Data processing in histopathology

M. K. ALEXANDER

From the South Warwickshire Hospital Group

'Flat and flexible truths are beat out by every hammer; but Vulcan and his whole forge sweat to work out Achilles his armour.' (Sir Thomas Browne: *Cyrus' Garden*, Ch. 5.)

A histopathologist performs in diagnosis at least five fundamental tasks: he abstracts and mentally integrates features of tissue structure; he relates these patterns to information drawn from memory and other records; he makes inferences from all these data; he then forms and finally communicates a description. In any consideration of the ways in which computer technology can affect this sequence, we are immediately faced with the fact that the primary process, the recognition of pattern, is an exceedingly complex and largely unanalysed cerebral activity. Automatic recognition of pattern is still in its infancy, and although a beginning has been made in the limited and relatively simple fields of cytology and chromosome analysis towards the removal of the human brain from the scene, it is proposed to confine the present discussion to the remaining aspects of the process, namely, the formation, storage, recall, and communication of descriptions.

The process of describing and making inferences from the patterns which we recognize, although open to logical analysis, is again a complex business. In practice, the simpler and more obvious the pattern, the fewer the words we employ. In the simplest cases of all we may proceed directly to the inference and apply a classifying label without more ado. On the other hand, where the pattern is not clearly seen, our descriptions are more consciously analytical and form the basis for an assessment of probabilities. This part of the process is comparable to clinical diagnosis and in principle is open to the techniques of propositional calculus (Feinstein, 1967), numerical taxonomy, and multivariate analysis (Baron and Fraser, 1965; Hayhoe, Quaglini, and Doll, 1964) which are being used experimentally in that field.

However, the more immediate benefits to histopathology which are likely to be obtained from the application of modern data processing methods lie in the direction of the storage and retrieval of data on a large scale.

Before discussing methods it seems pertinent to consider the reasons for wishing to embark on any

system of data collection. I would suggest that these are both local and general in nature. At the local level, the linking of the histopathology reports with the remainder of the patients' records is the primary need. The problem at this point becomes part of the much larger and more difficult problem of the organization of the medical record for data processing. The potential benefits to pathology of the capacity to make correlations with other laboratory and with clinical data need no emphasis, although we are as yet far from a successful solution. Secondly there is the indexing of the local collection of reports and sections: this is the traditional field for most of the data processing that has been practised in histopathology to the present time.

To turn to the motives which could lead to the collection of data on a larger scale, these include the formation of large data banks for reference and research, and the collection of data for statistical analysis of morbidity and mortality. Examples of this kind of use already exist in the shape of the various regional cancer collections, and a strong case can be made for the extension of the principle to other types of pathological data. Furthermore, I would suggest that the term 'pathological data' could in this context well be extended to include other classes of related information whose significance is enhanced when it is reviewed in the mass. Anthropometric data are of this kind: for example, it is possible to calculate total body fat and muscle and skeletal mass from simple postmortem measurements (Alexander, 1964). Again, the details of traumatic death could provide a valuable source of facts in the field of accident prevention. The problem is largely one of organization; at its heart lies the need for a free exchange of information between centre and periphery. As we know all too well, there is a natural and proper resistance to the labour of collecting and supplying information without any visible return. Before any pathologist could take part in a larger scheme with enthusiasm, he would reasonably expect free and rapid access to centrally stored information, this to include slides as well as documents. Given the necessary central organ-

ization for implementing such a proposal, there would seem no inherent difficulty in carrying it out by a postal system, perhaps on a regional basis.

When he comes to store and retrieve data the histopathologist is at an advantage over the clinician because the terminology which he uses is at once more precise and more formalized.

In essence, the problem is the practical one of transforming and systematizing a verbal terminology into a form which can more readily be manipulated by the computer. A choice of methods exists, the choice depending on the particular aim which is pursued and the facilities available.

A number of systems have now been devised which successfully employ the plain language of pathological terminology as the input for their data, *eg*, those of Smith and Melton (1965) and of Lamson (1966). Given certain restrictions on the organization of the language used, which in practice are at most a minor inconvenience, they permit the capture of data in computer-readable form as a byproduct of the writing of a report of a conventional kind. In these systems, coding into a numerical representation is made by the computer program itself, in this way removing the burden of manual coding from the pathologist.

Manual coding is of course a practical procedure, provided that it is pursued with sustained enthusiasm and without undue elaboration. There must be few pathologists who have not at some time used such a method, and for most there is no satisfactory alternative in the immediate future. Up to the present, the value of the information coded in this way has been severely limited by the fact that the codes employed have usually been private to the laboratory or institution concerned. In an effort to supply the need for a comprehensive coding of pathological data, the Systematized Nomenclature of Pathology (SNOP) was produced by the College of American Pathologists in 1965. This code classifies data under four main headings: topography, morphology, aetiology, and function. A four-digit coding system applies to each of these fields. The system was designed with the 80-column punch card in mind as, in addition to the digits 0 to 9, the letters X and Y are used in some locations, thus allowing for full coding with a single punch in each card column by taking advantage of all 12 rows of the Hollerith card.

The code systems for the four fields are partly branching in structure. For example, as regards topography, the successive digits from left to right identify the site of the lesion with increasing precision. It is thus possible to obtain sufficient accuracy for many purposes by using the first two digits of the topographical code.

A limited personal experience of SNOP for the coding of routine histopathology has been satisfactory. The sections used have been mainly the two-digit topographical and the four-digit morphological codes, although provision has been made in the method of recording for expansion to a full 16-digit coding as desired. The coded diagnoses are entered on punched cards and are then available for sorting and listing on local data processing equipment or for computer input. In practice it has been found useful to make a brief personal selection of the most commonly used codes: approximately 30 topographical and 50 morphological codes together cover the diagnoses of a large majority of the material routinely handled.

The vistas opened by the more widespread use of a coding system of this nature very quickly reveal the existence of rival claimants, which although designed for somewhat different functions, nevertheless demand consideration in so far as the pathologist's diagnosis may in some circumstances need to be expressed in their terms. The principal systems involved at present are the International Classification of Diseases (ICD) of the World Health Organisation and the Current Medical Terminology (CMT) and the Standard Nomenclature of Diseases and Operations (SNDO) of the American Medical Association.

Taking the wider view, any large-scale system of data collection for pathology on a national or international basis must recognize the problem of compatibility of coding if an intolerable duplication of effort is to be avoided. Whether the ideal of a universal and comprehensive terminology suited to the needs of all branches of medicine is attainable or even reasonable, may be questioned. It seems more probable that the problem will ultimately be solved by making use of the power of the computer to transform data from one code to another. The Current Medical Terminology will soon be available on magnetic tape, its codings matched with the ICD, SNOP, and SNDO code systems. An extension of this concept, as proposed by Minkler (1968), would carry the process a stage further. A standard input in plain language could be matched at will with a dictionary of the desired code held on file in the computer and the information transformed as required. An approach of this or of a similar kind is one important way in which the power and versatility of the computer could be used to greatest effect.

SUMMARY

Although the direct application of the computer to the diagnostic process in histopathology has yet to be properly explored, the technical and con-

ceptual difficulties are so great as to render its progress slow.

The more immediately practical approach would seem to lie in the field of large-scale data collection and analysis, and suggestions are made as to the objectives and methods which might be pursued.

The problems of reconciling the demands of different disciplines as regards satisfactory coding of data are discussed; it is suggested that this may best be achieved by computer translation.

REFERENCES

- Alexander, M. K., (1964). The postmortem estimation of total body fat, muscle and bone. *Clin. Sci.*, **26**, 193-202.
- American Medical Association, Chicago. *Current Medical Terminology*. 3rd ed., edited by B. L. Gordon (1966).
- Baron, D. N., and Fraser, P. M. (1965). The digital computer in the classification and diagnoses of diseases. *Lancet*, **2**, 1066-1069.
- College of American Pathologists, *Systematised Nomenclature of Pathology* (1965), edited by E. T. Thomson and A. C. Hayden, Chicago.
- Feinstein, A. R. (1967). *Clinical Judgment*, Williams and Wilkins, Baltimore, USA.
- Hayhoe, F. G. J., Quaglino, D., and Doll, R. (1964). *Spec. Rep. Ser. med. Res. Coun. (Lond)*, **304**.
- Lamson, B. G. (1966). *Data Processing in a Medical Center*, University of California, Los Angeles.
- Minkler, T. M. (1968). *Amer. J. clin. Path.*, **49**, 246 (abstract). (National Conference on Medical Nomenclature. (1961). *Standard Nomenclature of Diseases and Operations*, 5th ed., McGraw-Hill, New York.
- Smith, J. C., and Melton, J. (1965). *Data Control for Anatomic Pathology*, Western Reserve University, Cleveland, Ohio, USA.
- World Health Organization (1949). *Manual of the International Statistical Classification of Diseases, Injuries and Causes of Death*, (1949). 6th revision. HMSO, London.