DEVELOPMENT OF MICROCOMPUTER BASED SOFTWARE FOR
ENCODING AND RETRIEVING MEDICAL INFORMATION IN PATHOLOGY

by L. Hause, Ph.D., D. Rothwell, M.D., and C. Frey, M.D.

The Medical College of Wisconsin

Abstract

The development of a microcomputer-based software package for the entry, encoding, storage and retrieval of medical information in anatomic pathology is described. Computer assisted encoding of diagnostic language was based on SNOMED nomenclature and was found to automatically encode 92% to 98% of the cases in routine surgical pathology. This application package was based on CP/M operating system so it functions on a variety of microcomputers and can integrate with other common microcomputer packages such as word processing.

Introduction

The preparation, indexing and filing of medical information have long been necessary yet demanding tasks in anatomical pathology. The advent of microcomputers offers an economical means to significantly improve the processing of this information. Both word processing and data base packages are available on most common computer systems so that economical reporting and retrieval of medical information can be achieved with relative ease.

One remaining question is what to do with natural language in the diagnostic text? In natural language the same condition can be variably described which can limit its retrieval. An encoding system was developed to classify medical text based on a hierarchical nomenclature in pathology -- Systematized Nomenclature of Pathology (SNOP). This systematized encoding dictionary was subsequently extended in a multi-axial manner to all of medicine as the Systematized Nomenclature of Medicine (SNOMED). Medical information so encoded minimizes the variability of language and facilitates consistent information retrieval both of which are consistent with computerized applications. Encoding can, in itself, be labor intensive if done manually so that an important step in the formation of a practical medical information systems is computer assisted encoding.

Semantic techniques which were based on appropriate dictionaries and formulae have been developed to process natural language. Medical language has been encoded in this manner based on the SNOP encoding dictionary and processing algorithms. These algorithms required the power of large computer systems to comprehensively encode diagnostic language. The MUMPS computer language in a large computer system was successfully applied to the encoding and retrieval of information in surgical pathology.

Because of the requirement for large computers with specialized programming, such systems were not available to most users so that contemporary microcomputers have been applied to process medical information. Enlander applied a key-word cascade to encode the diagnosis in pathology into SNOP codes. We have developed a general microcomputer-based system using the SNOMED dictionary and a set of rules which facilitate the computer assisted encoding of diagnostic language. Thus, the applicability of microprocessors to medical encoding in a manner which approaches large system processing appears feasible and practical.

In this report, the development of a microcomputer-based software system to enter, encode, store and retrieve medical information in pathology is discussed. The system was designed to be independent of specific hardware and integrates with common microcomputer packages such as word processing and data base management.

Developmental Procedures

The design of this software was based on three overall objectives which include: a) a common denominator among hardware systems, b) integration of data files with other standard software, and c) the practical applicability to needs in surgical pathology. In order to achieve a common denominator in the fast changing world of microcomputers hardware, development was based on CP/M microcomputer operating system (Control Program / Monitor, Digital Research, Pacific...
Grove, CA). CP/M facilitated operation on multiple hardware systems and offered compatibility so that data files could be integrated with other CP/M-based software packages such as word processing and data base management.

Program development was done in interpreter BASIC (BASIC-80 V5.0, Microsoft, Bellevue, WA). This common interpreter easily accommodated the necessary algorithms of text orientated applications. Modular development of software was maintained where possible through the use of sub-routines since this facilitated the ease of revision during test phases. Final application programs were compiled with Microsoft's BASIC Compiler V 5.3. Support for software development was provided through The College of American Pathologists and the Department of Pathology, Medical College of Wisconsin.

Microcomputer hardware used for development was all 2-80 based for compatibility with CP/M-80 software. Since 8 bit codes fully represent all English text, it was found that 8 bit microcomputers effectively demonstrated the capability to process medical information with limits due to the size of memory or disk files in some cases. Primary development was carried out on a Radio Shack Model II system with Lifeboat (New York, N.Y.) CP/M V2.25. Several other microcomputer systems with CP/M were used to test the software including an Apple II computer with a Softcard module from Microsoft.

Software Development and Application

Initial steps in the software development included a pilot study of the feasibility of microcomputer assisted medical encoding based on the SNOMED nomenclature. This initial study used the SNOMED dictionary and evaluated encoding formulae, a semantic technique, with a relatively small Radio Shack Mod I microcomputer. It was determined that microcomputer assisted encoding of medical text was feasible based on a SNOMED microglossary (the dictionary) indexed to an encoding algorithm similar to "Soundex". The "Soundex" is routinely applied to encode names on state driver's licenses and was found applicable to medical text. Briefly, the algorithm forms a code based on the first letter of a word followed by numbers representing the consonants which follow. An abbreviated algorithm using only the first letter was applied here. The SNOMED dictionary keyed with an abbreviated "Soundex" algorithm was used to efficiently encode diagnostic language which was combined with demographic information to form patient files for subsequent applications.

Development included modules of text entry, encoding and editing. Specific searching modules for patient files were included in the development. The first prototype of the package required about three months for development. The package was tested and revised several times before application which spanned a period of about one year. On site testing and one of us (DR) directed the practical applicability of the software. -- objective c. Testing was found to be a key step in developing a user-practical package and represented approximately 60% of the development time which is consistent with functional new product development.

The final software package for encoding and filing or medical information in pathology contains the following primary modules:

1) Initialization of demography -- the format of demographic patient information can be defined (customized) by users to fit the need of their facility. Subsequent software operation and the structure of patient files are directed by this user-defined format.

2) Patient text entry and encoding -- patient text with demography and diagnosis can be entered either by keyboard or by CP/M-based word processing (i.e., Word Star). Microcomputer assisted encoding occurs after diagnostic text is entered and then a temporary patient file is formed with text and codes. This temporary text is reviewed and edited at the convenience of the operator. Edited patient information with demography and coded diagnosis is stored in a permanent patient file for subsequent retrieval. Entry steps remain separate from editing steps in order to optimize clerical efficiency.

A SNOMED surgical microglossary of 2700 terms (lines) selected by one of us (DR) was typically able to match about 90-95% of the routine medical text in surgical pathology. In the evaluation of the software system, Harry, et al., found a 92% match rate for 2691 tumor registry cases. He concluded that the majority of the unmatched cases were due to topographic sites originally coded in ICD-0 for the tumor registry. This basic dictionary can be further customized by adding selected terms at the users discretion which further improves the match rate at specific sites.

The encoding process includes two phases -- the first automated and the second interactive. The first, automated phase is limited to direct matches or near matches between the diagnostic text and the dictionary. It benefits from the formation of a good dictionary and optimizes automated match rate for routine terminology. The second, interactive phase is applied during editing the temporary file and is based on the selection of phrases with specified words in the dictionary.

3) Medical information retrieval -- Two cases of retrieval are available for patient files within the software. The first is the determination of previous visits by review of patient files. The second is the search of patient files for medical relevance based on demography and / or diagnostic information. For example, the package can be used to determine whether a certain patient has been seen before or to specify the number of patients with a given disease by age and sex who
have been diagnosed in the past several years.

This software package facilitates economic manipulation of medical information in Pathology and efficiently circumvents much of the routine tedium of data handling and encoding. Software is based on CP/M so that it is compatible with a variety of microcomputers. In addition, the CP/M base allows integration of the application software in a modular manner with other CP/M based packages such as Word Star and d-Base. This allows versatile manipulation of patient information through CP/M text files. Reports describing the software and evaluating the software have been written. This software package with comprehensive documentation is available through The College of American Pathologists (Travers City, MI).

References


