COMPUTERS AND BIOMEDICAL RESEARCH 15, 164–187 (1982)

Discovery, Confirmation, and Incorporation of Causal Relationships from a Large Time-Oriented Clinical Data Base: The RX Project

ROBERT L. BLUM

Department of Computer Science, Margaret Jacks Hall, Stanford University, Stanford, California 94305

Received July 31, 1981

The objectives of the methods and computer implementation presented here are (1) to automate the process of hypothesis generation and exploratory analysis of data in large nonrandomized, time-oriented clinical data bases, (2) to provide knowledgeable assistance in performing studies on large data bases, and (3) to increase the validity of medical knowledge derived from nonprotocol data. The RX computer program consists of a knowledge base (KB), a discovery module, a study module, and a clinical data base. Utilizing techniques from the field of artificial intelligence, the KB contains medical and statistical knowledge hierarchically organized, and is used to assist in the discovery and study of new hypotheses. Confirmed results from the data base are automatically encoded into the KB. The discovery module uses lagged, nonparametric correlations to generate hypotheses. These are then studied in detail by the study module which automatically determines confounding variables and methods for controlling their influence. In determining the confounders of a new hypothesis the study module uses previously "learned" causal relationships. The study module selects a study design and statistical method based on knowledge of confounders and their distribution in the data base. Most studies have used a longitudinal design involving a multiple regression model applied to individual patient records. Data for system development were obtained from the American Rheumatism Association Medical Information System.

1. INFORMATION

Every year as computers become more powerful and less expensive, increasing amounts of health care data are recorded on them. Motivation for collecting data routinely into ambulatory and hospital medical record systems comes from all quarters. Health practitioners require this data for clinical management of individual patients. Hospital administrators require it for billing and resource allocation. Government agencies require data for quality of health care assessments. Third party insurers require it for reimbursement. Data bases may also be used for performing clinical research, for assessing the efficacy of new diagnostic and therapeutic modalities, and for the performance of postmarketing drug surveillance.

The various uses for data bases may be grouped into two fundamentally distinct categories. The first category pertains to uses that merely require *retrieval of a set of data*. For example, we may wish to know the names of all patients who had a diastolic blood pressure greater than 100 for more than 6

0010-4809/82/020164-24$02.00/0

Reprinted from Computers and Biomedical Research
Volume 15, Number 2, April 1982
Academic Press, Inc.