Research Paper ■

Access to Data: Comparing AccessMed with Query by Review

GEORGE HRIPCSAK, MD, BARRY ALLEN, PHD, JAMES J. CIMINO, MD, ROBERT LEE

Abstract **Objective:** To evaluate the performance of tools for authoring patient database queries.

Design: Query by Review, a tool that exploits the training that users have undergone to master a result review system, was compared with AccessMed, a vocabulary browser that supports lexical matching and the traversal of hierarchical and semantic links. Seven subjects (Medical Logic Module authors) were asked to use both tools to gather the vocabulary terms necessary to perform each of eight laboratory queries.

Measurements: The proportion of queries that were correct; intersubject agreement.

Results: Query by Review had better performance than AccessMed (38% correct queries versus 18%, p = 0.002), but both figures were low. Poor intersubject agreement (28% for Query by Review and 21% for AccessMed) corroborated the relatively low performance. Subjects appeared to have trouble distinguishing laboratory tests from laboratory batteries, picking terms relevant to the particular data type required, and using classes in the vocabulary's hierarchy.

Conclusion: Query by Review, with its more constrained user interface, performed somewhat better than AccessMed, a more general tool. Neither tool achieved adequate performance, however, which points to the difficulty of formulating a query for a clinical database and the need for further work.

■ **JAMIA.** 1996;3:288–299.

One of the major successes of computers in health care has been the clinical information system's result reporting component,[1-3] which delivers data to providers when and where they need it, unencumbered by lost sheets, paper charts, and closed record rooms. Using these systems has nevertheless been a challenge for naive users. Most clinical information systems have adopted roughly the same system of menus to steer a provider to the appropriate data. New users rely on word of mouth, trial and error, and, occasionally, training to master the foreign-looking user interface. Most users persevere and prevail because of the sheer value of the data.

The use of these same data for decision support and clinical research has been on the rise, and this is the focus of our work. Knowledge-base authors and researchers must face yet another user interface to select the data they need. Several approaches to formulating queries for clinical databases have been reported. For example, some institutions have given users direct access to the underlying database, using the native tools of the database management system,[4] whereas others have created specialized query languages[4,5] and user interfaces[6] intended to better match the needs of clinical users. Much work in this area has concentrated on organizing the terms used to store and retrieve data into vocabularies based on simple hierarchies or semantic networks, which helps users select the terms needed in their queries.[7-11] Some research has focused on improving the understanding of the clinical domain, which should help in building tools that are more clinically relevant.[12-14]

Much has been published on the design, implemen-

Affiliations of the authors: Department of Medical Informatics, Columbia-Presbyterian Medical Center, New York, NY.

This work was supported by National Library of Medicine grant R29 LM05627, "Linking Knowledge-based Systems to Clinical Databases."

Correspondence and reprints: George Hripcsak, MD, Department of Medical Informatics, 161 Fort Washington Ave., DAP-1310, New York, NY 10032. e-mail: hripcsak@columbia.edu

Received for publication: 9/15/95; accepted for publication: 3/5/96.

tation, and theoretical advantages of such tools, but there have been few evaluations of whether the tools actually achieve adequate results; no evaluations have focused specifically on queries for automated decision support systems. ClinQuery is a tool for retrieving clinical data based on a set of menus. Users reported 58% definite or probable success when using the tool to retrieve data and answer clinical questions.[6] It is difficult to interpret survey results, however, because users are sometimes unaware of their errors and so judge the tools to be excellent.[15] The HERMES workstation is a tool that assists clinical researchers in retrieving data and performing analyses. A formal evaluation[15] revealed that, with the tool, clinicians answered 54% of research questions correctly and completely, and biomedical researchers answered 81% of research questions correctly and completely. While the tool addresses many steps in the clinical research process—selecting the data, picking the type of analysis, carrying out the analysis, and so on—the authors reported that selecting the data remained a difficult challenge. There has been work in the evaluation of clinical vocabularies,[7,16] but the focus has not been on their ability to select terms for generating queries to a clinical database.

At Columbia-Presbyterian Medical Center (CPMC), we have been using a tool called AccessMed to help knowledge-base authors and researchers get to the clinical data they need. It exploits a semantic network to steer the user to the correct terms needed to retrieve data. Because naive users have reported trouble using this tool, we created a new tool called Query by Review, which steers the user to the correct data by mimicking the institution's result review system. In this study, we measured the performance of both tools.

Background

Clinical Laboratory Data

At CPMC, a centralized patient database[17] contains data from a wide variety of areas: clinical laboratory, admit-discharge-transfer, pharmacy, discharge summaries, textual radiology reports, coded radiology findings, pathology, outpatient notes, and data from a number of ancillary departments. Laboratory data, which represent the largest volume and the most frequently queried data in the database, were the focus of our study because that is where our tools were the most mature.

The CPMC clinical laboratory assigns a unique code to all the tests it can perform. Most tests are grouped into batteries (panels). A single patient specimen usually undergoes a battery of tests, and the results are reported together. Examples of tests are the serum sodium test, a hemoglobin measurement, and a hepatitis B surface antibody titer. Examples of batteries are the chemistry panel (which includes the sodium test), an automated blood count (which includes the hemoglobin measurement), and a hepatitis panel (which includes the antibody titer).

The laboratory distinguishes among tests with a very fine level of granularity. For example, a serum potassium test performed at the main medical center has a different code than an otherwise identical test performed at a satellite hospital. Whenever the machines that analyze the specimen change, a new code is assigned. The same test performed as part of two different batteries is usually (but not always) given two different codes.

Clinical Database and Vocabulary

The central patient database[17] is based on a relational database management system. Laboratory data are organized into two tables: a header table and a component table. The header table contains information relevant to all tests, such as the medical record number of the patient, the time the test was received by the laboratory, the status, and the code for the battery that contains the tests. If an individual test is performed, then a dummy battery is stored. The component table contains the individual tests, including their codes, values, and, occasionally, subcomponents for nested data.

The codes stored in the database are defined in the institution's vocabulary, called the Medical Entities Dictionary (MED),[18] which is based on a semantic network. The vocabulary serves not only to define the codes but also to map the central database codes to the codes used in ancillary departments.

It is difficult to predict which of the laboratory's many distinctions are clinically relevant. For example, a change of equipment may result in a change of normal laboratory levels, which is clinically relevant. Therefore, the central patient database maintains all the laboratory distinctions by mapping the laboratory codes one-to-one to distinct central MED codes.

At the time of retrieval, all the codes relevant to a given query must be included. To facilitate this process, the vocabulary contains classes that group codes under clinically relevant concepts (e.g., the "serum potassium ion tests" class, which groups the serum potassium levels from several different hospital laboratories).

Automated Decision Support

At CPMC, a clinical event monitor[19] tracks events in the medical center, such as the storage of laboratory results, admission, discharge, transfer, pharmacy orders, and so on. Events trigger the execution of rules called Medical Logic Modules (MLMs), based on the Arden Syntax.[20] The MLMs read data from the patient database, evaluate a set of criteria, and, if appropriate, send messages to health care providers. The system has been used for clinical alerts, reminders, interpretations, and screening messages.

We have found that data retrieval is a critical challenge to the clinical event monitor in terms of knowledge-base authoring, maintenance, and performance[21] and to knowledge-base sharing.[22] The largest stumbling block to creating effective MLMs has been the writing of appropriate database queries. Knowing what is stored, where it is stored, and what it is called are the main challenges.

Creating a Query

There are several steps to creating a valid query. First, the user must specify what data are desired by selecting the appropriate terms (codes) for laboratory tests. A term may signify a particular test, or it may signify a set (class) of tests. As an option, the user may constrain the test result so that it is only retrieved if it is part of a particular laboratory battery. This is done where the same test may appear in two different batteries and the batteries are based on slightly different specimens (e.g., arterial and venous blood). In most cases, we assign different terms to tests when they appear in different batteries, which obviates the need to specify a battery constraint.

The next step is for the user to specify additional constraints (such as time constraints: "within the past month"), aggregation operators (e.g., last), and a choice of attributes to return (such as time, status, and value). The terms, constraints, and operators are handed to an interface program known as a "data access module" (DAM).[17] The DAM converts the information to a valid SQL query, which is applied to the patient database. In the final step, the result is returned to the user if it is an interactive query or to the MLM if the query is part of the decision-support system.

The two tools described here are intended to help the user carry out the initial step, selecting the relevant terms. Both tools employ a knowledge base of terms and a browser. The knowledge base can be described by the underlying formalism used to represent the terms and their relations (e.g., a semantic network) and by the general approach used to organize the terms (e.g., terms may be organized into medically relevant definitions). The browser allows the user to traverse the knowledge base and pick out the appropriate terms.

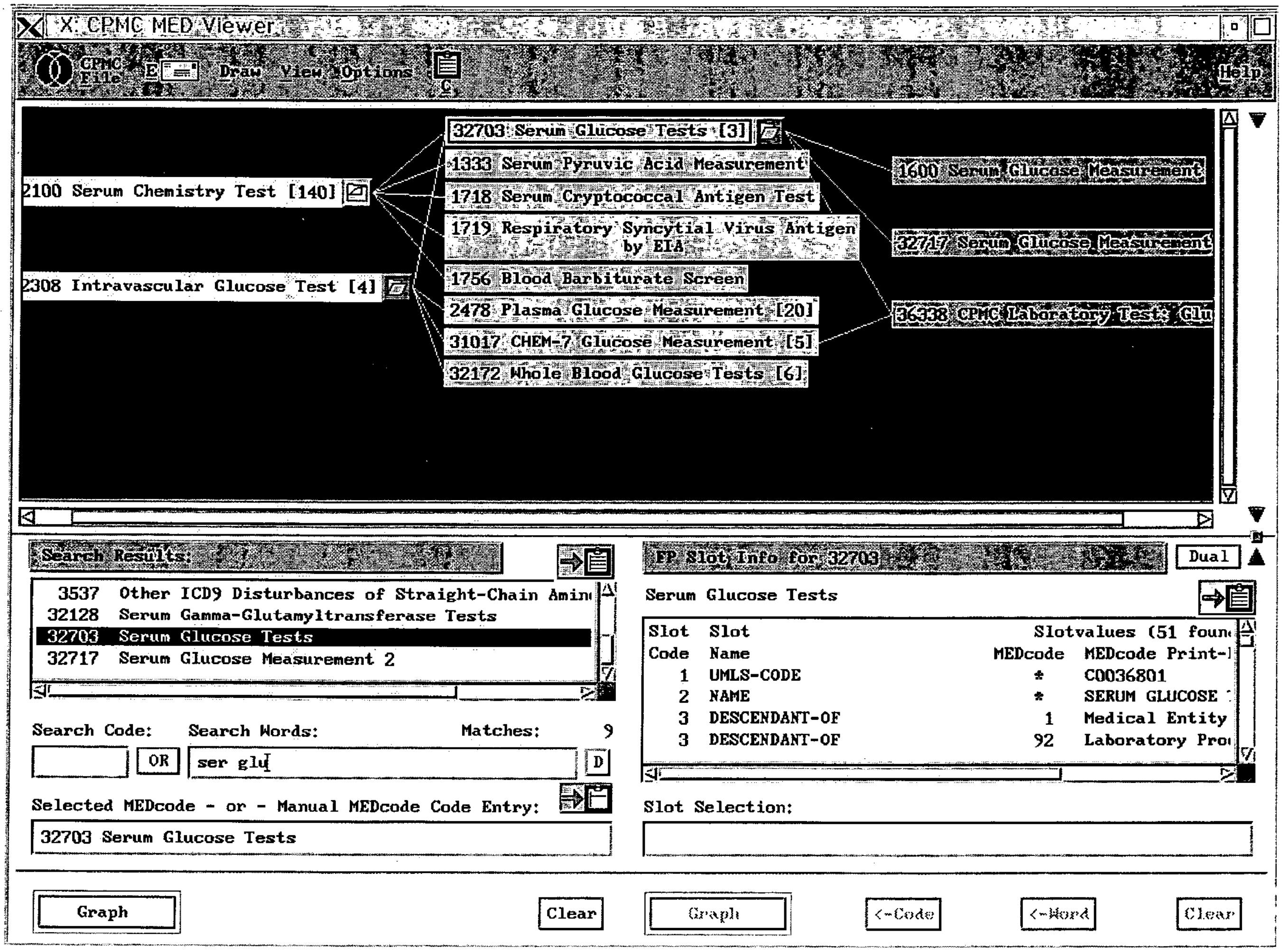
AccessMed

The first tool, AccessMed,[23] uses the Medical Entities Dictionary (MED) as its knowledge base. Its underlying formalism is a semantic network. Hierarchical (parent–child) links convey subclass ("is a") relations. Other links between nodes convey further semantic information, including "part of" links. In addition to these relations (links), nodes may have literal attributes and values.

Within the semantic network, terms are organized mainly by medically relevant definitions, which are conveyed through semantic links and attributes. For example, the term "serum glucose measurement" has as its parent the class "serum glucose tests," and as its grandparent the class "intravascular glucose test." There is a "part-of" link between it and the chemistry battery "Presbyterian SMAC"; there is a "substance-measured" link between it and the term "glucose"; and there is a "specimen" link between it and the term "clinical chemistry serum specimen." It has an attribute called "CPMC lab test name," which has the value "GLUC."

The AccessMed browser (Fig. 1) supports looking up terms by lexical matching (partial words and words that look alike), synonyms, hierarchical links (more general and specific terms), and semantic links (related terms). For example, a user who is looking for a test that measures glucose in the serum might type in the word "glucose," which would match the term "glucose" in the MED. The user could then follow the semantic link "measured-by" (which is the inverse of the "substance-measured" link) to find a number of tests that measure glucose. Based on the "specimen" links, the user might choose "serum glucose measurement" as one of several relevant terms.

Once a user has found one relevant term, the semantic network may guide the user naturally to related terms that are also relevant. For example, after finding "serum glucose measurement," the user could jump to its parent, "serum glucose tests," and then jump down to siblings of the original term, such as "serum glucose measurement 2," which is also relevant. If all the descendants of a given term are relevant, the user can simply select that term to signify both itself and all its descendants in the hierarchy. In this example, all the descendants of "serum glucose tests" are relevant, so the user might select this term and no longer


X: CPMC MED Viewer
CPMC File
Draw
View
Options
Help
2100 Serum Chemistry Test [140]
2308 Intravascular Glucose Test [4]
32703 Serum Glucose Tests [3]
1333 Serum Pyruvic Acid Measurement
1718 Serum Cryptococcal Antigen Test
1719 Respiratory Syncytial Virus Antigen by EIA
1756 Blood Barbiturate Screen
2478 Plasma Glucose Measurement [20]
31017 CHEM-7 Glucose Measurement [5]
32172 Whole Blood Glucose Tests [6]
1600 Serum Glucose Measurement
32717 Serum Glucose Measurement
36338 CPMC Laboratory Test: Glu
Search Results:
FP Slot Info for 32703
Dual
OR
D
Graph
Clear
Graph
<-Code
<-Word
Clear

```
  3537  Other ICD9 Disturbances of Straight-Chain Amin
 32128  Serum Gamma-Glutamyltransferase Tests
 32703  Serum Glucose Tests
 32717  Serum Glucose Measurement 2
```

Search Code: Search Words: ser glu Matches: 9

Selected MEDcode - or - Manual MEDcode Code Entry: 32703 Serum Glucose Tests

Serum Glucose Tests

```
Slot  Slot                        Slotvalues (51 foun
Code  Name                   MEDcode  MEDcode Print-l
  1   UMLS-CODE                 *     C0036801
  2   NAME                      *     SERUM GLUCOSE
  3   DESCENDANT-OF             1     Medical Entity
  3   DESCENDANT-OF            92     Laboratory Pro
```

Slot Selection:

Figure 1 AccessMed browser. AccessMed is based on the institutional vocabulary, a semantic network called the Medical Entities Dictionary (MED). The user can find terms by performing lexical searches (lower left corner of the browser); traversing the MED's "is_a" hierarchy (top half of the browser); or traversing the MED's other semantic links and reviewing literal attributes (lower right corner of the browser). These terms are then inserted into a query to perform a database retrieval. In this example, the user entered "ser glu" in the "Search Words" section, then selected "32703 Serum Glucose Tests" from the pick list and pressed the "Graph" button. As a result, the "is_a" hierarchy is shown centered on the MED code 32703. Its attributes are available in the lower right corner.

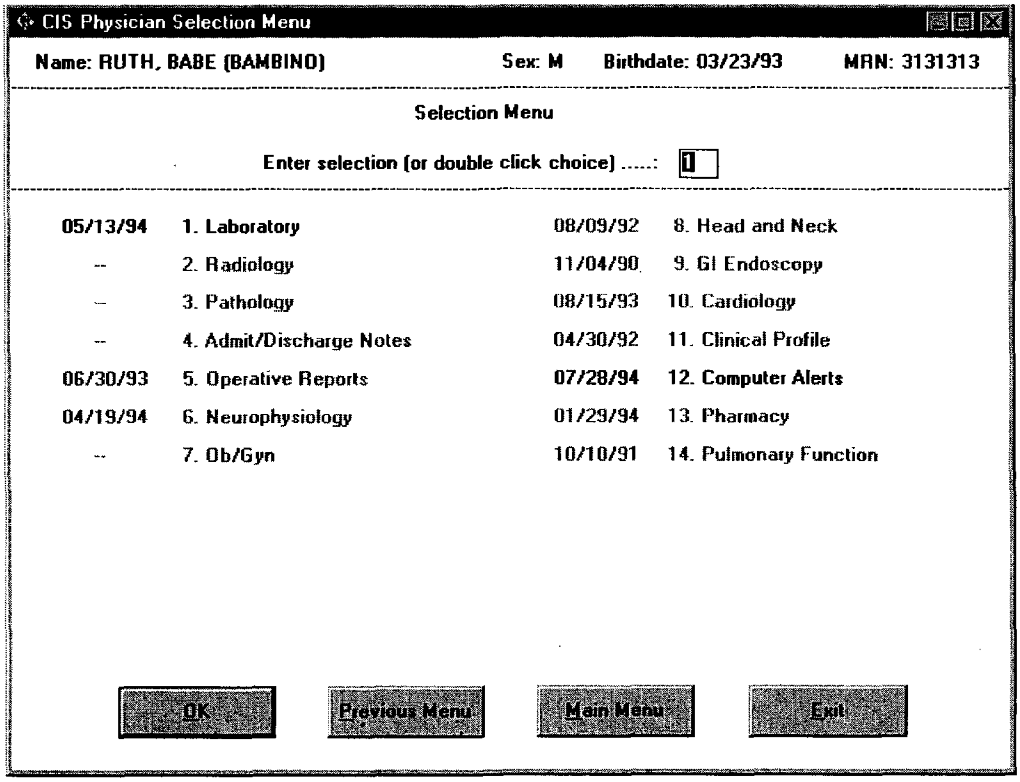
need to specify "serum glucose measurement" explicitly. Another way a user can find related terms is to traverse the semantic links. After finding "serum glucose measurement," the user could jump to its "substance-measured," which is "glucose," and then jump back to other tests that measure glucose via the "measured-by" inverse link.

AccessMed has been in use for two years. It supports many institutional functions, including the writing of queries for MLMs. There has been some concern that a good deal of training is necessary to use the tool effectively. The user needs to have some idea of how a semantic network works, how to traverse a hierarchy, and how the terms are organized within the network.

Query by Review

To address the needs of more naive users, we built Query by Review (QBR) to exploit the training that users have undergone to access information on the clinical information system's result-reporting function. The tool mimics the result review interface, offering users a familiar look and feel.

The QBR knowledge base mirrors the structure of the result-review system. The underlying formalism used by QBR is a simple hierarchy. There is no attempt within the formalism to identify the semantic meaning of the parent–child links. At different places in the hierarchy the links may signify "is-a" relations, "part-of" relations, and so on.



CIS Physician Selection Menu

Name: RUTH, BABE (BAMBINO) Sex: M Birthdate: 03/23/93 MRN: 3131313

Selection Menu

Enter selection (or double click choice) .....: 0

05/13/94	1. Laboratory	08/09/92	8. Head and Neck
--	2. Radiology	11/04/90	9. GI Endoscopy
--	3. Pathology	08/15/93	10. Cardiology
--	4. Admit/Discharge Notes	04/30/92	11. Clinical Profile
06/30/93	5. Operative Reports	07/28/94	12. Computer Alerts
04/19/94	6. Neurophysiology	01/29/94	13. Pharmacy
--	7. Ob/Gyn	10/10/91	14. Pulmonary Function

A

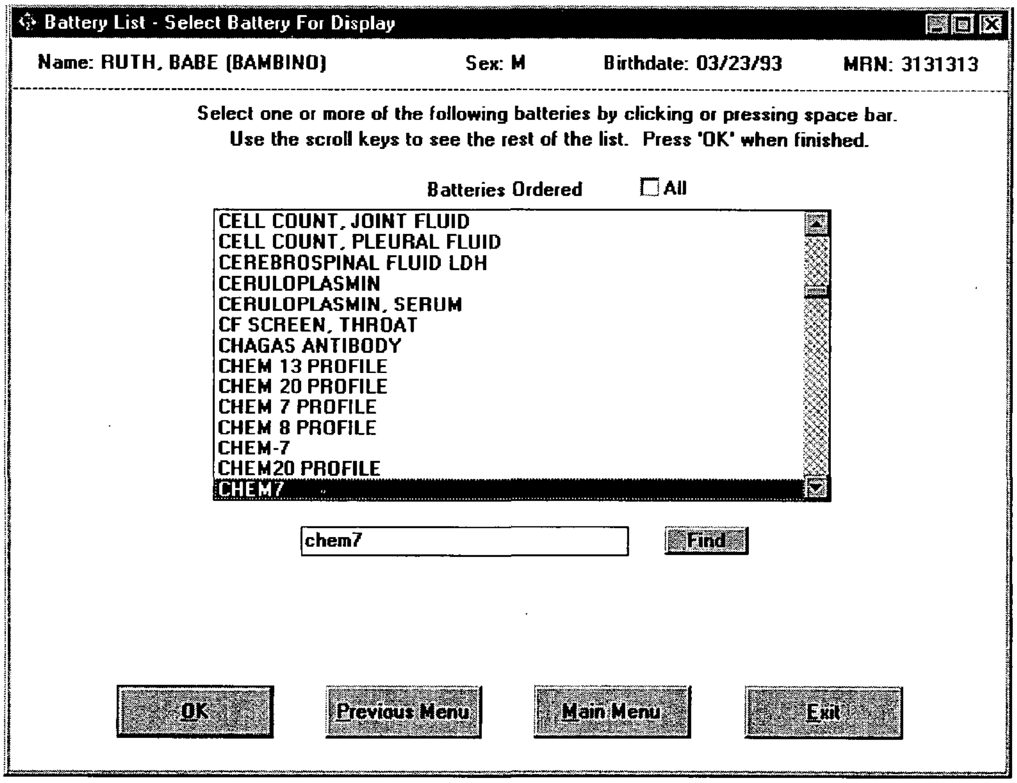
Figure 2 *Above and facing page:* Query by Review browser. The Query by Review user interface mimics the institutional result-review system in order to exploit existing training. The user follows a series of menus to select the desired terms of data retrieval. (A). In this example, the user selected laboratory tests. (B) Then the user selected the "chem7" battery. After several more menus that are not shown, the user selected a glucose test from the chem7 battery. (C) Based on the MED, Query by Review then suggests related terms that can be selected and included in the query.

The hierarchy is defined as follows. The terms that represent actual tests are at the bottom (leaves) of the hierarchy. Their parents are the laboratory batteries of which they are a part. For example, "serum glucose measurement" has as its parent the chemistry battery "Presbyterian SMAC." For those terms that are normally ordered individually rather than in a battery, a dummy battery is created in the hierarchy. At the higher levels of the hierarchy, terms are organized by the departments that produce data. For example, all terms related to data produced by the clinical laboratory are grouped together.

The QBR browser (Fig. 2) allows a user to move from the root of the hierarchy to the target test terms via a series of menus. The structure of the menus mimics the structure of the institutional result-review system: from departments to batteries to tests. The names for terms in the hierarchy are the names actually used in the result-review system, not the unfamiliar names used in the MED.

Once a user gets to the final menu, the real result-review system shows all the tests currently available for a given patient. Because real patients have only a few tests per day, it is feasible to scroll through all available tests to find other ones that are relevant. For QBR, there is no actual patient, and any number of the 1,774 tests might be relevant to the query. We could have added additional levels to QBR's hierarchy to make the 1,774 tests more manageable, but this would have required multiple classification schemes to handle different contexts (different classifications of tests are considered relevant in different contexts). We would have essentially duplicated AccessMed's function, and we would have had to train our users further.

Instead, we let the user select at least one relevant test with a simple lexical lookup. Most users can find at least one relevant test easily, based on their experience with the result-review system. The difficulty lies in finding *all* the relevant tests in a list of 1,774 entries; several different tests may measure the same substance, and all of these tests may be the target of a query. Given one test, QBR suggests other tests that may be relevant based on their having similar labo-



Battery List - Select Battery For Display

Name: RUTH, BABE (BAMBINO) Sex: M Birthdate: 03/23/93 MRN: 3131313

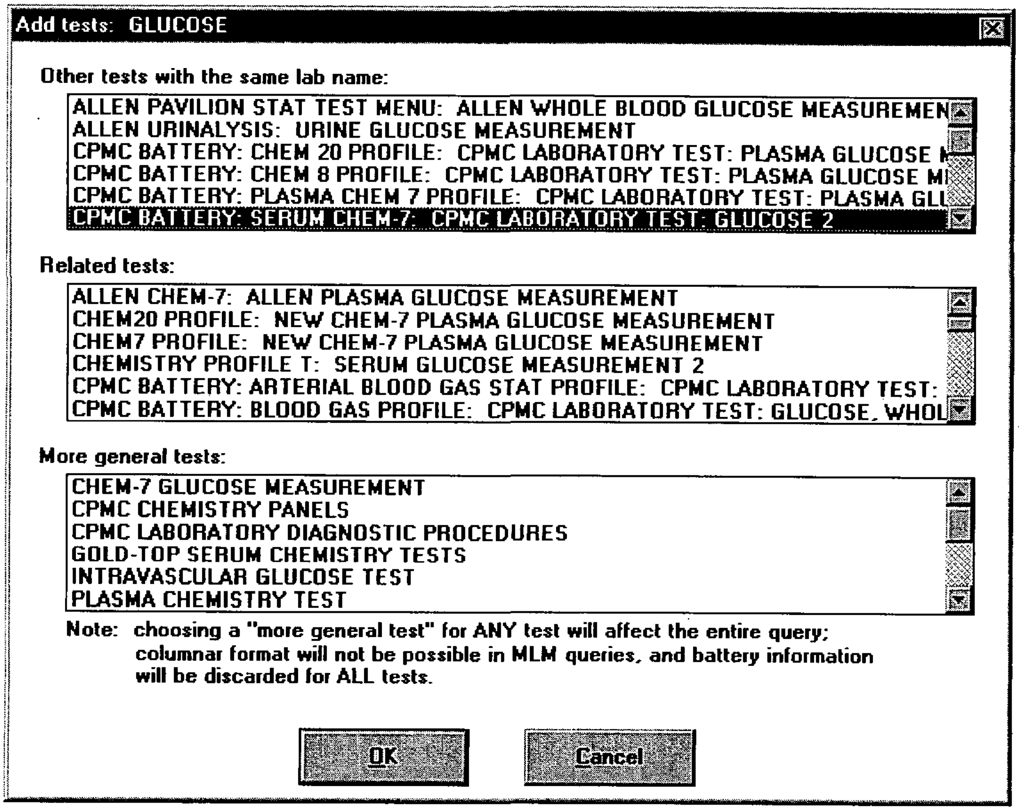
Select one or more of the following batteries by clicking or pressing space bar.
Use the scroll keys to see the rest of the list. Press 'OK' when finished.

Batteries Ordered □ All

CELL COUNT, JOINT FLUID
CELL COUNT, PLEURAL FLUID
CEREBROSPINAL FLUID LDH
CERULOPLASMIN
CERULOPLASMIN, SERUM
CF SCREEN, THROAT
CHAGAS ANTIBODY
CHEM 13 PROFILE
CHEM 20 PROFILE
CHEM 7 PROFILE
CHEM 8 PROFILE
CHEM-7
CHEM20 PROFILE
CHEM7

chem7

B



Add tests: GLUCOSE

Other tests with the same lab name:

ALLEN PAVILION STAT TEST MENU: ALLEN WHOLE BLOOD GLUCOSE MEASUREMEN
ALLEN URINALYSIS: URINE GLUCOSE MEASUREMENT
CPMC BATTERY: CHEM 20 PROFILE: CPMC LABORATORY TEST: PLASMA GLUCOSE M
CPMC BATTERY: CHEM 8 PROFILE: CPMC LABORATORY TEST: PLASMA GLUCOSE MI
CPMC BATTERY: PLASMA CHEM 7 PROFILE: CPMC LABORATORY TEST: PLASMA GLL
CPMC BATTERY: SERUM CHEM-7: CPMC LABORATORY TEST: GLUCOSE 2

Related tests:

ALLEN CHEM-7: ALLEN PLASMA GLUCOSE MEASUREMENT
CHEM20 PROFILE: NEW CHEM-7 PLASMA GLUCOSE MEASUREMENT
CHEM7 PROFILE: NEW CHEM-7 PLASMA GLUCOSE MEASUREMENT
CHEMISTRY PROFILE T: SERUM GLUCOSE MEASUREMENT 2
CPMC BATTERY: ARTERIAL BLOOD GAS STAT PROFILE: CPMC LABORATORY TEST:
CPMC BATTERY: BLOOD GAS PROFILE: CPMC LABORATORY TEST: GLUCOSE, WHOL

More general tests:

CHEM-7 GLUCOSE MEASUREMENT
CPMC CHEMISTRY PANELS
CPMC LABORATORY DIAGNOSTIC PROCEDURES
GOLD-TOP SERUM CHEMISTRY TESTS
INTRAVASCULAR GLUCOSE TEST
PLASMA CHEMISTRY TEST

Note: choosing a "more general test" for ANY test will affect the entire query; columnar format will not be possible in MLM queries, and battery information will be discarded for ALL tests.

C

Table 1 ■

Clinical Descriptions

Urine sodium concentration
Quantitative blood platelet count ("blood" includes serum, plasma, . . .)
Blood magnesium level ("blood" includes serum, plasma, . . .)
Syphilis antibodies (any specimen)
Blood ferritin level ("blood" includes serum, plasma, . . .)
Arterial PO_2
Blood gentamicin level ("blood" includes serum, plasma, . . .)
Serum bicarbonate level (serum only; **not** plasma, . . .)

ratory test names or based on their being siblings or parents of the original term in the MED. The user can then select the truly relevant terms from this filtered list. Therefore, like AccessMed, QBR does exploit the institutional vocabulary, but it guides the user in a way that minimizes the necessary training.

At present, QBR supports queries to laboratory chemistry, hematology, and serology data. It has not yet been deployed for real use.

Methods

We compared users' ability to select relevant terms using the two tools, QBR and AccessMed. We chose laboratory data (chemistry, hematology, and serology) as the medical domain because the institutional vocabulary is mature in the area, because it represents the majority of queries in our system, and because this was the first area implemented for QBR.

Eight descriptions of laboratory data were generated (Table 1), chosen from the laboratory's chemistry, hematology, and serology divisions. They were selected by one of the authors (GH) so that they represented relatively common tests and were not already coded in an MLM written by one of the subjects. Where relevant, the desired specimen was stated explicitly.

We chose as subjects all CPMC employees who were involved with writing MLMs, because this is the target audience for the tools. People who were involved with creating either tool were excluded from the study. Subjects were asked about their background and their experience with computers, the result-review system, decision-support queries (i.e., number of queries they have written for MLMs), and AccessMed (Query by Review had not been used).

Study subjects were asked to use Query by Review and AccessMed to find the terms appropriate to generate a query for each description. Subjects were asked to specify terms for the laboratory tests they were looking for. Although subjects were given the option of adding battery constraints, they were told that in most cases battery constraints were not needed.

Each subject used both tools on each description (this allowed a paired statistical test). To minimize carryover (transferring the answer obtained with the first tool to the answer with the second tool), the tasks were organized so that there was a time lapse and so that the subject did other work between two analyses of the same description. Subjects were randomly assigned to two groups. One subject group analyzed the first four descriptions using Query by Review first and then analyzed the second four descriptions using AccessMed first. The other subject group did the opposite. The time to use the tools was recorded for each subject.

Subjects' answers were compared with a standard that was assembled as follows. The pooled selections of all the study subjects plus those of one of the authors (GH, who helped design the database and the vocabulary and who had written about 100 decision-support queries) were assembled into eight queries, one for each description. The queries were run against the patient database. The retrieved data were checked for appropriateness (by the same author); terms that resulted in the retrieval of data that did not match the description were dropped from the standard.

The main performance measure was the proportion of queries that were correct (i.e., identical to the reference standard). In addition, recall and precision were calculated for each description for each subject. Recall was defined as the number of correct terms chosen by a *subject* for a description divided by the number of terms in the *standard* for that description. Precision was defined as the number of *correct* terms chosen by a subject for a description divided by the *total* number of terms chosen by a subject for a description. These calculations were done with respect to low-level terms (those that represent actual tests in the database). A term that represented a set of tests (i.e., classes) was converted to its corresponding set of low-level terms for the analysis. Those terms that could not have resulted in the retrieval of data (correct or otherwise) were not counted; that is, the actual class codes themselves were not counted in recall and precision (only their descendants were counted) because they do not appear in the database.

The subject was the unit of analysis in all statistical tests. That is, an aggregate measure was first calculated for each subject (e.g., average recall for a subject), and the overall study measure was then calculated from subjects' results, using N equal to the number of subjects; the result is a conservative confidence interval. A paired t-test was used to compare

Table 2 ■

Characteristics and Performance of Subjects (Subjects Sorted by Overall Performance)

Background	Years of using computers	Times used clinical info. system	Times written decision support queries	Times used AccessMed	QBR proportion correct	AccessMed proportion correct	Overall proportion correct
Programmer	4	1-5	1-5	21-100	0.250	0.000	0.125
Physician	5	>100	0	0	0.375	0.125	0.250
Physician	10	>100	0	1-5	0.375	0.125	0.250
Physician	15	>100	21-100	21-100	0.375	0.125	0.250
Physician	11	>100	21-100	21-100	0.375	0.250	0.313
Physician	10	>100	1-5	>100	0.500	0.250	0.375
Physician	12	>100	>100	>100	0.375	0.375	0.375

the proportion of correct queries, recall, and precision for the two tools (again using N equal to the number of subjects). We suspected that subjects' attempts to choose terms for batteries would cause confusion, so we also analyzed the data, ignoring battery constraints, to see if performance improved.

To corroborate the above analysis, pairwise intersubject agreement was also reported. This is simply the proportion of queries for which two subjects agreed. This result does not depend on the reference standard defined above.

We assessed whether previous experience with writing decision-support queries or AccessMed had an affect on subjects' performance (no one had used Query by Review before). Subjects were asked about ease of use and overall impression of both tools. A 1 (low)-to-10 (high) scale was used.

To better judge why errors occurred, we calculated the proportion of correct queries for each clinical description, and we manually reviewed the answers while looking for the most common errors.

Results

Seven CPMC employees qualified as MLM authors not involved with creating the tools. Five were veteran MLM authors, and two were new to the task. Six subjects had medical experience. The subjects had a range of 4 to 15 years with some kind of experience in using computers; this included word processor experience, and so on. Table 2 shows the number of times subjects have engaged in related activities.

Each subject took 2 to 3 hours to carry out the 16 analyses (there was no significant difference between the two tools). The standard was generated and checked against the patient database. Based on the final standard, two of the eight queries generated by the author (GH) required changes (each required one additional term), which corresponds to a 0.75 proportion of correct queries.

Subjects' individual performances are reported in Table 2. A comparison of the two tools is shown in Table 3. The proportion of correct queries was low for both tools, but it was about twice as high for Query by Review as for AccessMed, and the difference was significant. If battery constraints were ignored, the performance of both tools improved. Query by Review was still higher, but the difference was not significant. Query by Review had higher recall (not significant) but lower precision (significant) than AccessMed.

Intersubject agreement was poor, as shown in Table 4. It was slightly higher for subjects using Query by Review, but the range was wide. The highest agreement for any pair of subjects for either tool was 0.63.

Table 3 ■

Performance of QBR versus AccessMed (95% Confidence Intervals in Parentheses)

	QBR	AccessMed	Paired t-test
Proportion of correct queries	0.38 (0.23-0.52)	0.18 (0.07-0.34)	0.002
Proportion of correct queries Ignoring battery constraints	0.41 (0.27-0.57)	0.34 (0.16-0.50)	NS
Recall	0.74 (0.58-0.86)	0.68 (0.50-0.83)	NS
Precision	0.82 (0.71-0.91)	0.89 (0.81-0.95)	0.02

Table 4 ■

Intersubject Agreement

Tool	Average agreement	Minimum	Maximum
QBR	0.28	0.13	0.50
AccessMed	0.21	0.00	0.63

Table 5 ■

Effect of Experience on Performance

	≤20 times	>20 times	Paired t-test
Experience with decision support queries			
QBR (proportion correct)	0.38	0.38	NS
AccessMed (proportion correct)	0.13	0.25	NS
Experience with AccessMed			
QBR (proportion correct)	0.38	0.38	NS
AccessMed (proportion correct)	0.13	0.20	NS

As shown in Table 5, Query by Review showed no effect from experience in writing queries or in using AccessMed. AccessMed did show an improvement with experience, but the effect was not significant. Subjects' ratings of Query by Review and AccessMed are shown in Table 6. There were no significant differences between the tools.

The proportion of correct queries for each clinical description is shown in Table 7. The range was 0 to 0.79. The most common errors associated with each query are shown in the table.

Discussion

Overall Performance

The most striking result of this study is the poor performance of both tools. Even taking the confidence intervals into account, subjects got, at most, half of the queries correct; the actual fraction appears to be closer to one quarter or one third. Because it is finite, well-defined, and relatively unambiguous, the clinical laboratory should be an easy domain in which to write queries. Performance in the laboratory domain is likely to be much better than performance in a fuzzy area like the clinical history or physical examination. Therefore, these results have implications for all areas of clinical medicine.

The poor performance is not due to an inappropriate reference standard. Because the standard was verified by manually reviewing the result of database queries based on it, any errors in it are probably due to leaving terms out (resulting in false negative retrieval) rather than to including inappropriate ones (resulting in false positive retrieval). If terms have been left out, then the subjects' true performance must be lower than reported; the standard was based on part on subjects' answers, so the subjects must have omitted the same terms. Furthermore, the intersubject agreement was poor. Pairs of subjects agreed on the queries only one quarter of the time on average, and a little over half the time at the maximum. Even if the reference standard is wrong and one of the subjects actually has all the correct answers, then the other subjects still must have performed poorly.

One might conclude that the study simply shows that Query by Review is a failed experiment and that better tools would succeed. AccessMed, however, is an established tool at CPMC with many users. Its approach is similar to that of other tools reported in the literature.[8-10] The general feeling of users is that it is an accurate, helpful tool that they would not do without. We believe that the problem of retrieving data is simply more difficult than has been realized and that, even with good tools, retrieval is not very accurate.

Perhaps our metric, the proportion of correct queries, is too harsh; perhaps we should not expect queries to be completely accurate. We believe that accurate queries should be possible in a limited domain such as laboratory. When one considers that these queries are inserted into MLMs, which may run thousands of times per day, sending messages to dozens of clinicians, any small error will have an effect on many people.

Comparing Query by Review with AccessMed

The performance of Query by Review (0.38), albeit low, was significantly better than that of AccessMed (0.18). Query by Review differs from AccessMed in three ways: (1) much of Query by Review's user interface mirrors that of the institutional result-review system, whereas AccessMed's user interface is unfamiliar to new users; (2) Query by Review constrains the user to a fairly limited path, whereas AccessMed allows the user to jump around anywhere in the MED; and (3) Query by Review helps the user to select appropriate batteries after the tests are selected, whereas AccessMed offers no such help. The question is, Which of these factors is the most important?

Battery constraints limit the query so that data are returned only if the tests occurred as part of the spec-

Table 6 ■

Ease of Use and Overall Impression (Scale of 1 to 10)

	QBR (sd)	AM (sd)	Paired t-test
Ease of use	6.6 (5.9–7.3)	6.4 (5.3–7.3)	NS
Overall impression	6.4 (5.4–7.6)	6.2 (5.4–7.1)	NS

Table 7 ■

Performance and Errors by Clinical Descriptions

Clinical Description	Proportion of Queries Correct	Most Common Error	Appropriateness of MED Classes (for tests)
Urine sodium concentration	0.00	Included 24-hour total without volume (versus concentrations)	Class included concentration and 24-hour total
Quantitative blood platelet count ("blood" includes serum, plasma, . . .)	0.14	Included qualitative platelet tests (versus quantitative)	Appropriate class was available
Blood magnesium level ("blood" includes serum, plasma, . . .)	0.43	Missed one of 4 tests or one of 4 batteries	Serum magnesium class existed, which did not include a plasma test
Syphilis antibodies (any specimen)	0.07	Missed one or more of 13 tests that represent different forms of syphilis antibodies	The union of two classes covered all relevant tests
Blood ferritin level ("blood" includes serum, plasma, . . .)	0.57	Missed one of 3 tests or one of 4 batteries	Appropriate class was available
Arterial PO_2	0.14	Included venous PO_2 test; inappropriate or missing battery	Class included venous PO_2
Blood gentamicin level ("blood" includes serum, plasma, . . .)	0.79	Confused antibiotic sensitivity test (for cultures) with concentration	Appropriate class was available
Serum bicarbonate level (serum only; not plasma, . . .)	0.07	Missed "carbon dioxide" (versus "bicarbonate") or included plasma tests	Class missed one relevant test

ified batteries. Despite advice to the contrary, subjects attempted to specify appropriate batteries rather than choose no battery constraints. If we re-analyze the data, making believe the subjects used no battery constraints at all, then the performance of AccessMed improves far more than that of Query by Review, and their performance is no longer very different (but still below 42%). Therefore, this may have been the most important difference between the two tools, although statistical power is limited here. It appears that simply preventing users from trying to specify battery constraints can improve performance. There are some queries for which such constraints are necessary, but these queries are rare and could be eliminated by small modifications to the vocabulary (so that the same test is never part of two different batteries). This result points to the difficulty of offering sophisticated options to naive users: they invariably try to use them.

Experience with decision-support queries did seem to improve a subject's AccessMed performance (although not to a statistically significant extent), but this effect was probably confounded by AccessMed experience: those subjects with decision-support query experience also had AccessMed experience (sample size is too low to separate the effects). The fact that no one had experience with Query by Review and yet performance was better than that for AccessMed only strengthens the conclusion that Query by Review is good for naive users, perhaps in part because of its more constrained user interface.

Sources of Error

Subjects had a wide range of performance for different clinical descriptions (Table 7). The descriptions that caused subjects the most difficulty (less than 0.20 correct) were those for which there was some possible ambiguity over the type of data desired, in terms of dimensions (concentration versus 24-hour total), quantitative versus qualitative measures, and specimen. The subject may have understood what type was required but may not have known which terms were appropriate. For example, it is not directly apparent from the term's name whether platelet count are quantitative or qualitative. This information is represented in vocabulary's literal attributes, but subjects may not have known to look for it. For these descriptions, the major error was one of including inappropriate terms; missing terms represented a less common error.

Even those descriptions that should have been trivial—little ambiguity and few terms (e.g., blood magnesium level and blood ferritin level)—presented some difficulty. Subjects appeared to leave out some terms, although without a clear pattern. Thus, it appeared that leaving out terms caused moderate drops in performance for all descriptions, but including inappropriate terms caused major drops in performance for a few descriptions.

Some of the classes in the MED's classification hierarchy have been created to help query authors choose the correct terms by lumping related terms into single classes. Subjects used the classes to varying degrees. In almost every case, they included both the class and individual terms underneath it, indicating that they did not really understand the role of classes. In half the descriptions, an appropriate class or pair of classes was available in the MED (see the last column in Table 7). For the other half of the descriptions, the nearest (i.e., most relevant) class either missed appropriate terms or included inappropriate ones. These classes were not *incorrect*; they were merely inappropriate for the given descriptions. In some cases, subjects included an inappropriate class, resulting in an incorrect query.

One wonders whether Query by Review's result-review interface really added much, because its test-based performance (i.e., ignoring battery constraints) was similar to that of AccessMed. Apparently, result-review systems work well because a given patient undergoes only a few tests at a time. After only a few menus, it is possible to show all the patient's tests and let the user select the relevant ones. Query by Review has no real patient, so, even after several selection menus, there are still many tests to choose from (all the tests a given patient *could* have had). Because many tests may be relevant to any query, and because the tests' names may not be similar, the user is left searching for the correct set of tests. Our answer was to provide assistance similar to that provided by AccessMed (Figure 2, C), using the institutional vocabulary to steer the user to the correct related tests. Apparently, this is the critical point where many users fail. (Note also that this difference between true result review and Query by Review implies that our findings of poor performance cannot be extended to clinicians reviewing results for patient care.)

Implications and Future Directions

These results are highly relevant to other groups working on vocabularies and database access.[8-11, 16, 24-30] It is important to recognize that vocabulary representations that are well suited to medical record coding or mapping among ancillary coding schemes are not necessarily appropriate to help users find the terms they need for data retrieval. A vocabulary's ability to cover a domain is not a test of its usefulness; coverage is merely the first requirement. Browsers intended to help domain experts maintain a vocabulary are not necessarily the best tools for using the vocabulary. It is clear from our work and that of others[6,15] that further research is required in this area.

For now, we feel that the best approach to improving queries is to select the most common queries and let an expert pre-assemble them. A user may then pick the desired queries off a list. Based on a review of existing MLMs, it may be possible to cover 80% of a user's queries with a relatively small list. The rest of the queries would have to be written using a tool like AccessMed or Query by Review. The reference standard was verified by performing the queries and reviewing the results manually. Query authors should be encouraged to do this for their own queries, and perhaps tools that facilitate this process are the best current investment.

Conclusion

Producing an accurate query to a patient database is a difficult task, requiring knowledge of the data actually available, the database schema, relevant codes, the query language, and so on. Tools may help in this area by steering the user to the correct terms needed to form the query. It is important, however, not to overestimate the power of such tools, and evaluation plays a critical role in determining their true performance.

The authors thank Stephen B. Johnson, PhD, for reviewing this paper.

References ■

1. Safran C. Searching for answers on a clinical information system. Methods Inf Med. 1995;34:79-84.
2. Sengupta S, Clayton PD, Molholt P, et al. IAIMS and sharing. Int J Bio Med Comput 1994;34:339-48.
3. Stead WW, Hammond WE. Computer-based medical records: the centerpiece of TMR. MD Comput. 1988;5:48-62.
4. Hammond WE, Straube MJ, Blunden PB, Stead WW. Query: the language of databases. SCAMC Proc. 1989:419-23.
5. Morgan MM, Beaman PD, Shusman DJ, Hupp JA, Zielstorff RD, Barnett GO. Medical Query Language. SCAMC Proc. 1981:322-5.
6. Safran C, Porter D, Lightfoot J, et al. ClinQuery: a system for online searching of data in a teaching hospital. Ann Intern Med. 1989;111:751-6.
7. Cimino JJ. Use of the Unified Medical Language System in patient care at the Columbia-Presbyterian Medical Center. Methods Inf Med. 1995;34:158-64.
8. van Mulligen EM, Timmers T, van Bemmel JH. A new architecture for integration of heterogeneous software components. Methods Inf Med. 1993;32:292-301.
9. Huff SM, Rocha RA, Bray BE, Warner HR, Haug PJ. An event model of medical information representation. J Am Med Inform Assoc. 1995;2:116-34.
10. Leao B, Mantovani R, Rossi RI, Zielinsky P. Incorporating knowledge to databases—a solution to complex domains. SCAMC Proc. 1993;234-8.
11. Gouveia-Oliveira A, Salgado NC, Azevedo AP, et al. A unified approach to the design of clinical reporting systems. Methods Inf Med. 1994;33:479-87.

12. Clancey WJ. The learning process in the epistemology of medical information. Methods Inf Med. 1995;34:122–30.
13. Tuttle MS, Cole WG, Sheretz DD, Nelson SJ. Navigating to knowledge. Methods Inf Med. 1995;34:214–31.
14. Patel VL, Arocha JF. Cognitive models of clinical reasoning and conceptual representation. Methods Inf Med. 1995;34: 47–56.
15. van Mulligen EM, Timmers T, van Bemmel JH. User evaluation of an integrated medical workstation for clinical data analysis. Methods Inf Med. 1993;32:365–72.
16. Henry SB, Holzemer WL, Reilly CA, Campbell KE. Terms used by nurses to describe patient problems: can SNOMED III represent nursing concepts in the patient record? J Am Med Inform Assoc. 1994;1:61–74.
17. Johnson SB, Hripcsak G, Chen J, Clayton P. Accessing the Columbia Clinical Repository. SCAMC Proc. 1994;281–5.
18. Cimino JJ, Clayton PD, Hripcsak G, Johnson SB. Knowledge-based approaches to the maintenance of a large controlled medical terminology. J Am Med Inform Assoc. 1994; 1:35–50.
19. Hripcsak G, Clayton PD, Cimino JJ, Johnson SB, Friedman C. Medical decision support at Columbia-Presbyterian Medical Center. In: Timmers T, Blum BI, eds. Software Engineering in Medical Informatics. Amsterdam: North-Holland, 1991;471–9.
20. Hripcsak G, Ludemann P, Pryor TA, Wigertz OB, Clayton PD. Rationale for the Arden Syntax. Comput Biomed Res. 1994;27:291–324.
21. Hripcsak G, Johnson SB, Clayton PD. Desperately seeking data: knowledge base-database links. SCAMC Proc. 1994; 639–43.
22. Pryor TA, Hripcsak G. Sharing MLM's: an experiment between Columbia-Presbyterian and LDS Hospital. SCAMC Proc. 1994;399–403.
23. Barrows RC, Cimino JJ, Clayton PD. Mapping clinically useful terminology to a controlled medical vocabulary. SCAMC Proc. 1994;211–5.
24. Evans DA, Cimino JJ, Hersh WR, Huff SM, Bell DS. Toward a medical-concept representation language. J Am Med Inform Assoc. 1994;1:207–17.
25. Campbell KE, Das AK, Musen MA. A logical foundation for representation of clinical data. J Am Med Inform Assoc. 1994;1:218–32.
26. Friedman C, Cimino JJ, Johnson SB. A schema for representing medical language applied to clinical radiology. J Am Med Inform Assoc. 1994;1:233–48.
27. Bell DS, Pattison-Gordon E, Greenes RA. Experiments in concept modeling for radiographic image reports. J Am Med Inform Assoc. 1994;1:249–62.
28. O'Neil M, Payne C, Read J. Read Codes Version 3: a user led terminology. Methods Inf Med. 1995;34:187–92.
29. McCray AT, Nelson SJ. The representation of meaning in the UMLS. Methods Inf Med. 1995;34:193–201.
30. Rothwell DJ. SNOMED-based knowledge representation. Methods Inf Med. 1995;34:209–13.