Designation: E2891 – 13

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.

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),

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

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

1.2.4 Qualified and competent personnel, and

1.2.5 Life-cycle management of MVDA.

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 and health practices and determine the applicability of regulatory limitations prior to use.*

2. Referenced Documents

2.1 *ASTM Standards:*[2]

C1174 Practice for Prediction of the Long-Term Behavior of Materials, Including Waste Forms, 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

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

2.2 *ICH Standards:*[3]

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

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

Current edition approved Nov. 1, 2013. Published November 2013. DOI: 10.1520/E2891-13.

[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 Document Summary page on the ASTM website.

[3] Available from International Conference on 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.

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 enhance process understanding, decision making in process development, process monitoring and control (including product release), product life-cycle management and continual improvement.

4.2 MVDA is a widely used tool in various industries including the pharmaceutical industry. To generate a valid outcome, MVDA should contain the following components:

4.2.1 A predefined objective based on a risk and scientific hypothesis specific to the application,

4.2.2 Relevant data,

4.2.3 Appropriate data analysis techniques, including considerations on validation,

4.2.4 Appropriately trained staff, and

4.2.5 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 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. Concepts of MVDA Model and MVDA Method

5.1 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.

5.2 *MVDA Model:*

5.2.1 As defined in Practice 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 regression as well as latent variable-based techniques—such as, but not limited to, Principal Component Analysis (PCA) and Partial Least Squares (PLS) Regression. These models often relate observational data to a known property or set of properties from a process. The mathematical relationship is established for a sufficient number of cases—preferably derived from experimental designs. The model can then be applied to a similar set of observational data in order to predict the targeted property/properties.

5.2.2 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.

5.3 *MVDA Method:*

5.3.1 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.)

5.3.2 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

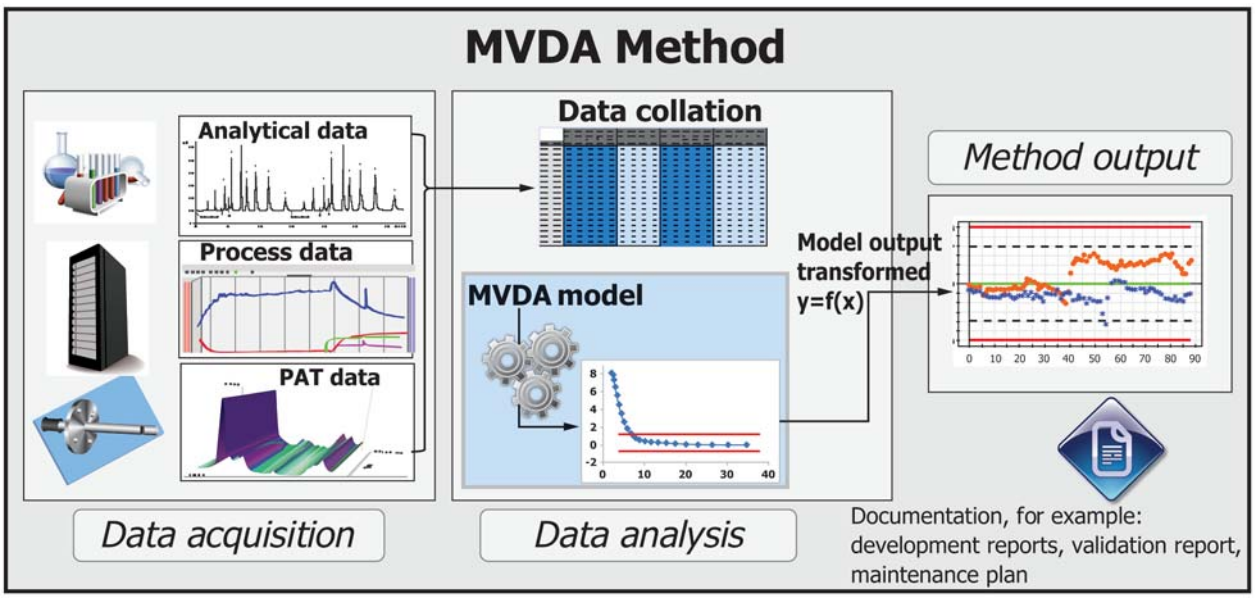

FIG. 1 Relationship Between an MVDA Method and an MVDA Model

pre-defined MVDA method output. The validation of the MVDA model and the MVDA method are two different activities. Section 9 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).

5.4 *Two-Phase Nature of MVDA:*

5.4.1 Data analysis usually, but not always, has two phases. In predictive analysis, 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 data to predict a value of the property. The first-phase analysis is usually called a multivariate calibration for 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 data and monitor the prediction diagnostics when using the model for prediction. Such diagnostics are often referred to as residual and score space diagnostics or inner/outer model diagnostics.

5.4.2 In tracking and trending analysis, the first phase is to establish data analysis parameters, trending limits, or a criterion for the end point of trajectory tracking. A model may be created in the first-phase trending analysis. The second phase is predicting 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. 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 Depending on the MVDA-defined objective, the data could come from designed experiments (DOE) or from routine development and manufacturing processes, 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 analysis. A predictive 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. Often, intentionally induced variations, preferably following a DOE, are created so that the data with a larger variation range is used as part of the training set to build the model.

7.2.2 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).

7.2.3 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 and model maintenance activities.

7.3 *Data Preprocessing:*

7.3.1 Data preprocessing (or pretreatment) is a critical step in the implementation of any MVDA application. 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).

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 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, 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 made by the model is evaluated against such model diagnostics.

7.4.3 *Outliers in Model Prediction Phase— Diagnostics:*

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

7.4.3.2 It is recommended to use both inner and outer model diagnostics to provide assurance that any new sample being predicted is adequately represented in the calibration and validation sets 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 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 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, 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) 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 does not necessarily imply causation when MVDA is used to explore the relationship 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.

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. Risk assessment, appropriately documented, should be performed to determine the variables and ranges that should be incorporated in the model.

8.2.2 Data for calibration 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.

8.3 *Trending Analysis for Process Monitoring or Control, or Both:*

8.3.1 The data used to establish the trending criteria should have common cause variations only. The purpose of trending analysis is to verify that the process is performing consistently but also to identify special cause variations (process deviations) and process shifts (for example, upwards or downward trends, changes in common cause variability).

8.3.2 The multivariate approach should be used for data containing multiple inputs with some level of collinearity and potential interaction terms. A multivariate approach to trending allows the detection of significant changes in the covariance structure of the data, and in turn leads to early fault detection and increased process understanding. Combining multiple entries into a summarized MVDA output will provide an overall review of the process, complementary to the individual univariate output.

9. Model Validation

9.1 The term validation in this guideline refers to the validation of the multivariate model (Section 5). Model validation is the process of determining the degree to which a calculation method is an accurate representation of the real world from the perspective of the intended uses of the calculation method. The fundamental strategy of validation is the identification and quantification of error and uncertainty in the conceptual and computational models with respect to intended uses (Guide E1355).

9.2 The extent of validation should be based on the objective and risk of the application. If MVDA is used as part of the control strategy, the level of validation has to reflect the potential risk associated with the application and is expected to be significant.

9.3 The validation of a model emphasizes the uncertainty and predictability of the model. Once the model is considered acceptable and it is used in an MVDA method, it may be subject to method validation.

9.4 Model internal validation can provide useful information if used properly but the validation of a model should not be solely based on the result of internal validation. Internal validation can be used to assess the appropriate parameters of the MVDA model, such as the model rank and provide a first estimate of a model's predictive ability. Internal validation can be performed by using an internal validation data set, cross validation or statistical resampling with the calibration data set.

9.5 The use of an external independent test set is strongly recommended whenever possible as part of the model validation process. Model external validation refers to using truly independent data from the one used to generate the model, to perform the validation. Truly independent data means the data is not in any way used in the model creation. For example, the internal validating data used to select an optimum model is not independent, nor are replicate data from the same sample from which the calibration or training data was collected.

9.6 The success of an external model validation is based on the performance of the model and assessed against acceptance criteria (fitness for use). The model performance is typically expressed as overall error, and specific statistics should also be considered for specific applications in quantitative models and qualitative models (Practice E2617).

9.7 For the exploratory data analysis that does not create a predictive model, some level of validation of the data analysis is needed to ensure the interpretation of the analysis is valid and consistent. Because of the open-minded nature of data exploration, the validation of the exploratory model is usually less rigorous compared to that of a predictive model. To ensure the validity of data analysis, the data should be screened carefully, and the data analysis should be conducted by an experienced user with necessary knowledge. External test set validation may be used to verify that the data analysis conclusions are consistent from multiple similar data subsets.

10. Subject Matter Expert (SME)

10.1 Qualified and competent personnel defined as SME should perform the MVDA and be responsible for the MVDA outcome. An SME may be an individual or a team with expertise in chemometrics, statistics, chemistry, and engineering, and is capable of understanding the nature of the data, selecting an appropriate data analysis technique, using relevant software, interpreting the result, and documenting the outcome.

10.2 The knowledge and skills of data analysis should not be underestimated even with the advance of software in which some data analysis activity becomes procedure driven.

11. Life-Cycle Management of MVDA

11.1 Life-cycle management of the MVDA application assures the consistency and validity of a data analysis approach throughout the life cycle of the product. Multivariate models require periodical review and potentially updates.

11.2 *Criticality of the Maintenance Activities:*

11.2.1 A strategy on the maintenance of the MVDA application should be considered as a core component of the overall activities, together with the establishment of associated standard procedures, as appropriate to the phase of the product life cycle and use of the MVDA application. The level of maintenance should be considered as part of the risk assessment activities and be fit for the purpose of the MVDA application. Depending on the criticality of the MVDA activities, the level of maintenance will vary, and an application supporting critical-to-quality decisions on attribute prediction or parameter controls will require a more stringent maintenance strategy (typically linked to change control procedures and documented as part of the quality management system) than an application used for information only.

11.2.2 All maintenance activities should be clearly documented to provide full traceability of any adjustment made, be it on the model itself, the data sources, or the control component of the MVDA application. The maintenance activities on MVDA should be an integral part of the existing change control procedures.

11.3 *Technical Considerations on Maintenance Activities:*

11.3.1 The multivariate model is subject to a review based either on a predefined time frequency or events such as changes

of raw-materials variability or manufacturer, changes in the upstream process, a drift in model prediction, or an out of specification result in the model output. The outcome of the review or investigation governs the need for model adjustment. Maintenance should be based on the risk associated with the application objective, and the triggers for maintenance should be considered as part of a risk assessment.

11.3.2 The maintenance of the MVDA model should not only consider the prediction output of the model but also include the model diagnostics, with acceptance criteria determined during validation. Periodical comparison of multivariate model predictions and reference method measurements is not always necessary but recommended. It should be done when there is a reason to suspect the model performance is deviating from the expected results.

11.3.3 It may not always be required to rebuild an MVDA model from scratch and adjustment of the training set (addition or removal of samples), together with a thorough review of the modeling conditions, can be considered, as long as it is scientifically justified and properly documented.

11.4 *Data Management*—Assurance of data integrity and corresponding documentation should be commensurate with the level of criticality of the application.

12. Keywords

12.1 chemometrics; chemometric model; data analysis method validation; data model; exploratory data analysis; model building; model validation; multivariate data analysis; outliers; predictive data analysis; process analytical technology