Research Paper ■

Generation and Evaluation of Intraoperative Inferences for Automated Health Care Briefings on Patient Status After Bypass Surgery

DESMOND A. JORDAN, MD, KATHLEEN R. MCKEOWN, PHD, KRISTIAN J. CONCEPCION, MS, STEVEN K. FEINER, PHD, VASILEIOS HATZIVASSILOGLOU, PHD

Abstract **Objective:** The authors present a system that scans electronic records from cardiac surgery and uses inference rules to identify and classify abnormal events (e.g., hypertension) that may occur during critical surgical points (e.g., start of bypass). This vital information is used as the content of automatically generated briefings designed by MAGIC, a multimedia system that they are developing to brief intensive care unit clinicians on patient status after cardiac surgery. By recognizing patterns in the patient record, inferences concisely summarize detailed patient data.

Design: The authors present the development of inference rules that identify important information about patient status and describe their implementation and an experiment they carried out to validate their correctness. The data for a set of 24 patients were analyzed independently by the system and by 46 physicians.

Measurements: The authors measured accuracy, specificity, and sensitivity by comparing system inferences against physician judgments, in cases where all three physicians agreed and against the majority opinion in all cases.

Results: For laboratory inferences, evaluation shows that the system has an average accuracy of 98 percent (full agreement) and 96 percent (majority model). An analysis of interrater agreement, however, showed that physicians do not agree on abnormal hemodynamic events and could not serve as a gold standard for evaluating hemodynamic events. Analysis of discrepancies reveals possibilities for system improvement and causes of physician disagreement.

Conclusions: This evaluation shows that the laboratory inferences of the system have high accuracy. The lack of agreement among physicians highlights the need for an objective quality-assurance tool for hemodynamic inferences. The system provides such a tool by implementing inferencing procedures established in the literature.

■ **J Am Med Inform Assoc.** 2001;8:267–280.

Affiliation of the authors: Columbia University, New York, New York.

This research was supported in part by contract R01 LM06593-01 from the National Library of Medicine and by the Columbia University Center for Advanced Technology in High Performance Computing and Communications in Healthcare (funded by the New York State Science and Technology Foundation).

Correspondence: Desmond A. Jordan, MD, Department of Anesthesiology, 622 West 168th Street, PH-5, New York Presbyterian Hospital, New York, NY 10032; e-mail: <daj@columbia.edu>. Reprints: Kathleen R. McKeown, 450 Computer Science Building, Department of Computer Science, 1214 Amsterdam Avenue, Columbia University, New York, NY 10027; e-mail: <kathy@cs.columbia.edu>.

Received for publication: 11/15/99; accepted for publication: 1/16/01.

When a caregiver needs to act quickly because of a patient's clinical status, a succinct overview highlighting important events (e.g., that the patient was hypertensive) can communicate information more efficiently than an exhaustive log of every vital sign, procedure, and laboratory result over a length of time. For example, a single sentence that mentions an inferred episode of hypertension occurring during a bypass operation could effectively summarize what would otherwise be an overwhelming number of low-level raw blood pressure readings gathered during the operation (1,080 readings for the average five-hour bypass operation).[1,2] Both brevity and

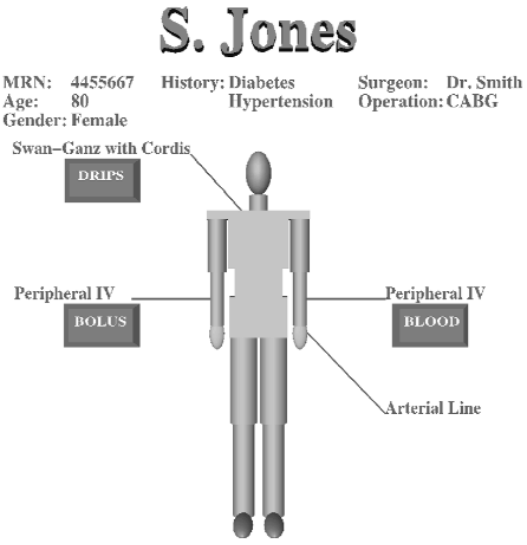

Figure 1 MAGIC-generated response including graphics and speech.

identification of important events can be achieved by recognizing relevant patterns in the patient's record that represent illness severity. However, if these inferences are wrong, the summary can be useless or even harmful. The research we describe here addresses the development and evaluation of inference rules that identify important information about the patient's status to include in a summary.

We are carrying out our work on inferences in the context of MAGIC (multimedia abstract generation of intensive care data), an experimental system that generates multimedia presentations automatically to explain a cardiac patient's postoperative status to caregivers.[1,2] These presentations are based on detailed medical data obtained during the patient's cardiac surgery and are intended to inform personnel in the cardiac intensive care unit (ICU) of the patient's status prior to their arrival from the operating room (OR). The transfer of an anesthetized and ventilated patient is a critical event that entails a high degree of risk. Notifying ICU personnel in advance of the patient status and impending transfer minimizes delays in therapy.[3]

Patient status at critical points[4] represents information that is necessary for continued care of the patient in the ICU and for improvement in individual patients' outcomes, as shown by mortality studies[5] and the resultant severity of illness measurements developed by the New York State Heart Association.[6] Without such information, care may be inappropriate or delayed.[3] Because patients are aggressively treated in the OR, those with derangements in physiologic values should be considered nonresponders to therapy, or "treatment failures," heightening the need to convey problems to subsequent care providers.[7]

A key component of MAGIC is its set of medical inference rules. These rules identify patterns in the patient's record, from which they infer information that can be used to describe the patient's status more concisely, as shown in Figure 1.[1,2] In this paper, we present the inference rules that we designed and describe an experiment to validate their correctness. Our inference rules operate in real time, identifying abnormal events from numeric data in current cases, but they can also be used on historical cases. Our experiment compares system performance on a set of historical cases with performance of a group of residents and attending physicians on the same cases. The physicians were provided with the same patient data given to the system and asked to identify whether the abnormal conditions that the system tracks occurred.

Background

While many researchers study the integration of individual abnormalities to judge the overall severity of patients' conditions,[8] our focus is on communication of abnormalities and severity to clinicians. When a cardiac patient arrives in the ICU after surgery, a variety of information about the patient's condition and status must be summarized for the ICU medical team. This summary is usually given orally by a physician from the OR, the anesthesia resident, to another physician and nurse in the ICU. Some critical information about the patient is provided by telephone during the operation, but this information is cursory and OR physicians are rushed. In current practice, the information that has been conveyed is not easily accessible to a clinician who is responsible for the patient's care but was not present at the briefing. Therefore, as nursing shifts change and new medical staff are added, these clinicians must review the anesthesia chart along with other material in the patient record.

Our goal in developing MAGIC is to replace the telephone call from the OR with an automatically generated briefing that provides the full information given in the ICU briefing. This will supplement, rather than replace, the ICU briefing, providing information earlier so that ICU clinicians have time to prepare.[3] MAGIC can also be used to replay the briefing for clinicians who missed it. MAGIC offers the potential to

provide a consistent, standard set of information for each patient, offsetting the possibility that a resident may forget to report to the ICU staff critical incidents that require postoperative follow-up.

MAGIC is unique in its ability to determine automatically the content and form of a briefing on patient status, including the sentence structure and wording of the language,[9] the graphic representations,[10] the intonation of the speech,[11] and the coordination of the different media in a single briefing.[1,12] Figure 1 shows a portion of MAGIC's multimedia output. In this context, inferencing plays a critical role in determining what is important to communicate. Given that the resulting briefing must be concise, MAGIC must select from a large quantity of information on the patient only information that is critical to the patient's ongoing care. This places demands on the inferencing process above and beyond those placed on traditional expert systems, requiring reasoning over time, limiting information, and linking it to events that can be communicated (i.e., critical time points). In the past, medical inferencing has been used primarily to suggest diagnoses, recommend practice, and provide decision support rather than to select information to communicate.[13,14]

Although communicating the existence and degree of abnormalities in physiologic parameters is important at any time during cardiac surgical procedures, there are several critical points at which transmitting the status of this information to subsequent caregivers is vital.[5,15] The four critical points investigated here are induction, skin incision, start of bypass, and end of bypass. *Induction* (or intubation) is when the patient is anesthetized and mechanically ventilated. *Skin incision* initiates the onset of surgical stimulation and stress. During cardiopulmonary bypass, the heart is stopped and blood pressure is controlled by a mechanical pump. *Start of bypass* refers to the point when this begins, and *end of bypass* is the time when the patient is taken off the pump. These milestones constitute reference points[4] at which surgical processes commonly induce lability in heart rate, blood pressure, and laboratory test results.

Other researchers have developed systems to scan electronic surgical records and report incidents.[15] However, their focus was on outcome instead of communication. They did not link incidents to critical points, and they used higher thresholds for abnormality than we do. In our experience, this approach misses events that, although more difficult to detect, should be communicated when they occur at a critical point. For example, hypocalcemia following bypass is far more important than hypocalcemia at any other point in the case.[16] An ionized calcium concentration below 0.7 mcg/dL is considered hypocalcemic.

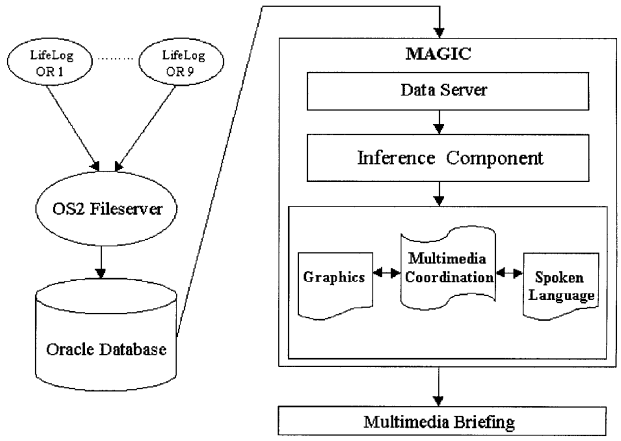

Figure 2 MAGIC system architecture.

To determine when an event is abnormal, we use severity-of-patient-illness scores developed previously, such as the Acute Physiologic and Chronic Health Status Evaluation (APACHE III),[5,8] Multi-organ System Illness Score (MSIS),[7] and the therapeutic Intervention Severity Score (TISS).[17] These scores have been used individually or in combination for many years for the prognostic scoring of surgical and critically ill patients as well as for stratification of patient illness severity.[18] In this paper, the existence and extent of physiologic derangement during bypass surgery was assessed using APACHE III, MSIS, and intraoperative thresholds[19] representing abnormality for blood pressure, heart rate, arterial blood oxygen (both PO_2 and saturation), pH, potassium, glucose, hematocrit, and ionized calcium (see Table 1). Our use of these multiple scores is similar to the use of composite outcome scales, such as the American Society of Anesthesiology (ASA) Physical Status[16] and the MSIS,[7] which allow quantification of complex clinical phenomena that cannot be adequately described by a single clinical or biochemical measure.

System Description

The system architecture for MAGIC, shown in Figure 2, shows that during the course of a cardiac operation, information is automatically collected using the LifeLog data acquisition system (Modular Instruments Inc.), which is part of the existing information infrastructure in the cardiac OR at New York Presbyterian Hospital. It polls medical devices (Hewlett Packard Merlin monitors, Ohmeda anesthesia machines, and saturation monitors) every 50 sec-

onds, recording indicators of patient status, including vital signs, inhaled anesthetics, and ventilation parameters. Bolus drugs, postoperative drugs, laboratory results, intravenous lines, information about devices such as a pacemaker, and data from echocardiograms are manually entered by the anesthesiologist using the LifeLog interface. Surgical events such as time of intubation, skin incision, and start and stop of bypass are also entered manually.

At the end of the operation, the collected data are downloaded into an Oracle database, for easy access using a standard query language. MAGIC's inference engine, which we developed for the purpose of generating multimedia briefings, scans heart rate and blood pressure readings, for which values are available at 50-second intervals, and laboratory results, for which values for one to ten tests are available before the start of bypass and after the end of bypass. The inference engine applies a set of inference rules to determine whether any abnormal events occurred within a 20-minute window before or after any of the four critical time points, determining the time interval of the event (i.e., start and stop times) and identifying drugs that were given. Any database entry labeled as DRUG is extracted; typical drugs include pressors (e.g., phenylephrine, ephedrine) and depressors (e.g., esmolol, nitroglycerine). This yields eight possible time periods (one before and one after each critical point) during which abnormalities are detected and reported, if they occur.

For the rare cases in which the patient goes on and off bypass multiple times, we currently use the first on-bypass and the last off-bypass times, although it would be relatively easy to include all on and off bypass times if we found that this were preferable. This inferred information is then stored in MAGIC's database, along with other extracted data such as demographics, medical history, and drugs given, to be used as the content for a multimedia briefing. At the time of the inferencing experiments reported here, there were no facilities for automatically transferring data from the OR to the Oracle database at the end of each operation. Instead, information for a set of patients could be transferred periodically.

For the experiment, information on a test group of a month's worth of concurrent patients was stored in the database once, and we evaluated MAGIC on this test set. MAGIC has since been integrated into the online information infrastructure in the cardiac OR at New York Presbyterian Hospital, although it is not yet deployed. It runs in a networked environment with full access to the OR database. Information on a patient is automatically stored in the database as soon as the operation is complete.

The inference engine can find two classes of abnormal events: those relating to hemodynamics and those indicated by laboratory results, both described below. Hemodynamic inferences identify episodes of hypotension, hypertension, bradycardia, and tachycardia. Laboratory inferences identify acidosis, alkalosis, hypercardia, hypoxia, low saturation, hyponatremia, hypernatremia, hypokalemia, hyperkalemia, hypocalcemia, hypercalcemia, anemia, hypoglycemia, and hyperglycemia. The algorithms used for each class are rule-based and use thresholds based on severity of illness scores developed and extensively tested in previous work,[5,7,8,19] as described below. However, the algorithms differ because of differences in the amount of information available. For hemodynamics (i.e., arterial blood pressure and heart rate), there is a tremendous amount of data, since heart rate and systolic, diastolic, and mean blood pressures are recorded every 50 seconds. In contrast, laboratory results are reported sporadically, usually before and after bypass. As a result, hemodynamic inference rules operate over temporal intervals, using temporal abstraction to determine abnormality from a set of frequent readings, whereas laboratory inference rules test a specific point in time.

Hemodynamic Inferences

These inferences look for intervals of time when blood pressure or heart rate rises above or falls below a predetermined threshold for a length of time. We developed rules that directly encode standard thresholds for bradycardia, tachycardia, hypotension, and hyper-

Table 1 ■

Threshold Values of the Inference Engine

	Hypo/Low	Hyper/High
Blood pressure	100	150
Heart rate	50(pre)/60(post)	120
Calcium	0.8	1.5
K+	3.5	5
Carbon dioxide	N/A	45
Oxygen	60	N/A
pH	7.35	7.45
Sodium	135	145
Saturated oxygen	90	N/A
Glucose	90	200
Hematocrit	32(pre)/30(post)	N/A

tension,[20,21] shown in Table 1. Berger et al.[20] and Block[21] based these thresholds on experiments using extensive data and statistical models. For example, our rules detect hypotension when blood pressure falls below 100 for 250 seconds (five 50-second intervals).

To smooth over temporal variations in data, we use a sliding scale average,[15] looking at a window of five consecutive values of blood pressure and heart rate. If the average does not meet the threshold, MAGIC drops the oldest value and slides forward in time to add a new value. If the average meets the threshold, the start of an abnormal episode is recorded and we continue calculating sliding averages across the window until the average returns to a normal value, marking the end of the episode. Once the time period for each episode has been calculated, MAGIC then records the drugs and amounts that were given so that the briefing can describe treatment. After all abnormal episodes have been found, MAGIC links each episode with one of the four critical time points (induction/intubation, skin incision, start of bypass, end of bypass), noting whether it occurred within a window of 20 minutes before or after that point. Since the anesthesiologist manually enters the critical time points during the operation, this window also helps compensate for errors in charting.

In almost all cases, we found artifacts in the data. For example, a spike may occur in the heart rate or blood pressure because of electric cautery, blood draws, catheter flushing, or other reasons. To avoid making false inferences, MAGIC automatically filters the data before beginning inferencing, to retain only data in cases both where values remain within valid ranges and where changes in one value (e.g., heart rate) are accompanied by an appropriate change in the other (e.g., blood pressure). More specifically, our algorithm for filtering artifacts is as follows:

1. Filter any values that fall within the following invalid ranges:
   A. All three blood pressures (mean, systolic, and diastolic) are equal. This usually occurs when the LifeLog controls are incorrectly set.
   B. Any systolic blood pressure greater than 250.
   C. Both blood pressure and heart rate are zero. This happens when the machine was not turned on immediately. Zeros are replaced by average heart rate and blood pressure, provided that the patient was not currently on bypass.
2. Remove cases in which one parameter's change is not accompanied by a change in another parameter. If the patient had a change in heart rate greater than 50 within a 50-second interval, MAGIC retains the spike if there is a corresponding change of 10 in blood pressure. If blood pressure did not change, then the spike is replaced with the last good heart rate value. The reverse is also true; spikes in blood pressure are retained when accompanied by changes in heart rate.

Laboratory Inferences

Thresholds for laboratory values were taken directly from APACHE scores.[8] To accurately report abnormal values, our system separately inspects data obtained before and after bypass. If laboratory tests were performed during bypass, we ignore the results because of the difference between "normal" values and "on-bypass" values. Results of laboratory tests taken during bypass are not known to be indicative of patient postoperative status and are used only to control bypass settings.[16] For each set of laboratory results, we apply the corresponding APACHE threshold (listed in Table 1) to calculate whether or not the results were abnormal.

Methods

The performance of an automated system for medical inferencing should, in principle, be evaluated against a set of correct decisions on the same input data. If such a gold standard were available, then measures such as specificity, sensitivity, and accuracy could be used to measure quantitatively system performance. It could be argued that the carefully calculated decision thresholds and associated rules that are part of methodologies such as APACHE[5,8] would provide such a standard for hemodynamic and laboratory inferences. This is problematic, though, since MAGIC itself incorporates these rules, so using the APACHE rules as a standard would give very high scores to our automated system and would not determine whether system output was useful in practice.

Instead of using APACHE, we relate the evaluation to actual physicians' decisions in the ICU. Consequently, we collected data from physicians on a set of historical patients and compared their decisions to those automatically produced by our system. One of the goals of our study was to establish whether the physicians' answers are consistent enough across different physicians to be used to evaluate the automated system, or whether the automated system should be used as a quality assurance tool in the face of significant physician uncertainty about the correct answer.

Selection of Human Judges

We obtained LifeLog data for input to MAGIC and the corresponding human-readable charts for a set of 24 concurrent adult patients, aged 36 to 78 years, at New York Presbyterian Hospital, who had undergone cardiac surgery, performed by a variety of surgical teams, during February and March 1998.

A standard questionnaire was prepared and handed out to physicians participating in the study, along with each patient record; this questionnaire is given in Appendix A. For each of the four critical time points discussed previously, the questionnaire asked the physician to determine, using a checklist, the presence or absence, within 20 minutes before or after that point, of each of the conditions that MAGIC can identify. For pre-bypass and post-bypass, a list of laboratory results was provided, each of which was to be marked by the subject if it was abnormal. Again, the physicians made a binary ("yes"/"no") decision on each potential abnormality. No definition of abnormal was provided; subjects used their own judgment. This deliberate design decision was made to avoid forcing physicians to use a definition that was not their own and to learn how abnormalities can be identified in practice. A final question asked the subject to identify any other abnormality not covered and to indicate its temporal relationship to the nearest critical time point.

Each individual patient's chart (see Figure 3) was given to three different physicians, yielding three responses per patient. Most physicians saw more than one chart. No physician reviewed their own patient's chart. A total of 46 physicians affiliated with New York Presbyterian Hospital participated in the experiment, of whom 18 were residents in anesthesiology and 28 were attending physicians in anesthesiology, ranging in age from 28 to 63 years. The residents were in their third or fourth year of training. Using attending physicians in addition to residents means that the accumulated responses may be of higher accuracy than responses in practice, where the residents may be the only ones who report on the patient's intraoperative course. Monitoring of a patient in the OR and the ICU is performed by anesthesiologists, and thus they may be more qualified than the other physicians involved in the case (e.g., the thoracic surgeon or cardiologist) to make decisions about laboratory and hemodynamic abnormalities. Anesthesiologists manually record these data in the OR and may have the highest skill level required to read the surgical record.

The experiment was conducted over a $2\frac{1}{2}$-day period, and assignment of cases to physician subjects was done on a first-come first-served basis. Each physician was allowed to spend as much time as desired on the questionnaire for a given patient (average, 20 minutes), and the patient's record was available during the entire time.

Both the physicians and the system produced judgments on 60 variables for each patient whose chart they examined. Each of these binary variables represents the presence or absence of a particular abnormal condition at a particular time (e.g., hypotension before the start of bypass). Hemodynamic inferences include four conditions and four critical time points, with time periods both before and after, giving a total of 32 variables. Laboratory inferences include 14 conditions and two time periods (before and after bypass), giving a total of 28 variables.

We collected judgments for 24 patients; however, it was impractical to have a physician produce judgements on all 1,440 combinations of patients and potential inferences (24 patients times 60 variables), because of the time required (about 20 minutes per review of each case). Instead, each of the 46 physicians processed the entire set of data for a limited number of patients (three or fewer per physician), and we ensured that each patient received three sets of judgments from three different physicians. In this way, we were able to create three composite judges, whose assessments of the patients' conditions we analyzed and compared with the system's output.

Measuring Agreement among Human Judges

In the analysis that follows, we first look at the average agreement between the three composite judges, as a means of determining the types of decisions for which the physicians can be considered correct and thus for which their responses can form a gold standard for evaluation. We measure the average rate of agreement between the three human judges in each case, which provides a measure of the validity of their responses as a gold standard.[22,23] The average rate of agreement between three binary decision sets is defined as the average of the three percentages of pairwise agreement, calculated on each of the three possible pairs of decision sets.

We also calculate the average agreement rate between the system and the judges when any one of the latter is replaced by the system. Three replacement operations are performed (in each case, one human judge is left out), the average agreement between the remaining two physicians and the system is calculated as described earlier, and the three

resulting numbers are averaged. Rates of agreement in these pools of three sets of decisions (two by humans and one by the system) that are similar to the inter-agreement rate in the original pool of three human judges validate that the system's performance is comparable with that of the physicians.

Comparing System Output with a Reference Standard

Whenever the agreement analysis provides evidence that the three physicians' judgments can form the basis for an objective gold standard, we need to create a single "best" set of responses, which becomes the reference standard and can be compared with the system's output. We considered two methods for constructing the standard:

- *Full agreement standard.* We consider only the cases in which all three physicians agree. These cases are most likely to be truly correct, but may also be the easiest ones to judge. Of the 1,440 patient–inference pairs, 1,156 fall into this category.
- *Majority evaluation standard.* We take the majority opinion as the ground truth in each case. This may increase the number of errors in the evaluation standard; two rather than three physicians must misinterpret the same data to cause an error. Since the cases in which disagreement arises are likely to be more difficult to judge, we expect a lower accuracy for the system if it is evaluated against that standard and increased uncertainty in the quality of the evaluation than what we get by use of the first method. On the other hand, this approach covers all 1,440 samples.

For each of these evaluation standards, we measured sensitivity, the percentage of abnormal situations correctly identified by the system among all abnormal situations in the reference model; specificity, the percentage of correctly avoided false positives among all non-abnormal situations in the model; and accuracy, the percentage of identical decisions between the system and the gold standard across all cases.

Results

Agreement between Human Judges

We measured the average agreement between the three composite judges for each type of inference, and for classes of inferences (hemodynamic vs. laboratory); the results are shown in Table 2. We note that the percentage of agreement is much higher in the case of laboratory inferences (91.47 percent) than in hemodynamic ones (82.81 percent). This can be attributed to two possible causes—laboratory inferences involve the assessment of a single number (rather than a curve on the chart) and thresholds for abnormal conditions are routinely reinforced in practice by the laboratory results. When the system takes the place of one of the physicians, the inter-agreement rate consistently decreases for hemodynamic inferences but increases for laboratory inferences. If the physicians were making their decisions at random according to the observed rate of "yes" answers

Table 2 ■

Average Agreement (%) Between Human Subjects and Between System and Human Subjects

Inference	Human Subjects	System and Human Subjects
Hemodynamic:		
Hypotension	75.35	71.41
Hypertension	86.11	85.42
Bradycardia	83.33	79.40
Tachycardia	86.46	85.07
AVERAGE ACROSS ALL HEMODYNAMIC INFERENCES	82.81	80.32
Laboratory:		
Acidosis	91.67	93.52
Alkalosis	83.33	83.80
Hypercardia	91.67	93.98
Hypoxia	98.61	99.07
Lowsaturation	91.67	94.44
Hypernatremia	100.00	100.00
Hyponatremia	98.61	98.61
Hyperkalemia	95.83	96.30
Hypokalemia	93.06	93.06
Anemia	80.56	81.02
Hyperglycemia	68.06	73.61
Hypoglycemia	100.00	100.00
Hypercalcemia	93.06	95.37
Hypocalcemia	94.44	95.83
AVERAGE ACROSS ALL LABORATORY INFERENCES	91.47	92.76

NOTE: Sample sizes were 192 for each hemodynamic inference, 48 for each laboratory inference.

Table 3 ■

Results for Laboratory Inferences

Inference	Full Agreement Reference Standard			Majority Reference Standard		
	Sensitivity (%)	Specificity (%)	Accuracy (%)	Sensitivity (%)	Specificity (%)	Accuracy (%)
Acidosis	100.00 (6)	100.00 (36)	100.00	100.00 (8)	95.00 (40)	95.83
Alkalosis	100.00 (3)	90.91 (33)	91.67	100.00 (6)	88.10 (42)	89.58
Hypercardia	100.00 (4)	100.00 (38)	100.00	100.00 (5)	97.67 (43)	97.92
Hypoxia	N/A (0)	100.00 (47)	100.00	100.00 (1)	100.00 (47)	100.00
Lowsat	N/A (0)	100.00 (42)	100.00	100.00 (1)	100.00 (47)	100.00
Hypernatremia	N/A (0)	100.00 (48)	100.00	N/A (0)	100.00 (48)	100.00
Hyponatremia	N/A (0)	100.00 (47)	100.00	N/A (0)	97.92 (48)	97.92
Hyperkalemia	N/A (0)	100.00 (45)	100.00	100.00 (1)	95.74 (47)	95.83
Hypokalemia	N/A (0)	97.67 (43)	97.67	0.00 (2)	97.83 (46)	93.75
Anemia	86.67 (15)	94.74 (19)	91.18	77.27 (22)	92.31 (26)	85.42
Hyperglycemia	66.67 (6)	100.00 (19)	92.00	56.25 (16)	100.00 (32)	85.42
Hypoglycemia	N/A (0)	100.00 (48)	100.00	N/A (0)	100.00 (48)	100.00
Hypercalcemia	N/A (0)	100.00 (43)	100.00	N/A (0)	100.00 (48)	100.00
Hypocalcemia	N/A (0)	100.00 (44)	100.00	50.00 (2)	100.00 (46)	97.92
Average (micro-averaged)	8.24 (34)	99.09 (552)	98.46	76.56 (64)	97.70 (608)	95.68
Average (macro-averaged)	90.67	98.81	98.04	78.36	97.47	95.69

NOTE: The sample size is 48 for all individual inferences under the majority standard and varies between 25 and 48 for inferences under the full agreement standard. The number of abnormal and normal events is listed in parentheses in the sensitivity and specificity columns respectively (their sum equals the sample size).

(13 percent for hemodynamic inferences and 11 percent for laboratory inferences), their expected rate of agreement would be much closer to their actual rate of agreement for hemodynamic inferences (77.2 vs. 82.8 percent) than for laboratory inferences (80.7 vs. 91.5 percent).

Given the lower overall values of agreement in the class of hemodynamic inferences, we conclude that the physicians are not reliable enough to be used as a gold standard for such inferences. This is further supported by the fact that our system (which applies a decision process established in the literature) agrees more with the average human judge than the other judges do in the case of laboratory inferences, but less so in the case of hemodynamic inferences.

Comparison Between Human Judges and Our System on Laboratory Inferences

Given the above analysis, it is possible, for laboratory inferences only, to compare the decisions of the three composite judges, taken collectively, with those of the system. We constructed the majority and full agreement models, as described earlier, and calculated measures of sensitivity, specificity, and accuracy for our system. Table 3 shows the results of this evaluation. Averages for each inference class are calculated by either micro-averaging, which gives each sample equal weight, or macro-averaging, which calculates the result for each inference and averages those, giving the same weight to each inference. Different inferences have different sample sizes in the case of the full agreement reference standard, because there are different numbers of patients for which all physicians agree. We report only micro-averaged results for each separate inference, in the interest of brevity.

The evaluation indicates that our system performs with high sensitivity and specificity compared with the physicians on laboratory inferences, for an average of 88 and 91 percent (micro- and macro-averaging, respectively) sensitivity and 99 percent specificity (both micro- and macro-averaging) against the full agreement model, and 77 and 78 percent sensitivity (micro- and macro-averaging, respectively) and 97 and 98 percent specificity (micro- and macro-averaging, respectively) against the majority model. In almost all cases of

individual inferences, the performance scores are in the high 90s; the few cases in which our system displayed poor sensitivity are associated with extremely low counts of abnormal findings (e.g., the system found none of the two cases of hypokalemia according to the majority model). The results in Table 3 confirm our expectation of slightly worse scores against the majority model, compared with the full agreement model (especially on sensitivity), since the former is likely to contain more marginal or harder cases.

Discussion

These results show that physicians are consistent in identifying abnormal events indicated by laboratory test results, and thus, both the full agreement and majority models can be used as a gold standard against which to evaluate the performance of MAGIC on laboratory inferences.

MAGIC performs quite well in comparison with physicians, achieving perfect accuracy on 10 of 14 inferences in the full agreement model and 7 of 14 in the majority model. Average accuracy is 98 percent for the full agreement model and 96 percent for the majority model.

For the purposes of our study, physicians cannot be used as a gold standard for the more difficult hemodynamic events. In fact, physicians agree only 83 percent of the time, while a chance assignment of results with the same proportion of abnormal results would yield a 77 percent rate of agreement. As we discuss below, our examination of discrepancies between system and physician performance revealed cases in which the physicians were clearly in error. Physicians did not identify abnormal events even when therapy was given to correct for the event. These are clear indications that the event was considered abnormal by the attending physician.

The lack of a viable gold standard in practice indicates the need for a quality assurance tool that can consistently identify and report hemodynamic problems. Given that MAGIC is based on the APACHE thresholds, it could provide a predictable means for reporting events over time. Currently, no existing tool can perform this same service. Once we have modified MAGIC's rules as suggested by our experiments and verified that they produce quality results, our plan is to install MAGIC as a quality assurance tool and do further testing.

Given that the consistency of physician decisions on hemodynamic inferences appears low, and only marginally better than chance, we carried out further analysis of the discrepancies to identify causes of the differences. We re-examined the charts of the patients for whom the physicians reported a hemodynamic anomaly that the system did not report (false negative results, if we consider the physicians a gold standard) and cases in which the system reported a hemodynamic anomaly but the physicians did not (false positive results). This was performed by the first author, as knowledge of MAGIC's inferencing procedure in addition to medical expertise was essential for this comparative analysis.

False Negative Findings

There are 46 cases in which the physicians report an abnormality that the system missed under the majority model (out of decisions for hemodynamic inferences), of which 8 also appear in the full agreement model. By examining the charts, we found five major causes for the discrepancies:

- *Artifact errors.* An artifact in the chart caused the physician to label an event abnormal when, in fact, it was not. The system correctly screened out these artifacts. (4 cases total, 0 in the full agreement model.)
- *Charting errors.* The relevant critical time point (e.g., intubation or skin incision) was not charted by the anesthesiologist in the OR. It was missing from the LifeLog data and was not shown on the chart. The system misses all inferences around such time points. Physicians, however, could sometimes infer the approximate time of such events by observing changes in the blood pressure and heart rate lines. (18 cases, 0 in the full agreement model.)
- *Window errors.* An abnormal event occurred and was reported by the physicians, but outside the 20-minute window around the critical time point. Despite directions that clearly instruct physicians to identify abnormal events within the 20 minutes before and after a time point (Appendix A), physicians did not always follow these instructions consistently. (7 cases, 0 in the full agreement model.)
- *Threshold errors.* The physicians used lower thresholds than those established in the literature and used by the system, usually by a small margin. (7 cases, 1 in the full agreement model.)
- *Corresponding changes.* The physicians used a lower threshold, as above, but were also influenced by another curve in the chart that also increased or decreased simultaneously (e.g., an increase in the heart rate along with a below-threshold increase in blood pressure may lead to

their reporting of hypertension). This merits further analysis, and may lead to the consideration of dependencies between curves in the chart for events that just missed the threshold, something MAGIC currently does not do. (10 cases, 7 also in the full agreement model.)

This analysis shows that 29 of the 46 discrepancies (63 percent) that fall into the categories of artifacts, charting errors, and window mismatches are ones in which the system should not be penalized. None of these errors occurred in the full agreement model. In the case of charting errors, if data are missing from the database, there is no way that the system can compensate. For future experiments, we may want to remove both artifact and charting discrepancies from the set of test cases for evaluation. In the case of window discrepancies, we may be able to provide better instructions to physicians. Instead of providing directions only once at the beginning of the set of instructions, we may want to highlight the appropriate time period around each critical time point.

The remaining 17 cases may indicate system omissions of abnormal conditions. For threshold errors, given the small differences between physician and system thresholds, it is possible that the system should use a small window around the APACHE thresholds, but it is equally possible that the physicians were incorrect. More experimentation is required.

Finally, correlated changes account for the largest discrepancy in missed abnormalities in the full agreement model, and these are second only to charting errors in the majority model. We will experiment with relaxing the thresholds when parallel changes in corresponding measurements occur, investigating the effect on both false negatives and false positives.

False Positive Findings

Analysis of the discrepancies that occurred when the system identified an abnormal event that the physician missed revealed seven categories of differences. Four of these—at threshold, above threshold, charting errors, and artifacts—are similar to problems identified for false negative results and are described again below. The new categories are related to the duration of an event and where exactly in relation to a critical time point it occurred. Of the total 102 false positive findings, 60 occurred in the full agreement model. In a few cases, there were two reasons for the discrepancy (e.g., artifact and short duration), and in these cases each reason was assigned 0.5 in determining counts.

- *Charting errors.* The critical time point was entered manually after it actually occurred. For example, if start of bypass is entered too late, the system detects hypotension in the interval before the start of bypass instead of after. (13 cases total, 11 in the full agreement model.)
- *Artifacts.* An artifact occurred, which the system did not screen out, whereas the physicians did. (1.5 cases total, also in the full agreement model.)
- *At threshold.* An event occurred right at the threshold specified by APACHE. The system caught these cases, whereas the physicians did not count them. (30.5 cases total, 18.5 of which occurred in the full agreement model.)
- *Above threshold.* These events were well above the APACHE threshold but, depending on duration, were missed by physicians. For example, when a parameter (e.g., blood pressure) remained low or high for a long duration, physicians often did not call it abnormal, perhaps reasoning that it must not be serious if the physicians on the case opted not to treat it. In contrast, if the parameter stayed at the same low or high rate but then had a quick rise or dip, physicians would label it abnormal. (24.5 cases total, of which 10 occurred in the full agreement model.)
- *Short duration at specific time points.* When systolic blood pressure or heart rate crossed a threshold for a short period of time at start or end of bypass, physicians ignored the abnormality. We suspect that physicians expect abnormalities briefly around bypass and do not report them. (28.5 cases total, 15 in the full agreement model.)
- *Fixed time points.* While the system always uses an interval of 20 minutes before and after a critical time point, it appears that physicians use different time intervals, depending on the critical time point. For example, they use a narrower time interval than the system for skin incision and a wider time interval for induction. (3 cases total, all in the full agreement model.)
- *Chart difficult to read.* The chart is difficult to read in cases where the hemodynamics change quickly, such as after going off bypass (e.g., distinguishing heart rate graph line from the blood pressure line can be difficult). Readability for physicians is made even more difficult by the presence of artifacts in the chart, but for the system these are screened out. A portion of a chart, illustrating these difficulties, is shown in Figure 3. (1 case total, also in the full agreement model.)

The system should not be penalized for charting errors. Both categories of threshold discrepancies show that MAGIC is in line with APACHE scores but in disagreement with the physicians. Furthermore, addressing "at threshold" discrepancies would conflict with addressing the "threshold" discrepancies identified for the false negative findings. We suggested that for false negative discrepancies, adding a window around the threshold would allow the system to identify missed cases, but this would cause an increase in the number of "at threshold" false positives. Conversely, adjusting the threshold for the false positive discrepancies would increase the number of false negatives. More experimentation with thresholds needs to be done.

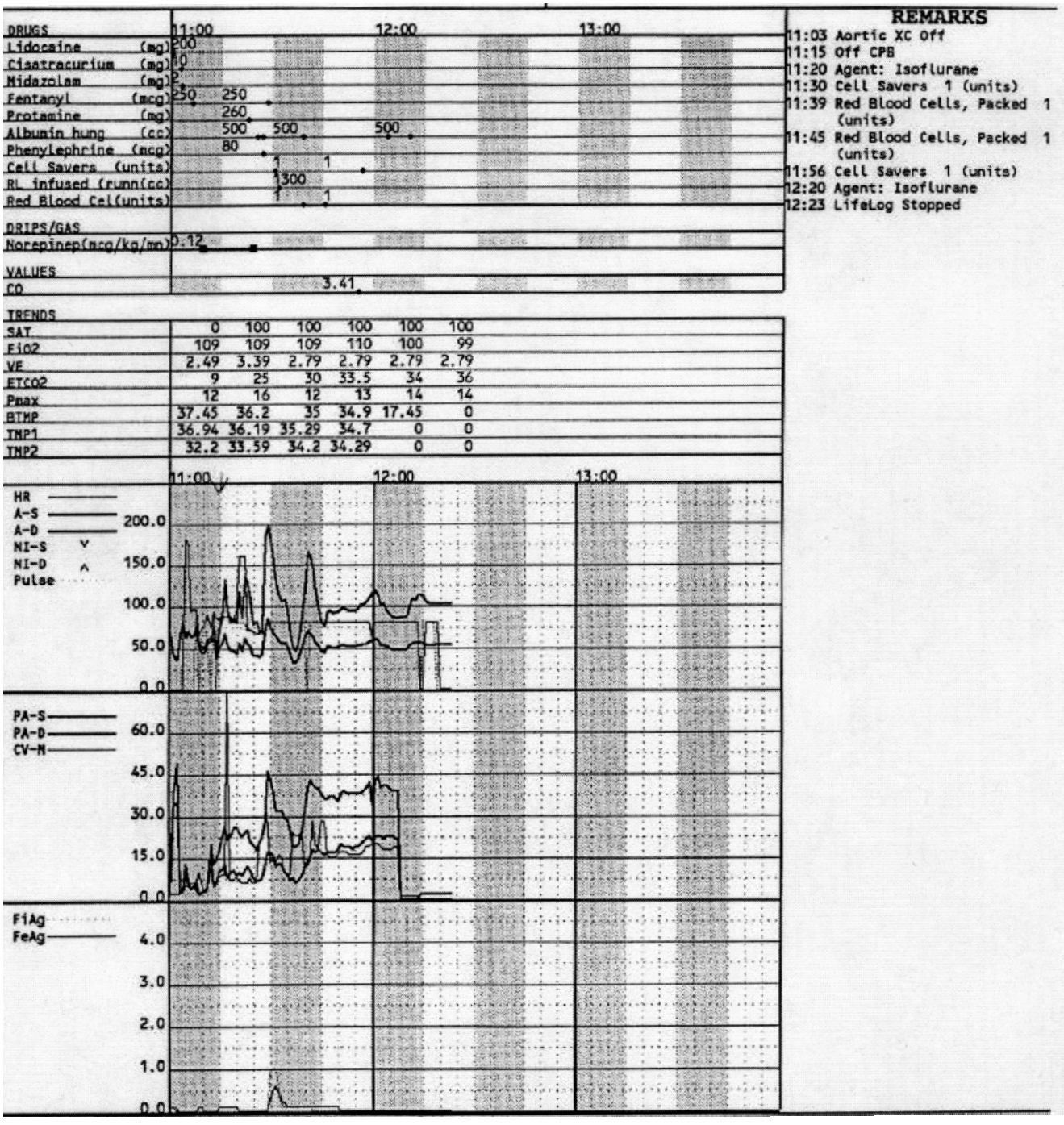

Figure 3 A portion of the chart showing end of bypass at 11:15.

In multiple cases in the set of discrepancies, the physicians were clearly in error. For example, in several cases, the patient experienced bradycardia or tachycardia before end of bypass. Therapy was given, and this was indicated on the chart. For bradycardic incidents, a pacemaker was placed, while for tachycardic incidents, cardioversion was given. In cases in which therapy has been given on the basis of abnormal physiologic parameters, this should be communicated to subsequent caregivers (e.g., the patient is on a pacemaker because he had bradycardia before the end of bypass).

On the other hand, the categories "short duration" and "fixed time point" suggest some changes that could be made to the system design. We may want to modify MAGIC so that it ignores abnormal events that occur for a short period of time immediately before or after bypass. Similarly, we may need to modify the length of the window in which we check for abnormal events around critical points such as skin incision and induction.

Study Limitations

The primary limitation of the study is the lack of data. A set of 24 patients is a small sample size. Furthermore, the interdependence of the inference decisions, which all involve the same set of patients, does not allow the computation of statistical significance levels for comparing the observed agreement with the agreement expected by chance. Nonetheless, this initial study allowed us to determine the viability of using physicians as a gold standard before going on to large-scale studies. It also allowed us to identify places where we can experiment with changes to MAGIC. In addition, it allowed us to critique and improve possible plans for conducting a real-time, prospective study.

It is extremely difficult to obtain adequate data without seriously interfering with normal physician practice in the stressful environment of the ICU. When we are ready to do a large-scale, prospective study with patients, it is important that our experimental methodology place minimal demands on their time and measure accuracy efficiently.

Future Work

Our analysis of discrepancies yielded some good insights into changes that we can institute in MAGIC and test in future studies. We found three rules used by the physicians that seem to us to be justified and that could be easily implemented. These include checking for correlated changes (when a parameter is close to a threshold and a corresponding parameter also changes, count this as an abnormal event), short duration (when a parameter crosses a threshold for a short duration immediately after going on or coming off bypass, ignore the abnormality, as it likely related to bypass), and flexible time periods (use different windows for different critical time points). Our study also shows the need for a follow-up study on the use of thresholds. It is unclear whether the physician or the system is correct when discrepancies in the use of thresholds occur, and we need more experimentation to determine when and how to change thresholds.

One major issue for future work, in particular when investigating disagreement about thresholds, is finding a good gold standard. Some alternatives that have been suggested include using experienced physicians only (but this seriously limits the supply of judges and does not reflect actual practice), using only physicians present during the operation (but this limits us to two or less per case), and using a panel of physicians who discuss the results and come to agreement among themselves as to what constitutes an abnormal hemodynamic event. Given that clear standards for abnormal hemodynamic events are not routinely taught or discussed, having a panel of physicians who spend time resolving disagreements seems the most promising alternative. Given time constraints, this would be most feasible if limited to the questionable cases identified by our current study.

Finally, we plan to test MAGIC as a quality assurance tool. Once it is used on a daily basis, we will conduct a study based on a task analysis and subjective questioning to determine whether use of MAGIC demonstrates the usefulness of the inferences. For example, we will study whether identification of abnormalities leads to differences in patient care and, through questioning, whether physicians find the identification of abnormalities useful in practice.

Conclusions

We have presented an implemented system that can detect abnormal events during cardiac surgery and, thus, can identify information that is critical to the provision of responsible care for patients arriving in the ICU. Furthermore, inferencing allows the system to summarize large amounts of collected but otherwise unexamined data in a meaningful way. Evaluation shows the system to be quite accurate for laboratory inferences, with an average accuracy of 98 percent (full agreement) and 96 percent (majority model). An

analysis of inter-rater agreement, however, showed that physicians do not agree on hemodynamic abnormal events; thus, we were left with no viable gold standard for evaluating hemodynamic events.

Examination of discrepancies between the system and physicians yielded several suggestions for future changes to MAGIC but also revealed cases in which physicians were clearly in error. For example, physician judges reported no abnormality when attending physicians on a case treated an abnormality. More important, the lack of a viable gold standard suggests that MAGIC should be tested as a quality assurance tool, providing a service that is currently lacking in practice. Such a tool could help physicians better learn how to identify abnormal events.

MAGIC is an ongoing group project that has benefited from the design and development work of Elizabeth Chen, Shimei Pan, James Shaw, and Michelle Zhou.

References ■

1. Dalal M, Feiner S, McKeown K, Jordan D, Allen B, alSafadi Y. MAGIC: an experimental system for generating multimedia briefings about post-bypass patient status. Proc AMIA Annu Fall Symp. 1996:684–8.
2. Dalal M, Feiner S, McKeown K, et al. Negotiation for automated generation of temporal multimedia presentations. Proceedings of the 4th ACM International Conference on Multimedia; Nov 18–22, 1996; Boston, Massachusetts, pp 55–64.
3. Insel J, Weissman C, Kember M, Askanazi J, Hyman AI. Cardiovascular changes during transport of critically ill and postoperative patients. Crit Care Med. 1986;14(6):539–42.
4. DiNardo JA, Schwartz M. Anesthesia for cardiac surgery. Norwalk, Conn: Appleton & Lang, 1990.
5. Becker RB, Zimmerman JE, Knaus WA, et al. The use of APACHE III to evaluate ICU length of stay, resource use, and mortality after coronary artery by-pass surgery. J Card Surg. 1995;36(1):1–11.
6. Dept. of Health (NY State), Cardiac Advisory Committee. Cardiac Surgery Report. Feb 1999. Report DOH-2243A.
7. Jordan D, Miller C, Kubos K, Rogers M. Evaluation of sepsis in a critically ill surgical population. Crit Care Med. 1987;15:897–904.
8. Knaus WA, Wagner DP, Draper EA. The APACHE III prognostic system: risk prediction of hospital mortality for critically ill hospitalized adults. Chest. 1991;100:1619–36.
9. McKeown K, Pan S, Shaw J, Jordan D, Allen B. Language generation for multimedia healthcare briefings. Proc Applied NLP. 1997:277–82.
10. Zhou M, Feiner S. Automated production of visualizations: from heterogeneous information to coherent visual discourse. J Intell Info Sys Dec. 1998;11(3):205–34.
11. McKeown K, Pan S. Prosody modeling in concept-to-speech generation: methodological issues. Phil Trans R Soc Lond. 2000; 358(1769):1419–31.
12. McKeown K, Feiner S, Dalal M, Chang S-F. Generating multimedia briefings: coordinating language and illustration. Artif Intell J. 1998;103:95–116.
13. Hripcsak G, Clayton PD, Jenders RA, Cimino JJ, Johnson SB. Design of a clinical event monitor. Comput Biomed Res. 1996;29:194–221.
14. Clancey WJ, Shortliffe EH (ed). Readings in Medical Artificial Intelligence: The First Decade. Reading, Mass: Addison-Wesley, 1984.
15. Sanborn KV, Castro J, Kuroda M, Thys DM. Detection of intraoperative incidents by electronic scanning of computerized anesthesia records. Anesthesiology. 1996;85(5): 977–87.
16. Miller P. Anesthesia. New York: Churchill Livingstone, 1981.
17. Cullen DJ, Keene R, Waternaux C, et al. Objective, quantitative measurement of severity of illness in critically ill patients. Crit Care Med. 1984; 5:137.
18. Zimmerman JE, Knaus WA, Sun X, Wagner DP. Severity stratification and outcome prediction for multisystem organ failure and dysfunction. World J Surg. 1996;20(4):401–5.
19. van Oostrom J, Gravenstein C, Gravenstein J. Acceptable ranges for vital signs during general anesthesia. J Clin Monit. 1993;9:321–5.
20. Berger J, Donchin M, Morgan L, van der Aa J, Gravenstein J. Perioperative changes in blood pressure and heart rate. Anesth Analg. 1984;63:647–52.
21. Block F. Normal fluctuation of physiologic cardiovascular variables during anesthesia and the phenomenon of "smoothing". J Clin Monit. 1991;7:141–5.
22. Hripcsak G, Kuperman GJ, Friedman C, Heitjan DF. A reliability study for evaluating information extraction from radiology. J Am Med Inform Assoc. 1999;6:143–50.
23. Friedman C, Hripcsak G. Evaluating natural language processors in the clinical domain. Methods Inf Med. 1998;37:334–44.

The Appendix appears on the following page.

Appendix

QUESTIONNAIRE GIVEN TO PHYSICIANS FOR THE EXPERIMENT

Your Name: _____

MRN of Patient: 00000000

1) Please look over the following patient record and become familiar with the case.

2) Given the following list of critical points, please check all abnormalities (blood pressure and heart rate) that occurred during the procedure before or after each critical point. Please check the "Nothing" slot if none of the listed abnormalities took place. Please check the "Not Documented" slot and continue if the critical point was not documented in the patient report. (NOTE: "Before" and "after" should be interpreted as "up to 20 minutes before" and "up to 20 minutes after," respectively. Please do not indicate abnormalities that occurred outside this range.)

Induction—There was: __Nothing __Not Documented

Hypotension	Hypertension	Bradycardia	Tachycardia
__before __after	__before __after	__before __after	__before __after

Skin Incision—There was: __Nothing __Not Documented

Hypotension	Hypertension	Bradycardia	Tachycardia
__before __after	__before __after	__before __after	__before __after

Start of Bypass—There was: __Nothing __Not Documented

Hypotension	Hypertension	Bradycardia	Tachycardia
__before __after	__before __after	__before __after	__before __after

End of Bypass—There was: __Nothing __Not Documented

Hypotension	Hypertension	Bradycardia	Tachycardia
__before __after	__before __after	__before __after	__before __after

3) Please indicate whether or not the following abnormal labs occurred during the time period listed, where pre-bypass means the entire time recorded before the start of bypass and post-bypass means the entire time after the end of bypass. Labs taken during bypass can be ignored.

Pre-bypass—There was: __Nothing __No labs documented

__Acidosis	__Alkalosis	__Hypercarbia	__Hypoxia
__Low saturation	__Hypernatremia	__Hyponatremia	__Hyperkalemia
__Hypokalemia	__Anemia	__Hyperglycemia	__Hypoglycemia
__Hypercalcemia	__Hypocalcemia		

Post-bypass—There was: __Nothing __No labs documented

__Acidosis	__Alkalosis	__Hypercarbia	__Hypoxia
__Low saturation	__Hypernatremia	__Hyponatremia	__Hyperkalemia
__Hypokalemia	__Anemia	__Hyperglycemia	__Hypoglycemia
__Hypercalcemia	__Hypocalcemia		

4) If there are any abnormalities, not covered above, that you feel are important, please list them here along with where they fall with respect to the critical points.