Free-Text Fields Change the Meaning of Coded Data

William R. Hogan MD, Michael M. Wagner MD, PhD
Center for Biomedical Informatics
University of Pittsburgh

Researchers have advocated the supplementation of coded fields with free-text fields in electronic medical records (EMRs) to provide clinicians with flexibility during data entry. They cite advantages of more complete data capture and improved clinician acceptance and use of the EMR. However, free text may have the disadvantage of changing the meaning of coded data, which causes lower data accuracy for applications that cannot read free text. We studied the free-text entries that clinicians made during the recording of medication data. We found that these entries changed the meaning of coded data and lowered data accuracy for the medical decision-support system (MDSS) in our EMR. We conclude that supplemental free-text entries made by clinicians frequently alter the meaning of coded data.

INTRODUCTION

Capturing medical data in a coded format directly from clinicians is an important problem in the design and development of an electronic medical record (EMR) (1). One approach is to require clinicians to code data, which is problematic because no coding system currently is sufficient enough to allow clinicians to describe even a single data type (e.g., diagnoses) completely and accurately (2-5). Another approach, advocated by some researchers, is to provide clinicians with the ability to supplement coded data in EMRs with free-text entries (6-8). A key advantage of this approach is that clinicians can enter data that they otherwise would not have been able to code, or would have been required to code inaccurately. A second advantage is that clinicians may be more receptive towards an EMR that allows them to use their own style of language to describe patient care (9). A key disadvantage is that free-text entries are not machine readable. Although natural-language processing applications may eventually ameliorate this problem somewhat (10-14), there are still significant barriers to their routine implementation in EMRs. A second disadvantage of free-text fields, which to our knowledge no one has studied, is that they may change the meaning and consequently diminish the accuracy of coded data. It is our experience that clinicians record supplemental free-text data that may significantly alter or even contradict the meaning of data that they have coded.

To study the extent of this problem, we analyzed the free text that clinicians entered when recording medication data. We examined how often clinicians make supplemental free-text entries, the categories into which these entries fall, how they change the meaning of coded data, how often they provide data that coincide with the coded fields in medication records, and how often they provide data that is additional to what the coded fields can represent. We then compared the accuracy of medication data for applications that cannot interpret free text to the accuracy of data for clinician users. The purpose of the latter analysis was to measure the "cost" of supplementing coded data with free text.

METHODS

Setting

Benedum Geriatric Center (BGC) is a multidisciplinary geriatrics clinic at a university medical center serving a patient population of approximately 2,000. The BGC EMR is a locally designed and developed system running as a Windows application on 22 workstations in the clinic. The EMR reproduces the functionality of the paper record and appointment schedule with electronic forms that resemble a face sheet, an appointment status board, and visit records. The EMR captures patient data by direct clinician entry of data at computer terminals at the time of a visit or during phone contacts with patients and providers, and by data entry by licensed nurse practitioners from structured encounter forms. Regardless of who is entering data, he or she enters the first four letters of the brand or generic name into a standard electronic form, and from the resulting pick-list, selects the desired formulary item (e.g., HCTZ 25 mg tab). The user then enters the schedule of administration and comments. There are validation procedures for each field, but no field by itself is required. The only requirement when entering a medication record is that the clinician either code the formulary item or make a free-text entry in the comments field.

Inclusion criteria

This study is an analysis of data that we collected in a previous study (15). In that study, we included any patient with an appointment for either a medical nurse practitioner or geriatrician during a three-week

0195—4210/96/$5.00 © 1996 AMIA, Inc.

Table 1. Medication-Record Schema in BGC EMR

	Field name	Type	Description
1	medication_id	Number (Long)	Primary key
2	pid	Number (Long)	Foreign key for the patient record
3	med_ancestor	Number (Long)	Medication_id of the record from which this record was created due to a dose change or correction
4	created_date	Date/Time	Date record was created
5	created_who	Alphanumeric	Name of provider who created this record
6	creator_id	Number (Long)	Foreign key to provider table
7	created_reason	Alphanumeric	Whether the event was a dose change or correction
8	created_encounterID	Number (Long)	Foreign key to the encounter table
9	archived_date	Date/Time	Date record became inactive
10	archived_reason	Alphanumeric	Reason why record made inactive
11	archived_who	Alphanumeric	Name of provider who made record inactive
12	archived_encounterID	Number (Long)	Foreign key to encounter table
13	archiver_id	Number (Long)	Foreign key to provider table
14	formulary_id	Number (Long)	Foreign key to formulary
15	sig_value	Alphanumeric	Number of units of medication
16	sig_units	Alphanumeric	Units (e.g., tablet)
17	sig_route	Alphanumeric	Route of administration
18	sig_interval	Alphanumeric	Interval of administration
19	sig_prn	Alphanumeric	Whether medication is prn
20	sig_comments	Alphanumeric	Free-text comments
21	who	Alphanumeric	Name of provider
22	provider_id	Number (Long)	Foreign key to provider table
23	disp_number	Number (Int)	Quantity dispensed
24	disp_units	Alphanumeric	Units dispensed (e.g., tablets)
25	disp_refills	Number (Int)	Number of refills
26	disp_date	Date/Time	Date of last dispensing
27	disp_encounterID	Number (Long)	Foreign key to encounter table
28	disp_who	Alphanumeric	Name of provider who last dispensed
29	dispenser_id	Number (Long)	Foreign key to provider table
30	disp_p	Yes/No	Whether the medication requires a prescription
31	verified_date	Date/Time	Date the medication was last checked with the patient
32	verified_who	Alphanumeric	Name of provider who checked medication
33	verified_encounterID	Number (Long)	Foreign key to encounter table
34	verifier_id	Number (Long)	Foreign key to provider table

study period. We excluded patients with an empty medication list.

Data collection

In our EMR, clinicians use fields 14 through 19 to code medication data (Table 1). These fields include a key to the formulary item, and the dose and schedule of administration of a medication. The free-text comments field that clinicians use when entering medication data is field 20 (Table 1). The remainder of the fields are either filled in by the EMR, relate to refills, and so on. We identified all medication records from our previous study that did not have a blank comments field.

Categories of free-text entries

We calculated the proportion of all medication records with comments. We classified each free-text entry according to the data that it provided. In our analysis, we distinguished between (1) free-text entries that provide similar information to that which is provided by coded fields and (2) free-text entries that provide information for which there is no representation in the EMR. It is possible for free-text entries to provide both types of information. In this case, we counted the free-text entry twice—once for each analysis. We defined eight main categories (Table 2) based on the first classification—free-text entries that provided information corresponding to the coded fields or that modified the meaning of coded data. We designed these eight categories to be mutually exclusive, and developed them based on a preview of the medication records included in this study. We considered the free-text field as being the source of the data specified in the category

Table 2. Categories of free-text entries

	Category name	Category description	Example
1	*medID+comp sched & dose*	States medication identity and complete schedule and dose data	HCTZ 25 mg 1 tab po qd
2	*medID+part sched & dose*	States medication identity and partial schedule and dose data	HCTZ 1 tab
3	*medID+changes sched & dose*	States medication identity and data that changes meaning of coded schedule and dose	HCTZ, and taking prn
4	*medID only*	States medication identity only	HCTZ
5	*comp sched & dose*	States complete schedule and dose data for coded medication identity	1 tab po qd
6	*part sched & dose*	States partial schedule and dose data for coded medication identity	q 6 hrs
7	*changes sched & dose*	Provides data that changes meaning of schedule and dose for coded medication identity	Pt also takes 1/2 tab prn
8	*not taking*	States that patient is not taking medication	Pt. not taking

description if it was the *sole* source of such data. That is, if a free-text field merely duplicated coded data, then we did not count it as having provided the data.

We defined 5 categories based on the second classification of free-text entries—data for which there are no coded fields in the EMR (Table 3). These categories describe types of data that clinicians recorded in free text because there were no coded fields in the EMR to represent such types of data.

Table 3. Additional categories for free-text entries

	Category name	Category description	Example
1	*special*	special instructions	with meals
2	*indication*	indication for medication	edema
3	*freq of prn*	frequency of prn use	occasional
4	*duration*	duration of therapy	10 days
5	*uncertain*	uncertainty	?dose

Data accuracy

In our previous study, we measured accuracy as follows. Clinicians determined at the time of a visit whether each medication on a patient's list was an accurate representation of what the patient was actually taking. Clinicians also recorded missing medications. We defined a correct medication as one from which a clinician could determine the correct identity, dose, and schedule of a medication that the patient was actually taking. We used measures of correctness and completeness to describe data accuracy, consistent with standard methodology (1). Our measure of completeness was the mean number of missing medications per patient list and our measure of correctness was the proportion of all medication records that were a correct representation of the medication identity, dose, and schedule of administration.

To determine the data accuracy that applications which cannot read free text might experience, we adjusted the measures of accuracy from our previous study based on our analysis of the contents of free-text fields. We highlight an MDSS as an example of applications that cannot read free text because the MDSS in our EMR sends reminders on the basis of coded medication data, and we wanted to estimate its level of incorrect advice and missed reminding opportunities. We thus counted as errors the medication records that fell into our eight main categories. Since clinicians had already marked some of these records inaccurate, we determined which ones clinicians had marked as accurate to obtain the number of *additional* errors imparted by the data in free-text fields. The first four categories (where the medication identity is only available in free text) correspond to errors of omission, since an MDSS would not be able to determine the medication identity. That is, these medications would be missing from patients' lists from the viewpoint of an MDSS. We therefore used the additional errors imparted by free-text fields in these categories to adjust our measure of completeness. The second four categories (*comp sched & dose*, *part sched & dose*, *changes sched & dose*, and *not taking*) correspond to errors of commission, since in all of these cases, the medication identity is coded and therefore available to an MDSS. We used additional errors from these categories to adjust our measure of correctness.

RESULTS

Of 663 medication records studied, 234 had free-text comments (35%). One-hundred ninety of the 234 free-text fields (81%) contained data that altered the meaning of the medication record (Table 1, fields 14-19) in some fashion (Table 4). The most common data element recorded was the medication identity (*medID+comp sched & dose*, *medID+part sched &*

dose, *medID+changes sched & dose*, and *medID only)*, which was provided by 116 free-text fields (17.5%). The free-text field was the sole source of complete dose and schedule data (*medID+comp sched & dose* and *comp sched & dose*) in 74 medication records (11%), and it altered the meaning of the dose and schedule information (*medID, changes sched & dose* and *changes sched & dose*) in 34 records (5%). Five medication records had free-text fields that indicated the patient was not taking the medication (0.8%). Of these five records, clinicians marked four as being correct representations.

Table 4. Free-text comments that altered meaning of coded data

Category name	# (%) of records	# marked correct
medID+comp sched & dose	43 (6.5)	29
medID+part sched & dose	14 (2.1)	6
medID+changes sched & dose	1 (0.2)	1
medID only	58 (8.7)	41
comp sched & dose	31 (4.7)	28
part sched & dose	5 (0.8)	5
changes sched & dose	33 (5.0)	28
not taking	5 (0.8)	4
Totals	190 (28.7)	142

Of the 234 free-text fields, 54 (23%) provided data that was supplemental to the coded fields in our EMR database schema. The breakdown of categories for these additional data types reveals that clinicians most often entered an indication for a medication (Table 5).

Table 5. Additional data provided by free-text

Category name	# (%) of records	# marked correct
special	11 (1.7)	6
indication	21 (3.2)	16
freq of prn	10 (1.5)	10
duration	4 (0.6)	3
uncertain	8 (1.2)	3
Totals	54 (23.0)	38

Of the 190 medication records whose free-text fields altered their meaning, clinicians marked 142 (75%) as being correct representations of what the patient was actually taking. When counting these 142 records as 77 additional errors of omission and 65 additional errors of incorrectness (Table 4), completeness worsened from 0.37 medications missing per patient medication list to 1.02 missing per list, and our correctness decreased from 83% of medication records correct to 73% (Table 6). Thus, an MDSS would experience about three times the missing data rate and over one and one half times the incorrect data rate that a human user would experience. This decrease in data accuracy for a MDSS represents the "cost" of providing free-text comments fields in our EMR.

Table 6. Comparison of data accuracy

	Correctness[a]	Completeness[b]
Clinician	0.83 (549/663)	0.37 (43/117)
MDSS	0.73 (484/663)	1.02 (120/117)

[a] proportion of records that are correct entries
[b] number of missing medications per list

DISCUSSION

The primary conclusion of this study is that the prevalence of supplemental free-text entries in an EMR that provides such capability may be high and that most of these entries will alter the meaning of the coded data in the EMR. The implications of this study are twofold. First, the change in meaning lowers the accuracy of EMR data for automated users such as an MDSS. Second, developers of EMRs should use supplemental free-text fields cautiously.

We found that the accuracy of data for applications such as MDSSs that cannot (as yet) read free text is likely to be substantially lower than data accuracy for human users. Although supplemental free-text fields may improve data content and accuracy for human users, they also may lower data accuracy for MDSSs. Designers of EMRs should consider the tradeoff between the potential benefits and costs of providing free-text capabilities when developing their systems.

Our EMR requires that users either code the medication identity or make a free-text entry, but does not disallow both. Although eliminating a user's ability to make free-text entries for coded medication data would reduce free-text entries that change the meaning of data, it also would limit the utility of medication lists for clinicians, since they could not then record special schedules of administration, indications, and so on, for their own use.

Our EMR also permits clinicians to enter many data types in one comments field. A better scheme would be to ask users for free-text entries separately for each coded field left blank. For example, if the medication identity, but not the schedule of administration of a medication were coded, the EMR could then prompt the user to enter a free-text schedule and code the schedule field as 'other.'

Representational limitations of the EMR database schema led to problems with clinicians' ability to code special cases of the dose and schedule of administration of medications. For example, clinicians frequently recorded insulin dosing and prednisone tapering regimens in free text because no mechanism exists to code this type of data.

One potential limitation of our study is that we did not study clinicians' motivations for entering data in free text. For example, in our EMR clinicians must pick a formulary item when coding a medication. Thus, if a clinician knew that a patient were taking, for example, Amoxicillin 500mg four times a day, but did not know whether the patient had 250mg or 500mg tablets, then the clinician may have been inclined to record the medication in free text rather than guess the size of the tablets. Our scheme for classifying free-text entries did not account for this possibility, and thus our rate of free-text entries and rates of error may have been artificially high. Modifying the EMR so that clinicians can code medications without specifying exact formulary items may increase the amount of coded data in the EMR.

These observations suggest that if an EMR does have free-text capabilities, the contents of supplemental free-text fields should be monitored to identify potential data dictionary and EMR database-schema extensions. The benefits of such extensions may include the improved ability of clinicians to code data accurately and improved clinician acceptance of the EMR. Cimino et al. have enumerated the difficulties of maintaining a controlled medical vocabulary (16-18). Review of free-text entries made by clinicians could serve as the basis by which researchers set their priorities.

CONCLUSIONS

In systems that provide for the capture of data in supplemental free-text fields, the prevalence of such entries is likely to be high, and they are likely to change the meaning of coded data. MDSSs in EMRs with free-text capabilities will potentially experience lower data accuracy than human users. Monitoring of free-text entries can identify data dictionary and EMR database extensions that would allow for a greater proportion of data to be captured in a coded format.

Acknowledgments

We would like to thank the clinicians at BGC for their participation in this study.

References

1. Shortliffe E, Perreault L, eds. Medical Informatics. Computer Applications in Medical Care. Reading, MA: Addison-Wesley Publishing Co., 1990.
2. Payne TH, Martin DR. How useful is the UMLS metathesaurus in developing a controlled vocabulary for an automated problem list? Proc 17th SCAMC 1993:705-9.
3. Rosenberg KM, Coultas DB. Acceptability of Unified Medical Language System terms as substitute for natural language general medicine clinic diagnoses. Proc 18th SCAMC 1994:193-7.
4. Campbell JR, Payne TH. A comparison of four schemes for codification of problem lists. Proc 18th SCAMC 1994:201-5.
5. Evans DA, Cimino JJ, Hersh WR, et al. Toward a medical-concept representation language. The Canon Group. JAMIA 1994;1(3):207-17.
6. Scherpbier HJ, Abrams RS, Roth DH, Hail JJ. A simple approach to physician entry of patient problem list. Proc 18th SCAMC 1994:206-10.
7. Safran C, Rury C, Rind DM, Taylor WC. Outpatient medical records for a teaching hospital: Beginning the physician-computer dialogue. Proc 15th SCAMC 1991:114-8.
8. Wilton R. Non-categorical problem lists in a primary-care information system. Proc 15th SCAMC 1991:823-7.
9. Safran C, Rury C, Rind D, Taylor W. A computer-based outpatient medical record for a teaching hospital. MD Computing 1991;8(5):291-9.
10. Friedman C, Alderson PO, Austin JH, et al. A general natural-language text processor for clinical radiology. JAMIA 1994;1(2):161-74.
11. Lenert LA, Tovar M. Automated linkage of free-text descriptions of patients with a practice guideline. Proc 17th SCAMC 1993:274-8.
12. Lin R, Lenert L, Middleton B, Shiffman S. A free-text processing system to capture physical findings: Canonical Phrase Identification System (CAPIS). Proc 15th SCAMC 1991:843-7.
13. Sager N, Lyman M, Bucknall C, et al. Natural language processing and the representation of clinical data. JAMIA 1994;1(2):142-60.
14. Hripcsak G, Friedman C, Alderson PO, et al. Unlocking clinical data from narrative reports: A study of natural language processing. Ann Int Med 1995;122(9):681-8.
15. Wagner MM, Hogan WR. The accuracy of medication data in an outpatient electronic medical record. JAMIA 1996;3(3):234-44.
16. Cimino J, Hripsack G, Johnson S, Clayton P. Designing an introspective, multipurpose, controlled medical vocabulary. Proc 13th SCAMC 1989:513-8.
17. Cimino JJ, Clayton PD. Coping with changing controlled vocabularies. Proc 18th SCAMC 1994:135-9.
18. Cimino JJ, Clayton PD, Hripcsak G, Johnson SB. Knowledge-based approaches to the maintenance of a large controlled medical terminology. JAMIA 1994;1(1):35-50.