

## Six Sigma<sup>®</sup> can be dangerous to your health

Jan S. Krouwer

Received: 29 April 2008 / Accepted: 4 August 2008  
© Springer-Verlag 2008

**Abstract** Six Sigma<sup>®</sup> is a popular quality program that has been applied to clinical laboratory assays. Sigma is calculated as  $(TEa - \text{bias})/CV$  and these sigma numbers are often claimed as a sole measure of quality based on medical requirements. But these sigma calculations do not account for all results. An additional set of wider limits must be added, such that all data are accounted for. This gives a minimum of three zones: zone A, where 95% of the data should be; zone C, where there should be no data; and zone B, where 5% of the data is allowed. An additional problem is that sigma calculations are often based on valid analytical data, meaning that pre-analytical and post-analytical errors are excluded. Also, samples that are flagged by the system and produce no results are, of course, excluded, but delayed results can cause patient harm. A better measure of assay quality can be provided by a failure mode effects analysis (FMEA), which attempts to assess the probability of failure and its severity for every possible failure mode. In this paper, an example of what is entailed is described for two failure modes and the overall process is outlined. The amount of effort required for a full FMEA is beyond virtually any clinical laboratory. Some compromises are suggested. Calculating sigma values, which have little meaning about patient harm, is not recommended.

**Keywords** Six Sigma · FMEA · Error grid · Total error · Clinical laboratory assay

---

Papers published in this section do not necessarily reflect the opinion of the Editors, the Editorial Board and the Publisher.

---

J. S. Krouwer (✉)  
Krouwer Consulting, 26 Parks Drive, Sherborn, MA 01770, USA  
e-mail: jan.krouwer@comcast.net

At a recent conference [1], there were several presentations about Six Sigma for clinical laboratory assays. To recall [2, 3], sigma is calculated as  $\text{sigma} = (TEa - \text{bias})/CV$  where:

TEa is the total allowable error

Bias is the inaccuracy of the measurement procedure

CV is the imprecision of the measurement procedure

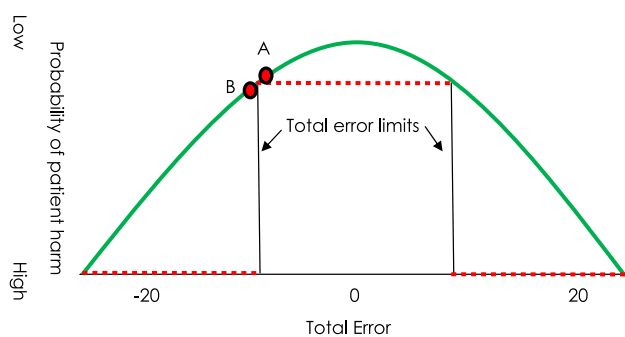
The problem with Six Sigma is that it is often presented as a *sole measure of quality*—that is, a high sigma value ( $>6$ ) is said to guarantee high quality for an assay. This article explains issues with this claim and suggests alternatives.

The first problem with Six Sigma is that the TEa limits are often equated with *medically acceptable* limits. For example, the ISO 15197 standard for glucose [4] states that the limits are:

“... the minimum acceptable accuracy criteria are based on the medical requirements for glucose monitoring.”

The implied meaning of medically acceptable limits is shown in Fig. 1 (dashed line), whereby all values within the limits are totally acceptable (have a low probability of causing patient harm) and all values outside of the limits are totally unacceptable (have a high probability of causing patient harm).

The solid line, based on Taguchi [5], is a more realistic model, which shows that the probability of causing patient harm is lowest for zero error and, as one observes error, there is a quadratic increase in the probability of causing patient harm. Thus, data points A and B, which fall on different sides of the total error limits, are close in the amount of error and are almost equally likely to either cause or not cause problems. It is only when one retreats

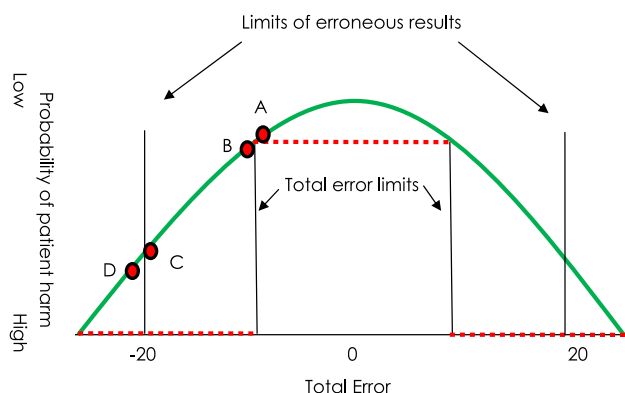


**Fig. 1** Total allowable error limits for a clinical assay. The *dashed line* is implied by some authors who calculate sigma, the *solid line* is more realistic

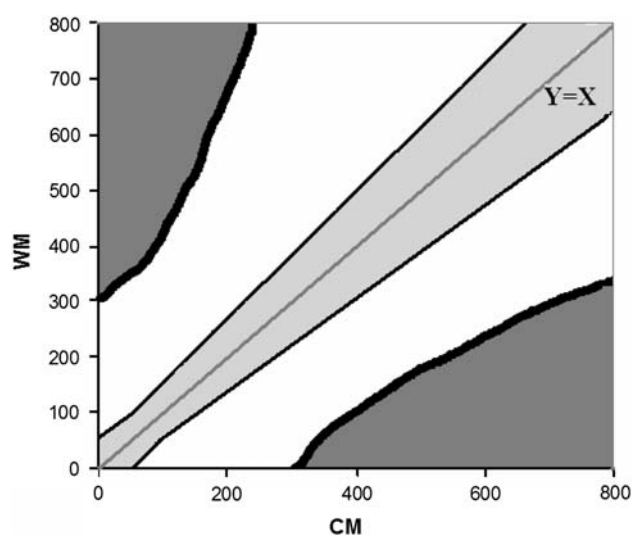
farther away from the limits in Fig. 1 that one is almost certain to have results that can cause harm. This is shown with points C and D in Fig. 2, which are at these limits, called the limits of erroneous results (LER). Note that there is always a “gray zone.” Thus, points B and C are in this gray zone, although point C is much more likely to cause patient harm than point B.

Figure 3, taken from FDA guidance regarding CLIA waiver methods [6], is a way of representing an error grid and also incorporates the concepts in Fig. 2. In the error grid, the traditional TEa limits are shown as the boundary between the light gray and white areas. This boundary is often called the zone A region (gray), where 95% (or in some cases 99%) of the results should be. Zone B (white) can contain up to 5% of the results and zone C (dark gray) should contain no results. The error grid contains more information than the TEa alone, since the two sets of limits in an error grid can be different for each concentration. Error grids have been used for glucose assays [7, 8], but are otherwise uncommon.

Thus, sigma only accounts for zone A, but patients are harmed by values in zone C! Now, one might argue that



**Fig. 2** Total allowable error limits for a clinical assay. The new lines, limits of erroneous results (LER), where point D is just outside of, are limits that, when exceeded, are likely to cause medical harm



**Fig. 3** An error grid for a clinical assay. The error grid divides assay results into at least three zones. Most results are desired to fall in the innermost zone and, ideally, no results should be in the outermost zone. WM (waiver method) is the test method and CM (comparative method) is the reference method

there is, nevertheless, a relationship between sigma and the three zones, meaning that assays with high sigma values are unlikely to have values in zone C. This is also not necessarily true for the following reasons:

1. Often, incorrect models are used to assess the total error [9].
2. In estimating bias and CV, outliers—the very values that cause harm—are often discarded.
3. Sigma calculations are based on the assumption that the data are normally distributed. Most data do not fulfill this criterion. There are often more values in the tails of the distribution (zone C) than that expected by calculations based on the normal distribution.
4. And, perhaps most important of all, values can occur in zone C that have nothing to do with the analytical process. If a patient sample mix-up is detected, these values are excluded from virtually all analytical evaluations, as one argues that one is evaluating the analytical properties of the assay.

For an example to illustrate these points, consider two assays for which sigma has been calculated. Assay A has 95% of observations within the TEa limits. This provides a sigma of about 3.1, which would not be considered as a good assay, as it is much lower than a sigma of 6. But if the remaining 5% of the results were close to the TEa limits with no results in the LER zone, then it is unlikely that this assay would cause patient harm. Assay B also has 95% of results within the TEa limits, but 0.5% of the results are in the LER zone. This is equivalent to 5000 dangerous results per million—a much worse condition than assay A, even

though both assays have the same sigma. Even for assays with sigma greater than 6, one must consider the possibility of shifted results, since, by definition, sigma includes a 1.5-sigma shift. Klee has shown how small shifts in the mean can have clinical consequences [10]. The bottom line is that calculating the percentages of observations in each zone in an error grid (such as that of Fig. 3), as well as plotting the data in the error grid, is more informative than calculating sigma.

Six sigma is used in other industries to measure the number of defects. Perhaps this is one of the problems. The distance from the target required for a defect is not as easy to establish for a diagnostic assay. Given these problems with six sigma, some alternative methods to estimate the quality of an assay are suggested, using hCG (human chorionic gonadotropin) as an example assay.

First, when the total analytical error is calculated to estimate the percentage of values in zones A–C in an error grid, one should use conservative methods, such as the empirical distributions suggested by the CLSI EP21A method [11], and where no data are deleted. Assume that a clinical laboratory has carried out this method comparison evaluation with 40 patient samples for a new method and a reference method and found no results in zone C for an hCG assay. What can one conclude? Although there are 0% of the values in zone C, the 95% confidence interval extends to 7.2%. This means that, for every million hCG results performed, up to 72000 results could be in zone C. Thus, these types of evaluations do not prove much, although one suspects that the 7.2% rate is unlikely (because if this rate occurred, it would be noticed).

Failure mode effects analysis (FMEA) is an approach that will provide a more complete answer to the quality question, but in its complete form, FMEA requires considerable effort. To complete an FMEA analysis, one has to start by postulating all possible reasons as to why a result could fall into zone C or a result that could be delayed (if delay is a severe event). To get an idea of what is involved, take two possible failure modes, human antimouse antibodies (HAMA) interference and a patient sample mix-up.

**HAMA interference** To estimate the likelihood of a zone C result from HAMA interference, one needs to know the level of HAMA that will cause erroneous results in the assay and the probability of such levels in the population being sampled. Contacting the manufacturer might yield the level of HAMA that is problematic for the assay. I am unaware of data that informs about the distribution of HAMA in patient samples. Yet, one knows that HAMA interference occurs and causes patient harm [12].

**Patient sample mix-up** There are some data available for patient sample mix-ups [13]. However, it seems that

these cases are caught within the laboratory. One would need to determine how many cases are actually not caught within the laboratory. One could then model the likelihood of a zone C result by sampling from the empirical distribution of hCG results that are observed in the laboratory in order to see the likelihood of a mix-up causing a zone C result.

Because there are so many existing data in a clinical laboratory, one may also have the opportunity to use existing data to help with this modeling.

One must then continue with the FMEA:

- With each other possible failure mode, calculate the probability of zone C results
- Calculate the overall probability of zone C results (from all failure modes) and determine if that risk is acceptable
  - Special algorithms are typically used to perform these calculations to combine probabilities [14]
- Construct a Pareto table, which ranks for all potential errors, the combination of severity, and the likelihood of error
- If the overall probability of zone C results is too high, propose control measures to lower the overall risk to an acceptable level for items at the top of the Pareto table:
  - The control measures must, of course, be affordable

At this point, one can see the idea that this level of effort is out of reach for clinical laboratories, since the level of expertise and work needed to just estimate the likelihood of a zone C result is huge. Even if a clinical laboratory could perform this task, it makes no sense to require every clinical laboratory to do so.

One possibility is to have a standards group tackle such a task, although this too has limitations, as was shown for a (universal) control measure to prevent wrong site surgery [15].

Whereas a complete FMEA may be out of reach of most clinical laboratories, every clinical laboratory could perform the FMEA by qualitatively (instead of quantitatively) assessing probabilities such as likelihood on a scale of 1–4, where the numbers correspond to the frequency of occurrence, as suggested by the Veterans Administration [16]. There is a great deal of actual failure data to help with these qualitative assessments. The steps in this type of FMEA have been reviewed [17].

There are no easy answers to preventing severe, low-frequency failures that cause patient harm, but estimating “sigma” for an assay is also not the answer. And nor is doing nothing.

## References

- 13th Conference on Quality in Medical Laboratories, Antwerp, Belgium, March 2008. Home page at: <http://www.qualityspotlight.com/>
- Westgard JO, Klee GG (2006) Quality management. In: Burtis CA, Ashwood ER, Bruns DE (eds) Tietz textbook of clinical chemistry and molecular diagnostics. Elsevier Saunders, St. Louis, MO
- Westgard S (2005) From method validation to Six Sigma metrics: evaluating a new instrument. Available online at: <http://www.westgard.com/qcapp32.htm>
- International Organization for Standardization (ISO) (2003) ISO 15197: in vitro diagnostic test systems—requirements for blood-glucose monitoring systems for self-testing in managing diabetes mellitus. Available online at: [http://www.iso.org/iso/iso\\_catalogue/catalogue\\_tc/catalogue\\_detail.htm?csnumber=26309](http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=26309)
- Taguchi methods. Wikipedia entry at: [http://en.wikipedia.org/wiki/Taguchi\\_methods](http://en.wikipedia.org/wiki/Taguchi_methods)
- Center for Devices and Radiological Health (CDRH), US Food and Drug Administration (FDA) (2008) Guidance for industry and FDA staff: recommendations for clinical laboratory improvement amendments of 1988 (CLIA) waiver applications for manufacturers of in vitro diagnostic devices. Available online at: <http://www.fda.gov/cdrh/oivd/guidance/1171.pdf>
- Clarke WL, Cox D, Gonder-Frederick LA, Carter W, Pohl SL (1987) Evaluating clinical accuracy of systems for self-monitoring of blood glucose. *Diabetes Care* 10:622–628
- Parkes JL, Slatin SL, Pardo S, Ginsberg BH (2000) A new consensus error grid to evaluate the clinical significance of inaccuracies in the measurement of blood glucose. *Diabetes Care* 23:1143–1148
- Krouwer JS (2002) Setting performance goals and evaluating total analytical error for diagnostic assays. *Clin Chem* 48:919–927
- Klee GG (1995) Analytic performance goals based on direct effect of analytic bias on medical classification decisions. *CDC, 1995 Institute: Frontiers in Laboratory Practice Research*, pp 219–226. Available online at: <http://www.phppo.cdc.gov/dls/pdf/institute/klee.pdf>
- Clinical and Laboratory Standards Institute (CLSI), formerly National Committee for Clinical Laboratory Standards (NCCLS) (2003) Estimation of total analytical error for clinical laboratory methods; approved guideline NCCLS EP21A. CLSI, Wayne, PA. Available online at: <http://www.clsi.org/source/orders/free/ep21-a.pdf>
- Butler SA, Cole LA (2001) Use of heterophilic antibody blocking agent (HBT) in reducing false-positive hCG results. *Clin Chem* 47:1332–1333
- Wagar EA, Tamashiro L, Yasin B, Hilborne L, Bruckner DA (2006) Patient safety in the clinical laboratory: a longitudinal analysis of specimen identification errors. *Arch Pathol Lab Med* 130:1662–1668
- Vesely W (2002) Fault tree handbook with aerospace applications. Available online at: <http://www.hq.nasa.gov/office/codeq/doctree/fthb.pdf>
- Kwaan MR, Studdert DM, Zinner MJ, Gawande AA (2006) Incidence, patterns, and prevention of wrong-site surgery. *Arch Surg* 141:353–357
- US Department of Veterans Affairs. Healthcare Failure Mode and Effect Analysis. Available online at: <http://www.patientsafety.gov/SafetyTopics.html#HFMEA>
- Krouwer JS (2004) An improved failure mode effects analysis for hospitals. *Arch Pathol Lab Med* 128:663–667. Available online at: [http://arpa.allenpress.com/arpaonline/?request=get-document&doi=10.1043%2F1543-2165\(2004\)128%3C663:AIFMEA%3E2.0.CO%3B2](http://arpa.allenpress.com/arpaonline/?request=get-document&doi=10.1043%2F1543-2165(2004)128%3C663:AIFMEA%3E2.0.CO%3B2)