Quantitative Analysis of the QMS for Pharmaceutical Manufacturing

Guoxu Wang^{1,2} · Weibing Wang³ · Qiang Zheng^{1,2}

Accepted: 30 July 2021 / Published online: 12 September 2021
© The Author(s) 2021

Abstract

Purpose To propose a statistical methodology for quantitative analysis of the quality management system (QMS) of pharmaceutical manufacturing.

Methods (1) Based on the manufacturing data from two established active pharmaceutical ingredient (API) manufacturers in China from 2010 to 2019, the linear regression with Pearson correlation coefficient is used to find the correlations between the proposed QMS operation indicators and performance indicators. (2) A stepwise multiple linear regression is used to identify the independent operation indicators with the biggest impact on a given performance indicator. (3) The Akaike Information Criterion is used to predict the performance indicators based on the operation indicators.

Results (1) Correlation: the right-first-time rate correlates strongly with various changes and deviations; the customer complaints correlate with changes, deviations, and CAPAs; the deficiency rate of foreign inspections correlates with deviations and CAPAs; and the CAPA on-time completion rate correlates with changes, deviations, and the ratio of employees in quality. (2) Impact: the right-first-time rate and the customer complaints are mostly impacted by the total deviations; the deficiency rate of foreign inspections is mostly impacted by deviations in equipment and instrument, and deviations due to human error; the CAPA on-time completion rate is mainly impacted by deviations in facility and utilities. (3) Predictability: the right-first-time rate, the customer complaints, the deficiency rate of foreign inspections, and the CAPA on-time completion rate can all be predicted based on the existing data with statistical significance.

Conclusions Deviations emerge as a key leading indicator for the performance of QMS. The proposed statistical methodology provides a basis for the data-driven quality management and regulation, whose visibility and predictability are likely to progress as the data accumulates.

Keywords QMS · Pharmaceutical manufacturing · Statistical modeling · Performance · Deviations

Introduction

Quantitative characterization of the quality management system (QMS) for pharmaceutical manufacturing has been a research subject of academia, industry, and regulator for years. A successful characterization usually requires a large

amount of real-world manufacturing quality data and a systematic statistical approach.

Macher of Georgetown University and Nickerson of Washington University studied manufacturing quality risks based on the US Food and Drug Administration (FDA) site inspection data for the purpose of prioritizing GMP inspections to manufacturing sites with high-quality risks [1]. They also included site operation data of manufacturing facility, human resources, deviation management, and manufacturing performance. Data was collected by using survey questionnaires to 50 active pharmaceutical ingredient (API) and finished dosage form (FDF) manufacturing sites. The data was product-based, i.e., data was collected for each product for each site, on a monthly basis from 1999 to 2003. The statistical analysis using linear regression and Pearson correlation coefficient is site-based by aggregating data of all the products for each site. They found that larger sites

✉ Qiang Zheng
zhengqiang@pku.edu.cn; qiangzheng@hotmail.com

¹ Department of Industrial Engineering and Management, College of Engineering, Peking University, Beijing 100871, China

² Center for Pharmaceutical Information and Engineering Research, Peking University, Beijing 100871, China

³ Academy for Advanced Interdisciplinary Studies, Peking University, Beijing 100871, China

correspond to more deviations. For API manufacturing sites, the site size correlates with raw materials deviation, production equipment deviation, and product and process deviation, with p -values less than 0.01, 0.05, and 0.1, respectively.

In the spirit of you can't improve what you can't measure, and for the desire of knowing the state of pharmaceutical quality, the FDA initiated the effort to study quality metrics [2]. International Society for Pharmaceutical Engineering (ISPE), in collaboration with McKinsey & Company, conducted two waves of large-scale quality metrics study for the purpose of identifying metrics that characterize the state of pharmaceutical quality [3, 4]. Data was collected by using survey questionnaires to 83 API and FDF sites on a monthly basis from 2015 to 2016, including both product-based and site-based descriptors, such as complaints, acceptance lots, invalidated out-of-spec (OOS) rate, and CAPA effectiveness. The study found, by using linear regression, a number of correlations with a p -value < 0.05, including between the lot acceptance rate and the invalidated OOS rate ($R^2 = 0.78$), between the deviation rate and lots pending 30 days for disposition ($R^2 = 0.52$), and between deviations with unknown root cause and lots pending 30 days for disposition ($R^2 = 0.35$).

Friedli of the University of St. Gallen and his collaborators have continued quality metrics study, in the context of operational excellence (OPEX) and quality system maturity [5]. They studied the relationship between quality control (QC) lab and OPEX, based on the survey data of 53 QC labs of pharmaceutical manufacturers, where the annual site-based data was collected between 2016 and 2018 [6, 7]. Using the independent t -test and the multiple linear regression, they found that more OPEX implementation leads to higher QC labs effectiveness, which is characterized by indicators of customer complaint investigation rate, invalidated OOS, delayed CAPAs, QC lab deviations, and their reoccurring deviations. They also studied the relationship between quality culture behavior and quality system maturity based on the survey data of 336 manufacturing sites in 2018 and found that higher-performing OPEX leads to more timely release of products, higher level of quality culture behavior, and quality system maturity [8]. They also found, based on a survey of 61 manufacturing sites from 2009 to 2019 and using t -test, a positive correlation between higher-performing OPEX and better FDA inspection outcome in terms of inspection classification, which includes no action indicated (NAI), voluntary action indicated (VAI), and official action indicated (OAI), with a significance level at 5% of t -test [9].

McKinsey & Company used its Pharma Operational Benchmarking Tools (POBOS) to compare the performance of more than 100 indicators of quality results and quality costs with its internal database of more than 600 pharmaceutical manufacturing sites over 40 countries to identify factors affecting quality, efficiency, customer satisfaction,

and potential risks, especially the commercial impact of the quality problems [10]. Using linear regression, McKinsey found correlation between the deviation rate and right-first-time rate ($R^2 = 0.96$) and the batch rejection rate ($R^2 = 0.91$), respectively, and between the customer complaint and the batch rejection rate ($R^2 = 0.71$) and the product recall ($R^2 = 0.56$), respectively [11].

Peking University's Pharmaceutical Information and Engineering Research Center led a team of eight leading pharmaceutical manufacturers in China to conduct a quality metrics pilot study from 2014 to 2015 [12]. The data was collected from eight API and FDF manufacturing sites, on an annual basis from 2011 to 2014, including both product-based and site-based. The study met challenges in that some of the sites did not have a matured QMS and lacked a meaningful system to track drug adverse effects, recalls, and customer complaints. Some of the data was not clearly defined or incomplete. Despite repeated efforts to standardize the metrics interpretation, some of the study outcomes appear better than reality, as shown in Table 1.

Most of the above quantitative studies concern with finding correlations among QMS indicators of many manufacturers, which have also been broadly named as elements, variables, descriptors, and metrics. The effectiveness of these studies relies on the condition that different sites share a similar mature QMS so that it makes sense to compare data across sites. In reality, this condition may not always be met and is definitely problematic in China. Thus, it would be helpful to first study correlations among the QMS attributes for a given site, where the QMS is stable, the data interpretation and collection are likely to be consistent, and the statistical analysis of the data is likely meaningful. A good quantitative characterization of QMS within one site would provide useful insight for the quantitative QMS comparison of multiple sites.

While correlations among QMS attributes are quantitative by nature, they may not provide quantitative information on the impact of any particular indicator. For instance, suppose the right-first-time rate is found in correlation with the deviations due to human error and the total deviations. Since the two deviations are dependent, it is hard to say quantitatively how much the right-first-time rate is impacted by the deviations due to human error. In fact, this is a general

Table 1 Comparison between Peking University and ISPE quality metrics studies

	Peking university study	ISPE study
Quality complaint rate (ppm)	0.03	26
Deviation rate	2.80%	19.20%
Recurring deviation rate	0.65%	15.00%
CAPA effectiveness	100.00%	89.00%

problem for correlation studies, when multiple indicators are introduced without consideration of their dependency.

Moreover, the ultimate goal of the quantitative study of QMS is to achieve some kind of predictability. For instance, ideally, if a quantitative negative correlation is established between the deviations due to human error and the right-first-time rate, then the latter can be increased by controlling the former. To authors' knowledge, a rigorous statistical approach with prediction power of pharmaceutical QMS has not been reported in literature.

The present work addresses two issues. One is to propose a statistical methodology to systematically study correlation, independence, and prediction among QMS indicators. Another is to cross-validate the methodology and the data from two established API manufacturers in China. The belief here is that, given the methodology's generality and rigorosity, if the calculated correlations and predictions are consistent with the professional common sense, the data is likely to be meaningful, and this, in turn, may indicate the usefulness of the methodology.

Methodology

A. Data Collection, Definition, and Exclusion

Two large manufacturing sites of two established domestic pharmaceutical firms participated in the present study. Per their request, the site names are referred to as site A and site B in the present work.

Site A is a large API and FDF manufacturer in China, exporting most of its products to 70 countries and regions, including the USA and EU. It received its first FDA inspection in 1992 and has been frequently inspected by regulators and audited by customers. It produces some 90 API products and 15 FDF products, covering therapeutic areas of anti-tumor, anti-infection, anti-depression, cardiovascular, endocrine, immune-suppression, and orthopedics. Its quality team has some 600 full-time employees. Since site A's FDF products are relatively new with incomplete data, only its API products are included in the present study.

Site B primarily specializes in manufacturing of cephalosporin APIs, including sterile and oral APIs and pharmaceutical intermediates. As one of the largest cephalosporin API manufacturers in China, it has 26 listed products, exporting to over 40 countries and regions, including the USA and EU, and receives frequent foreign inspections and audits. Its quality team has some 400 full-time employees.

The data was collected on an annual basis from sites A and B, covering all their API products from 2010 to 2019. Data from site A was derived directly from the operation record by authors, whereas data from site B relied on survey questionnaires. In both cases, extensive effort was made to verify the consistency, authenticity, and accuracy of data,

with site visits and site managements' support. The data of Site A was collected first, and the data interpretation was rather straightforward. This is because the full manufacturing record was accessible to the authors. Based on the experience of collecting and processing Site A's data, the questionnaires for Site B were designed. The task of completing the questionnaires was extensive and iterative, as the data definition and interpretation took time and effort to reach a consistently mutual understanding. It took a nearly 7000-word questionnaire guidance and three months to finish the data collection work. It should be cautioned that the integrity of the original data from sites A and B should ultimately be on par with the integrity of the sites and firms, i.e. the data used in the present study is only truthful and complete, to the best knowledge and effort of the authors through numerous rounds of discussion with sites and by site visits, to the manufacturing record. i.e. the data used in the present study is only truthful and complete, to the best knowledge and effort of the authors through numerous rounds of discussion with sites and by site visits, to the manufacturing record.

While the data was collected for 10 consecutive years from 2010 to 2019, not all parts of the data exist for all 10 years. For instance, site A did not start to collect data on deviation recurring rate until 2017. Some data has little or no variations over the years. For instance, the customer audit pass rate was always 100% for both sites A and B. The customers of API products of these two firms are usually large overseas FDF manufacturers, and their audits are usually professionally done and the outcome data tends to be reliable. Nonetheless, this makes the data not conforming to the normal distribution and thus not readily for statistical analysis that is often normal distribution based.

For purposes of characterizing QMS and making statistically meaningful predictions, the QMS indicators are assigned to two groups, the QMS operation indicator group and the QMS performance indicator group (Table 2). The former includes 25 indicators representing the operation aspect of QMS, such as changes and deviations. The latter includes 16 indicators representing the outcome of QMS, such as audit and inspection outcomes. Each dataset has $410 = ((25 + 16) \times 10)$ individual data elements, including 25 for operation indicators and 16 for performance indicators for 10 years.

The selection and assignment of operation and performance indicators are based on literature precedent, although the names used vary broadly, including attributes, descriptors, drivers, elements, enablers, factors, indicators, metrics, and variables [1, 3–7, 9, 11, 12]. Several indicators are new, including total products, deficiency rate of internal audit, customer audit pass rate, inspection pass rate, deficiency rate of inspections, and complaint recurring rate. The indicators in Table 2 are by no means a complete characterization of QMS.

Table 2 Definition of QMS operation and performance indicators. All numbers (#s) are counted for a site within a calendar year**QMS Performance Indicators (16 Indicators)****Product (4 Indicators)**

Right-First-Time Rate	Percentage of lots that need no reprocess nor rework
Total Complaints	# of complaints for all products
Complaint Rate	Total # of complaints over # of dose units released
Total ADRs	# of adverse drug reactions (ADRs) received for all products

Audit and Inspection (9 Indicators)

Customer Audit Pass Rate	# of customer audits passed over # of customer audits conducted
Inspection Pass Rate	# of regulatory inspections passed over conducted
Deficiency Rate of Domestic Inspections	# of deficiencies cited in domestic inspections over # of domestic inspections
Deficiency Rate of Foreign Inspections	# of deficiencies cited in foreign inspections over # of foreign inspections
Warning Letters	# of warning letters issued by FDA
OAI 483s	# of Official Action Indicated 483 forms issued by FDA
VAI 483s	# of Voluntary Action Indicated 483 forms issued by FDA
NAI 483s	# of No Action Indicated 483 forms issued by FDA
Lots Recalled	# of lots recalled for all products

Continuous Improvement (3 Indicators)

Deviation Recurring Rate	# of recurring deviations over total # of deviations
Complaint Recurring Rate	# of recurring complaints over the total # of complaints
CAPA On-Time Completion Rate	# of CAPAs completed on time over the total # of CAPAs

QMS Operation Indicators (25 Indicators)**QMS Maturity (20 Indicators)**

Changes of Materials and Containers	# of changes of raw and packaging materials, containers, and suppliers
Changes of Product Formulation	# of changes of excipient and their usage amount
Changes of Manufacturing Process	# of all manufacturing process changes, according to SOPs including those not subject to regulatory oversight
Changes of Product Specs and Testing Methods	# of changes of specs for materials, intermediate and finished product, and testing methods, including also # of changes in testing equipment and facilities, as well as testing lab renovations
Changes of Utilities and Warehouse	# of changes of the water system, the air conditioning system, and storage facilities
Changes of Documentation	# of changes in the text in the SOPs and lot/batch record, not related to process parameter changes, only to improve readability
Changes of Outsourcing Services	# of changes of services providers to material and finished product manufacturing and testing, rodent control, etc
Misc Changes	# of changes of, e.g., contact information of vendors or service providers
Total Changes	The sum of #s of all above changes
Deviations in Confirmed OOS	# of out of spec (OOS) deviations not due to laboratory errors
Deviations in Manufacturing Process	# of formulation or operational errors, process parameters outside the defined range, and occurrences of new impurities
Deviations in Equipment and Instruments	# of equipment and instrument failures and shutdowns during production, and passed the time for re-calibration
Deviations in Facility and Utilities	# of deviations of, e.g., the pressure difference in the clean area exceeds the range, the water system exceeds microbial spec; the environmental monitoring result exceeds the limits
Deviations in Documentation	# of deviations, e.g., a lack of production record, a poorly recorded or altered documentation
Deviations due to Human Error	# of deviations caused by human errors
Misc Deviations	# of miscellaneous deviations caused by other reasons
Total Deviations	The sum of all #s of deviations
Total CAPAs	Total # of corrective and preventive actions being conducted
Deficiency Rate of Internal Audit	# of deficiencies cited in internal audits over the # of internal audits conducted
On-Time Rate of Annual Review	# of annual product quality reviews completed within 30 days of the deadline over the total # of products manufactured

Table 2 (continued)**QMS Performance Indicators (16 Indicators)****Manufacturing (3 Indicators)**

Lot Acceptance Rate	# of lots accepted over the # of lots manufactured
Invalidated OOS Rate	# of invalid OOSs, due to testing negligence, over total # of OOSs
Total Products	# of all products manufactured

Personnel (2 Indicators)

Ratio of Employees in Quality	# of employees working in quality function over the total # of employees
Ratio of Highly Educated Employees in Quality	# of quality employees with a master's degree or above over the total # of quality employees

The partition of QMS indicators into the operation and performance groups, under the assumption that the former impacts the latter, is not always a clear cut. For instance,

consider the total number of CAPAs and the total number of complaints, which are treated operation and performance indicators, respectively. In reality, opening a CAPA is often

Table 3 Selected QMS indicators for further statistical analysis

	Site A		Site B	
	Shapiro–Wilk <i>p</i> -value	Years with missing data	Shapiro–Wilk <i>p</i> -value	Years with missing data
QMS Performance Indicators (5 Indicators)				
Product (3 Indicators)				
Right-First-Time Rate	0.321	3	0.070	0
Total Complaints	0.071	0	0.175	0
Complaint Rate	0.443	3	0.898	0
Audit and Inspection (1 Indicator)				
Deficiency Rate of Foreign Inspections	0.024	0	0.667	4
Continuous Improvement (1 Indicator)				
CAPA On-Time Completion Rate	0.463	0	0.203	0
QMS Operation Indicators (17 Indicators)				
QMS Maturity (14 Indicators)				
Changes of Materials and Containers	0.228	0	0.032	1
Changes of Manufacturing Process	0.849	0	0.968	1
Changes of Product Specs and Testing Methods	0.427	0	0.122	1
Changes of Utilities and Warehouse	0.602	0	0.059	1
Misc Changes	0.659	0	0.078	1
Total Changes	0.318	0	0.397	0
Deviations in Manufacturing Process	0.099	0	0.379	0
Deviations in Equipment and Instruments	0.350	0	0.113	0
Deviations in Facility and Utilities	0.500	0	0.069	0
Deviations due to Human Error	0.566	0	0.081	0
Misc Deviations	0.120	0	0.220	0
Total Deviations	0.145	0	0.916	0
Total CAPAs	0.124	0	0.492	0
Deficiency Rate of Internal Audits	0.060	0	0.227	0
Manufacturing (2 Indicators)				
Lot Acceptance Rate	0.321	3	0.012	0
Total Products	0.931	3	0.016	0
Personnel (1 Indicator)				
Ratio of Employees in Quality	0.152	0	0.025	0

a way to address a complaint, so the performance indicator impacts the operation indicator, not the other way around. This case will be discussed later in the Results Section of the present work.

Table 2 lists two groups of 16 performance indicators and 25 operation indicators, which are further assigned to three subgroups, respectively. For the performance indicator group, the three subgroups are product indicators, audit and inspection indicators, and continuous improvement indicators. For the operation indicator group, the three subgroups are QMS maturity, manufacturing, and personnel. The number counting follows the same rule for all products and for both sites. For instance, for the indicator “deviation recurring rate,” a recurrence was always added to the year when the deviation first occurred. Suppose a deviation occurred in 2017 for the first time, and in 2018 the second time, then the deviation was counted for both 2017 and 2018. However, the reoccurring deviation was only counted for 2017. Moreover, a reoccurring deviation was counted only when it occurred within 12 months of the deviation first occurred. Otherwise, the deviation was treated as a new one. For another instance, a change was counted, only if it was considered as a change according to the standard operating procedures (SOPs), and in the year when it was implemented, not when it was proposed or closed.

The data of sites A and B were collected according to the above defined QMS indicators, followed by statistical analysis to determine whether the data of each indicator follows a normal distribution.

To see whether a distribution is normal, a combination of qualitative and quantitative methods is used. First, a histogram and a Q–Q plot are used to visualize the distribution shape. Then, the Shapiro–Wilk test is applied [13]. If the

significant level of the test (the Shapiro–Wilk p -value) is larger than 0.01, then the data is deemed to follow a normal distribution. The following three Q–Q plots provide examples of the typical appearance of Q–Q plots and their associated Shapiro–Wilk p -values (Fig. 1).

If the Shapiro–Wilk p -value is larger 0.01, the data is considered as following a normal distribution. For convenience of statistical analysis in the present work, only indicators with normally distributed data for both sites A and B are included for further analysis. Moreover, indicators with more than 4 years of missing data are also excluded from further analysis. The “missing data” refers to data that was not collected by the site(s), usually during the early years, not data that was collected but missing from the record. These two requirements exclude 11 out of the 16 performance indicators and eight out of 25 operation indicators, and the remaining indicators are tabulated in Table 3, along with their Shapiro–Wilk p -values and data missing year information.

B. Data Analysis and Model Building

(1) Correlation

The Pearson correlation coefficient and p -value are used to find correlations between QMS operation indicators and QMS performance indicators. In general, three prerequisites for using the Pearson correlation coefficient hypothesis test are as follows: data is normally distributed; the gap between the data is not too large, and each group of data is independently sampled [14]. The first prerequisite is automatically met due to the data exclusion requirement discussed in the previous section. The last two prerequisites are generally met, because data from the same site

Fig. 1 Q–Q plots of changes of manufacturing process (Shapiro–Wilk p -value=0.849), deviations due to human error (Shapiro–Wilk p -value=0.566), and total complaints (Shapiro–Wilk p -value=0.071)

does not vary much year over year, and each year's data for each indicator was indeed separately recorded and counted.

For the Pearson correlation coefficient r , the considerations are as follows. A positive r indicates a positive correlation, and a negative r indicates a negative correlation. An absolute value of r between 0.8 and 1.0 is considered as a strong correlation. An absolute value of r between 0.5 and 0.8 is considered as a significant correlation. An absolute value of r between 0.3 and 0.5 indicates is considered as a weak correlation [15].

To ensure the integrity and rigorousness of analysis, in correlation analysis, none of the indicators is weighted, no data is arbitrarily discarded, and all statistically significant (p -value ≤ 0.1) results are included for analysis. Please note this p -value is different from the previous Shapiro–Wilk p -value in the previous section.

(2) Impact Analysis

On the basis of correlation analysis, for each of the five performance indicators in Table 3, a stepwise multiple linear regression is applied to the 17 operation indicators in Table 3. The operation indicator dependency and the linear regression's multi-collinearity problems are overcome by using the stepwise regression method that starts with one operation indicator and adds or removes operation indicator(s), while checking and removing dependency [16–18]. The objective of this analysis is to identify the leading independent operation indicators that impact the performance indicators the most.

(3) Predictability

An optimal statistical analysis is performed by using the Akaike Information Criterion (AIC),

$$\text{AIC} = -2 \ln(L) + 2k \quad (1)$$

where $\ln$ is the natural logarithm function, L is a likelihood function, and k is the number of operation indicators used for model building [19]. The objective is to optimize (minimize) AIC by maximizing L and minimizing k .

Here is how the AIC method works. Consider $k=2$, and a linear random equation:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \xi \quad (2)$$

where Y represents a performance indicator; X_1 and X_2 represent two randomly selected operation indicators; β_0 , β_1 , and β_2 are three parameters to be determined; and ξ represents a random variable. Equation (2) says that performance indicator Y is dependent to operation indicators X_1 and X_2 , and the dependency is probabilistic in nature. ξ is assumed to follow a normal distribution with a variance of σ . The likelihood function L is a probability density function of three variables

of Y , X_1 , and X_2 . Parameters β_0 , β_1 , β_2 , and σ are determined by the assumption that L is maximized under the condition of a given set of values of $\{Y^*, X_1^*, X_2^*\}$, which are drawn from the 10 years of data of the indicators. Variance σ is assumed to be the same throughout the years. The rationale of the AIC optimization is that the observed events $\{Y^*, X_1^*, X_2^*\}$ are the most likely to happen in nature and that's why they are observed, and this likelihood should be reflected in the likelihood probability function L .

The above selection of performance indicator Y , the number $k=2$, and operation indicators X_1 and X_2 are arbitrary. In practice, the optimization is repeated for each of the five performance indicators in Table 3 for all possible combinations of $k=1, 2, \dots, 17$. If k is small, L may not capture all the correlation information between a performance indicator and the k operation indicators. If k is large, L may reach saturation of correlation information, but the penalty term $2k$ in Eq. (1) is large. So, to minimize AIC, it is necessary to maximize L and minimize k . The essence of the AIC method is to select the fewest variables to make the model best fit the data.

For the AIC method to work, a large number of data is needed, so that part of the data can be used as the training set, and the remaining part as the test set. The predictability is determined by predicting the data of the test set based on the data of the training set. The overall predictability is measured by the average fitness of many random partitions of training and test sets. The AIC method usually requires the training set have at least ten data points, corresponding to the 10 years of the data. But the present study only has 10 years of data for each manufacturing site, thus leaving no data for the test set. To compromise, the data of sites A and B are pooled together by introducing a dummy variable, whose value is set to 0 for site A and 1 for site B, to form a single set of data of 20 years. This is the only part of the present work that the data from two separate sites is pooled together for statistical analysis. In consequence, the outcome of the AIC analysis is the average predictability of the performance indicators based on the operation indicators of the combined data set.

Results

1. Correlation

(i) Product Indicators

Site A The right-first-time rate and the total changes show a strong positive correlation ($r=0.909$), suggesting that changes were effective in increasing the right-first-time rate. The types of changes most closely related to the increased right-first-time rate are misc and material and container

Table 4 Correlations between product performance indicators and operation indicators

Operation Indicators	Site A		Site B	
	<i>p</i> -Value	Correlation Coefficient (<i>r</i>)	<i>p</i> -Value	Correlation Coefficient (<i>r</i>)
Right-First-Time Rate				
Total Changes	0.0045	0.909		
Misc Changes	0.0229	0.823		
Changes of Materials and Containers	0.0656	0.724		
Changes of Utilities and Warehouse			0.0886	0.599
Total Deviations	0.0035	−0.919	0.0028	−0.832
Deviations in Manufacturing Process	0.0021	−0.933	0.0040	−0.816
Deviations in Equipment and Instruments	0.0146	−0.853	0.0379	−0.660
Deviations in Facility and Utilities			0.0104	−0.762
Deviations due to Human Error	0.0062	−0.897		
Misc Deviations	0.0461	−0.763		
Total Complaints				
Misc Changes			0.0851	−0.604
Total Deviations	0.0129	0.748		
Deviations in Manufacturing Process	0.0264	0.693	0.0972	0.553
Deviations in Equipment and Instruments	0.0249	0.698		
Deviations due to Human Error	0.0111	0.758		
Total CAPAs	0.0387	0.658		
Complaint Rate				
Changes of Product Specs and Testing Methods	0.0481	−0.759		
Total CAPAs	0.0306	0.800		

changes. The reason may be that packaging is the last process in manufacturing and changing the packaging materials had a direct impact on product release. The right-first-time rate and the total deviations show a strong negative correlation ($r = -0.919$), echoing the experience that more deviations lead to a decreased right-first-time rate. Similar strong correlations are seen across all deviations in Table 4.

The total complaints and the total deviations show a significant correlation ($r = 0.748$), suggesting that more deviations led to more complaints. The types of deviations that correlate most with the complaints are deviations in manufacturing process, equipment and instrument, and due to human error.

The complaint rate is also significantly and negatively correlated with the changes of specs and testing methods ($r = -0.759$). This is conceivable, as e.g., tightening the product specs and improving testing methods, is often an important means to address customer complaints. A strong positive correlation between the complaint rate and CAPA ($r = 0.800$) is also expected, as a manufacturer often addresses a complaint by opening up a CAPA.

Site B The correlations between the performance and operation indicators are somewhat similar to site A, but there are notable differences.

The correlation between the right-first-time rate and changes is much less pronounced, with only the changes of utilities and warehouses showing a nearly significant correlation ($r = 0.599$). This weakened correlation may be due to the following reasons. Since 2017, the revisions of SOPs at site B are no longer managed by the change control system and are counted separately. As a result, since 2017, the annual number of changes of “documentation” has dropped sharply from 223 in 2016 to 48 in 2017 and has stayed more or less at this level. Since the changes of documentation of site B accounted for more than 40% of the total changes before 2017, such a drop reduced the total number of changes, leading to the weakened correlation. The correlation between the right-first-time rate and the total changes is too weak ($r = 0.391$, p -value = 0.263) to be included in Table 4.

A significant negative correlation ($r = -0.604$) is observed between the total complaints and the misc changes, but no such correlation is observed for site A. Given that the misc changes are rather non-specific, the present work has no rationale for this correlation.

The correlation between the total complaints and deviations is narrower and weaker than site A, with only deviations in manufacturing process showing a significant

correlation (0.553) but with a much larger p -value (0.097). The present work has no rationale for the narrow and weak correlation. The same is true for the lack of correlation between the complaint rate and the changes of product specs and testing methods.

The lack of correlation between complaints and total CAPAs is probably due that site B runs a centralized CAPA system that assigns a unique CAPA id for the entire site, whereas site A runs a decentralized CAPA system that tends to over-count CAPAs. For example, if a CAPA is opened for addressing a customer complaint, it is counted as one CAPA by the unit responsible for complaint management. Later, a change is initiated for addressing the same complaint, the change management unit would count it as another CAPA. When the total CAPAs of site A are counted, the same CAPA is counted twice. For reference, annually and on average, site B has 314 CAPAs, while site A has 1243 CAPAs.

In summary, all the correlations shown in Table 4 are consistent with common sense. What is inconsistent is that similar correlations do not always show up across sites A and B. Since the correlation analysis itself is a straightforward mathematical calculation, the inconsistent outcome may suggest that either the original recording of data is flawed, or the simple correlation analysis does not capture the actual complex relationship between the performance and operation indicators. In fact, similar inconsistency happens throughout the following sections. The present work does not know which possibility is right.

(ii) Audit and Inspection Indicator

Site A The number of cited deficiencies in a foreign inspection is a crude characterization of inspection outcome, as more deficiencies do not necessarily translate into a more severe compliance problem, and the number tends to vary from one inspector to another. It is therefore interesting to see the significant positive correlation between the deficiency rate and a range of deviations, even though it is conceivable that more deviations tend to lead to more inspection deficiencies.

The significant positive correlation with the total CAPAs ($r=0.585$) is not surprising, as an inspection deficiency is usually a good reason for the site to open a CAPA. However, the relationship is different in this case. Usually, it is the operation indicators that impact the performance indicators, but here, it is the other way around. This shows that the assignment of QMS indicators into two groups of performance indicators and operation indicators, and the assumption that the latter impact the former, are not always appropriate. But as far as correlations

themselves are concerned, this is fine, as correlation does not imply causation.

Site B Compared to site A, site B has a more focused product range, fewer customers, and fewer customer audits and regulatory inspections. For instance, between 2010 and 2019, site B did not have any foreign regulatory inspection in four of these years. Consequently, there are fewer inspection-related correlations. The one exception is the strong correlation ($r=0.901$) and high statistical significance (p -value=0.01) with the deviations due to human error, which is absent from site A. This is probably due to the fact that site A's classification of the deviations due to human error is rather broad that some of the deviations should have been classified into other types of deviations [20]. The broad classification increased the noise of data, thus reducing the signal of correlation.

(iii) Continuous Improvement Indicator

Site A When the number of CAPAs goes up, the CAPA on-time completion rate comes down, as shown by the significant negative correlation ($r=-0.775$) and low p -value (0.0084). This can be explained that more CAPAs would stretch thinner the human resources, thus lowering the CAPA on-time completion rate.

The lack of correlation with changes and deviations, coupled with significant correlation with the total CAPAs, seems to suggest that the opening and completion of CAPAs were not closely related to changes and deviations, indicating a poor CAPA management, probably caused by the decentralized CAPA management system of site A as discussed previously.

Another explanation for the observed significant correlation between the CAPA on-time completion rate and the total CAPAs is as follows. Consider that the opening of a CAPA is double-counted by two different units. Once a CAPA is closed by one unit, the respective CAPA is also closed by the other unit, resulting in a double-counted completion of CAPA. When the CAPA on-time completion rate is calculated, these two double counts cancel each other, such that the doubling counting of CAPAs does not affect the correlation between the total CAPAs and the CAPA on-time completion rate.

Site B The correlation with operation indicators is broad and significant, with the strongest correlation to deviations in facility and utilities. Similar to site A, a significant negative correlation ($r=-0.662$) is observed for the total CAPAs. Taken together, these correlations suggest that site B has a better CAPA management than site A, as site B's CAPA on-time completion rate is closely correlated to changes and deviations, as they should.

Table 5 Correlations between an audit and inspection performance indicator and operation indicators

Operation indicators	Deficiency rate of foreign inspections			
	Site A		Site B	
	<i>p</i> -Value	Correlation coefficient (<i>r</i>)	<i>p</i> -Value	Correlation coefficient (<i>r</i>)
Total deviations	0.0613	0.610		
Deviations in equipment and instrument	0.0180	0.724		
Deviations in facility and utilities	0.0517	0.628		
Deviations due to human error			0.0141	0.901
Total CAPAs	0.0756	0.585		

The observed significant correlation ($r=0.558$) with the ratio of employees in quality suggests that by adding employees to the quality team can help to speed up the CAPA completion process. If this is true, further suggestions can be made as follows. The fact that this correlation is absent from site A but not site B may be due to the ratio of employees in quality of site A is lower than site B (11% vs 18%). It is also possible that for site A, few employees in quality were assigned to work on CAPAs, when the total number of employees in quality was increased. Again, these suggestions may signal a poor CAPA management system of site A.

2. Impact Analysis

The previous section shows a complex correlation pattern between QMS operation indicators and QMS performance indicators, where one operation indicator correlates with multiple performance indicators and one performance indicator correlates with multiple operation indicators. For the latter, one may ask: for a given multiple to one correlation, assuming the operation indicators impact the performance indicator, which operation indicators make the biggest impact, and by how much?

These are fundamental questions to any quantitative analysis, where an outcome variable depends on multiple

source variables that may not be independent. One natural tendency is to weigh on these source variables, but how to assign the weights poses a challenge, as the weights should really be an outcome of the analysis, not input parameters that bias some source variables to others. If an equal weight is used, it may appear unbiased, but it is not necessarily true. This is because the source variables' dependency affects the weights. Consider the following extreme case, if two nearly identical source variables are used, it is equivalent to using just one of them but with a weight factor of two. So, it is important to identify the independent source variables and find their independent impact to the outcome variable [21, 22].

Based on correlation findings between operation and performance indicators in the previous section, for each of the five performance indicators in Tables 4, 5, and 6, a stepwise multiple linear regression method is applied [16–18]. This method avoids multi-collinearity problem of the regular linear regression method, selects the independent operation indicator(s) with the biggest impact to the correlation, and calculates the probability that how much the calculated impact accounts for the true relationship by the value of adjusted R^2 , which measures the goodness of fit of the stepwise linear regression to data. In general, adjusted R^2 values of 0.75, 0.50, or 0.25 are considered as substantial, moderate, and weak impacts, respectively. For example, $R^2=0.75$

Table 6 Correlations between continuous improvement performance indicators and operation indicators

Operation indicators	CAPA on-time completion rate			
	Site A		Site B	
	<i>p</i> -Value	Correlation coefficient (<i>r</i>)	<i>p</i> -Value	Correlation coefficient (<i>r</i>)
Changes of manufacturing process			0.0725	0.624
Total deviations			0.0122	−0.751
Deviations in manufacturing process			0.0744	−0.587
Deviations in equipment and instrument			0.0201	−0.715
Deviations in facility and utilities			0.0020	−0.827
Total CAPAs	0.0084	−0.775	0.0372	−0.662
Ratio of employees in quality			0.0940	0.558

Table 7 The regression result of the right-first-time rate and the total complaints, based on data in Table 4

	Site A	Site B
Right-First-Time Rate (%)		
Deviations in Manufacturing Process	−0.021***	
Total Deviations		−0.0028***
Adjusted R^2	0.8453	0.6544
Total Complaints		
Deviations due to Human Error	0.054**	
Adjusted R^2	0.5208	

indicates that 75% of the impact (correlation) from the operation indicator to the performance indicator is captured by the stepwise linear regression [23].

For brevity, only regression results with high (p -value < 0.01, noted by ***) and moderate (p -value < 0.05, noted by **) statistical significance in are discussed in the following paragraphs.

(i) Product Indicators

(a) Right-First-Time Rate For site A, Table 7 shows that the deviations in manufacturing process are the operation indicator that has the biggest impact on the performance indicator right-first-time rate. When the deviations in manufacturing process increase by one, the right-first-time rate decreases 0.021%. For site B, the biggest impact operation indicator is the total deviations. When the total deviations increase by one, the right-first-time rate decreases 0.0028%. While these regression values are quantitatively too small to have any practical significance, the negative signs are qualitatively correct, i.e., deviations reduce the right-first-time rate. Moreover, these values are of high statistical significance (p -value < 0.01) as noted by ***.

For site A, $R^2 = 0.8453$ shows that the deviations in manufacturing process, as a regression variable, captures 85% (substantial) of its correlation with the right-first-time rate. For site B, $R^2 = 0.6544$ (moderate to substantial) shows that the total deviations, as a regression variable, capture 65% of its correlation with the right-first-time rate.

Compared to Table 4, in which the right-first-time rate is shown to correlate strongly ($r > 0.8$) or significantly ($r > 0.5$) with eight operation indicators of various changes and deviations for site A, the stepwise linear regression reduces the seemingly complex one-to-many correlations to a simple one-to-one correlation, by removing the dependency among the eight operation indicators. As far as correlation to the right-first-time rate is concerned, all eight operation indicators are statistically dependent. The biggest impact operation indicator happens to have the biggest (absolute value) correlation coefficient $r = -0.933$ and smallest p -value = 0.0021.

The case for site B is similar, where the right-first-time rate is shown to correlate strongly ($r > 0.8$) or significantly ($r > 0.5$) with four deviations and one change related operation indicators, the stepwise linear regression method reduces them to a simple one-to-one correlation to the total deviations, by removing the dependency among them. The total deviations happen to have the biggest (absolute value) correlation coefficient $r = -0.832$ and the smallest p -value = 0.0028.

Please note that the correlation to the total deviations is very close to the correlation to the deviations in manufacturing process ($r = -0.919$ vs -0.933 , p -value = 0.0035 vs 0.0021) for site A in Table 4. If the total deviations are chosen as the regression variable, the regression results will likely be similar. Likewise, for site B, the correlation to the deviations in manufacturing process is very close to the correlation to the total deviations ($r = -0.816$ vs -0.832 , p -value = 0.0040 vs 0.0028). These observations suggest that, while Table 7 shows that the biggest impacts come from different operation indicators for sites A and B, the reality is that the total deviations can effectively be the biggest impact operation indicator for both sites. The apparent discrepancy is caused by the strictness of the stepwise linear regression itself, which rejects the addition of any operation indicator that does not substantially improve the regression fit.

(b) Total Complaints For site A, when the deviations due to human error increase by one, the total complaints increase by 0.054, with a moderate statistical significance (p -value < 0.05) as noted by **. $R^2 = 0.5208$ shows that the deviations due to human error, as a regression variable, capture 52% of its correlation with the total complaints. As far as correlation to the total complaints is concerned, all five operation indicators that show significant correlations in Table 5 are statistically dependent. Among them, the deviations due to human error happen to have the strongest correlation ($r = 0.758$) and the smallest p -value (0.001). Please note that the correlation coefficient and the p -value of the total deviations are very close to those of the deviations due to human error ($r = 0.748$ vs 0.758 and p -value = 0.0129 vs 0.0111), suggesting that, if the total deviations are used for linear regression, the outcome will likely be similar to that of the deviations due to human error.

For site B, possible explanations for the no show of moderate (p -value < 0.05) or strong (p -value < 0.01) impact from operation indicators in Table 7 are as follows. Site B's product range is narrower, and its customers are less than site A, so its customer complaints are less. The total complaints per year are, on average, 25 for site A, and 15 for site B. Moreover, the correlations to the misc changes and the deviations in manufacturing process in Table 4 are borderline weak ($r = -0.604$ and 0.553) and their statistical significances are

Table 8 The regression result of the deficiency rate of foreign inspections, based on data in Table 5

Deficiency rate of foreign inspections	Site A Site B	
	Site A	Site B
Deviations in equipment and instrument	0.15**	
Deviations due to human error		0.56**
Adjusted R^2	0.4639	0.7655

low (p -value = 0.0851 and 0.0972, and please note that these are the p -values of the correlation analysis in Table 4, not of the stepwise linear regression in Table 7).

(c) Complaint Rate The stepwise linear regression finds no statistically significant impact from the changes of product specs and testing methods and the total CAPA, whose rather strong correlations to the complaint rate are shown in Table 4. A probable explanation is as follows. The large p -values of these correlations (0.048 and 0.031) suggest that the statistical significance of the correlations is low, and upon doing the impact analysis by using the stepwise linear regression method, these correlations do not rise to the level of strong (p -value < 0.01) or moderate (p -value < 0.05) statistical significance.

(ii) Audit and Inspection Indicator

For site A, deviations have a moderate impact (p -value < 0.05) as noted by ** on the deficiency rate of foreign inspections, as shown in Table 8. Site A has a wide variety of API exports, and the equipment and instruments involved are complex and diverse, which were likely the focus of foreign inspections. For an increase of one deviation in equipment and instrument, the number of deficiencies cited in foreign inspections increases on average by 0.15. $R^2 = 0.4639$ shows that the deviations in equipment and instrument, as a regression variable, captures only 46% of its correlation with the deficiency rate of foreign inspections. In contrast to the previous section, where the total deviations have a nearly biggest impact on the product performance indicators, the total deviations' impact is not close to the impact of the deviations in equipment and instrument, as indicated by their respective correlation coefficients ($r = 0.610$ vs 0.724) and p -values (0.0613 vs 0.0180) in Table 5.

For site B, the stepwise linear regression has only one correlation, to begin with, as shown in Table 5, namely, the deviations due to human error. So, it is only natural that this operation indicator is the one with the biggest impact,

Table 9 The regression result of CAPA on-time completion rate, based on data in Table 6

CAPA on-time completion rate (%)	Site A Site B	
	Site A	Site B
Total CAPAs	−0.00092***	
Deviations in facility and utilities		−0.54***
Adjusted R^2	0.5513	0.6824

which as a regression variable, captures nearly 77% of the correlation with the deficiency rate of foreign inspections, with a moderate statistical significance (p -value < 0.05), as noted by **. For an increase of one deviation due to human error, the number of deficiencies cited in a foreign inspection increases on average by 0.56.

(iii) Continuous Improvement Indicator

For site A, the CAPA on-time completion rate is impacted by the total number of CAPAs. When the latter increases by one, the former decreases by 0.00092%. Quantitatively, this number is too small to have any practical significance, but its statistical significance is high (p -value < 0.01), as noted by ***. Qualitatively, however, the negative sign is confirmative, i.e., more CAPAs tend to drive down the CAPA on-time completion rate. $R^2 = 0.5513$ shows that the total CAPAs, as a regression variable, capture 55% of its correlation with the CAPA on-time completion rate.

For site B, the CAPA on-time completion rate is mainly impacted by the deviations in facility and utilities. When the latter increases by one, the former decreases by 0.54%. Again, quantitatively, this number is too small to have any practical significance, but its statistical significance is high (p -value < 0.01), as noted by ***. Qualitatively, the negative sign is confirmative, i.e., more deviations in facility and utilities tend to drive down the CAPA on-time completion rate. $R^2 = 0.6824$ shows that the deviations in facility and utilities, as a regression variable, capture 68% of its correlation with the CAPA on-time completion rate. Similar to the previous sections, the stepwise linear regression reduces the seven correlations to a single one, the deviations in facility, and utilities, which happens to have the strongest correlation ($r = -0.827$) and the smallest p -value (0.002) in Table 6. Please note that, while the total deviations appear to have the second biggest impact, its differences to the deviations in facility and utilities are significant, ($r = -0.751$ vs -0.827) and p -values (0.0122 vs 0.0020).

Please note that the stepwise linear regression does not always select just one regression variable. The reason that the regression results for each of the five performance indicators in Tables 7, 8, and 9 include just one operation

indicator is two folds. One is that the operation indicators are highly dependent. Another is the limit of the method itself, which adds or removes operation indicator(s) at each step of the way by a rather strict criterion.

3. Fitness and Predictability

For each of the five performance indicators in Table 3, any given value k ($1 \leq k \leq 17$), and any given set of k distinct operation indicators $X_1, X_2, \dots, X_k$ chosen from 17 operation indicators in Table 3, the linear random dependency expression can be written as

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k + \xi \quad (3)$$

where $\beta_0, \beta_1, \beta_2, \dots, \beta_k$ are parameters to be determined, ξ is a random variable that follows a normal distribution. The corresponding AIC in Eq. (1) is minimized by using the combined 20 years of data of the performance indicator and the 17 operation indicators from sites A and B. The minimum AIC of all the minimized AICs is selected as the best fit. The corresponding k is the best-fit k_b . The corresponding $\beta_0, \beta_1, \beta_2, \dots, \beta_k$ are the best-fit parameters $\beta_{0b}, \beta_{1b}, \beta_{2b}, \dots, \beta_{kb}$. The corresponding $X_1, X_2, \dots, X_k$ are the best-fit operation indicators, $X_{1b}, X_{2b}, \dots, X_{kb}$. The “best fit” means that, for the given 20 years of data of performance and operation indicators, the dependency between the performance indicator and the operation indicators is best described by the following AIC best model:

$$Y = \beta_{0b} + \beta_{1b} X_{1b} + \beta_{2b} X_{2b} + \dots + \beta_{kb} X_{kb} \quad (4)$$

Next, the AIC best model’s predictability is determined by splitting the data into training set and test set with a ratio of 3:1, i.e., training set has 15 years of data, and the test set has 5 years of data. This is a common practice of using AIC method [19, 24]. Since parts of data are missing certain years, as shown in Table 3, a training set may have less than 15 years of data, but a test set is always ensured with 5 years of data. The combined data of sites A and B are ordered as 2010–2019 of site A followed by 2010–2019 of site B, like for instance, [A2010, A2011, ..., A2019, B2010, B2011, ..., B2019]. The selection of 5 years for the test set is random, in the form of, for example, [..., A2012, ..., A2014, ..., A2018, ..., B2013, ..., B2015, ...], where the five arbitrarily selected years are explicitly shown.

The premise of the AIC method is that, if the best model fits the existing data well, the dependency expression of Eq. (4) should hold true for new data. For instance, let’s select the last 3 years of site A and the last 2 years of site B, i.e. A2017, A2018, A2019, B2018, and B2019 for testing. This amounts to using the first 7 years of data from site A and the first 8 years of data from site B to establish a dependency expression and apply it, as the prediction, to over the

last 3 years for site A and the last 2 years of site B. This is a typical forward prediction. Since the selection of the five data years is arbitrary, the prediction can be backward too, i.e., the prediction of the dependency relation over five “missing years.”

The predictability of the AIC best model is measured by the average of predictability of all possible test set selections. For a given test set, the root mean square error (RMSE) is calculated between the predicted values and the actual data values at the 5-year points. A smaller RMSE indicates a closer fit between the predicted and the actual values. Averaging RMSEs over all possible test set selections give rise to the RMSE of the AIC best model.

(i) Product Predictors

In Fig. 2, the AIC best model fitness and the predictability of the three product performance indicators in Table 3 are shown. On the left are the plots of the goodness of fit of the AIC best models. The abscissa represents the data years of sites A (2010–2019) and B (2010–2019) from left to right. The solid lines represent the actual data plots. The dotted lines represent the AIC best models. On the right are the plots of actual data (solid lines) and representative predictions (dotted line), along with the RMSE for the shown representative predictions and the RMSE averaged overall predictions.

(a) Right-First-Time Rate As shown in Table 3, site A misses data for the first 3 years (2010 to 2012), so the solid line starts from year mark 4. For the AIC best model, $k_b = 5$, and the five selected operation indicators are the lot acceptance rate, the deviations in the manufacturing process, the deviations in equipment and instrument, the misc deviations, and the dummy variable. Because site A’s lot acceptance rate also misses data for the first 3 years as shown in Table 3, the dotted line starts from year mark 4 as well. Both the AIC best model fitness and its predictability are good, qualitatively and quantitatively.

The biggest impact operation indicator for site A in Table 7, the deviations in manufacturing process, is selected by the AIC best model, but not the biggest impact operation indicator for site B, the total deviations. This is likely due to that the AIC best model is for the combined data of both sites, and the site A’s impact and adjusted R^2 (−0.021, 0.8453) are larger than site B’s (−0.0028, 0.6544). Consequently, site A has a stronger influence on the AIC best model than site B. In this regard, the result of AIC is consistent with the result of the stepwise linear regression.

The selection of the dummy variable indicates that there is a large difference between the data of sites A and B, and a dummy variable is needed to bridge the data. This is evident

Fig. 2 The fitness and predictability of the product performance indicators

in Fig. 2, where the average right-first-time rate for site A is visibly lower than for site B.

(b) Total Complaints For the AIC best model, $k_b=3$, and the three selected operation indicators are the lot acceptance rate, the total deviations, and the deficiency rate of internal audits. Again, due to the missing data of the lot acceptance rate, the dotted line starts from year mark 4. Both the AIC best model fitness and its predictability are good, qualitatively and quantitatively.

The biggest impact operation indicator for site A in Table 7, the deviations due to human error, is not selected by the AIC best model. However, as previously discussed, the impact of the total deviations is similar to the deviations due to human error, and the operation indicator of total deviations is selected by the AIC best model. In this regard, the result of AIC is consistent with the result of the stepwise linear regression.

(c) Complaint Rate Site A misses data for the first 3 years, the solid line starts from year mark 4. For the AIC best model, $k=4$, and the selected operation indicators are the total CAPAs, the deviations due to human error, the total changes, and the dummy variable. Because none of these four operation indicators misses data, the dotted line starts from year mark 1. The AIC best model fitness is good, both qualitatively and quantitatively. The predictability is good qualitatively, but less good quantitatively.

The selection of the dummy variable indicates that there is a large difference between the data of sites A and B, so that a dummy variable is needed to bridge the data. This is evident in Fig. 2, where the average complaint rate for site A is visibly higher than for site B.

The stepwise linear regression method yields no statistically significant (p -value < 0.05) information on impact from operation indicators to the complaint rate. In contrast, the AIC best model selects four operation indicators. Table 4 lists two correlating operation indicators. One of them, the total CAPAs is among the selected by the AIC best model. Another, the changes of product specs and testing methods, while not among the selected by the AIC best model, its encompassing one, the total changes are selected by the AIC best model. In this regard, the result of AIC is consistent with the result of the stepwise linear regression.

(ii) Audit and Inspection Indicator

In Fig. 3, the fitness and predictability of the deficiency rate of foreign inspections are shown for the AIC best model, where $k_b=3$, and the selected operation indicators are the deviations in equipment and instrument, the total deviations, and the total CAPAs. The overall fitness of the AIC best model is qualitatively good but is off quantitatively at several-year marks. The overall prediction, with the exception of year mark 3, is qualitatively good, i.e., following the same up and downtrend. The year mark 3 has no particular meaning, as what is shown is just an illustrated one from many predictions. The shown RMSE (14.24) and RMSE ave (14.56) are close, suggesting the shown prediction is, on average, representative for all the predictions.

The biggest impact operation indicator for site A in Table 8, the deviations in equipment and instrument, is also selected by the AIC best model. The deviations due to human error are not selected by the AIC best model, although its encompassing operation indicator, the total deviations, is. It is interesting to see that all three AIC selected operation

Fig. 3 The fitness and predictability of the deficiency rate of foreign inspections

Fig. 4 The fitness and predictability of the CAPA on-time completion rate

indicators are listed in Table 5. In this regard, the result of AIC is consistent with the results of the stepwise linear regression and the simple correlation analysis.

(iii) Continuous Improvement Performance Indicator

In Fig. 4, the fitness and predictability of the CAPA on-time completion rate are shown for the AIC best mode, where $k_b = 6$, and the selected operation indicators are the deviations in facility and utilities, the ratio of employees in quality, the deviations due to human error, the deviations in manufacturing process, the changes of product specs and testing methods, and the dummy variable. The overall fitness of the AIC best model is both qualitatively and quantitatively good, and so is the model's prediction.

The selection of the dummy variable points to a large difference between the data of sites A and B, so that a dummy variable is needed to bridge the data. This is evident in Fig. 4, where the average CAPA on-time completion rate for site A is visibly higher than for site B.

The biggest impact operation indicators, respectively, for sites A and B, the total CAPA, and the deviations in facility and utilities in Table 9, are both selected by the AIC best model. Among the seven correlating operation indicators in Table 6, three of them are selected by the AIC best model, namely, the deviations in facility and utilities, the ratio of employees in quality, the deviations in manufacturing process. AIC finds two new operation indicators that are even absent from the correlations listed in Table 6, namely, the deviations due to human error, and the changes of manufacturing process.

(iv) Summary

The AIC best model fitness and predictability are better for the product and continuous improvement performance indicators (the right-first-time rate, the total complaints, the complaint rate, and the CAPA on-time completion rate) than for the audit and inspection performance indicator (the deficiency rate of foreign inspections). This may be attributed to the fact that the former is more directly affected by the product quality and the effectiveness of QMS than the latter, as the outcome of a regulatory inspection is often subject to various uncertain factors. Moreover, the definition of the deficiency rate in Table 2 by itself is a rather crude characterization of an inspection outcome, as more cited deficiencies do not necessarily lead to a worse inspection classification such as an OAI classification.

Conclusion and Discussion

(i) Methodology

A four-step statistical methodology has been presented for quantitative analysis of the QMS for API manufacturing.

1. Use the Q–Q plot and the Shapiro–Wilk test to determine whether a set of data follows a normal distribution that is convenient for further statistical analysis.
2. Use the Pearson correlation coefficient to identify all the statistically significant correlations between the QMS performance indicators and operation indicators.

3. Use the stepwise multiple linear regression to determine the operation indicators with the biggest impacts to the performance indicators.
4. Use AIC to predict the most likely dependency relationship between performance and operation indicators.

Step 1 is not necessary for statistical analysis, but a normally distributed dataset allows convenient use of the statistical techniques and tools in steps 2–4. For this reason, the present work excludes all the datasets that are not normally distributed, resulting in the exclusion of 11 of the 16 performance indicators, and eight of the 25 operation indicators. The excluded data may contain useful information, but its characterization is beyond the scope of the present work.

Step 2 is perhaps the most commonly used method to study the relationships between one variable Y and a number of variables $\{X\}$. It is easy to do and produces a full account of all the 1–1 correlations, which are helpful for interpreting results in step 3. Its limit is that these correlations may be dependent, and therefore, it is difficult to quantify which X independently impacts Y by how much.

Step 3 is good in identifying which of the operation indicators have the biggest impact to a given performance indicator, but tends to produce a “winner takes all” type result. For instance, in Table 7, for site A, the impact of the deviations in manufacturing process is marginally bigger than the impact of the total deviations, but this information does not show up in Table 7. To get a more complete understanding of the stepwise linear regression result, one should take into consideration of the full list of correlations identified in step 2.

Step 4 takes a very different approach than steps 2 and 3, by focusing on making the best use of the existing data for prediction, without paying attention to the dependency of the operation indicators. The AIC method finds the best “teamwork,” a linear combination of a number of operation indicators, even though individually they may not correlate strongly to a given performance indicator. For instance, the AIC best model selects the lot acceptance rate and deficiency rate of internal audits for the total complaints, which do not even show up in Table 4.

In summary, a 4-step statistical methodology has been proposed to systematically study the quantitative relationship among the QMS indicators. While the present work focuses API manufacturing, the applicability of the proposed methodology is not limited to API manufacturing, as the methodology is not specific to product type or process step. While the present work focuses API manufacturing, the applicability of the proposed methodology is not limited to API manufacturing, as the methodology is not specific to product type or process step.

(ii) Data Quality

Judging by the rigorousness of the proposed statistical methodology and the results of correlation, stepwise linear regression, and AIC, many of which are consistent with professional common sense and among themselves, it is reasonable to conclude that the data in Table 3 from sites A and B is in good enough quality for further statistical analysis in steps 2–4. However, there are also numerous places where the results are inconsistent across sites A and B, or from one method to another. This may be caused by the poor quality of data, or by the complex nature of the relationship between the QMS performance and operation indicators, which cannot be fully characterized by the statistical methodology proposed in the present work. For the latter case, a better quantitative QMS model would be needed, as statistical analysis alone can only go so far. For the latter case, a better quantitative QMS model would be needed, as statistical analysis alone can only go so far.

If the data quality is indeed fine, it would imply that sites A and B each run a partially functional QMS that generated the data for the present work. This modest implication is consistent with the authors’ experience with the sites.

(iii) Characterization of QMS and Implications

For the five QMS performance indicators, the right-first-time rate, the total complaints, the complaint rate, the deficiency rate of foreign inspections, and the CAPA on-time completion rate, the biggest impact comes from deviations, as the selection of any stepwise linear regression and the AIC best models always includes at least one type of deviations. Among all types of deviations, the total deviations are the leading contributor, followed by the deviations of equipment and instrument, as they show up in all correlation results, all AIC selections, and some of the stepwise linear regression results. These findings are in agreement with the findings of the strong correlation ($R^2=0.96$) between the right-first-time rate and the deviations by McKinsey & Company [11].

The total deviations are a simple and broad characterization of deviations, as they are just a number count of all the deviations. An increase of this number will lower the right-first-time rate, increase the customer complaints, increase the deficiencies cited in foreign inspections, and slow the CAPA completion. These results suggest that the total deviation, using the language of quality metrics, is a leading indicator for the performance of QMS.

(iv) Predictability and Implications

The AIC model in the present work is based on the combined data from sites A and B out of the necessity for having

at least 10 years of data for the training set. The data mixing is likely to introduce artificial noise to the data, reducing predictability. For a manufacturing site that has 20 or more years of data, the AIC predictability is likely to be higher. In fact, the more data a site has the higher the predictability. This would lead to the following implications.

For a manufacturer that collects the performance and operation data year after year, by using the presented statistical methodology, it will be able to make increasingly better QMS performance predictions and better risk management, as the data accumulates. For a regulator interested in the state of quality and data-driven regulation, an increased predictability of QMS' performance will increase the visibility of a manufacturer's compliance risk.

Acknowledgements The authors thank the quality teams of sites A and B for providing data and assisting data interpretation, Xiaogang Wang, Jinglin Sun, Xi Zhang, and Jianguo Wu for helpful suggestions, and PKU-Hisun QbD Lab, PKU-Siyao GMP Lab, PKU-Hepalink Research Fund, PKU-Merck Serono Research Fund, and PKU-MSD Research Fund for support.

Funding PKU-Hisun QbD Lab, PKU-Siyao GMP Lab, PKU-Hepalink Research Fund, PKU-Merck Serono Research Fund, and PKU-MSD Research Fund. PKU (Peking University).

Declarations

Conflict of Interest The authors declared that they have no conflict of interests.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Macher J, Nickerson J. Pharmaceutical manufacturing research project. Final Benchmarking Report, Olin School of Business. 2006.
- Food and Drug Administration Drug Shortages Task Force and Strategic Plan; Request for Comments, Federal Register, Feb 12, 2013.
- ISPE Quality Metrics Initiative: A Report from the Pilot Project—Wave 1. 2015.
- ISPE Quality Metrics Initiative: A Report from the Pilot Project—Wave 2. 2016.
- Friedli T, Köhler S, Buess P, Calnan N, Basu P. Outlook—The St. Gallen pharmaceutical production system model and its contribution to the FDA quality metrics initiative. In: Friedli T, Basu P, Calnan N, Mänder C, editors. 21c Quality management in the pharmaceutical industry. Aulendorf: ECV; 2018. pp. 279–283.
- Köhler S, Friedli T. Operating context determines enabler performance relation: three distinct patterns in pharmaceutical quality control labs. *J Pharm Innov*. 2020.
- Köhler S, Friedli T, Basu P. Operational excellence in pharmaceutical quality control labs: driver of an effective quality system. *J Pharm Innov*. 2020;15:138–54.
- Friedli T, Buess P, Köhler S, Chen C, Mendivil S, Baker D. The impact of quality culture on operational performance—an empirical study from the pharmaceutical industry. *PDA JPST*. 2018;72(5):531–42.
- Friedli T, Köhler S, Buess P, Eich S, Basu P, Calnan N. FDA quality metrics research final report year 3. 2019. https://www.researchgate.net/publication/341180336_FDA_Quality_Metrics_Research_-_3rd_Year_Report Accessed Dec 2019.
- Mckinsey & Company. POBOS benchmarking tool. 2006. <https://pobos.mckinsey.com/pharma-quality>
- George K. Quality metrics: learnings from Mckinsey's "POBOS" benchmarking, measuring pharmaceutical quality through manufacturing metrics and risk based assessment. The Brookings Institution Conference. 2014. Washington DC.
- Dong K. Quality metrics as a tool to characterize quality system performance for chinese generic drug manufacturers. PhD Thesis, Peking University, 2019.
- Royston JP, Shapiro-Wilk W statistics. In Wiley StatsRef: Statistics reference online (eds N. Balakrishnan, T. Colton, B. Everitt, W. Piegorisch, F. Ruggeri and J.L. Teugels). 2014. <https://doi.org/10.1002/9781118445112.stat01826>
- Shevlyakov GL, Oja H. Robust correlation: theory and applications. 1st ed. Wiley Series in Probability and Statistics; 2016.
- Adler J, Parmryd I. Quantifying colocalization by correlation: the Pearson correlation coefficient is superior to the Mander's overlap coefficient. *Cytom Part A*. 2010;77(8):733–42.
- Streiner DL. Regression in the service of the superego: the do's and don'ts of stepwise multiple regression. *Can J Psychiatry*. 1994;39(4):191–6.
- Zitzmann S. A computationally more efficient and more accurate stepwise approach for correcting for sampling error and measurement error. *Multivar Behav Res*. 2018;53(5):612–32.
- Hauck WW, Miike R. A proposal for examining and reporting stepwise regressions. *Stat Med*. 2010;10(5):711–5.
- Burnham KP, Anderson DR. Multimodel inference: understanding AIC and BIC in model selection. *Sociolog Met Res*. 2004;33(2):261–304.
- Berry C, Chviruk B, Hernandez D, et al. What does it really look like to properly address a "human error problem" in biopharma? The human performance blue-sky description that will help improve industry performance. *PDA J Pharm Sci Technol*. 2019;73(4):391–400.
- Marill KA. Advanced statistics: linear regression, part II: multiple linear regression. *Acad Emerg Med*. 2004;11(1):94–102.
- Kim JH. Multicollinearity and misleading statistical results. *Korean J Anesthesiol*. 2019;72(6):558–69.
- Jawlik A A. r, Multiple R, R², R², R square, R² adjusted: statistics from A to Z. John Wiley & Sons, Ltd, 2016.
- Schüürmann G, Ebert R, Chen J, et al. External validation and prediction employing the predictive squared correlation coefficient — test set activity mean vs training set activity mean. *J Chem Inf Model*. 2008;48(11):2140–5.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.