Research Paper ■

Monitoring Expert System Performance Using Continuous User Feedback

MICHAEL G. KAHN, MD, PHD, SHERRY A. STEIB, WILLIAM CLAIBORNE DUNAGAN, MD, VICTORIA J. FRASER, MD

Abstract **Objective:** To evaluate the applicability of metrics collected during routine use to monitor the performance of a deployed expert system.

Methods: Two extensive formal evaluations of the GermWatcher (Washington University School of Medicine) expert system were performed approximately six months apart. Deficiencies noted during the first evaluation were corrected via a series of interim changes to the expert system rules, even though the expert system was in routine use. As part of their daily work routine, infection control nurses reviewed expert system output and changed the output results with which they disagreed. The rate of nurse disagreement with expert system output was used as an indirect or surrogate metric of expert system performance between formal evaluations. The results of the second evaluation were used to validate the disagreement rate as an indirect performance measure. Based on continued monitoring of user feedback, expert system changes incorporated after the second formal evaluation have resulted in additional improvements in performance.

Results: The rate of nurse disagreement with GermWatcher output decreased consistently after each change to the program. The second formal evaluation confirmed a marked improvement in the program's performance, justifying the use of the nurses' disagreement rate as an indirect performance metric.

Conclusions: Metrics collected during the routine use of the GermWatcher expert system can be used to monitor the performance of the expert system. The impact of improvements to the program can be followed using continuous user feedback without requiring extensive formal evaluations after each modification. When possible, the design of an expert system should incorporate measures of system performance that can be collected and monitored during the routine use of the system.

■ **JAMIA.** 1996;3:216–223.

The evaluation of medical expert system performance is difficult and time-consuming. Wyatt and Spiegelhalter[1] and Engelbrecht et al.[2] define complex multistage evaluation models, with each stage designed to address different aspects of expert system performance. A key insight described by Engelbrecht et al. is that once a system has matured to the operations and maintenance phase, a continuous process of performance revalidation is required. Because medical knowledge is changing rapidly, the specialized knowledge embedded in medical expert systems is likely to require frequent revalidation.

Because of the high costs involved, most medical expert systems undergo minimal evaluation, and few systems incorporate in-the-field concurrent perfor-

Affiliations of the authors: Section of Medical Informatics, Division of General Medical Sciences (MGK, SAS), Section of Quality Management, Division of General Medical Sciences (WCD), and Division of Infectious Diseases (WCD, VJF), Department of Medicine, Washington University School of Medicine, St. Louis, MO.

Supported by grants R29-LM05387 and UO1-LM05845 from the National Library of Medicine, grant UO1-CA60267 from the National Cancer Institute, and funds provided by the Department of Pharmacy, Barnes Hospital.

Correspondence and reprints: Michael G. Kahn, MD, Division of General Medical Sciences, Campus Box 8005, 660 South Euclid Avenue, St. Louis, MO 63110.
e-mail: kahn@informatics.wustl.edu

Received for publication: 11/13/95; accepted for publication: 1/4/96.

mance monitoring. Continuous retesting and revalidating an expert system's performance would require substantial time and resources. Yet the need to revalidate expert system performance is critical because seemingly simple modifications can result in dramatic performance changes.[3]

We developed a rule-based medical expert system called GermWatcher, which attempts to identify potential nosocomial infections using inpatient microbiology culture results.[4] GermWatcher supports hospital-wide nosocomial infection surveillance activities by:

- reducing the time required by infection control nurses to review and categorize all daily positive microbiology cultures as "likely" or "unlikely" nosocomial infections, and
- imposing consistency on the categorization of positive microbiology cultures to ensure meaningful long-term historical trends.

Although GermWatcher's categorization of a specific microbiology culture result is not used to modify individual patient care, the historical database generated by GermWatcher is used to detect outbreaks of new infections and rising endemic rates of preexisting infections that may result in modifications to nursing practices or other preventive clinical interventions.

Prior to full-time field deployment, GermWatcher underwent an extensive formal validation. The evaluation procedures required significant time commitments from three infection control nurses, one infectious disease physician (the reference standard), and two members of the development staff. Although the evaluation results indicated that the system was acceptable for deployment, the analysis also highlighted areas that needed improvement. However, performing an extensive evaluation before deploying each modified version of GermWatcher would have been unacceptable to the expert clinical evaluators.

In this paper, we describe the use of continuous user feedback as a means of monitoring the impact of successive GermWatcher modifications. The feedback metrics provided indirect evidence that iterative releases of the program were addressing the deficiencies noted in the initial evaluation. The performance improvements implied by the continuous user feedback were confirmed by a second formal evaluation.

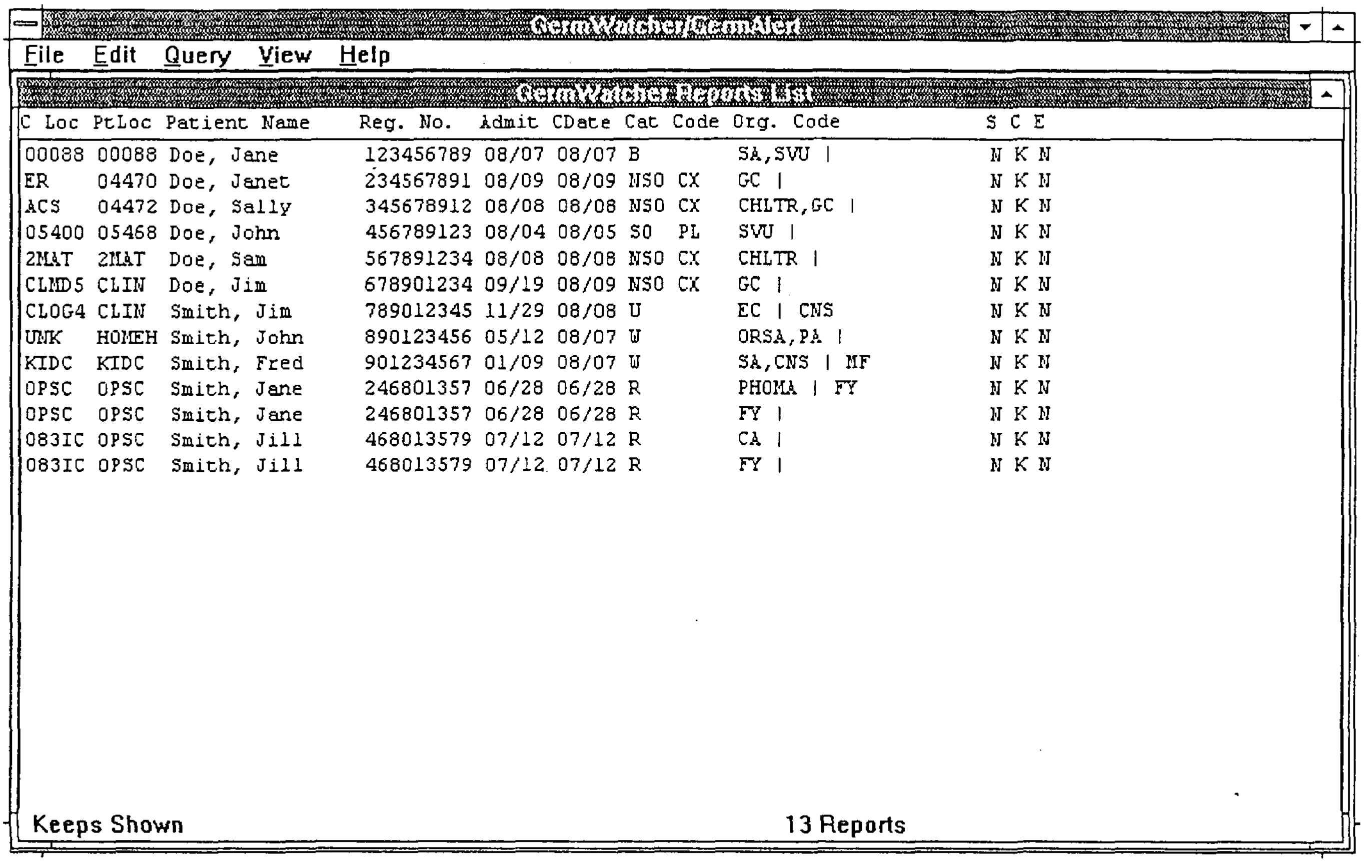
Because the monitored metrics exploit statistics that are generated during the daily routine use of the deployed expert system, we continue to assess the performance impact of additional modifications to GermWatcher without engaging in expensive or time-consuming formal evaluations. Based on these findings, we examine user feedback statistics as an inexpensive and efficient indirect measure of GermWatcher's performance.

Methods

The basic design, architecture, and first evaluation of GermWatcher are described elsewhere.[4] In brief, GermWatcher receives all final positive microbiology culture results every morning from the hospital microbiology system and assigns each result to one of three classifications: keep, discard, or watch. GermWatcher classifies a culture as "keep" if it meets the culture-based criteria of the Centers for Disease Control and Prevention's (CDC's) National Nosocomial Infection Surveillance System (NNIS) definition for a potential nosocomial infection[5,6]; as "discard" if it does not meet any NNIS definition; and as "watch" if it requires a second confirmatory culture. Over time, all cultures initially classified as "watch" are reclassified as "keep" if a confirmatory culture appears within the required time frame (usually 24–48 hours from the time of the first culture) or as "discard" if no confirmatory culture appears. One of three infection control nurses reviews and approves each culture classification before the result is stored in a long-term historical database (Fig. 1). If a nurse disagrees with GermWatcher's classification, an editing function permits changes to be made and recorded (Fig. 2) along with the stated reason for the change. For unchanged results, the name of the nurse, the time and date of the approval, and the approved classification are stored; for changed results, the original classification by the expert system and the nurse's modified classification are also stored in the database.

Figure 3 illustrates the formal evaluation methodology. In the first evaluation of GermWatcher's performance, 2,161 consecutive cultures were independently classified by three infection control nurses, an infectious disease physician, and GermWatcher. The infectious disease physician then reassessed cultures that had any discrepant classification among the five classifiers. The physician was blinded to the source of the discrepancy. The physician's second set of classifications formed the reference standard from which the initial performance metrics were computed.

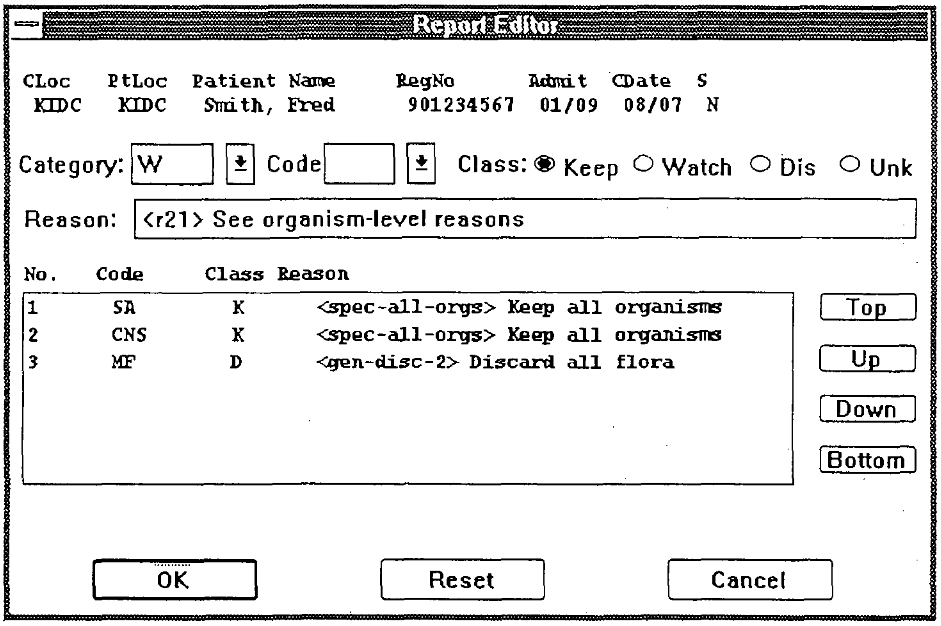
For each culture, an agreement was recorded when the reference standard and the expert system assigned the same classification code; otherwise, a disagreement was recorded. The following performance measures were defined:



```
C Loc PtLoc Patient Name     Reg. No.  Admit CDate Cat Code Org. Code            S C E
00088 00088 Doe, Jane        123456789 08/07 08/07 B        SA,SVU |             N K N
ER    04470 Doe, Janet       234567891 08/09 08/09 NSO  CX  GC |                 N K N
ACS   04472 Doe, Sally       345678912 08/08 08/08 NSO  CX  CHLTR,GC |           N K N
05400 05468 Doe, John        456789123 08/04 08/05 SO   PL  SVU |                N K N
2MAT  2MAT  Doe, Sam         567891234 08/08 08/08 NSO  CX  CHLTR |              N K N
CLMD5 CLIN  Doe, Jim         678901234 09/19 08/09 NSO  CX  GC |                 N K N
CLOG4 CLIN  Smith, Jim       789012345 11/29 08/08 U        EC | CNS             N K N
UNK   HOMEH Smith, John      890123456 05/12 08/07 W        ORSA,PA |            N K N
KIDC  KIDC  Smith, Fred      901234567 01/09 08/07 W        SA,CNS | MF          N K N
OPSC  OPSC  Smith, Jane      246801357 06/28 06/28 R        PHOMA | FY           N K N
OPSC  OPSC  Smith, Jane      246801357 06/28 06/28 R        FY |                 N K N
083IC OPSC  Smith, Jill      468013579 07/12 07/12 R        CA |                 N K N
083IC OPSC  Smith, Jill      468013579 07/12 07/12 R        FY |                 N K N

Keeps Shown                                          13 Reports
```

Figure 1 The "line-listing" window of cultures classified by GermWatcher. This window is used by the infection control nurses to approve GermWatcher's classification prior to saving the culture into a historical database. This figure shows a portion of one day's listing of "keeps" for one nurse (one culture per line). The full-text culture report is available by double-clicking the line.



```
CLoc  PtLoc  Patient Name   RegNo      Admit  CDate  S
KIDC  KIDC   Smith, Fred    901234567  01/09  08/07  N
```

Category: W Code Class: ● Keep ○ Watch ○ Dis ○ Unk

Reason: <r21> See organism-level reasons

```
No.  Code   Class Reason
1    SA     K     <spec-all-orgs> Keep all organisms
2    CNS    K     <spec-all-orgs> Keep all organisms
3    MF     D     <gen-disc-2> Discard all flora
```

Figure 2 A screen snapshot of GermWatcher's Edit Culture window. GermWatcher's results can be modified by the nurses if they disagree with the expert system's classification via this window.

- *False-positive classification:* a culture classified by the expert system as a keep but classified by the reference standard as a discard;
- *False-negative classification:* a culture classified by the expert system as a discard but classified by the reference standard as a keep;
- *Strong false-positive classification:* a culture classified by the expert system as a keep but classified by all four human evaluators as a discard;
- *Strong false-negative classification:* a culture classified by the expert system as a discard but classified by all four human evaluators as a keep.

Examination of GermWatcher's misclassifications revealed a number of coding errors, mistaken interpretations, and previously unrecognized standard practices. Within six months, a second version of the system, which corrected most of the deficiencies discovered by the first evaluation, was released. A second evaluation using 1,851 independent consecutive cultures was then performed using the same study design as the first evaluation.

During the six months in which the second version of the program was developed, interim versions that addressed high-impact problems were released without a formal evaluation. Because interim versions of the program were being released into production in rapid succession, a method of determining the positive or negative impact on performance of each new version without engaging in additional resource-intensive and time-consuming evaluations was required.

Beginning with GermWatcher's initial deployment, we monitored the number of cultures that were reclassified by the infection control nurses when they reviewed GermWatcher's output. The nurses' monthly disagreement rates, as an aggregate and by each individual nurse, were used as an indirect measure of the program's performance. We hypothesized that a significant increase in the nurses' disagreement rate would suggest a deterioration in the expert system's true performance, while a drop in the disagreement rate would reflect improvement in true performance. We used the results of the second formal evaluation to test this hypothesis and thus to evaluate the use of the nurses' monthly disagreement rate as an indirect indicator of the expert system's performance.

The z-test was used to compare proportions between the two formal evaluations and the nurses' unanimous agreement rate. The chi-square contingency test for independence was used to compare the relative false-positive and false-negative rates, and the relative strong false-positive and strong false-negative rates. Logistic regression was used to analyze temporal changes in the nurses' disagreement rate. All analyses were performed using JMP Version 3.1 by SAS Institute, Inc. (Cary, NC).

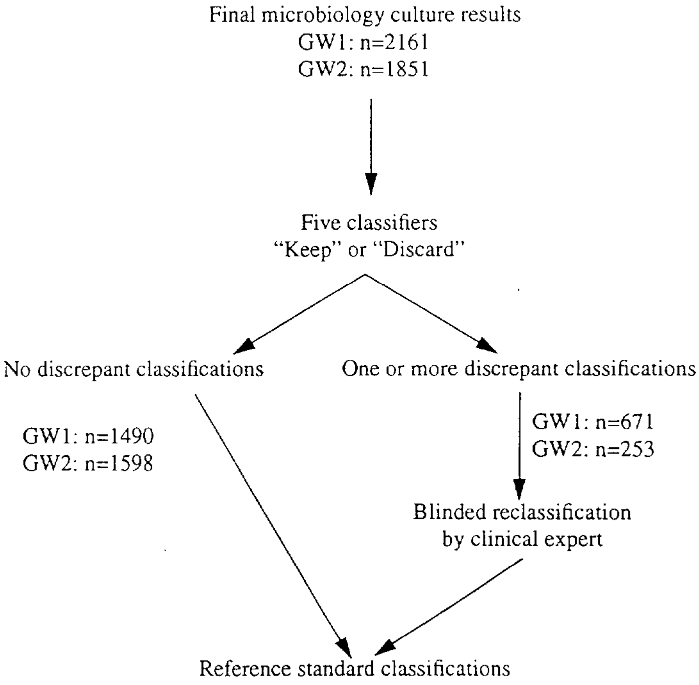


Figure 3 The evaluation methodology used for both versions of GermWatcher (GW1 and GW2). For Version 1, n = 2,161 cultures; for Version 2, n = 1,851 cultures. The five classifiers were GermWatcher, an infectious disease physician, and three infection control nurses. The clinical expert was blinded to the source of a discrepant classification during the reclassification step.

Results

Of the 2,161 cultures used in the first evaluation, 671 (31%) were reassessed by the infectious disease expert due to one or more discrepant classifications. Of the 1,851 cultures used in the second evaluation, 253 (13.6%) were reassessed (Fig. 3).

Table 1 shows the results of the formal evaluations of the two versions of GermWatcher. The first version of GermWatcher misclassified 16% of cultures—12% false positives and 4% false negatives. The second version of GermWatcher misclassified 3.5% of cultures—2.8% false positives and 0.7% false negatives. Table 1 shows a marked increase in the agreement rate between Version 1 and Version 2 ($z = -13.04$, $p < 0.001$). In addition, there was a significant decrease in the

Table 1 ■

Agreement and Disagreement Rates: GermWatcher Version 1 and Version 2 vs the Reference Standard

		Disagreement Rate			
	Agreement Rate	False Positives	False Negatives	Strong False Positives	Strong False Negatives
GermWatcher Version 1 (n = 2,161)	84.0%	12.0%	4.0%	5.3%	1.2%
GermWatcher Version 2 (n = 1,851)	96.5%	2.8%	0.7	1.5%	0.3%
	$z = -13.04$ $p < 0.001$	$z = 10.86$ $p < 0.001$	$z = 6.70$ $p < 0.001$	$z = 6.49$ $p < 0.001$	$z = 3.22$ $p < 0.001$
		$\chi^2 = 0.73$ $p = 0.39$		$\chi^2 = 3.59$ $p = 0.31$	

false-positive rate ($z = 10.86$, $p < 0.001$) and false-negative rate ($z = 6.70$, $p < 0.001$). The *relative* proportion of false positives and false negatives was unchanged between the two evaluations ($\chi^2 = 0.73$, $p = 0.39$), indicating that the observed improved agreement rate resulted from similar relative improvements in both the false-positive and the false-negative rates. Table 1 also shows a marked decrease in the strong false-pos-

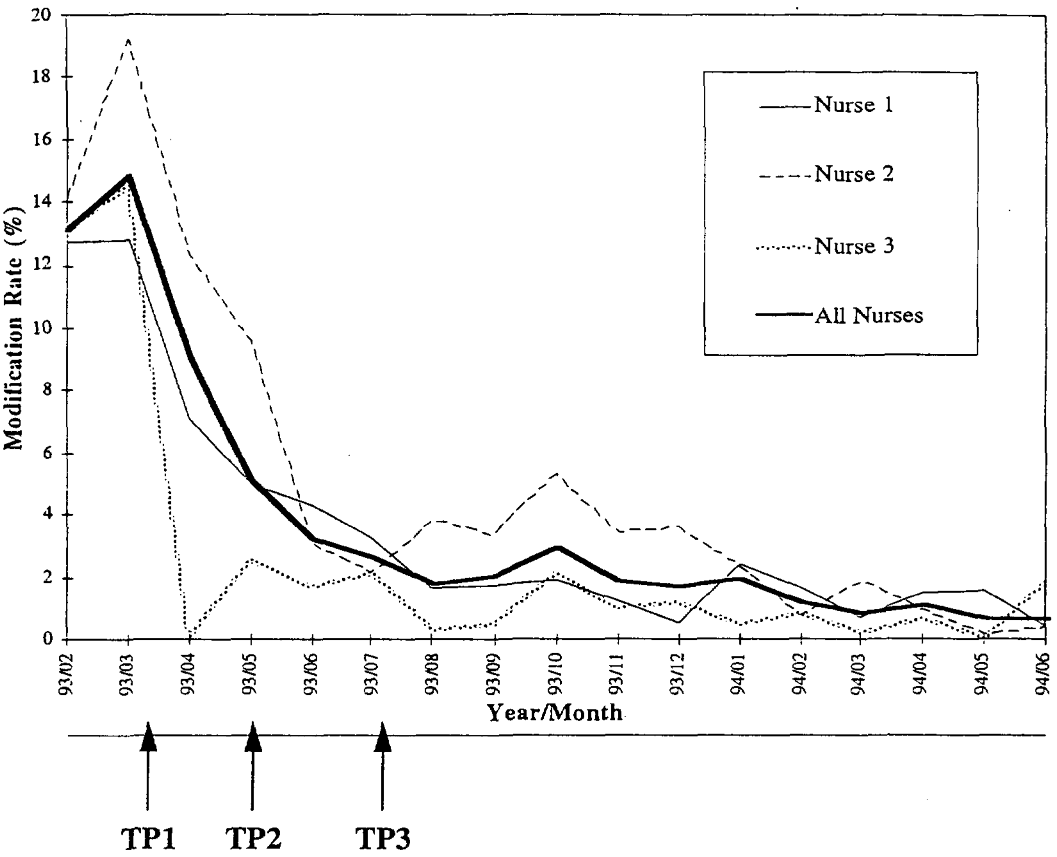


Figure 4 A graph of monthly disagreement rate, by nurse and in aggregate. Key events in the development of successive versions of GermWatcher are also noted. Time point 1 (TP1) denotes elimination of duplicate organisms and a new algorithm for identifying contaminated cultures, time point 2 (TP2) denotes organism-specific reasoning and new rules for yeast, and time point 3 (TP3) denotes new temporal processing and the release of GermWatcher Version 2.

itive rate ($z = 6.49$, $p < 0.001$) and the strong false-negative rate ($z = 3.22$, $p < 0.001$) for the two versions. As with the total false-positive and false-negative rates, the *relative* reduction in the strong false-positive and strong false-negative rates was unchanged between the two evaluations ($\chi^2 = 3.6$, $p = 0.31$), indicating similar relative improvements in the strong error rates. These data document a significant improvement in the performance of the system and confirm that many of the deficiencies noted in the first evaluation had been corrected.

Figure 4 graphs the monthly disagreement rates for each of the three infection control nurses, the aggregate disagreement rate, and the key events in the deployment of interim versions of GermWatcher. This figure shows a steady drop in the nurse disagreement rate as new versions of GermWatcher were released. Anecdotal comments from the nurses over this period confirmed this trend.

To examine the rate of decline in the nurses' disagreement rate after the deployment of GermWatcher Version 2, a logistic regression was performed using data from August 1993 to June 1994. The log-likelihood odds estimate was 0.11/month ($\chi^2 = 51.7$, $p < 0.001$), indicating a significant drop in the disagreement rate even after the release of GermWatcher Version 2.

Discussion

Jorgensen describes various sources of data that could be used to evaluate the effects of new information technology.[7] He notes:

> There are limited resources for every technology assessment, and all the desired information may either not be available or be too time-consuming to collect. One usually has to consider whether it is worthwhile doing a relatively "quick and dirty" assessment, or whether a more thorough study using all available information is necessary. Available time and resources are important criteria in the choice of method.

Jorgensen also notes that automated data acquisition based on system activity and usage logs can provide useful information for technology assessment, yet this data collection method "seems to be under-utilized for this purpose."[7]

Many systems provide a mechanism for users to send comments to the system designers. This information can provide useful anecdotal information about a system's acceptance and performance. However, the information gathered from this technique is highly biased to the most vocal users and is employed only infrequently in actual practice. For example, Hripcsak and Clayton describe the use of voluntary online comments as a means of obtaining ongoing performance monitoring for a clinical alerting system.[8] However, only 0.5% of all user interactions result in a comment and not all comments pertain only to the expert system.

Table 2 ▪

Unanimous Classification Rate among Three Infection Control Nurses and the Reference Standard

	Unanimous Classifications
GermWatcher Version 1 (n = 2,161)	78.2%
GermWatcher Version 2 (n = 1,851)	87.8%
	$z = -7.99$ $p < 0.001$

We describe a monitoring method that collects performance metrics as a by-product of the daily use of the system. Because the nurses routinely review all expert system output and the system has been designed to capture information about each modification of results by a nurse, no additional work on the part of the nurses or the developers is required to monitor the nurses' monthly disagreement rates.

Because the nurses are not the reference standard, the observed agreement/disagreement rates are only an indirect measure of the system's true performance. We can estimate the degree to which the indirect metric reflects the true performance that we would expect to observe in a more formal evaluation by examining the performance of the nurses relative to the reference standard in our two formal evaluations. Table 2 illustrates the overall agreement rate among the three infection control nurses and the reference standard during the first and second evaluations. The high rate of agreement between the nurses and the reference standard evaluation suggests that the indirect measure does indeed reflect the system's true performance. A more detailed analysis of these interrater agreement rates appears elsewhere.[9]

A key goal of GermWatcher is to maintain a *consistent* set of criteria for classifying cultures as potential nosocomial infections. Consistency is critical so that infection rates measured over long periods reflect true changes in rates rather than fictitious changes caused by temporal variations in classification criteria. We refer to the classifications provided by the infectious disease physician as a "reference" standard rather than as a "gold standard" to emphasize that Germ Watcher seeks to impose a consistent set of classification criteria that are thought to represent *potential* nosocomial infections. Without performing a clinical study, we cannot be certain about the accuracy of either our reference standard or our expert system to detect *true* nosocomial infections from inspection of culture results only.[10,11] A clinical validation that compares the expert system's culture-based classification compared with a bedside evaluation of each potential nosocomial infection currently is in progress.

GermWatcher replaced a manual, paper-based culture surveillance system. Prior to GermWatcher's deployment, each nurse was responsible for reviewing positive culture reports for a specific set of wards. There had never been an evaluation of the comparative performances of the three infection control nurses with the manual system. In view of this prior inability to compare internurse agreement rates, it is particularly noteworthy that there was a significant reduction in the variability among the classifications of the three nurses and the reference standard between the first and second GermWatcher evaluations (Table 2; $z = -7.99$; $p < 0.001$). The infection control nurses have attributed this reduction to their daily use of GermWatcher, which provides education in the form of frequent exposure to a consistent set of classification rules. Under the manual system, each nurse developed his or her own set of heuristics to classify unusual cultures. Unlike GermWatcher, the manual system provided no mechanism to monitor changes in the performance of an infection control nurse.

Although we focus here on the use of metrics for continuously monitoring the ongoing performance of the GermWatcher expert system, the same metrics can be used to analyze the nurses' use of the system. Each culture result is approved by a nurse. The nurse's name and the time and date of each approval are stored. With this information, we can calculate nurse-specific disagreement rates and time-to-approval rates. This information could be used to identify acceptance issues, training needs, or workload imbalances.

It is notable that the nurses' disagreement rate continued to drop even after the deployment of GermWatcher Version 2. We have attributed this sustained drop to the additional changes that have occurred to the program during this period. Because the indirect performance metrics have not shown an increase in the disagreement rate after the deployment of these post-Version 2 changes, we have not engaged in a third formal evaluation study.

A limitation inherent in our approach is the inability to establish causality. Thus, the changes observed in the nurses' disagreement rates could simply be due to a contemporaneous trend unrelated to rule changes. For example, an alternative interpretation of the drop in nurses' disagreement rates in Figure 4 is that the nurses gradually became less diligent in critiquing GermWatcher's output. The drop in the disagreement rate then could reflect either a loss of critical review of GermWatcher's classifications or the impact of GermWatcher's "training" the nurses rather than a true improvement in performance. An examination of only the nurses' reclassification rates in Figure 4 cannot distinguish between these competing interpretations. However, the second formal evaluation provides evidence that both the program (Table 1) and the nurses' performance (Table 2) were simultaneously improving relative to the reference standard, who was not using GermWatcher during this time, a finding that would not be expected if the nurses were simply ignoring GermWatcher's output. Also, it is unlikely that all three nurses would ignore GermWatcher to the same degree. As GermWatcher reaches expert-level performance, it will be necessary to revalidate the nurses' disagreement rates periodically by repeating formal validation experiments using the reference standard. In a companion paper, we describe the use of statistical process control charts and statistical inspection sampling methods to ensure that high-quality expert system output is sustained without the need to repeat large-scale resource-intensive formal evaluations.[12]

Accuracy is only one aspect of a software system's performance. Other aspects such as dependability, robustness, adaptability, and consistency are also critical to the successful deployment and adoption of new information systems technology. In addition to the ability to capture continuous user feedback with routine system use, GermWatcher incorporates other features to detect performance issues. For example, if the expert system is unable to classify a culture (usually due to the addition of a new organism name that is not recognized), GermWatcher notifies the developers via electronic mail. In addition, after each expert system run, a report is generated that contains key metrics that could reflect failures in the system. These error indicators lead to additional improvements to be incorporated, resulting in continued decreases in the nurses' disagreement rate after the deployment of Version 2. By incorporating multiple performance metrics into the basic design and daily use of expert systems, it is possible to monitor the impact of changes in the performance of deployed medical expert systems.

Carl Carpenter, Tom Corcoran, and Stephen Fly contributed to the first expert system prototype. Bridget Spitznagel, Edward Spitznagel, and Scott Levitt contributed to GermWatcher, Version 1. Marilyn Jones, RN, Susan O'Rourke, RN, and Beth Owens, RN, contributed to both GermWatcher evaluations. William Shannon, PhD, and Edward L. Spitznagel Jr., PhD, provided statistical expertise.

References ■

1. Wyatt J, Spiegelhalter D. Evaluating medical expert systems: what to test and how? Med Inf (Lond). 1990;15:205–17.
2. Engelbrecht R, Rector A, Moser W. Verification and validation. In: van Gennip EMSJ, Talmon JL (eds). Assessment and Evaluation of Information Technologies in Medicine. Amsterdam, The Netherlands: IOS Press, 1995;51–66.
3. Hripcsak G. Monitoring the monitor: automated statistical tracking of a clinical event monitor. Comp Biomed Res. 1993;26:449–66.
4. Kahn MG, Steib SA, Fraser VJ, Dunagan WC. An expert system for culture-based infection-control surveillance. SCAMC Proc. 1993:171–5.
5. Garner JS, Jarvis WR, Emori TG, Horan TC, Hughes JM. CDC definitions for nosocomial infections, 1988. Am J Infect Control. 1988;16:128–40.
6. Emori TG, Culver DH, Horan TC, et al. National Nosocomial Infections Surveillance System (NNIS): description of surveillance methods. Am J Infect Control. 1991;19:19–35.
7. Jorgensen T. Methods for data acquisition. In: van Gennip EMSJ, Talmon JL (eds). Assessment and Evaluation of Information Technologies in Medicine. Amsterdam: IOS Press, 1995:111–6.
8. Hripcsak G, Clayton PD. User comments on a clinical event monitor. SCAMC Proc. 1994:636–40.
9. Kahn MG, Steib SA, Spitznagel EL, Dunagan WC, Fraser VJ. Improvement in user performance following development and routine use of an expert system. In: Greenes RA, Peterson HE, Protti DJ (eds). MEDINFO '95. Edmonton, Alberta, Canada: International Medical Informatics Association/Healthcare Computing & Communications Canada, Inc., 1995;1064–7.
10. Centers for Disease Control. Public health focus: surveillance, prevention, and control of nosocomial infections. MMWR. 1992;41/42:783–7.
11. Broderick A, Mori M, Nettleman MD, Streed SA, Wenzel RP. Nosocomial infections: validation of surveillance and computer modeling to identify patients at risk. Am J Epidemiol. 1990;131:734–42.
12. Kahn MG, Bailey TC, Steib SA, Fraser VJ, Dunagan WC. Statistical process control methods for expert system performance monitoring. JAMIA. 1996;3:in press.