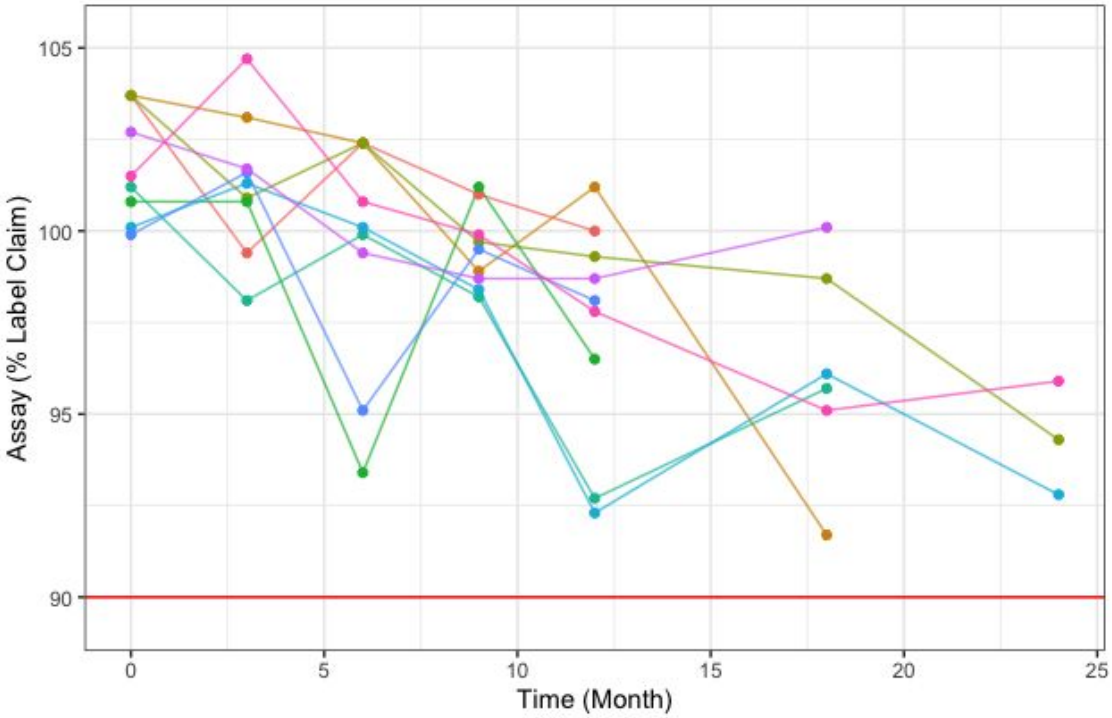
Genentech
A Member of the Roche Group

Cristian Oliva-Aviles

Shelf-life estimation through tolerance intervals under linear mixed models

10th IABS Statistics Workshop
University of Maryland, IBBR
November 13th, 2024

Motivation: unbalanced stability data



Given that stability data is frequently modeled via linear mixed models (LMMs), **how can I compute a tolerance interval for the response distribution over time**?

Source: PQRI Industry Example Data Set (modified)

Outline

1. **Tolerance intervals (TIs)** have become an important statistical tool across various industries due to their ability to provide information about a specified **proportion of a population with high certainty**.
    - Often used to make informed risk-based decisions regarding **product quality** (e.g., shelf-life).

2. However, there is a **lack of statistically-justified TI methods** for general unbalanced linear mixed models (LMMs).
    - TI methods for unbalanced data have been mainly focused on *one-way random-effects models*.
    - Sharma and Mathew (2012) proposed a method that relies on small-sample asymptotics, and assumes independence between random effects.

Outline

3. Thus, we **developed a method** -based on Generalized Pivotal Quantities- to compute TIs for a large class of LMMs.

- Particularly, this method can be used for random–intercepts random-slopes models, oftenly used to analyze stability data

The method was recently published in **Technometrics**:

Tolerance Intervals Under a Class of Unbalancer Linear Mixed Models.

Oliva-Aviles, C. and Hauser, P. (2024)

Paper available here.

Preliminaries

Types of Tolerance Intervals

In general, a tolerance interval is an interval that **contains 100β% of the population**.

Guttman (1970) defined two types of tolerance intervals:

1. **(β,γ)-tolerance interval**
   - contains at least 100β% of the population **with probability 100γ%.**

2. **β-expectation tolerance interval (dual interpretation)**
   - covers 100β% of the population on **average**,
   - equivalent to prediction intervals for a single future observation.

Approaches for (β,γ)-tolerance intervals

One-sided TI for W~N(θ,τ^2)

Equivalent to one-sided confidence limits on quantiles of Fx.

For instance, a lower one-sided (β,γ)-TI is equivalent to a γ-confidence lower limit for the parameter

$$\delta = \theta - z_\beta \tau$$

Two-sided TI for W~N(θ,τ^2)

Liao et al. (2005) considered an approach for computing approx two-sided (β,γ)-TIs based on the following setting: there exists an independent statistic T~N(θ,σ^2).

Define $\zeta = \sqrt{(\tau^2 + \sigma^2)}$. Then, an approximate two-sided TI is given by

$$\left(T - z_{(1+\beta)/2} R_{\zeta,\gamma}, T + z_{(1+\beta)/2} R_{\zeta,\gamma}\right)$$

where $R_{\zeta,\gamma}$ denotes the upper generalized γ-confidence limit for ζ.

Approaches for β-expectation tolerance intervals

The computation of β-expectation tolerance intervals is commonly tied to their **equivalency with prediction intervals for a single future observation** with confidence level 100β%.

Note: *while the computation of these intervals is not discussed in this presentation, the proposed method can be easily adapted to compute them.*

Generalized Pivotal Quantities

We developed a method to compute tolerance intervals under unbalanced LMMs that is based on **Generalized Pivotal Quantities** (GPQs).

The concept of GPQs was first introduced by *Weerahandi (1993)*.

Let R = r(X;x,v), where v=(θ,δ). Then, R is said to be a (fiducial) GPQ for θ if:

1. The distribution of R is free of unknown parameters.
2. The observed value of R, say r(x;x,v) is exactly θ.

Example: Let X1, X2, …, Xn be a random sample from N(μ,θ). Also, let S^2 be the sample standard deviation, and s^2 its observed value. Then, the following is a GPQ for θ:

$$\frac{s^2\theta}{S^2} = [(n-1)s^2]\left[\frac{\theta}{(n-1)S^2}\right]$$

The random-intercepts random-slopes model

Random-intercepts random-slopes

The measured value y_{ij} for batch i at time t_{ij} follows the model

$$y_{ij} = (\mu + \mu_i) + (\tau + \tau_i)t_{ij} + e_{ij}$$

where, for each i,

$$\begin{pmatrix} \mu_i \\ \tau_i \end{pmatrix} \overset{iid}{\sim} N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_\mu^2 & \sigma_{\mu,\tau} \\ \sigma_{\mu,\tau} & \sigma_\tau^2 \end{pmatrix} \right]$$

and

$$e_{ij} \overset{iid}{\sim} N(0, \sigma_e^2)$$

Matrix Form, random-intercepts random-slopes

Toy example (2 batches)

$$\begin{pmatrix} y_{11} \\ y_{12} \\ y_{21} \\ y_{22} \\ y_{23} \end{pmatrix} = \begin{pmatrix} 1 & 30 \\ 1 & 90 \\ 1 & 30 \\ 1 & 90 \\ 1 & 180 \end{pmatrix} \begin{pmatrix} \mu \\ \tau \end{pmatrix} + \begin{pmatrix} 1 & 30 & 0 & 0 \\ 1 & 90 & 0 & 0 \\ 0 & 0 & 1 & 30 \\ 0 & 0 & 1 & 90 \\ 0 & 0 & 1 & 180 \end{pmatrix} \begin{pmatrix} \mu_1 \\ \tau_1 \\ \mu_2 \\ \tau_2 \end{pmatrix} + \begin{pmatrix} e_{11} \\ e_{12} \\ e_{21} \\ e_{22} \\ e_{23} \end{pmatrix}$$

Generally, for batch i, the model can be expressed in matrix form as follows:

$$\mathbf{Y}_i = \mathbf{X}_i \begin{pmatrix} \mu \\ \tau \end{pmatrix} + \mathbf{Z}_i \begin{pmatrix} \mu_i \\ \tau_i \end{pmatrix} + \mathbf{e}_i \qquad \text{with} \quad \mathbf{Z}_i = \mathbf{X}_i$$

Matrix Form, random-intercepts fixed-slopes

Toy example (2 batches)

$$\begin{pmatrix} y_{11} \\ y_{12} \\ y_{21} \\ y_{22} \\ y_{23} \end{pmatrix} = \begin{pmatrix} 1 & 30 \\ 1 & 90 \\ 1 & 30 \\ 1 & 90 \\ 1 & 180 \end{pmatrix} \begin{pmatrix} \mu \\ \tau \end{pmatrix} + \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} + \begin{pmatrix} e_{11} \\ e_{12} \\ e_{21} \\ e_{22} \\ e_{23} \end{pmatrix}$$

In this case, for batch i, the model can be expressed in matrix form as follows:

$$\mathbf{Y}_i = \mathbf{X}_i \begin{pmatrix} \mu \\ \tau \end{pmatrix} + \mathbf{Z}_i \mu_i + \mathbf{e}_i, \quad \text{with} \quad \mathbf{Z}_i = \mathbf{X}_i \begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

The Distributions of Interest

Distributions

Recall that tolerance intervals are computed for distributions, unlike confidence intervals, which are computed for parameters.

Given c≥0, we're interested in computing tolerance intervals for the **batch response distribution at time t=T**:

$$(\mu+\mu_i)+(\tau+\tau_i)T+c\epsilon_{ij} \sim N\left(\mu+\tau T, \sigma_\mu^2+2T\sigma_{\mu,\tau}+\sigma_\tau^2T^2+c^2\sigma_e^2\right)$$

Mean value at t=T

Batch-to-batch variability at t=T

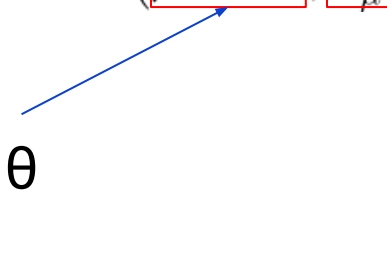
Error variability

Distributions

Recall that tolerance intervals are computed for distributions, unlike confidence intervals, which are computed for parameters.

Given c≥0, we're interested in computing tolerance intervals for the **batch response distribution at time t=T**:

$$(\mu+\mu_i)+(\tau+\tau_i)T+c\epsilon_{ij} \sim N\left(\mu+\tau T, \sigma_\mu^2+2T\sigma_{\mu,\tau}+\sigma_\tau^2T^2+c^2\sigma_e^2\right)$$



Connection with shelf life

$$(\mu+\mu_i)+(\tau+\tau_i)T+c\epsilon_{ij} \sim N\left(\mu+\tau T, \sigma_\mu^2+2T\sigma_{\mu,\tau}+\sigma_\tau^2T^2+c^2\sigma_e^2\right)$$

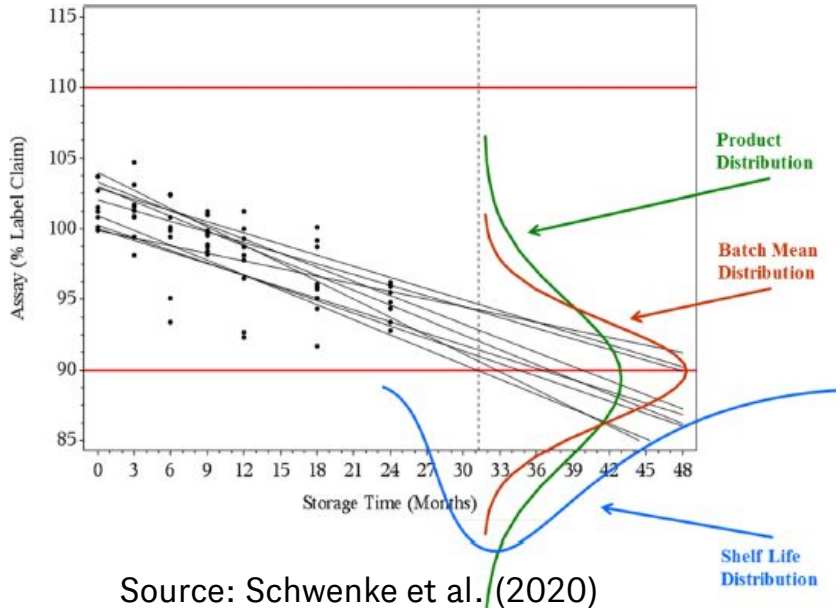
Schwenke et al. (2020) introduced three distributions of interest for shelf life estimation.

1) **Batch mean distribution**: c=0

2) **Product distribution**: c=1

3) **Shelf life distribution**

1) and 3) have been shown to lead to equivalent shelf life definitions (under certain conditions).



Source: Schwenke et al. (2020)

The Proposed Method

Searching for GPQs

Following Liao et al. (2005), the strategy to compute tolerance intervals consists of finding GPQs for the required parameters:

$$N(\theta, \rho + c\sigma_e^2)$$

General idea of the method: use a specific set of independent pivotal quantities to generate a system of equations with unique solutions (which are numerically obtained).

The GPQ method

$$N(\theta, \rho + c\sigma_e^2)$$

The developed GPQ method to compute TIs consists of three main steps:

1. Compute $R_{\sigma_e^2}$ through a GPQ for σ_e^2
2. Compute R_ρ through a GPQ for ρ (given $R_{\sigma_e^2}$)
3. Compute R_θ through a GPQ for θ (given $R_{\sigma_e^2}$ and R_ρ).

GPQs

Let SSE be the residual sum of squares of the model obtained by ignoring the random effects. Then, the method is based on the following pivotal quantities:

1. $U = \frac{SSE}{\sigma_e^2} \sim \chi^2(N - 2K)$

2.

3.

GPQs

Let SSE be the residual sum of squares of the model obtained by ignoring the random effects. Then, the method is based on the following pivotal quantities:

1. $U = \frac{SSE}{\sigma_e^2} \sim \chi^2(N - 2K)$

2. $Q = \sum_{i=1}^{K-1} \frac{Q_i}{\rho + \lambda_i \sigma_e^2} \sim \chi^2(K - 1)$ → **ρ is allowed to take negative values**

3.

GPQs

Let SSE be the residual sum of squares of the model obtained by ignoring the random effects. Then, the method is based on the following pivotal quantities:

1. $U = \frac{SSE}{\sigma_e^2} \sim \chi^2(N - 2K)$

2. $Q = \sum_{i=1}^{K-1} \frac{Q_i}{\rho + \lambda_i \sigma_e^2} \sim \chi^2(K - 1)$

3. $Z = \sqrt{\mathbf{1}_K'[\mathbf{G}(\sigma_\ell^2, \sigma_e^2)^{-1}]\mathbf{1}_K}(\tilde{\theta} - \theta) \sim N(0, 1)$

GPQs

Let SSE be the residual sum of squares of the model obtained by ignoring the random effects. Then, the method is based on the following pivotal quantities:

1. $U = \frac{SSE}{\sigma_e^2} \sim \chi^2(N-2K)$

2. $Q = \sum_{i=1}^{K-1} \frac{Q_i}{\rho + \lambda_i \sigma_e^2} \sim \chi^2(K-1)$

3. $Z = \sqrt{\mathbf{1}_K'[\mathbf{G}(\sigma_\ell^2, \sigma_e^2)^{-1}]\mathbf{1}_K}(\tilde{\theta} - \theta) \sim N(0,1)$

Note: These independent pivotal quantities do not require dealing with the variance of the random components separately.

Monte Carlo sampling algorithm

For every set (U, Q, Z) of simulated values, we obtain a set of realizations for the parameters of interest:

$$(R_{\sigma_e^2}, R_\rho, R_\theta)$$

This process is repeated M times.

Then, for each realization set, we compute the quantity required to get the TI of interest. For instance:

- For an upper one-sided (β,γ)-TI, we compute

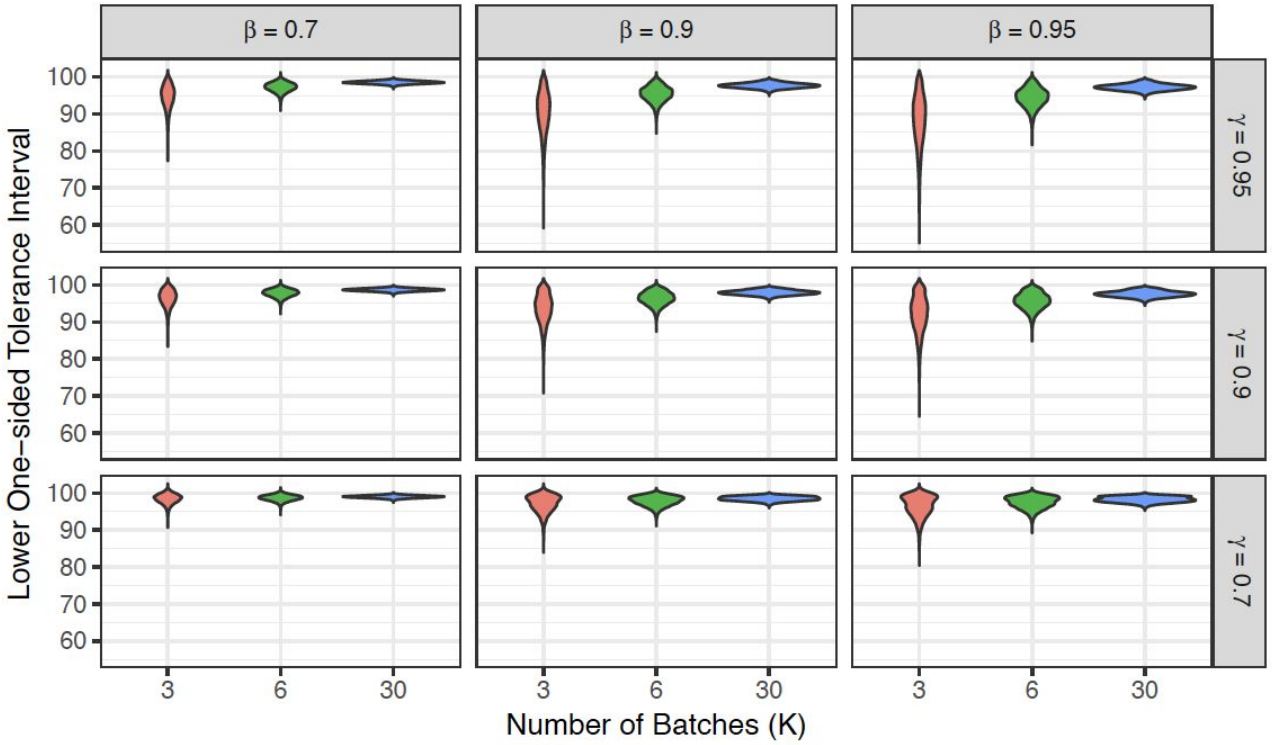
$$R_\delta = R_\theta + z_\beta \sqrt{\max(0, R_\rho + cR_{\sigma_e^2})}$$

- Then, the TI is obtained by computing the γth quantile of the R_δ set.

Simulations

Size of TIs

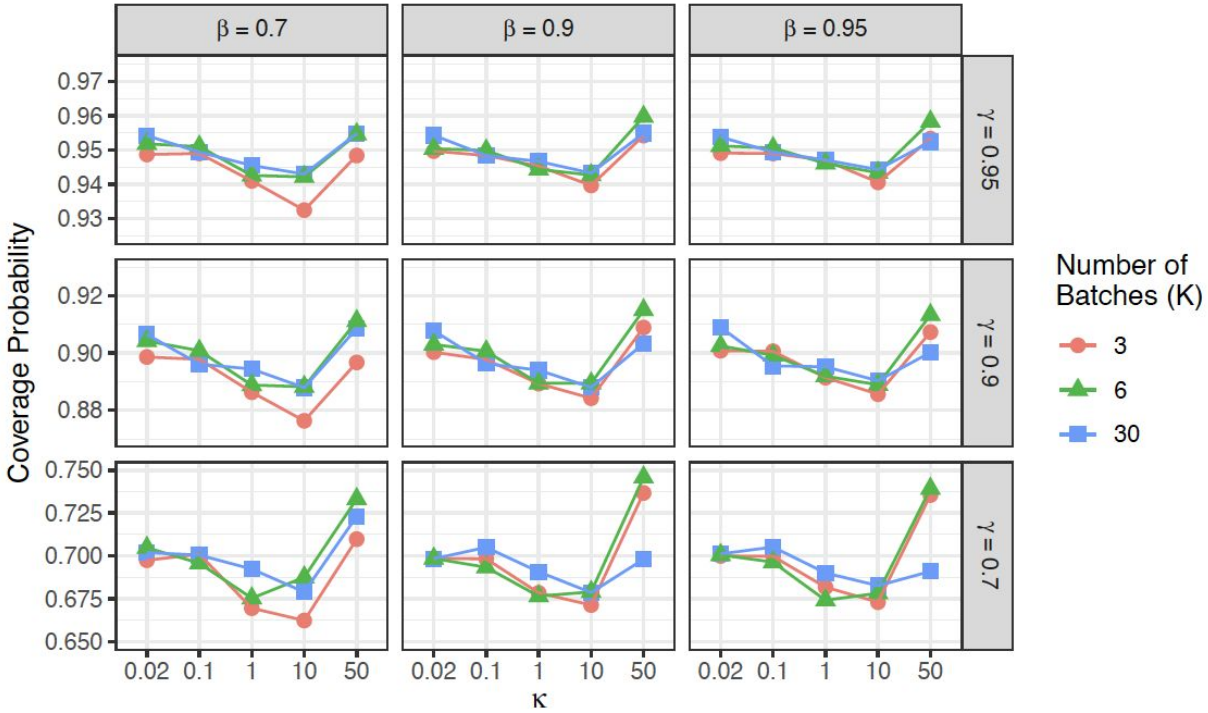
Simulate 10k datasets



Coverage Probabilities are accurate (c=0)

The proposed method worked as expected.

- Coverage probabilities matched their nominal levels.



Genentech
A Member of the Roche Group

Case Study

PQRI Industry Example Data Set (unbalanced)

Some timepoints in the data were removed to build an unbalanced dataset:

Batches 1,4,7:

T=(0,3,6,9,12)

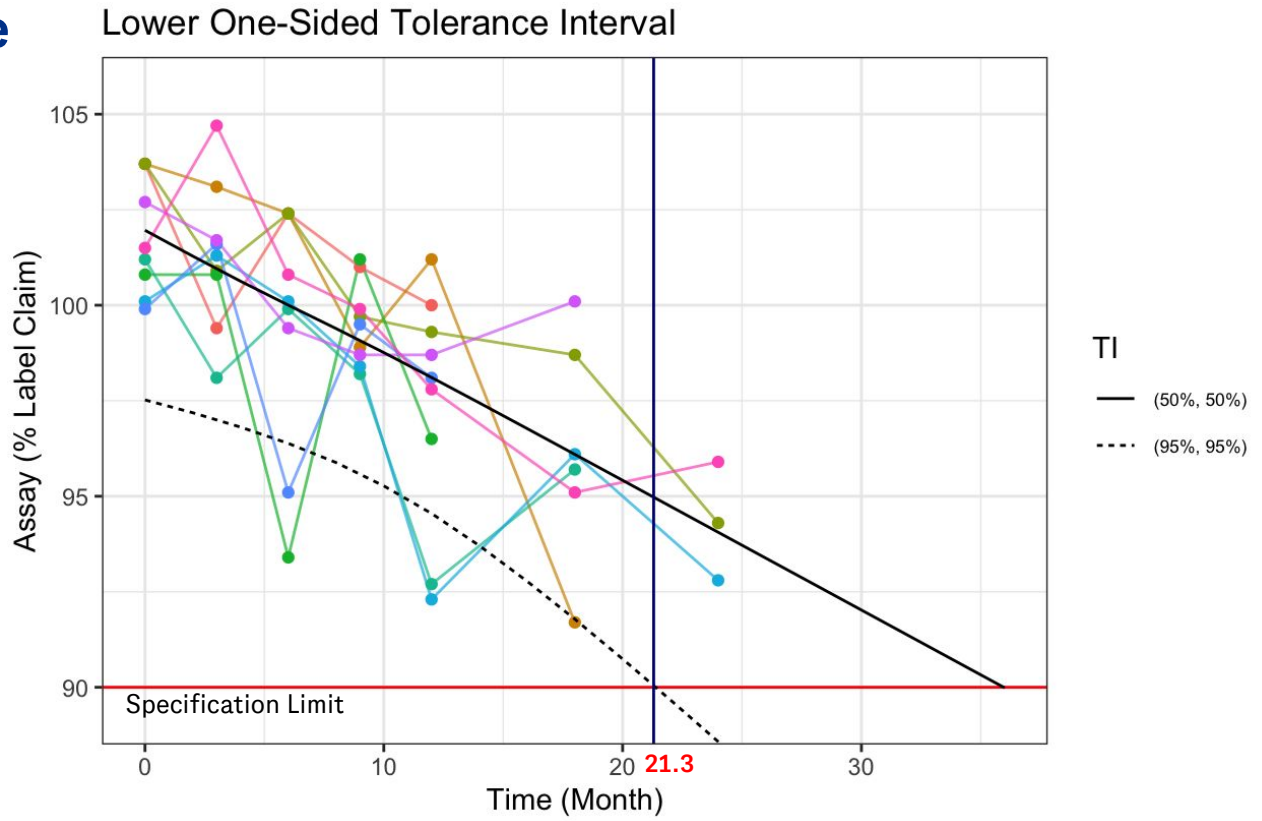
Batches 2,5,8:

T=(0,3,6,9,12,18)

Batches 3,6,9:

T=(0,3,6,9,12,18,24)

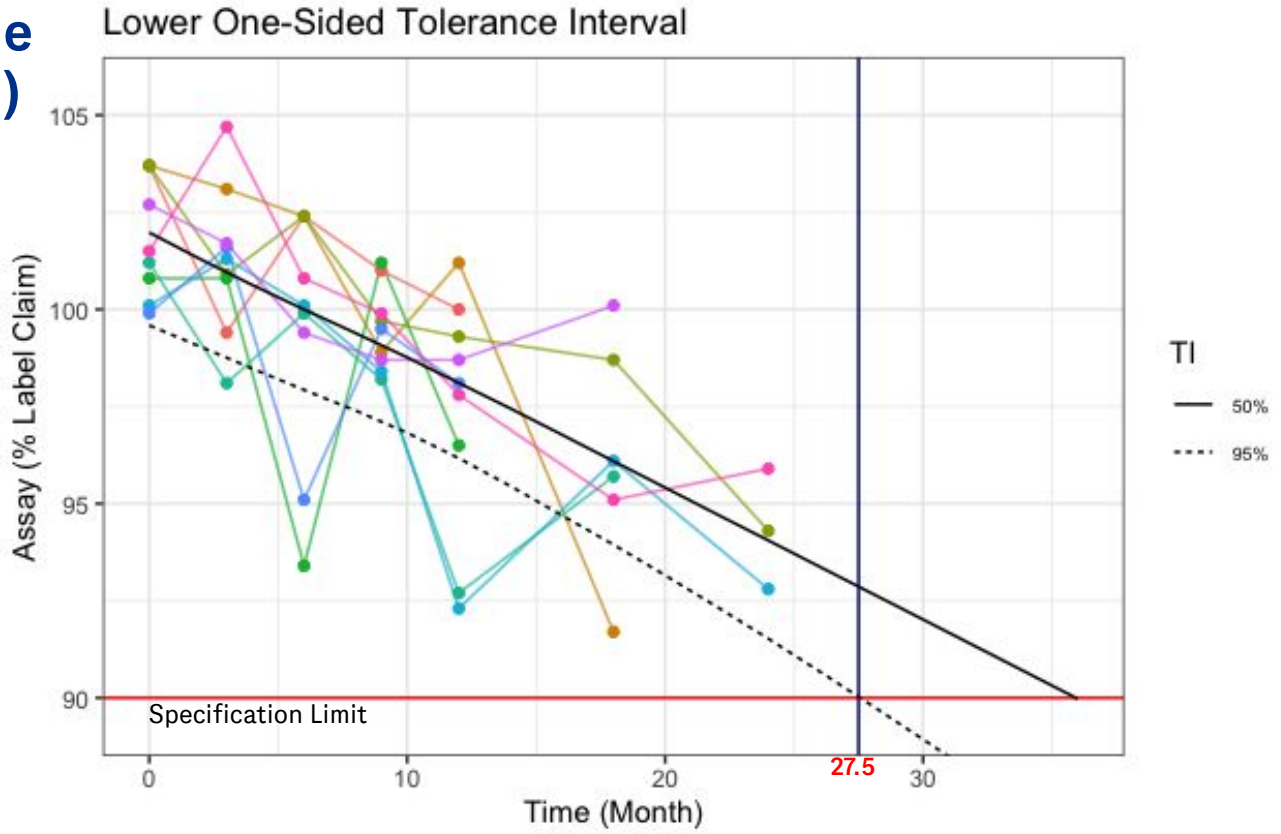
Shelf-life estimate based on (95%, 95%)-TI for batch mean distribution: **21.3m**



PQRI Industry Example Data Set (unbalanced)

β-expectation tolerance intervals were also computed using the proposed method.

β = 50% and 95%.



Shelf-life estimate based on 95%-expectation TI for the batch mean distribution:

27.5m

Conclusions

Summary

We **developed** a GPQ-based method to compute TIs for a class of unbalanced LMMs.

- It relies on obtaining **realizations** of the parameters of interest. The computational burden is **light**.
- It covers the standard LMMs frequently used to analyze **stability data.**
- The method does **not** require estimation of each of the variance components.

Thanks for your attention!

Doing now what patients need next

Genentech

A Member of the Roche Group