Inhalation Toxicology, 18:901–910, 2006
Copyright © Informa Healthcare
ISSN: 0895-8378 print / 1091-7691 online
DOI: 10.1080/08958370600822631

informa
healthcare

Assessment of the Beryllium Lymphocyte Proliferation Test Using Statistical Process Control

Daniel J. Cher
Exponent, Inc., Menlo Park, California

David C. Deubner
Brush Wellman Inc., Elmore, Ohio, USA

Michael A. Kelsh, Pamela S. Chapman, and Rose M. Ray
Exponent, Inc., Menlo Park, California

Despite more than 20 years of surveillance and epidemiologic studies using the beryllium blood lymphocyte proliferation test (BeBLPT) as a measure of beryllium sensitization (BeS) and as an aid for diagnosing subclinical chronic beryllium disease (CBD), improvements in specific understanding of the inhalation toxicology of CBD have been limited. Although epidemiologic data suggest that BeS and CBD risks vary by process/work activity, it has proven difficult to reach specific conclusions regarding the dose-response relationship between workplace beryllium exposure and BeS or subclinical CBD. One possible reason for this uncertainty could be misclassification of BeS resulting from variation in BeBLPT testing performance. The reliability of the BeBLPT, a biological assay that measures beryllium sensitization, is unknown. To assess the performance of four laboratories that conducted this test, we used data from a medical surveillance program that offered testing for beryllium sensitization with the BeBLPT. The study population was workers exposed to beryllium at various facilities over a 10-year period (1992–2001). Workers with abnormal results were offered diagnostic workups for CBD. Our analyses used a standard statistical technique, statistical process control (SPC), to evaluate test reliability. The study design involved a repeated measures analysis of BeBLPT results generated from the company-wide, longitudinal testing. Analytical methods included use of (1) statistical process control charts that examined temporal patterns of variation for the stimulation index, a measure of cell reactivity to beryllium; (2) correlation analysis that compared prior perceptions of BeBLPT instability to the statistical measures of test variation; and (3) assessment of the variation in the proportion of missing test results and how time periods with more missing data influenced SPC findings. During the period of this study, all laboratories displayed variation in test results that were beyond what would be expected due to chance alone. Patterns of test results suggested that variations were systematic. We conclude that laboratories performing the BeBLPT or other similar biological assays of immunological response could benefit from a statistical approach such as SPC to improve quality management.

Received 25 April 2005; accepted 24 April 2006.

Financial support was provided by Brush Wellman Inc. A version of this analysis was presented at the Expert Panel on Application of the Beryllium Blood Lymphocyte Test, held at Exponent, Inc., February 2002. One of the authors (DCD) is a full-time employee of Brush Wellman. Exponent is a scientific research and consultation firm, specializing in scientific analysis of engineering, environmental, and health matters. Exponent scientists Cher, Kelsh, Chapman, and Ray, who are epidemiologists, statisticians and clinicians, designed and conducted the analyses of these data independently of the perceptions of Brush Wellman medical directors. The authors thank Marc Kolanz and Charlene Edinboro for their review of this article.

Current addresses for authors no longer at Exponent are: Daniel Cher, Kyphon, Inc., 1221 Crossman Ave., Sunnyvale, CA 94089, USA; and Pamela Chapman, ChemRisk, 25 Jessie St. at Ecker Square, San Francisco, CA 94105, USA.

Address correspondence to Michael A. Kelsh, PhD, MPH, Exponent, Inc., 149 Commonwealth Dr., Menlo Park, CA 94025, USA. E-mail: mkelsh@exponent.com

Cell-mediated immunoassays include skin tests for latex or other allergen sensitization, patch testing for contact eczematous dermatitis, histamine release tests for particular allergens or mastocytosis, lymphocyte toxicity tests for drug hypersensitivity (Lalla & Virolainen, 1974), and lymphocyte proliferation tests (LPT) for chemical hypersensitivity to metals (beryllium, (Kreiss et al., 1989; Deubner et al., 2001; Barna et al., 2003), nickel (Rasanen et al., 1991), chromium (Rasanen & Tuomi, 1992), mercury, gold, cadmium, and palladium (Stejskal et al., 1999) and other compounds. While some LPTs are in common clinical use, others have been shown to be inaccurate (Cederbrant et al., 1997; Laine et al., 1997).

The beryllium blood lymphocyte proliferation test (BeBLPT) and its predecessor, the beryllium blood lymphocyte transformation test, have been used for more than 30 years to test for the presence of beryllium (Be)-sensitized lymphocytes in human subjects. Brush Wellman Inc., a leading manufacturer of beryllium-based products, has used the BeBLPT for more than a decade in worker surveys (Deubner et al., 2001, Kolanz, 2001). During the course of these surveys, the company medical directors have observed periods during which the performance of a specific BeBLPT laboratory appeared to decline in comparison to other laboratories. This was manifest as changes in the frequency of positive results and the sensitivity, specificity, and positive predictive value of the test for the presence of granulomatous inflammation on transbronchial biopsy.

The purpose of the workplace surveys has been to learn more about the dose-response relationship between workplace air levels and beryllium sensitization (BeS) or subclinical chronic beryllium disease (CDB), as well as the possible relationship to other workplace factors, such as skin rashes. While a consistent outcome of these studies has been the observation of higher rates of BeS and subclinical CBD in specific jobs (Kreiss et al., 1996, 1997), more specific quantitative analyses of beryllium exposure and BeS and subclinical CBD occurrence have rarely identified consistent and significant dose response trends between measures of cumulative or peak beryllium exposure. Limitations in exposure assessment may explain the studies' inabilities to demonstrate dose response. Another potential explanation may be the variability of the BeBLPT, which has resulted in misclassification of workers with respect to their true BeS status. These issues continue to be important with respect to establishing an occupational exposure limit for beryllium (Borak, 2006).

Previous work has described interlaboratory and intraindividual variability of the BeBLPT among Brush Wellman employees (Deubner et al., 2001). Other researchers have examined interlaboratory variability among Department of Energy medical surveillance data (Stange et al., 2004). The current study uses statistical process control (SPC) methods to examine the reliability of BeBLPT measurements reported by four United States laboratories on employees at a beryllium manufacturer. Although the primary purpose of this analysis is descriptive in nature, one the main hypotheses of this study was that variation observed in laboratory performance was beyond expected random variation, and that this unexpected statistical variation is associated with clinical impressions of periods when inconsistent results were noted. To test this hypothesis, researchers who designed and applied the analytical procedures to the 10 years of BeBLPT survey data were not informed about the specific periods of time when clinical impressions indicated inconsistent test results.

METHODS

Data Sources

BeBLPT testing has been used at Brush Wellman as both a surveillance tool, as part of planned research studies of asymptomatic workers, and as a diagnostic aid for workers with symptomology, which may be due to a beryllium-related health effect. Over the past decade, four laboratories have been used. The goal of this study was to examine variation in the BeBLPT stimulation index (SI) results among those without obvious disease. Typically, blood from an asymptomatic employee undergoing testing was split and sent to two different laboratories. If either laboratory reported a borderline or technically inadequate result, another sample of blood was sent to that laboratory for repeat testing. If one of the two laboratory results was abnormal, testing was repeated at both laboratories with the aim of confirmation of abnormal results. All employees with abnormal tests were encouraged to undergo flexible fiber-optic bronchoscopy with bronchoalveolar lavage and lung biopsy to detect lung inflammation characteristic of CBD. Diagnosis of CBD was made on the basis of granulomatous inflammation in the biopsy specimen and either two abnormal BeBLPTs or an abnormal beryllium bronchoalveolar lavage lymphocyte proliferation test (BeBALLPT). All employees who underwent testing with the blood BeBLPT between 1992 and 2001 were included in the analysis presented in this report. Employee participation rates in facility surveys ranged from 85 to 100%.

Results for these analyses did not include any personal identifying confidential information; therefore, this research was considered exempt from a formal institutional review board (IRB) review and approval process.

The laboratories used for BeBLPT testing were selected on the basis of the scientists participating in a particular survey, geographic location of the laboratory relative to the beryllium facility, and in some cases perception regarding laboratory performance.

Statistical Process Control

Statistical process control (SPC) is typically used to monitor the stability of a process or processes (Doty, 1996; Montgomery, 2001). SPC uses statistical methods to identify periods of time during which a process goes from "in control" to "out of control." If a process is "in control," the distribution of individual results or summary statistics of the process should behave according to predictable patterns based on standard statistical theory. If a process is "out of control," the statistics of interest shift significantly. For any given process, several different statistics (e.g., mean, variance for continuous variables, proportion for binary variables) can be monitored for unexpected values. In addition, unexpected patterns of results (e.g., trends, runs of points below the mean, etc.) can indicate that a process is "out of control." Published use of SPC in laboratory testing appears to be limited to Pathfinder, an instrument used for monitoring the mechanical aspects of cytology screening (Berger, 1996). It has also been applied in quality-of-care assessments (Carey & Lloyd, 1995; Kaminsky et al., 1998; Boggs et al., 1999), and other health monitoring (McAree et al., 1998; Schwab et al., 1999; Quesenberry, 2000). In this study, SPC was used to determine periods of time during which the reported values produced by the four national laboratories were unexpected, given their overall performance.

Description of BeBLPT

During the BeBLPT, replicates of equal numbers of lymphocytes are coincubated with different concentrations of the soluble salt beryllium sulfate (typically 1, 10, and 100 μM), positive control substances such as phytohemagluttinin A (PHA) that nonspecifically stimulate the cells, and recall antigens such as tetanus toxoid or *Candida albicans* that cause specific immune responses. Other cell aliquots are incubated only with growth media as a negative control. After 4 to 7 days, tritiated thymidine is added to the mixtures, and cell growth is estimated by measuring the amount of thymidine, reflected by radioactivity counts, incorporated into cells. Results are usually expressed in terms of a stimulation index (SI), which reports the ratio of mean counts among wells with active substances (Beryllium or positive control) divided by mean counts in negative control wells (Kreiss et al., 1989; Frome et al., 1996, 2003; Deubner et al., 2001; Barna et al., 2003). Each laboratory typically reports 6 SIs for beryllium (2 poststimulation days at 3 concentrations each). The test is classified as normal, borderline, or abnormal, depending on whether zero, one, or two or more SIs equal or exceed a predetermined cut point, respectively. Three of the laboratories used 3.0 as the cut point, whereas one (Lab A) used a variable cut point based on the performance of individual laboratory technicians conducting the test.

Application of SPC to BeBLPT Data

Most subjects with CBD have elevated BeBLPT SI values. Since the goal of this study was to examine variation in BeBLPT SI results among those without obvious disease, subjects known to have CBD or who had elevated BeBALLPT results were excluded from all analysis. In addition, we also performed sensitivity analyses where subjects who underwent biopsy testing were excluded, and an analysis where all subjects were included to evaluate the impact of inclusion/exclusion criteria on SPC analyses results.

As a result of company efforts in workplace monitoring and long durations of employment for some workers, many subjects had repeated testing, especially among those who had borderline elevated BeBLPT results. Because results from such subjects are likely to be correlated, raw mean values during months with a high degree of retesting would therefore be biased. Also, the monthly sample size varied, resulting in means with different standard errors, which prevents direct comparison. To address these issues, statistical modeling procedures, described later, were used to account for repeated measurements among individuals and varying monthly sample sizes. The goal of this modeling was to focus on the effect of time while factoring out the effect of repeated observations over individuals and monthly variations in sample sizes. A 1 month grouping interval was selected as a compromise between length of the study period (about 10 years) and overall sample size.

SPC assumes that, in the absence of special causes of variation, parameters are distributed randomly (Doty, 1996). The observation of a period of time with results that are statistically unexpected suggests the occurrence of a special cause of variation. Due to random variation alone, the testing of many periods might identify a small number of "out of control" points due to chance alone. However, a large number of time points outside of the expected range would be unusual for a process that is "in control," resulting in the conclusion that the process was "out of control."

Standard SPC software does not allow SPC analysis accounting for correlated, repeated observations nor for differing sample sizes. We therefore used the SAS MIXED procedure (PC SAS version 8.02, SAS Institute, Inc., Cary, NC) (Littell et al., 1996), which enabled mixed modeling that included random effects and hierarchical nesting. These models accounted for repeated observations within individuals over time (typically as an additional evaluation of an elevated response) with appropriate covariance structures. The result was an analysis of calendar month (January 1992 to May 2001) as a fixed effect, factoring out the effect of repeated observations among individuals and varying sample sizes.

The procedure just described was performed separately on mean day 5 and mean day 7 results. Most laboratories reported day 5 and day 7 values. During some periods of time, some laboratories reported day 6 values. For efficiency of presentation, these were assumed to represent probable results on day 7 and were recoded as representing day 7. Regression coefficients (*t* values) versus time were plotted for each laboratory and each day of testing. Periods of time with very high or low *t* values or with trends or other positive "run tests" were identified on an ad hoc basis by visual inspection. Table 1 provides a list of common run tests. The run tests are based on defining zones, which are results that are a certain number of standard deviations away from the mean value; for example "Zone C" is one standard deviation away from the mean, "Zone B" is >1–2 standard deviations away, and "Zone A" is >2–3 standard deviations away from the mean value. A pattern of runs is then defined by the number of results that fall consecutively within a certain zone (Table 1).

As BeBLPT SI values, as well as *Candida*-positive control SI values, were highly right-skewed, all test values were converted

TABLE 1
Zone tests performed by SAS Proc Shewhart

Index	Pattern description
1	One point beyond Zone A (outside the control limits)
2	Nine points in a row in Zone C or beyond on one side of the central line
3	Six points in a row steadily increasing
4	Fourteen points in a row alternating up and down
5	Two out of three points in a row in Zone A or beyond
6	Four out of five points in a row in Zone B or beyond
7	Fifteen points in a row in Zone C on either or both sides of the central line
8	Eight points in a row on either or both sides of the central line with no points in Zone C

Proc Shewhart is part of the SAS QC module used for performing standard SPC analyses.

Note. Zone A extends from 2σ to 3σ and -2σ to -3σ. Zone C extends from $-\sigma$ to σ. Zone B is between Zones A and C. σ is a measure of the standard deviation of the process, which is either estimated with the entire sample or determined during a period of time in which the process is known to be in control.

with a log transformation. The statistical analysis was performed separately by two of the study authors without knowledge of which laboratories or which time periods were clinically suspected of performance degradation.

Missing Rates

Laboratories occasionally reported some test values as "missing." Reasons for missing values are typically not provided. However, it is known that laboratories use various methods to delete observations thought to be outliers or to censor values due to excessive variation in the individual tritiated thymidine count values. Monthly missing rates were calculated as:

$$\text{Missing rate} = (\text{number of test results missing})/ \\ (\text{number of possible test results})$$

As each test consisted of 6 possible beryllium SI results, if 10 subjects were tested in 1 month and 1 value was reported as missing for each of 4 people, the missing rate that month was $4/(10 \times 6) = 6.7\%$. The relationship between probit (monthly missing rates) and the monthly t values was examined using Pearson's R product-moment correlation.

RESULTS

Overall, 8808 tests were performed on 2213 subjects from 1991 to 2001 (Table 2). Peaks of testing occurred in 1993 and 1999/2000, reflecting intensive survey periods of the occupational health surveillance program at Brush Wellman Inc. Many subjects were tested repeatedly. Subjects who were ever biopsied had more repeated tests and had higher mean BeBLPT results (which prompted biopsy) (Table 3). In this workforce, 166 individuals (7.5%) underwent biopsy; their BeBLPT results were excluded from SPC analyses. Results of sensitivity analyses, where all tested subjects were included, produced only slight differences in SPC findings and are not presented.

TABLE 2
Number of BeLPT tests by year and laboratory

	Laboratory				
Year	A	B	C	D	Total
1990	1	0	0	0	1
1991	3	2	0	2	7
1992	174	165	1	10	350
1993	460	188	613	2	1263
1994	326	37	350	1	714
1995	57	33	95	1	186
1996	0	84	101	0	185
1997	0	78	83	2	163
1998	1	225	236	23	485
1999	0	921	811	127	1859
2000	3	1610	197	392	2202
2001	0	1043	343	7	1393
Total	1025	4386	2830	567	8808

Regression Results

For each laboratory the deviation of the mean day 7 monthly value from the overall mean, standardized for the number of tests performed and adjusted for repeated testing among individuals (i.e., the regression-based t value) was plotted against calendar month (see Figure 1). Overlaid on each plot is probit(P_{missing}), a relative measure of each laboratory's monthly missing rate. In all regression models and at all laboratories, the model intercept values did not differ significantly from 1.0, reflecting the expected centering of all stimulation indices (SIs) around unity for a population of workers with a low rate of sensitization.

Visual inspection of the figures shows substantial evidence of nonrandom variation, with several periods of time during which values were consistently higher or lower than expected (Figure 1). In most cases, mean day 5 values correlated strongly with mean day 7 values; therefore, unexpected values were also strongly correlated. Several periods were observed during which some of the other SPC "zone rules" or run tests were triggered (e.g., two of three points in a row in Zone A or beyond). Zone rules were triggered in SPC charts of three of the four labs. Of 38 total months in which testing was done, Lab A had 8 (day 5 results) and 10 (day 7 results) months with unexpected results (t value >3 or <-3), Lab B had 18 and 25 of 88 months, Lab C had 7 and 24 of 92 months, and Lab D had 5 and 4 of 28 months.

TABLE 3
Mean BeBLPT results for day 5 and day 7 by biopsy status

	Day 5			Day 7		
Biopsy ever done, result	*n*	Mean (mean log)	Standard error (log)	*n*	Mean (mean log)	Standard error (log)
No	7751	0.79 (−0.10)	(0.01)	7684	0.55 (−0.26)	(0.01)
Yes, negative	684	2.57 (0.41)	(0.04)	678	1.44 (0.16)	(0.04)
Yes, positive	356	6.31 (0.80)	(0.06)	355	3.09 (0.49)	(0.07)

Note. Results are expressed in natural scale, with log scale in parentheses.

The high frequency with which each lab showed unexpected values was highly unlikely to be due to chance alone. Finally, trends in some laboratories (e.g., upward trend in Labs B and C beginning in mid-1999) were observed. These trends and other "zone rule" tests in SPC charts provide evidence that laboratory processes were not in control. Importantly, the regression results served to confirm the results from clinical impressions (Table 4).

Missing Values

Lab A did not report values for SIs 18% of the time on day 5 and 32% of the time on day 7 (Table 4). The missing value rates from Lab D were 3% and 9%, respectively. In contrast, the missing values rates on either day from Labs B and C were less than 1%.

In Lab A, missing rates varied inversely with mean monthly values, with a Pearson correlation of $-.40$ ($p = .01$) for day 5 and $-.72$ ($p < .0001$) for day 7 (Table 5). That is, when missing rates were higher than average, the mean reported SI values were lower than expected, suggesting that not reporting values biases results downward. A similar phenomenon was observed at Lab D, with a Pearson correlation of $-.558$ ($p < .002$) and $-.637$($p = .0003$) for days 5 and 7, respectively. However, at Labs B and C, missing values rates were not related to mean reported SI values.

DISCUSSION AND CONCLUSIONS

We used regression-based techniques to perform a statistical process control (SPC) analysis of BeBLPT results among employees at a large manufacturer of beryllium materials. To our knowledge, application of similar statistical evaluation has not been applied to the BeBLPT or other lymphocyte proliferation tests. The analysis was restricted to tests among exposed but apparently unsensitized employees. However, analyses that included all subjects or also excluded any subjects who were biopsied produced similar findings to the analyses that were restricted to unsensitized subjects. Unexpectedly high and/or low values were observed, along with trends and other zone-rule violations that strongly suggested periods of time at each laboratory when BeBLPT results were "out of control." In any test, monthly variation in mean observed values is to be expected over time. However, the amount of variation observed was shown to be statistically excessive. This excess variability was observed at both the start and end of the 10-yr testing period, suggesting no improvement over time. These results indicate the need for careful scrutiny of BeBLPT data when used in occupational medical surveillance or in epidemiologic studies of sensitization patterns as they relate to exposure. Our findings are not suggestive of recent improvements in laboratory performance as reported by Stange et al. (2004) in their analysis of BeLPT data collected from Department of Energy (DOE) workers.

TABLE 4
Correlation between clinical suspicion and regression results

Laboratory	Clinical observation	Regression
A	1993–1994: Increased rate of positive tests, lower positive predictive value	Decrease of mean values early 1993, higher mean values late 1993/early 1994
B	8/2000–2001: Decrease in PHA values	Increase in Be values most of 2000–2001 Decrease in PHA values
C	6/1999–10/1999: Decreased rate of positive tests, increase in rate of *Candida* failures, increase in control counts; 10/2000–1/2001: Decrease in PHA and *Candida* values	Decrease in *Candida* values, rising Be values
D	2–5/2000: Increased rate of positive tests, lower positive predictive value	Increase in Be values

Potential Causes of Out-of-Control Periods for BeBLPT

For BeBLPT testing of a population of workers over time, sources of variation that could cause shifts in mean monthly laboratory values include:

1. Nonrandom distribution of individuals who are sensitized among the entire population tested.
2. Changes in cell reactivity in a given person over time (e.g., development of sensitization, resulting in increasing responses, or loss of sensitization, resulting in decreased responses). This has been observed during retesting of individuals who had originally tested abnormal and then tested normal several years later (Deubner et al., 2001).
3. Changes in cell reactivity due to extraneous biological factors, such as viral illnesses or drugs that suppress immune responses.
4. Changes in laboratory conditions under which the test is performed. Such changes could include different technicians

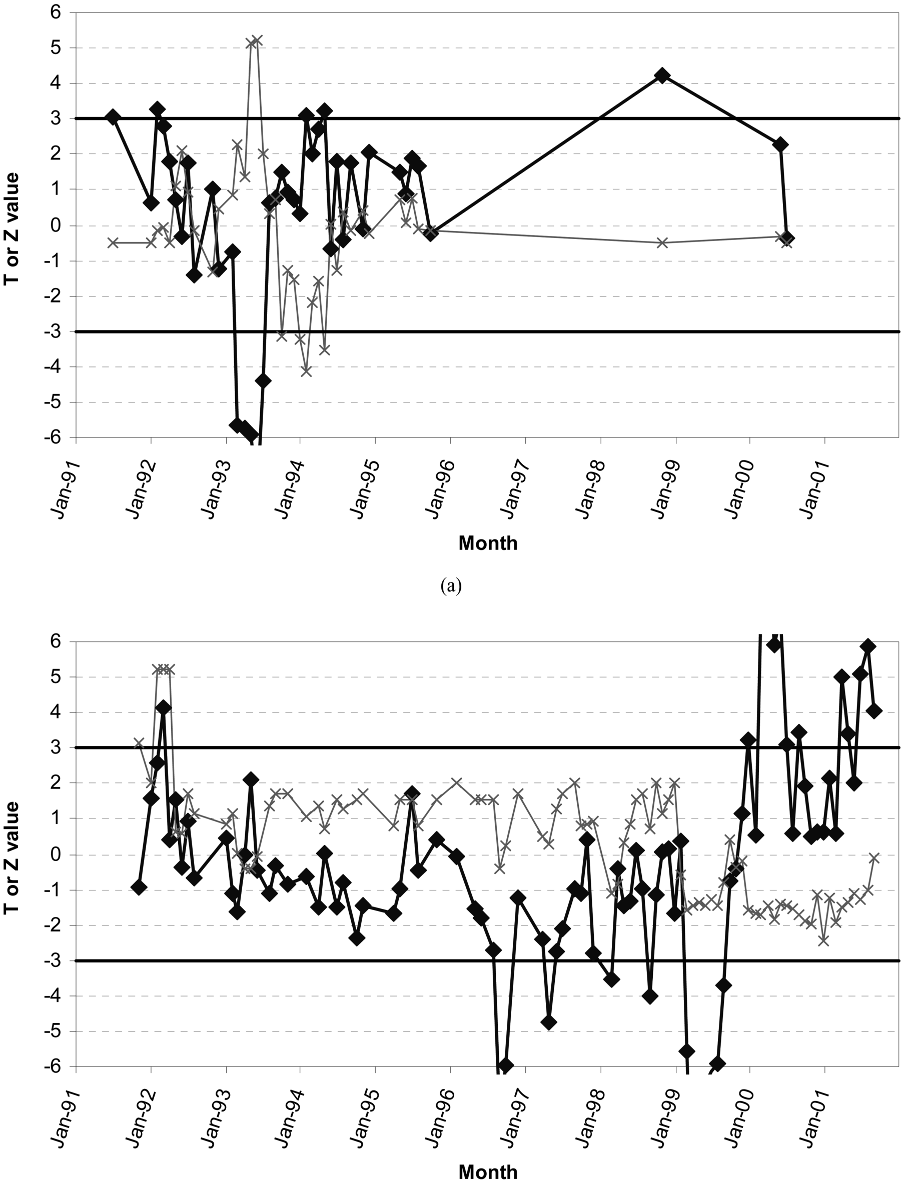
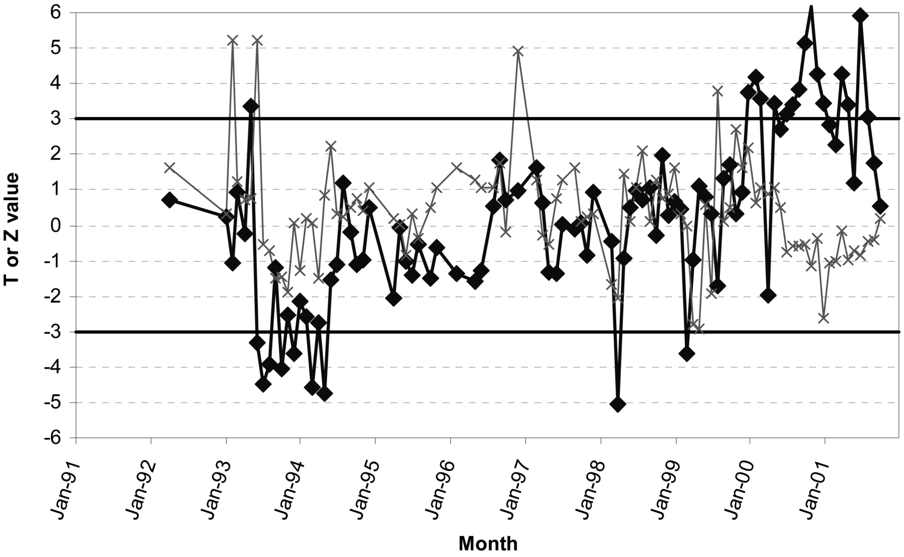
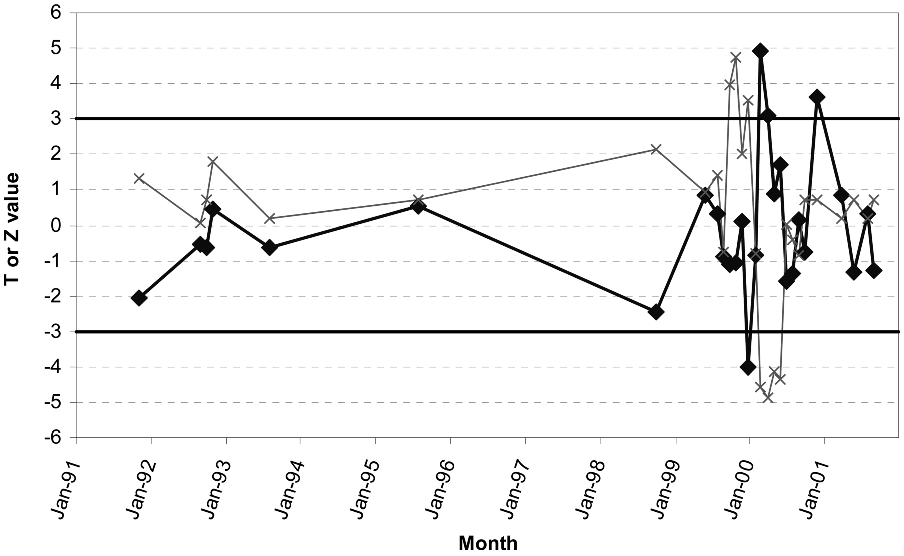


FIG. 1. BeBLPT regression test results for each of the four separate laboratories (labeled as a–d). Each plot shows the standardized deviation from the grand mean (i.e., *t* value for regression coefficient) for each month of testing (diamonds) and the standardized missing rate (×). Values higher than 3 or lower than −3 represent unexpected values. Low missing rate values indicate a relatively lower number of missing values. Only day 7 results are shown; day 5 results were very similar. *(Continued)*



(c)



(d)

FIG. 1. (Continued).

performing the test, day-to-day reliability of the technician, characteristics of test substances (e.g., changes in serum lots or degradation of serum over time, differences in reagents used), or laboratory errors (e.g., adding the test substance twice to a well or bacterial contamination). The fact that there is no positive control available to calibrate this test on a routine basis may contribute to laboratory variation.

5. Changes in workplace exposures, either levels of beryllium exposure or type of beryllium exposure (chemical form, particle size changes).
6. The size of control groups among laboratories varies and is generally small. In addition, laboratories tend to use the same people repeatedly to generate the baseline value for calculating the stimulation index. This procedure may contribute to the instability of the simulation index.
7. Some laboratories have a more stringent protocol for excluding extreme SI values, which may reduce laboratory variability relative to other labs.

Of the potential causes just outlined, laboratory variation seems the most reasonable. There were no obvious seasonal

TABLE 5
Correlation of probit(missing rate) with standardized regression coefficients from SPC analyses by month

Lab	Missing value rate (%); day 5, day 7	Correlation of probit (missing rate) with monthly mean value: Pearson *R* (*p* value)	
		Day 5	Day 7
A	18, 32	−.40 (.01)	−.72 (<.0001)
B	<0.01, <0.01	.08 (.43)	−.07 (.54)
C	<0.01, <0.01	.17 (.10)	−.07 (.51)
D	3, 9	−.56 (.002)	−.64 (.0003)

patterns, ruling out viral illnesses or drugs used to treat them as a potential cause of observed results. Indeed, many of the "out of control" points were values that were higher than expected, which is unlikely to be attributable to response-*suppressing* viral illnesses. Systematic variation in the set of workers tested would likely have given rise to unexpectedly high or low values at more than one laboratory simultaneously; this did not occur. Changes in beryllium exposure may have occurred during the 10 year study period; however, one would have to hypothesize that such changes in exposure were very rapidly associated with changes in immune responses among the population of workers tested. It has been shown that among patients never previously exposed to beryllium, a hypersensitivity reaction can occur within 1 to 2 weeks (Curtis, 1951). However, this is at much higher beryllium concentrations than current workplace exposures, so although this is possible, it appears unlikely.

SPC analysis was undertaken both to examine laboratory variability and to determine whether clinical suspicion of laboratories being "out of control," based on an ad hoc analysis by medical staff at Brush Wellman, agreed with the statistical evaluation. As shown in Table 4, each case of clinical suspicion of unexpected results or trends was confirmed in regression testing. Note that all statistical testing was performed without specific information regarding "out-of-control" periods suspected by company clinicians. Of interest, there were also periods of time not suspected clinically that appeared "out of control" by statistical analysis. Thus, this analysis confirmed that clinical suspicion can be a useful monitoring tool of laboratory processes and that statistical testing can provide quantitative confirmation of clinical impressions as well as identify other periods that were not identified by clinical impression. However, the fact that laboratories were not chosen randomly could bias results, although the direction of this potential bias is unknown.

A relationship was observed in two of the four laboratories between monthly laboratory test missing value rates and mean test result (SI) values. It is known that laboratories performing the BeBLPT use particular heuristics to remove some observations thought to represent outliers. For example, Laboratory C used a method in which replicate values resulting in excessive variation (coefficient of variation >30%) were removed until the remaining points showed less variation. The test was then interpreted using the remaining points. Statistical methods for BeBLPT that are insensitive to outliers have also been reported (Frome et al., 1996, 2003) and are being used by some laboratories. The goal of all such analyses is to remove from consideration values that are highly likely to represent outliers due to laboratory error, as opposed to outlying values that represent true cellular responses. In theory, removal of an outlier representing laboratory error should not "adversely" affect the distribution of the remaining points. That is, the resultant distribution after removal of the outlier should resemble the distribution of values in the previous month, the subsequent month, and all other months. However, if outlier values not representing laboratory error were removed, the resultant distribution would be shifted away from the removed, but "real," values; the more values that are removed, the more the resultant shift. This phenomenon, which was observed to a statistically significant and substantial degree in Lab A and Lab D (Pearson correlation values ranging from .4 to .7), suggests that the values removed may reflect actual biological processes as opposed to errors. The removal of values would itself constitute an error, and interpretation of the remaining values would be suspect. This relationship highlights difficulties with some of the outlier rejection methods. Alternate statistical methods such as the least absolute value and statistical biological positive have been described (Frome et al., 1996, 2003), but these methods have not been proven to be less sensitive to outliers. In the setting of BeBLPT testing, erroneous removal of high values would result in more subjects having (potentially false) "normal" tests; erroneous removal of results with low values would result in more subjects having (potentially false) "abnormal" tests.

Note that removal of data points deemed to represent error or excessive variation probably resulted in the remaining data set having less variability than the "unedited" (but unavailable) data set. Despite this removal of excessive variability, our analyses identified many time periods that were "out of control." Thus, our analysis may have been conservative in estimating the number of time points that were "out of control."

Limitations

Although SPC is an accepted method to examine processes with a long history of use, several assumptions of SPC may not be met in this dataset.

First, process parameters (e.g., mean, standard deviation) are typically determined during a period in which the process is known to be "in control." Since laboratories do not typically publish such parameters, and periods of time during which the laboratory measurements were "in control" were not known or published, it was assumed that measurements from the entire study period could be used to estimate these unknown parameters.

Second, SPC assumes that a process is governed by a single set of parameters for a distribution (e.g., a single mean, variance). It is possible that the heterogeneous nature of exposure and potential sensitization status among Brush Wellman employees tested during the past decade results in a heterogeneous underlying distribution. The combination of such heterogeneous distributions might produce a "flattened out" or widened distribution of test results. This was not observed; moreover, even if the raw distribution were "flat," mean values tend toward normality (central limit theorem). This finding suggests that the mild amount of heterogeneity of the data did not negatively impact our analysis. Third, if subjects from the presumed different subpopulations are tested nonrandomly, biases could have occurred. There was no way to directly evaluate this potential bias, though it seems highly unlikely; furthermore, this bias would not explain the observed relationship between missing rates and mean reported values. Fourth, all day 5 and all day 7 measurements were grouped prior to calculating monthly means; this grouping produces a type of smoothing that is likely to have reduced the overall variation examined in comparison to an analysis of individual concentrations on individual days. An analysis of individual concentrations would therefore be more likely to show variation over time.

Implications for Clinical Practice

The impact of month-to-month variations in test measurements and deletion criteria is evident in persons who test positive one month and negative or borderline the next month, or positive at one laboratory and negative at another. The actual impact of variation in test results on time to diagnosis, treatment patterns or clinical impairment among sensitized workers is difficult to estimate. For example, a worker may have a positive result, followed by a negative result, followed by another positive result, followed by a biopsy showing granulomatous changes. Overall, the negative result, while possibly a laboratory error, may have delayed diagnosis by only a small period of time, typically producing no clinical consequences. The potential for clinical consequence is low because asymptomatic or subclinical CBD is not medically treated and the long-term effect of removal from beryllium exposure is unknown. One concern would be decreases in test specificity, which would result in increased false positives, a lower positive predictive value, and therefore a larger number of unnecessary and more invasive clinical procedures such as biopsy.

In addition to the potential medical harm in false positive results are the anxiety and life decisions resulting from receipt "abnormal" test results. Even in the absence of demonstrable lung disease, workers with abnormal results are often advised to cease exposure to beryllium and are warned that they are at higher risk of developing CBD in the future. Active beryllium workers may make decisions to change careers or employers as a result of falsely abnormal test results, with significant social and economic consequences, and may consider themselves affected and impaired for their entire lifetime.

Implications for Occupational Studies

BeBLPT testing is commonly used as a surveillance tool to gauge the risk of sensitization due to various beryllium exposures or industrial hygiene practices. The laboratory variability documented here could have substantial impact on the measured prevalence or incidence of abnormal tests. In the setting of low prevalence of abnormal results (on the order of a few percent), variations in test accuracy over time or across laboratories could overwhelm true differences in abnormal test rates, producing unusual or uninterpretable results. This might especially be the case when comparisons are made based on asynchronous testing across locations. The variability of laboratory performance over time may make it difficult to compare results before and after industrial hygiene intervention. Newer methods that attempt to both account for day-to-day variation in well counts as well as to adjust for (or at least be insensitive to) outliers are available (Frome et al., 1996, 2003). These methods do not rely on a specific fixed cut point but rather measure statistical deviation of test well counts from control well counts. The ability of this method to increase sensitivity, specificity, and positive predictive value has not been demonstrated. Moreover, the extent to which the approach can adjust for trends or excessive variation in laboratory accuracy is not known.

We believe that use of the BeBLPT test could be improved by an enhanced effort in continuous laboratory self-evaluation. Immediate changes that could be made include: (1) repeated testing, within laboratories, of the same individuals (sensitized and unsensitized) over time in small well-designed experiments; (2) plots of mean results over time among all individuals tested; (3) further examination and standardization of methods used to remove outliers; (4) nationwide use of a standardized set of test reagents and sera; and (5) nationwide, blinded, pairwise testing of samples from a small number of blood samples at the major laboratories. Calculation of receiver operating characteristics (ROC) curves (Stokes & Rossman, 1991) on a regular basis might be part of this testing protocol. Small, designed experiments, consisting of the repeated testing of a small number of individuals known to be unexposed, could prove very informative, given the large effects we observed. This may be of particular importance since confirmed abnormal BeLPTs have been found in unexposed control groups and in workers prior to being exposed to beryllium in an occupational setting (Yoshida et al., 1997; Kolanz, 2001; Barna et al., 2003; Stange et al., 2004).

The BeBLPT was first studied for clinical use over 20 years ago (Williams & Williams, 1982). However, publications testing the accuracy of the test over time are not available. In one study, a set of positive BeBLPTs had a high predictive value for CBD (Kreiss et al., 1993); in our previous report, the predictive value was significantly lower (Deubner et al., 2001). Unfortunately, there is no other standard test for beryllium sensitization accepted for use in surveillance in the United States. Rather, laboratories may have to rely on more indirect methods (such as across-time agreement) to ensure the quality of the results they produce.

The data presented in this article suggest that the occupational community should proceed cautiously with generalized use of LPTs for screening, surveillance, or monitoring of exposed workers or community populations. Our analysis studied the interlaboratory performance of a specific LPT, the blood BeBLPT, to evaluate potential reliability and performance problems. The statistical process control methods used to perform this evaluation proved to be both consistent with general clinical impressions and a useful analytical procedure to identify additional time periods not suspected on the basis of clinical impression to be "out of control." We recommend these approaches be incorporated into quality control/quality assurance aspects of medical monitoring programs involving biological assays. Laboratories offering LPTs should provide users with the results of carefully designed experiments that validate their use. Although the goal of protecting worker health may promote the use of more sensitive biological assays in the workplace, it is important to recognize that well-designed studies, which allow for an assessment of test reliability, sensitivity, specificity, and predictive value of a positive result, should precede more widespread application of new technologies.

REFERENCES

Barna, B. P., Culver, D. A., Yen-Lieberman, Y., Dweik, R. A., and Thomassen, M. J. 2003. Clinical application of beryllium lymphocyte proliferation testing. *Clin. Diag. Lab. Immunol.* 10:990–994.

Berger, B. M. 1996. Statistical quality assurance in cytology. Use of the Pathfinder to continuously assess screener process control in real time. *Acta Cytol.* 40:97–106.

Boggs, P. B., Hayati, F., Washburne, W. F., and Wheeler, D. A. 1999. Using statistical process control charts for the continual improvement of asthma care. *J. Commun. J. Qual. Improv.* 25:163–181.

Borak J. 2006. The beryllium occupational exposure limit: Historical origin and current inadequacy. *J. Occup. Environ. Med.* 48(2):109–116.

Carey, R. G., and Lloyd, R. C. 1995. *Measuring quality improvement in healthcare: A guide to statistical process control applications*. New York: Quality Resources.

Cederbrant, K., Hultman, P., Marcusson, J. A., and Tibbling, L. 1997. *In vitro* lymphocyte proliferation as compared to patch test using gold, palladium and nickel. *Int. Arch. Allergy Immunol.* 112:212–217.

Curtis, G. H. 1951. Cutaneous hypersensitivity due to beryllium: A study of thirteen cases. *Arch. Derm. Syphilol.* 64:470–482.

Deubner. D. C., Goodman, M., and Iannuzzi, J. 2001. Variability, predictive value, and uses of the beryllium blood lymphocyte proliferation test (BLPT): Preliminary analysis of the ongoing workforce survey. *Appl. Occup. Environ. Hyg.* 16:521–526.

Doty, L. 1996. *Statistical process control*. New York: Industrial Press.

Frome, E. L., Smith, M. H., Littlefield, L. G., Neubert, R. L., and Colyer, S. P. 1996. Statistical methods for the blood beryllium lymphocyte proliferation test. *Environ. Health Perspect.* 104S:957–968.

Frome, E. L., Newman, L. S., Cragle, D. L., Coyler, S. P., and Wambach, P. F. 2003. Identification of an abnormal beryllium lymphocyte proliferation test. *Toxicology* 183:39–56.

Kaminsky, F. C., Maleyeff, J., and Mullins, D. L. 1998. Using SPC (statistical process control) to analyze measurements in a healthcare organization. *J. Health Risk Manage.* 18:36–46.

Kolanz, M. 2001. Introduction to beryllium: Uses, regulatory history, and disease. *Appl. Occup. Environ. Hyg.*16:559–567.

Kreiss, K., Newman, L. S., Mroz, M. M., and Campbell, P. A. 1989. Screening blood test identifies subclinical beryllium disease. *J. Occup. Med.* 31:603–608.

Kreiss, K., Wasserman, S., Mroz, M. M., and Newman, L. S. 1993. Beryllium disease screening in the ceramics industry. Blood lymphocyte test performance and exposure–disease relations. *J. Occup. Med.* 35:267–274.

Kreiss, K., Mroz, M. M., Newman, L. S., Martyny, J., and Zhen, B. 1996. Machining risk of beryllium disease and sensitization with median exposures below 2 μg/m^3. *Am. J. Ind. Med.* 30:16–25.

Kreiss, K., Mroz, M. M., Zhen, B., Wiedemann, H., and Barna, B. 1997. Risks of beryllium disease related to work processes at a metal, alloy, and oxide production plant. *Occup. Environ. Med.* 54:605–612.

Laine, J., Happonen, R. P., Vainio, O., and Kalimo, K. 1997. *In vitro*lymphocyte proliferation test in the diagnosis of oral mucosal hypersensitivity reactions to dental amalgam. *J. Oral Pathol. Med.* 26:362–366.

Lalla, M., and Virolainen, M. 1974. Blood lymphoblast proliferation *in vivo* in cutaneous drug hypersensitivity reactions. *Int. Arch. Allergy Appl. Immunol.* 46:289–299.

Littell, R. C., Milliken, G. A., and Stroup, W. W. 1996. *SAS system for mixed models.* Cary, NC: SAS Institute, Inc.

McAree, P. W., Bauer, K. W., Jr., Louis, D. J., and Jackson, J. A. 1998. Use of statistical process control for surveillance of pulmonary dysfunction in groups in the workplace. *Health Care Manage. Sci.* 1:53–59.

Montgomery, D. C. 2001. *Introduction to Statistical Quality Control.* New York: John Wiley.

Quesenberry, C. P. 2000. Statistical process control geometric Q-chart for nosocomial infection surveillance. *Am. J. Infect. Control* 28:314–320.

Rasanen, L., and Tuomi, M. L. 1992. Diagnostic value of the lymphocyte proliferation test in nickel contact allergy and provocation in occupational coin dermatitis. *Contact Dermatitis* 27:250–254.

Rasanen, L., Sainio, H., Lehto, M., and Reunala, T. 1991. Lymphocyte proliferation test as a diagnostic aid in chromium contact sensitivity. *Contact Dermatitis* 25:25–29.

Schwab, R. A., DelSorbo, S. M., Cunningham, M. R., Craven, K., and Watson, W. A. 1999. Using statistical process control to demonstrate the effect of operational interventions on quality indicators in the emergency department. *J. Health Qual.* 21:38–41.

Stange, A. W., Furman, F. J., and Hilmas, D. E. 2004. The beryllium lymphocyte proliferation test: Relevant issues in beryllium health surveillance. *Am. J. Ind. Med.* 46:453–462.

Stejskal, V. D., Danersund, A., and Lindvall, A. 1999. Metal-specific lymphocytes: Biomarkers of sensitivity in man. *Neuroendocrinol. Lett.* 20:289–298.

Stokes, R. F., and Rossman, M. D. 1991. Blood cell proliferation response to beryllium: Analysis by receiver-operating characteristics. *J. Occup. Med.* 33:23–28.

Williams, W. R., and Williams, W. J. 1982. Development of beryllium lymphocyte transformation tests in chronic beryllium disease. *Int. Arch. Allergy Appl. Immunol.* 67:175–180.

Yoshida, T., Shima, S., Nagaoka K., Taniwaki, H., Wada, A., Kurita, H., and Morita, K. 1997. A study on the beryllium lymphocyte transformation test and the beryllium levels in the working environment. *Ind. Health* 35:374–379.