A COMPARISON OF DIAGNOSTIC CODING BY MANUAL AND COMPUTERIZED METHODS

R. W. HOWELL

British Steel Corporation, London, S.W.1

Howell (1968) and Howell and Loy (1968) have described the 'fruit machine' method of diagnostic coding which, using a large computer, will code at the rate of 2,000 diagnoses a minute. More recently, the Medical Research Council's Computer Unit has undertaken reprogramming and extension of this 'fruit machine' method to enable coding to be carried out on a medium-sized machine, currently at the rate of about 60 diagnoses a minute. The data from the present series were processed by courtesy of the Medical Research Council.

The Research and Intelligence Unit of the Scottish Home and Health Department suggested that a comparison be made, using Scottish morbidity data, of the merits of machine and manual coding.

The objects of the study were to examine approximately 25,000 hospital discharge summaries:

(1) to assess the present accuracy of manual coding over various diagnostic ranges for the different regional boards;

(2) to compare the accuracy of machine and manual coding;

(3) to assess the feasibility of future machine coding of Scottish morbidity data, i.e., to establish that automatic coding is operational on 'real data'; and

(4) to provide further information which would be used to update the fruit machine dictionaries, i.e., to make them more comprehensive.

METHODS

The Scottish Home and Health Department randomly selected some 25,000 hospital discharge forms (S.M.R.1) from the first seven months of 1968. Mental hospitals were not included in these returns nor were obstetric cases. Basic information, including up to three diagnoses in plain language and the manual coders' code numbers (ICD 8th revision), was transferred to punched cards. Space was left on the punched cards to permit the computer insertion of the code numbers given by the machine after its reading of the diagnoses. A further computer run produced tables by regional hospital board area and by diagnostic grouping for those patients for whom there was agreement be-

tween the manual and machine codes. Where the machine failed to code or where there was a difference between the manual and machine code numbers, the punched cards were scrutinized manually as the machine was unable to decide where the error lay. Where the machine and manual codes differed, error detection was normally not difficult since the machine had generated a list of almost 10,000 different diagnostic phrases which had been individually checked by the Office of Population Censuses and Surveys. A small number of doubtful differences was referred to the OPCS for a second opinion.

RESULTS

Section 1 of Table I shows quite clearly that machine coding in this survey was nearly 10% more

TABLE I
COMPARISON OF COMPUTER AND MANUAL RESULTS

	Machine Coding		Manual Coding	
	No. of Patients	% of Total	No. of Patients	% of Total
Coded correctly				
1a. Machine and manual in full agreement	18,710	74.3	18,710	74.3
1b. Machine only correct	3,984	15.8		
1c. Manual only correct			1,665	6.6
Subtotal	22,694	90.1	20,375	80.9
1d. Difference between machine and manual probably not significant	461	1.8	461	1.8
Total section 1	23,155	92.0	20,836	82.8
Coded incorrectly				
2a. Linkage failures	88	0.3	32	0.1
2b. Sex- or age-dependent errors	62	0.2	23	0.1
2c. Abbreviations	256	1.0		
2d. All other miscoding	464	1.8	4,286	17.0
Total section 2	870	3.5	4,341	17.2
Failure to code				
3a. Poor diagnosis	171	0.7		
3b. Abbreviations	158	0.6		
3c. Sex-dependent diagnoses	23	0.1		
3d. 'Non' diagnoses: 'pregnancy', 'social reasons', etc.	140	0.6		
3e. Inadequate machine dictionary	660	2.6		
Total section 3	1,152	4.6		
Grand total	25,177	100	25,177	100

effective than manual coding in obtaining correct results. Of even greater importance, however, was the finding in section 2 that the manual error was 17.2%, almost fivefold that of the machine. Furthermore, computer error will fall to almost nil with experience, for the errors arise largely in new diagnostic phrases encountered. With suitable updating of the machine dictionaries these errors will be eliminated, though one would expect additional, but numerically fewer, errors to arise on the next run of 25,000 patients from the same source.

Whereas the manual coder coded all diagnoses, the computer failed to code about 5%. This proportion could be expected to diminish with experience as the dictionaries are updated. However, there may be some merit in deliberately letting the machine fail to code some diagnoses.

For example, in 171 cases the diagnosis was too poor for adequate coding and there might be some merit in making further enquiries before coding. Most of the 158 cases which failed to code on abbreviations could be taken care of by updating. The 140 'diagnoses', such as social admission, menstruating, pregnancy, and so on, are readily machine codable once definite code numbers can be agreed. Although these 311 cases have been credited as 'manual only correct' (Section 1c, Table I), in fact manual coders selected a great variety of individual code numbers for these terms and they cannot all be right. Machine coding proved to be entirely consistent, with no unexpected or unexplained failures of method.

Section 1d of Table I shows 461 cases (1.8% of all patients) for whom manual and machine coding differed, but where it would have been unfair to assume that the manual coder was wrong though the machine was correct according to the manual. Many of these were diagnoses of 'tonsillitis' which the machine correctly coded as 463-. It may be that in some cases manual coders knew that patients had been on waiting lists for long periods and therefore coded cases as chronic tonsillitis (500-).

There was little variation in the manual coders' error rate for primary, secondary, and tertiary diagnoses, nor was there any substantial variation between regional hospital board areas; no region had a particularly low error rate.

Table II gives the proportion of manual coders' errors by sections of the ICD classification (8th revision). This table shows a particularly high error rate when coding complications of pregnancy, childbirth, and the puerperium. The main source of error arises from the coding of abortions, particularly in the fourth digit. Incomplete abortion (644.9) was invariably coded by some hospitals as

with sepsis (644.0) and by others as spontaneous abortion (643.9). It may be that coders had not adjusted to the differences in the 7th revision where mention of sepsis involved the third digit rather than the fourth. Abortion without mention of sepsis then fell under 650 and abortion with sepsis under 651. The position may have further changed because of the introduction of the Abortion Act, which was not operative at the time of these hospital discharges.

There was a high error rate in the coding of mental disorders but the number of cases was small, and no particular diagnosis was consistently misconstrued. Diseases of the circulatory system had a high error rate and some coders did not appear to appreciate the difference between coronary ischaemia and coronary infarction. The fourth digit appeared to present difficulties to some coders, and it was not unknown, where a patient had two circulatory diagnoses, to find one coded as with hypertension and the other coded without hypertension. Proapsed intervertebral disc was frequently coded as 725.1 rather than 725.9 at some hospitals.

Re-scrutiny of the data to compare manual and machine coding using the first three coding digits only showed a substantial fall in the manual coders' error rate from 17% to 11%. The original manual error rate was roughly halved on a three-digit comparison in the following diagnostic groups:

- Mental disorders
- Nervous system and sense organs
- Circulatory system
- Genitourinary
- Complications of pregnancy, childbirth, and the puerperium
- Congenital abnormalities.

TABLE II
PROPORTION OF MANUAL CODERS' ERRORS BY DISEASE GROUPING—PRIMARY DIAGNOSIS

Disease Group	Percentage of errors
Infective and parasitic	13.8
Neoplasms	16.0
Endocrine, nutritional, and metabolic	9.4
Blood and blood-forming organs	19.1
Mental disorders	24.1
Nervous system and sense organs	17.1
Circulatory system	22.9
Respiratory system	12.0
Digestive system	11.4
Genitourinary	11.2
Complications of pregnancy, childbirth, and the puerperium	48.2
Skin and subcutaneous tissue	15.2
Musculoskeletal system and connective tissue	19.8
Congenital anomalies	16.0
Perinatal morbidity and mortality	4.2
Symptoms and ill-defined conditions	6.2
Accidents, poisonings, and violence (N)	16.0
All causes	15.4

DISCUSSION

MACHINE CODING

Howell and Loy (1968) described two types of machine coding error which might arise; firstly, errors through incompleteness in the dictionary and, secondly, errors arising through mistakes by human coders when compiling the machine dictionary. They provided an error detection system but they suggested that these errors would diminish with experience and with more complete dictionaries. In the event, in this trial these types of error gave rise to miscoding in less than 1% of the total.

Advantage was taken originally of the availability of a large machine (half a million bytes of core store) to test the underlying concept of the fruit machine method. Emphasis was less on the best way of programming than on early results. For example, although in theory better methods existed, in practice binary search by word length was arbitrarily chosen because this approach could be readily programmed. Coding speeds of 2,000 diagnoses a minute were subsequently obtained.

However, it was clear that most hospitals could not expect a machine of this size to be available to them, and more sophisticated programs, conserving storage, would be necessary, once feasibility of the basic concepts had been demonstrated. When reprogramming the original Codioc version for a medium-sized computer, Greenwood (1971) not only looked at other programming methods for his Kodioc but also relaxed one important condition in the original program. Originally this required the identification of a code number common to all individual diagnostic words, other than connective words such as 'from', 'and', 'the', and so on, which were deleted by the program, or words such as 'acute', 'chronic', 'left', and 'right' which only occasionally affected the code number when only the code numbers which were relevant were stored. Greenwood therefore chose a 'best fit' method rather than the original 'exact fit'; this relaxation meant some storage savings at the expense of an increase in the proportion miscoded (miscoding often restricted to the fourth digit), which in any case would be far less than the error rate of manual coders. Those who regard this error rate (3.5%) as too high have an option to revert to the original method requirements. In any event, the error rate will drop as the dictionary is updated to become more comprehensive. The favourable machine error rate in this survey may well be accepted as justification for Greenwood's concept.

No attempt has been made so far to code diagnoses which are age or sex dependent, such as pelvic inflammation (unless the sex is stated in the diag-

nosis, e.g., female sterility). The machine has been left to 'fail safe' by not coding. Since age and sex are normally included in the administrative data, it would be possible to read this in conjunction with the diagnosis. However, the failure rate for these two causes was very low.

A third form of error occurs when a linkage is required between two or more diagnoses. For example, angina with hypertension is correctly machine coded as 413-0, but where this phrase is presented as two diagnoses, the computer will code angina 413-9 and hypertension 401-. There are reasonable ways of resolving this problem in an operational system, but the manual coder, too, missed some of these linkages (section 2a, Table I). One underestimated source of error in this survey was where both machine and manual coder made the same mistake and thus agreed. In practice it was found that this error was virtually restricted to linkage failures which the machine at this stage does not purport to cover. Since a high proportion of manual coders ignored the linkage rules, and because electronic data processing will be increasingly used in the future, it may be that future revisions of the ICD should ignore or simplify such linkages and rely on machine methods for retrieving linkages and combinations.

The number of cases where the computer failed to code because of shortages in the dictionary will clearly diminish with further updating but there are some rare or unorthodox diagnostic expressions which may not be worth updating because the storage involved is not used frequently enough. Operational experience will clarify this situation, but it is always anticipated that expert manual assistance will be required in a small proportion of cases, leaving bulk routine coding to the machine, which, incidentally, will also code patients' occupations etc. from the same dictionary. No occupation encountered so far is confused with a diagnosis.

In this trial, one difficulty which was encountered was that of abbreviations such as MVDVD (mitral valve disease, aortic valve disease), partly because this specific terminology was not in the dictionary, but also because there are some difficulties in coding this expression (using the Greenwood 'best fit' concept) which the machine regards as six one-letter words. DOWNDATING the dictionary will ensure that the miscoding which arose from abbreviations (Table I, section 2c) does not occur again. About half of those miscodings related to the term CCF which usually meant congestive cardiac failure, though exceptionally coders used or interpreted it as congenital club foot or congenital cardiac

failure. If abbreviations such as C T E V (congenital talipes equino varus) were submitted as one word, either as CTEV or C.T.E.V., then these difficulties would disappear. Clearly some simple rules are necessary for abbreviations in a machine system and also for ensuring consistency in usage among manual coders.

MANUAL CODING

Probably the most striking finding of the survey was the high failure rate of manual coders. This figure of 17% may well be higher than expected, partly because the coder has knowledge from the notes which cause him to modify some diagnoses, but there are many instances where the coder has not used this knowledge. For example, there are cases admitted with symptoms such as abdominal pain where the diagnosis is retained and coded after definitive surgery such as hysterectomy. Where the coder uses extra knowledge to amend the diagnosis for coding he should be required to alter the diagnosis on the summary and S.M.R. 1 sheets so that this is apparent. Certainly extra knowledge was not the major factor in this high failure rate, for there were many examples where the coder was completely wrong. Duodenal ulcer with perforation was consistently coded by some hospitals as without perforation and all herniae were coded by some coders as with obstruction. One or two coders were under the impression that carcinoma *in situ* of the cervix should be coded as a malignancy, and frequently patients admitted for sterilization were coded as 'sterility'—a very different diagnosis!

Greenwood (1972) found a manual coding error of 8% with central coding and 20% with local hospital coding; he found that machine errors were roughly one-fifth of those produced by manual coders. Howell (unpublished data), in two small surveys, found error rates of 6% and 8% for central coders. The Oxford Record Linkage Study (1971) recently found in a sample of input data that their highest error coding rates were for diagnosis (7.0%) and operation (4.9%). Half of the errors were incorrect basic codes, the remainder being omissions or errors of qualification usually associated with the fourth digit. Lockwood (1971), on a smaller sample also using S.M.R.1, but for 1969, found a manual coder's error of less than 2%.

This study and that reported by Lockwood (1971) are different in their intentions and are not strictly comparable. Lockwood's study was concerned, among other things, with abstraction of diagnoses from the clinical notes and their coding according to the ICD. An omission of a secondary or tertiary diagnosis or part of the relevant information in a

primary one, or a misplacement, could conceivably have altered the coding. Part of the discrepancy between these two studies may have arisen from different sampling methods; Lockwood's sample was taken as 'representative of all specialties' from hospitals of 100 or more beds, whereas this study derived from a random sample of hospital discharges during the first seven months of 1968. Nevertheless, the differences found between these two studies appear to be too large to be entirely explained in this way. As a result, with the co-operation of Dr. M. A. Heasman, a further study was carried out. A subsample of 500 S.M.R. 1 forms drawn from the 25,000 used in this study, and a similar-sized sample of 1971 forms was examined and checked by an experienced manual coder. (Forms relating to intervening years were not available as they are destroyed after punching.) Discrepancies between the hospital codes and those found on checking were then categorized (Table III).

Included in the group 'Correctly coded (non-standard ICD)' are cases

(a) which are difficult to code but are probably correct, e.g., swelling in the groin coded to 782.7 (enlarged lymph node); and

(b) where there have been 'local' departures from ICD coding or where the diagnosis alone was incorrectly coded but taken with the stated operation becomes correct (e.g., femoral fracture with plate removal operation correctly coded to Y32.0).

Included in the 'possibly correctly coded' group are 11 1968 cases recorded as cerebro- or cardiovascular disease *with* hypertension although there was no record of hypertension on form S.M.R. 1. In 1971 only one case of this type was found in the sample.

Cases in the incorrect group vary from those with trivial errors, which just qualify for this group, to serious and obvious miscodings or misplacement of digits. There is clearly a subjective element in this type of classification, and there is a need for a large, current, detailed study of the types and severity of errors made by manual coders together with some method of subsequent feedback to obtain greater standardization. On the other hand, machine coding

TABLE III
COMPARISON OF HOSPITAL AND CENTRAL CODING ON
1968 AND 1971 DATA

Diagnoses	1968 No.	1968 %	1971 No.	1971 %
Correctly coded (ICD)	503	81.4	600	92.4
Correctly coded (non-standard ICD)	5	0.8	15	2.3
Possibly correct	42	6.8	13	2.0
Probably or certainly incorrect	68	11.0	21	3.2
Total no. of diagnoses	618	100	649	100
Total no. of S.M.R. 1 forms	500		500	

gives standardized coding which can be monitored automatically.

Of the cases with incorrect coding, one case in 1968 and five in 1971 involved a three-digit code number which was inadvertently right-adjusted in a four-digit punched card field. These cases would subsequently be rejected by the standard computer feasibility check.

If the 'possibly correct' and 'probably incorrect' items from Table III are regarded pessimistically as all demonstrating coding errors and are combined, then a maximum sample error for manual coders is obtained (Table IV). Similarly, scrutiny of these maximum sample errors for classification into serious errors and minor errors (e.g., fourth digit errors, poor descriptions, etc.) then the number of major errors was only 26 in the 1968 subsample and 8 in the 1971 sample (Table IV). The 1971 sample of major errors reveals a situation very similar to that described by Lockwood (1971). Equally, the computer error rate is largely made up of fourth digit errors which will rapidly decline with operational experience and with dictionary updating.

Between 1968 and 1969 (and hence 1971) two important changes occurred in Scottish hospital inpatient statistics. Computer printout on feasibility and validity checks was returned to hospitals together with appropriate forms for correction; secondly, data relating to individual consultants were distributed (Heasman, 1970; Heasman and Carstairs, 1971). These changes could be expected to have had a salutary effect on hospital coding. Further, training courses for medical records staff were instituted by the Scottish Administrative Staffs Committee in 1969.

However, a very probable reason for differences between the main part of this study and Lockwood's findings is that this survey used data from the first seven months of 1968, the year of introduction of the 8th revision of ICD. During this period coders were using the tabular list only, since the index volume was not available and did not reach Scottish hospitals until July 1969. It was, however, in use during the period of Lockwood's study. This explanation must be somewhat speculative, and would require a more definite study to ascertain the present standard in Scottish hospitals, but it is almost

certain that machine coding will always be more accurate and consistent than manual coding. As the Oxford Record Linkage Study (1971) puts it: 'It matters little what (code) numbers are assigned, provided the same numbers are always assigned to the same category whoever codes, and that the correct numbers are known to those responsible for retrieval'. The machine has the ability to allocate known numbers consistently. Manual coding clearly needs frequent monitoring.

Costs

It is difficult to make any realistic assessment of costs for the two different methods, partly because the cost of manual coding is difficult to ascertain. Trout (1971) has shown that in one regional hospital board area only 35 cases per day are coded per coding clerk; this would be equivalent to one full-time coder for 7,500 to 10,000 discharges per year, which is quite a standard ratio. Manual coding is thus an extremely costly business but in many hospitals the coder also arranges the individual sheets in a record, after discharge, in a laid-down sequence. In many hospitals, the cost of pulling and filing records before and after coding is carried out by filing clerks and is not costed against coding. In central S.M.R.1 operations, to be added to the cost of local diagnostic coding is the cost of coding the consultant's name and the patient's occupation plus the cost of card punching and verifying. Experimental work has shown that in a computer coding system occupation can be coded from a combined dictionary of occupations and diagnoses. There is no reason why the consultant's name should not be coded on the same run; the whole operation can be combined in a card-to-tape sequence. On the other hand, there is considerably more card punching involved (partly offset by the ability to retrieve the original plain language) in entering the basic material in the system instead of just numeric data. It must also be stressed that in this present survey approximately 10% of the entries failed at the first attempt at machine coding because of spelling errors. It is fair to add that these unverified cards were punched by operators (some trainees) with absolutely no experience of medical terminology and that some of the writing on the forms was quite atrocious. However, it is equally clear from this survey that given some control over the standard of input, particularly of spelling and of abbreviations, the fruit machine system will code 95% of diagnoses of the variety encountered, with a progressively low error rate after operational experience. Howell (1970) has said 'I think that automatic coding will really prove its worth when optical character

TABLE IV
SUMMARY OF MANUAL CODERS' ERRORS

	1968		1971	
	Subsample No.	%	Sample No.	%
Maximum estimate of coders' errors	110	17.8	34	5.2
Serious coding errors	26	4.2	8	1.2

recognition (OCR) is developed a little further'. The experience of this survey tends to re-inforce this view. If the S.M.R.1 were typed by a medical secretary, then not only could carbon copies provide common information as part of the hospital summary sheet and the letter to the general practitioner, but the original would give OCR input, thus obviating not only the cost of manual coding but also the cost of punching and verifying cards. However, the cost of OCR in this type of exercise is not yet established, but consistency of the machine system may be the paramount consideration. Baldwin (1971) has found with hospital discharge data and manual coding that OCR is economic provided that both the OCR machine and the application are carefully chosen.

SUMMARY

A comparison has been made of the effectiveness of computerized and manual methods of coding diagnoses in a sample of just over 25,000 discharges from Scottish hospitals. Machine coding gave correct coding in 92% of discharges compared with 83% by manual methods. Of all cases 17% were wrongly coded by manual methods as opposed to a comparable figure of 3½% for machine coding. Since 1968 there is evidence that this manual error rate in Scotland has fallen considerably, and reasons for this are discussed. The computer failed to code nearly 5% of the diagnoses.

The results obtained suggest that for large-scale coding of hospital morbidity data, central machine coding has distinct advantages over local manual coding. The value of automatic monitoring in a machine system should not be overlooked. The advantages of optical character recognition input are also discussed.

There were no unexpected or unexplained failures of method in machine coding, which, unlike manual coding, was entirely consistent in returning the same code number for a particular form of diagnostic terminology.

I am grateful to Dr. M. Heasman and Dr. J. Donnelly for suggesting this study and for their subsequent

assistance and encouragement. Thanks are also due to their department (Research and Intelligence Unit, Scottish Home and Health Department) for a grant to cover the cost of card punching, and for the sample of S.M.R. 1 forms. Particular thanks are also due to the M.R.C. Computer Unit (London); to Dr. C. C. Spicer, the Director, for machine time and every assistance; to Dr. R. M. Greenwood for the use of his Kodiak program and considerable discussion; and to Mrs. Angela Mott for programming modifications and advice. Miss R. M. Loy, of the Office of Population Censuses and Surveys, very kindly agreed to assess doubtful diagnostic code numbers produced by manual coders and computer.

This study was possible only through the tolerance and encouragement of Dr. K. P. Duncan, Chief Medical Officer, British Steel Corporation.

Finally, in spite of all the encouragement and assistance which almost did the work for me, any opinions expressed are entirely personal.

REFERENCES

- BALDWIN, J. A. (1971). Personal communication.
- GREENWOOD, R. M. (1972). Kodiak, a system for disease coding. *Int. J. biomed. computing*, in press.
- HEASMAN, M. A. (1970). Scottish consultant review of in-patient statistics (SCRIPS). *Scot. med. J.*, **15**, 386.
- and CARSTAIRS, V. (1971). Inpatient management. Variations in some aspects of practice in Scotland. *Brit. med. J.*, **1**, 495.
- HOWELL, R. W. (1968). In *Proc. 5th int. Congr. Medical Records*. Swedish Medical Records Association, Stockholm.
- (1970). In *Proc. 2nd European Conf. Medical Records*. Association of Medical Records Officers, Great Britain and Nederlandse Vereniging Van Medische Administrateurs, The Hague.
- and Loy, Ruth M. (1968). Disease coding by computer; the 'fruit machine' method. *Brit. J. prev. soc. Med.*, **22**, 178.
- LOCKWOOD, E. (1971). Accuracy of Scottish hospital morbidity data. *Brit. J. prev. soc. Med.*, **25**, 76.
- Oxford Record Linkage Study (1971). A medical information system for general hospitals inpatients. Research Report No. 1.
- TROUT, K. (1971). Personal communication.