Contents

Introduction: CVD Epidemiology Leads the Way

Ancel Keys computing circa 1958

Senior workers in CVD epidemiology recall the chatter of mechanical, then electric calculators as hand-written data were entered from field and laboratory notebooks. Before long, those working in the new field were able to put aside their calculators and slide rules and start sorting punched cards as they handled the increasing volume of data from the first generation of prospective studies. Then, in the 1960s, not long after workers became comfortable with cross-classifications and simple statistical tests advanced by card-sorting, digital electronic computers burst upon the scene. The result was a powerful facilitation of data storage and management, computation of correlations and regressions, and the facile handling analytically of multiple variables simultaneously with programs for the multiple logistic and hazard ratios (Truett, Cornfield, and Kannel 1967; Cox 1972). These advances were technically and emotionally akin to another 1960's phenomenon then under way: man's dramatic blasting off into space with flights to the moon!

While parallel development of computer technology had a powerful impact on design and analysis in epidemiological studies, the discipline of CVD epidemiology itself made major and early contributions to biomedical data handling. It was the pressure of dealing with massive field data collected in the many CVD surveys and cohort studies of the 1950s and 60s that generated an intense motivation to devise ways of managing and analyzing data on a new and immense and more complex scale for the biosciences.

Anecdotal examples of the evolution of data entry and storage, and of computing problems, and the techniques developed to cope with them, provide a flashback to those times. The effort here is not to

treat this important technological history in exhaustive detail, which is well considered elsewhere, but to tell a few stories central to our topic of early prevention research with population data.

Here are mainly stories about the transition to modern computing, mentioning a few of the giants who resolved the issues with their wits and famed machines.

Data-processing and computing at one early CVD epidemiology center

For forty years, John Vilandre was associated with or directed the data-processing and analysis operations at the Laboratory of Physiological Hygiene, which pioneered the field of CVD prevention research, and then at the Division of Epidemiology, University of Minnesota School of Public Health. His depiction here of historic transitions in data management in that center reflect the technical developments common to many North American institutions involved with epidemiological studies over the same period from the 1960s to the 2000s. In a 2005 interview, he began by describing mechanical calculators used from the outset and well into the 1960s:

> These calculators were able to accumulate sums and calculate sums of squares and cross-products of the two variables, values used to compute descriptive statistics such as means, standard deviations, and correlation coefficients. The Monroes and Friedens (brands of mechanical and later electrical calculators) were at first hand-operated but by the time I arrived they were electric. Such mechanical machines placed practical limitations on the sizes of studies because someone had to sit there and add up all those data by hand and do the squares. And, of course, if you have a column of 100 or more numbers and punch it in twice, you are often going to come up with different answers. So you must do it a third time to get the thing rectified.
>
> Eventually a common technique for handling large samples was to use [Hollerith[card-sorting machinery to order the cards on a particular variable, then determine quintile, decile, or centile cut points to use as entry values for the calculating machines. (Vilandre 2005)

Rose Hilk, another long-term LPH data analyst tells of the time "there was a huge card sort going on in the Laboratory and we had a large bin on the wall that held the cards as they came out in order. Dr. Keys was eager to get this analysis done, so he came in late one evening and loaded the card stacker with his finished work. It fell off the wall later that night and I discovered it in the morning. Needless to say, I had to pick up all the cards and resort them!"(Hilk 2005).

Such near-disasters were a common experience of the punch-card era. Kalevi Pyörälä, recounted a harrowing episode in the life of the Helsinki Policemen Study:

> We had a frightening accident in 1970, when all the data files of the study, including the paper archives and large packages of punch cards, had to be transferred from the Institute of Occupational Health, where the first round of the study was done, to the new location provided

> by the Finnish Heart Association. Pirkko Parviainen (later Siltanen), who was our study nurse, ordered a lorry from a transport enterprise to take them to the new location, but they did not arrive at the given address.
>
> For some weeks nothing was heard about them, but then Pirkko got a phone call that the lost packages had been found among other items in the storeroom of the Helsinki main railway station. Thereafter we were more careful and always kept copies of the data in several places. This became easier with the advent of magnetic tapes (Pyörälä, pers.comm.).

As cumbersome as the punch-card system was, it represented a giant step forward in data analysis--a leap into the computer age. Old-timers compare notes about the excitement generated by each technological advance that seemed revolutionary at the time but then was viewed as laughably primitive only a short time later. Vilandre recalled how precious drawers-full of Minnesota data on punched cards were at first transported to the site of large Univac and Control Data computers on and near the Minneapolis campus. This continued for some years, even after the LPH got its first very limited-capacity computer, the Digital Equipment Company's "PDP 8" ("PDP" for "Programmed Data Processor")--introduced in 1963 by Bill Parlin, actuary and data manager at a Twin Cities life insurance company, who became the Lab's first computer programmer. That machine began the transition to computer autonomy:

> Everyone has a computer these days and talks about megabytes and gigabytes of memory. That machine [the PDP 8] had 4 kilobytes of memory! It had no storage device of any kind except paper tape, which isn't really internal. There was no disk drive, no tape drive, no display CRT. Each time you ran a program you had to load it in from the paper tape, and the numbers were punched out on a teletype machine of the kind used by the news media in those days . . . With its attached punched paper tape, it allowed [small] programs and data sets to be stored for later use.
>
> The first thing [Parlin] did was to write a little piece of code called a "handler" (today we call them "drivers") that would interface a card-reading machine with the computer. So now we could actually read the numbers from the punch cards into our computer. You couldn't store anything. All the statistics were done by numbers stored in memory for that run only. Once you turned the machine off you had to start all over.

So, although the PDP 8 allowed us to perform some data analysis in-house, its limitations (small memory, no internal data storage) meant that larger problems still needed to be taken outside (Vilandre 2005).

Programmers at the time who were accustomed to working with high-level programming language had to learn to do machine-language programming, Vilandre said, because trying to use Fortran with the limited memory available on the new, smaller computers resulted in codes that were inefficient. "If you really wanted to get the maximum use of those 4,000 units of PDP memory, you had to write using the computer's basic instruction set. It was great fun, much like doing puzzles" (ibid.).

The PDP 8 was replaced in 1973 by a PDP 11, which Vilandre called “a nice little machine.” Its auxiliary hardware floating-point number processor, now an integral part of any PC computer chip, was then a “6-foot-tall cabinet full of cards with resistors and capacitors--just to do basic multiplication and division.” But its magnetic tape storage allowed data exchanges, back-up, and more sophisticated programming, though still only a single-user machine. The department was growing rapidly and analytical staff had to work in shifts.

The next technological leap, five years later, was to the Lab’s first multi-user system, the PDP 11/34. “Each disk drive was a separate, floor-standing unit,” Vilandre said, “about the size of a small dishwasher. When you think of what's in your little laptop now it seems crazy. But the nice thing about the PDP 11 was its capacity for multiple, concurrent users.” And direct key-to- disk data entry now allowed bypassing punch cards. Vilandre described the next transition:

> The second massive data media conversion, this time from punch cards to magnetic tape, was performed during this period. We had literally millions of punched cards by the mid-1970s. Rose Hilk, over a period of months, fed the cards through a reader attached to a computer and then we stored the data on magnetic tape. We had big bins in the hallway to dump the cards into once they had been read and when the card stock was sold to a recycling company, we got enough money to buy a microwave oven for the kitchen! (ibid.)

Subsequent versions of the PDP 11, each faster and with greater capacity, formed the nucleus of departmental networks of thirty to forty users, in which the machines would perform sluggishly at peak-use hours. Then, with gigantic trials, surveillance, and community projects coming along in the 1970s, the data needs were accommodated by a new kind of DEC computer, the VAX series, still in use in many places. Macintosh computers began to be used as a VAX terminal, and progress since then in operating systems and networks has been steady.

> Before the VAX, it was still necessary to put data on the PDP 11 magnetic tape and take them to the University’s CDC 6600 for batch processing and for-a-fee storage. “Some of the jobs that we were running otherwise would have run for days,” Vilandre said, “while on the CDC machine they could run in a matter of hours. Today, they run on our own desktops in a few minutes” (ibid.).

About 1985, with computing demand increasing, the first VAX computer, a model 8600 was purchased. This new, 32-bit virtual memory machine removed most limitations on program size that were inherent in the PDP series. The new processor had sufficient capacity to run a relational database in software, allowing retirement of the Britton-Lee machine and supporting the SAS analysis system, which quickly supplanted the use of BMDP.

The VMS (Virtual Memory System) operating system is now supported on newer, high speed, 64-bit RISC processors. While the longevity of VMS has meant less retraining of programmers and other staff, today most users in the division at Minnesota use personal computers for the bulk of their work. The

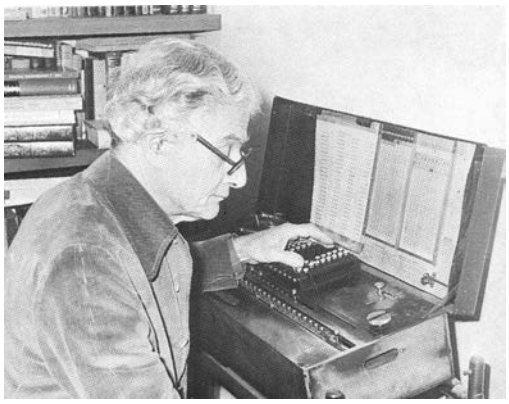
VMS systems are service providers, or servers, to the desk and laptop systems, providing data storage, database capabilities, e-mail service, etc. to the end users.

For all that has been gained in the fast-moving history of high-tech computation, Vilandre noted that something has been lost, as well. With all its limitations, an old-fashioned mechanical calculator at least assured that "you had to understand the process. Now you put it in SAS and get a number out. Sometimes I think people don't even know if their chosen statistic is the appropriate one to use. Anyway, that's why we have statisticians to check the work. Basically, still true today, we human beings are the single weakest point--because we make errors" (ibid.).

Minnesota Epidemiology is now, like most such academic centers, a large, computer-capable, demanding family of faculty and staff requiring servers and intimate support services for computation and communication among many PCs and Macs.

Sir Ronald Fisher and his 'Millionaire' Calculator

R. A. Fisher, often called the father of modern biometry, became a close friend and colleague of William Gosset, statistician at Dublin's Guinness brewery, largely through correspondence. They laboriously and separately calculated the distributions of t; Gosset, for example, calculated all values for t=1 from N=2 to N=30--out to seven decimal places--and found that his were almost identical to Fisher's calculations Gosset's t-test became famous under his pseudonym "Student" that he used to evade the brewery's proprietary policy. Fisher's daughter, Joan Fisher Box, described the scene in a biography of her father:



> One image is [of] Fisher working his "Millionaire" [his self-designed mechanical calculator] at [his laboratory in] Rothamsted, a large machine on which one turned a crank to set the number and inserted a plunger to start its noisy operation at each step; one imagines Gosset putting his hand-operated "Baby Triumphator" into his work sack and carrying it home to work on the tables in the evenings, calculating t by his formula, checking with Fisher's results, and recalculating doubtful values" (Box 1978, 116-117).
>
> Box went on to describe her father's passion for his calculator:
>
> [Fisher] liked his Millionaire calculating machine and was disdainful of the up-to-date desk calculators which were in plentiful supply by the mid-1930s. The Millionaire stood on its own stand. It was clumsy to move and operate and it made a noise like an old threshing machine. [But] Fisher defended it because it did multiply instead of doing multiplication by repeated addition like the other calculators. He could make it perform complicated calculations both

> quickly and accurately. Others also worked much with the machine and appreciated its virtues, like Frank Yates at Rothampsted, who . . . kindly permitted himself to be photographed at the Millionaire still in his office at Rothampsted in 1974. (ibid.)

Fisher devoted much of his life to solutions of genetic and evolutionary questions using statistical methods. Using variance components, he demonstrated that human inheritance was entirely consistent with Mendelian principles. But he is mostly known for provision of tabular distributions of statistics for common use that became popular. With the small experimental samples used at the time, it was essential to know the theoretical distribution to assess the results. He was one of the first to develop the idea of designing experiments to gain more precise information for a given amount of experimental effort, thus developing the methodology of modern biometry. (A short biography of R. A. Fisher, detailing his career and complex relationship with the famed mathematician Karl Pearson, is found at: http://www.morris.umn.edu/~sungurea/introstat/history/w98/RAFisher.html])

Risk-assessment tools for preventive cardiology

The multitude of tables and devices developed for estimating an individual's risk of a future coronary event exceeds capacity for description here, but most of these derive directly or indirectly from the Coronary Risk Handbook prepared from the multivariate analyses and cross-tabulated risk of a coronary event by Tavia Gordon and William Kannel from the Framingham Study as published by the American Heart Association in 1973. The handbook featured tabulations of absolute six-year risk per hundred for coronary events among men and women ages 35 to 65, by class of smoker or non-smoker, with or without non-specific ECG findings, and in detailed tables according to increasing systolic blood pressure and the presence or absence of glucose intolerance. It also provided tables of ideal weight and a management form for treatment and follow-up.

The AHA reproduced and distributed this popular handbook by the tens of thousands until more sophisticated systems appeared for practitioners' use. Now individuals can enter their personal data and receive their risk estimates via the internet. The Framingham Risk Score, based on the study's 10-year and 30-year data, for men and women, for primary or recurrent events and mortality, is the most-used base for calculating risk for individuals or, appropriately or not, for population comparisons, and can also be downloaded from the internet (http://www.framinghamheartstudy.org/risk/index.html).

Pooling epidemiological data: The American Heart Association Pooling Project

Waiting seemed interminable for the first solid results from the first-generation prospective studies. Even when they began to trickle in, in the 1950s, first and mainly from the Framingham Study, there began to be enthusiasm for early data pooling among U.S. cohort studies. Pooled data, with more cases, should help confirm by dent of numbers the weakness of the few events in individual studies, and allow more detailed questions of the data.

But the enthusiasm for pooling seemed proportional to the distance of a study from the Framingham-Albany Study consortium. According to Framingham stalwart, Bill Kannel, it was supposedly the clamor of the "Big E" epidemiologists (theoreticians rather than "real docs") and of those involved in "little studies" (thus, with inadequate numbers) among the epidemiological fellowship who plugged for the AHA to create the Pooling Project. He also explained in our interview another source of his discomfort with the formal pooling process, indicating something of Framingham attitudes about the epidemiological universe:

> You [Blackburn] and I were the younger members [of the AHA Pooling Project], and I was beginning to notice that I'd try and say something to no avail, [for example] when they decided, "By golly we're going to use diastolic pressure," I said, "You know we've got some data about systolic" [as a superior predictor] but nobody paid any attention. And I finally said, "OK, guys." And I just left and never came back and nobody missed me for two days when I rejoined them. So it was apparent that they weren't going to listen. I said to myself "most of the data was Framingham data, so what-the-hell, let it go." And so I did. (Kannel 2000).

The American Heart Association pooling group of the 1970s included Co-principal Investigators Felix Moore and Fred Epstein and members Henry Blackburn, John Chapman, Len Cook, Joseph Doyle, Alan Dyer, Tavia Gordon, William Kannel, Ancel Keys, Patricia McNamara, Paul Oglesby, Rick Shekelle, Jerry Stamler, and Henry Taylor.

In any case, from the time of the First National Conference on Cardiovascular Diseases in 1950 it was the hope and plan of some early investigators to pool data from their studies to reach earlier and firmer conclusions about the influence of the CVD risk characteristics that all were studying, as well as to address issues no one study could soon answer. A plan was set in motion formally in 1961 by colleagues in the AHA Subcommittee on Criteria and Methods. The historical background of their effort is explained in the Final Report of the Pooling Project:

> The incentive to pool data arose from the recognition of the requirement of large numbers, beyond those available in any one study, to evaluate with a reasonable level of certainty the

> relationship of several traits to risk of coronary heart disease (CHD). In 1964, therefore, the principal investigators of several studies and their biostatistical colleagues agreed to collaborate in pooling their data and undertook to elaborate the necessary criteria and procedures. The investigations initially included were those underway in Albany, Chicago, Framingham, Los Angeles and Minnesota. The [Minnesota] northwestern U.S. railroad and Tecumseh community studies were added in 1969. The effort was undertaken as a project of the Council on Epidemiology of the American Heart Association, with support from the American Heart Association, the Heart Disease Control Program, and the National Heart Institute of the U.S. Public Health Service. Pre-requisites for pooling data are comparability of methods of measurement and criteria for defining disease (American Heart Association 1978, 202).

Looking forward to the establishment of a pooling project as an important end in itself, the Committee on Epidemiological Studies (now the Council on Epidemiology) of the AHA, established a Subcommittee on Criteria and Methods (now a committee of the council) in 1961. It first undertook a survey of on-going epidemiological studies of CHD in the U.S. to assess their comparability and to promote agreement on criteria and methods among investigators. The results of this survey were published in Circulation in 1964 with a covering editorial. Following completion of this survey, a coordinating center for a pooling project was established at the University of Michigan in Ann Arbor, under the direction of Fred Epstein and Felix Moore.

The seventeen-year interval between the planning for the project and publication of its final report speaks appropriately to the struggles the AHA Pooling Project encountered in handling the organization and analysis of the data, voluminous for those times, on 8,422 men, ages 40 to 64, with 72,011 person-years of experience (PYE), and 658 first major coronary events. The meager staff at Michigan was eventually swamped, then finally rescued, by the data operation with Alan Dyer in Jerry Stamler's Chicago Health Research Foundation.

The potential to address questions unanswerable in single studies elicited intellectual excitement in the pooling undertaking. The pooling study goals included such issues as the joint relationship of multiple characteristics to CHD when there were huge numbers of potentially combinable variables; or, the Aristotelean logic of blood pressure, which is related to CHD risk, and relative weight, which is related to risk of high blood pressure, where relative weight is little related to CHD risk; and to the more important question of all: the effect on CHD risk of change in risk factors and their trends in the population, using the pooled interval data.

Representing the view of Framingham investigators, Framingham co-investigator, Bill Kannel, wrote to Felix Moore, agreeing with Stamler's stated goals of pooling data, but observing: "It is my opinion that the kinds of analysis undertaken thus far do little to exploit the potential of pooled data. What has been done thus far can be better done, and with greater validity, by each individual participant in the project using their full follow-up" (Kannel 2002).

Framingham may have felt it had settled the issues of "traditional" risk factors and the other study findings were, at best, confirmatory. But Kannel went on to echo Stamler's exciting reasons for pooling:

to broaden the biologic base, to increase exposure to uncommon factors, to make predictions in one population from data of another, to stimulate more sophisticated analyses, to look for consistency, and to get information for uncommon end-points such as silent myocardial infarct.

As Stamler paraphrased his fellow Chicagoan, architect Daniel Burham: "Make no little plans!" (Burnham, a Chicago-based big-city architect and urban planner at the end of the nineteenth century, was responsible for, among other colossal projects, the construction of the World's Columbian Exposition of 1892. He was known for his ambitious pronouncement: "Make no little plans. They have no magic to stir men's blood and probably will not themselves be realized.")

But Framingham, Kannel protested, was in the throes of trying to survive by shifting to a base in Boston University. It had neither the time nor staff to address the complexity of producing the interim data required.

Ultimately, inaction won the day. After a lifespan of seventeen years, the Pooling Project's final report in 1978 contained nothing about those more interesting but never-pursued questions. Instead it repeated the oft-cited relationship of the baseline "traditional" risk factors to coronary risk and thus seemed to bear out Kannel's wondering whether pooling was worth the effort, when it had reached only this conclusion:

> . . . the Pooling Project results presented here add further to the large body of epidemiologic data demonstrating that the relationships between the three major risk factors and premature atherosclerotic disease meet the criteria of consistency, strength, graded relationship, independence, temporal relationship and predictive capacity. . . . [They] also meet the criterion of coherence. Therefore these relationships are almost certainly cause-and-effect, i.e. etiologic, in nature (American Heart Association 1978, 266).

In fact, the AHA Pooling Project confirmed and rendered more powerful and generalizable the findings of many other studies, including international studies, that came to the fore during the long period taken to carry out the pooling analyses. It can be said, at least, to have "locked down" the evidence of the independent effect of the traditional risk factors, concluding that 70 percent of events were related to "suboptimal" risk factor levels, based on findings in the lowest quintile of multivariate risk.

The Pooling Project authors qualified their conclusions by suggesting that prediction would be improved by a knowledge of time trends and by recording more than one baseline measurement, as well as by collecting lifetime (diet and activity) data. But their report recommended unqualifiedly that the findings "may be applied with confidence in the practice of preventive medicine and public health to identify the more vulnerable segments of the population who are especially candidates for preventive management. Their widespread use should be consistently and vigorously encouraged" (ibid.).

The pooling data were made available, pre-publication, and became influential in deliberations of the Joint Commission Report (Wright and Frederickson 1970) and in design of NIH trials that arose in the

early 1970s. But it was finally only Framingham multivariate risk scores, not the Pooling Project estimates, that made up the widely circulated AHA risk-screening handbook (American Heart Association 1973). Framingham also provided the precise screening criteria NIH used at this period to screen and select, from a population of some 370,000 U.S. screenees, a high-risk group to participate in the NHLBI-sponsored Multiple Risk Factor Intervention Trial (MRFIT), a decision that produced an unusual-risk population.

Subsequent strategies of summarizing data (e.g. meta-analysis) have utilized the variability of results among multiple studies rather than the pooled values of those studies to arrive at causal inference and to quantify the contributions of risk factors to prediction. Present-day perspective on meta-analysis, an historic ideal for combining the data of similar studies to strengthen the evidence, comes from Curtis Meinert, clinical trials pioneer, who suggests that a plan to combine studies should exist, not at their completion, but from their outset:

> One thing I believe in firmly--but I'm in a minority--I think we ought to have planned meta-analyses, in other words, design studies [to be] amenable to later meta-analysis. It seems to me that when you're monitoring there ought to be certain classes of studies in which you are, by definition, pooling results . . . (Meinert 2002).

(The American Heart Association pooling group of the 1970s included co-principal investigators Felix Moore and Fred Epstein and members Henry Blackburn, John Chapman, Joseph Doyle, Alan Dyer, Tavia Gordon, William Kannel, Ancel Keys, Patricia McNamara, Oglesby Paul, Rick Shekelle, Jerry Stamler, and Henry Taylor. Len Cook was AHA administrative staffer and Thomas Karunas the AHA statistician.)