Designation: ~~E2891 – 13~~ E2891 – 20

Standard Guide for Multivariate Data Analysis in Pharmaceutical Development and Manufacturing Applications[1]

This standard is issued under the fixed designation E2891; the number immediately following the designation indicates the year of original adoption or, in the case of revision, the year of last revision. A number in parentheses indicates the year of last reapproval. A superscript epsilon (ε) indicates an editorial change since the last revision or reapproval.

1. Scope

1.1 This guide covers the applications of multivariate data analysis (MVDA) to support pharmaceutical development and manufacturing activities. MVDA is one of the key enablers for process understanding and decision making in pharmaceutical development, and for the release of intermediate and final ~~products.~~ products after being validated appropriately using a science and risk-based approach.

1.2 The scope of this guide is to provide general guidelines on the application of MVDA in the pharmaceutical industry. While MVDA refers to typical empirical data analysis, the scope is limited to providing a high level guidance and not intended to provide application-specific data analysis procedures. This guide provides considerations on the following aspects:

1.2.1 Use of a risk-based approach (understanding the objective requirements and assessing the fit-for-use ~~status),~~status);

1.2.2 Considerations on the data collection and diagnostics used for MVDA (including data preprocessing and ~~outliers),~~ outliers);

1.2.3 Considerations on the different types of data ~~analysis and model validation,~~analysis, model testing, and validation;

1.2.4 Qualified and competent ~~personnel,~~personnel; and

1.2.5 Life-cycle management of ~~MVDA.~~MVDA model.

1.3 *This standard does not purport to address all of the safety concerns, if any, associated with its use. It is the responsibility of the user of this standard to establish appropriate ~~safety~~ safety, health, and ~~health~~environmental practices and determine the applicability of regulatory limitations prior to use.*

1.4 *This international standard was developed in accordance with internationally recognized principles on standardization established in the Decision on Principles for the Development of International Standards, Guides and Recommendations issued by the World Trade Organization Technical Barriers to Trade (TBT) Committee.*

2. Referenced Documents

2.1 *ASTM Standards*:[2]

C1174 Guide for Evaluation of Long-Term Behavior of Materials Used in Engineered Barrier Systems (EBS) for Geological Disposal of High-Level Radioactive Waste

E178 Practice for Dealing With Outlying Observations

E1355 Guide for Evaluating the Predictive Capability of Deterministic Fire Models

E1655 Practices for Infrared Multivariate Quantitative Analysis

E1790 Practice for Near Infrared Qualitative Analysis

E2363 Terminology Relating to Process Analytical Technology in the Pharmaceutical Industry

E2474 Practice for Pharmaceutical Process Design Utilizing Process Analytical Technology (Withdrawn 2020)[3]

E2476 Guide for Risk Assessment and Risk Control as it Impacts the Design, Development, and Operation of PAT Processes for Pharmaceutical Manufacture

E2617 Practice for Validation of Empirically Derived Multivariate Calibrations

---

[1] This guide is under the jurisdiction of ASTM Committee E55 on Manufacture of Pharmaceutical and Biopharmaceutical Products and is the direct responsibility of Subcommittee E55.01 on Process Understanding and PAT System Management, Implementation and Practice.

Current edition approved ~~Nov. 1, 2013~~July 1, 2020. Published ~~November 2013~~July 2020. Originally approved in 2013. Last previous edition approved in 2013 as E2891 – 13. DOI: ~~10.1520/E2891-13.~~10.1520/E2891-20.

[2] For referenced ASTM standards, visit the ASTM website, www.astm.org, or contact ASTM Customer Service at service@astm.org. For *Annual Book of ASTM Standards* volume information, refer to the ~~standard's~~standard's Document Summary page on the ASTM website.

[3] The last approved version of this historical standard is referenced on www.astm.org.

2.2 *ICH ~~Standards:~~Publications:*[4]
ICH Q2(R1) Validation of Analytical Procedures: Text and Methodology
ICH-Endorsed Guide for ICH Q8/Q9/Q10 Implementation ICH Quality Implementation Working Group Points to Consider (R2)
~~ICH Q2(R1) Validation of Analytical Procedures: Text and Methodology~~

3. Terminology

3.1 *Definitions*—Common term definitions can be found in Terminology E2363 for pharmaceutical applications and some terms can be found in other standards and are cited when they are mentioned.

4. Significance and Use

4.1 A significant amount of data is ~~being~~ generated during pharmaceutical development and manufacturing activities. The interpretation of such data is becoming increasingly difficult. Individual examination of the univariate process variables is relevant but can be significantly complemented by multivariate data analysis (MVDA). ~~Such methodology has been shown to be particularly efficient at handling large amounts of data from multiple sources, summarizing complex information into meaningful low dimensional graphical representations, identifying intricate correlations between multivariate datasets taking into account variable interactions. The output from MVDA will generate useful information that can be used to~~ MVDA may be particularly appropriate for exploring and handling large sets of heterogenous data, mapping data of high dimensionality onto lower dimensional representations, exposing significant correlations among multivariate variables within a single data set or significant correlations among multivariate variables across data sets. MVDA may extract statistically significant information which may enhance process understanding, decision making in process development, process monitoring and control (including product release), product life-cycle management, and ~~continual~~continuous improvement.

4.2 MVDA is ~~a~~ widely used ~~tool~~ in various industries including the pharmaceutical industry. To ~~generate~~achieve a valid outcome, an MVDA model/application should ~~contain~~incorporate the ~~following components:~~ following:

4.2.1 A predefined ~~objective based on a risk and scientific hypothesis~~risk-based objective incorporating one or more relevant scientific hypotheses specific to the ~~application,~~application;

4.2.2 ~~Relevant data,~~Sufficient relevant data of requisite quality covering the variance space encountered during intended use, that is, pharmaceutical development, or pharmaceutical manufacturing, or both;

4.2.3 Appropriate data analysis ~~techniques,~~ and model utilization practices including considerations on ~~validation,~~testing, validation, and qualification of all new data prior to using a model to analyze it;

4.2.4 Appropriately trained ~~staff, and~~staff;

4.2.5 Appropriate standard operating procedures; and

4.2.6 Life-cycle management.

4.3 This guide can be used to support data analysis activities associated with pharmaceutical development and manufacturing, process performance and product quality monitoring in manufacturing, as well as for troubleshooting and investigation events. Technical details in data analysis can be found in the scientific literature and standard practices in data analysis are already available (such as Practices E1655 and E1790 for spectroscopic applications, Practice E2617 for model validation, and Practice E2474 for utilizing process analytical technology).

5. Risk-Based Approach for MVDA

5.1 A risk-based approach requires consideration of two aspects: the risk associated with the use of MVDA for a specific objective and the justifications and rationales during the data analysis to ensure the model is fit for use. Aspects of general risk assessment and control are described in Guide E2476 and more specific model considerations are discussed in ICH-Endorsed Guide for ICH Q8/Q9/Q10 Implementation.

5.2 The risk level is considered high when the data analysis is an integral part of the control strategy, is used directly for the product or intermediate product release or is used to directly control the process. The risk is considered low when the output of the data analysis does not have significant impact on the assessment of the product quality.

5.3 In assessment of fitness for use of data analysis, several aspects should be considered:

5.3.1 *Criteria for Acceptable Data Analysis*—Criteria for the data analysis are defined by user requirements and project objectives.

5.3.2 *Data Source*—Relevant data should be collected and used in MVDA.

5.3.3 *Data Integrity*—Confirmation of accuracy, consistency, and traceability of the data from the source to the analysis.

5.3.4 *Data Analysis Practice (Technique and Procedure)*—In data analysis practice, numerous options are available and different options may generate similar results, all of which may be deemed fit for use. The data analysis process is an iterative

[4] Available from International ~~Conference on~~Council for Harmonisation of Technical Requirements for ~~Registration of~~ Pharmaceuticals for Human Use (ICH), ICH Secretariat, ~~c/o IFPMA, 15 ch. Louis-Dunant, P.O. Box 195, 1211 Geneva 20, Switzerland, http://www.ich.org.~~Route de Pré-Bois, 20, P.O Box 1894, 1215 Geneva, Switzerland, https://www.ich.org.

approach; in case of an unsatisfactory result, a different data analysis technique may be used or it may be necessary to obtain additional data or data of higher quality, or both, until a valid model can be established which is deemed fit for use.

6. Concepts of MVDA Model and MVDA Method

6.1 When implementing MVDA it is important to understand the differentiation between a multivariate model and a multivariate method. This is especially true as an MVDA application reaches the validation stage.

6.2 *MVDA Model:*

6.2.1 As defined in ~~Practice~~Guide C1174, a model is a simplified representation of a system or phenomenon with multiple variables based on a set of hypotheses (assumptions, data, simplifications, or idealizations, or ~~a~~ combinations thereof) that describe the system or explain the phenomenon, often expressed mathematically. In the context of this guidance the term MVDA model is to be taken in a broad sense covering, ~~for example~~ multivariate exploratory, regression as well as ~~latent variable-based techniques—such~~ dimension reduction techniques — such as, but not limited ~~to, Principal Component Analysis (PCA) and Partial Least Squares (PLS) Regression.~~to latent variable-based, principal component analysis (PCA), principal component regression (PCR) and partial least-squares (PLS) regression. These models often relate observational data to a known property or set of properties from a ~~process.~~ process, or a summarized measure of the process state that can be used for statistical process control (SPC) approach, as described in 8.3. The mathematical relationship is established for a sufficient number of ~~cases—preferably~~ cases — preferably derived from experimental designs. The model can then be applied to a similar set of observational data in order to ~~predict~~estimate the targeted property/properties.

6.2.2 MVDA is not limited to such multivariate calibrations and predictions, and similar considerations as the ones described in this guidance are applicable to direct and indirect calibration, as well as PCA-based approaches used for example for exploratory data analysis.

6.3 ~~*MVDA Method:*~~*Analytical and Process Control Methods, Including One or More MVDA Elements:*

6.3.1 The MVDA method uses the output from the MVDA model to define the targeted and predefined process characteristic of interest. The MVDA model is one component of the broader concept that is an MVDA method. Such method should typically be characterized by the collection of data, the input data to the calculation, the data analysis, and some potential transformation from the MVDA model output to generate the pre-defined MVDA method characteristic of interest. (See Fig. 1.)

6.3.2 Note that an MVDA method can incorporate multiple MVDA models (for example, across multiple unit operations, from multiple pieces of equipment, etc.) that can be running in parallel or feeding sequentially into one another to provide the pre-defined MVDA method output. MVDA methods can also incorporate mechanistic and univariate models. The validation of the MVDA model and the MVDA method are two different activities. Section 97 of this guideline provides an overview of the MVDA model validation. The validation of an MVDA method should follow the same overarching principles as for any method validation, such as the ones described in ICH Q2(R1).

6.3.3 Method development comprises the creation of a model, its testing, and validation

6.3.4 Method validation consists in the validation of not only the model but also all the aspects of data acquisition, analysis and reporting outlined in Fig. 1. The level of method validation will depend on the intended model impact as defined in ICH-Endorsed Guide for ICH Q8/Q9/Q10 Implementation. A low impact model will require a fit for purpose model calibration, testing and validation but a lower consideration for method validation. A medium or high impact model will require a higher consideration for method validation.

6.4 *Two-Phase Nature of MVDA:*

6.4.1 Data analysis usually, but not always, has two phases. ~~In predictive analysis, the~~ The first phase is the creation of a model from acquired data with a corresponding known property, and the second phase is the application of the model to newly acquired independent data to ~~predict~~estimate a value of the property. The first-phase analysis is usually called a multivariate calibration for

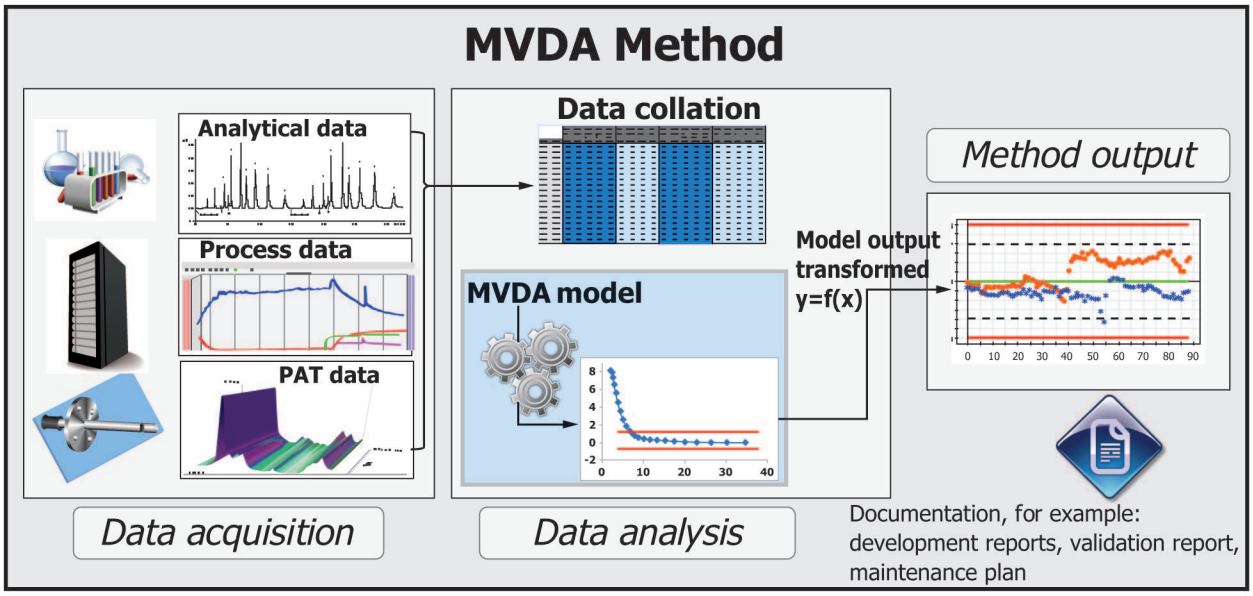

FIG. 1 Relationship Between an MVDA Method and an MVDA Model

a regression process or training for a learning process. The emphasis is usually on the model building phase in practice: how to design cases properly, how to process the data to build a model, and how to test the model to see whether the model is fit for use. The model prediction phase, however, should be emphasized equally. A valid model ~~does not always generate a valid result; it~~ will generate a valid result only if the input ~~data~~ is valid too. It is important to screen the input data and monitor the prediction diagnostics when using the model for prediction. ~~Such~~ For latent variable-based models, such diagnostics are often referred to as residual and score space diagnostics or inner/outer model ~~diagnostics.~~diagnostics (see Section 7). In addition, a strategy for life-cycle management of the MVDA method is required (see Section 11).

6.4.2 In ~~tracking and trending~~ process monitoring analysis, the first phase is to establish data analysis parameters, trending limits, or a criterion for the end point of trajectory ~~tracking.~~monitoring. A model may be created in the first-phase trending analysis. The second phase is ~~predicting~~estimating the new values based on the established parameter set (including a possible model) and assessing the trajectory based on the established criteria.

~~6. Risk-Based Approach for MVDA~~

~~6.1 A risk-based approach requires consideration of two aspects: the risk associated with the use of MVDA for a specific objective and the justifications and rationales during the data analysis to ensure the model is fit for use. Aspects of general risk assessment and control are described in Practice E2476 and more specific model considerations are discussed in ICH-Endorsed Guide for ICH Q8/Q9/Q10 Implementation.~~

~~6.2 The risk level is considered high when the data analysis is an integral part of the control strategy, is used directly for the product or intermediate product release, or is used to directly control the process. The risk is considered low when the output of the data analysis does not have significant impact on the product quality or the assessment of the product quality.~~

~~6.3 In assessment of fitness for use of data analysis, three aspects should be considered:~~

~~6.3.1 *Criteria for Acceptable Data Analysis*—Criteria for the data analysis are defined by user requirements and project objectives.~~

~~6.3.2 *Data Source*—Appropriate and relevant data should be collected and used in MVDA.~~

~~6.3.3 *Data Analysis Practice (Technique and Procedure)*—In data analysis practice, numerous options are available and different options may generate similar results, all of which may be deemed fit for use. The data analysis process is an iterative approach; in case of an unsatisfactory result, a different data analysis technique may be used or it may be necessary to obtain additional data and/or data of higher quality.~~

7. Data Collection and Diagnostics

7.1 Relevant data properly representing all factors impacting the MVDA objective should be used for data analysis. Data gathered from various sources should be screened for errors, appropriate data preprocessing should be used, and data should be screened for ~~outliers.~~ outliers and for irrelevant or accidental correlations which may confound attempts to find exploitable correlations. All processing of data, exclusion of outliers, selection of samples or variables, or both, and other analysis parameters need to be justified and documented.

7.2 *Data Source:*

7.2.1 Data can be continuous, discrete, or categorical and from multiple sources. The most common sources are input/raw material properties, process parameters, *in situ*/PAT data and intermediate/finished product properties. Data should be gathered with acceptable quality (free of any obvious human or machine errors but properly representing a typical noise level likely to be present in such data), with appropriate significant figures. Outlier detection is strongly recommended (see 7.4).

7.2.2 Depending on the MVDA-defined objective, the data ~~could~~can come from designed experiments (DOE) specifically for developing the MVDA model, or from ~~routine~~other development runs and ~~manufacturing processes,~~routine manufacturing, or both. Data originating from a DOE on input/process parameters has inherent variation (special cause variability), while data obtained from routine operations may reflect smaller variation within the acceptable operational ranges, tighter than ranges studied during process development (common cause variability). The data collected from a routine process may be used for trending, process monitoring, identification of atypical behavior but rarely for predictive regression-based multivariate analysis. A predictive empirical model built from the data that has small variation will typically have a very small range limited by the combination of specification, constrained incoming material variation and routine process parameter variation (operating ranges). Model diagnostics should be used to ensure the model is ~~predicting a meaningful result.~~ providing reliable outputs. Often, intentionally induced variations, preferably following a DOE, are created so that the data with a larger variation range and non-confounded process conditions is used as part of the training set to build the model.

7.2.3 ~~Data can be continuous, discrete, or categorical and from multiple sources. The most common sources are input/raw material properties, process parameters, in situ/PAT data and intermediate/finished product properties. Data should be gathered with acceptable quality (free of any obvious human or machine errors but properly representing a typical noise level likely to be present in such data), with appropriate significant figures. Outlier detection is strongly recommended (see~~The manufacturing scale at which data collection is performed should be carefully considered. While collecting full-scale manufacturing data at target conditions is necessary, it is often prohibitive (time and cost) to perform DOEs at full-scale. Small scale DOEs are often preferred. However, if the MVDA model is to be used as part of the control strategy or the final release of the commercial manufacturing scale, or both,

the commercial manufacturing scale should be considered during the MVDA data collection phase. When appropriate and the MVDA model is to be used at full-scale manufacturing, the relevance and equivalence of the data collected at other scales than full-scale (scale, operation, raw materials, process dynamics, impact on the signal, batch run time, etc.) should be discussed and documented to support the overall data collection strategy, such that the impact of manufacturing process scale on the MVDA model's ~~7.4).~~applicability can be addressed in a rational manner.

7.2.4 Data review is highly recommended and should be aligned with the risk level identified for the MVDA activity. Appropriate documentation of the data review activity should be available as part of the model ~~development~~ development, model validation, and model maintenance activities.

7.3 *Data Preprocessing:*

7.3.1 Data preprocessing (or pretreatment) ~~is~~ can be a critical step in the implementation of any MVDA ~~application.~~ applications, especially for chemometric datasets involving spectroscopic data. The approach chosen in the preprocessing of the data may have a significant impact on the output of the multivariate analysis and should be considered carefully. The preprocessing of the data should aim at reshaping the data structure to enhance the key features targeted by the MVDA objectives. Appropriate data preprocessing depends on the nature of the data, the MVDA technique used, and the purpose of the data analysis. Multiple preprocessing steps, or chained preprocessing, can sometimes be applied to achieve the desired objective but should be considered carefully, particularly as the order chosen for the individual preprocessing steps is likely to have significant impact on the data analysis outcome. It may take several iterative cycles to optimize the preprocessing steps to ensure the necessary, yet sufficient level of preprocessing is applied to the data set to enable the MVDA model objectives to be achieved.

7.3.2 Even though preprocessing can reduce or eliminate some unwanted variations in data, this must not be aimed to transform data that are not fit for purpose (for example, measurement errors) into usable data. If the data is unusable, a new data collection step should be considered to improve the quality of the data.

7.4 *Outliers:*

7.4.1 An outlier means an outlying observation that appears to deviate markedly in value from other members of the sample in which it appears (Practice E178). Outliers typically originate from either a measurement error (clerical, sampling, sensor) or a process error (process ~~deviation).~~deviation), or from extreme samples outside the model's space.

7.4.2 *Outliers in Model Building Phase:*

7.4.2.1 The purpose in identifying outliers in the model building phase or data exploration phase is to ensure that the model is not distorted by the inclusion of a few non-representative data points. Justification and documentation of assignable cause to suspected outliers is recommended prior to the removal of any point in the dataset. There are a variety of statistical tools and visualization techniques (such as Hotelling's ~~T2~~T^2 plots, histograms, distance to model plots, control charts, cluster analysis) available to the MVDA practitioners, but one must recognize that the knowledge of the process and the measurement is fundamental to make a decision on excluding a sample from the set of analysis.

7.4.2.2 Caution should be exercised when an outlier is to be removed from the data set. Potential outliers do not have to be removed automatically in the model building phase. Excessive sample removal may reduce the model space and sacrifice the robustness of the model. Some outlying observations may be due to the inherent true variability of the data or process, or both, therefore the decision to remove any outliers should always be based on a thorough data review, justified using expert knowledge of the data or process, or both, and documented.

7.4.2.3 In ~~latent variable based analysis,~~ dimensionality reduction techniques, the common outlier detection approach is based on residual space and score space diagnostics. The criteria or limits for the diagnostics should be established from the data set used to build the model so that each ~~prediction~~estimation made by the model is evaluated against such model diagnostics. Thus, it is critical that model building phase data are representative of future process variations (8.2).

7.4.3 *Outliers in Model Prediction ~~Phase— Diagnostics:~~ Phase—Diagnostics:*

7.4.3.1 The purpose in identifying outliers in the MVDA model ~~prediction~~estimation phase is to ensure that the data in the ~~prediction~~estimation set is comparable to the data used in the model building phase. The ~~prediction~~estimation set should be within the model boundaries established while testing or validating the model, so that the prediction set can be reliably used by the model.

7.4.3.2 It is recommended for latent variable-based models to use both inner and outer model diagnostics to provide assurance that any new sample being ~~predicted~~estimated is adequately represented in the ~~calibration and validation sets and~~ data which was used to validate the model and not an outlier. This will provide assurance that the model is relevant for the predicted value and the prediction is reliable and valid.

7.4.3.3 Note that there are robust data analysis approaches that have high resistance to outliers. The resultant model may be robust in terms of attenuating outlier influence to the creation of the model. However, when the model is used to predict new data, appropriate techniques are still needed to ensure the outliers are screened and the model ~~results~~outputs are reliable.

7.4.3.4 Outliers and out of specifications (OOS). An outlier observed during data analysis means that ~~the predictive model used will produce an invalid result,~~ an invalid output is produced by the multivariate model because the sample is outside the model range, but it does not give a reliable indication of an OOS result. OOS results are obtained when a multivariate model output is outside the acceptance criteria, but the model diagnostics are within acceptable predefined limits. In multivariate identification (qualitative) models, ~~nonconforming results (not positively identified as any entry defined in the model)~~observations not attributed to one of the classes used during model development should be treated as outliers.

8. Data Analysis Process

8.1 *Exploratory Analysis:*

8.1.1 Exploratory data analysis should be implemented when initiating an MVDA project to review the available data set. Exploratory analysis is intended to reveal the structure or patterns in data and can aid in finding the relationship between the measured data and quality attributes, and establishing the preferred algorithmic methods to be applied to the measured data and outlier detection.

8.1.2 The correlation within data ~~set~~sets does not necessarily imply causation when MVDA is used to explore the ~~relationship~~ empirical relationships in the data. In a multivariate model, even though the cause-effect relationship may be revealed, necessary knowledge of physics, chemistry and engineering should be applied, and consideration should be given to demonstrate or confirm a causal relationship by running structured or designed experimentation. Whenever applicable, the underlying mechanisms of the process being investigated should be taken into consideration so as to generate an understanding of the expected dynamics (for example, relationships between inputs and outputs and the expected MVDA model estimations) for improved process control and product quality.

8.2 *Data Modeling:*

8.2.1 The data for model calibration should cover sufficient variations that might be encountered during implementation to ensure model robustness. The sources of calibration data as stated in 7.2 could be a ~~DOE or process data.~~ DOE, process data or a combination of both. Risk assessment, appropriately documented, should be performed to determine the variables and ranges that should be incorporated in the model. Small scale experiments can be used to confirm risk-based variable selection and identify non-relevant variables. When variables selected from risk-based small scale are applied to full-scale manufacturing, certain adjustments may be needed to address the effect of scale depending on the nature of the variables, the scale change, and the manufacturing process unit operation involved.

8.2.2 Data for ~~calibration~~ calibration, test, and validation should be collected in a manner assuring it is comparable to the data collected during routine use.

8.2.3 The modeling approach including associated data preprocessing should be justified and documented with sufficient details so that the data analysis is readily repeatable. The justification can be based on the knowledge of the MVDA practitioner or an optimization approach.