Designation: 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 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);

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, model testing, and validation;

1.2.4 Qualified and competent personnel; and

1.2.5 Life-cycle management of 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, health, and 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

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 Publications:*[3]

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)

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 generated during pharmaceutical development and manufacturing activities. The

[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 July 1, 2020. Published July 2020. Originally approved in 2013. Last previous edition approved in 2013 as E2891 – 13. DOI: 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 Document Summary page on the ASTM website.

[3] Available from International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH), ICH Secretariat, Route de Pré-Bois, 20, P.O Box 1894, 1215 Geneva, Switzerland, https://www.ich.org.

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). 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 continuous improvement.

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

4.2.1 A predefined risk-based objective incorporating one or more relevant scientific hypotheses specific to the application;

4.2.2 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 and model utilization practices including considerations on testing, validation, and qualification of all new data prior to using a model to analyze it;

4.2.4 Appropriately trained 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 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 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 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, multivariate exploratory, regression as well as dimension reduction techniques — such as, but not limited 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, 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 derived from experimental designs. The model can then be applied to a similar set of observational data in order to 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 *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 7 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. 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 estimate 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 will generate a valid result only if the input is valid too. It is important to screen the input data and monitor the prediction diagnostics when using the model for prediction. For latent variable-based models, such diagnostics are often referred to as residual and score space diagnostics or inner/outer model 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 process monitoring analysis, the first phase is to establish data analysis parameters, trending limits, or a criterion for the end point of trajectory monitoring. A model may be created in the first-phase trending analysis. The second phase is estimating the new values based on the established parameter set (including a possible model) and assessing the trajectory based on the established criteria.

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 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

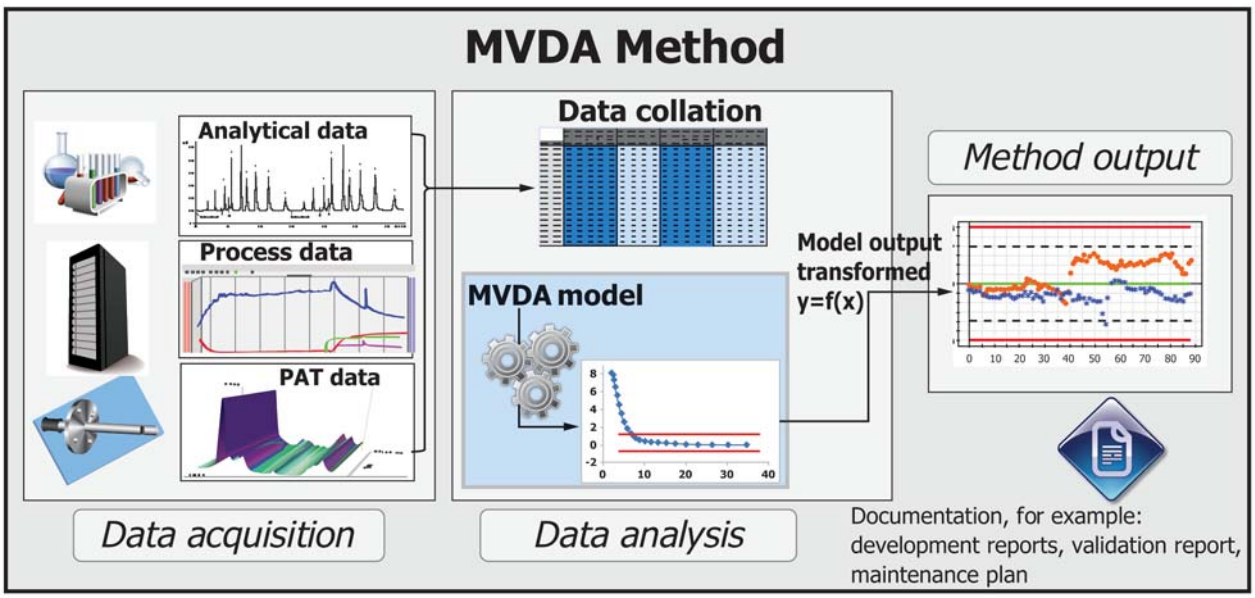

FIG. 1 Relationship Between an MVDA Method and an MVDA Model