Basis of data control for anatomical pathology

J. CHANDLER SMITH

From the George Washington University, Washington, DC, USA

A principal function of the morbid anatomist is the generation of tissue diagnoses and the recording of them. This activity requires skill and produces large amounts of data. At the rate of 600 necropsies per year, and 20 diagnoses per necropsy, a single department may generate over 12,000 morphological terms in a year's time. These terms are of considerable importance; they express the frequency of various disease states, they reveal the relationships among different lesions, and, as the record continues, they disclose the changing incidence of human illnesses. It is, therefore, important to medicine and to the welfare of communities and perhaps nations, that these data be recorded and stored in such a way that retrievability is enhanced, and analyses are possible. As the record enlarges, and the usefulness of it increases, the need for generating a retrievable account becomes imperative, and the responsibility for satisfying this need rests squarely on the shoulders of the morbid anatomists.

While the anatomical record grows rapidly, the methods for retrieving those data have not been overlooked. There are simple methods and complex methods. There are code-by-hand methods (Systemized Nomenclature of Pathology, 1965), there are retrieve-by-machine methods (Systemized Nomenclature of Pathology, 1965; Carpenter, 1962), and there are methods that both code and retrieve by machine alone (Smith and Melton, 1964; Lamson and Dimsdale, 1966; Paplanus, Shepard, and Zvargulis, 1969). This welter of methods, however, does not entirely solve the problem before pathologists. Between the mountain of records to be retrieved on the one side, and the welter of methods for accomplishing the retrieval process on the other, lies the bottleneck of translating the English words of the former into the code signals of the latter. The success of any system will depend on the ease with which this step is accomplished. The bottleneck of coding is depicted in Figure 1.

It is the coding procedure that requires special attention. Coding begins with a definition of what the retrieval process is expected to accomplish. The retrieval process should be expected to accomplish two main objectives: the first is the identification of any record that lists a particular diagnosis, and the second is the enumeration of the anatomical site involved and the pathological process independently of each other, and on hierarchical levels progressing from generality to specificity. Thus, enumerating the pathological involvement of anatomical systems regardless of the lesions, and on the other hand, enumerating pathological processes irrespective of their sites, may be easily undertaken. For example, the frequency of abnormalities of the central nervous system may be of interest regardless of what the abnormalities were. And in the same way, the incidence of pathological processes such as carcinoma in a necropsy population is a figure of interest irrespective of the sites at which they arose. The ability to scan each of these categories of information independently of the other is thus useful. The dual objectives of identifying specific diagnoses and of scanning processes and sites independently of each other, and on different hierarchical levels, thus provides for virtually unlimited control of the data. This is especially true when other aspects such as age, weight, race, sex, and organ weights are also included as independent variables.

CODING

It cannot be denied that these capabilities are desirable; it can only be asked whether they can be achieved. The key lies in the process of coding. The key to coding is not agreement on numbers or symbols to represent words, for the variety of words has no limit, the key is instead the ex-
pression of the anatomical diagnoses in a regular format. Coding thus begins with the structure of the diagnosis, and it is this constant structure that is the basis of data control for morbid anatomy. It is this regularity of expression that eases the bottleneck of coding, and it is this format that allows all retrieval systems to fulfill the dual objectives of the retrieval process. Coding, which is the critical step in data control, thus requires a concession on the part of morbid anatomists, and the concession is that their diagnoses be written in a constant format. The concession is small, but once granted, the rewards are substantial.

The constant format of the anatomical diagnosis is the conjunction of one pathological process with one anatomical site. The reason for this format is that the words expressing each component of the diagnosis are classifiable in independent and separate categories. To give the most elementary example, the following three diagnoses consist of single process-site conjunctions: haemosiderosis of lung, carcinoma of liver, cyst of kidney. It is apparent from these diagnoses that the three structures involved represent three different anatomical systems of the body. Moreover, the abnormalities listed represent three different pathological processes. Each of these word types is classifiable independently of the other. Codification and storage in this form thus permits scanning of the respiratory system for cases of haemosiderosis of the lung, and also for searching of the pathological process of pigmentation for examples of haemosiderosis in any other system. In the same way, cases of carcinoma of the liver are specifically identifiable under the gastrointestinal tract, while all cases of malignant epithelial tumour are recoverable under the process of malignant neoplasia irrespective of where they occur.

Not only are the process and site words separately classifiable, they are also subject to hierarchical arrangement. Thus, the stomach, liver, and colon are part of one anatomical unit, the digestive system; and the adenomas, carcinomas, and lymphomas are part of one pathological process, that of neoplasia. These subdivisions allow for hierarchies within each category of information and hierarchies provide for searching on several levels of increasing specificity. It is accordingly as easy to search for lesions of the entire respiratory tract, or only the bronchi, or only the lungs, or only the pleura, as it is on the other hand to search for all cases of neoplasia, all cases of malignant neoplasmia, or all cases of malignant neoplasia of squamous type.

In order to realize these benefits, three rules should be followed. The first is that each diagnosis be written as a single process-site conjunction. As an example, ‘infarct of kidney’ conjoins one process with one site. The second rule is that the most specific words be used, ‘occlusion of coronary artery’, for example, is less specific than ‘thrombus of coronary artery’. The third rule is that the diagnoses be registered according to anatomical systems.

The effect of the first rule is to quantify the data, i.e., each diagnosis consists of two discrete units, one a pathological process, the other an anatomical structure. The second rule provides for words with specific meanings. These two rules allow for classification and subclassification of two different kinds of information in separate categories. The third rule segregates the diagnosis according to anatomical systems so that searching by systems and comparisons among systems may be made. The general effect of the three rules, however, is to quantify the data and express the information in a consistent pattern. This allows man or machine to deal with it effectively.

With respect to machines, it has been pointed out that the computer may turn medicine into a mathematical science, and that if this is to happen physicians will need to pay more attention to the precision with which they use words. Mechanism is bound to occur, and pathologists may lead the way in this process because they are fortuitously provided with a highly disciplined vocabulary, and the mechanism for dealing with that vocabulary is based on an uncompromising demand for clear instructions. With respect to pathological diagnoses, it is the words used and the format followed that constitute these instructions. A format for these rules is shown in Figure 2.

**ADVANTAGES**

I should like now to suppose that an institution has recorded its tissue diagnoses for one year according to these rules, that is, only single process-site conjunctions have been registered, the most specific words have been used, and the diagnoses have been recorded according to the anatomical systems of the body. What are the advantages to be gained from this self-imposed discipline?

The first advantage is that the dual objectives of the retrieval process are fulfilled, that is to say any process-site conjunction may be identified, for example, the cases of cirrhosis of the liver, and scanning of the entire record for the involvement of any anatomical site, or the occurrence of any pathological process may also be undertaken. Thus, it is easily possible to find out how often the
Basis of data control for anatomical pathology

INSTITUTION

Name:            Necropsy No.:            Date:

Demographic Data:               Date:

Principal Diagnoses
Recent thrombus of left coronary artery
Recent infarct of myocardium
Bronchopneumonia of right lung

Diagnoses by Anatomical Systems
Cardiovascular
Recent thrombus of left coronary artery
Recent infarct of myocardium
Acute fibrinous epicarditis

Respiratory
Bronchopneumonia, lower lobe of right lung
Acute fibrinous pleuritis, right

Digestive
Acute passive hyperaemia of liver

Genito-urinary
Arteriolar nephrosclerosis

Haematopoietic
Passive hyperaemia of spleen

Endocrine
Cortical adenoma of left adrenal

Musculoskeletal
No pathological diagnosis

Cranio-spinal
Permission for examination not granted

Integumentary
Scars of the skin

FIG. 2. Tissue diagnoses expressed as process-site conjunction, one per line, according to anatomical systems. In some instances a diagnostic word includes both process and site, eg, nephrosclerosis, epicarditis.

pancreas was affected by any lesion, and it is equally possible to find out how often carcinomas were recorded in the entire record. In addition, of course, it is possible to identify all of the carcinomas of the pancreas. It is perhaps unnecessary to emphasize that these three principal searches may also be adjusted according to several additional factors such as age, sex, race, chronological intervals, and weights of various organs. These prospects are substantially in advance of most retrieval systems which tend to allow only for the recovery of process-site conjunctions, and do not provide for scanning of processes and sites independently of each other or on different hierarchical levels, or in conjunction with a variety of demographic features. The first advantage of the method is thus substantial.

The second advantage is that the necropsy record is improved. It is improved because the tissue diagnoses are registered as morphological observations by anatomical systems. The diagnoses are not registered in speculative combinations or in accordance with presumptions of importance. They are registered instead as anatomical facts independently of speculations regarding significance or probable relationships. Thus, objectivity is advanced and value is increased. The necropsy examination, which reveals altered structure, and the experimental investigation, which seeks reproducible results, are both scientific procedures, and it would be incongruous to require objectivity in one and allow subjectivity in the other. Thus, the record improves as the diagnoses are registered in discrete units, and as these factual units are kept separate from opinions and conclusions about them.

This is not to say that a dead-house tabulation of facts is an end in itself. On the contrary, it is to say that a 'living pathology', ie, the interpretation of clinical events and the understanding of clinicopathological correlations, depends entirely on an objective and systematic registration of the facts. Such facts place no restriction whatsoever on the interpretation of them, now or later, but they do provide the basis for the best possible interpretations whenever they are made. It is, of course, apparent that interpretations cannot be better than the facts supporting them. Others may also speculate on clinicopathological correlations, but only the pathologist can provide the morphological facts, and the subsequent clinical speculation depends on the reliability and completeness of them. The primary function of the pathologist is thus to glean these facts and record them as such. That practice improves the record for the present and enhances its value for the future as well.

A third advantage of the format is that the training of morbid anatomists is improved. They also learn in their formative years the distinction between fact and opinion. For example, the observation of a myocardial infarct is a fact, and the observation of an acute fibrinous epicarditis is a fact, but it is opinion that they are related. The two conditions may be written together, ie, 'recent infarct of myocardium with acute fibrinous epicarditis' and this suggests that the former is the cause of the latter. It is quite possible, however, that uraemia was the cause of the epicarditis and it would have been present in the absence of the infarct. To register these two conditions as associated may thus be, in fact, wrong. There are many examples of this practice that purports to show such a cause-
and-effect relationship. 'Fracture of the skull with subarachnoid haemorrhage' implies that the former was the cause of the latter, while in fact, the haemorrhage may have caused the patient to fall and sustain the fracture. Therefore, it is better that each observation be registered independently of the other, and that the two observations be separated from opinions about them.

The practice of combining diagnoses is not only fraught with error, it also necessitates an unsound principle; it encourages the morbid anatomist to register not what he sees, but what he presumes about what he sees. In the former instance, the record becomes a compilation of the objective data, while in the latter it is a subjective hybrid of observation, presumption, bias, and speculation. The training of the morbid anatomist thus improves as he distinguishes between lesions that are verifiable and incorruptible facts, and interpretations of them that are subjective and prejudicial opinions.

The last main advantage of the format recommended is that it is equally amenable to simple, intermediate, and complex methods of data storage. Moreover, all of these methods fulfill the retrieval objectives of specific lesion recall and scanning search function. A simple storage method is represented by hand coding, card filing by hand, and recall by searching the compiled cards by hand. While all of these steps are performed manually, the registration of anatomical site words and pathological process words independently of each other allows for searches that are specific or of a scanning nature and may be directed at either the sites involved or the processes observed.

Intermediate methods are also possible: transfer of coded data to punch cards allows for automated searching of the stored data (Systemized Nomenclature of Pathology, 1965). Optical coincidence systems are also in use, in which the anatomical and pathological data are stored in the form of drill holes in plastic cards through which light shows (Carpenter, 1962). The lights, which may be counted automatically, reveal the case numbers and the number of cases, in which any conjunction of process and site is to be found on either a specific or generic level. Also, overlaying the cards reveals the association with other aspects of the data such as demographic factors and organ weights. The storage of data in independent categories on such cards is shown in Figure 3. The method is now in early operation at the George Washington University (Smith, 1969).

The most sophisticated of the data storage methods relies on the computer alone for both the coding and retrieval of the data. This is possible because the format recommended is constant, the process and site words are for the most part specific, and these words may be classified in hierarchical schemes (Smith and Melton, 1965). The original writing of the diagnoses is performed on a tape typewriter, the paper tape is accepted by the computer, and each word is recorded on magnetic tape. By automatic routines, the words are then matched against a previously written dictionary, and when a match is found between the incoming word and the dictionary word, the assignment of a code to that word is automatically performed. Later, the English word is dropped, the code is reassembled in order of word accession numbers,

FIG. 3. Storage of data cards by anatomical sites and pathological processes.
and the final tape is stored in the form of a sequence of code signals. When recall is desired, it is necessary to consult the dictionary for the code of the words sought, and to instruct the computer to search the tape for a match between the code of the question and the code of the tape. When a match is found, the computer prints out the number of the case in which that diagnosis is recorded. The system has been found to be accurate and serviceable (Smith and Melton, 1964).

The key factor in the computer method is the dictionary which fixes the hierarchical classification of the process and site words. The responsibility for this classification is not yet assigned. Nevertheless, it is perhaps better for pathology organizations to supervise the classification of such words than it is for them to devise numerical codes which function as substitutes for those words. The former allows the pathologist, or his designate, to work with words and utilize machines for the process of coding, while the latter consigns the pathologist to the unending chore of coding by hand and the needless task of transposing words into numbers.

While computers hold allure because of their enormous computational power, it should not be presumed that their use for the storage of pathological diagnoses is either simple or easy. The method is complex, time consuming, and expensive. The disadvantages of the method have been described (Smith, 1969). Nevertheless, the storage capacity of the method is virtually unlimited, the variety of search options is extremely broad, and the speed of searching is unexcelled. Moreover, input requires only a paper tape of the final diagnoses so that centres for the storage of data from hospital units without such equipment bid fair to become a practical possibility. It is in the context of a centre for the accumulation of large amounts of data covering a broad territory that the computer method is likely to find its most useful service.

SUMMARY

In summary, it can be seen that pathology now stands, along with other medical specialities, on the threshold of an age of rapid technical advance. A feature of this age is instrumentation capable of manipulating large amounts of data. It is important that pathology should avail itself of these advances. The key to data retrieval for morbid anatomy, and the connexion with the technical advances of the modern day, is the efficiency of the coding procedure. That is accomplished by expressing the diagnoses in a constant format, i.e., single process-site conjunctions, one to the line, according to the anatomical systems of the body, and by using the most specific words to express the two main categories of information. Registration in this form precludes no storage method, makes available the data to all machine methods, minimizes or automates the coding procedure, and allows for generic to specific recall of the anatomical site involved or pathological process. Registration in this form also improves the anatomical record and advances the training of pathologists.

Perhaps the most important effect of this practice, however, is to quantify the data. It has been remarked that '... all reality can be reasoned about, and not to quantify that which can be quantified is only to be content with something less than the full range of reason' (MacNamara, 1968). Let us avail ourselves of the fullest range of reason.

REFERENCES