

## Field Methods for the New Discipline of CVD Epidemiology

### Contents

[Introduction](#)

[Clinical Methods Converted to Field Use](#)

[Specific Field Methods](#)

[Rose and Blackburn: The Making of WHO Technical Manual No. 56](#)

[The Electrocardiogram in Population Studies: The Minnesota Code](#)

[Indirect Blood-Pressure Measurement](#)

[Anthropometry: Measures of Body Mass and Body Build](#)

[Blood Lipids and Lipoproteins: A Call for Standards](#)

[Dietary Assessment and Food Analysis](#)

[Tobacco Exposure and Smoking Habit](#)

[Physical Activity and Fitness Measures](#)

[References](#) (opens a separate PDF with references for entire Methods section)

### Introduction

The need for valid, reliable, standard, and practical methods for field work in chronic disease epidemiology is obvious to investigators today, who apply them without imagining a time when every new study required reinvention of the wheel. But part of the uniqueness of CVD epidemiology lies in the situation confronting researchers who initiated studies in mid-20th century. Mainly clinical and laboratory experts without experience in epidemiological methods, they were for the first time addressing questions among populations and departing from familiar clinical environs and practices.

Tests of observer variation in diagnosis, for example, were unknown to clinicians of the time, and, when first attempted by the novice epidemiologists, caused great wonder and chagrin. Repeat observations became an essential ceremony of the medical researcher's baptism in, and conversion to epidemiology.

Some younger recruits in those early days made career marks with their assignments to translate clinical impressionism into repeatable, more objective recording of symptoms, signs, and diagnostic labels. They set out to devise improved ways to measure and classify the health history, symptoms, blood pressure, electrocardiograms, and laboratory tests.

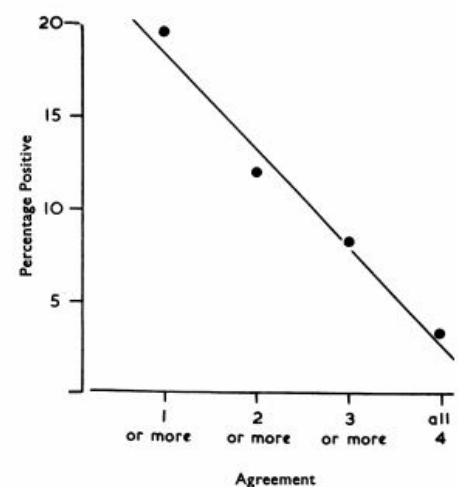


FIG. 2.—Percentage electrocardiograms read by four readers as compatible with coronary disease.

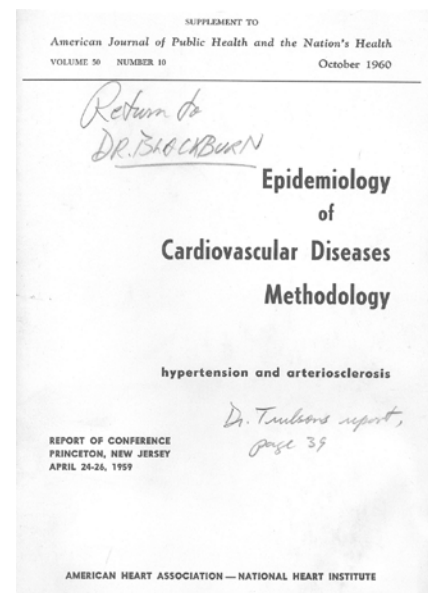
Here we offer summaries of some of the early developments converting clinical observations, measurements, and diagnoses to quantities applicable in field conditions, along with stories about those where we had direct experience. (*Henry Blackburn*)

## **Clinical Methods Converted to Field Use**

Central agencies that developed and encouraged sound field methods for the population studies launched at the beginnings of formal CVD epidemiology included the U.S. National Heart Institute (NHI), the American Heart Association (AHA), the World Health Organization (WHO), the London School of Hygiene and Tropical Medicine (LSHTM), and the Laboratory of Physiological Hygiene (LPH) at Minnesota. The AHA and WHO were particularly active in setting up continuing institutional mechanisms to foster improved methods.

The AHA Subcommittee on Criteria and Methods of the Community Service Committee, formed in 1960 and led by its chairman, Fred Epstein, effectively addressed standards for recording indirect blood-pressure (Weinstein and Epstein 1964). WHO, in Geneva, called methods conferences in the 1950s and in 1964, under Zdenek Fejfar, commissioned the first edition of *Cardiovascular Survey Methods* (published in 1968), an internationally circulated technical manual. It included guides for design, measurement, classification, and analysis of CVD surveys, with detailed forms and criteria and even a bare-bones survey philosophy.

These innovations in survey methods came after several “enforced” confrontations among U.S. investigators in a series of methods conferences sponsored by NIH and AHA in the 1950s, where many issues were resolved while others were expressly avoided. Even then, not all investigators were “on board” about the need for validation and standardization of clinical methods in field studies. Some were resistant to the very idea of a systematic approach to replace their traditional masterly impressionism. These conferences helped fix the agenda, nevertheless, and stimulated groups to address their methodological issues, while providing rationale, guidelines, and motivation for improved methods and for on-going quality control systems for data collection. (A series of down-to-earth, sleeves-rolled-up, hard-fought work sessions in the U.S. during the 1950s focused on population survey and laboratory methods and was held at the Arden House in upstate New York (Conference on the Epidemiology of Atherosclerosis and Hypertension) 1956; at Brookline, Massachusetts (Conference on Longitudinal Cardiovascular Studies) 1957; and in Princeton, New Jersey (Conference on Methodology in Epidemiological Studies of Cardiovascular Diseases 1959.)





In the long run, most of the new epidemiological research centers came around to testing and improving their methods, preparing standard operating manuals, and training staff to follow protocol. A few included blinded duplicate evaluations, while all participated eagerly in a national, then an international program for standardization and quality control of blood lipid determinations carried out by CDC.

## **Specific Field Methods**

When commonly-used clinical methods, such as history-taking and blood pressure and electrocardiographic recording, were first taken into the field for CVD studies in populations, they were soon found wanting. All required redefinition, reassembly, and codification, followed by focus on the apparatuses, techniques, and procedures themselves, along with the field conditions under which they were applied.

The early challenge was to render the clinical measurements more valid and reliable (repeatable), more objective and quantitative, with mutually exclusive classifications. But major new systems also were created for automated laboratory techniques and data management and analysis. Of the active early foci of methods development and testing, WHO and AHA became the continuing agencies to insist on and disseminate guidelines and to foster training.

### **Questionnaires were among the first needs**

An early challenge was the “simple” issue of recording medical history and examination data in efficient standard formularies for later conversion to numeric files. British pulmonologist Charles Fletcher’s “cough and spit” questionnaire used in the immediate post-war period was among the first “standard” survey instruments of chronic disease epidemiology (Fletcher et al. 1959).

(Charles Fletcher tagged on to the national program of chest X-ray screening for tuberculosis in post-war UK and pioneered field surveys of chronic bronchitis, the so-called “British disease.” His time-tested cough questionnaire began: “Do you cough at all when you get up or first thing in the morning?” (Fletcher et al. 1959).

In Norway, similar tuberculosis screening programs during the same period provided investigators with the cohorts eventually employed for prospective studies and clinical trials in cardiovascular diseases (Westlund and Nicolaysen 1972).

Fletcher’s questionnaire became a template for the chest pain-claudication questionnaire, called the London School of Hygiene Cardiovascular Questionnaire, a standard survey instrument expanded and tested thoroughly by Geoffrey Rose and colleagues (Rose 1962).

(Geoffrey Rose and colleagues at the LSHTM took the lead of Fletcher’s respiratory disease questionnaire to find questions that effectively identified angina pectoris on effort, myocardial

infarction, and intermittent claudication, utilizing the characteristics of ischemic pain. These were tested against clinical diagnosis, the electrocardiogram, and follow-up events. The lead-off question, followed by dissection into its detailed components, is: “Have you ever had any pain or discomfort in your chest?” (Rose, 1962).)

Fletcher was among the first to address the problems of standardizing diagnoses using clinical methods that included administered questionnaires. With both physician diagnoses and standard questionnaires, random and systematic errors were large, depending on the skill and form and order with which the questions were pursued, matters intensely considered from earliest times in chronic disease epidemiology (Payne 1951; Fletcher 1963; Rose and Blackburn 1968).

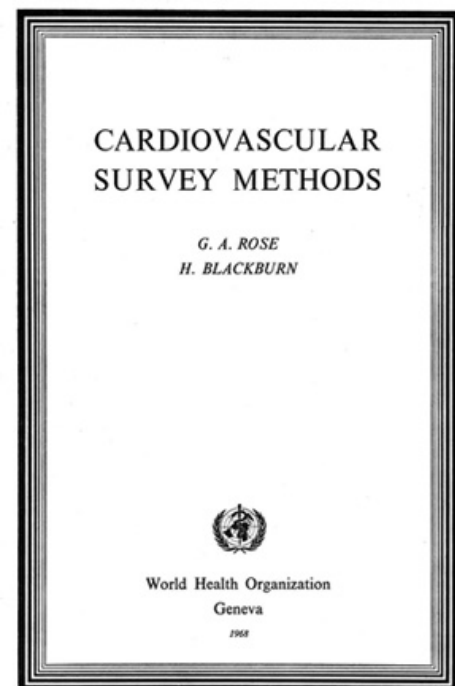
Meanwhile, parallel to the early evolution of survey methods and of guidelines for the conduct of epidemiological research, ethical issues were also being addressed; that is, issues of privacy, informed consent, and human rights, which soon began to affect all medical research, especially epidemiological.

## **Rose and Blackburn: The Making of WHO Technical Manual No. 56**

In late fall of 1964, during the period when WHO Cardiovascular Unit Director Zdenek Fejfar was recruiting “young chaps” to build a strong bank of standard field methods, he asked Geoffrey Rose and me (Henry Blackburn) to write a primer of CVD epidemiological methods. The resulting technical monograph, *Cardiovascular Survey Methods*, was intended to provide survey approaches, common forms, and diagnostic criteria for the flourishing new field. The manual came to be known as *Rose and Blackburn*.

During this season, when Lake Geneva was covered by unrelenting fog, Geoffrey and I worked together in an old League of Nations building, confined to a cell with bare walls and furnished only with two desks and two chairs, and assisted by a staff of one pallid English secretary. The following monastic protocol remains even today the favored WHO method for producing expert reports: first, prepare well the groundwork, with exchange of manuscripts and correspondence over many months, then incarcerate the experts in Geneva until they come up with a draft.

We worked well as a team. Geoffrey handled the “big picture,” the didactics, principles, and computations, while I concentrated on details of field methods, definitions, and criteria. In some areas where our expertise overlapped, we comfortably traded tasks. For example, Geoffrey took over





methods in the sections on chest pain, smoking habit, and respiratory symptom questionnaires and I wrote about principles of survey organization, operations, and recruitment.

We soon were able to decipher each other's handwriting, illegible in different ways, and then edited each other's drafts. I rarely marked up his crystalline copy. Geoffrey whittled away mightily and tactfully on mine until eventually I came to acquire some of his parsimony of language. Alex Burgess, a senior WHO reviewer of the monograph, remarked, for what it's worth, that he had never found, "save in classic sonnets, so much information concentrated in such limited space" (personal comm.)

Following our initial Geneva "retreat" of ten days to produce the first draft, and further labors by correspondence, the manual was subjected to several years of arcane, oppressive, maddening WHO bureaucratic procedure but was finally published in 1968, after which it was widely disseminated by WHO. The manual had a rebirth of application in the 1980s in preparations for the massive WHO MONICA surveillance program (Tunstall-Pedoe 2003). Subsequently it was revised in a second edition led by Ronald Prineas and Richard Gillum (Rose et al. 1982) and in a third edition by Russell Luepker and an international team (Luepker et al. 2004).

An anecdotal sequel to the writing of the draft of *Rose and Blackburn* in Geneva has to do with the order of authorship. In Geoffrey's mind, our effort for WHO was primarily to produce a manual of method and procedure. He had no doubt, therefore, that the order should be Blackburn and Rose, since my prior contributions to the field had been largely methodological and I had taken early initiative on the manual. Similarly, there was never any doubt in my mind--because of his major contributions to survey methods and outlining of epidemiological principles, plus the stamp of his crisp style and language on the whole monograph--that Rose *must* be the first-listed author.

We left Geneva still at an impasse on the matter. By correspondence thereafter we quietly lobbied Zdenek Fejfar, each insisting on the priority of the other for senior authorship. In this unique case of "Alphonse and Gaston," the overwhelming superiority of my argument won out. Typically, Geoffrey was dignified in defeat. He reluctantly but graciously accepted to be first author of *Rose and Blackburn*. (Henry Blackburn)

## **The Electrocardiogram in Population Studies: The Minnesota Code**

The electrocardiogram (ECG) records variations in electrical current over the body produced by the beating heart that provide information on cardiac anatomy and function essential to the diagnosis of cardiovascular diseases in their many manifestations. (Modern electrocardiography dates from the Nobel Prize work of Willem Einthoven from a physiology laboratory in the Netherlands in the early twentieth century. He developed the critical instrument, a string galvanometer, to amplify the tiny potential differences between points on the skin produced with each heartbeat. Clinical applications were made rapidly after the visit to Einthoven by Britain's leading clinical investigator, Thomas Lewis, who then led in clinical recognition of the ECG manifestations of cardiac ischemia, hypertrophy, conduction defects, and arrhythmias. It was these that allowed the eventual synthesis, by James



Herrick of Chicago in 1912, of the clinical syndrome of cardiac infarction *with survival*, which, within a decade or so, dramatically and widely enhanced the capability to diagnose coronary heart disease and to perceive it as a mounting epidemic (Burch and DePasquale 1990; Fye 1994).

For modern-era epidemiological studies, completely automated measurement and classification are available with NOVACODE, a system for processing analog and digital records developed by Rautaharju and colleagues in the EPICARE laboratory. (Rautaharju et al. 1990). The history of forty years of technical developments in applications of computerized ECG measurement and analysis is presented in a monograph by Pentti and Farida Rautaharju (Rautaharju and Rautaharju 2007).

But transfer of modern electrocardiography from the clinic to the field, and central analysis of the data, were not simple matters. Some of the early drama of these encounters is found in the following story of the Minnesota Code for electrocardiograms in population studies:

During the late 1940s and the 1950s, when most systematic population studies of heart disease began, it became obvious that diagnoses recorded on death certificates, or made by physicians in the field having different backgrounds of training, custom, and language, might easily result in spurious findings of population differences. Even independent diagnoses within studies, by specially trained physicians, were found to be highly variable. It was important to find standard, objective, and quantitative means to compare CVD rates.



The four "Minnesota Coders" (Rautaharju, Crow, Blackburn, & Prineas)

At Minnesota's Laboratory of Physiological Hygiene (LPH), Ancel Keys and Company were preparing to launch two major studies--the U.S. Railroad Employees Study and the Seven Countries Study--both of which would begin in 1957 and would involve large-scale measurement in populations for comparison of heart disease rates and risk. The need for comparability in clinical assessments led the group to consider use of the electrocardiogram as an objective measure. Keys assigned me (Henry Blackburn)--then a young colleague of Ernst Simonson in electrocardiology, and newly on Keys's staff as a research associate--the task of developing a classification system for the ECG, along with preparing and testing survey questionnaires, forms, and diagnostic criteria to translate clinical observations into quantities.

The electrocardiogram (ECG) indicates cardiac manifestations of greatest interest in epidemiology, that is, evidence of a heart attack, as scars of infarcted heart muscle (large Q waves), signs of ischemia or inadequate blood supply to the heart (displaced S-T segments and negative T-waves), increased muscle mass or hypertrophy (high amplitude R-waves and S-T segment depression), and disturbances of cardiac rhythm or conduction. The ECG seemed particularly promising for surveys as an objective



graphic record amenable to standard procedures of collection, measurement, and classification. It also was ideal because it was acceptable, painless, simple, and relatively inexpensive.

Hurdles were soon encountered, however, among reading ECGs taken in surveys. Physicians were found to differ widely in their blinded readings of a record, as did a cardiographer reading the same record at different times. Moreover, the literature held few tested, objective, accepted criteria for departures from the norm; only diagnostic labels of gross pattern-recognition of abnormalities (e.g. “evolving anterior infarction”). The early challenge for epidemiology was to come up with valid, repeatable, standard criteria and measurements so that prevalence or incidence rates of cardiac “events” in populations could be reliably assessed and compared.

Pentti Rautaharju and Sven Punsar from the Finnish population studies of Marti Karvonen, Gunnar Blomqvist from the hospital studies of Gunnar Björck in Sweden, and me (Henry Blackburn), all trainees resident in the LPH at Minnesota, began by compiling the few existing electrocardiographic criteria and sorting them into quantitative classes and rankings. We devised and validated new criteria based on data from specific cases and controls, and then assembled them with unambiguous descriptions, rank-ordered them by magnitude, and assigned them code numbers. The resulting classes were quantitative, mutually exclusive within classes, and directly relevant to common disease states. The codes described prominent Q-QS waves, or negative T-waves, etc. of specific magnitudes rather than giving them clinical labels of “probable or possible infarction, ischemia or hypertrophy.” Continuous measurements were recorded of a few relevant ECG amplitudes and intervals.

Finally, records from different living populations were classified and evaluated for “reasonableness” of the population distributions obtained, including sensitivity-specificity as applied to general populations and to patients. Test-retest reliability among the group improved gradually.

Between 1958 and 1959, early versions of the classification were circulated for comments from leading cardiologists and investigators involved in population studies. Ian Higgins at Cardiff, Fred Epstein in Michigan, and Geoffrey Rose in London became the more active and enthusiastic collaborators in evaluating the criteria and suggesting revisions. Higgins recalled later that it was Aubrey Kagan of London who, while on temporary assignment to WHO in Geneva in the late 1950s, dubbed the embryonic system the “Minnesota Code” (Higgins 2003). The name stuck.

Despite this use of unambiguous and quantitative criteria, with mutual awareness of standard procedure, coding variation among our staff remained great. Clear definitions and objective criteria, therefore, failed to guarantee similar application at different times or by different observers. Consequently, we gave greater attention to the conditions of electrocardiographic recording and measurement, and then to coding procedure detail, particularly rules for dealing with ambiguities and observer differences. Magnifying devices were devised for assessing Q-wave duration and other intervals, which improved agreement for classifications around the borderlines of criteria cut-points. New rules accounted for pattern variation among heart beats, and procedure was further shored up by quality-control with duplicate, independent readings and adjudications by a third party.

Nevertheless, after publication of the Minnesota Code in *Circulation* in 1960 (Blackburn et al. 1960), a major burden of coding continued to weigh heavily on the small LPH post-doc group. Increasing numbers of ECGs arrived in Minnesota for standard coding from multiple survey fields. The high level of interest and collegiality among us physicians-in-training during the developmental phase of the system was replaced by boredom with the tedious routine of an incessant and “unpaid” chore.

### **Clerical Coding**

At about the same time, and independently, Rose and Prineas in London and those of us in Minnesota hit upon the idea of employing technicians to do the coding. Rose had approached the issue logically and began training London School clerks. The Minnesota resolution was more serendipitous, arising during a special experience with ECG coding at home in the summer of 1962.

That summer, I was solicited by the National Health Examination Survey of the National Center for Health Statistics to apply the Minnesota Code to 6,000 electrocardiograms from a representative U.S. sample. This I agreed to do as a summertime avocation. It offered the meager reward of 25 cents a record, but seemed otherwise a worthwhile undertaking “for science,” and it indirectly acknowledged the Minnesota Code as a new national “standard.” In any case, the project seemed worthwhile--until the actual shipment of paper electrocardiograms arrived. They came as unmounted strips, stowed in tiny cardboard cubicles as tightly wrapped cylinders some two to three meters long. Each record had to be teased from its case, unrolled, held flat, read, measured, codified, tabulated, rerolled, and reinserted in the packing. The entire process took many times longer than reading, classification, and tabulation alone.



*First class in clerical Minnesota coding of ECGs for population studies, 1965*

With the inducement of one cent per record, I was able to enlist the enthusiastic aid of two alert, non-technical persons--my 6- and 7-year-old daughters Heidi and Katia--who were close at hand that summer. While one would extract and unroll records, the other would hold them flat until I coded them. Then together they would reroll and refile the tapes.

The girls’ summer wage came to 6,000 cents each, and mine to 6,000 times 23 cents. Their work reduced mine substantially; plus the fact that father and daughters had the pleasure of working together much of the summer.

After only a few days’ experience, my young assistants could identify P-, Q-, R-, S-, and T-waves. After a little more time, they were eager to point out to their father when a P-wave or a T-wave was “upside down” or when a Q-wave was “big and fat” or an R-wave was “too tall,” and so on. After a few more





days, by which time they had spontaneously associated the code numbers I assigned with tall, flat, inverted, or prolonged waves, it was clear that they were capable of becoming excellent ECG coders!

In those days, it wasn't unusual for LPH staff to spend precious time reading up to 30,000 routine records a year from the field. The summer experience convinced me that it was time to change the protocol. In fall of 1962, I began recruiting part-time university students and instructing them in the coding procedure. Subsequently, over the decades, many dozens of student coder and civil service alumni have provided the coding service to national and international population studies, while at the same time putting themselves through school.

Geoffrey Rose and Ronald Prineas, meanwhile, had set up a successful clerical coding system in London for several U.K. and continental studies. Soon, lay coders in Minnesota, London, Glasgow, and Prague were providing central ECG reading services to epidemiological studies and trials worldwide.

Eventually, much of this service came to be provided by automated computer programs analyzing both analog and digital ECG data, using both the Minnesota Code and NOVACODE, in the Winston-Salem laboratory of Pentti Rautaharju, and in other centers. (*Henry Blackburn*)

## **Indirect Blood-Pressure Measurement**

The historic drama of Reverend Hale's experiment in direct measurement of arterial pressure, by height of the blood column in a glass tube inserted into the neck or the crural artery of his mares, was later matched, more delicately, with sterile catheters in human arteries. In the early 20<sup>th</sup> century, the mercury manometer indicated pressure indirectly in an artery-occluding arm cuff by the appearance of the pulse by palpation or auscultation. This became the standard clinical tool for indirect sphygmomanometry and the one later applied in population studies. Its validity representing intra-arterial pressure was well demonstrated.

But the chief methodological issue for epidemiology was repeatability and bias in recognition of the sounds over the brachial artery (sounds labeled by the Russian, Korotkov), which represented the systolic and diastolic pressure levels in the vessel. Other issues were the posture and metabolic state of the subject, particularly the relationship of the arm to the heart level, and how to deal with outsized, scrawny or large, upper arms in the workings of the cuff. And finally, the more important issue, what pressure, systolic, or diastolic 4<sup>th</sup> or 5<sup>th</sup> phase, is the more "valid," representative or predictor?

Many clinicians felt that indirect arterial pressure should be measured serially and only the lowest pressure recorded; others preferred an average of several recordings. Field investigators focused on standard conditions and the optimal number of repetitions to smooth out measurement variability. Eventually, plots of precision against the number of repetitions indicated that most of the information obtained from an infinite number of blood pressure determinations was achieved with three readings.



Moreover, it was found that greater physiologic stability of the indirect pressure could be achieved by allowing a few minutes to pass following *any* change in posture, whether seated or supine. Conditions might be further standardized for temperature and lighting and noise level of the room, and for raiment chosen to avoid constraint of the artery, and by appropriately standardizing the time interval after changing posture, and after exercising, smoking, or eating. In fact, the variables involved to improve accuracy and reliability of indirect arterial blood pressure seemed endless.

In systematic study of blood pressure measurements it was found early on that observers tend to perceive and record levels of pressure with certain preferences for digits, often a bias toward zeros. Various approaches were sought to reduce this bias, including training for reading to the nearest 2 millimeters of mercury or, more effectively, using a random zero “muddler.” One such device was developed by Geoffrey Rose at the LSHTM, with which an eccentric cog was spun between recordings to reset the baseline level randomly and blindly (Curb et al. 1983). A clumsy box, it was balefully accepted and used by technicians in the field.

In usual practice, observer variability and digit preference turned out to be so resistant to training that great efforts were made to develop wholly automatic devices free of such artifacts yet mechanically reliable and giving valid data. Despite hundreds of careful comparative observations, no generally accepted device has come to be considered for field use in epidemiology over the well-trained blood pressure technician using a well-maintained mercury manometer. Nevertheless, automated apparatuses have replaced mercury manometers, mainly because the latter are no longer made or repaired by the original makers.

Ongoing problems of blood pressure measurement for surveys included identifying outliers among blood pressure technicians, retraining them, or, when found incorrigible, transferring them to another duty. Absent a laboratory standard, frequent checks were made to determine random and systematic bias and time trends of pressure recordings by individual technicians and by entire survey teams, then applying adjustments to any systematic errors found. Fortunately, a few investigators find these mundane methodological questions as fascinating as they are essential. Despite these attempts, and agreement on repeated standard pressure recordings in the field, systematic error in blood pressure survey data is a persistent problem for dissection.

## **Anthropometry: Measures of Body Mass and Body Build**

In not-so-older times, the nutritional state and economic value of slaves on the paddock were assessed by pinching up the skin and fat layer at various body sites. The panniculus of subcutaneous fat, as well as the muscle mass below it, gave witness to the health and prior treatment of the slave. Variations of that crude process are still used today in cattle markets. But the traditional finger-pinch procedure was quantified by students of *Arbeitsphysiologie* in pre-war Germany and by post-war U.S. physiologists to assess body composition in proportion of fat.



To study the role of obesity and body build in the risk of CVD, first-generation epidemiological surveys began to use subcutaneous skin-fold measurement by caliper as a surrogate for obesity (as distinguished from overweight by weight for height, or BMI). The standard pinch instrument became a Harpenden-type, constant-pressure caliper in the hands of trained observers, in which skinfold thickness as a measure of obesity was measured at subscapular, triceps, and abdominal sites and often summed.

A series of investigations of body composition used correlations of multiple measures to estimate body fat. In-air and underwater weighing (with correction for residual lung volume) was made with apparatus still in use in Minnesota's Laboratory of Physiological Hygiene. Several publications on these measurement techniques became classics in physical anthropometry and nutritional assessment from the same era as the beginnings of CVD epidemiology (Keys and Brozek 1953; Brozek and Keys 1956).

Later efforts to define the CVD risk associated with obesity were stimulated by findings about abdominal versus truncal adiposity. Girth (a marker of omental adiposity), and the ratio of waist to hip circumference, were found superior to body mass (BMI) as a predictor of future CVD risk. This field advanced through specialized applications of impedance plethysmography or Doppler sonograms to characterize the subcutaneous abdominal panniculus, or especially the volume of omental fat, which is increasingly regarded as an endocrine and possibly toxin-producing organ.

For the simpler measure of body mass, Alphonse Quetelet, a Belgian astronomer of the early nineteenth century, developed the idea of the statistically "average man," and among his many insights calculated an index that is now called the Body Mass Index (BMI) by which he related actual to an ideal weight for height (weight in Kg/height in meters squared) (Winkelstein 2012). More than a century later, BMI became the standard method of reporting relative weight, superseding the Metropolitan Life Insurance Tables (percentile values of an actuarially ideal weight) used in the 1930s and '40s.

The physiologist's penchant for metabolic relevance and for precision in measurement, particularly in the 1940s and '50s, led to all sorts of other explorations of body composition, including dilution methods (plasma distribution of tritiated water or tissue concentrations of potassium ions). Fortunately, their correlations with simple skin-folds are such that the latter have been sufficient for field survey use. In any case, the technology and technique have improved even as the generation of physiologists obsessed with precise physical anthropometry has died out.

In 1956, a landmark conference established the recommended measures and techniques that remain standards today, at least for an agreed minimum number of measurements that appear essential for assessment of nutriture in adults. They include body weight, body dimensions, adipose tissue, and muscular tissue (Keys 1956). In 1971, Ancel Keys and colleagues at Minnesota put Quetelet's index at the forefront of metrics today, renaming it the Body Mass Index, and providing evidence of its modest superiority at categorizing obesity, independent of height, and its simplicity of computation, where BMI remains today (Blackburn and Jacobs, 2014) (In his last formal publication, Ancel Keys, at age 95, addressed the prediction of survival in relation to BMI and its change during the middle-age years of

men (Keys 2000). His greatest frustration in his final years was the inability to complete a similar study of body linearity-laterality, that is, “male-female” form measured as biachromial-bicristal diameters, in regard to survival. These anthropometric data rest in a Minnesota archive awaiting the curiosity and enterprise of others.)

## **Blood Lipids and Lipoproteins: A Call for Standards**

Field operations in epidemiological surveys are rarely free of complications, usually unforeseen. For the first round of Seven Countries Study surveys, which began in Yugoslavia in 1958, a simple method of collecting and shipping serum cholesterol samples was devised by Joseph Anderson and Ancel Keys for use under primitive field conditions (Anderson and Keys 1956). Blood samples were centrifuged on the spot and the serum separated and pipetted in four 0.1 milliliter aliquots onto Watzman #2 filter paper, labeled with an identity in pencil, and hung to dry. The dried strips were then cut and duplicates analyzed in the laboratory in Zagreb for local use, while the other two were placed in separate glassine envelopes and sent air mail for central analysis at the University of Minnesota.



*Thomas Strasser processing serum samples in the field,  
Dalmatia 1958*

Early in the survey, Zagreb technicians found inexplicable differences in the duplicates from the field laboratory. After careful scrutiny, it was attributed to lipid in the fly-specks deposited on one or the other of the drying filter paper spots! When field conditions could not be upgraded with window screens, the problem was corrected in large part by engaging a youth to fan the flies away.

Automated technology for field surveys would not arrive until the 1970s, when the measurement of blood cholesterol and blood lipid-lipoprotein became a major laboratory industry, bringing with it an expanding academic community of “lipidologists” and their far-flung studies (e.g. NHLI Lipid Research Centers).

The larger and more fundamental issue about the comparative predictive adequacy of total serum cholesterol vs. lipoprotein patterns had been settled, if only for a time, by the Cooperative Lipid-Lipoprotein Study begun in 1952. That complex and fractious study addressed the predictive power of simple serum cholesterol measurement against the complex, costly, but functionally relevant ultracentrifugal lipoprotein fractionation. They were finally considered equal in predictive power, with Gofman writing a minority report (Gofman et al. 1956). For a time, LP fractions were abandoned in many surveys.

## **The NIH/CDC Lipid Standardization Laboratory**

In the midst of changing technology and an overwhelming influx of epidemiological data, many researchers felt that, no matter which system of lipid measurement ultimately prevailed, there was an immediate need for standardization of total serum cholesterol determinations both within and among laboratories. Gerald Cooper, a clinical pathologist at the CDC in the early 1960s, recalled that the appeal for standards arose insistently every year at the annual meetings on CVD epidemiology in Chicago. When NHI director James

Watt found that epidemiological grant proposals were failing for lack of lipid-measurement standards, he called on Cooper to develop the procedural, technological, and statistical models for quality-control (Cooper 2005).



*Gerald Cooper, founder and long-term director of the CDC Lipid Reference Laboratory, joins Parvin Waymack, current chief of the Laboratory, and Shelton Stribling, researcher, in the lab.*

Soon Cooper's CDC laboratory was serving numerous NIH-sponsored grants by monitoring their study's laboratory performance. Then, at a WHO Meeting in Geneva in the late 1960s, Watt spontaneously volunteered CDC services to develop international standardization of lipid laboratories. WHO promptly made a formal request for this in a letter to Cooper, who remembered making this reply:

I wrote back and said we were working on a national program. Let us [first] work up our problems in it and then we'll be glad to do it [for you].

Boy, right back came a letter to the Surgeon General saying that I didn't want to work with them! I'm telling you, the next day after CDC got the letter I was in Geneva! (ibid.)

The main stumbling block in setting up an international program, Cooper realized, would be a lack of specific funding. "So what we decided to do was to superimpose the international program . . . on top of our national NHI/CDC program and just go on and do it" (ibid.).

Gerald Cooper estimated that eventually forty-five international laboratories were standardized under the WHO collaborating center. He had skillfully found ways to provide an international outreach for standardization of lipid laboratories within a limited budget and a rigid U.S. bureaucracy. It functions today as a major service to prevention studies worldwide.

After Cooper, who remained the central laboratory's chief consultant, died in 2009, the National Heart, Lung, and Blood Institute (NHLBI) removed its support for this major contribution to international epidemiology, which support shifted wholly to CDC.

The engaging history of laboratory lipid measurements is treated in greater and fascinating detail in Daniel Steinberg's series on the "cholesterol controversy" in the *Journal of Lipid Research* (Steinberg 2004; 2005; 2006). The need eventually arose again to measure the lipoprotein fractions described by





John Gofman, including LDL, HDL, and VLDL, due to the evidence from several disciplines of their different functional roles and predictive import. The practical solution for large studies was offered by the Friedewald Equation, as treated in the following story told by Friedewald himself.

### **Origin of the Friedewald Equation for Estimating LDL**

In the late 1960s, when knowledge emerged of the different function and pathogenic importance of lipoprotein (LP) subfractions, as LDL, HDL, and VLDL cholesterol, a measure of the major “pathogen,” LDL, was needed for risk assessment and preventive practice; one that was less costly and more accessible than the Gofman ultracentrifugal fractions. The Atherosclerosis Index, developed by John Gofman and tested in Framingham data, required measurement of total, LDL, and HDL cholesterol in serum or plasma with computation of the ratio of total or LDL to HDL cholesterol. The chemical process for measuring total and HDL cholesterol were both straightforward. LDL was another matter.

A novel and valid, if rough-and-ready indirect estimate of LDL cholesterol was found at the National Heart and Lung Institute (NHLI, later NHLBI) more or less serendipitously, according to former NHLBI staffer, William Friedewald:

This was in 1969 going into '70, and Bob Levy [lipid researcher at NHLI] had this idea that there was a stoichiometry, an actual necessary relationship, between triglycerides and cholesterol in the VLDL portion . . . in a constant 5 to 1 ratio, so he thought. A statistician there, to whom Levy had given the problem, decided that he couldn't make it work out with VLDL triglyceride divided by 5. The correlation [of VLDL divided by 5] with total plasma cholesterol wasn't that good; so Bob gave it to me and said, “I think there's really something here.”

I said, “We're really trying to measure LDL; we don't really care about VLDL cholesterol, so let's look at that.” Although there was a fair amount of error in the estimate of VLDL, because VLDL cholesterol is such a small percentage of the total [cholesterol] it didn't make a big difference. When, in fact, we used [total plasma] triglycerides divided by 5 we got a really strong correlation between LDL [cholesterol] estimated by the equation and that measured [by ultracentrifuge].

Bob said, “Why don't you write it up, then?” So my name ended up in first position on the paper.

I saw Don Frederickson [former director of NIH] about five or six years ago and he said, “Out of all that stuff that we did [in lipoprotein classification], the only thing that seems to have survived really is the Friedewald Equation.”

It should, however, have been the Levy Equation! (Friedewald 2003).

The Friedewald equation is still used for the routine laboratory estimate of LDL cholesterol because of the constancy of cholesterol content in plasma triglycerides and the relative ease of measurement of

VLDL, HDL, and total cholesterol; at least, that is, in chylomicron-free (fasting) plasma and among “normals” and “run-of-the-mill” dyslipidemias.

## **Dietary Assessment and Food Analysis**

From the outset, measurement of diet components was a central need in the testing of the role of diet in atherosclerosis, the diet-lipid-heart disease hypothesis. The U.S. Department of Agriculture had reported on the net disappearance of foods at the national level since 1909. Many countries collect these sorts of data, which are published in food balance sheets of the Food and Agriculture Organization of the United Nations. Ancel Keys, for example, prepared the first ecologic correlations of diet and cardiac deaths from these tables in six cohorts for his 1953 Mt. Sinai article that was later vilified by Yerushalmy and Hilleboe (Yerushalmy and Hilleboe. 1957; Blackburn and Labarthe, 2012).



*Processing aliquot food samples in the field for chemical analysis centrally, Serbia 1973*

These periodic international reports give overall values for classes of nutrients consumed per capita and do not pretend to be dietary data; they provide nothing about the distribution of the food consumed in relation to age, sex, body size, or caloric needs. As an average (median) figure, half the people would be over-consumers and the other half short on rations, so that the method provides little information of interest to nutritionists or physicians.

Nevertheless, certain public health assumptions may be derived from them. For example, when per capita food consumption decreases over the years in which average body weight rises or does not change, this is an ecologic indication that energy expenditure is diminishing and that attention be given to physical inactivity as a social factor in obesity.

Keys looked at these data on disappearance of food fats in respect to heart disease mortality rates as one piece of many bits of evidence brought together in his early arguments about a diet-heart hypothesis. Meanwhile, back in Minnesota, he initiated rigorous dietary survey methods (a seven-day journal compared with home food weighing) to characterize individual and group patterns in the Minnesota Business and Professional Men Study, his prospective epidemiological investigation beginning in 1947. He also called attention to the need to collect survey data on individuals well prior to coronary events. In surveys based on actual records of the subjects' intake, supplemented by trained interviewers' queries, the agreement of repetitions of the diet recall methods and the handier food frequency methods is generally “acceptable,” but ascertainment by different observers falls to unacceptable levels ( $r = .1$  to  $.3$  agreement). Thus, all methods based on memory, while useful for



classifying groups, came to be considered unreliable for characterizing individuals unless based on multiple repetitions to reduce variance (Keys, 1979).

Twenty-four-hour recall is a frequently used method in CVD epidemiology and generally under-reports food intake. In the early 1950s, senior government nutritionist Sadie Adelson came out from USDA to supervise collection of a detailed seven-day diet history in 123 participants in Keys's Business and Professional Men Study. The seven-day recalls, aided by the men's wives and coaching by nutritionists, were compared with daily patterns based on weighed food aliquots of foods consumed in the home. The group comparisons agreed within 1 to 2 percent. But for individuals, only a third of the meat and milk intake recalls were within 20 percent of the recorded-weighed amounts and only half of the recalls of fats and oils and fruits and vegetables were within 20 percent of those amounts. And this was the experience among highly intelligent, unusual men who had been cooperating in a long-term study for some years (Adelson and Keys 1962). The situation has not changed. The partial and difficult solution is multiple repeat individual assessments to reduce variance (Liu et al. 1978).

### **Food Frequency Questionnaires**

Beyond the reliability and validity of dietary survey methods is the issue of translation of their raw data into nutrients consumed, the composition of which depends on the season, production, storage, preparation, and packaging of the foods. Thus, nutrient composition of foods in one country cannot be applied to those in another. Regional food composition tables must be adjusted to the whole situation and that depends on chemical analyses of local foods, which, in turn, are also in constant flux.

Within families, the composition of diet varies according to the work level and customs of the individuals, particularly the head of the household. Dietary variation within the individual is the dominant factor in risk factor analytical strategy, among distributions of individual intake in nutrient proportions of calories. That position is then correlated with some health phenomenon such as relative body weight, blood pressure, or the frequency of coronary events, which depends on the stability of the individual in his distribution at any given time and over time.

As Keys wrote, "Obviously the current diet, even if accurately estimated, may be rather irrelevant. Can it be shown that it is in fact representative of the long-term diet? There are essentially no data on this point for individuals" (Keys 1979, 16).

Moreover, all these methods fail to take into account the effect of the survey itself on what subjects eat during the survey and on their reporting of it, which reflect their image to the interviewer. The reported consumption becomes unrepresentative of the person's diet.

Keys rather weakly concluded his formal review of diet survey methods thusly: "All that can be said is that it is of the utmost importance to persuade the subjects to make absolutely no changes from what they will be eating in the absence of the intrusion on their lives of the survey" (ibid., 20). Bonne chance!



Today, little has improved since these beginnings except for the calculation of the optimal numbers of recalls needed to characterize individuals usefully (Liu et al. 1978), and the development of blood, tissue, and urine markers of nutrients, vitamin, and electrolyte consumption (Stamler et al. 2003). Emphasis is increasingly on the relevance to CVD and to health of particular foods, food groups, and eating patterns rather than on deconstruction of human diets into nutrients. Much research also focuses on the rapid trends in consumption of calories, and on calorie density, the ever-diminishing cost of ever-larger portion sizes of “fast foods,” that is, sugary, salty, fatty food and drink.

## **Tobacco Exposure and Smoking Habit**

Problems in the measurement of tobacco exposure were encountered early in modern chronic disease epidemiology. Paradoxically, despite a twenty-fold greater risk of lung cancer in cigarette smokers among subjects of the classic British Doctors Study by Doll and Hill, those who reported they inhaled had a lower risk than reported non-inhalers.

This purportedly led Doll to exclaim: “There *must* be a measurement finding!”—that is, an artifact.

In the 60-odd years since this finding the trend has steadily been toward more objective measures reflecting valid exposure to mainstream smoke, and with greater interest in effects of environmental tobacco smoke (ETS) or passive smoking, to lower concentrations of smoke products. Because of lower participation of smokers allowing biomedical markers, the search nevertheless goes on for more valid self-report survey measures and models.

More detailed questionnaires were first attempted to assess the “active” smoker, in which some counted cigarette-butt lengths and empty packs collected from survey subjects, all of which explained more variance than the reported number of cigarettes a day, alone. But quantitative measures began in earnest in the 1960s with exhaled carbon monoxide, assuming blood-breath equilibrium, followed by plasma carboxyhemoglobin levels, then saliva, urine, or plasma thiocyanate level (a metabolic product of cyanide in the smoke), then saliva, urine, and blood cotinine (a more readily measured metabolite of nicotine), then the addictive agent and more specific smoke component, nicotine itself, and minor tobacco alkaloids, again in saliva, urine, or plasma.

The short half-life of nicotine and its markers, and of CO and carboxyhemoglobin creates great variability. Thiocyanate in body fluids, in turn, with a longer half-life after exposure, is importantly confounded with food consumption of produce rich in thiocyanates. Thus, total cotinine and its glucuronides in plasma has become the more utilized survey method because of its representation of nicotine exposure and its half-life up to a day.

More recently, hair and toenail nicotine biomarkers, mainly cotinine, are found to provide long-term exposure information, smoothing out the daily variations in smoking due to long half-life and presumably representing the overall “burden of tobacco exposure” over recent months (Al-Delaimy and Willette 2008).



## **Physical Activity and Fitness Measures**

Jeremy Morris of London first compared occupations differing obviously in activity levels, then progressed to the study of individuals based on their questionnaire responses about leisure-time exercise, which questionnaire he also first developed. Paffenbarger and others soon began to seek insights into the relation of physical activity and CVD risk with innovative study designs using simple “instruments” such as questions about participation in sports (Paffenbarger, Wing, and Hyde 1978). Similarly, Robert Bruce early began asking about breathlessness or sweating during exercise (Bruce, Kusumi, and Kosmer 1973).

But these instruments were worrisome, as Ralph Paffenbarger bemoaned in our 2001 interview “The questionnaire *is* very important. How it was prepared, after trials and tribulations; how it was redesigned; where the questionnaire is put at the time of interview, or even the order of the questions, [both] are very important. Boy, you learn the hard way! All the mistakes we made! Trial and reversal and try again--and it still doesn’t work!” (Paffenbarger 2001).

Henry Taylor at the Minnesota Laboratory of Physiological Hygiene began the tedious development of an hours-long questionnaire that detailed dozens of activities “recently engaged in” (Taylor et al. 1978), while Jerry Morris created equally monstrous questionnaires about activity carried out on a given work day and holiday (Morris et al. 1973). They both computed continuous scores based on energy expended in physical activities *outside* of work. They assumed the future would see a universal diminution of occupational activity among workers in industrial societies.

These increasingly complex evaluations of an individual’s activity habits had been necessitated by the pioneers’ finding of selection bias in comparative studies of occupations. Body build, other risk-factor levels, psychology, and living habits, it turned out, are all associated with the choice of and persistence in a particular job, which, in turn, has an independent influence on the risk and course of cardiovascular disease. Thus, the investigator’s problem was to assess whether differences found in coronary disease risk by activity class were due to these other biologic and socio-cultural factors rather than to habitual physical activity level itself. They found it difficult to separate the inter-related effects.

The early investigators using questionnaire methods tended to get discouraged over the psychic energy spent and the weak results achieved in developing, testing, scoring, and analyzing such cumbersome indicators of personal behavior. They turned gradually from the late 1960s, and more recently almost entirely, to two methods: objective, dynamic, quantitative reflections of habitual activity (e.g. accelerometers) and assessments of “fitness” (the metabolically relevant resultant of habitual exercise) by standard graded exercise tests of work capacity or physiological responses like heart rate and oxygen consumption. (Validity testing of single and multiple-axial accelerometers has provided a guideline to this non-invasive approach now widely applied in epidemiological studies of physical activity and energy expenditure in children and in elderly and sedentary populations) (Westerterp 1999).





## **Fitness and Graded Work Performance Tests**

Applied physiologists from early in the twentieth century posed the idea that human function might best be studied realistically and more revealingly under the demands of stressful activity than in the resting state. Physiologists and a few physicians in the U.S. also came to that idea early, including Arthur Masters, Manhattan cardiologist, with his 1920s Two-Step Test, and Robert Bruce, Seattle cardiologist, in the 1960s, with a particular graded treadmill protocol bearing his name and now a clinical standard for induction of ECG signs of cardiac ischemia (Bruce, Kusumi, and Kosmer 1973).

Fitness is variously characterized as maximal work capacity or maximal oxygen consumption (Max  $\text{VO}_2$ ); or the maximal heart rate or time achieved on the treadmill during a standardized graded stress test. In addition to being more quantitative and more reproducible, these tests are also closer to the biological phenomena that both reflect and depend on the nature of habitual activity and its functional effects.

Whatever the case, the measurement of work capacity has substantially increased the understanding of the role of physical activity in CVD and other health risks in affluent industrial society (Blair et al. 1996; Bruce, Kusumi, and Kosmer 1973; Kornitzer et al. 1980).

The more recent methodological issue for CVD epidemiology has been a purist concern to measure work capacity in its maximal state, as maximal oxygen consumption (defined as: oxygen intake during intense exercise in relation to maximal work capacity and in a physiologically steady state). This difficult ideal--achieving a steady-state "Max  $\text{VO}_2$ "--has been replaced by a more pragmatic goal: performance in a standard but submaximal workload, sufficiently stressful that it reliably predicts maximal capacity. Such a measure is used primarily for screening for cardiac ischemia or for characterization of fitness of healthy populations. The focus for epidemiology and for the clinic is a standard exercise test that is accessible, economical, practical, sensitive, and safe in eliciting ischemic changes in the ECG, either as a predictor of risk in cohort studies or as an intermediate endpoint in preventive or therapeutic trials.

## **Population Comparisons**

Even the field imposition of standard graded workload testing does not necessarily provide valid comparisons of fitness and disease among cultures with different work habits or perceptions about work and workload. For example, Finnish men in the Seven Countries Study surveys appear genuinely challenged by the exercise tolerance test imposed in its surveys. But they appear to subjectively grade a given effort as consistently lower, on a scale of *perceived* exertion, than do men of Mediterranean populations, as noted in a journal from the 1964 Seven Countries Study survey in Finland:

The Finnish men appear genuinely challenged by our exercise tolerance test, 'giving it their all.' Yet, in responding to Alan Barry's scale of perceived exertion they appear to us to grade their effort consistently lower here than do participants in other Seven Countries surveys. On a scale of 1 to 15, the same level of activity perceived as an 8 by the Finns is more like a 13 for the



Italians. Admittedly, these comparisons have not yet been made systematically (Blackburn 1995, pg.70-71).

Similarly, with pulmonary function tests, the Finn technicians were more vigorous and the participants more responsive with “all-out” expiratory efforts. As Geoffrey Rose found, and as was confirmed using his chest-pain questionnaire in regions of the Seven Countries Study surveys, sizable cultural differences exist in responses among men--differences unrelated to the population burden of coronary disease.

Thus, even carefully standardized comparative population studies do not resolve *all* fundamental issues about physical performance as a causal influence in disease risk, or about the relative roles of “nature versus nurture.” The more repeatable and presumably valid measures of the effect of habitual exercise on actual performance and on the level of risk factors have nevertheless enhanced understanding of the role of physical activity in disease, without entirely eliminating the effect of “constitution,” or “natural athletic endowment.” (*Henry Blackburn* )