Published in final edited form as:

Epidemiology. 2008 July ; 19(4): 621–627. doi:10.1097/EDE.0b013e318176bfb2.

Monitoring Quality Control:

Can We Get Better Data?

Ofer Harel^a, Enrique F. Schisterman^b, Albert Vexler^b, and Marcus D. Ruopp^b

^a*Department of Statistics, University of Connecticut, Storrs, CT*

^b*NICHD/NIH/DHHS, Rockville, MD*

Abstract

Quality control is important in many fields, especially industrial production. Major research has been developed with regard to industrial quality control to ensure reliable and consistent products. We adapt and develop methodology in quality control to monitor data collection in epidemiologic studies. There are no procedures currently used by epidemiologists to evaluate quality control during the actual process of data collection; methods are implemented only after the data have been collected. We focus on procedures that can be used during data collection: instrument calibration and population sampling. For the first, we propose methods utilizing Shewhart control charts and Westgard stopping rules. For evaluating population sampling, we present methods utilizing regression analysis. We provide a motivating example to highlight the utility of these methods. The proposed methodology may help investigators to identify data quality problems that can be corrected while data are still being collected, and also to identify biases in data collection that might be adjusted later.

Data collection for epidemiologic investigations is a dynamic process susceptible to variations in quality. Lack of quality control can be detrimental to analysis and conclusions in unexpected ways. Epidemiologic tools have been developed to evaluate data either a priori (through the use of pilot studies), or post hoc through data cleaning and basic data analysis (eg, comparison of the characteristics of participants vs. nonparticipants, stratified analysis, whisker plots). Little has been written about methods to evaluate quality control during the actual process of data collection. This concept has been applied extensively in industrial production and has led to industrial standards for quality control to ensure a reliable and consistent product.¹ An extension of this concept to epidemiologic research would allow investigators to monitor on-going data collection to assess data quality.

There are 2 main processes to be evaluated. The first is calibration of data collection instruments, including study questionnaires and biomarker measurements. For example, are incomplete responses in the questionnaires missing at random? Are there systematic variations in measurements obtained by different interviewers? The second consideration is whether study participants are an unbiased sample of the target population. Examining these 2 processes during real-time data collection allows investigators to assess the quality of data they are collecting and, if necessary, to adjust low quality data that has already been collected. In this paper, we attempt to define the 2 main processes affecting quality control and to propose quality control methods for use during actual data collection. These methods allow real-time monitoring of data quality, assure that the sample is representative of the target population, and suggest clear stopping rules in case of bias. These methods can be used in most types of epidemiologic studies, including case-control studies, cohort studies, and clinical trials. In the next section we introduce sources of variability that require quality control. We subsequently

elaborate on the background of quality control and data collection, and offer methods to evaluate the effects of instrument calibration and population sampling on quality control. We present an example of these methods, and conclude with recommendations for future epidemiologic quality-control standards.

Sources of Variability That Require Quality Control

Quality control consists of evaluating 2 main processes during data collection. One is instrumentation control, which monitors the instrument used to collect data. The other is population quality control, which monitors the sample to make sure it is representative of the target population.

Instrumentation control can itself be divided into 2 parts. The first part is to ensure that instruments measuring biologic variability (eg, blood pressure measurement devices, scales to weigh patients) are calibrated and give precise and accurate data. The second part is to evaluate how well the standard calibrated instruments, such as questionnaires, are capturing information of interest. It is common epidemiologic practice to evaluate the calibration of these instruments before beginning a study and to analyze responses of participants and nonparticipants after a study; however, little is done to monitor quality control during a study. Similarly, the evaluation of population sampling (eg, the effectiveness of recruitment) is typically done after all subjects have been recruited.

These processes can all be improved with data monitoring during the actual process of data collection. For example, if one is interested in recruiting babies weighing 1500–3000 g (eligibility criteria) and the recruited sample consists mostly of babies 2000–3000 g, the eligibility criteria will be met, but the study sample would be a biased sample of the target population. Based on the selected sample, the estimate of the mean birth weight would be too high and the variance in birth weight would be too low. Monitoring the process would allow the investigators either to change the recruitment strategies or to collect auxiliary information to correct for the bias later.

Quality Control Methods

One way to monitor the processes of variability is a method called statistical process control. This procedure, pioneered by Shewhart^{2,3} and taken up by Deming,^{4,5} uses statistical tools to observe the performance of the production line to predict significant deviations that may result in rejected products. The underlying assumption is that any production process will produce products whose properties vary slightly from their designed values, even when the production line is running normally, and these variances can be analyzed statistically to distinguish random variation from error. Through statistical monitoring, a researcher can discover changes in data collection (type of missingness, instrument variation, interviewer variation) and correct the problem.

Instrumental Quality

The primary statistical tool for such monitoring is the Shewhart control chart (Fig. 1).² The control chart, also known as the “Shewhart chart” or “process-behavior chart,” is designed to assess the nature of variation in a process. A control chart is a sequence of quantitative data with 5 horizontal lines: a center line at the mean; 2 “warning limit” lines at 2 standard deviations above and below the mean; and 2 “control limit” lines at 3 standard deviations above and below the mean.

Natural variations in the process will appear in the data plots as an irregular pattern on the chart, mostly within the control limits. Observations outside the limits (or disrupted patterns

within) can signal problems in the process that are damaging the data quality. The control chart provides a context in which to interpret signals affecting data quality. Limits are established for a given control chart. For example, Shewhart² set 3 standard-deviation limits on the following basis: for any probability distribution, the probability of an outcome greater than k standard deviations from the mean is at most $1/k^2$. When a point falls outside of these limits, they must be investigated.

Even when a process is in good control, there is approximately a 0.27% probability of any given point exceeding 3 standard-deviation control limits (based on normally distributed data). Since the control limits are evaluated each time a point is added to the chart, it follows that every control chart will eventually signal the possible presence of a special cause, even though this has occurred by chance. For a Shewhart control chart with 3 standard-deviation limits, this false alarm occurs on average once every $1/0.0027$ or 370 observations. Therefore, the in-control average run length of a Shewhart chart is 370 observations, during which time one false positive is expected and can be ignored. If a “special cause” (or true deviation) does occur, it may not be of sufficient magnitude for the chart to produce an immediate alarm condition. If a special cause occurs, one can describe that cause by measuring the change in the mean or variance.

Consequently, to evaluate the presence of special causes of variance, stopping rules for an in-control average run length can be established. These are termed the Westgard rules.⁶⁻⁹ For convenience, we adopt short-hand notations to abbreviate different decision criteria or control rules (eg, 1_{2s} to indicate 1 control measurement exceeding 2-standard-deviation control limits). Combinations of rules are indicated by using a “slash” mark (/) between control rules (eg, $1_{3s}/2_{2s}$).

1. 1_{3s} —A run is rejected when a single control measurement exceeds the mean plus 3s, or the mean minus 3s control limit.
2. 2_{2s} —A run is rejected when 2 consecutive control measurements both exceed the mean plus 2s, or both exceed the same mean minus 2s control limit.
3. R_{4s} —A run is rejected when 1 control measurement in a group exceeds the mean plus 2s and another exceeds the mean minus 2s.
4. 4_{1s} —A run is rejected when 4 consecutive control measurements all exceed the mean plus 1s, or all exceed the mean minus 1s control limit.
5. $10_{\bar{x}}$ —A run is rejected when 10 consecutive control measurements fall on one side of the mean.

These rules (or others developed by the investigation) can help to monitor the performance of instruments measuring exposure or outcome.

Population Quality Control

To evaluate quality control in sampling the target population, we propose a two-phase procedure. First, sequentially regress the population variable of interest on a variable that represents the order in which the subjects entered the study. In some cases, the identification (ID) number will reflect this counting process. One would expect that the subjects enrolled into the study will be independent of this variable. If the slope of this simple regression is zero, we can infer that the study recruitment is consistent in its selection from the target population. On the other hand, if the slope is not zero, we can infer that some selection bias is occurring.

To accompany this tool, we propose to sequentially follow the t value of the slope estimate. By applying the previously proposed Westgard rules to these values, we take into consideration the sequential nature of this data and thereby are able to flag for possible distortions on

recruitment. For example, if a study is targeting men and women between the ages of 18 and 40 years, then using this tool the investigator will be able to detect a drift in the age mean of the recruited population. At this point, the researcher can decide whether to continue the course or change the recruitment strategy. If women's average age at enrollment is increasing, the researcher can limit the recruitment of the older women while targeting more of the younger women. As a result, the slope estimates would give us an idea for the direction of the bias and its magnitude. The t values would allow us to test whether it is a significant deviation from the targeted population or a chance event. The flags indicate the need to stop the process and evaluate. If no problems are found, the researcher can continue data collection. Clearly, in this case the monitoring will lead to unnecessary stopping and some possible bias. However, given that most survey data and epidemiologic studies are subject to varying participation rates, nonresponse, instrument variability and other sources of error, this method should be useful in any data-gathering process.

In addition, we are able to test for changing trends in the recruited populations of sequential data by modifying a method developed by Chia-Shang and White.¹⁰ To accomplish this, consider a trend process under the null H_0 where:

$$Y_t = \alpha_0 + \varepsilon_t, \varepsilon_t \sim N(0, \sigma^2), t \geq 1.$$

Using this process, the goal is to sequentially detect a change in the trend function at some unknown point in time resulting in accepting an alternative hypothesis H_1 . The alternative H_1 is that of a single shift at an unknown change point v where:

$$H_1: Y_t = \alpha_0 + \varepsilon_t, t \leq v; Y_t = \alpha_1 + \beta t + \varepsilon_t, t > v,$$

with parameters $\alpha_0, \alpha_1, \beta$ and σ unknown. Following Chia-Shang and White,¹⁰ we are able to define a test statistic where:

$$T_n = \frac{n^{1/2}}{6\hat{\sigma}_n} \max_{k \leq n} k^3 |\hat{\beta}_k|,$$

with the maximum likelihood estimator $\hat{\beta}_k$ having the form:

$$\hat{\beta}_k = \frac{1}{k \sum_{t=1}^k t^2 - \left(\sum_{t=1}^k t \right)^2} \left[k \sum_{t=1}^k t Y_t - \sum_{t=1}^k t \sum_{t=1}^k Y_t \right], \left(k \sum_{t=1}^k t^2 - \left(\sum_{t=1}^k t \right)^2 \right) = (k^4 - k^2)/12$$

and $\hat{\sigma}_n$ being the maximum likelihood estimator of σ . Having adopted this form, the sequential sampling rule is the least smallest subject to $T_N \geq c$ with sample size N and test threshold c . Thus, we have:

$$N(c) = \min \{n: T_n \geq c\}$$

and when this holds, we report that a change in the trend has occurred. This sequential method is powerful because it is based on the likelihood ratio method.¹¹ To determine the test-threshold, c , we have that the α level sequential test is a stopping rule τ under the null with $P\{\tau < \infty\} \leq \alpha$ (the probability of a false alarm resulting in a stopping of sampling). To find c we choose:

$$c_\alpha: P\{N(c_\alpha) < \infty | H_0\} \leq \alpha.$$

Equations for calculating the test threshold are given in the Appendix.

If a flag is raised by the proposed tools for detecting bias in the sampling distribution, the researcher should evaluate whether the recruiting is actually appropriate for the target population. If this assumption is not met, the researcher can alter the recruitment as needed. For example, if the ages of the participants are older than anticipated, the researcher can target younger recruits for the study. Alternatively, the researcher can decide to continue with the recruiting effort without changing the protocol and then to use weights during the analysis. An illustration for this method will be covered in the next section.

Motivating Example

Consider a study that enrolled a population-based sample of randomly selected residents of Erie and Niagara Counties, ranging from 35 to 79 years of age. The New York State Department of Motor Vehicles drivers' license database was used as the sampling frame for adults between the ages of 35 and 65 years while the elderly sample (ages 65–79 years) was randomly selected from the Health Care Financing Administration database. The overall participation rate was 47% ($n = 1171$) of eligible participants.

After the exclusion of participants with a history of chronic diseases ($n = 135$), 899 subjects were selected for these analyses. Participants provided a 12 hour fasting blood specimen for biochemical analysis. (Additional information can be found in a previous paper.¹²) For simplicity, we will focus on the quality of 4 variables: body mass index (BMI), age, high density lipoprotein (HDL), and low density lipoprotein (LDL).

To perform quality control, we use the first 15% of the subjects as training data to establish basic parameters to construct the control chart. One can assume that these parameters are known and then modify them as the study progresses. Beyond the first 15%, each subject who comes into the study is evaluated and will be flagged if there is a violation of the Westgard rules described above.

Instrumental Quality

The 2 variables most at risk for instrumental bias are HDL and LDL levels. (We will not expect instrumental bias for measuring age or BMI.) Although we concentrate on HDL and LDL, plotting the control chart for all variables is highly beneficial.

We calculated the mean and variance for the training data subjects and used these values for the “control limit” and “warning limit” lines. We then plotted subjects joining the study in sequential order on the appropriate control chart. Each figure shows the raw data, the mean line, the “control limit” lines 2 standard deviations above and below the mean, and the “warning limit” lines 3 standard deviations above and below the mean. Figure 1(A) displays the distribution of HDL as subjects enter the study. The solid line represents the mean of the training data, while the dotted lines represent 2 or 3 standard deviation differences from the mean. Notice that in Figure 1(B), which displays the distribution of LDL, there is no “control limit” line below the mean due to the distribution of the data.

The HDL results present no obvious flags. There are some measurements away from the mean but they are distributed evenly in the plot. There are 2 measurements that are located on the “warning limit” line, but none beyond it. There are few observations above and below the 2 standard deviation lines; but again, none of the observations falls into the Westgard rules.

We do see several flags in the LDL results. There are 3 values 3 standard deviations above the mean; and in general, there are more measurements above the mean than below. It would have

been beneficial at that point in the study to make sure the instrument was not biased. This is a stopping and evaluation point, meaning data collection should be stopped and instrumentation verified (reagents, volume, etc.).

The age results are presented in Figure 1(C). One observation is far outside the 3 standard deviation line. This observation is likely to be an error, and probably mis-coded. There are several individuals with values greater than the 2 standard deviation line but only one below it. This merits more detailed investigation.

The BMI results are presented in Figure 1(D). No flags are raised for values under the mean, but several flags are raised above the mean. There are 8 values that are 3 standard deviations above the mean and several more values 2 standard deviations above. Based on these plots, it seems that the population is aging and that the BMI increases over time. Since there is no real instrument quality control when testing the BMI, these flags may be indicators of bad data, outliers, or mis-coding.

Population Quality

To test whether the population of the training data is representative of the population being enrolled, and that the sampling distribution of the data does not depend on time, we explore the sequential regression analysis of the variable in mind (eg, age, BMI) given the identification number or the time the subject enters into the study. As explained above, we expect the slope coefficient to be close to zero. A flag would be raised if the t value exceeds 1.96 or is lower than -1.96.

Figure 2(C) shows the t value of the slope coefficient for the regression of age over time (represented by ID number). After the 150 subjects of training data are enrolled, subjects subsequently enrolled have a different average age. There are differing waves of subjects based on the time-point of collection. There are change points (eg, 150, 300, 350) that represent the change of the population. In Figure 2(A), we present the slope coefficients for the same regression presented in Figure 2(C). In Figure 2(A), we can see that the regression value is initially not zero but converges to zero with time. In general, after the flag was raised at the end of the training data, the researcher will have to decide whether to continue collecting data.

To correct for selection bias after data are collected, one can use sample weights. In Figure 3, we can see the sequential slope estimate with the complete data, and 2 additional plots that represent additional weights for younger adults. The dashed curve represents 0.5 weights for each older adult, while the double-dashed curve represents 0.25 weights. It is obvious that by using the weights and the fact that we have a large data set there is a convergence to the true population. Another possibility is to eliminate part of the sample based on the weight characteristics.

There are 2 main tactics for the problem of an aging population. First, we can add restrictions to the data collection protocol. Second, we can complete the data collection and correct the bias at the end. It is important to consider the advantages and disadvantages of each tactic. In our example, there is a positive correlation between age and the time entering into the study. A relevant question is whether this is due to selection bias or due to the length of the period of data collection and the aging population. The researcher needs to evaluate these reasons and determine the appropriate action.

Considering similar plots for the BMI variable (Fig. 2(D) for the t values and Fig. 2(B) for the actual slopes), we can see that there is a risk for this variable as well. Unlike the age variable, a flag is raised approximately 80 subjects after the training data. Due to the correlation between

age and BMI, the age selection might be the cause of the BMI. If corrections are needed, the same procedure as described for the age variable should be used.

DISCUSSION

A continuous process for monitoring data collection can improve data quality. Otherwise, adjustments for low-quality data are confined solely to post hoc analysis. We present tools that are easy to implement for monitoring instrumental and population quality control, and that have a long history of applicability. Software is readily available to implement the Westgard rules, and the population quality control procedures are easily programmable. Code can be requested from the authors.

The methods developed in this paper for instrumental and population quality are adaptable to different distributional assumptions such as log-normal. These methods can also be used to evaluate multiple processes simultaneously, allowing multivariate control charts to be constructed. In epidemiologic research, as in other types of research, findings are only as strong as the data collected. Implementing rigorous and well-defined quality control procedures may help to ensure sound results.

APPENDIX

Calculating the test threshold c

To specifically calculate the test threshold c , we have:

$$\begin{aligned} P\{N(c_\alpha) < \infty | H_0\} &= \lim_{N \leftarrow \infty} P\{\max_{n \leq N} T_n \geq c_\alpha | H_0\} \\ &= \lim_{N \leftarrow \infty} P\left\{\frac{N^{1/2}}{6\sigma} \max_{k \leq N} k^3 |\hat{\beta}_k| \geq c_\alpha | H_0\right\} \\ &= 1 - \sum_{j=\infty}^{-\infty} (-1)^j P\left\{(2j-1) \sqrt{3}c_\alpha < \xi < (2j+1) \sqrt{3}c_\alpha\right\}; \xi \sim N(0,1) \end{aligned}$$

(For details, see pp. 290–291 of Chia-Shang and White.¹⁰)

For a fixed α (eg, 0.05), the test threshold can be calculated as a root of the equation.

$$1 - \sum_{j=T}^{-T} (-1)^j P\left\{(2j-1) \sqrt{3}c_\alpha < \xi < (2j+1) \sqrt{3}c_\alpha\right\} = \alpha, \\ T \gg 1,$$

or $2[1 - \Phi(\sqrt{3}c_\alpha)] = \alpha$, where Φ is the cumulative distribution function of a normal random variable.

References

1. Cembrowski, GS.; Carey, RN. Quality control procedures. In: Cembrowski, GS.; Carey, RN., editors. Laboratory Quality Management. Chicago: ASCP Press; 1989. p. 59-79.
2. Shewhart, WA. Economic Control of Quality of Manufactured Product. New York: Van Nostrand; 1931.
3. Shewhart, WA. Statistical Method from the Viewpoint of Quality Control. New York: Dover; 1986.
4. Deming WE. On probability as a basis for action. Am Stat 1976;29:146–152.
5. Deming, WE. Out of the Crisis: Quality, Productivity and Competitive Position. 2. Cambridge: Cambridge University Press; 1982.

6. Westgard JO, Barry PL, Hunt MR, et al. A multi-rule Shewhart chart for quality control in clinical chemistry. *Clin Chem* 1981;27:493–501. [PubMed: 7471403]
7. Westgard, JO.; Barry, PL. Improving quality control by use of multirule control procedures. In: Westgard, JO.; Barry, PL., editors. *Cost-Effective Quality Control: Managing the Quality and Productivity of Analytical Processes*. Washington, DC: AACC Press; 1986. p. 92-117.
8. Westgard, JO.; Klee, GG. Quality management. In: Westgard, JO.; Klee, GG., editors. *Fundamentals of Clinical Chemistry*. 4. Philadelphia: WB Saunders Company; 1996. p. 211-223.
9. Westgard, JO.; Klee, GG. Quality management. In: Westgard, JO.; Klee, GG., editors. *Textbook of Clinical Chemistry*. 2. Philadelphia: WB Saunders Company; 1994. p. 548-592.
10. Chia-Shang JC, White H. A direct test for changing trend. *J Business Econ Stat* 1992;10:289–299.
11. Lai TL. Sequential changepoint detection in quality control and dynamical systems. *J Roy Stat Soc* 1995;57:613–658.
12. Schisterman EF, Faraggi D, Browne R. TBARS and cardiovascular disease in a population-based sample. *J Cardiovasc Risk* 2001;8:219–225. [PubMed: 11551000]

FIGURE 1. Control charts of high density lipoprotein (HDL), low density lipoprotein (LDL), age, and body mass index (BMI) for our motivating example.

FIGURE 2.
Slope regression coefficients and t values for the slope regression of age and BMI for motivating example.

FIGURE 3.
Slope estimates of age for complete data and weighted data.