DISEASE CODING BY COMPUTER

THE "FRUIT MACHINE" METHOD

BY

R. W. HOWELL, A.M.R.

Authority Health and Safety Branch, U.K. Atomic Energy Authority, Harwell, Berkshire

AND

RUTH M. LOY

World Health Organization Centre, General Register Office, Somerset House, London, W.C.2

Increased use is being made of computers in the processing, retrieval, and manipulation of medical data, but too frequently this is a refinement added to an existing system. There is a tendency for punched cards to be produced after laborious reduction of basic data by protracted manual coding and checking. There are now many applications where data can be submitted directly to the computer in uncoded alpha-numeric form; elimination of hand coding would give both saving in time and greater consistency. Preferably output should appear in plain language (subject to ethical and confidential requirements) rather than the coded mass so frequently encountered. There are some applications which would best be served by a mixture of machine-coded and plain language storage.

Coding usually encompasses such items as sex, marital status, consultant, specialty, occupation, social class, source of admission (waiting list, emergency, etc.), result of treatment, diagnoses, operations, and other similar data. Length of stay is often calculated manually from admission and discharge dates. Of these items, only diagnosis normally gives rise to difficulty in automatic coding; the large number of diseases encountered, extended by a formidable array of syndromes, eponyms, and synonyms, makes this a serious problem. Considerable machine storage is necessary, and comparison within the computer may prove an expensive item in an inefficient system.

THE "FRUIT MACHINE" METHOD

In the conventional fruit machine, the "jackpot" is obtained when the lemons appear in line. Simi-

larly, in this method of diagnosis coding, the "jackpot" (the correct code number) is obtained when a code number appears which is common to all words in the diagnosis (Fig. 1). In the main fruit machine dictionary each significant word of a diagnosis is stored with all the code numbers with which it has been associated.

Although acute appendicitis with perforation is shown pictorially in Fig. 1 as a series of separate punched cards for each word and number represented, storage is far more compact in the computer dictionary. The dictionary was, however, built up by using punched cards in this manner.

If, for example, the unqualified diagnosis of appendicitis is submitted to the machine, it will select the correct code number of 551 from among the several numbers linked with this term because an asterisk has been placed against the appropriate code number during the compilation of the dictionary (Fig. 1). This modification applies only to diagnoses consisting of a single word. If a one-word diagnosis (e.g. asphyxia) is considered inadequate for coding, then no asterisk is placed against any of the code numbers associated with this term and the computer rejects it for further amplification.

It was understood before programming started that in a small proportion of diagnoses a permutation or "cross-over" could occur; in such a case the fruit machine dictionary would be inadequate, for two or more numbers would meet the requirements of the basic method. An example of a cross-over which occurs in practice is shown in Fig. 2. As "myocardial insufficiency" is coded as 422.2 and "myocardial infarction" as 420.1, both

FIG. 1.—Computer coding of the diagnosis "Acute Appendicitis with Perforation"

code numbers are stored in the dictionary with "myocardial". Similarly, 420·1 and 422·2 for "insufficiency" are derived from coronary insufficiency and, of course, from myocardial insufficiency. This combination permits a cross-over.

FIG. 2.—Coding of "Myocardial Insufficiency" by the fruit machine dictionary.

Fig. 2 shows that "myocardial insufficiency" will provide both 420·1 and 422·2 as suitable code numbers. Where more than one code number is indicated by the main dictionary, the complete diagnosis is automatically referred to a sub-dictionary which stores each "cross-over" diagnosis together with its correct code number. Fortunately, the proportion of cross-overs is not so high that formidable storage problems are encountered. Some storage could probably be saved by replacing the asterisk in the fifth column with an extra overpunch in the fourth column of the diagnosis field. This fifth column could be used for a coding system to replace the cross-over dictionary, but the present method has simplicity in updating to recommend it.

The programme also provides for a warning print-out on occasions when the presented diagnosis is not in the dictionary. The dictionary can then be updated to include the new diagnosis for this and all subsequent encounters. This print-out also indicates diagnoses where more than one code number is encountered yet reference to the cross-over dictionary fails to give a preferred code number. The print-out of "update cross-over dictionary" against the diagnosis is an indication for suitable updating.

It was also recognized that a diagnosis could be miscoded because it failed to come to the cross-over dictionary through incompleteness; this is particularly likely in the building-up process of the dictionaries. For example, if 422·2 (the correct code number) had not previously been encountered and incorporated in the fruit machine dictionary, it might well be possible for myocardial insufficiency to be wrongly coded as 420·1 (see Fig. 2). During the building up of the dictionary it therefore appeared necessary to check manually each number coded; this could have been a tedious and time-

consuming process. To reduce checking it would have been desirable that only diagnoses not previously encountered were submitted to the machine, but this would have been difficult to ensure as the dictionary grew into thousands of words. In practice thousands of diagnoses were fed into the computer (without previous screening) for the purpose of building up and testing the dictionary.

To avoid this tedious checking of each one of the thousands of cards machine-coded, a further sub-dictionary was set up—the master diagnosis file. As a diagnosis was accepted, it was checked by the computer against this new sub-dictionary. If the diagnosis had not been previously submitted it was updated in the master diagnosis file and, at the same time, details were printed out so that manual checking could be undertaken. Diagnoses which have been met before are not so listed. A print-out can be obtained of all separate diagnoses so far submitted to the machine. It was also thought that there might be some merit in ascertaining the frequency of diagnoses in their various forms of terminology, and a counting device was therefore incorporated. Obviously the frequency of diagnoses will be affected by any bias in the material processed. For example, if regional obstetric data are processed the impression might be gained that female sterility is a non-existent condition!

The programme ignores such expressions as "with", "and", "to", "from", and other connective terms not directly related to the diagnostic code and also any descriptive material (such as whether or not a *post mortem* examination was carried out; external causes of injury; suicide) which can be stored in parentheses on the card being entered. Such information may be part of systems other than coding. This dictionary (the "compare" file) also includes words which would otherwise appear many times in the storage in many guises—"malignant neoplasm", "carcinoma", "ca", "ca.", "carcinomatous", etc.

Correct spelling is currently required (though such alternatives as "haemorrhage" and "hemorrhage" could be stored in the fruit machine dictionary or dealt with in the compare dictionary), but this could be readily amended, with some increase in storage requirements, to provide for a lower operating standard by the acceptance of one or more unexpected spelling errors. For example, "cholecistitis" could be accepted for "cholecystitis" without the former spelling having to be separately stored. The system could also code surgical operations, and an attempt will be made to do this in the near future.

The full experimental system used in the building-up

process is shown in Fig. 3. In practice the compare and fruit machine dictionaries are integrated, but are probably better represented pictorially as shown.

DISCUSSION

It was recognized initially that complete coverage of all possible terms and diagnoses would require a computer with considerable storage and that the dictionary could never hope to be completely comprehensive. It was thought that 90 to 95 per cent. coverage of all terms would provide a reasonable and practical operating standard. In such circumstances, in order to restrict the size of the dictionary, most of the small proportion of terms not in the dictionary could be re-submitted in an alternative form. For example, if "Besnier-Boeck-Schaumann disease" (ICD 138.0) were rejected as not in the dictionary it could be re-submitted as "sarcoidosis" rather than increasing the storage requirements.

At the moment, the full building-up programme (Fig. 3) requires a computer with storage of about half a million characters. However, it seems reasonable to believe that with further development some economies in storage may be achieved in addition to that made possible by using a standard "developed" dictionary which omits the numerous checks used at present. With just over 20,000 card equivalents in the fruit machine dictionary, 2,000 diagnoses in the cross-over dictionary, and about 2,500 single

word diagnoses marked with an asterisk, the present system will code over 1,000 diagnoses a minute. The fruit machine method has the advantage that the diagnosis may be submitted in any word order without increasing the main dictionary storage. Quite clearly, too, the present method could be readily made to encompass a subsequent retrieval and statistical system for all material entered. The expense of a large computer may thus be justified, and clearly specialty coding (*e.g.* cancer registration, regional maternity and obstetric data) could be carried out on machines with smaller storage. The present storage is currently giving 95 per cent. coverage when judged by several series of randomly selected causes of death provided by the Registrar General for England and Wales; updating is continuing using morbidity data.

One question which must be considered in both a manual and a mechanized system is the level of accuracy required; a complementary consideration must be the standard and consistency of coding which is actually attained. It has been shown that many diseases, such as carcinoma of the lung, cerebral accidents, and other important causes of death, are often considerably under or over-diagnosed—a leader in the *Lancet* (December 1966) discussed this and gave a number of references. It may therefore be questionable whether the material to be coded is worthy of 100 per cent. accuracy, or indeed whether this can be achieved even if expense is not of primary concern. Accuracy

FIG. 3.—Diagnosis coding—The system during development.

in manual coding varies from coder to coder and from country to country (*Wld Hlth Org. Chron.*, 1967); rates based on broader groupings may be reliable when those for specific rubrics are not (Reid and Rose, 1964). It therefore seems unlikely that machine coding will be less satisfactory than manual methods, and the computer will certainly prove more consistent if a standard dictionary is used.

Machine coding is particularly suitable for multiple-cause coding, but sequential single-cause coding is possible provided that the rules can be logically expressed for programming, and that the pattern of sequences encountered, both "acceptable" and "unacceptable" according to the rules, is not too varied.

The future will bring more and more information entered into multi-access systems to give regional on-line computing. With this type of computing there are great advantages in methods which will eliminate manual coding; the normal fixed-field 80-column limitation need no longer apply. For example, the machine will not only code the diagnosis but will also store both the rubric and the plain language diagnosis. With present methods in general, if cases of bronchospasm are required, all cases in ICD 527.2 are retrieved and reference is made manually to the original records. The more flexible method described will produce a print-out showing visually cases of bronchospasm among patients with oedema of the lung, hernia of the lung, mediastinitis, stenosis of bronchus and trachea, etc., who also share this code number.

It would appear that automatic coding should prove at least as accurate as manual methods and certainly more consistent, though it would be prudent to adopt a sample manual check of machine coding. Whether this proportion should be 1 per cent. or, say, 5 per cent. will depend on experience with this method and upon desirable standards. Rare diseases or unlikely sequences may also be worth checking. Accuracy in the preliminary dictionary coding is obviously of vital importance. The system will work in languages other than English although the dictionary will have to be built up systematically in each.

Although the code numbers referred to are derived from the Seventh Revision of the International Classification of Diseases, the system is currently operating only in the Eighth Revision.

SUMMARY

There is a tendency for punched cards (or other input media) to be produced only after laborious reduction of the basic data by protracted manual

coding and checking. Many applications arise where uncoded alpha-numeric data could be submitted directly to the computer both for coding and for subsequent storage or processing.

In the present system of computer coding, each different word encountered in many thousands of diagnoses is stored with all the code numbers with which it has been associated. The number common to all words in the diagnosis submitted is normally the correct code number. This system has the advantage that the diagnosis can be submitted in any word order without adding to the size of the main dictionary. However, single-word diagnoses frequently have a large number of attached code numbers and the correct rubric is indicated by the inclusion of an asterisk. The system produces more than one code number for diagnoses of more than one word in about 6 per cent. of all cases. In this event, the complete diagnosis is referred to a sub-dictionary in which is stored the correct code number.

Other difficulties arising in the development of the system are discussed as are the measures taken to overcome them. A large computer has been necessary to develop the system which codes over 1,000 diagnoses a minute. With a fully-developed standard dictionary not requiring all the checking stages, the system could be run with a small computer with disc backing. Certainly specialty coding (*e.g.* cancer registration, regional obstetric statistics) could be carried out with a smaller machine if the storage requirements of the remainder of the system were not too high.

The advantages of automatic coding for both multiple and sequential diagnoses are discussed.

We are glad to acknowledge our indebtedness to Dr R. D. Weir (University of Aberdeen), who suggested research into systems based on diagnostic words rather than on the diagnosis as a whole.

Our thanks are due to Mr S. Y. Duncan (U.K.A.E.A.) who has carried out the programming and will be reporting details of this elsewhere.

We are also grateful for the help and support given by the World Health Organization, the Registrar General's department (England and Wales), and the U.K. Atomic Energy Authority.

REFERENCES

- Lancet* (1966). Leader, 2, 1349 (Accuracy of death certificates).
- Reid, D. D., and Rose, G. A. (1964). *Brit. med. J.*, 2, 1437 (Assessing the comparability of mortality statistics).
- World Health Organization (1967). *Wld Hlth Org. Chron.*, 21, 11 (The accuracy and comparability of death certificates).