2017

# Evaluating Current Practices in Shelf Life Estimation

Robert Capen
*Merck & Co. Inc.*, robert_capen@merck.com

David Christopher
*CMC Statistics*

Patrick Forenzo
*Analytical Research and Development*

Kim Huynh-Ba
*Pharmalytik Consulting and Training Services*

David LeBlond
*Private Statistical Consultant*

***See next page for additional authors***

*AAPS PharmSciTech* (© 2017)
DOI: 10.1208/s12249-017-0880-4


*Research Article*

# Evaluating Current Practices in Shelf Life Estimation

**Robert Capen,[1,2,13] David Christopher,[1] Patrick Forenzo,[3] Kim Huynh-Ba,[4] David LeBlond,[5] Oscar Liu,[6] John O'Neill,[7] Nate Patterson,[8] Michelle Quinlan,[9] Radhika Rajagopalan,[10] James Schwenke,[11] and Walter Stroup[12]**

*Received 17 June 2017; accepted 13 September 2017*

***Abstract.*** The current International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH) methods for determining the supported shelf life of a drug product, described in ICH guidance documents Q1A and Q1E, are evaluated in this paper. To support this evaluation, an industry data set is used which is comprised of 26 individual stability batches of a common drug product where most batches are measured over a 24 month storage period. Using randomly sampled sets of 3 or 6 batches from the industry data set, the current ICH methods are assessed from three perspectives. First, the distributional properties of the supported shelf lives are summarized and compared to the distributional properties of the true shelf lives associated with the industry data set, assuming the industry data set represents a finite population of drug product batches for discussion purposes. Second, the results of the ICH "poolability" tests for model selection are summarized and the separate shelf life distributions from the possible alternative models are compared. Finally, the ICH methods are evaluated in terms of their ability to manage risk. Shelf life estimates that are too long result in an unacceptable percentage of nonconforming batches at expiry while those that are too short put the manufacturer at risk of possibly having to prematurely discard safe and efficacious drug product. Based on the analysis of the industry data set, the ICH-recommended approach did not produce supported shelf lives that effectively managed risk. Alternative approaches are required.

**KEY WORDS:** stability; shelf life estimation; FDA; ICH Q1A/Q1E; managing risk.

## INTRODUCTION

### Overview

In 2006, the Product Quality Research Institute (PQRI) established a Stability Shelf Life Working Group (referred to as the "Working Group" in this article) with the mandate to investigate current statistical methods for estimating shelf life based on stability data. The Working Group is composed of pharmaceutical, regulatory, and statistical scientists from industry, government, and academia. As one of its first actions, the Working Group reviewed available literature and applicable guidance documents, and discussed current industry and regulatory practices related to determining the shelf life for pharmaceutical products. Different issues with the current practices were discussed along with possible statistical approaches to resolve them. The Working Group engaged in discussions to review and summarize available descriptions of shelf life, evaluating their benefits, drawbacks, and consequences to better target the appropriate research questions for statistical discussions. Key results from these discussions are published in the Working Group's first paper (1). The concepts and terminology set forth in that paper are used throughout this review. In particular, the product shelf life is the true but unknown limit on the period of storage time during which the pharmaceutical or drug product remains within specifications and is therefore considered effective and fit for use. Any suitably conservative estimate of the product shelf life, as supported by statistical methods, is called the supported shelf life.

This second paper is a report on the continuation of those discussions which again are meant to raise public

[1] CMC Statistics, Merck, West Point, Pennsylvania, USA.
[2] Merck & Co. Inc., WP 35-240, 770 Sumneytown Pike, West Point, Pennsylvania 19486, USA.
[3] Analytical Research and Development, Novartis, East Hanover, New Jersey, USA.
[4] Pharmalytik Consulting and Training Services, Newark, Delaware, USA.
[5] Private Statistical Consultant, Wadsworth, Illinois, USA.
[6] Research and Development, Insys Therapeutics, Chandler, Arizona, USA.
[7] Nagano Science, Albany, New York, USA.
[8] Bayer U.S., Berkeley, California, USA.
[9] Clinical Pharmacology Biostatistics, Novartis Oncology, East Hanover, New Jersey, USA.
[10] FDA/CDER/OPQ/OLDP/Quality Assessment Lead, Silver Springs, Maryland, USA.
[11] Applied Research Consultants, New Milford, Connecticut, USA.
[12] Statistics, University of Nebraska-Lincoln, Lincoln, Nebraska, USA.
[13] To whom correspondence should be addressed. (e-mail: robert_capen@merck.com)

Published online: 25 September 2017

awareness of the existing different interpretations of shelf life and to stimulate a broader public discussion on these topics, which are relevant for drug products, drug substances, clinical supplies, *etc*. In this process, the Working Group has considered existing guidelines but sometimes taken the liberty to question elements of these for the purpose of potentially developing an alternative methodology.

The International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH) brings together regulatory authorities and the pharmaceutical industry to discuss scientific and technical aspects of drug registration. For stability testing, ICH Q1A (2) specifies that the supported shelf life for a drug product be set as the storage time during which drug batches are expected to remain within specification. Similarly, ICH Q1E (3) notes that the strategy for estimating the supported shelf life of a drug product is to determine the storage time during which the critical attributes of the drug product remain acceptable for all future batches, manufactured, packaged, and stored under similar conditions. Both guidance documents highlight the batch as the focal point of interest. The current statistical methods for determining a supported shelf life outlined in ICH Q1A and Q1E assume that a simple linear regression model is adequate to describe the observed data obtained from a stability study. In addition, the ICH guidance documents allow other statistical models to be used, *e.g*., a quadratic polynomial or nonlinear model, when appropriate.

The ICH statistical methods for estimating the shelf life have long been the subject of discussions, publications, and conference presentations, in which the Working Group has participated. However, to the Working Group's knowledge, Kiermeier *et al.* (4) and Quinlan *et al*., (5) using simulated data, provide the first review of the results of these ICH methods. The current paper continues the review of these methods through the analysis of a real-life industry data set provided to the Working Group by a PQRI-member pharmaceutical company. This industry data set is used for all discussions, examples, and computations and is comprised of 26 individual stability batches with most batches having 24 months of stability data. Each batch represents the results of a stability-limiting product characteristic; here, percent assay or potency measured over storage time. As in the ICH guidance documents, a simple linear regression model is assumed and is appropriate for the industry data set. While our focus is on the ICH methods for estimating shelf life, the statistical summaries and figures described herein, as well as the emphasis on risk, provide a novel framework for determining if a particular method produces a "good" estimate of the true product shelf life. This will be further explored in a future paper.

The primary purpose of this paper is to evaluate the ICH methods for determining the supported shelf life in terms of their distributional properties and in terms of the proportion of batches conforming to specifications at the supported shelf life. To do so, an evaluation of the ICH stability shelf life estimation methods is considered from three perspectives to highlight different aspects of the estimation methods. First, the industry data set is used assuming the 26 individual stability batches represent a common pharmaceutical product from a production process under control. Here, for the purpose of discussing and evaluating the ICH methods, the 26 stability batches of the industry data set are treated as if they represent the entire population of stability batches for the pharmaceutical product. Second, a model selection procedure is outlined in the ICH guidance documents, commonly referred to as the "poolability" tests, and is a major component to the ICH methods for estimating shelf life. To evaluate the poolability testing process in terms of the ICH methods, comparisons are made among the supported shelf life distributions for each of the variations of the simple linear regression model. Third, as stated in both ICH Q1A and Q1E, a supported shelf life of a drug product must define the storage time during which drug product batches are expected to remain within specifications. This storage time is applicable to all future batches that are manufactured, packaged, and stored under similar conditions. While not mentioned as such in these guidance documents, this defines a type of quality statement. Using the industry data set, the capability of a typically sized stability study (usually three batches but sometimes more) to generate a supported shelf life that conforms to this quality statement is investigated with respect to both patient risk and business risk.

## A Brief Review of the ICH Methods

The current ICH methodology for determining the supported shelf life from a stability study assumes a fixed batch analysis. In a fixed batch analysis, the variation among batches is interpreted as differences in batch characteristics or differences in the fixed intercept and slope parameters which uniquely characterizes each batch's response over storage time, measured traditionally in months. Typically, three batches (occasionally more) of drug product are represented in a stability study for a New Drug Application (NDA) or Marketing Authorization Application (MAA) filing and are referred to as the registration batches.

The ICH shelf life analysis strategy begins with assuming that a simple linear regression model is adequate to characterize the response data from each of the batches, allowing for differences in the estimates of the intercept and slope parameters among the batches. Next, a series of poolability tests are conducted using a regression model selection procedure to determine how best to statistically characterize the batches. The poolability tests consider three variations of a simple linear model to determine the best fitted regression model to characterize trend across storage time in the set of stability batches: (1) allowing separate intercepts and separate slopes for the regressions fitted to each stability batch; (2) allowing separate intercepts with a common slope; and (3) a single regression fitted with a common intercept and common slope to the set of stability batches. In practice, for some stability-limiting responses, a fourth model is included that allows for (4) a common intercept with different batch slopes. Depending on the selected statistical model, a 95% confidence interval about the overall batch mean response or the worst case individual batch mean response is derived and the supported shelf life is defined as that storage time in which the confidence interval first intersects the acceptance criterion. In the next section, an industry data set is presented in which the response decreases over time. The remainder of this paper will utilize this data set.

## MATERIALS AND METHODS

### An Industry Data Set

To facilitate and to make the discussion as relevant as possible to the determination of the supported shelf life of a drug product, an industry data set is used for all computations and examples. These data were offered to the Working Group for their use from an anonymous PQRI member company. The data were blinded from the Working Group in terms of drug product and measurement units on the response. The industry data set represents 26 individual batches of a common pharmaceutical or drug product, where most batches remained on stability for a 24-month period and came from a manufacturing process assumed to be under statistical control. The response measure is percent assay (potency). For this discussion, it is assumed that (1) a simple linear regression model is appropriate to describe the stability data from each of the 26 batches, (2) the stability data themselves represent true batch mean responses, and (3) the determination of the overall regression line provides the true response trend among the 26 batches. That is, the regression line obtained for a particular batch represents the true regression line for that batch. This in turn allows determination of the true batch shelf life for each stability batch in the assumed finite population of 26 batches. The distribution of the true batch shelf lives corresponding to the 26 stability batches of the industry data set is used as a reference for evaluating the ICH approach for determining a supported shelf life. Consistent with the traditional ICH stability shelf life study scenario, sets of three or more stability batches are randomly sampled without replacement from the industry data set such that each set is sampled only once. Each set of randomly selected batches is used to estimate the regression parameters and determine a corresponding supported shelf life following ICH methods. The distribution of the supported shelf lives from the random sampling of the industry data set is then compared to the reference distribution of true batch shelf lives corresponding to the industry data set treated as a population.

The batch response data for the industry data set are shown in Fig. 1. The vertical axis is the percent assay or potency response which represents a stability-limiting response variable for the drug product. Acceptance criteria are set at 90 and 110%, as is typical for a potency specification in the USA. The horizontal axis is stability storage time in months, which is terminated at 48 months for ease of presentation. The actual range of storage times at which the batch lines intersect the lower acceptance limit ranges from 18 months (as shown) to 169.5 months.

The 26 estimated batch regression lines derived from the industry data set (solid black lines in Fig. 1) are assumed to represent the true regression lines for each batch. Also, the 26 estimated batch mean responses derived from the regression analysis at a particular time point are assumed to represent the 26 true batch mean assay responses at that time point. If the batch response data are first averaged and a regression analysis is performed on these averages, the resulting estimated line (the dashed line in Fig. 1) is assumed to represent the true regression trend among the overall batch mean responses. This overall true or population mean line intersects the lower acceptance criterion at 37.8 months (vertical reference line in Fig. 1), which, for our purposes, will define the true product shelf life.[1] In practice, of course, only a sample of batches is included in the stability study and variation is present both within and among batches. The variation depicted in Fig. 1 among the data and the intercepts and slopes of the 26 batch regression lines is typical of what can be observed in stability study results. This variation along with a consideration of risk leads to supported shelf lives that can be much shorter than the true product shelf life of a drug product. This is described in detail in the Results and Discussion Section.

### Product, Batch Mean, and Shelf Life Distributions

There are three distributions that are important to the discussion of shelf life: The product distribution, the batch mean distribution, and the shelf life distribution. Any method used to estimate the product shelf life has to account for both the risk associated with patients using a potentially subpotent product (patient risk) and the risk associated with a manufacturer having to set a short shelf life, which could lead to discarding good product (business or industry risk). These distributions provide the necessary framework to characterize the properties of a good estimation procedure. The product distribution and batch mean distribution are both defined with respect to the stability-limiting response variable, depicted as the vertical axis in Fig. 1. The shelf life distribution is defined with respect to the storage time, depicted as the horizontal axis in Fig. 1.

While it is recognized that stability testing is often performed on composite samples (multiple dosage units ground or dissolved together) such that unit dose uniformity is averaged out to some extent, the assumption is made here that each measured percent assay result in the industry data set represents the true percent assay associated with a single dosage unit. Under this assumption, the product distribution describes the range of product percent assay or potency available for patient usage at any given storage time

[1] The true product shelf life is an elusive concept to discuss and equally elusive to represent precisely with any mathematical rigor. Historically, for practical purposes, the true product shelf life has been defined as the storage time at which the mean of the product distribution, following some response trend across storage time, intersects the acceptance criteria. This is the basis for referring to 37.8 months as the true product shelf life of the industry data set. In practice, when there is between batch and within-batch variation, if the mean of the product distribution is used to define the true product shelf life, half of the product distribution will fall out of specification at expiry, which appears to be in conflict with the ICH philosophy. This paper is a step towards developing a coherent framework to discuss the true product shelf life. While some author's (*e.g.*, Kiermeier, *et al.* (6)) have proposed defining the true product shelf life to ensure that a predefined proportion of dosage units meet the acceptance criterion at expiry, until an agreed upon framework is developed, it is necessary to continue to apply the historic interpretation of true product shelf life discussed previously.

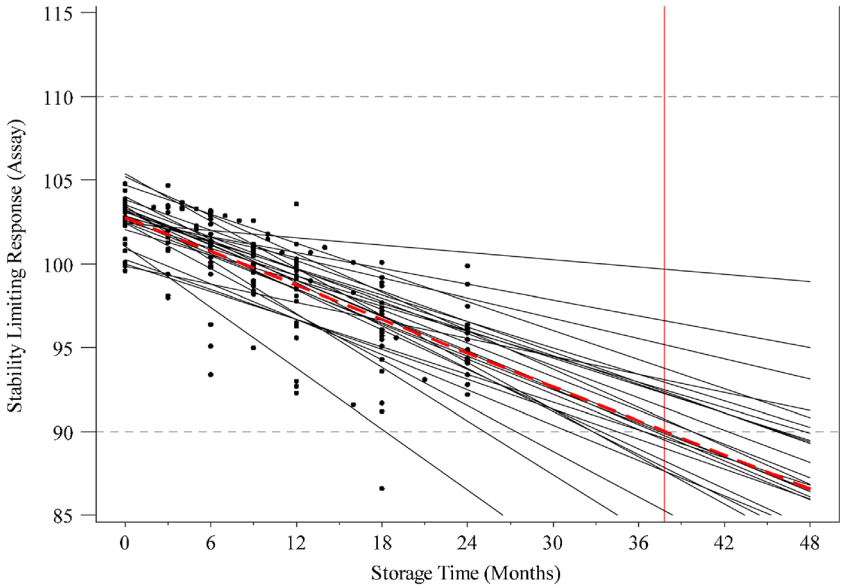


**Fig. 1.** True batch mean responses and corresponding regression lines for 26 batches of a common pharmaceutical product measuring percent assay *versus* storage months

(conceptually) although usually just expiry is of interest. It includes both between-batch and within-batch variation and is represented by the dots in Fig. 1. It is the product distribution that directly confronts the patient taking the drug product.

The batch mean distribution describes the variation among the batch mean responses derived from the regression analysis at any given storage time although, like the product distribution, usually just expiry is of interest. It is the batch mean distribution that is of primary interest to regulatory agencies when setting stability shelf life limits and to the pharmaceutical industry when determining the stability profile and overall acceptability of the drug product. The batch mean distribution shown in Fig. 2 is defined at a storage time of 37.8 months. The horizontal axis represents product potency expressed as percent assay where the midpoint of each bracket is labeled. The vertical axis is the frequency or number of batches with mean potency response in each bracket. Thirteen (50%) of the stability batches have estimated means below (outside) the lower acceptance criterion of 90% of product strength at this storage time and are thus considered nonconforming.

The shelf life distribution is defined across the horizontal axis representing the storage time corresponding to where each batch mean response intersects the lower acceptance criterion (in this case). Typically, the shelf life distribution is a positively skewed distribution unlike the product distribution and batch mean distribution, which are generally assumed to be normally distributed. The shelf life distribution for the industry data set is summarized in Fig. 3. The horizontal axis represents storage time in months where the midpoint of each bracket is labeled. The vertical axis describes the frequency or number of batches with the indicated shelf life in each bracket.

There is an interesting relationship between the batch mean distribution and the shelf life distribution. For a given storage time, $T$, the fraction of batches with shelf lives less than or equal to $T$ is the same as the fraction of batches with batch means less than or equal to the lower acceptance criterion. For example, in Fig. 3, the 5th quantile of the shelf life distribution is 25.1 months. From Fig. 1, at 25.1 months, only 2 out of the 26 batch means (7.7%) fall at or below the 90% specification limit. Similarly, the 75th quantile of the shelf life distribution is 47.6 months where 20/26 (76.9%) of the batch means fall below the 90% specification limit.

Unlike the industry data set, where the true batch means and true shelf lives are assumed known, in practice, these quantities are unknown and need to be estimated from the available stability data. Because the intent is to obtain conservative estimates, the relationship described earlier no longer holds. In general, quantiles of the individual distributions do not correspond to each other unless there is no variation among the batch slopes (Quinlan *et al.* (5)). The medians (50th quantiles) of the two distributions do correspond to each other. A schematic of this relationship is depicted in Fig. 4, where the vertical axis represents the values of a stability-limiting response variable and the horizontal axis is storage time. The solid line is the population overall mean batch response trend for a stability-limiting response variable which is increasing over storage time. The dashed line represents a plausible relationship between the distributions which depends on the shape and variances of the respective distributions. Note that by using the overall mean batch response, the batch mean distribution centers on the acceptance limit where half the distribution is out of specification.

## RESULTS AND DISCUSSION

### Applying ICH Methods for Estimating the Product Shelf Life to the Industry Data Set

#### *Overall Summary*

To evaluate the ICH methods for estimating the product shelf life, the industry data set is used to sample sets of three and six batches to reflect the typical range in sample size (number of batches) of most stability studies. Following ICH

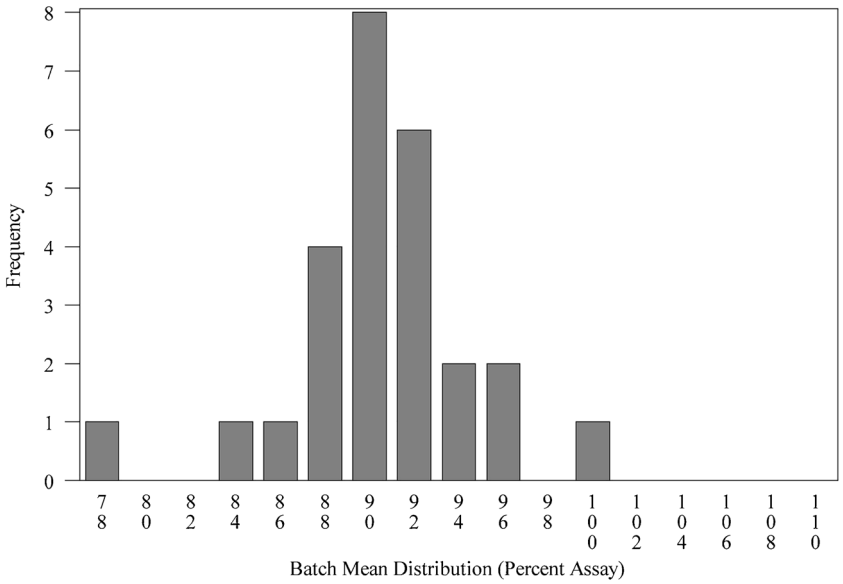


**Fig. 2.** The batch mean distribution for the industry data set as defined by the batch mean responses at 37.8 months of storage derived from the individual (true) regression lines. Thirteen (50%) of the batches have batch mean responses below 90%

guidance, the sampled sets of batches are used to estimate the product shelf life. A simple linear regression model is used to characterize batch assay response over storage time. The poolability testing strategy is then conducted, allowing for all four possible regression models as described previously. Based on the resulting regression model from the poolability tests, the estimated (*i.e.*, supported) shelf life is obtained based on when the appropriate 95% confidence interval crosses the lower acceptance criterion. While similar results are expected for the two cases (three batches *vs* six batches), for the sake of completeness, the full details of both are presented with important differences noted when they occur.

There are 2600 unique combinations of sets of three batches from the 26-batch industry data set. Each combination set of three batches represents a stability study. Following ICH methods, a supported shelf life is determined through a fixed batch analysis for each of these 2600 combination sets. Figure 5 is the summary distribution of the 2600 supported shelf lives truncated to 48 storage months for ease of presentation. The spike in frequency for supported shelf lives between 15 and 17 months is a result of including one or more rapidly degrading batches in the sample (see also Fig. 1). In these cases, the three sampled batches are usually not allowed to be pooled in the sense that a common slope

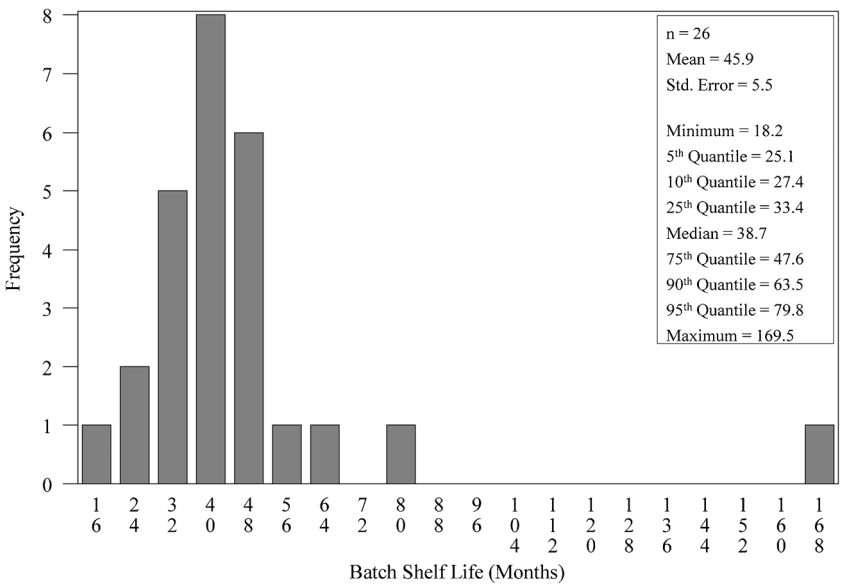


**Fig. 3.** The shelf life distribution defined by the storage time corresponding to when the true regression line for each stability batch in the industry data set intersects the acceptance criterion

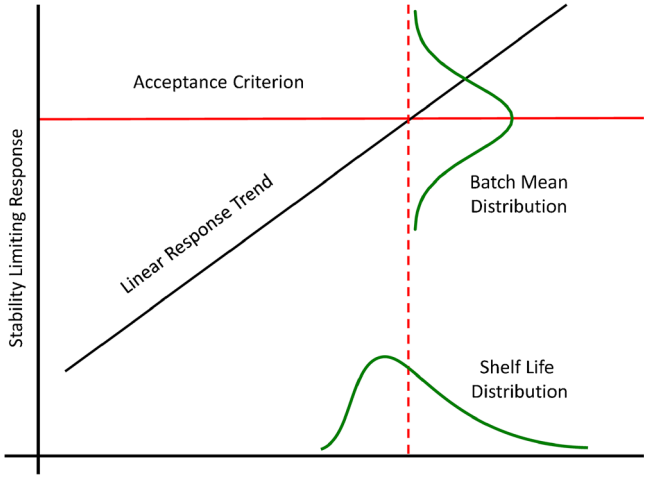


**Fig. 4.** Schematic of the relationship between the batch mean distribution and the shelf life distribution. The median of the shelf life distribution (red vertical dashed line) defines a storage time at which 50% of the batches are nonconforming

and/or common intercept estimate among the batches is not achievable. When this occurs, the data and information in the other two batches are only used in obtaining an estimate of the common variance, and not directly used in deriving the supported shelf life. In this case, a 95% confidence band about the worst case batch regression line leads to a short supported shelf life.

Correspondingly, if batches demonstrating slower degradation than typical are included in the sample, this can also prevent the pooling of the slopes and/or intercepts. In this case, though, even if pooling cannot be achieved, the effect on the supported shelf life is not as dramatic because the worst case batch still demonstrates a typical rate of degradation. However, if the data supports a common intercept and slope, the supported shelf life is then based on the overall mean response and can therefore be much longer. This is why the shelf life distribution tends to be positively skewed. Note that neither excessively short nor excessively long supported shelf life estimates are desirable.

Figure 6 is a similar summary of the supported shelf lives from combination sets of six batches. There are 230,230 unique combinations when selecting sets of six batches from the 26-batch industry data set. Analyzing all the possible combinations of six batches is not manageable. A randomly selected representative sample of 20,000 combination sets of six batches is used to produce Fig. 6 and for the following discussions.

Similar to Fig. 5, the erratic spikes in the lower section of the shelf life distribution shown in Fig. 6 are due to the increase in the number of batches from three to six which then increases the probability of sampling one or more of the rapidly degrading stability batches. This in turn increases the probability of not being able to pool the batches which then results in a shorter supported shelf life. The mean supported shelf life for sets of three batches is 26.0 months (Fig. 5), whereas the mean supported shelf life for sets of six batches is 23.4 months (Fig. 6). Similarly, a decrease in corresponding quantiles can be seen comparing Figs. 5 and 6 indicating a negative shift in the supported shelf life distribution with increase in the number of batches. This relationship is also depicted in Fig. 7 which displays the cumulative distributions of the ICH-supported shelf lives for both three and six batches. The cumulative distribution of the true batch shelf lives is also shown in Fig. 7 for comparison purposes. Note that the cumulative distribution corresponding to six batches is almost entirely to the left of the cumulative distribution corresponding to when only three batches are included in the stability study. With the current ICH methodology, increasing the number of batches in a stability study does not necessarily result in a longer supported shelf life.

In Fig. 7, the cumulative distribution of the true batch shelf lives is to the right of the corresponding cumulative distributions of the supported shelf lives as derived based on the ICH methods. This is expected since any reasonable estimation procedure will necessarily have to produce estimates of the product shelf life at which future batches are expected to remain within specification (ICH Q1A).

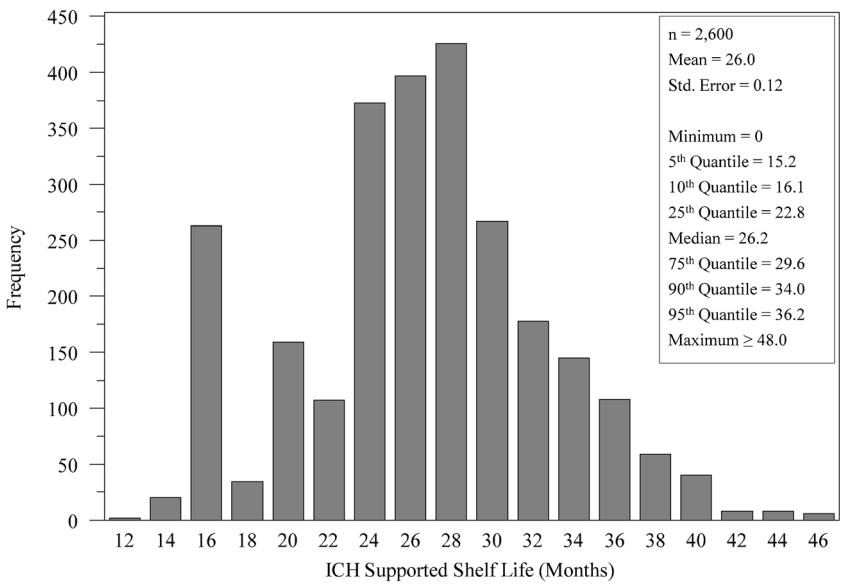


**Fig. 5.** The supported shelf life distribution based on ICH guidance methods using three batches

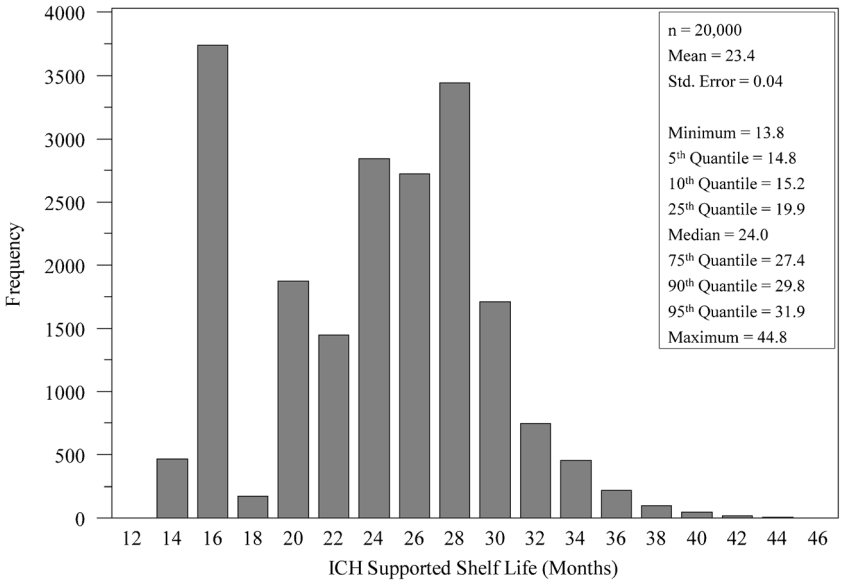


**Fig. 6.** The supported shelf life distribution based on ICH guidance using six batches

*Model-Specific Results*

The ICH guidance specifies that a poolability assessment should be conducted to determine the best fitted simple linear regression model to characterize the trend demonstrated by the batches. The poolability procedure allows for modeling the batches using either a common regression line for all batches or for separate intercept and/or slope parameters for each batch. The previous section focused on the distribution of the supported shelf life using the ICH methodology without regard of the results from a poolability assessment. Through the poolability assessment, the best fitted simple linear regression model or system of models characterizing each individual batch was determined, allowing for all four possible alternative models. Those four possible alternative models to best characterize response trend across storage time for the batches are as follows:

- Model 1: separate intercepts and separate slopes estimated for each batch,
- Model 2: common intercept and separate slopes estimated among batches,
- Model 3: separate intercepts and a common slope estimated among batches,
- Model 4: common intercept and a common slope estimated for all batches.

Using the industry data set and considering all 2600 possible combination sets of three batches, statistical summaries of the distributions of supported shelf lives corresponding to each specific model are provided in Fig. 8a–d. The poolability analysis was conducted as would be for the typical stability study regression analysis following the ICH analysis strategy.

There is a dependence of the distributions of the supported shelf life on the outcome of the poolability testing. For those models defined by separate slope estimates for each batch, models 1 and 2 summarized in Fig. 8a, b, respectively, the mean supported shelf life and related distributional quantiles for each model are similar. For those models defined by a common slope estimate for each batch, models 3 and 4 summarized in Fig. 8c, d, respectively, the mean supported shelf life and related distributional quantiles for each model are also similar. However, there is an overall shift to shorter storage times in the supported shelf life distributions comparing models 1 and 2 (separate slopes) to models 3 and 4 (common slopes).

To elaborate on this, suppose there is a business need to have a supported shelf life that is no less than 24 months. The above results imply that if a separate slope model is chosen, there is only about a 50% chance of attaining the required shelf life (median storage time ≈ 24 months in Fig. 8a, b). If a common slope model is chosen, then there is about a 90% chance of attaining a shelf life of 24 months (10th quantile ≈ 24 months in Fig. 8c, d). For the industry data set, while there is almost equal chance of the poolability test resulting in a separate slope model compared to a common slope model (48% (466 + 788)/2600) compared to 52% ((983 + 363)/2600), there is a greater industry or business risk associated with having to use a separate slope model instead of a common slope model.

Using the industry data set by considering a random sample of 20,000 of the 230,230 possible combination sets of six batches, statistical summaries of the distributions of supported shelf lives categorized by the results of the poolability tests to determine the best fitted regression model are summarized in Fig. 9a–d.

With six batches, similar to using three batches, there is a dependence of the results of the poolability tests for selecting the best fitted regression model on the distributions of the supported shelf life for each regression model. However, with six batches, there is a lesser chance of concluding that a common slope model is appropriate for characterizing the batches *versus* a model where the common slope assumption is not justified, 45 *versus* 55%, respectively. Using six batches, the mean supported shelf life and distributional quantiles for models 1 and 2 are similar, as shown in Fig. 9a, b, respectively. This pattern is not as evident for models 3 and 4, as shown in Fig. 9c, d, respectively, with the mean supported shelf life and the distributional

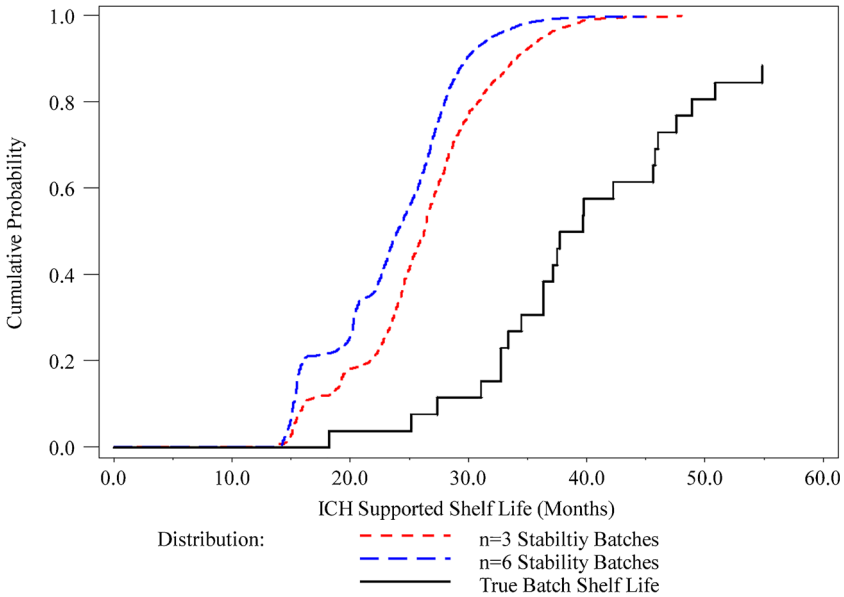


**Fig. 7.** Cumulative distribution functions of the supported shelf lives for three and six batches following ICH guidance and the corresponding true batch shelf lives from the industry data set

quantiles of model 3 being generally smaller than those for model 4. This is because the increase in the number of batches increases the probability that at least one rapidly degrading batch is included in the set of six batches preventing the poolability tests to conclude that a common intercept/slope model is adequate to characterize the overall response trend over storage time. Note that when comparing common slope models, 94.4% (8451/(8451 + 499)) of the combination sets from the industry data set resulted in model 3 best characterizing response trend and only 5.6% of the combination sets allowed for both a common intercept and a common slope.

Again, assume there is a business need to have a supported shelf life that is no less than 24 months. The above results imply that if a separate slope model is chosen, there is only about a 25% chance of attaining the required shelf life (75th quantile ≈ 24 months in Fig. 9a, b). If a separate intercept/common slope model is chosen, then there is about a 90% chance of attaining a shelf life of 24 months (10th quantile ≈ 24 months in Fig. 9c). For a common intercept/common slope model, all of the supported shelf lives were at least 24 months (Fig. 9d). As for the case of three batches, the choice of model can significantly impact business risk. Of course, whether or not a shelf life of 24 months (or longer) is satisfactory from a patient risk perspective would also have to be factored into the manufacturer's deliberations regardless of how many batches are placed on stability.

## Evaluating the Risk/Benefit of Estimating a Product Shelf Life Using the ICH Approach

The primary objective of any shelf life estimate is to mitigate the patient's risk through the assurance of a safe and efficacious drug product throughout the drug product's storage life. For a given storage time, usually taken to be the expiration date of the drug product, patient risk can be defined in terms of the product distribution, the batch mean distribution, or the shelf life distribution. In terms of the product distribution, patient risk is defined as the proportion of the individual dosage units that fall outside of the acceptance criteria (*i.e.*, out of specification) at expiry. Because it is often not feasible to adequately characterize the product distribution at expiry, for both regulatory and industry reasons, a surrogate definition of patient risk is used, namely, the proportion of the batch mean distribution outside the acceptance criteria at expiry, as shown in Fig. 10. Recall that the batch mean distribution is defined by the intersection points of the individual batch mean responses, derived from a regression analysis, at expiry. The shelf life distribution is defined by the distribution of shelf lives obtained through the batch mean distribution corresponding to each batch's shelf life. Because of the duality between the batch mean distribution and the shelf life distribution, patient risk can also be assessed through the latter distribution as the proportion of batches with shelf lives that are less than the estimated product shelf life, as shown in Fig. 10. Industry risk is the counterpart to patient risk and is of concern when the product shelf life is sufficiently short to force the batch mean distribution to be well within the acceptance criteria or specification at expiry. Having a shelf life that is too short may necessitate early discard of still effective product and not meet the business needs of the manufacturer. A "good" estimate of shelf life is therefore one that, in some sense, balances these two competing risks.

To validate any shelf life estimate and quantify the risk to the patient, how well the product distribution and batch mean distribution are managed with respect to the acceptance criteria must be considered. Traditionally, this is done at expiry as defined by the supported shelf life. The schematic in Fig. 4 is a depiction of an example where 50% of the distribution of batch means is out of specification at a storage time corresponding to when the overall mean response trend intersects the acceptance criteria. Figure 10 is a depiction of an example where the supported shelf life is derived to minimize risk to the patient while ensuring that it still meets the business needs of the manufacturer. Note that the supported shelf life delineated at the vertical dashed line in

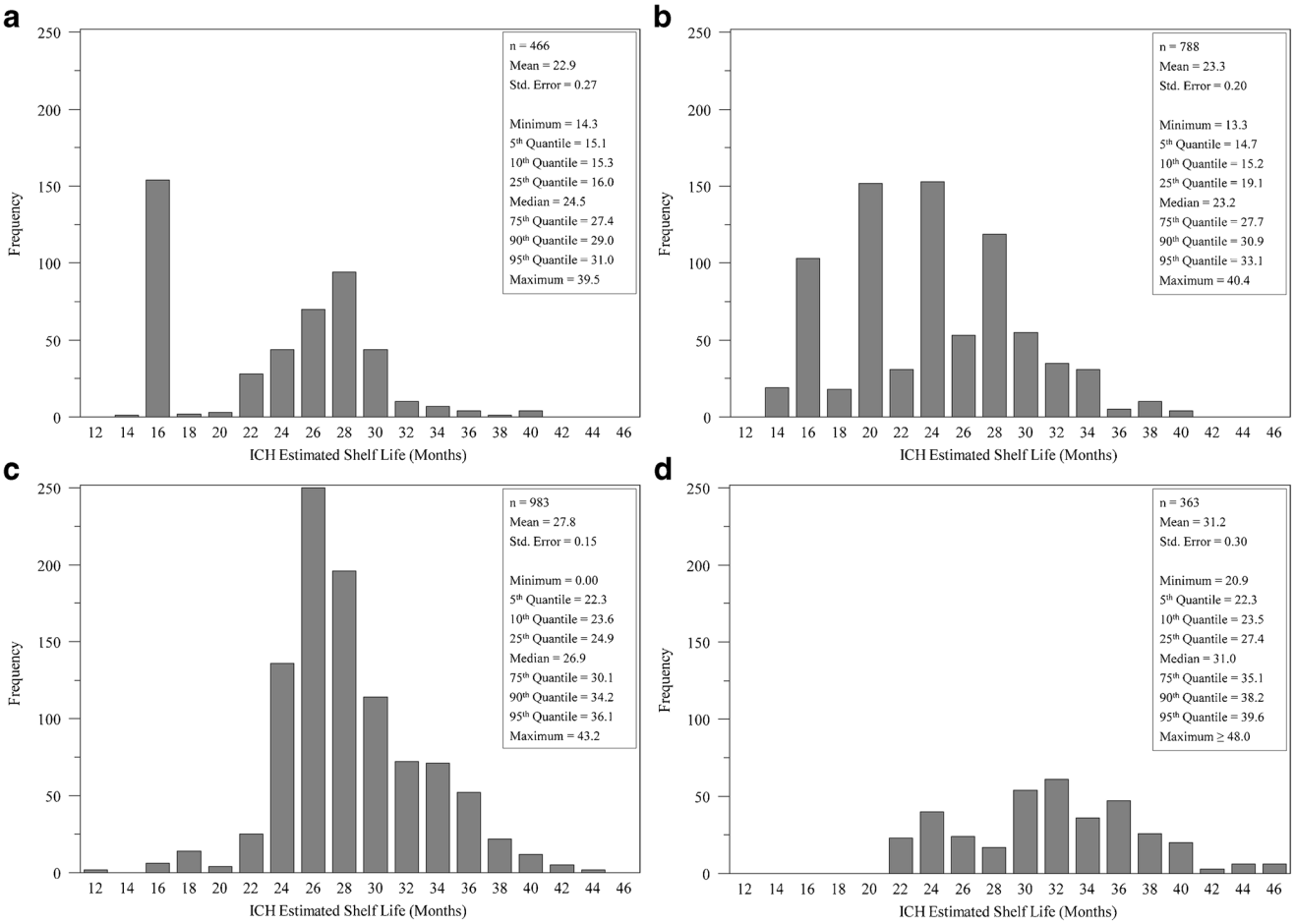


**Fig. 8.** Distribution of supported shelf lives based on three batches corresponding to **a** model 1: separate intercepts and separate slopes, **b** model 2: common intercept and separate slopes, **c** model 3: separate intercepts and common slope, **d** model 4: common intercept and common slope.

Fig. 10 is not defined by where the mean response intersects the acceptance criteria. This will be further explored in a future paper.

### *Assessing the ICH Approach Using the Batch Mean Distribution*

To assess how well the ICH methods perform in regard to managing risk, the industry data set is again used to represent a finite population of product stability batch data for discussion and comparison purposes. Using a sampling of all possible combinations of three (or random sample of all possible combinations of six) batches, the supported shelf life is computed for each combination set following ICH guidance methods. As mentioned previously, because shelf life often exceeds the duration of the stability study, it is difficult to correctly define the proportion of the product distribution that would be out of specification with respect to an acceptance criteria at expiry. A pragmatic solution is to characterize each batch by the predicted mean response of the stability-limiting variable as determined from a regression analysis thereby defining the batch mean distribution at expiry. For each combination set of batches, a batch from the industry data set will be considered nonconforming if the predicted mean response for that batch is out of specification at expiry as determined by the supported shelf life. The proportion of the 26 batches from the industry data set that are nonconforming is recorded for each combination set of batches. In addition, the number of occurrences when all of the 26 batches from the industry data set are within specification bounds (conforming), as defined by each batch's predicted response being within specification bounds, is also recorded. These results are summarized in Table I categorized by the regression model type best fitted to the three batches through the poolability tests and for all combinations of three combination batch sets.

As was concluded in previous discussions, the ability to fit, or not fit, a regression model with a common slope estimate for all batches in each combination set of batches influences the summarized results. Here, all supported shelf life estimates are based on 95% confidence intervals about the mean pooled response or about the mean response of the worst case batch results as determined by the poolability tests. From Table I, if the mean response trends are characterized by individually estimated batch slopes, models 1 and 2, the average proportion of nonconforming batches at expiry is 6.1 and 7.0%, respectively. Recall from Fig. 8a, b that the mean (range) of the supported shelf lives was 22.9 (14.3–39.5)

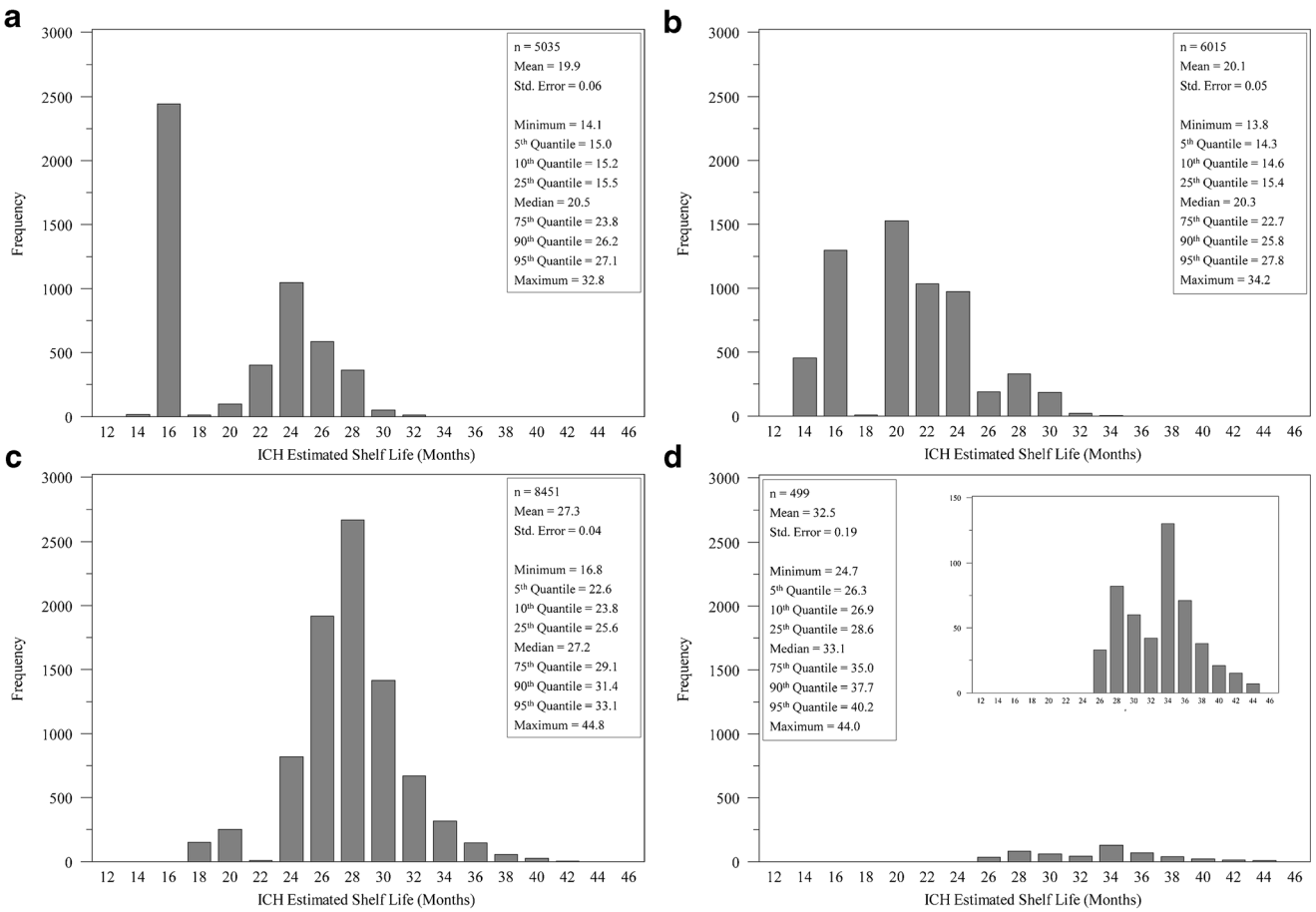


**Fig. 9.** Distribution of supported shelf lives based on six batches corresponding to **a** model 1: separate intercepts and separate slopes, **b** model 2: common intercept and separate slopes, **c** model 3: separate intercepts and common slope, **d** model 4: common intercept and common slope (to better exhibit the relationship among the bars, the inset displays the same data but on a frequency scale of 0 to 150).

months and 23.3 (13.3–40.4) months, respectively. Fitting a regression model with a common slope estimated among batches results in a larger average proportion of nonconforming batches, 11.9 and 20.6% for models 3 and 4, respectively. From Fig. 8c, d, the mean (range) of the supported shelf lives was 27.8 (0.00–43.2) months and 31.2 (20.9–≥ 48) months, respectively. The overall range in the proportion of nonconforming batches among the 2600 combinations is quite broad at 0.0 to 76.9%.

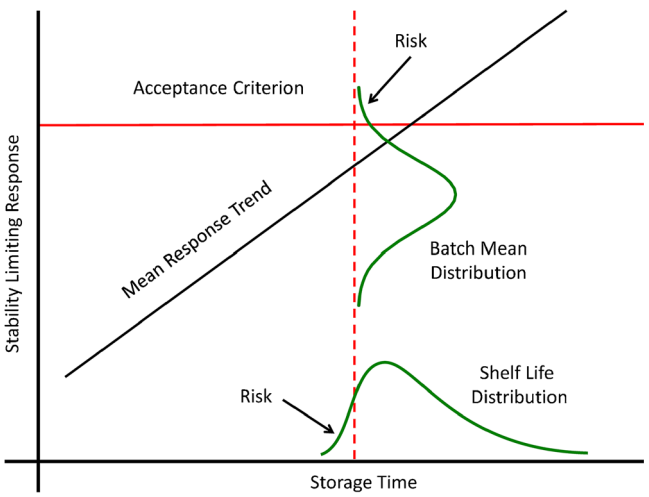


**Fig. 10.** Batch mean distribution and the shelf life distribution with vertical red dashed line representing a supported shelf life that effectively manages patient risk while satisfying business needs

As also displayed in Table I, the proportions of combination sets corresponding to batch mean distributions being completely within acceptance criteria are 33.7% and 17.0% for models 1 and 2, respectively, assuming individual slope estimates among batches. Among those combination sets, the corresponding mean (range) of the supported shelf lives is 15.6 (14.3–16.5) months and 15.1 (13.3–18.1) months, respectively. For those batch combination sets where a common slope is fitted, models 3 and 4, the corresponding proportions are 2.2% and 0.0%, respectively. For the combination sets fit by model 3, the corresponding mean (range) of the supported shelf lives is 14.9 (0.0–17.6) months. Note that all of the combination sets associated with model 4 contained at least one nonconforming batch. From Figs. 1 and 2, it is not surprising that all of the maximum supported shelf lives listed in the last column in Table I are approximately 18 months or less since the minimum true shelf life is 18.2 months. In the industry data set, if a standard is set that requires the entire batch mean distribution to conform to acceptance criteria at expiry, which reflects the spirit of the ICH guidance documents, then the supported shelf life cannot be larger than about 18 months.

**Table I.** Summary of Proportion of Nonconforming Batches and Proportion of Combination Sets Corresponding to the Batch Mean Distribution Being Completely Within the Acceptance Limits at Expiry for ICH-Supported Shelf Lives Determined Using Three Batch Combinations from the Industry Data Set

| Regression model | Number of combinations | Nonconforming batches based on the predicted mean batch response at expiry | | Combinations sets corresponding to the batch mean distribution being completely conforming to acceptance criteria | | |
|---|---|---|---|---|---|---|
| | | Average percent | Percent range | Number of combinations | Mean supported shelf life | Range of supported shelf lives |
| Model 1 | 466 | 6.1% | 0.0–50.0% | 157 (34%) | 15.6 | 14.3–16.5 |
| Model 2 | 788 | 7.0% | 0.0–57.7% | 134 (17%) | 15.1 | 13.3–18.1 |
| Model 3 | 983 | 11.9% | 0.0–61.5% | 22 (2%) | 14.9 | 0.0–17.6 |
| Model 4 | 363 | 20.6% | 3.8–76.9% | 0 (0%) | – | – |
| Overall | 2600 | 10.6% | 0.0–76.9% | 313 (12%) | 15.4 | 0.0–18.1 |

Table II is a summary using a random sample of combination sets of six batches. Comparing Table II with Table I, the increase from three to six batches decreased the overall average percent of nonconforming batches from 10.6% to 6.5% and increased the proportion of batch mean distributions completely conforming to the acceptance criteria from 12.0% to 23.3%. These results are expected due to the decrease in the supported shelf life when increasing the number of batches from three to six, reported in Fig. 8a–d and Fig. 9a–d, respectively. The maximum supported shelf lives reported in the last column of Table II are also similar to those reported in Table I. As for the case of three batches, to satisfy a standard that requires the entire batch mean distribution to conform to acceptance criteria at expiry, the supported shelf life cannot be larger than about 18 months.

*Assessing the ICH Approach Using the Shelf Life Distribution*

The earlier discussion described how the current ICH methods manage the batch mean distribution in quantifying patient risk. But beyond just characterizing the distribution of batch means at the supported shelf life, it is also important to consider the distribution of supported shelf lives when evaluating any method that estimates the product shelf life. Tables III and IV summarize the supported shelf life distribution using all possible three batch combinations (Table III) or a random sample of all possible six batch combinations (Table IV) from the industry data set. For example, for the 466 combinations that fell under model 1 in Table III, the corresponding mean supported shelf life is 22.9 months with a range of 14.3 to 39.5 months. Each of these 466 supported shelf lives corresponds to a particular quantile of the overall distribution of supported shelf lives that was depicted in Fig. 3. The average of these 466 quantiles is 37.8% with a range of 0.8% to 98.7%. Out of the 2600 supported shelf lives derived from all possible selections of three batches from the industry data example, 37.8% of them were less than or equal to 22.9 months, 0.8% of them were less than or equal to 14.3 months and 98.7% of them were less than or equal to 39.5 months. Similar interpretations can be made for the other models.

From Table III, it is apparent that there is a dependence of the results of the poolability tests for selecting the best fitted regression model on not only the mean of the supported shelf lives (models 1 and 2 *versus* models 3 and 4), but also on the corresponding quantiles of the overall supported shelf life distribution. In addition, this dependence is more specific when comparing quantiles for separate *versus* common intercept estimates for common slope models, 58.5% *versus* 72.8%, respectively.

Table IV is a summary of the results for a sampling of 20,000 combinations of six batches from the 230,230 possible combinations using the industry data set and following ICH methods for estimating shelf life. The results are similar and consistent with those results summarized in Table III for sets of three batches.

For either three or six batches, if a regression model with separate slopes among batches is selected, the supported shelf life tends to be in the lower tail of the overall supported shelf life distribution with mean supported shelf lives that are less than the overall mean. Although such models do a better job of satisfying the intent of the ICH guidance, when the estimated true product shelf life is too short, there is an increased risk that safe and efficacious batches will have to be discarded at expiry. If a common slope model is selected, the supported shelf life tends to be in the upper tail of the overall supported shelf life distribution with mean supported shelf lives that are greater than the overall mean. However, these estimates correspond more to situations where a relatively high proportion of the distribution of batch means is not in conformance at expiry, potentially putting patients at risk.

*Assessing the ICH Approach: Final Thoughts*

No realistic shelf life can assure that the critical attributes of the drug product can meet the ICH criterion with certainty. Thus, when evaluating the performance of any estimation procedure, the proportion of nonconforming drug product batches at expiry must be considered in order

**Table II.** Summary of Proportion of Nonconforming Batches and Proportion of Combination Sets Corresponding to the Batch Mean Distribution Being Completely Within Specification at Expiry for ICH-Supported Shelf Life Determined Using Six Batch Combinations from the Industry Data Set

| Regression model | Number of combinations | Nonconforming batches based on the mean batch response at expiry | | Combinations sets corresponding to the batch mean distribution being completely conforming to acceptance criteria | | |
|---|---|---|---|---|---|---|
| | | Average percent | Percent range | Number of combinations | Mean supported shelf life | Range of supported shelf lives |
| Model 1 | 5035 | 2.7% | 0.0–23.1% | 2474 (49%) | 15.5 | 14.4–16.6 |
| Model 2 | 6015 | 3.4% | 0.0–26.9% | 1752 (29%) | 14.8 | 13.8–18.1 |
| Model 3 | 8451 | 10.0% | 0.0–61.5% | 131 (2%) | 17.5 | 16.8–18.1 |
| Model 4 | 499 | 23.7% | 3.8–61.5% | 0 (0%) | – | – |
| Overall | 20,000 | 6.5% | 0.0–61.5% | 4357 (23%) | 15.3 | 13.8–18.1 |

to manage the risk/benefit ratio of estimating a product shelf life. The manufacturer and regulatory agency must come to an agreement on allowing an *a priori*-defined small proportion of nonconforming batches at expiry providing balance between patient and industry risk. The conclusion reached by considering Tables I, II, III, and IV is that the supported shelf lives, as determined from the ICH approach, vary greatly in both storage time and in the corresponding quantile of the supported shelf life distribution and, as a result, vary greatly in their ability to manage the business needs of the manufacturer while also conforming to the intent of the ICH guidance. Alternative approaches are required.

## CONCLUSIONS AND FURTHER RESEARCH

The evaluation of the ICH methods for estimating shelf life presented in this paper is based on an industry data set. The data set is composed of 26 batches of a common drug product on stability for 24 months. Using the industry data set as a reference, representing a finite population of commercial production batches, allowed evaluation of the ICH methods as would be applied to the typical analysis of stability data for estimating a product shelf life. Following ICH methods, shelf life estimates were computed using poolability tests to select the appropriate regression model then constructing 95% confidence intervals about the overall batch mean response or the worst case individual batch mean response depending on the regression model selected. It was shown for the industry data set that there is approximately equal chance of selecting one of the common slope models as compared to one of the separate slope models for both sampling three and six batches. The result is that the supported shelf life for those models with separate slopes among batches is substantially less than the supported shelf life obtained from common slope models. Similarly, the supported shelf life based on models with separate intercepts among batches is less than those estimates from models with a common intercept, although the difference is less dramatic. A regression model that assumes both a common intercept and slope among batches results in the longest estimates of the supported shelf life. However, the common intercept/slope regression model is the model selected least often through the poolability tests.

Several issues were highlighted throughout this evaluation. The batch mean distribution at expiry and the shelf life distribution were defined and utilized to evaluate the ICH strategy in terms of managing risk. A relationship between the two distributions was described through the quantiles of each. ICH puts emphasis on the batch mean distribution to define an appropriate shelf life estimate since according to the guidance, the objective of a stability study is to give assurance that future batches will be safe and efficacious at expiry. However, consideration of both the batch mean distribution and shelf life distribution is

**Table III.** Summary of the Shelf Life Distribution for ICH-Supported Shelf Lives Using Three Batch Combinations from the Industry Data Set

| Regression model | Number of combinations | Mean supported shelf life (months) | Range of supported shelf life (months) | Average of the corresponding quantiles of the overall supported shelf life distribution | Quantile range |
|---|---|---|---|---|---|
| Model 1 | 466 | 22.9 | 14.3–39.5 | 37.8% | 0.8–98.7% |
| Model 2 | 788 | 23.3 | 13.3–40.4 | 37.1% | 0.1–99.2% |
| Model 3 | 983 | 27.8 | 0.0–43.2 | 58.5% | 0.1–99.7% |
| Model 4 | 363 | 31.2 | 20.9–48.0 | 72.8% | 18.7–100.0% |
| Overall | 2600 | 26.0 | 0.0–48.0 | 50.3% | 0.1–100.0% |

**Table IV.** Summary of Shelf Life Distribution for ICH-Supported Shelf Lives Using Six Batch Combinations from the Industry Data Set

| Regression model | Number of combinations | Mean supported shelf life (months) | Range of supported shelf life (months) | Average of the corresponding quantiles of the overall supported shelf life distribution | Quantile range |
|---|---|---|---|---|---|
| Model 1 | 5035 | 19.9 | 14.1–32.8 | 33.2% | 0.4–96.3% |
| Model 2 | 6015 | 20.1 | 13.8–34.2 | 31.2% | 0.0–97.9% |
| Model 3 | 8451 | 27.3 | 16.8–44.8 | 71.9% | 21.1–100.0% |
| Model 4 | 499 | 32.5 | 24.7–44.0 | 90.7% | 54.2–100.0% |
| Overall | 20,000 | 23.4 | 13.8–44.8 | 50.4% | 0.0–100.0% |

needed to properly manage risk. Results for sampling both three and six batches from the industry data set ranged from a nominal to a high percentage of the batch mean distribution being out of specification at expiry. In addition, results from the industry data set showed a high percentage of the batch mean distribution being completely conformable to acceptance criteria at expiry. Neither result is optimal from either the patient or manufacturer perspective. That is, utilizing a standard that requires the entire batch mean distribution to conform to acceptance criteria at expiry, which is the intent of the ICH guidance, may require the manufacturer to set the product shelf life to a storage time that is less than estimated by their data and, potentially, so short that the product does not satisfy business needs.

The current ICH methods are based on considering batches as a fixed effect in the statistical regression analyses, similar to considering batches as a fixed factor in a typical analysis of variance. The ramification of a fixed batch effect analysis is that inference is restricted to only those batches included in the stability study because only a single variance estimate is obtained to describe both the variation among the observed response data within batches and the variation between batches. An alternative to the fixed batch analysis is the so-called random batch analysis that has been discussed for a number of years within the statistical community and among those in other drug development areas interested in stability issues. The random batch analysis offers two major advantages over the fixed batch analysis. First, by incorporating batch-to-batch variability in the model, it allows for a broader inference to all future batches, which is more in line with the stated objectives of the ICH guidance. Second, the random batch analysis does not require poolability testing. Batch-to-batch differences in the intercepts and/or slopes contribute to the overall variability in the model. The random batch analysis strategy was applied to the industry data and demonstrated advantages over the ICH-fixed batch analysis results but also did not effectively manage risk. This leaves open further discussion and research on how to estimate a product shelf life that does not fall in the extreme left tail of the shelf life distribution but is also not so long that an unacceptable percentage of the product and/or batch mean distribution exceeds acceptance limits at expiry. Research continues on addressing these remaining issues. Current efforts center on how to best incorporate methods based on tolerance intervals and/or calibration techniques. Work is also progressing on applying Bayesian methods to estimate product shelf life and manage risk. Results of this continuing research will be reported in a future paper.

## ACKNOWLEDGEMENTS

The authors are the current members of the PQRI Stability Shelf Life Working Group and thank all past members and colleagues of the working group. The Working Group also wishes to acknowledge the helpful comments of the reviewers which improved the presentation of the material.

**Funding Information** This research is supported by the Product Quality Research Institute (PQRI) of Arlington, Virginia.



## REFERENCES

1. Capen R, Christopher D, Forenzo F, Ireland C, Liu O, Lyapustina S, et al. On the shelf life of pharmaceutical products. AAPS PharmSciTech. 2012;13(3):911–8.
2. Anonymous. Stability Testing of New Drug Substances and Products Q1A(R2) Step 4. In International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use. 2003. Available from URL: http://www.ich.org. Last accessed 21 Sep 2017.
3. Anonymouse. Evaluation of Stability Data Q1E Step 4. In International Conference On Harmonisation Of Technical Requirements For Registration Of Pharmaceuticals For Human Use. 2003. Available from URL: http://www.ich.org. Last accessed 21 Sep 2017.
4. Kiermeier A, Jarrett R, Verbyla A. A new approach to estimating shelf-life. Pharm Stat. 2004;3:3–11.
5. Quinlan M, Stroup W, Schwenke J, Christopher D. Evaluating the performance of the ICH guidelines for shelf life estimation. J Biopharm Stat. 2013;23(4):881–96.
6. Kiermeier A, Verbyla A, Jarrett R. Estimating a single shelf life for multiple batches. Aust NZ J Stat. 2012;54(3):343–58.