Review ■

Statistical Process Control Methods for Expert System Performance Monitoring

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

Abstract The literature on the performance evaluation of medical expert systems is extensive, yet most of the techniques used in the early stages of system development are inappropriate for deployed expert systems. Because extensive clinical and informatics expertise and resources are required to perform evaluations, efficient yet effective methods of monitoring performance during the long-term maintenance phase of the expert system life cycle must be devised.

Statistical process control techniques provide a well-established methodology that can be used to define policies and procedures for continuous, concurrent performance evaluation. Although the field of statistical process control has been developed for monitoring industrial processes, its tools, techniques, and theory are easily transferred to the evaluation of expert systems. Statistical process tools provide convenient visual methods and heuristic guidelines for detecting meaningful changes in expert system performance. The underlying statistical theory provides estimates of the detection capabilities of alternative evaluation strategies.

This paper describes a set of statistical process control tools that can be used to monitor the performance of a number of deployed medical expert systems. It describes how p-charts are used in practice to monitor the GermWatcher expert system. The case volume and error rate of GermWatcher are then used to demonstrate how different inspection strategies would perform.

■ **JAMIA.** 1996;3:258–269.

The literature on the evaluation of information systems and expert systems in medicine is extensive. Stead proposed a framework that relates five levels of system development to five types of evaluation studies.[1] This model explicitly recognizes routine use as a phase in system development that requires validation and efficacy studies. Although Stead identified randomized trials, inception cohorts, and impact studies as appropriate evaluation techniques for this phase, the need to incorporate continuous performance monitoring in response to modifications, updates, or program fixes was not described. In a follow-up article, Stead examined 39 manuscripts in the context of his evaluation framework.[2] Six papers focus on systems in routine use; none deals with the issue of post-deployment performance monitoring.

Van Gennip and Talmon provide a comprehensive compendium of the evaluation experiences from the European Advanced Informatics in Medicine program.[3] In this collection, Talmon and van der Loo list 684 evaluation studies related specifically to medical expert systems.[4] None of the studies in the compendium addresses the problem of the continuous evaluation of expert systems after they have been deployed into routine daily use.

Estimates from traditional information systems development literature suggest that the maintenance phase

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

Supported by grants R29-LM05387 and U01-LM05845 from the National Library of Medicine, grant U01-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.

Received for publication: 11/13/95; accepted for publication: 3/15/96.

of a deployed system consumes as much as 60% of information systems resources within large companies.[5] In describing a series of experiments in the long-term maintenance of the QMR knowledge base, Giuse states:

> Computerized medical knowledge bases must be revised constantly, and can never be considered completely finished.... Consequently, even the best medical knowledge bases are subject to obsolescence unless a careful maintenance and updating process is implemented.[6]

For knowledge-based systems, Giuse's findings suggest that continued knowledge-base maintenance can itself require significant additional maintenance resources.

If the on-going maintenance of an expert system is an unending, resource-intensive task, techniques to minimize these expensive activities are highly desired. However, the need to minimize the cost of long-term maintenance should not compromise the need to ensure that an expert system is performing adequately. Since deployed systems continue to undergo constant changes in response to errors, new requirements, or additional knowledge, continuous performance-monitoring techniques must be implemented and supported. The question we examine here is: What tools and techniques are available to design efficient monitoring schemes that result in a quantifiable level of confidence in an expert system's current level of performance?

Statistical process control provides both a theoretical framework and a set of practical techniques for addressing this question. Statistical process control was developed to interpret sources of variability in manufacturing processes and outputs. Because of this heritage, statistical process-control texts and practitioners use manufacturing terms such as producer, buyer, and defect. These terms can be readily translated into expert system (producer), clinician (buyer), and expert-system error (defect). Because of this correspondence, statistical process-control techniques developed for concurrent monitoring of the quality and acceptability of manufactured products can be used to monitor the quality and acceptability of expert-system output.

In this overview paper, we describe the use of statistical process-control techniques to monitor the performance of expert systems that are in daily use. We show how process control charts can be used as a tool for monitoring and detecting significant changes in expert-system performance over time. In addition, we show how statistical sampling inspection plans can be used to determine the frequency and intensity of post-deployment re-evaluation studies that must be performed to estimate an expert system's current error rate. We conclude by providing some initial guidance on the proper selection of these techniques in various clinical settings.

The GermWatcher expert system is used to illustrate these techniques. GermWatcher is an expert system designed to evaluate positive microbiology cultures for potential nosocomial infections.[7,8] GermWatcher encodes the culture-based criteria of the Centers for Disease Control and Prevention's National Nosocomial Infection Surveillance System. GermWatcher is in full production at two academic hospitals; the p-chart example in this review is in routine use at Washington University; the inspection sampling examples have not been implemented.

Statistical Process Control Charts

All complex processes and systems exhibit variation. Detecting and explaining variation are the fundamental concepts in statistical process control. Based on fundamental statistical concepts combined with practical heuristics, a wide variety of tools have been created to examine variation in manufacturing processes. Statistical process-control methods have been developed to detect two dissimilar sources of variation:

- Common-cause or controlled variation refers to the ever-present, small, random changes due to unknown causes. Common-cause variation produces a stable and consistent pattern of variation over time. A system or process that exhibits only common-cause variation is said to be in the state of statistical control.
- Special-cause or uncontrolled variation refers to a sustained significant change that may have one or more "assignable" causes. Special-cause variation produces a pattern of variation that changes over time. A system or process that exhibits special-cause variation is said to be out of the state of statistical control.

For systems in statistical control, experience can be used to predict future system behavior with a level of confidence determined from basic statistical principles; for systems not in statistical control, experience cannot be used to predict future system behavior using any well-founded analytic methods. Note that a system in statistical control still may be producing output that is unacceptable according to performance specifications or tolerances. The existence of statistical control only implies that the system's behavior is sufficiently stable that future system output should remain within a statistically predictable range.

Distinguishing between common-cause and special-cause variation is essential for improving the behavior of the observed system or process. When special-cause variation is present, efforts to discover an assignable source of variation must occur before further process improvements can be achieved. But when common--

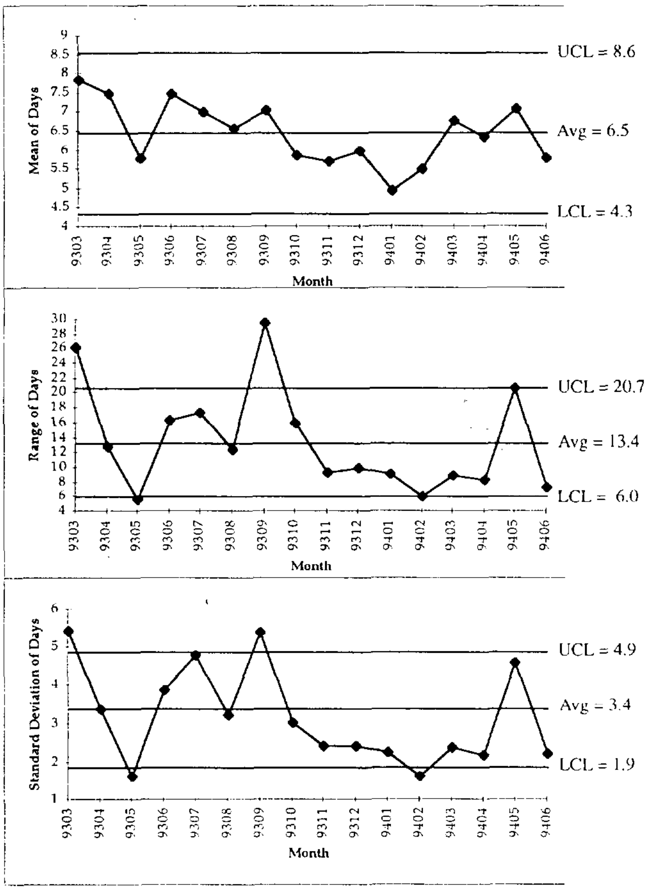

Figure 1 $\bar{X}$/R and $\bar{X}$/s Shewhart plots of the number of days from the time that a culture specimen is collected until the culture result is finalized by the microbiology laboratory and is processed by GermWatcher, from March 1993 to June 1994. Three standard deviations were used for all control limits. UCL = upper control limit, LCL = lower control limit.

cause variation is present, new processes or re-engineered systems must be developed to further improve system behavior. Thus, the tools used to improve processes or system output differ greatly in the presence of special-cause or common-cause variation.

The Shewhart control chart is the visualization tool most widely used to distinguish common-cause variation from special-cause variation. Different versions of the Shewhart chart are used to plot continuous measurement data versus discrete counts or values derived from discrete counts, such as percentages. The Shewhart chart versions most widely used in manufacturing are the $\bar{X}$/R (average and range) chart for monitoring continuous data measurements and the p-chart for monitoring the fraction of product rejected during inspection, a derived value based on discrete counts. A less common variation of the $\bar{X}$/R chart is the $\bar{X}$/s (average and standard deviation) chart.

In this section, we describe the construction and use of the $\bar{X}$/R, $\bar{X}$/s, and p-charts. We then illustrate the use of p-charts to monitor the concurrent performance of the GermWatcher expert system. P-charts, control limits, and tests for special-cause variation were analyzed using JMP Version 3.1 by SAS Institute, Inc. (Cary, NC).

$\bar{X}$/R and $\bar{X}$/s Charts

Consider a manufacturing process that produces an item with measurable dimensions. Due to known and unknown differences in materials, machines, operators, and processes, each manufactured item will have slightly different measurements. The $\bar{X}$/R and $\bar{X}$/s Shewhart charts divide the continuous output from the manufacturing process into subgroups or samples. Measurements taken from each subgroup are summarized by a subgroup mean and range (highest value minus lowest value) for $\bar{X}$/R plots or by a subgroup mean and standard deviation for $\bar{X}$/s plots. These summary statistics are plotted on a control chart (Fig. 1). As long as the process remains stable, the subgroup statistics also remain fairly constant. However, if the process is not stable, a plot of the subgroup statistics reveals one of a small number of detectable patterns. Later in this review, we describe a set of common patterns used to detect processes that are not in statistical control (see Detecting Changes: The Western Electric Rules, below).

Control limits, calculated using the same data that generate the subgroup statistics, help determine when a process is markedly out of control. Control limits are expressed as an upper control limit (UCL) and a lower control limit (LCL). Three standard deviations are traditionally used for calculating UCL and LCL. The choice of the size of the control limits is a balance between statistical theory and practical experience. Narrower limits have an increased ability to detect when a process is out of control (increased sensitivity) but also have an increased risk of erroneously inferring a stable process to be out of control (decreased specificity). The calculation of control limits for means, ranges, and standard deviations is straightforward and has been presented elsewhere.[9,10] Most statistical process control books provide tables of the constants used in calculating control limits. These tables reduce the statistical computations to simple equations that can be solved easily on the manufacturing floor.

Equations for UCL and LCL are derived from formulas that depend on the normal distribution. However, extensive empiric simulation studies have shown that the use of these formulas in many markedly non-normal distributions, such as the uniform, triangular, and exponential distributions, does not alter the performance of the charts significantly as long as subgroups contain 5 or more measurements.[11] Thus, $\bar{X}$/R and $\bar{X}$/s process control charts are extremely robust methods to monitor even data that do not follow a normal distribution pattern.

Figure 1 illustrates the use of $\bar{X}$/R and $\bar{X}$/s process control charts to monitor the number of days from the collection of a culture specimen until the culture result is finalized by the microbiology laboratory and is processed by GermWatcher. This plot does not provide information on GermWatcher's performance directly, but it does provide information on the timeliness of the expert-system output in the field. Although the mean values are in statistical control, the range and standard deviation exhibit many patterns that indicate a lack of statistical control (see Detecting Changes: The Western Electric Rules, below).

P-Charts

$\bar{X}$/R and $\bar{X}$/s charts are used for monitoring changes in continuous variables. A second class of measurements based on counts of the occurrence of some attributes, or attribute data, is more commonly used in medicine[12-16] and is more appropriate for monitoring expert-systems performance. Measurements based on counts are monitored using the p-chart, np-chart, c-chart, or u-chart; each variation has a specific setting in which it is best used. We focus on the p-chart because it is the most versatile and most widely used.

The p-chart was developed to monitor the rate of rejected product, also called the fraction defective. If inspecting an output is to result in classifying a product as accepted or rejected, then the p-chart is the most appropriate process-control chart. The fraction rejected, denoted by p, is the ratio of the number of rejected articles divided by the total number of inspected articles. The percent rejected, denoted as 100p, is 100 times the fraction rejected. Although calculations require the use of fraction rejected (p), it has become customary to plot p-charts using the percent rejected (100p).

P-charts are constructed by inspecting a sample for defects or nonconformance with specifications. The most efficient strategies for defining samples for inspection are described below (see Inspection Sampling Plans). For each sample *i*, the number of defective articles divided by the total sample size is p_i, the fraction rejected for that sample. P-charts plot p_i versus *i* (Fig. 2). In general, p-charts do not require that the number of articles inspected in each sample be equal.

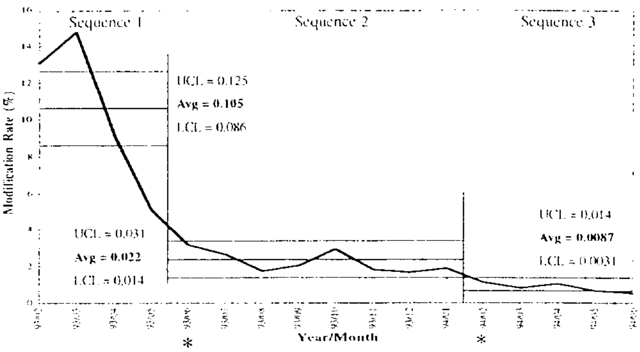

Figure 2 Shewhart p-chart of the monthly disagreement rate between GermWatcher output and the infection control nurses from February 1993 to June 1994. Three standard deviations were used for all control limits. Asterisk denotes where the data set was divided into independent sequences based on two or more points greater than 3 standard deviations from the mean (Western Electric Rule 2 in Table 2). Averages, UCLs, and LCLs for each sequence were calculated separately.

However, since the calculation of control limits is simplified with equal sample sizes, many authors use the average sample size as a simplifying assumption when different sample sizes are present.

The binomial distribution forms the statistical basis for calculating the UCLs and LCLs in p-charts. Like $\bar{X}$/R and $\bar{X}$/s charts, three standard deviations from the mean fraction rejected rate is a frequently used limit. As in the previous charts, this limit is a balance between competing goals. Other limits may be used to balance the probability of incorrectly inferring the existence of assignable causes of variation when none exists (reduced specificity) versus the cost of not detecting assignable causes of variation when they are present (improved sensitivity).

Detecting Changes: The Western Electric Rules

Control charts are used to detect processes that exhibit special-cause variation. Different deficiencies in a process cause different recognizable patterns to appear in a control chart. Although many statistical and heuristic rules for detecting special causes have been set forth, the most widely used set of basic rules was developed by managers of the Western Electric Telephone Company (now AT&T) in the 1950s.[17] Table 1 lists the 15 most common control-chart patterns described by the Western Electric investigators. Table 2 lists operational definitions for the eight basic Western Electric rules. If multiple rules are used, *any* positive rule indicates a significant change or loss of process control in the measured process. The Western Electric rules do not require complex pattern recognition and therefore can be easily applied by nontechnical personnel.

Table 1 ■

The Western Electric Control Chart Patterns

Pattern Name (alphabetical order)	Pattern Description
Cycles	Short trends that occur in repeated patterns
Freaks	Presence of a single measurement greatly different from the others
Gradual change in level	Progressive and sustained difference in measurements in a single direction
Grouping or bunching	A sudden clustering of measurements all or most quite close together
Instability	Unnaturally large fluctuations with erratic ups and downs
Interaction	Tendency of one variable to alter the behavior of another
Mixtures	Measurements that tend to fall near the high and low edges with an absence of normal fluctuations near the middle
Natural pattern	Stable pattern without trends, sudden shifts, erratic ups and downs; balanced
Stable forms of mixture	Presence of more than one distribution, each being in balance
Stratification	A form of stable mixture characterized by an artificial constancy that hugs the centerline
Sudden shift in level	A positive change in one direction
Systematic variables	Any predictable pattern; natural fluctuations are unpredictable
Tendency of one chart to follow another	Point-to-point or level-to-level correspondence in changes between two supposedly unrelated control charts
Trends	A continuous movement up or down; a long series of points without a change of direction
Unstable forms of mixture	A form of mixture caused by several distributions that are shifting or changing with respect to each other

Modified from AT&T Statistical Quality Control Handbook. Charlotte, NC: Delmar Printing, 1956.

Using P-Charts to Monitor GermWatcher Performance

GermWatcher is an expert system that uses the culture-based definitions of the Centers for Disease Control and Prevention's National Nosocomial Infection Surveillance System to classify positive microbiology cultures as potential nosocomial infections. Two extensive performance validations of the program have been described elsewhere.[7,8]

GermWatcher's design includes a means of tracking cultures when an infection control nurse disagrees with the classification assigned to those cultures by the expert system. In a previous paper, we described the use of the rate of nurse disagreement as a useful, inexpensive, indirect measure of GermWatcher's performance.[8] We use the monthly nurse disagreement rate to illustrate the use of p-charts and the Western Electric rules.

The problem we address in this use of statistical process-control charts is how to ensure that the program modifications constantly being made to the GermWatcher expert system do not result in a deterioration of its performance. Although new functionality is extensively tested in controlled laboratory conditions before release, we seek sound methods for determining if some unforeseen interaction in the deployed setting will result in reduced rather than improved expert-system performance.

Figure 2 plots the monthly nurse disagreement rate for the first 17 months of GermWatcher's deployment. This period is illustrated because improvements to early versions of the program were being rapidly developed and implemented.

During the period of time plotted in Figure 2, Western Electric Rule 2 (Table 2) was triggered at two time points. When the rule was triggered, a new segment or interval was created, and new mean and upper/lower control limits were calculated, resulting in three intervals. Based on the information presented in Figure 2 and the results of the Western Electric rules for special causes, we conclude that the GermWatcher expert system had a sustained period of special-cause variation caused by multiple releases of the software, which resulted in significant improvement in expert-system performance. As of August 1994, the program has remained in statistical control.

Based on two extensive formal evaluations, GermWatcher's performance (3.5% error)[8] is well within the acceptable specification range (less than 15% error) required by our domain experts. Because our continuous monitoring of the nurses' disagreement rate remains in statistical control, we have provided a significant level of confidence that the program's performance continues to remain well within the acceptable range.

Sampling Inspection Plans

The nurses' disagreement rate is an indirect indicator of GermWatcher's true performance. Only a more formal blinded evaluation using an infectious disease physician as the reference standard can provide information on the expert system's true performance. However, formal evaluation studies using highly skilled personnel are extremely resource intensive. Therefore, we require a sound methodology to determine the most efficient procedures for monitoring the true performance characteristics of the expert system. We introduce statistical inspection sampling as one approach for examining the trade-offs between minimizing the use of expensive resources and maximizing the error-detection rate provided by different performance-monitoring schemes.

The evaluation of quality by inspecting 100% of the product is called *screening* inspection; inspecting some but not all of the product is called *sampling* inspection.[18] In sampling inspection, the sequence or methods used to determine which product to inspect is called an acceptance plan. A perfect acceptance plan would enable the buyer to accept all nondefective items and to reject all defective items. In medical expert systems terms, a perfect acceptance plan would ensure that the clinician receives only correct results and that all incorrect results would be either rejected or corrected. Although 100% inspection is the only way to ensure complete separation of correct from incorrect results, total-inspection plans are difficult to realize in actual practice. If automated review is not possible, 100% inspection is extremely expensive and frequently subject to error due to human fatigue.

Sampling inspection plans examine only a portion of the available product. Two major classes of inspection plans are lot-by-lot plans and continuous plans.[19] Lot-by-lot plans are used when items to be inspected are produced as meaningful aggregates or lots. A sample is drawn from the lot to determine the acceptability of the entire lot. In continuous sampling inspection, current inspection results are used to determine whether sampling inspection or 100% inspection is to be used for the next set of articles. In using these techniques for monitoring expert system performance, the ability to use statistical theory to calculate the average rate of defects (expert system errors) that can be detected by alternative inspection plans is of particular interest. The goal of these techniques is to use sampling inspection plans that provide a quantifiable confidence level in estimating system performance and that minimize the cost of system monitoring.

An Overview of Acceptance Sampling

Acceptance inspection refers to the examination of a sufficient quantity of product to provide a basis for action. Key actions are accepting the product for the user, rejecting the product, or repairing the defects. Since all manufacturing processes produce some defective product, the goal of acceptance inspection is to detect when a predefined acceptable quality limit has been exceeded. A sampling plan establishes a sample size, denoted by n, and a limit called the acceptance number, denoted by c, which defines the maximum allowable number of defects in a random sample drawn from a lot. The acceptance number is chosen according to the likelihood that this number of defects would be seen in a sample drawn from a lot of size N with a true defect rate less than or equal to the predefined quality level. As n/N decreases (the sample size becomes a smaller proportion of the lot size), the two key elements of a sampling plan, the sample size and the acceptance number, become independent of the lot size.

Table 2 ▪

The Basic Western Electric Rules

Test 1	A single point falls outside of the 3 sigma limit
Test 2	Two out of three successive points fall greater than 2 sigma or beyond
Test 3	Four out of five successive points fall greater than 1 sigma or beyond
Test 4	Eight successive points fall on one side of the centerline
Test 5	Six points in a row steadily increasing or decreasing
Test 6	Fourteen points in a row alternating up and down
Test 7	Fifteen points in a row above and below the center line, all within 1 sigma
Test 8	Eight points in a row on both sides of the center line, none within 1 sigma

Modified from AT&T Statistical Quality Control Handbook. Charlotte, NC: Delmar Printing, 1956, and from Sall J, Ng K, Hecht M, Tilley D, Potter R. JMP, Cary, NC: SAS Institute, 1994.

A plot of the likelihood of rejecting a sample versus the true defect rate for various acceptance plans is called an operating characteristic (OC) curve. Figure 3 illustrates a set of OC curves for four plans with different sample sizes (n = 77,130,177,222) and different acceptance numbers (c = 0, 1, 2, 3).

The perfect sampling plan would accept all lots when the true defect rate was less than the prespecified limit and would reject all lots that exceeded this limit (the "ideal" plan in Figure 3). The probability that the observed number of defects in a sample exceeds the acceptance number when the sample is drawn from a lot with a known defect rate is calculated exactly by the binomial distribution. When the true defect rate is small, this probability is approximated using the Poisson distribution.[20] It is from the OC curves that one can assess the degree of protection against product defects for specific sampling plans. All four plans illustrated in Figure 3 are designed to reject lots drawn from a population with a true defect rate of more than 3.0% with probability = 0.90. The equivalent interpretation is that all four plans incorrectly accept lots drawn from a population with a true defect rate of more than 3.0% with probability = 0.10.

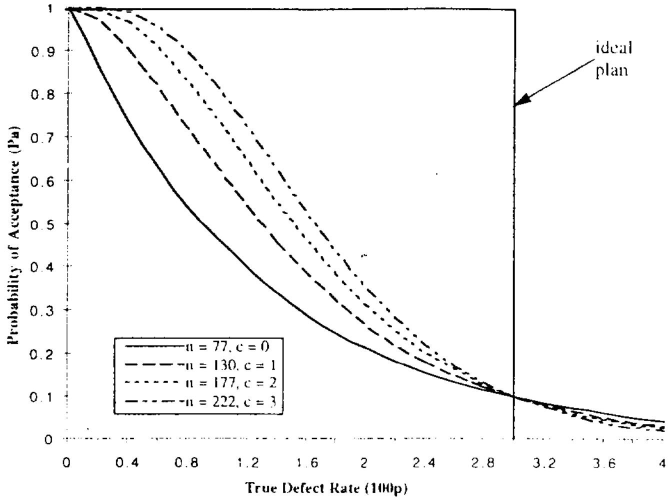

Figure 3 Operational characteristic (OC) curves for four sampling acceptance plans. For each plan, the probability of accepting a lot (P_a) based on inspecting a sample of *n* specimens drawn randomly from a population with a true percent defective rate (100p) is approximated by a Poisson distribution. Four plans which accept lots with $p > 3.0\%$ with $P_a < .05$ are shown. An ideal sampling plan always accepts lots ($P_a = 1$) when $p \leq p_{\text{acceptable}}$ and always rejects lots ($P_a = 0$) when $p > p_{\text{acceptable}}$. The ideal plan for $p_{\text{acceptable}} = 3.0\%$ is illustrated, although no such sampling plan can be created in practice.

The shape of the OC curve for a given sampling plan leads to two additional concepts: the producer's risk and the consumer's risk.[17] The producer's risk is defined as the probability or risk of rejecting product when the sample quality actually is acceptable. A plan that incorrectly rejects product when the true defect rate is acceptable causes unnecessary rework or waste by the producer, which increases the cost of producing product. The consumer's risk is defined as the probability or risk of accepting product when the sample quality actually is unacceptable. A plan that incorrectly accepts product when the true defect rate is unacceptable does not provide the quality protection desired by the consumer. The goal of inspection-sampling plans is to minimize both quantities. Table 3 illustrates the difference in producer's risk and consumer's risk for the four plans in Figure 3.

Lot-by-Lot Inspection Plans

Published sampling plans have been described in the following ways[19]:

- Acceptable quality level (AQL)—the highest percent defective that is acceptable as a process average. AQL describes the maximum percent defective that will be accepted regularly by the sampling plan or the maximum percent defective for which the probability of acceptance is very high. A plan described as having a 3% AQL means that a process with a true defect rate of 3% will result in samples that will be accepted in most cases (usually 95% acceptance). AQL embodies a producer's risk perspective.
- Lot tolerance percent defective (LTPD)—the quality above which there is a small chance that a lot will be accepted. LTPD describes the maximum allowable percent defective for which the probability of acceptance is very low. A plan described as having a 3% LTPD means that a process with a true defect rate of 3% will result in samples that would be rejected in most cases (usually 95% rejection). LTPD embodies a consumer's risk perspective.
- Point of control—the quality rate in which a sample has a probability of acceptance of 0.50. Point of control is infrequently used as a description of a sampling plan.
- Average outgoing quality limit (AOQL)—the upper limit on quality that may be expected in the long run when all rejected samples are subjected to 100% inspection, with all defective articles removed and replaced by good articles. AOQL plans require that rejected samples can have their defects removed or corrected. AOQL defines the worst average quality that can exist in outgoing product, and it attempts to combine both the producer's and consumer's risk perspectives. It can only be used in settings in which all defective product in a rejected sample can be either removed or repaired by the producer before releasing that sample to the consumer.

Single sampling rejects or accepts a lot based on one sample from a set of product. Double sampling defers rejecting a lot if the number of defects in a first sample is insufficient for acceptance but not large enough for outright rejection; a second sample is then obtained, and the total number of defects in both samples is used to make a final decision. Thus, there are four possibilities with double-sampling plans[10]:

1. Acceptance after the first sample
2. Rejection after the first sample
3. Acceptance after the second sample
4. Rejection after the second sample

Multiple-sampling plans generalize the potential to defer the acceptance or rejection of a sample until three or more samples are obtained. Table 4 illustrates the sampling procedures for single-, double-, and higher-order lot-by-lot sampling plans.[19] On average,

Table 3 ■

Producer's and Consumer's Risk in the Four Single Lot-by-Lot Acceptance Sampling Plans Shown in Figure 3. Assumes a Prespecified Quality Limit (Lot Tolerance) = 3.0%

	Inspection Plan	n = 77	n = 130	n = 177	n = 222	Ideal Plan
		c = 0	c = 1	c = 2	c = 3	
Producer's risk	Lots 1.0% defective rejected	.54	.37	.26	.19	0
	Lots 2.0% defective rejected	.79	.73	.69	.65	0
	Lots 2.5% defective rejected	.88	.84	.81	.80	0
Consumer's risk	Lots 3.5% defective accepted	.067	.058	.054	.050	0
	Lots 4.0% defective accepted	.046	.034	.028	.023	0
	Lots 5.0% defective accepted	.021	.011	.007	.005	0

for the same level of quality protection, multiple-sampling plans require less inspection than do single- or double-sampling plans. However, multiple-sampling plans are more difficult to put into operation and can cause unpredictable variability in the inspection workload.

Continuous Inspection Plans

Continuous-sample plans are appropriate when there is no natural aggregation of product into lots, such as in conveyor-line production. Because most continuous-inspection procedures remove defective articles from production, nearly all continuous-inspection plans are of the AOQL type. Four major approaches form the basis for most continuous-sampling plans. Figure 4 illustrates these four standard approaches. Single-sampling plans such as CSP-1 are easy to describe and simple to implement, but they can reject a significant number of samples that actually are within the desired performance limits, thus resulting in more 100% inspection than is necessary. Double-sampling plans, and their extension to higher order sampling plans, are more complex to implement and may occasionally require larger sampling sizes; however, they are more efficient on average and are more discriminating in their ability to accept samples that truly meet specifications and to reject samples that truly do not. CSP-2 differs from CSP-1 in that 100% inspection is not initiated when a single defect is found but is invoked only if a second defect occurs within the next *i* units. CSP-3 refines CSP-2 by including an inspection of the next four units to protect against a sudden run of unacceptable output. If none of these units is defective, sampling continues as in CSP-2. CSP-M allows for successive reduction in sampling frequency if previous sampling does not reveal any defects. When a unit is rejected, inspection occurs at the previous sampling frequency. Any number of sampling frequency levels may be provided. As with lot-by-lot inspection plans, differences in the discriminating ability of alternative sampling plans are described using OC curves.

Using Inspection-Sampling Concepts to Monitor GermWatcher's Performance

Due to their historical development in the manufacturing sector, statistical process control techniques such as inspection sampling are concerned with detecting defects in manufactured goods. If expert-system output is considered to be the "finished product," the theory and techniques of inspection sampling can be used without modification in expert-system performance monitoring. We use these techniques to determine an inspection plan that could be used by an infectious disease physician (our reference standard) to ensure that GermWatcher's performance remains acceptable. The goal is to efficiently use the physician's time to review GermWatcher output while providing some level of certainty that the expert system's output has not exceeded a predetermined error rate.

From previous studies, GermWatcher's error rate was determined to be 3.5%.[8] For this example, we examine both lot-by-lot and continuous sampling methods. Lot-by-lot plans consider each week or month to be a single lot that contains "defects" in the form of misclassified cultures. Continuous plans consider the steady daily stream of cultures to come from an ongoing production process for which the creation of

Table 4 ■

Basic Outline of Multiple Lot-by-Lot Sampling Plans. An Infinite Number of Levels Are Possible

		Combined Samples		
Sample	Sample Size	Size	Acceptance Number	Rejection Number
First	n_1	n_1	c_1	r_1
Second	n_2	$n_1 + n_2$	c_2	r_2
Third	n_3	$n_1 + n_2 + n_3$	c_3	r_3
Fourth	n_4	$n_1 + n_2 + n_3 + n_4$	c_4	r_4
Fifth	n_5	$n_1 + n_2 + n_3 + n_4 + n_5$	c_5	r_5

Note: $c_1 < c_2 < c_3 < c_4 < c_5$ and $c_i < r_i$, for all i

Modified from Bowker AH, Lieberman GJ. Engineering Statistics, 2nd ed. Englewood Cliffs, NJ: Prentice-Hall, 1972; 511.

weekly or monthly batches is an artificial aggregation of results. Since inspection corrects identified errors, we describe our alternative sampling plans in terms of AOQL performance.

The microbiology laboratory processes approximately 2,500 positive cultures per month; approximately 350 cultures are finalized each day. Using GermWatcher's current average error rate of 3.5% and assuming that we seek to maintain no less than 5.0% AOQL from our inspection sampling plan, we have the information required to use the Dodge & Romig sampling inspection tables to calculate the most efficient single and double sampling plans that match these performance characteristics.[21]

Table 5 presents the values for sample size (*n*) and acceptance number (*c*) for single and double sampling plans based on lots formed from monthly and weekly culture results. Table 5 also includes the lot tolerance per cent defective (p_t) with a consumer's risk probability of 0.10—the error rate that has a 10% chance of being incorrectly accepted given the proposed sampling plan.

Table 5 illustrates a number of key issues in selecting effective and efficient inspection plans. Although single inspection plans are the easiest to implement, they require more inspection *on average* than do double or multiple inspection plans. For example, the monthly single inspection plan requires that the infectious dis-

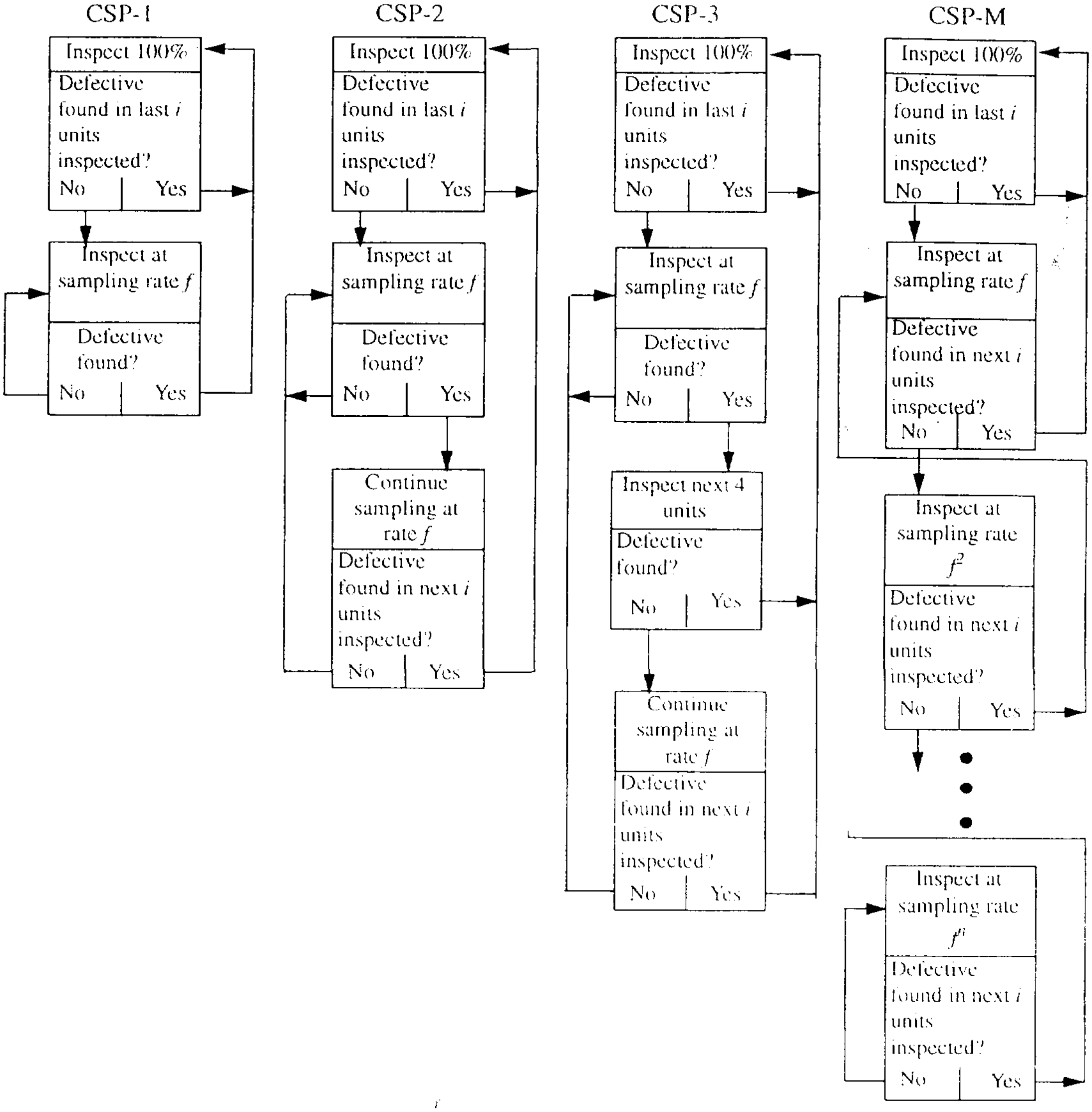
CSP-1
Inspect 100%
Defective found in last *i* units inspected?
No
Yes
Inspect at sampling rate *f*
Defective found?
No
Yes
CSP-2
Inspect 100%
Defective found in last *i* units inspected?
No
Yes
Inspect at sampling rate *f*
Defective found?
No
Yes
Continue sampling at rate *f*
Defective found in next *i* units inspected?
No
Yes
CSP-3
Inspect 100%
Defective found in last *i* units inspected?
No
Yes
Inspect at sampling rate *f*
Defective found?
No
Yes
Inspect next 4 units
Defective found?
No
Yes
Continue sampling at rate *f*
Defective found in next *i* units inspected?
No
Yes
CSP-M
Inspect 100%
Defective found in last *i* units inspected?
No
Yes
Inspect at sampling rate *f*
Defective found in next *i* units inspected?
No
Yes
Inspect at sampling rate f^2
Defective found in next *i* units inspected?
No
Yes
Inspect at sampling rate f^n
Defective found in next *i* units inspected?
No
Yes

Figure 4 Four major classes of continuous sampling plans. Multi-level plans, such as CSP-2, CSP-3, and CSP-M, can be extended to multi-level plans of arbitrary depth. Modified from Grant EL, Leavenworth RS. Statistical Quality Control, 6th ed. NY: McGraw-Hill, 1988.

Table 5 ■

Single and Double Lot-by-Lot Sampling Inspection Plan for GermWatcher. Assumes Average Process Error = 3.5% with Desired AOQL = 5.0%

		Single Sampling			Double Sampling				
Lots	N	n	c	p_t	n_1	c_1	$n_1 + n_2$	c_2	p_t
Monthly	2500	75	6	13.9%	50	2	180	14	12.0%
Weekly	350	26	2	20.0%	27	1	65	6	16.6%

N = lot size; n, n_1, n_2 = sample size; c, c_1, c_2 = acceptance number; P_t = lot tolerance percent defective with consumer's risk = 0.10.

ease physician examine 75 cultures from each monthly lot, whereas the double inspection plan usually requires the physician to examine only 50 cultures but occasionally to examine 180 cultures from each monthly lot. In addition, double inspection plans offer more protection against accepting lots with higher-than-acceptable defective rates. For example, the weekly single inspection plan has a 10% chance of accepting a lot with 20.0% defects, whereas the weekly double inspection plan has the same chance of accepting a lot with only 16.6% defects.

For continuous sampling, Table 6 illustrates the sample sizes required using a two-, three-, and four-level sampling plan at three AOQL levels using a fixed sample fraction f = 0.5. An inspection plan with AOQL = 4.0% ensures that the inspection process will allow no more than 4.0% defects, over the long run. In this example, if a three-level sampling plan were implemented, the infectious disease physician initially would inspect 19 consecutive cultures. If no cultures were misclassified, the physician would randomly sample only 50% cultures (f = 0.5). If none of the next 19 sampled cultures were misclassified, the physician would randomly sample only 25% cultures (f^2 = 0.25). If none of the next 19 sampled cultures were misclassified, the physician would randomly sample 12.5% cultures (f^3 = 0.25). When a misclassified culture was found in a sample, the physician would reinstitute sampling at the next higher sampling frequency until 19 cultures were seen without misclassification.

Selecting the Appropriate Monitoring Methodology

Table 5 illustrates the use of single and double lot-by-lot inspection sampling for GermWatcher. Table 6 illustrates the use of multi-level continuous monitoring. Given these results, is the choice of lot-by-lot or continuous sampling arbitrary? If not, how should one choose the approach to use?

Inspection sampling theory provides no guidance for selecting the most appropriate sampling plan. Domain considerations usually favor one approach over the other. In settings in which the cost of inspection is high or inspected items must be destroyed, plans that minimize the number of items inspected are favored, usually resulting in multi-stage sampling plans. In settings in which the impact of accepting substandard output is high, plans that minimize the consumer's risk usually are preferred. In our setting, clinical considerations determine if lot-by-lot or continuous sampling plans are appropriate.

GermWatcher's role is to provide data for a historical database of potential nosocomial infections that can be used to institute new infection-control policies or to investigate infectious outbreaks. In its current implementation, GermWatcher's output is not used to impact daily patient care, although its findings could lead to changes in nursing practices or other process changes. GermWatcher analyzes cultures that have been finalized by the microbiology laboratory, a process that usually takes approximately one week. By that time, the patient has received therapy based on preliminary culture results. GermWatcher does not attempt to recommend antibiotic therapies or any other aspect of concurrent patient care. The infectious disease physician could easily follow a monthly lot-by-lot inspection plan because it is possible to correct any misclassified cultures days or weeks after the initial classification by GermWatcher. Although we have also presented a weekly inspection plan in Table 5, a once-a-month inspection plan would allow the physician to schedule a portion of only one day per month for inspection sampling while ensuring a high level of

Table 6 ■

The Number of Cultures to be Inspected for Two-, Three-, and Four-Level Continuous Sample Plans with Sampling Fraction f = 0.5. AOQL levels of 4.0%, 5.0%, and 7.5% are shown

AOQL % defective	2 levels	3 levels	4 levels
4.0%	14	19	22
5.0%	11	15	18
7.5%	6	9	11

From Grant EL, Leavenworth RS. Statistical Quality Control, 6th ed. New York: McGraw-Hill, 1988; 526.

expert-system performance monitoring. If a sample is rejected, the physician could then review the entire month's output at that time. In our setting, we have proposed, but not implemented, the monthly double sampling inspection plan to our domain expert.

Unlike GermWatcher, many medical expert systems are used to provide timely patient-specific information that may be used to alter patient management. If these systems begin to produce unacceptable rates of incorrect output, this situation must be detected and corrected immediately. In this setting, continuous daily inspection sampling plans are preferred.

In both lot-by-lot and continuous inspection, the amount of expert-system output (cultures for GermWatcher) that must be examined for evaluation purposes is only a small fraction of the total. Hence, great improvement in efficiency can be gained by implementing sampling inspection plans during the maintenance phase of deployed clinical expert systems. In addition, the statistical theory behind inspection sampling allows accurate quantification of the risks involved in a particular sampling plan. The producer's risk quantifies the likelihood that a correctly functioning expert system will produce a sample that is rejected during inspection; the consumer's risk quantifies the likelihood that an incorrectly functioning expert system will produce a sample that is accepted during inspection. Based on the intended purpose of the expert system output and the clinical setting in which the expert system's output is used, these risks can be discussed openly and frankly with the intended consumers of the expert system. Alternative plans, with greater or lesser detection characteristics, can be examined based on the dangers associated with incorrect output versus the cost of more extensive output monitoring.

The statistical process control charts and inspection-sampling techniques introduced here cannot improve the impact of incorrect rules, bad assumptions, or other knowledge-base errors. Current manufacturing practices emphasize incorporating quality into all phases of product design and manufacturing. Inspection sampling and process monitoring are no longer used as the main guardians of final quality. As in manufacturing, modern software engineering practices have been developed to embed quality-improving processes in the design, development, and implementation of complex software systems.

If an expert system did not undergo continuous minor modifications and if medical practices did not change, on-going monitoring of the quality of the expert system's output would not be required once formal evaluations demonstrated acceptable system performance. Like most deployed medical expert systems, neither GermWatcher nor medical practices have remained static. Thus, some method of ensuring continued quality performance must be developed to ensure patient safety. Statistical process-control methods provide a sound and rigorous framework in which to monitor and describe the degree of safety and protection provided to patients, physicians, and other consumers of deployed clinical expert-systems technology.

Charlene Abrams, Carl Carpenter, Janette Coble, Tom Corcoran, Stephen Fly, Scott Levitt, Keith Marrs, Paul Schoening, Bridget Spitznagel, Edward Spitznagel, and Michael Spitznagel contributed to multiple versions of GermWatcher. The extensive comments of the *Journal of the American Medical Informatics Association* reviewers are appreciated.

References ■

1. Stead WW, Haynes RB, Fuller S, et al. Designing medical informatics research and library-resource projects to increase what is learned. J Am Med Inform Assoc. 1994;1: 28-33.
2. Stead WW. Matching the level of evaluation to a project's stage of development. J Am Med Inform Assoc. 1996;3:92-4.
3. van Gennip EMSJ, Talmon JL, eds. Assessment and Evaluation of Information Technologies in Medicine. Amsterdam: IOS Press, 1995.
4. Talmon JL, van der Loo RP. Literature on assessment of information technology and medical KBS evaluation: studies and methodologies. In: van Gennip EMSJ, Talmon JL, eds. Assessment and Evaluation of Information Technologies in Medicine. Amsterdam: IOS Press, 1995:283-327.
5. Banker RD, Datar SM, Kemerer CF, Zweig D. Software complexity and maintenance costs. Communications ACM, 1993;36:81-94.
6. Giuse DA, Giuse NB, Miller RA. Evaluation of long-term maintenance of a large medical knowledge base. J Am Med Inform Assoc. 1995;2:297-306.
7. Kahn MG, Steib SA, Fraser VJ, Dunagan WC. An expert system for culture-based infection-control surveillance. SCAMC Proc. 1993;171-5.
8. Kahn MG, Steib SA, Dunagan WC, Fraser VJ. Monitoring expert system performance using continuous user feedback. J Am Med Inform Assoc. 1996;216-23.
9. Wheeler DJ, Chambers DS. Understanding Statistical Process Control. Knoxville, TN: SPC Press, 1992.
10. Grant EL, Leavenworth RS. Statistical Quality Control, 6th ed. New York: McGraw-Hill, 1988.
11. Burr IW. Statistical Quality Control Methods. New York: Marcel Dekker, 1976.
12. Chamberlin WH, Lane KA, Kennedy JN, Bradley SD, Rice CL. Monitoring intensive care unit performance using statistical quality control charts. Int J Clin Monit Comput. 1993; 10:155-61.
13. Sellick JA, Jr. The use of statistical process control charts in hospital epidemiology. Infect Control Hosp Epidemiol. 1993;14:649-56.
14. Lee LT. Statistical process control charts [letter]. Infect Control Hosp Epidemiol. 1994;15:223-4.

15. Leet J. Statistical process control charts. Infect Control Hosp Epidemiol. 1994;15:223–4.
16. Hand R, Piotek F, Klemka-Walden L, Inczauskis D. Use of statistical control charts to assess outcomes of medical care: pneumonia in Medicare patients. Am J Med Sci. 1994;307: 329–34.
17. AT&T. Statistical Quality Control Handbook. Charlotte, NC: Delmar Printing, 1956.
18. Freeman HA, Friedman M, Mosteller F, Wallis WA. Sampling Inspection: Principles, Procedures, and Tables for Single, Double; and Sequential Sampling in Acceptance Inspection and Quality Control Based on Percent Defective. New York: McGraw-Hill, 1948.
19. Bowker AH, Lieberman GJ. Engineering Statistics, 2nd ed. Englewood Cliffs, NJ: Prentice-Hall, 1972.
20. Wadsworth GP, Bryan JG. Introduction to Probability and Random Variables. New York: McGraw-Hill, 1960.
21. Dodge HF, Romig HG. Sampling Inspection Tables: Single and Double Sampling, 2nd ed. New York: John Wiley & Sons, 1959.
22. Sall J, Ng K, Hecht M, Tilley D, Potter R. JMP, 1994 ed. Cary NC: SAS Institute, 1994.