

# BMJ

## Randomised crossover trial comparing the performance of Clinical Terms Version 3 and Read Codes 5 byte set coding schemes in general practice

Philip J B Brown, Victoria Warmington, Michael Laurence and A Toby Prevost

*BMJ* 2003;326:1127  
doi:10.1136/bmj.326.7399.1127

---

Updated information and services can be found at:  
<http://bmj.com/cgi/content/full/326/7399/1127>

---

*These include:*

### References

This article cites 9 articles, 5 of which can be accessed free at:  
<http://bmj.com/cgi/content/full/326/7399/1127#BIBL>

5 online articles that cite this article can be accessed at:  
<http://bmj.com/cgi/content/full/326/7399/1127#otherarticles>

### Rapid responses

3 rapid responses have been posted to this article, which you can access for free at:  
<http://bmj.com/cgi/content/full/326/7399/1127#responses>

You can respond to this article at:  
<http://bmj.com/cgi/eletter-submit/326/7399/1127>

### Email alerting service

Receive free email alerts when new articles cite this article - sign up in the box at the top left of the article

---

### Topic collections

Articles on similar topics can be found in the following collections  
[General practice / family medicine](#) (8098 articles)

---

### Notes

---

To Request Permissions go to:  
<http://group.bmj.com/group/rights-licensing/permissions>

To order reprints go to:  
<http://journals.bmj.com/cgi/reprintform>

To subscribe to *BMJ* go to:  
<http://resources.bmj.com/bmj/subscribers>

# Information in practice

## Randomised crossover trial comparing the performance of Clinical Terms Version 3 and Read Codes 5 byte set coding schemes in general practice

Philip J B Brown, Victoria Warmington, Michael Laurence, A Toby Prevost

### Abstract

**Objective** To determine whether Clinical Terms Version 3 provides greater accuracy and consistency in coding electronic patient records than the Read Codes 5 byte set.

**Design** Randomised crossover trial. Clinicians coded patient records using both schemes after being randomised in pairs to use one scheme before the other.

**Setting** 10 general practices in urban, suburban, and rural environments in Norfolk.

**Participants** 10 general practitioners.

**Source of data** Concepts were collected from records of 100 patient encounters.

**Main outcome measures** Percentage of coded choices ranked as being exact representations of the original terms; percentage of cases where coding choice of paired general practitioners was identical; length of time taken to find a code.

**Results** A total of 995 unique concepts were collected. Exact matches were more common with Clinical Terms (70% (95% confidence interval 67% to 73%)) than with Read Codes (50% (47% to 53%)) ( $P < 0.001$ ), and this difference was significant for each of the 10 participants individually. The pooled proportion with exact and identical matches by paired participants was greater for Clinical Terms (0.58 (0.55 to 0.61)) than Read Codes (0.36 (0.33 to 0.39)) ( $P < 0.001$ ). The time taken to code with Clinical Terms (30 seconds per term) was not significantly longer than that for Read Codes.

**Conclusions** Clinical Terms Version 3 performed significantly better than Read Codes 5 byte set in capturing the meaning of concepts. These findings suggest that improved coding accuracy in primary care electronic patient records can be achieved with the use of such a clinical terminology.

### Introduction

The capture of data in electronic health records is expected to improve clinical effectiveness, governance, and outcomes. However, the data collected must be accurate and consistent.<sup>1</sup> To help satisfy this quality requirement, the use of a standardised clinical terminology (a large knowledge based coding scheme) has been advocated.<sup>2</sup> The NHS developed such a

terminology called Clinical Terms Version 3, which was first released in 1994,<sup>3</sup> and is currently completing an evaluation of SNOMED Clinical Terms, which is an enhanced product merging Clinical Terms Version 3 with SNOMED RT (a terminology produced by the College of American Pathologists).<sup>4</sup> Although it shares some features with the earlier Read Codes coding scheme, Clinical Terms offers advantages of unlimited hierarchical depth, multiple relationships, pure concepts (concepts represented by unambiguous preferred terms, semantically correct synonyms with no duplication), and the opportunity to add detail with qualifiers (table 1).<sup>5</sup> Despite high levels of computerisation in UK general practice, surprisingly few developers of electronic patient record systems have adopted Clinical Terms; most still use the Read Codes 5 byte set. Implementing the use of a standard clinical terminology is a key element of the shared electronic record of the NHS information strategy, itself a critical component of the government's commitment to modernise the NHS.<sup>6</sup> The costs of such an implementation are likely to be considerable, yet there is little evidence that using a standard clinical terminology in primary care will accrue benefits.

Comparisons of different clinical coding schemes have mainly been conducted by coding experts looking at the schemes' coverage in relation to existing lists of terms.<sup>7-11</sup> No study has examined whether a clinical terminology improves the performance of coding electronic patient records by practising doctors in primary care. The main aim of this crossover study was to determine whether Clinical Terms Version 3 provides greater accuracy and consistency than Read Codes 5 byte set for coding electronic patient records by general practitioners.

### Methods

#### Setting and participants

The study was conducted by 10 general practitioners recruited from practices in urban, suburban, and rural environments in Norfolk. The general practitioners were a convenience sample recruited by personal contact from a local network of practices interested in research (SuNet). The 10 participants (one woman, nine men) had a median age of 47 (range 40-55) and had been qualified for a median length of 24 years

Editorial by  
Gardner

School of  
Information  
Systems, University  
of East Anglia,  
Norwich NR4 7TJ  
Philip J B Brown  
honorary lecturer in  
healthcare informatics

Humbleyard  
Practice, Hethersett,  
Norfolk NR9 3AB

Victoria  
Warmington  
research associate

Bacon Road  
Medical Centre,  
Norwich NR2 3QX  
Michael Laurence  
general practitioner

Department of  
Public Health and  
Primary Care,  
University of  
Cambridge,  
Institute of Public  
Health, Cambridge  
CB2 2SR

A Toby Prevost  
medical statistician

Correspondence to:  
P J B Brown,  
Humbleyard  
Practice, Hethersett,  
Norfolk NR9 3AB  
Pjbb@hicomm.  
demon.co.uk

bmj.com 2003;326:1127

**Table 1** Comparison of the main features of Clinical Terms Version 3 and Read Code 5 byte set

Feature	Description	Read Codes	Clinical Terms
Coded	Clinical concept labelled with a code	Yes	Yes
Preferred term	Authorised preferred unambiguous label for a concept	Yes	Yes
Unique concept based	A concept only has one code	(Yes)*	Yes
Synonyms	Facility to label a concept with extra synonymous terms	(Yes)†	Yes
Hierarchical	Concepts arranged in a tree form according to meaning (as "is a" or "type of" relationships)	Yes	Yes
Authorised	Produced by recognised authority	Yes	Yes
Available	Available free of charge to the NHS	Yes	Yes
Updatable	Regular release of updated scheme	Yes	Yes
Cross mapped	Explicit mappings to other coding schemes including ICD and OPCS-4	Yes	Yes
Multiple relationships	A concept can have one or more "parent relationships"	No	Yes
Unlimited hierarchy depth	No structural limitations in the depth of hierarchy to accommodate detail	No	Yes
Qualifiers‡	Availability of mechanism to add additional detail to core concept (such as specific site of excision of a skin lesion)	No	Yes
Size	Approximate number of concepts (1000s)	125	220

\*The fixed hierarchy in Read Codes 5 byte set requires that a small number of concepts have two codes to allow placement in more than one hierarchy (such as tuberculous meningitis (A130. and F004.)).

†The limited hierarchical space dictates that some synonyms are not semantically pure.

‡The use of qualifiers greatly enhances the expressivity of the terminology but is more complicated to implement consistently and was therefore not included in this study.

(range 11-33). They self ranked their computer literacy using a simple Likert scale: three reported being "very computer literate" and seven reported a "working knowledge." For their frequency of using coding schemes to record consultations, one reported "always," four reported "mostly," and five reported "occasionally." All 10 used the Read Codes 5 byte set; half had not heard of Clinical Terms Version 3, and the rest were aware of it but had no knowledge of its structure and content. None refused to participate in our study, and none dropped out after recruitment.

### Design

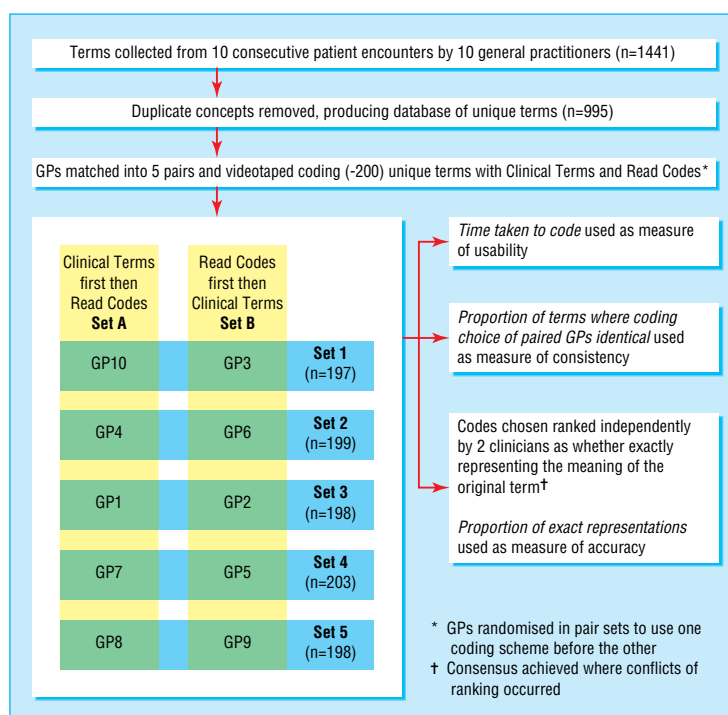
Each general practitioner manually recorded the consultation details of 10 consecutive patients in an arbitrarily chosen consultation session. A simple framework of headings was provided (reason for

encounter, diagnosis, treatment, and medical history) to prompt entry of details, but there was no restriction in the terms that could be recorded. The terms from these 100 records were then entered verbatim (except for correction of spelling mistakes) into an Access (Microsoft) database. We used random number tables<sup>12</sup> to group the general practitioners into five pairs and to randomly select one of each pair to code terms with Read Codes 5 byte set (termed Read Codes in this paper) first and then to code with the Clinical Terms Version 3 (termed Clinical Terms in this paper), and the other doctor in each pair to use the Clinical Terms first followed by Read Codes. We asked the clinicians to code the terms collected from their own records and those from the other doctor in their pair. Before this exercise, we identified and removed any duplicate concepts in the Access database, providing a body of about 200 terms for coding for each doctor (fig 1).

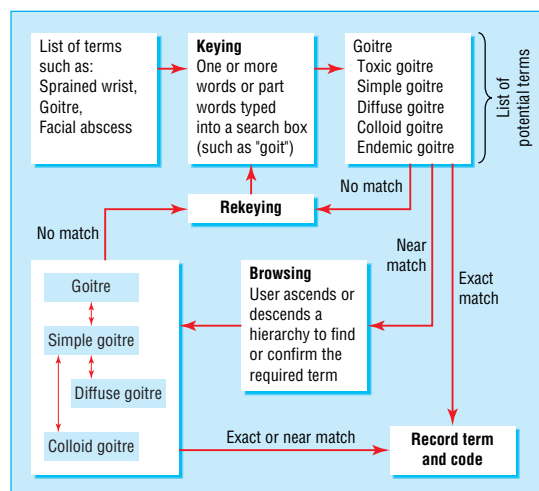
We videotaped each doctor coding his or her allocated file of terms using both Clinical Terms and Read Codes. To minimise any confounding variables from the human-computer interface, all the participants used the same software (NHS Information Authority Clinical Terminology Browser, March 2001 release) and laptop computer (Sony Vaio) to search the two coding schemes. Participants were given standardised instructions and training to identify a code for each term that would be "an acceptable match if the coded record were the only documentation of the concept in a paperless practice." They were encouraged to identify a match by searching for an appropriate term by keying in one or more words or part words and, when necessary, to browse the hierarchies of the coding scheme until a suitable equivalent was found, and then to record their match electronically by pasting their choice from the browser into an Access database (fig 2).

A researcher reviewed each video and recorded the time taken to code each term. Two researchers then independently examined the coded choices made by each general practitioner and ranked each match as exact or non-exact in representing the meaning of the original term. Differences in ranking between the two researchers were resolved by consensus reviews.

We estimated the accuracy of each coding scheme by calculating the proportion of coded choices ranked



**Fig 1** Summary of study design for comparing coding performance of Clinical Terms Version 3 and Read Codes 5 byte set



**Fig 2** Outline of main processes involved in coding medical records with a coding scheme

as being exact semantic representations of the original terms. We estimated consistency by identifying the proportion of cases where the coding choice of the paired general practitioners was the same. The length of time taken to find a code was used as a measure of usability of each scheme.

### Sample size

In a pilot study, the records of five patients generated 40 unique terms from a total of 42 terms. With one coding scheme 60% of the terms were matched exactly, and with the other scheme 70% were matched exactly, with 67% of terms being discordant pairs. Thus, in order to detect a 10% difference in exact matches between the schemes, 704 unique terms would provide 90% power using McNemar's test and 5% significance level. One hundred patients ought to generate 840 terms, which, after removal of duplicates, should be sufficient to provide 704 unique terms or, more certainly, 527 unique terms required for 80% power. The calculation was conservative in that, rather than being assessed by a single general practitioner, each term was assessed by two randomly paired clinicians and the average of their outcomes was used in the analysis.

### Statistical analysis

We chose statistical methods of analysis to be consistent with the paired nature of the design. For all outcomes, we performed stratified analyses within each of the five pairs of participants and then pooled the within pair estimates using weights proportional to the number of terms coded by each pair. We used Cohen's  $\kappa$  coefficient to assess agreement among participants in the exactness of coding under each scheme, with a value of  $\geq 0.6$  indicating good agreement.<sup>13</sup> We calculated the proportion of coded choices ranked as exact matches for both doctors in each pair and averaged the proportions as a repeated measures summary statistic before pooling them across the pairs. We calculated the 95% confidence interval from standard errors for a pair's averaged proportion, which was derived as half of the standard error of a difference in two paired proportions used in McNemar's test,<sup>14</sup> with results verified using the bootstrap method. We also used these methods to assess consistency between doctor pairs, defined as the

proportion of concepts coded identically and as an exact match by both doctors.

We calculated the difference between the two schemes in the time taken to code each entry for both doctors in each pair, and we used the average of the two differences as a repeated measures summary statistic in the analysis.<sup>13</sup> For those entries where only one doctor's time difference was available, this was analysed instead of the average. We used the bias corrected accelerated non-parametric stratified bootstrap method<sup>15 16</sup> with 5000 replications within S-Plus 2000 software to estimate 95% confidence intervals and P values for mean time differences because the time differences were inconsistent with a normal distribution. We also used this method to test for period effects and for carryover effects (scheme by period interaction), stratified by participant pairs using the testing approach based on difference measures for period effect and average measures for carryover effect.<sup>17</sup> All tests were two tailed and assessed at the 5% level of significance.

## Results

The 100 consultations generated a total of 1441 terms, providing 995 unique concepts after removal of duplicates. Findings (such as impetigo, upper respiratory tract infection) accounted for 730 of these concepts, with the remaining 265 representing procedures (such as hysterectomy, psychotherapy). As table 2 shows, the agreement among doctors in the exactness of coding under each scheme was good, with pooled  $\kappa$  coefficients of 0.69 (95% confidence interval 0.64 to 0.74) for Clinical Terms and 0.65 (0.60 to 0.69) for Read Codes.

### Accuracy of coding schemes

The proportion of concepts ranked as exact semantic representations with Clinical Terms ranged from 0.60 to 0.74 (pooled proportion 0.70) for the 10 participants, with seven of the doctors being in excess of 0.7. By contrast, the proportion of concepts ranked exact with Read Codes ranged from 0.37 to 0.58 (pooled proportion 0.50). All 10 doctors coded significantly more concepts as exact matches with Clinical Terms than with Read Codes ( $P < 0.001$  for each doctor). The excess proportion of concepts ranked exact with Clinical Terms ranged from 14% (95% confidence interval 7% to 21%) to 27% (19% to 34%) for the 10 participants. The excess proportion of concepts exactly matched with Clinical Terms was similar in the doctors who used this scheme before Read Codes (22%) and in those who used the scheme after using Read Codes (18%), although this relatively small difference represented a significant period effect. We also found a significant carryover effect, with proportions

**Table 2** Performance of 10 paired general practitioners using Clinical Terms Version 3 and Read Code 5 byte set to code terms extracted from 100 consultations. Values are pooled averages over the general practitioner pairs and proportions (95% confidence intervals) unless stated otherwise

Performance measure	Clinical Terms	Read Codes
$\kappa$ coefficient of interclinician agreement	0.69 (0.64 to 0.74)	0.65 (0.60 to 0.69)
Codes ranked as exact representations of original term	0.70 (0.67 to 0.73)*	0.50 (0.47 to 0.53)
Terms consistently coded	0.58 (0.55 to 0.61)*	0.36 (0.33 to 0.39)
Mean coding time (seconds)	30.2 (28.6 to 31.9)*	36.1 (34.3 to 37.9)

\* $P < 0.001$  for difference in performance between schemes.

of exact matches in the first period being 69% for Clinical Terms and 52% for Read Codes (excess 17% (95% confidence interval 13% to 19%)), and in the second period being 71% and 47% respectively (excess 23% (20% to 26%)).

#### Consistency of coding schemes

The percentage of concepts ranked consistent (that is, exact matches and coded identically by both members of a pair) ranged from 53% to 63% for Clinical Terms and from 31% to 43% for Read Codes. The excess in proportion ranked consistent with Clinical Terms ranged from 21% to 23% and was significant for each of the general practitioner pairs. The pooled proportion of consistent matches by general practitioner pairs was 0.58 for Clinical Terms and 0.36 for Read Codes, with a pooled difference in proportion of 0.22 (0.19 to 0.25) ( $P < 0.001$ ). A further 48 concepts in Clinical Terms and 80 concepts in Read Codes were coded identically by general practitioner pairs but not as an exact match of the original terms.

#### Usability of coding schemes

The median coding time for each of the 10 participants ranged from 14 to 27 seconds for Clinical Terms and from 18 to 49 seconds for Read Codes. For 989 terms (99%), either both (85%) or one (14%) of the general practitioner pairs had timing data recorded. Compared with Read Codes, the mean excess time taken to code with Clinical Terms ranged from -29 to 12.3 seconds for the pairs of participants. The mean time taken to code with Clinical Terms was shorter by a mean of 5.9 seconds (4.0 to 7.9), being significantly shorter in four pairs (by 13, 6, 3, and 7 seconds) and not significantly different in the remaining pair (0.5 seconds longer). However, on the basis of the 850 terms with full data available, there were significant period and carryover effects. In the first period mean coding times were 28.1 seconds for Clinical Terms and 42.1 seconds for Read Codes, and in the second period they were 30.5 seconds and 29.3 seconds respectively. Compared with Clinical Terms, mean coding time with Read Codes was significantly longer in the first period, by 14.0 seconds (11.2 to 17.0), and non-significantly shorter in the second period, by 1.2 seconds (-1.3 to 3.8).

## Discussion

Clinical Terms Version 3 performed significantly better than Read Codes 5 byte set in capturing the meaning of concepts required to describe the electronic records of 100 patients in primary care by 10 general practitioners. This superiority is partly a function of the larger size of Clinical Terms (about 220 000 concepts, compared with about 125 000 concepts in Read Codes), but the better accuracy and consistency were also accompanied by a shorter average browsing time to find the required coded terms. The 10 participants had little or no prior knowledge of Clinical Terms, and so it is particularly impressive that they should achieve better coding performance with an unfamiliar coding scheme. These findings suggest that improved coding accuracy in primary care electronic patient records can be achieved with the introduction of a clinical terminology.

#### Strengths and limitations of our study

We compared the content and usability of the two coding schemes in a practical setting, where clinicians

had a variable degree of competency in coding. While formulating the study, we considered videotaping the coding process during live patient consultations. We rejected this in favour of a randomised crossover trial as consistency between experimenters would have been difficult to assess and confounding variables such as time constraints on searching would have been difficult to control. Coding performance and times are therefore only proxy estimates of use in real patient encounters. Further improvement of data entry might be achieved with more sophisticated software than was used in our study—such as by using templates for data entry and menus to access commonly used terms.

We compared Clinical Terms Version 3 with Read Codes 5 byte set rather than the earlier Read Code 4 byte set, which is still in use, because an earlier study of coding performance in secondary care had indicated that Read Codes 5 byte set was superior in coverage than the earlier scheme when tested against a set of 2624 concepts.<sup>18</sup> Clinical Terms Version 3 also has the ability to support the construction of more detailed concepts using a mechanism of qualifiers; for example, the core notion of “skin abscess” can be qualified by its exact site with reference to a detailed anatomy chapter. This functionality provides great expressivity, and we excluded it from consideration in this study as it would have afforded an unfair comparative advantage and its influence would be heavily dependent on software implementation and user familiarity and skill. The value added by use of this qualifying mechanism merits further investigation.

We reduced confounding variables by using a randomised crossover trial and the same browser for searching both schemes. External validity was improved by involving several general practitioners. We identified carryover effects using tests that were sensitively based on within pair comparisons of general practitioners coding the same terms. The carryover effect in the proportion of terms exactly matched was small compared with the size of the difference between the two coding schemes in each period. The carryover effect for coding time reflected the change from the first period to the second period in the difference between the schemes, from 14 seconds shorter to 1 second longer for Clinical Terms compared with Read Codes. Four of the five doctors who coded first with Read Codes took more than 10 seconds longer to code each term than they did with Clinical Terms; review of the video comments of the remaining participant suggested that the longer coding time in the second period related to user fatigue (including remarks about the doctor’s uncertainty of meaning of the original term and technical difficulties in using the notebook keypad). Only one of the participants who coded first with Clinical Terms took more than 10 seconds longer to code with this than with Read Codes, and this may be accounted for by the doctor’s familiarity with the content of Read Codes. The small number of participants limits our ability to explain such differences with certainty, and we have cautiously interpreted them to say that the time taken to code with Clinical Terms was not significantly longer than that with Read Codes.

We did not try to measure the potential clinical importance of the non-exact matches. Clearly the absence of a detailed variant of a concept (such as

### What is already known on this topic

Clinical terminologies such as Clinical Terms Version 3 have been shown to offer greater coverage than earlier coding schemes when used by experts for coding data in electronic patient records, mainly by virtue of their larger size

Implementing the use of a terminology in a health service will be costly, and evidence is needed that this will improve the quality of medical data

### What this study adds

When used by general practitioners Clinical Terms Version 3 performed significantly better than Read Codes 5 byte set in consistently coding the meaning of concepts for electronic patient records in primary care

Despite the larger size of Clinical Terms, the improvement in accuracy was achievable without an increase in the average browsing time to find the required coded terms

Improved coding accuracy and consistency in primary care electronic patient records can be achieved with Clinical Terms Version 3

“lipoma of neck” as opposed to just “lipoma”) is less important than the complete absence of a suitable concept (such as “monoclonal gammaglobulinaemia of uncertain significance”). Judging the importance of non-exact matching would have introduced a further subjective element, requiring further checks of inter-rater reliability that were outside the scope of the study, although our data provide valuable material for further study.

### Comparison with other studies

In previous reports assessing the content of Clinical Terms Version 3, terminology experts coded lists of pre-existing concepts and generated rates of completeness of 73%, similar to our findings.<sup>7-9</sup> Our study examined the performance of a clinical terminology against an established coding scheme by general practitioners (non-expert coders).

Cimino et al used videotaping to study 238 coding events in secondary care (using a terminology known as Medical Entities Dictionary): 71% of the codings captured the exact meaning of the required concept, with a mean coding time of 40.4 seconds.<sup>19</sup> These findings are similar to our results. Cimino et al also evaluated the reasons for suboptimal performance and described problems in the terminology content (13%), representation (10%) and usability (6%). These aspects were not part of our current study, which concentrated on assessing practical performance, but further work in identifying the reasons for failing to achieve an exact match in our sample could provide useful information for improving the content of Clinical Terms Version 3 and software design.

### Conclusion

The coding of clinical records is an important aspect of medical audit, research, epidemiology, management of resources, and the direct care of patients. For

information technology to be fully adopted, clinical notions that are often complex must be accurately and easily represented as coded concepts that are “user friendly” and easily retrievable. Our study suggests that substantial advantages may be achieved by investing in the implementation of Clinical Terms Version 3 or a similar terminology.

We thank Ian Harvey (Health Policy and Practice, University of East Anglia) for his advice in designing the study and Doreen Cochrane (SuNet facilitator, Health Policy and Practice, University of East Anglia) for invaluable input during the preparation of this proposal. The following clinicians participated in the study: Robert Bawden, Peter Burrows, Jamie Dalrymple, Stephen Daykin, Christopher Hand, Carly Hughes, Andrew Leaman, David Munson, Tony Press, and Rob Stone.

Contributors: PJBB formulated the study and ATP, ML, and VW contributed to the design, ATP through the Cambridge Research Development and Support Group. ML recruited the participating clinicians, and VW coordinated the project, created the database of terms, and performed the video analysis. PJBB and ML ranked the matches. ATP performed the statistical analysis. PJBB wrote first draft, and all authors contributed to the final draft of the paper. PJBB is the guarantor of the paper.

Funding: The study was funded by the NHS Eastern Region R&D grant No RCC33031.

Competing interests: PJBB advises the NHS Information Authority on coding and terminology.

Ethical approval: The study was granted ethical approval by the Norfolk and Norwich Ethical Committee.

- 1 Brown PJB, Sönksen P. Evaluation of the quality of information retrieval of clinical findings from a computerised patient database using a semantic terminological model. *J Am Med Inform Assoc* 2000;7:401-12.
- 2 Evans DA, Cimino JJ, Hersh WR, Huff SM, Bell DS, for the CANON Group. Towards a medical-concept representation language. *J Am Med Inform Assoc* 1994;1:207-17.
- 3 Severs MP. The clinical terms project. *Bull R Coll Physicians Lond* 1993;27(2):9-10.
- 4 Stearns MQ, Price C, Spackman KA, Wang AY. SNOMED clinical terms: overview of the development process and project status. In: Bakken S, ed. *Proceedings of the 2001 AMIA Fall Symposium*. Philadelphia: Hanley and Belfus, 2001:662-6.
- 5 O'Neil M, Payne C, Read JD. Read codes version 3—a user led terminology. *Methods Inf Med* 1995;34:187-92.
- 6 NHS Executive. *Information for health: an information strategy for the modern NHS*. London: Department of Health, 1998.
- 7 Hausam RR, Hahn AW. Representation of clinical problem assessment phrases in US family practice using Read version 3.1 terms: a preliminary study. In: Gardner RM, ed. *Proceedings of the 1995 AMIA Fall Symposium*. Philadelphia: Hanley and Belfus, 1995:426-30.
- 8 Chute CG, Cohn SP, Campbell KE, Oliver DE, Campbell JR, for the Computer-Based Patient Record Institute's Work Group on Codes and Structures. The content of clinical classifications. *J Am Med Inform Assoc* 1996;3:224-33.
- 9 Mullins HC, Scanland PM, Collins BS, Treece L, Petruzzi P, Goodson A, et al. The efficacy of SNOMED, Read codes, and UMLS in coding ambulatory family practice clinical records. In: Cimino JJ, ed. *Proceedings of the 1996 AMIA Fall Symposium*. Philadelphia: Hanley and Belfus, 1996:135-9.
- 10 Campbell JR, Carpenter P, Sneiderman C, Cohn S, Chute CG, Warren J. Phase II evaluation of clinical coding schemes: completeness, taxonomy, mapping, definitions and clarity. *J Am Med Inform Assoc* 1997;4:238-51.
- 11 Humphreys BL, McCray AT, Cheh ML. Evaluating the coverage of controlled health data terminologies: report on the results of the NLM/AHCPR large scale vocabulary test. *J Am Med Inform Assoc* 1997;4:484-500.
- 12 Fisher RA, Yates F. *Statistical tables for biological, agricultural and medical research*. 6th ed. Edinburgh: Oliver and Boyd, 1963.
- 13 Altman DG. *Practical statistics for medical research*. London: Chapman and Hall, 1996.
- 14 Fleiss JL. *Statistical methods for rates and proportions*. 2nd ed. Chichester: John Wiley, 1981.
- 15 Efron B. Better bootstrap confidence intervals. *J Am Stat Assoc* 1987;82:171-200.
- 16 Carpenter J, Bithell J. Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians. *Stat Med* 2000;19:1141-64.
- 17 Pocock SJ. *Clinical trials: a practical approach*. Chichester: John Wiley, 1983.
- 18 Brown PJB, Odusanya L. Does size matter?—Evaluation of value added content of two decades of successive coding schemes in secondary care. In: Bakken S, ed. *Proceedings of the 2001 AMIA Fall Symposium*. Philadelphia: Hanley and Belfus, 2001:71-5.
- 19 Cimino JJ, Patel VL, Kushniruk AW. Studying the human-computer-terminology interface. *J Am Med Inform Assoc* 2001;8:163-73.

(Accepted 5 March 2003)