Clinical Research Databases—A
Historical Review

Morris F. Collen, M.D.

The increasing importance of computer-stored databases for clinical research prompted a historical review of their evolution over the past three decades. The special problems associated with the computer processing of clinical research data were reviewed, and the various types of clinical research registers and databases were described.

INTRODUCTION

In order to improve their knowledge of disease, clinicians have always followed the practice of studying their patients and the medical records documenting their care. The process of manually collecting, storing, and retrieving data from many traditional paper-based patient records was always very cumbersome and time consuming. In the 1960s, with the advent of computer applications to medicine, this process was dramatically changed. This review describes the evolution from the 1960s through the 1980s of computer-stored databases for the purpose of supporting clinical research; and how different investigators dealt with the problems of entering, storing, and retrieving their research data for their various clinical specialties.

RESEARCH DATA PROCESSING

Research Data Entry

The data entered into a clinical research database either were collected as part of the patient care process or were abstracted from patient medical records. The research data were then entered and stored in computer databases organized and operated by a database management system, a specialized software program for entering and retrieving the data.
For long-term chronic diseases, an important functional requirement of a clinical database was the capability of record linkage; that is, the ability to collate data for a patient which had been collected during several separate episodes of care. In addition, it was important to collect selected relevant data from many patient records and to collate the data by disease or by procedure in order to make the database useful to investigators. If the data were to be collected from multiple medical centers, it was necessary to share a common patient identifier, to agree upon a uniform mode for data entry, and to assure continuing patient confidentiality. Thus a major requirement for a clinical database was that data collected from various sources had to be sufficiently standardized to permit its aggregation and valid data combinations.

To achieve some standardization of the data collected from different sources, Pryor (1982) "... concluded that a database used for research should incorporate a coded data entry format." By the early 1980s, a group at the University of California in San Francisco proceeded to develop a system which automatically encoded patient data from the medical record.

... most databanks dedicated to the storage and analysis of clinical research data attempt to capture data in a fully coded form. This process is extremely laborious. ... It is difficult to maintain the accuracy and completeness of the data. ... The alternative to this approach is to capture data electronically as it occurs naturally in a clinical practice, and have a computer do the work of coding. ... A computer program has been developed which extracts fully coded data from the 'partially coded' patient information gathered by the Summary Time Oriented Record (STOR) system—an ambulatory record system. ... The first machine encoded data was passed from the STOR system to the ARAMIS data bank in July 1983. This first bolus included the data of 224 patients with systemic lupus erythematosus, covering 30 months of clinic visits. (Whiting-O'Keefe, 1983)

**Research Data Storage**

The high cost of computer storage in the 1960s restricted collections of patient data for clinical research to relatively simple registers which contained only patient descriptors (such as age, sex, and case codes which maintained patient confidentiality), diagnoses, and a limited amount of clinical data. In the 1970s and 80s, as computer storage became cheaper, increasing amounts of patient record data were stored in computer data banks or databases, specifically for the purpose of facilitating clinical research. These databases proliferated in the 1970s and became common in the 1980s. The great utility of computer databases resided in their capacity for storing huge volumes of information collected from large numbers of patients, and for permitting multiple researchers to rapidly search and retrieve their selected items for study.

Whereas a single patient's medical record contained all the information collected about all of the problems for one patient, a clinical database established for research purposes contained selected information from all the available patients' records for a specific medical problem or technology under investigation. Thus, clinical research databases tended to be disease- or population-specific, or procedure-, therapy-, or device-specific. Clinical research databases were soon developed for specific medical problems such as cancer, heart disease or rheumatism, or for medical technologies such as diagnostic tests, drugs, or surgical procedures.