

A Study of Communication in the Cardiac Surgery Intensive Care Unit and its Implications for Automated Briefing

Kathleen McKeown, Ph.D.*, Desmond Jordan, M.D.†, Steven Feiner, Ph.D.*,
James Shaw*, Elizabeth Chen†, Shabina Ahmad, M.D.†,
Andre Kushniruk, Ph.D.‡, Vimla Patel, Ph.D.‡

*Department of Computer Science
Columbia University
New York, NY 10027, USA

†Departments of Anesthesiology
and Medical Informatics
Columbia University
College of Physicians and Surgeons
New York, NY 10032, USA

‡Cognitive Studies in Medicine
Centre for Medical Education
McGill University
Montreal, CANADA

We present a study of the information transferred among caregivers in the context of cardiac surgery and use the study to evaluate a system, MAGIC, that we are developing for automated generation of briefings. Our framework integrates cognitive and quantitative evaluation methods and features three standards that reflect current practice in the Cardiothoracic Intensive Care Unit (CTICU). Using experimental design to compare human-generated and machine-generated briefings, we show that MAGIC's current level of performance is useful. Moreover, MAGIC could help improve information flow in the CTICU by providing a consistent set of information earlier than in current practice. The separate standards are also consistent in suggesting specific modifications that may be necessary for iterative design and further system development.

INTRODUCTION

When responsibility for a patient is handed off from one caregiver or team to another, specific information must be communicated to ensure continued quality of medical care. Exactly what information should be transferred and the level of detail at which it should be provided is an important question. In this paper, we address this problem in the context of cardiac surgery. We present a study of information transferred among caregivers from the point at which a patient's surgery has been completed in the OR (Operating Room) through the patient's admission to the CTICU (Cardiothoracic Intensive Care Unit). We are using this study to inform further development of MAGIC (Multimedia Abstract Generation of Intensive Care data) [1], an experimental system that we have been developing to produce briefings automatically of patient status after CABG (coronary artery bypass graft).

When a cardiac patient reaches the CTICU, a variety of information about the patient's condition and status must be summarized for the CTICU medical team, including existing medications, ventilation parameters, laboratory results, demographics, and past medical history. This summary is usually given orally by a physician, the anesthesia resident, to another physician and nurse in the CTICU. Because the participating caregivers are extremely busy, the information is provided only once and only after the patient

arrives. The report's quality varies with the experience and preferences of the reporting physician and the questions asked by the recipient. Personnel currently provide some critical information about the patient by telephone during the operation, but this information is cursory and they are rushed. We believe that communicating an interim postoperative status report to the CTICU personnel prior to the patient's arrival could have a beneficial effect on patient care by giving CTICU personnel more lead time to address otherwise unanticipated patient needs, improving the quality of information transmission, and minimizing drug administration errors. Our goal in developing MAGIC is to produce a full interim report automatically, eliminating the need for the interim telephone call.

While MAGIC's language generation component [2] is robust, the components for graphics generation [3], speech generation [4] and coordination among media [5] are not yet ready for production use. Since MAGIC is designed to spread its information across complementary media, we decided that evaluation of the language alone in a purely textual report would not be particularly useful for informing our design. Therefore, we developed the study described in this paper to allow us to evaluate the content produced by MAGIC, to determine both whether information selected for the briefing should be included and whether any information was missed.

Given the difficulties in subject participation and identifying good gold standards for communication in the busy CTICU environment, part of our focus is on the development of an evaluation framework that can provide us with multiple standards for comparison, reflecting best practice. Previous work on evaluation for natural language systems within clinical environments [6] also suggests the use of multiple standards, among other criteria. Most evaluations to date focus on systems that extract information for later medical applications from free text (e.g., [7, 8, 9]). In contrast, in MAGIC, our focus is on evaluating the provision of essential information to subsequent caregivers.

METHODS

Experimental Setup

We selected three standards that capture normal practice at the New York Presbyterian Hospital CTICU

against which to compare the content of MAGIC's briefing: the verbal briefing for CTICU personnel, the opinion of the attending OR physician, and the admission note.

The verbal briefing given by a resident when the patient arrives in the CTICU provides a model for MAGIC's target output. This briefing is standard practice and, in combination with the chart, is the best source of information that CTICU caregivers currently have. However, these briefings vary greatly in length and content, depending on time demands, patient care demands, questions asked by the CTICU caregivers, and resident experience. Our second standard provides a check against the residents' briefings. The attending anesthesiologist in the OR is the physician most knowledgeable about patient status, but since he is responsible for multiple cases in the operating room, he cannot give the briefing in the CTICU. The attending normally provides a handwritten supplement to the electronic record and this is placed in the chart that accompanies the patient to the CTICU.

To encourage subject participation, we collected the standards in as unobtrusive a manner as possible. We tape recorded for later transcription resident briefings as they were given in the normal day to day environment, thus capturing briefing content as it was naturally provided. To encourage attending physician's participation in capturing what are otherwise hard to read handwritten notes, we provided an online textual version of MAGIC's output for the patient just completed and asked them to edit the text. This edited text could be printed and serve as the postoperative note to be placed in the patient's chart. This provides a service to the attending and thus helps, rather than hinders, normal practice. Given the extreme time demands placed on these physicians, any time taken from normal practice would have made their participation impossible.

These two standards allow us to model the automated briefing after the physician briefing. To determine whether we could improve on the physician briefing, we turned to a third standard: the *Open Heart Admission Note*, a form designed by a team of physicians from cardiac surgery, cardiac anesthesia and cardiology [10] who identified important information that should be available when the patient arrives in the CTICU. The admission note is filled out by the CTICU resident when the oral briefing is given at patient admission. It is placed in the patient's chart and consulted by CTICU caregivers. The set of information represented by the admission note constitutes an established set of criteria that should serve as a yardstick for measuring briefing content. We used the admission note as a check for cases in which there were discrepancies between MAGIC and the other standards.

Modification of MAGIC for the Experiment

MAGIC had previously been tested offline on a database of historical cases. To perform this study, we moved it to the live environment, and modified its generation facilities to produce text only. At the end of the

John Doe is a 71 year-old male patient of Dr. Oz undergoing CABG. His weight is 89 kg and his height 165 cm. His infusion lines include a Swan Ganz in the right IJ, an arterial line in the right radial and an IV in each arm.

...
He had an easy intubation. Before start of bypass, he had bradycardia and received Heparin, Albumin hung, Midazolam, Fentanyl, Cisatracurium and Nitroglycerine. At start of bypass, he had alkalosis and relative anemia. ...

Figure 1: A segment from a MAGIC briefing.

<One liter of albumen; cat=colloid-total>, <six units of platelets; cat=plts-total>, <2 units of exogenous blood; cat=prbc-bypass> on pump. And two, that's it. Urine output was about <2 liters; cat=uo-total>. Uhm, <8.5 mg of Fentanyl; cat=med-total>, <10 mg of Versed; cat=med-total>.

Figure 2: A segment from an annotated resident briefing.

case, when the attending physician prints a record from the LifeLog data acquisition system (Modular Instruments Inc.) in the OR, a script is automatically run that initiates generation of the briefing. We also provided user interface facilities to allow editing of a briefing to create a postoperative note, to access past briefings, and to create automatically a checklist of generated content that could be used to compare MAGIC's briefings with our standards. The attending can edit the MAGIC briefing using a subset of MS Word editing commands. An example of MAGIC's output is shown in Figure 1.

Collection and annotation of data

We collected 24 recordings of briefings over a two month-period, including one briefing done by an attending physician and the rest by 10 different residents. We transcribed the briefings verbatim and tagged each phrase that conveyed content from a set of 223 tags developed by one of the physicians on our team to characterize information that should be conveyed, which we categorized into 39 different categories (e.g., the category cardiac-output includes the tags c.o.-postbypass and c.o.-prebypass). We measured tagging agreement between annotators on four transcripts, which yielded an average kappa agreement that was significant at 0.81. A portion of an annotated briefing is shown in Figure 2. This same tag set was used in the automatically produced checklist of MAGIC's output, to allow comparison. We collected eight edited briefings of MAGIC's output from the 24 operations studied. Six attending anesthesiologists who regularly work in the New York Presbyterian Hospital cardiac ORs participated. It is typical for cognitive studies to focus on in-depth qualitative analyses of few subjects, in order to obtain detailed data on underlying processes involved in information processing [11]. Each physician was asked to sit in front of a computer screen after the operation was completed and (1) edit the MAGIC briefing for their case so that it provided an accurate and understandable summary and (2) "think-aloud" [12], or verbalize their thoughts,

while they edited the briefing. These think-aloud protocols were audio recorded for later analysis.

Data analysis of the think-aloud protocols involved annotating the transcripts with coding categories to characterize the physicians’ reaction to the computer-generated reports for their cases. We used categories modified from our previous work in assessing the usability of a wide range of medical information systems [13], including categories for identifying problems about the computer summary, areas where a physician would change system output, as well as suggestions for improving system output. We then coded the transcriptions of physician verbalizations for the presence of the categories and summarized the results.

Analysis of content

We evaluated the content of MAGIC briefings using a two-phase comparison:

1. *MAGIC briefings vs. briefing transcripts.* This tells us information reported by the residents but not included in MAGIC, and information included in MAGIC but not reported by the residents.
2. *Omissions or extra information from the previous phase vs. the admission note.* This tells us whether errors found in the first phase are consistent with the admission note.

We measured how well MAGIC did in comparison with the transcripts by computing *recall* and *precision*[14]. Recall measures the amount of information in the standard that MAGIC also included (i.e., how much of the information communicated by the residents was also included by MAGIC), while precision measures how much of the information included by MAGIC was also said by the residents. We used the following formulae:

$$\text{recall} = \frac{|M \cap T|}{|T|}, \quad \text{precision} = \frac{|M \cap T|}{|M|}$$

where T is the set of categories in the transcript, M is the set of categories in MAGIC briefings, and $M \cap T$ is the set of categories that occurred in both.

We then filtered these results using the admission note. Where MAGIC and the residents disagreed, we used the admission note as the ultimate gold standard. If the residents (T) included information that MAGIC did not (\bar{M}), and that information did not appear in the admission note (\bar{A}), it was discounted. The set of such cases is $T \cap \bar{M} \cap \bar{A}$, where A is the set of categories in the admission note. If MAGIC (M) included information that the residents did not (\bar{T}), and that information did appear in the admission note (A), the error was also discounted. The set of such cases is $M \cap A \cap \bar{T}$.

$$\text{recall} = \frac{|M \cap T| + |M \cap A \cap \bar{T}|}{|T| + |M \cap A \cap \bar{T}| - |T \cap \bar{M} \cap \bar{A}|}$$

$$\text{precision} = \frac{|M \cap T| + |M \cap A \cap \bar{T}|}{|M|}$$

	Recall	Precision
MAGIC vs. filtered transcript	78%	100%
MAGIC vs. transcript	60%	45%

Table 1: Comparison of MAGIC vs. transcript

RESULTS

In this section, we first report on results of the two-phase comparison, and then present the results of analyzing the critiques. Finally, we look at the agreement between these two analyses, which suggests changes that we should make in MAGIC.

Comparison

Table 1 shows recall and precision for MAGIC in comparison with the residents and after filtering by the admission note. These results show that MAGIC’s briefing includes 78% (misses 22%) of the information that *should* be conveyed for a specific patient, based on the two-phase comparison. Furthermore, the results show that *all* the information in MAGIC’s briefing is indicated as necessary by either the resident or the admission note. In contrast, the resident misses 55% of the information included in MAGIC’s briefing, all of which the admission note indicates is important.

It is not surprising that MAGIC includes no information that is not present on the admission note. In designing MAGIC, we relied on the expertise of experienced attending physicians in determining what information to convey [1]. While we did not consult the note at that point, the attending physicians we consulted were aware of the guidelines it suggests.

Looking at the discrepancy between the residents and MAGIC, a major cause for MAGIC’s omission of information is lack of online availability. We determined that 34% of the missing information (out of 22% of what residents present, Table 1) is not available online. This includes charting errors for information that was not recorded for a specific patient, information that is not represented online for any patient, information that is difficult to extract and negative results. For example, for some patients, the laboratory results of glucose (missing in 2 out of 24¹), hematocrit (19), or potassium taken after bypass (5) were not available in the database, while for other patients they were present. However, information such as creatinine (2) is not stored online. Furthermore, some information (e.g., medical history (21), allergies (20) or social history (8)) is only present in the free-form text notes made by the attending at the start of the case. To include such information in MAGIC’s briefing, the system must be able to “understand” the natural language text notes, extracting the needed information, a hard research problem that we are currently working on. Finally, negative results (e.g., “no problems during induction”) are not represented in the database and it could

¹Included in parentheses following each attribute is the number of occurrences out of a total of 24 in which the information was present in resident briefings but missing in MAGIC’s.

be difficult for the system to infer which negatives are important to mention.

A second cause for information omissions occurs because MAGIC reports information at a finer level of granularity than the residents. The residents usually mentioned totals across the operation of fluids (e.g., crystalloids (17), colloids (14), urine output (17)), exogenous red blood cells (13), cell savers (6) and medications (9). In contrast, MAGIC gave incremental values as they were given at different time points during the operation. This discrepancy caused errors in both recall (missing totals) and precision (incremental values) in comparison with the transcripts.

An examination of information included by MAGIC (and the admission forms) but omitted by the residents shows three main categories of problems: demographics, laboratories, and identified problems. Demographic information (e.g., MRN (24²), surgeon (24), weight (19), height (19), name (18), date of birth (11)) is often omitted by the resident; this may be because information appears on a written form available to people in the room and thus, does not need to be repeated verbally. MAGIC includes an inference component that can infer abnormal intraoperative events from information such as labs (19) or hemodynamics (8). It identifies problems such as hypertension, hypotension, bradycardia, tachycardia, or any of a number of problems indicated by laboratories (e.g., anemia, hypercalcemia, hypocalcemia). Residents occasionally mention such problems in their briefing but rarely at the same level of granularity as MAGIC. For example, MAGIC may indicate that a patient has hypertension before and after start of bypass, while residents tend to summarize such events without repetition, referring to overall changes and indicating longer term trends. Furthermore, in cases where MAGIC indicated problems, the residents often specifically stated that there were no problems (e.g., saying the patient was “stable”, “no problem with blood pressure”, “pressure stable”). Other specific pieces of information that MAGIC includes which the residents often do not, include cross clamp time (15), information about placement of lines (17), and medications given on induction (15).

Attending Critiques

Using methods from cognitive science, we identified potential sources of problems in the generated briefings as well as opportunities for improving the content of MAGIC’s briefings. The objective is to use the results of the analysis to improve the design of MAGIC.

Our analysis revealed problems with MAGIC’s inferencing, apparent in three coding categories: wording, relevancy of information, and redundancy of information. While the wording problems category (18 occurrences, $\bar{x}=2.25$, $s=2.33$)³ included comments that

²These numbers refer to occurrences included in MAGIC but missing in the transcripts.

³ \bar{x} = average number of occurrences across the eight sessions; s = standard deviation.

referred to the actual phrasing, it also included references to statements the physician would not have included (i.e., extra information). These were often statements generated by MAGIC’s inference engine (e.g., in response to the statement, “the patient had bradycardia”, in one computer-generated report, a physician indicated “I would not have said that”). Comments such as these indicate MAGIC uses a more sensitive measurement to detect incidents than do physicians. This surfaced under the category of relevance also, where another physician stated: “In terms of these things with the bradycardia that’s not useful. If this patient was very prone to developing very significant bradycardia, that would be something I’d want to know post-operatively.” In a number of protocols, subjects indicated that the fine granularity of the inferencing data (i.e., heart rate and blood pressure) was unnatural and too detailed. They found repeated references to problems were problematic and these were captured under the category of redundancy.

The category accuracy (16 occurrences, $\bar{x}=2$, $s=1.32$) included comments about discrepancies between MAGIC’s briefing and other sources of information (e.g. differences in the age of a patient mentioned in the briefing and in MAGIC’s summary). In some cases, the physicians stated that dosage of drugs specified in MAGIC’s briefing was incorrect, but physicians may have interpreted as totals doses provided on an ongoing basis by MAGIC.

The category relevancy of information (12 occurrences, $\bar{x}=1.5$, $s=3.28$) included indications by subjects that information they considered very relevant to the case was missing from MAGIC’s briefings. For example, a number of subjects indicated that they wanted to know the total amounts of anesthetics given during the operation. Redundancy of information (10 occurrences, $\bar{x}=1.25$, $s=0.97$) also related to medications. Subjects did not understand that MAGIC provides medications incrementally with each inference, as one subject incorrectly inferred that the same medications were repeated: “OK, I see what its doing, every time we hang albumen, or every time we change the medication, it writes it down. So it keeps giving the same medications over and over again.”

In addition to reporting problems, physicians also made suggestions during the recorded sessions. The majority of suggestions dealt with information that physicians wanted in the computer-generated summary (57 occurrences, $\bar{x}=7.13$, $s=3.48$). In particular, specific aspects of the patient’s medical history were indicated as being relevant and useful to have in the computer reports (e.g. information about the patient’s allergies). The other categories of information most requested were those corresponding to items on the CTICU admission note which were not included in the MAGIC briefings (e.g. total urine output, last hematocrit value). Other suggestions dealt with rephrasing of specific statements generated from MAGIC inferences (4 occurrences).

Discussion

Our two-phase analysis of content and the attendings’ critiques of MAGIC consistently supports two prob-

lems in MAGIC's output: detailed inferencing and incremental reporting of drugs, blood products and fluids. Both analyses found that the physicians preferred to have totals rather than individual doses as they are given. For inferencing, the content analysis showed that residents do not often report details of inferences, such as problems that occur repeatedly at each critical time point during the operation. Such problems are typically only reported once by the residents. It also showed that residents sometimes explicitly state there were no problems when MAGIC identifies problems. The attendings also tended to remove references to problems that they did not feel were significant. They also removed repeated references to inferences and replaced them with summary phrases. These findings suggest that we should further evaluate the thresholds used in MAGIC for detecting inferences, looking at the effect on patient outcome, and develop techniques to further generalize repeated references to events (e.g., generating "blood pressure was labile" in place of referring to alternating incidents of hypertension and hypotension).

CONCLUSIONS

We provide an integration of cognitive and quantitative evaluation methods in a framework for analysis of communication and automated briefing in the CTICU environment. Our analysis of MAGIC demonstrates that it functions as a useful tool in its current state. Its briefing includes the majority of information (78%) that residents currently provide in theirs and provides information beyond the resident which the admission note indicates is relevant. Given that MAGIC is intended to *supplement*, as opposed to replace, the current resident briefing in the CTICU, providing information before the patient arrives, we believe MAGIC's current level of performance should help in improving information flow in the CTICU by providing a consistent set of additional information early on.

Finding a good standard for comparison is challenging. Human-human communication is variable and conditions in the CTICU make it difficult to find multiple expert standards. Our studies raise the possibility that in some cases the computer may be able to do better than existing communication. For example, a computer can identify problems at finer granularity during an operation and remember more details afterwards than people can. If it could be shown that abnormal events identified at fine granularity affect patient outcome, that would argue for retaining information in MAGIC that residents currently omit. When the literature does not provide information on adverse outcomes, personal preference plays a large role in the content of the briefings. This suggests that further study may be needed to demonstrate possibilities for improvement over current information flow through evaluation of patient outcomes; that is our next step.

Acknowledgments

This research is supported in part by NLM Contract R01 LM06593-01 and the Columbia University Center for Advanced Technology in Information Manage-

ment (funded by the New York State Science and Technology Foundation). The authors would like to thank Hillary Schmidt for her extensive input and discussion on experimental design.

References

1. M. Dalal, S. Feiner, K. McKeown, D. Jordan, B. Allen, and Y. alSafadi. MAGIC: An experimental system for generating multimedia briefings about post-bypass patient status. In *Proc. 1996 AMIA Fall Symp*, pages 684–688, Washington, DC, October 26–30 1996.
2. K. McKeown, S. Pan, J. Shaw, D. Jordan, and B. Allen. Language generation for multimedia healthcare briefings. In *Proc. Applied NLP*, pages 277–282, 1997.
3. M. Zhou and S. Feiner. Automated production of visualizations: From heterogeneous information to coherent visual discourse. *J of Intell Info Sys*, 11(3):205–234, December 1998.
4. K. McKeown and S. Pan. Prosody modeling in concept-to-speech generation: methodological issues. In K. Sparck Jones, G. Gazdar, and R. Needham, editors, *Computers, Language and Speech: Formal Theories and Statistical Data*. The Royal Society, London, UK, 2000.
5. M. Dalal, S. Feiner, K. McKeown, et al. Negotiation for automated generation of temporal multimedia presentations. In *Proc. ACM Multimedia*, 1996.
6. C. Friedman and G. Hripcsak. Evaluating natural language processors in the clinical domain. *Methods of Information in Med*, 37:334–344, 1998.
7. N. Sager, C. Friedman, and M. Lyman. *Medical language processing: Computer management of narrative data*. Addison-Wesley, Reading, MA, 1987.
8. P. Haug, D. Ranum, and P. Frederick. Computerized extraction of coded findings from free-text radiologic report. *Radiology*, 374:543–8, 1990.
9. G. Hripcsak, C. Friedman, P. Alderson, S. Johnson W. DuMouchel, and P. Clayton. Unlocking clinical data from narrative reports. *Ann of Int Med*, 122(9):681–8, 1995.
10. S. Lenz, S. Myers, S. Nordlund, D. Sullivan, and V. Vassista. Benchmarking: finding ways to improve. *Jt Comm J Qual Improv*, (20)5:250–9, 1994.
11. V. Patel, F. Arocha, and D. Kaufman. Diagnostic reasoning and medical expertise. In D. Medin, editor, *Psychology of Learning and Motivation*, volume 31, pages 187–252. Academic Press, New York, 1994.
12. K. Ericsson and H. Simon. *Protocol analysis: Verbal reports as data*. MIT Press, Cambridge, MA, 1993.
13. A. Kushniruk, V. Patel, and J. Cimino. Usability testing in medical informatics: Cognitive approaches to evaluation of information systems and user interfaces. *Proc. 1997 Fall AMIA*, 1997.
14. G. Salton and M. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, New York, 1983.