidemiologic principles

Farr's role in innovating population-based epidemiologic principles has been lauded by many prominent sources (Susser & Adelstein 1975; Langmuir 1976; Lilienfeld & Lilienfeld 1977). Examples of epidemiologic principles discussed in *"On prognosis"* include references to the spectrum of disease and "the epidemiologic iceberg" (Last 1963), a clear distinction between risks and rates (Elandt-Johnson 1975; Vandenbroucke 1985), and exploration of rates by "person, place, and time" variables. In reference to the epidemiologic iceberg, Farr writes on page 224 (all page references will be to the re-publication as it appears in *Soz.-Präventivmed*, Hill GB 2003a, b):

> Attacks of disease differ in intensity; some are so slight as to obtain no attention.

In reference to the distinction between risk and rates, Farr writes (p. 223):

> Disease may be examined (1) in their tendency to destroy life, expressed by the deaths out of a given number of cases; and (2) in their mean relative force of mortality, expressed by the deaths out of a given number sick at a given time.

And later, on page 279, the mortality rate is described as the ratio of the number of deaths divided by the central (mean) number exposed to risk:

> The *relative force of mortality* is an unusual term; but it implies the rate of dying – the number of deaths out of a given number, living a given time – and is required by the present state of science.

In sections labeled "a. The patient, b. External Circumstances, c. Diseases, and d. Periods of Disease" (pp. 220–224), Farr explores morbidity and mortality rates by person, place, and time variables. For example, an interesting section on recovery from insanity is presented on page 222 in which Farr notes diminishing rates of recovery with age, but constant rates of mortality.

Longitudinal analysis

Farr's application of current ("cross-sectional") life table methods to vital statistics is well-known, but his analysis of clinical cohorts is less frequently cited (Hill 1997). On page 223, Farr writes:

> To determine the mortality of diseases they should be followed from the beginning to the end; every death or recovery should be recorded; and this though exceedingly simple, has been rarely done. The mortality of cases can only be accurately ascertained by practitioners, who see

Soz.- Präventivmed. 48 (2003) 285– 289
0303-8408/03/050285–05
DOI 10.1007/s00038-003-3001-y

cases as they occur, slight or severe, and seldom lose sight of them to the end.

The importance of not losing sight of individual experiences "from the beginning to the end" underlies all longitudinal analyses and time-failure models, and forms the basis of cohort and case-control study methods (Miettinen 1976). That Farr was ahead of the time in applying time-failure principles in measuring human health and disease is illuminated in his *Table of Sickness* that appears on page 280. By working from first principles, and using methods most likely learned from the British actuary Thomas Rowe Edmonds (Eyler 1980; 2002), Farr determines the extent to which two separate states (life and illness) persist in individuals. This allowed him to estimate probabilities of their complementary states (death and recovery) by describing the cohort's survival experience at its essence. In his *Table of Sickness,* Farr shows the number recovering (column A), the number dying (column B), and the number remaining sick (column C) in a cohort of 5268 smallpox patients. On page 281 Farr states

> The probability of recovery is shown by the numbers of column A and column B, in juxtaposition.

By this, the autodidact Farr means the *odds* of recovery. Notwithstanding this semantical error, Farr properly calculates probabilities of survival and death, occasionally referring to them "fractions":

> at the 30th day the probability of recovery has risen; it is [2956] to 104 (29 to 1). The fraction expressing the probability of recovery is 2956/3060; the probability of dying is 104/3060; both probabilities added together make certainty (p. 281).

Based on these calculations, Farr derives a remarkably modern inference:

> The probability of dying constantly decreases in acute disease; as the deaths take place at an earlier period than the recoveries.

Today we would record this as evidence of *non-constant hazard*, a fact that would have been lost if data had been reduced to a too-simple comparison of proportions without accounting for time. Had graphical computing been available at the time, Farr might have expressed these results as an empirical survival curve (Fig. 1) or, better yet, as incidence rates plotted over time (Fig. 2).

Farr was also able to calculate the expected ("mean future") duration of disease for people who were to recover and for those who were to die, and for both groups combined (p. 281). His methods are akin to summing the number of person-days remaining in each subcohort ($_nT_x$ in current life table notation) and dividing this quantity by the number of people who began the interval. This allowed the calculation of the expected (mean future) duration of the state, or e_x in today's notation. (See Elandt-Johnson and Johnson (1980) for a review of computational methods). It should be pointed out that these expectations are superior to simple averages since they separate out survivors and decedents and deal with censored data in a relatively unbiased way (Morabia 1996; Hill 1997).

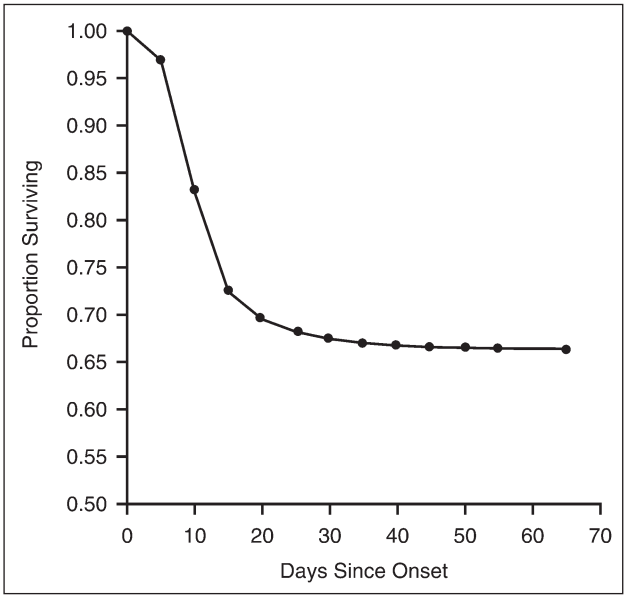

Figure 1 Survival following smallpox, based on Farr's *Table of Sickness* [Table 3]

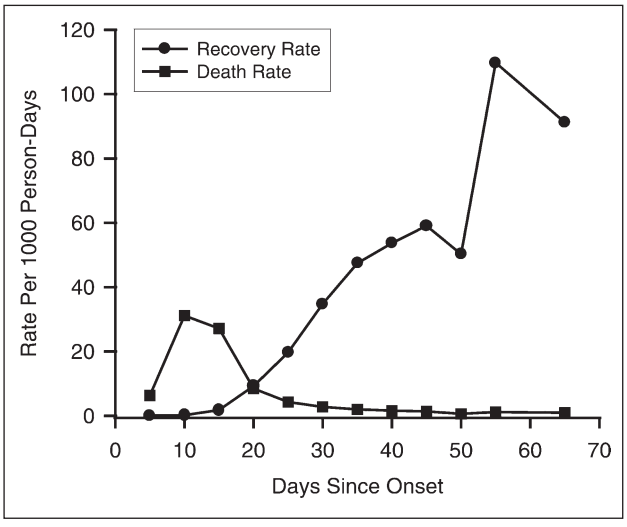

Figure 2 Rates of recovery and death in smallpox patients over time

Scrutiny of Farr's *Table of Sickness* ([Table 3], p. 280) and *Table of deaths and recoveries* ([Table 6], p. 282)

Farr's *Table of Sickness* (p. 280) is a modified life table that follows outcomes in a cohort of 5268 smallpox cases. The table begins on day 5 following hospitalization, for the reason explained elegantly in the original article (p. 281), with the cohort divided into those who ultimately recover (column A) and those who die (column B). Data are complete for 5-day follow-up intervals up to day 55. This is followed by two 10-day follow-up intervals (55–64 days; 64–4 days) before a final interval of indeterminate length is reported. There are no apparent losses before the final interval. Because data are incomplete for this last interval – the status of these individuals into the future is censored – Farr makes several assumptions so that he may complete his analysis. One of these assumptions is referenced in the form of a footnote to the table, where it is called an "interpolation". This interpolation assumes that, following the 55th day of illness, 0.76 of those bound to recover remain ill for at least an additional five days (i.e., 24% of those due to recover will do so within 5 days)* and that deaths (Column B) occur uniformly at 1 for every five days. Column C ("To Recover or Die") contains those remaining alive but ill as of the beginning of the interval (simply the sum of column A and column B). Column D ("Recover") and column E ("Die") are the actual data from which all other columns are derived. Column F ("Terminate") is simply the sum of column D and column E.

The aforementioned *Table of Sickness* is used to build the *Table of the deaths and recoveries* on page. 282. Column *a* in the *Table of the deaths and recoveries* contains the number of people *effectively* exposed to risk during the interval ("Constantly sick" in Farr's terminology). This is simply the average number of people who persisted in the state of illness until the next interval, assuming a uniform (or symmetrical) resolution of the current state. Farr makes an elegant description of how he calculated this number on page 281, but unfortunately makes an arithmetic error in two of his calculations. During the first interval, the average of 5268 and 5102 is 5185 (not 5135). During the second interval, the number of people effectively exposed to risk should read 4731 (not 4681). Otherwise, Column *a* in the *Table* is clean, apart from rounding errors.

Column *a* data in the *Table of the deaths and recoveries* serves as denominator data for the death rate (column *d*) and recovery rate (column *f*) in the table. For me, it is easier to see how these rates were calculated if we first determine the number of person-days in the interval as the product of the effective number exposed to risk (column *a*) and length of the interval (value in column $a \times 5$ for 5-day intervals, value in column $a \times 10$ for 10-day intervals). In Table 1 of this commentary, I show how these death rates and recovery rates can be calculated.

"Simple laws susceptible to calculation"

Column *e* in Farr's *Table of deaths and recoveries* (labeled "calculated") contains hypothetical mortality rates based on an assumed geometric progression of occurrence. Three stages of mortality are posited: for days 5–10, for days 10–15, and from day 15 onward. Farr concentrates his effort on predicting the mortality progression from day 15 onward. When the observed rates flatten out, Farr superimposes an additional rate on the predicted rates ("a rate regulating these rates") to allow his model to better fit the data. This, too, Farr, declares to act "according to a determined law" (see pp. 282–283). Farr then goes on to apply these law of progression to other populations (e.g., Paris) and other diseases (e.g., cholera, phthisis).

At this point it may be worth noting an important change in statistical reasoning since Farr's time. We no longer view statistical models as representing fixed *laws* of nature. Rather, statistical models are viewed as *tools* of science. Whether a particular model is true is irrelevant (Zeger 1991). What counts is whether we obtain correct scientific conclusions if we believe in the fiction of the model. "The hallmark of good science is that it uses models and "theory" but never believes them" (Martin Wilk cited in Tukey (1962) on p. 7). We have no reason to believe that Farr's laws of morbidity may be applied uniformly across populations and diseases. We may even speculate that the cohort studied by Farr was a select one, not necessarily representing the experience of smallpox cases in the source population. Thus, the generalizability of results is tenuous. Health outcomes are influenced by multiple causal factors acting together, and the incidence of any outcome can be assumed to vary across populations, depending on the prevalence of its complementary factors (Rothman 1976). Attempts to reduce rates of even a single disease to universal constants would today be considered short-sighted.

* Although Farr does not fully explain the source of this interpolation, it can be noted that approximately three-quarters of those in Column A "survive" (that is, stay in their current state of illness) until the next interval starting at about day 40. For example, the survival proportion from day 40 to 45 is 1468/1934 = 75%; the survival proportion from day 45 to 50 is 1084/1468 = 74%; and so on.

Soz.- Präventivmed. 48 (2003) 285–289

Table 1 Reconstruction of Farr's analysis of smallpox mortality and recovery

First day of interval		k	Cohort[a]	Recoveries[b] r	Deaths[c] d	Persistent illness[d] N	Effective number[e] N′	Person-days[f] T	Death rate[g] per 1000 p-days	Recovery rate[h] per 1000 p-days
5-day risk	0	0	–	–	–	–	–	–	–	–
periods	5	1	5268	2	164	5268	5185.0	25925.0	6.33	0.08
	10	2	5104	5	737	5102	4731.0	23655.0	31.16	0.21
	15	3	4367	37	552	4360	4065.5	20327.5	27.16	1.82
	20	4	3815	167[i]	153	3771	3611.0	18055.0	8.47	9.25
	25	5	3662	321	70	3451	3255.5	16277.5	4.30	19.72
	30	6	3592	487	39	3060	2797.0	13985.0	2.79	34.82
	35	7	3553	535	23	2534	2255.0	11275.0	2.04	47.45
	40	8	3530	466	14	1976	1736.0	8680.0	1.61	53.69
	45	9	3516	384	9	1496	1299.5	6497.5	1.39	59.10
	50	10	3507	246	3	1103	978.5	4892.5	0.61	50.28
10-day risk	55	11	3504	367	4	854	668.5	6685.0	0.60	54.90
periods	65	12	3500	179	2	483	392.5	3925.0	0.51	45.61
Remainder	75 – ?	13	3498	292	10	302				
TOTALS			3488	1780			160180.0			

[a] Initial cohort minus number of deaths from prior interval. Not reported by Farr, but needed to perform routine survival analysis and track the cohort
[b] From Column D in Farr's *Table of Sickness*
[c] From Column E in Farr's *Table of Sickness*
[d] This is the number of people in which illness persisted. You can think of it as the number who remained hospitalized as of the first day of the interval
$N_{k+1} = N_k - d_k - r_k$, where N_k = the number with persistent illness as of the first day of interval k, d_k = the number of deaths that occurred during this interval, and r_k = the number of recoveries that occurred during this interval. For example, $N_1 = 5268-64-2 = 5102$. This is reported as Column C in Farr's *Table of Sickness*.
[e] $N_k' = N_k - .5(r_k) - .5(d_k)$. Assumes mean time of recovery and death is mid-interval; Similar to column a in Farr's *Table of deaths and recoveries*
[f] $T_k = N_k' \times 5$ for 5 day intervals; $T_k = N_k' \times 10$ for 10 day intervals
[g] Death rate per 1000 person-days = $d_k/T_k \times 1000$. This information is the same as column d of Farr's *Table of deaths and recoveries* (corrected)
[h] Recovery rate per 1000 person-days = $r_k/T_k; \times 1000$. This information is the same as column f of Farr's *Table of deaths and recoveries* (corrected)
[i] "*On prognosis*" lists this value as 176 – a typesetting error

Acknowledgements

The authors expresses his appreciation to Dr. G. Hill and Dr. J. Eyler for their assistance with historical references and documents. He is also grateful to Dr. A. Morabia for suggesting the idea of a statistical review of Farr 1838 article and for publishing Farr's article in its original form, and to Dr. J. P. Vandenbroucke and Dr. J. Katz for their thoughtful and penetrating reviews of this manuscript.

References

Elandt-Johnson RC (1975). Definition of rates: some remarks on their use and misuse. Am J Epidemiol *102*: 267–71.

Elandt-Johnson RC, Johnson NL (1980). Survival models and data analysis. New York: J. Wiley.

Eyler JM (1980). The conceptual origins of William Farr's epidemiology: numerical methods and social thought in the 1830s. Time, places, and persons. A. M. Lilienfeld. Baltimore: The Johns Hopkins University Press: 1–21.

Eyler JM (2002). Constructing vital statistics: Thomas Rowe Edmonds and William Farr, 1835–1845. Soz Praventiv Med *47*: 6–13.

Farr W (1838). On prognosis. British Medical Almanack Suppl: 199–216.

Hill GB (1997). RE: "P. C. A. Louis and the birth of clinical epidemiology". J Clin Epidemiol *50*: 1187–8.

Hill GB (2003). Typed and edited transcript of *"On Prognosis"* by William Farr. Part I. Soz Praventiv Med *48*: 219–224.

Hill GB (2003). Typed and edited transcript of *"On Prognosis"* by William Farr. Part II. Soz Praventiv Med *48*: 279–284.

Humphreys NA (1885). Biographical sketch of William Farr. Vital Statistics: a memorial volume of selections from the Reports and Writings of William Farr. London: Office of the Sanitary Institute: vii–xxiv.

Langmuir AD (1976). William Farr: founder of modern concepts of surveillance. Int J Epidemiol *5*: 13–8.

Last JM (1963). The iceberg: "Completing the clinical picture" in general practice. Lancet *6:* 28–31.

Lilienfeld DE, Lilienfeld AM (1977). Epidemiology: a retrospective study. Am J Epidemiol *106*: 445–59.

Miettinen O (1976). Estimability and estimation in case-referent studies. Am J Epidemiol *103*: 226–35.

Morabia A (1996). P. C. A. Louis and the birth of clinical epidemiology. J Clinical Epidemiol *49*: 1327–33.

Rothman KJ (1976). Causes. Am J Epidemiol *104*: 587–92.

Susser M, Adelstein A (1975). Introduction. Vital statistics: a memorial volume of selections from he Reports and Writings of William Farr (1885). Metuchen, NJ: Scarecrow Press: iii–xiv.

Tukey JW (1962). The future of data analysis. Ann Math Stat *33*: 1–67.
Vandenbroucke JP (1985). On the rediscovery of a distinction. Am J Epidemiol *121*: 627–8.

Zeger SL (1991). Statistical reasoning in epidemiology. Am J Epidemiol *134*: 1062–6.

Address for correspondence

B. Burt Gerstman
Dept. of Health Science
San Jose State University
San Jose, California, USA 95192–0052
e-mail: gerstman@email.sjsu.edu