Paper 194-2023

Quality Control - Defining an Acceptable Quality Standard without Achieving Perfection

William Coar

Axio, a Cytel Company

ABSTRACT

In statistical programming, we spend vast amounts of time trying to have a zero-error rate. But time and time again we see that errors still occur even with the gold standard of independent programming. Risk based approaches to quality control have been proposed to emphasize focusing on what matters most. It is human to make mistakes even though there is an inherent drive to achieve perfection. We often spend time trying to be perfect even with the understanding that some issues may be minor and/or inconsequential. Mistakes will continue to happen, but that does not necessarily imply poor quality.

In Juran's Quality Handbook, quality is defined as "fitness for purpose" [1]. The author suggests that to be fit for purpose, products or services must meet two criteria: (1) the product/service must have the right features to satisfy their needs, and (2) they must be free from failure. While quality control has been a topic of many programming presentations, focus tends to be on (2) without any mention of (1).

The purpose of this presentation is to further expand on the concept of "fitness for purpose" and how it can be applied in our day-to-day environment as statisticians and programmers in the pharmaceutical industry. The end goal should be to develop acceptable quality standards to ensure a high quality product or service recognizing that perfection is not achievable. To date, objective criteria have not yet been identified. Our hope is to begin discussion with other industry leaders recognizing that perfection is not required to have a product or service that allows our customers to use our products or services with confidence.

BACKGROUND

The following discussion is a result of many years of day-to-day work analyzing and presenting interim data in support of Data Monitoring Committees (DMC). DMCs are a small group of expert clinicians and statisticians that are charged with assessing risk/benefit of ongoing clinical trials. DMC members are independent of the companies developing the drug and conducting the trial. They are not secret committees as portrayed by the media during the COVID pandemic. They are not required in all clinical trials. The need for a DMC depends on the population of patients, the disease and/or severity, and ethics.

Critical decisions are made from recommendations of DMCs with respect to safety of clinical trial participants. Such decisions are made based on interim, incomplete, sometimes erroneous data analyzed by a Statistical Data Analysis Center (SDAC). In this setting, nothing is perfect. Even though (extreme) efforts are sometimes made to match 100% in a QC process, there is no guarantee of perfection.

It is well understood that the interim nature of the data is considered during decision making. DMC members are aware that the data are imperfect. We strive for excellence to ensure that data, rules, and results are well investigated for critical endpoints used by the DMC in making recommendations. This involves analyzing data and providing a ***fit-for-purpose*** report to the DMC based on interim data. This often differs from a final study report that is typically based on a clean/locked set of data.

FIT FOR PURPOSE

With the help of the worldwide web, the following bullets are all definitions of fit-for-purpose:

- Capable of meeting its objectives or service level

- Good enough to do the job it was designed to do
- Something that does what it is meant to do
- Services provided by qualified and trained personnel and that they are provided while respecting standards generally accepted within the industry.

All help paint a consistent picture of what is implied by fit-for-purpose, and are also consistent with Jurgon. To be fit for purpose, products or services must meet two criteria: (1) the product/service must have the right features to satisfy their needs, and (2) they must be free from failure.

FEATURES TO SATISFY NEEDS

Having the right features to satisfy the needs of the customers is typically defined by a scope of work, or contract, signed between both parties. To be usable, the product or service must address pre-specified needs by the end user. If this is not clearly specified, the customer will ultimately be dissatisfied with the feeling that the product or service was of poor quality.

Complications arise when there are multiple end-users each with different needs, or a product or service from one use is repurposed for a different user (with different needs). This is common in the contract research organization. Programming of TLFs for a study report are often used for data review and interim monitoring of data quality during the conduct of the study. However, the specifications associated with these outputs may not be appropriate for analysis and decision making associated with interim data [4]. As we will demonstrate in our example, the programming may not be suitable for use by a DMC.

FREE FROM FAILURE

How do we define a “failure”? If our tables, listings, and figures generally describe the current data and correct decisions are being made based on an imperfect report, is it a failure? We argue that our products and services can still maintain excellence in quality provided the level of accuracy allows for correct decisions with a low risk of error.

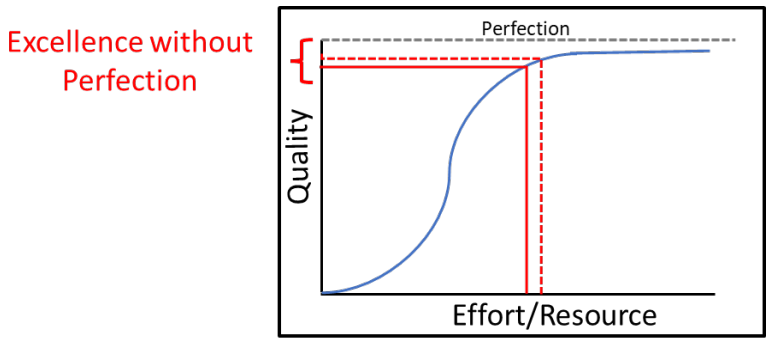
EXCELLENCE VERSUS PERFECTION

In his podcast titled “Breaking up with Perfectionism” (Worklife, May 2022) Adam Grant suggests that perfectionism is an impossible goal. Everyone of us likely has received gentle reminders of this throughout our careers. Grant states that perfectionism can be unhealthy and that humans are known to spend vast amounts of time to be perfect. At times, we continue to do so knowing it will not make a difference. Does this sound familiar? Far too many of us have spent hours programming to accommodate unexpected data or erroneous data trying to be perfect knowing there will be little or even no impact on the end result.

Whether perfectionism is part of culture, or how we were brought up, or engrained through social media, we should acknowledge that it exists and accept that we can still demonstrate excellence without being perfect.

QUALITY

It is well known that an increase in quality requires an increase in effort/resource/cost. However, at some point an inverse relationship is observed where large increases in effort result in a smaller and small increments in quality.



Even with infinite resources, perfection is not attainable. Even with the best processes and procedures, data are imperfect, and humans are imperfect. Errors will be missed, and possibly be even undetectable. If we acknowledge that perfection is not attainable, then where is the cut point where we can maintain excellence in quality with fixed resources?

We typically rely heavily on risk-based quality control procedures to ensure our product or service is accurate with a low but acceptable error rate. With risk-based approaches, more emphasis is placed on features that matter the most and provides for a balance between cost, time, and quality as discussed by Randall et. al [3]. This addresses **freedom from failure** part of fit-for-purpose.

Defining a zone of excellence without perfection should go beyond simply assessing a low error rate. We believe that usability, or maintaining **features that satisfy needs** of the end user, should also be considered. When faced with decisions regarding allocation of additional resources and/or delays, usability should never be ignored. When both are used to together, a fit-for-purpose product or service can be maintained with fixed resources, a constraint that applies to all of us.

Excellence, which is defined by the customer, can be achieved when we repeatedly provide a product or service using agreed upon quality standards that are defined to ensure both a low error yet highly usable product or service.

Concepts of low error rate and highly usable product or service are further discussed below.

QUALITY CONTROL

We define quality control as the collection of efforts made to ensure and document that programs produce expected results. This primarily focuses on (2) in Juran's definition implying that expected results are free from failure, or correct. However, we argue that the usability of the results should be considered during the QC process.

Risk based approaches to quality control in programming have been investigated [3] and are widely used in practice today. These approaches place more emphasis on what matters most (the more critical areas used for decision making) which is consistent with risk-based monitoring in data collection. This suggests that some errors might occur, but the consequences are acceptable.

While many standard operating procedures or work instructions may not formally state the approaches are risk based, operationally they are risk based any time a programmer is allowed to select *"an appropriate QC method*".

Best practices as defined by the PHUSE Best Practices for Quality Control states that customer satisfaction is a component in defining Quality Control. [2]. Yet, this is rarely ever discussed in a Standard Operating Procedure or Work Instruction or Guideline. The reference also states, *"Bugs and defects in the programs used to analyze clinical trial data can result in incorrect analyses which can have damaging consequences to the public health and trust."* We fully endorse this but suggest that not all

bugs or defects are equal, and not all will result in damaging consequences. We believe these need to be considered when defining a quality standard.

Whether your work environment requires 100% independent programming for everything, or it takes advantage of risk-based approaches, mistakes will continue to happen. Furthermore, not all errors will be detected. But does that suggest that the overall quality of the product or service is poor? We believe this is not the case as long as the end result (ie, analysis data and TLFs) contains the right features to satisfy the needs of the end user with a low risk of errors in decision making based on the results.

FAILURE

Suppose we define a "failure" as a case where an incorrect decision is being base. We can focus of the concept of accuracy with respect to programs generating the expected results where expected results are fully dependent on the rules, or specifications.

ACCURACY

We tie accuracy to the concept of usability. There are cases such as presentation of inferential statistics mandate an extremely high level of accuracy, but this is not the case in general.

We design experiments to estimate true parameters of a statistical model used to ***estimate*** a treatment benefit. As George Box quoted, "all models are wrong but some models are useful". In every clinical trial, data have variation due sampling distribution in addition to measurement error. If repeat the experiment, we are likely to observe different but consistent results, provide the sample size and study conduct are appropriate. Results could exactly represent the data, but does that mean the analyses are exact or guarantee appropriate decisions are made?

Statistics is not an exact science even if we believe we are perfectly accurate. Is it ever appropriate to indicate that the results reasonably reflect reality when we are extrapolating beyond the experiment?

PRECISION

Suppose we were to ask you your age? You are likely to respond with an integer approximation, and the approximation may be faulty simply because we wait until the entire year is complete before increasing our age. Similarly, if we were to ask conference attendees how far they travel to San Diego, we would likely get approximations.

Whether or not the approximation (or precision) is good enough depends on what decisions are to be made when interpreting the results.

Even in the presence of these measurement errors, adequate sample sizes ensure that the results are still interpretable. Matching 100% during a QC process does not make the results more 'exact'.

SPECIFICATIONS

Even though there may be a multitude of ways to program a given rule, many of which are equivalent once all data are cleaned and collected, different programming logic may result in different results especially with interim data. Even though they differ, they may all be viewed as "correct" given the logic in place.

Conder an example where we are asked to determine the number of subjects that received at least one dose of study drug. Rules generally suggest looking for evidence that a dose was taken using exposure data. Consider the following four programming statements:

- If EXYN=Y
- If EXDOSE>0
- If missing(EXSTDT)=0
- If length(EXSTDT)=10

At the conclusion of a study when the database is clean and locked, all four statements would likely yield identical results. However, in the interim, they may not. While this may be true, all four approaches provide reasonable estimates of the number of patients that received at least one dose. None are perfect but all are reasonable.

Some rules may be more appropriate depending on the nature of the interim data review. For example, without a date of first dose other data may subsequently be excluded from analysis.

When we think of accuracy, we need to fully understand the rules leading to programming logic. We should understand that rules (which are used to assess accuracy) are very dependent on the usability and features needed by the end user. A final study report has a different end user than an interim safety report provided to a DMC. They both server different purposes. Considerations associated with the impact of analysis using interim data was addressed by Coar [2022]. Different programming logic can lead to different results. When this happens, we view it as a QC process that is working. Reconciling the rules with the observed data can lead to updates to rules to ensure appropriate ***usability*** of the results.

APPLICATION OF FIT-FOR-PURPOSE & DMCS

As stated earlier, DMCs review interim data and make critical decisions to ensure patient safety while maintaining trial integrity. To do this, the SDAC generates unblinded analysis datasets and tables, listings, and figures that the DMC uses to make recommendations regarding the conduct of the study.

FIT-FOR-PURPOSE AT INITIAL DEVELOPMENT

A fit-for-purpose DMC report starts with the initial selection of data to be presented to the DMC. The table of contents are often defined in a DMC Charter, though most lack the specificity required for fit-for-purpose. Most DMC reports will contain information about:

- Disposition
- Demographics & Baseline Characteristics
- Disease History
- Exposure
- Adverse Events
- Laboratory Data
- ECGs
- Vital Signs
- Efficacy

At a high level, the contents are straightforward. However, how should the data be presented to allow the DMC members to easily review interim data and make decisions? Consider the following approaches that are often proposed because they are consistent with a final study report or a function of vendor reporting tools. We provide our assessment on whether or not each leads to a fit-for-purpose DMC report.

The end user in this setting is not the sponsor. Though they will indeed do some of their own QC of the outputs delivered to the DMC, the reports themselves are designed for the DMC. The DMC is thus the end user. More often than not, sponsor QC focuses on their interpretation of the accuracy of the output based on their understanding of the rules. This is very reasonable, but we argue that the ***usability*** by the DMC is rarely considered.

	Fit-for-Purpose	Not Fit-for-Purpose
DMC report in excess of 1000 pages		X
Inclusion of every possible combination of AE tables (exacerbated when there are multiple drugs in treatment regimen) • Related AEs resulting in treatment discontinuation • Related AEs resulting in dose delays • Related AEs resulting in dose reductions • Related SAEs • Related AE of Special Interest		X
Mean +/- SD or box plots over time	X	
Continuous statistics and change from baseline over time for all lab parameters		X
Lab shift tables by visit		X
Lab shift from baseline to worst post-baseline	X	
Potentially clinically significant (PCS) values for lab parameters of clinical interest	X	
Worst post-baseline grade for parameters that worsen on-treatment	X	
Medical history		X
Inclusion/Exclusion violations		X
Major protocol deviations	X	
Prior and Concomitant meds		X
Concomitant medications of special interest • Vaccinations • Rescue therapy	X	
Continuous statistics of vital sign measures by position over time		X
Potentially clinically significant (PCS) vital signs such as: • Change in SBP>20 from sitting to standing • Change in DBP>10 from sitting to standing • Large changes from baseline	X	
Continuous statistics of ECG measures over time		X
Potentially clinically significant (PCS) ECG	X	
Numbers of patients with dose delays or reductions, by reason	X	
Continuous statistics of total daily dose, average daily dose, dose intensity		X
Listings of all subject data		?
Listings of SAEs	X	
Listings of Deaths	X	
Listing of Subjects with PCS laboratory values	X	

In general, we try to provide a focused report that enables the DMC to assess critical safety data. Remember that the end-user are DMC members that are tasked with reviewing the data in a short amount of time, generally a few hours. Reviewing pages and pages of continuous statistics and changes from baseline, or 100s or 1000s of pages of listings can hinder the DMCs review rather then help. Such reports may be required for the final summary of the clinical trial, but they are not fit-for-purpose when we think if DMC reports.

This emphasizes the importance of (1) in Juran's definition of fit-for-purpose where the product or service must have the right features for the DMC to perform their review of the data.

FIT-FOR-PURPOSE DURING QC AND REVIEW PROCESS

Consideration of fit-for-purpose during the QC process is a topic we feel needs to be further discussed with industry leaders. This revolves making decisions with respect to usability of the report by the DMC while considering risk of incorrect decision making.

Assessing usability of the DMC report and likelihood of impacting a decision (ie, recommendation) should be part of a quality standard. However, this needs to be done without the perception that an imperfect report is of poor quality.

Let's revisit the concept of accuracy with respect to programming logic which was previously discussed. Throughout the course of a study, specifications for analysis datasets almost surely will need to be updated. Even with the best defensive programming, the unexpected will always occur with accumulating data forcing updates to programming rules. Furthermore, erroneous data values are impossible to predict. Often, statisticians and programmers will spend hours trying to modify rules to program around dirty data for small number of records. Does doing so improve the quality of the report?

The answer depends on the endpoint and the number of records. We suggest the programmers and statisticians collaborate to assess the impact on the usability of the report if the algorithms remained as is with the understanding that erroneous data will be fixed in subsequent data transfers. If the discrepancies identified do no impact the interpretability of the report and are well documented, we believe that quality is maintained and there may be little or no added value to updates to accommodate dirty data.

Again, not all 'bugs' are equal. Sometimes it may be necessary to consider updating specifications and programming but quality standards that include usability and assessment of risk should provide guidance.

Internal Review

Programmers and statisticians at the SDAC undoubtedly spend time reviewing data and TLFs with the DMC review in mind. Remember, the DMC is in fact the end-user and needs appropriate information to make informed decisions. The review will often reference prior reports and focus on areas identified by the DMC. Issues identified are typically addressed internally. These reviews are more often focused on a **fit-for-purpose** report for the DMC.

Sponsor Review

It is common for sponsors to request blinded versions of a DMC report for them to perform their own QC. We find this tends to be sponsor programmers and statisticians using study report programming to perform this. In doing so, discrepancies are often identified leading to a perception of poor quality. Common issues are:

- Sort of categorical values is incorrect
- Force rows with "missing" category
- Force rows of 0's
- Identification of unexpected data values and programming to address them
- Identification of different interpretation of rules leading to different programming logic

- Addition of variables or updates to presentation based on sponsor reviewer feedback
- Updates so a DMC report is consistent with sponsor internal reporting standards.

When the sponsor QC process has identified such issues, we can argue that the QC process actually worked. However, many of these updates are related to sponsors own internal standards without regards to the usability of the report to the DMC. The question is whether or not such updates improve the usability of the reports by the DMC and the likelihood of errors. Establishment of (industry accepted) quality standards focused on fit-for-purpose that are communicated to the sponsor should help address issues that may or may not require modifications to existing programming.

ACCEPTABLE QUALITY STANDARD

We believe no quality standard in our industry exists that addresses both usability and low error rates for programming. A major roadblock will continue to be the perception that 100% QC implies perfection, and that perfection improves the quality of a product or service. We believe that an acceptable quality standard and objective criteria should be established and should extend (far) beyond the so-called gold standard of 100% independent programming. We believe a quality standard should:

- Clearly define the end-user and their requirements
- Allow for risked based QC approaches
- Define object criteria that focus on usability of the product or service as well as freedom from failure
- Provide a framework for decision making around imperfections

CONCLUSION

We believe that excellence in quality can be maintained without striving for perfection. This can be done when we approach quality from a fit-for-purpose mindset. When we consistently provide programming that results a highly usable product or service with a low error rate, excellence is obtained.

Quality standards with objective criteria are needed that go far beyond our typical quality control processes currently in place. Quality standards should emphasize that excellence can be achieved without perfection provided the products and services are:

- Must be fit-for-purpose, and usable by the end user.
- Minimize the likelihood of failure, which we define as incorrect decisions and subsequent actions that were made based on the decision.
- Demonstrated consistently

When applying these principles to providing DMC support, we recognize that interim versions of data may be used to support activities such as manual data review, data cleaning, annual safety updates required by regulatory agencies, and support of a DMC. Programming is critical to ensure patient safety, that the experiment goes as expected, and to identify unforeseen circumstances that may require mitigation. But it does not always need to be perfect. Excellence in quality can still be maintained.

REFERENCES

[1] Juran's Quality Handbook: The Complete Guide to Performance Excellence

[2] PHUSE Best Practices for Quality Control

[3] Randall, A. and Coar, W. [2018] Risk-based Validation in Clinical Trial Reporting: Focus on What Matters Most, PharmaSUG 2018

[4] Coar, W. [2022] Cautionary Notes when Working with Interim Data, WUSS 2022

ACKNOWLEDGMENTS

The author would like to his colleagues supporting statistics and programming at Cytel as well as Joshua Sanders, Meredith Alm, and Patti Arsenault from QA for their continual guidance.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

William Coar
Axio Research, A Cytel Company
William.coar@cytel.com