Brit. J. prev. soc. Med. (1973), **27**, 59-62

AUTOMATIC DISEASE CODING: THE 'FRUIT-MACHINE' METHOD IN GENERAL PRACTICE

H. P. DINWOODIE

General Practitioner, Edinburgh

R. W. HOWELL

British Steel Corporation, Grosvenor Place, London S.W.1

Howell and Loy (1968) have described a working system of automatic coding of medical diagnoses (which Howell (1968) first described as the 'fruit-machine' method) and its use on hospital morbidity and mortality data. At about the same time Scott (1968) described experimental work based on a similar approach to diagnostic phrase analysis. The present paper attempts to ascertain whether comparable results can be achieved with general practice data, and to assess the potential of the method in this field.

Automatic disease coding permits a diagnosis to be recorded in plain language; its allocation to a numerical code is then performed automatically, instead of manually by the customary clerical coding process.

In the conventional fruit machine the 'jackpot' is obtained when the lemons appear in line. Similarly, 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. Failure to obtain a common code number means failure to code. In the main fruit-machine dictionary each significant word of a diagnosis is stored together with all the code numbers with which it has been found to be associated.

Objectives

Four objects were set out by Howell (1971) in his analysis of 25,000 Scottish hospital discharge summaries:

(1) to assess the present accuracy of manual coding over various diagnostic ranges;
(2) to compare the accuracy of machine and manual coding;
(3) to assess the feasibility of future machine coding on 'real' data; and
(4) to provide further information which would be used to update, and make more comprehensive, the machine 'dictionaries'.

These same objectives were taken as guidelines for the present small study in relation to general practice data; the second object formed the focus.

Material

A small collection of Scottish general practice morbidity data (recently collected for another purpose) (Dinwoodie, 1972) was utilized as this had the advantage of being readily available, multi-observer in origin, already manually coded, and unbiased in the sense that it was not collected specifically with this present study in mind.

Sixteen doctors—eight from different practices in the south-east of Scotland and eight from the south-west of Scotland—had recorded all attendances at both consultations and visits over a short period (eight days each), so that a reasonable variation in individual terminology was available. These diagnoses had already been coded manually, by a single clerk, using the Eighth Revision of the International Classification of Diseases (ICD). The first half of the survey data was used for this study.

Method

The full diagnosis was punched in plain language into a standard 80 column punched card, together with the manually assigned code number, and was then submitted for automatic coding, and comparison with the manual code, as described by Howell (1971). Where there was a difference between the machine and manual code numbers, individual scrutiny allowed a decision as to which was correct. One prior exclusion was made—of data manually coded to a 'Y' classification of the ICD, since the original machine dictionaries did not, as constructed, include these. There is, however, no technical difficulty in incorporating this section of the code. The 216 diagnoses thus excluded were largely of five main categories: life insurance and

Table

COMPARISON OF COMPUTER AND MANUAL RESULTS
CODING OF HOSPITAL DISCHARGE AND GENERAL PRACTICE MORBIDITY DATA

	Machine Coding				Manual Coding			
	Hospital Discharge Data		General Practice Data		Hospital Discharge Data		General Practice Data	
	No. of Patients	% of Total	No. of Patients	% of Total	No. of Patients	% of Total	No. of Patients	% of Total
Coded correctly								
1a Machine and manual in full agreement ..	18,710	74·3	1,251	82·7	18,710	74·3	1,251	82·7
1b Machine only correct	3,984	15·8	183	12·1				
1c Manual only correct					1,665	6·6	63	4·2
Subtotal ..	22,694	90·1	1,434	94·8	20,375	80·9	1,314	86·9
1d Insignificant differences between machine and manual	461	1·8	12	0·8	461	1·8	12	0·8
Total section 1.. ..	23,155	92·0	1,446	95·6	20,836	82·8	1,326	87·7
Coded incorrectly								
2a Linkage failure ..	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	3	0·2	4,286	17·0	186	12·3
Total section 2.. ..	870	3·5	3	0·2	4,341	17·2	186	12·3
Failure to code								
3a Poor or inadequate diagnoses	171	0·7	20	1·3				
3b Operations—diagnoses not given..	**	**	14	0·9				
3c Abbreviations ..	158	0·6	5	0·3				
3d Sex-dependent diagnoses	23	0·1						
3e 'Non-diagnoses' ('pregnancy', 'social reasons', etc.) ..	140	0·6	*	*				
3f Inadequate machine dictionary ..	660	2·6	24	1·6				
	1,152	4·6	63	4·2				
Grand total	25,177	100·0	1,512	100·0	25,177	100·0	1,512	100·0

* These data (coded by manual coder by the 'Y' supplementary classification of the ICD) were not supplied to the machine as the dictionary did not yet include this section (216 patients).
** Included in 3a.

other medical examinations (25); hay fever injections, smallpox vaccination, etc. (41); antenatal examinations (80); and examinations related to oral contraceptive prescriptions (24). None of the terminology used would present difficulties in machine coding. Some errors and inconsistencies were noted in the manual coding but were not evaluated as these cards were excluded from the survey.

Results

The results from this study are set out in the accompanying Table together with the original hospital discharge data.

Section 1 of the Table suggests that machine coding is of the order of 8% to 10% more effective than manual coding, in obtaining correct results, for both sets of data.

Section 2 suggests that manual coding errors can run between 10% and 20% of the total in both sets of data. This is similar to the findings of Greenwood (1972b) in other surveys.

Section 3 suggests that the machine's 'failure to code' rate is no greater (and may even be lower) on general practice data than on hospital discharge data. The manual coding error rate was fairly consistent throughout the whole diagnostic range and showed no great weakness in any particular section.

Discussion

Much the same problems arise in the coding of general practice and hospital discharge data. Coding is a repetitive, voluminous, tedious, error-prone clerical chore which seems—at least in theory—ripe for automation; it is a costly and time-consuming extra step in 'traditional' data processing which could with advantage be simplified.

A correct machine coding rate of over 95% has been obtained with this (admittedly small) sample of data, thus confirming Howell's (1971) contention that it is almost certain that machine coding will be more accurate and consistent than manual coding.

'Probably the most striking feature of the (original) survey was the high failure rate (17%) of manual coders'—Howell (1971). The present figure (12%) shows a distinctly higher standard of coding, perhaps because practice terminology was less 'complicated', with a higher percentage of some (but not all) common conditions, but almost certainly in part because the manual coding was performed centrally rather than peripherally. Encouraging also were the similar rates for 'poor or inadequate diagnosis', 'abbreviations', etc. When the first three digits only of the normal four-digit coding system were considered, the manual coding error fell to just over 7% compared with 11% in the hospital discharge data.

Greenwood (1972a) has also studied the problems and advantages of automatic disease coding, using death certificates and hospital discharge data. He, too, favours a method based on the breakdown of phrases, such as the fruit-machine method, rather than a straightforward 'table look-up' procedure, as used by Kennedy *et al.* (1968). These present results suggest that the same considerations apply for general practice morbidity data.

The use of automatic disease coding disposes of any controversy as to whether the doctor or his secretary should undertake manual coding: it disposes also of much of the controversy surrounding the relative ease of use of different disease classifications (College Code or ICD ?).

The fruit-machine method can be envisaged with any means of data input, whether by punched cards, paper tape, teletype, visual display unit, or optical character recognition (of documents typed in the surgery) either centrally or peripherally. Greenwood (1972b) has shown that the method can be adapted from the original large computer (2,000 diagnoses a minute coded) to a medium-sized machine (of the order of 16K available store with 24 bit words), still keeping the speed of coding within the bounds of practical usage (60 diagnoses a minute). Greenwood feels there is room yet for improvement in coding speed on the medium-sized machine.

Costs

It is difficult to make any comparison in coding costs with the two methods, partly because the cost of manual coding is difficult to ascertain for reasons given by Howell (1971); in hospitals it is certainly expensive. The cost of machine coding will vary according to the type of input and whether or not the coding is part of a data storage computer system. The cost of the actual machine coding (at less than ¼p per diagnosis) is far less than by manual methods. The main cost in the machine system arises from the greater volume of input, with conventional punching methods, when the plain language diagnosis is entered rather than the code number. The subsequent ability to retrieve the full diagnostic phrase rather than just a code number has some distinct advantages. With optical character recognition input this extra cost would largely be eliminated (Howell, 1970). In general practice recording of the present manual type, machine coding would probably be more expensive than manual coding. Where, however, accuracy and consistency are important, such as research in general practice, there may nonetheless be a strong case for automatic coding. 'Quality of data' would be considerably improved. In the sometimes contemplated general practitioner display unit recording system, manual coding would appear an anachronism.

Conclusion

It is concluded, therefore, that the problems and benefits of automatic coding of general practice data are similar to those of hospital discharge summaries and are centred on machine costs versus the undoubted efficiency and speed of machine coding. Nevertheless it seems desirable to illustrate the principle of automatic disease coding as applied to general practice.

Summary

The fruit-machine method of automatic disease coding has been used on a small sample (1,512 patients) of Scottish general practice morbidity data. Results comparable with those from a larger sample of Scottish hospital discharge data have been obtained; 95·6% of diagnoses were correctly coded by the machine as opposed to 87·7% by a manual coder.

The machine coding system failed to code 4·2% of the diagnoses presented, a proportion considered acceptable in an operational system, and one which would diminish as dictionaries were updated and became more complete. Miscoding by the machine, at 0·2% of the total diagnoses, was a rare occurrence.

We are grateful to the Research Committee of the South-East Scotland Faculty of the Royal College of General Practitioners and to those doctors in both the West of Scotland and South-East Scotland Faculties who participated in the original morbidity recording exercise. We are also indebted to Dr. C. C. Spicer and to Dr. R. M. Greenwood (MRC Computer Unit, London) for machine time and valuable discussion.

REFERENCES

DINWOODIE, H. P. (1972). Morbidity recording (faculty report) *J. roy. Coll. Gen. Practit.*, **22**, 417.

GREENWOOD, R. M. (1972a). Statistical distributions of diagnostic phrases of morbidity and mortality. *Int. J. Biomedical Computing*, **3**, 43.

—— (1972b). 'Kodiac', a system for disease coding by a medium-sized computer. *Int. J. Biomedical Computing*, **3**, 123.

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.

—— (1971). A comparison of diagnostic coding by manual and computerized methods. *Brit. J. prev. soc. Med.*, **25**, 225.

——, and LOY, R. M. (1968). Disease coding by computer; the 'fruit-machine' method. *Brit. J. prev. soc. Med.*, **22**, 178.

KENNEDY, F., COX, A. G., GLEN, A. I. M., ROY, A. D., and SUNDT, C. E. (1968). A computer-based system for handling clinical data. In *Computers in the Service of Medicine*, p. 85. Vol. 1. Edited by G. McLachlan and R. A. Shegog. Oxford University Press, London.

SCOTT, A. M. (1968). Automatic coding of a diagnosis. In *Computers in the Service of Medicine*, Vol. II, p. 89. Edited by G. McLachlan and R. A. Shegog. Oxford University Press, London.

Requests for reprints to R. W. H.