PharmaSUG 2022 – Paper MD-174

Measuring Reproducibility and Repeatability of an AI-based Quantitative Clinical Decision Support Tool Having a Medical Decision Point

Douglas A. Milikien, Accudata Solutions, Inc.

ABSTRACT

Artificial Intelligence and machine-learning based methods have led to a rapid expansion of software products used in medical diagnostics. These tools are most frequently intended not directly for diagnosis, but as supporting information for the clinician to arrive at a diagnosis and referred to as Clinical Decision Support Tools. Methods for quantifying the measurement agreement of these applications to gold standard measurement is well known. What is less well-known are methods for quantifying the reproducibility and repeatability of quantitative assessments when those assessments involve measurement variation due to 1) case difficulty, 2) operator skill level or judgment, and 3) stability over time. Although generic guidelines exist (CLSI EP05-A3) for designing and measuring precision experiments, these guidelines assume that there are multiple replicates for each combination of experimental conditions. The practicalities of using these machine-learning platforms often prevent the collection of more than one replicate per experimental condition. This paper illustrates visual methods for exploring the contributions of Case, Operator, and Time to measurement variability and estimates Variance Components and their standard error using a Mixed Model. Of particular interest are values close to a medical decision point and the classification of cases by the software as positive or negative based on that medical decision point. Also of interest is the repeatability of those binary classifications.

INTRODUCTION

Artificial Intelligence and machine-learning based methods have led to a rapid expansion of software products used in medical diagnostics. These tools, referred to as Clinical Decision Support Tools or Clinical Decision Support Systems, are most frequently intended to provide supporting information for the clinician to arrive at a diagnosis – not directly for diagnosis. Some CDS provide a binary classification regarding the presence or absence of a condition, such as qER (Qure.ai Mumbai, India) that identifies cranial bleeds, fractures, and other abnormalities from a head CT scan. Other CDS automatically provide an image-based quantitative measurement to use as an aid in diagnosis, e.g. HeartFlow FFR_{CT} (HeartFlow, Redwood City, CA) computes coronary blood flow (fractional flow reserve) in coronary arteries and vessels using coronary CT angiography (CCTA). These AI-based automated systems offer savings of time when compared to physician diagnosis alone, as well as potentially spare the patient from an invasive diagnostic procedure.

The accuracy of these CDS when compared to the diagnostic standard are of paramount importance in obtaining regulatory clearance. Specifically, sensitivity, specificity, negative predictive value, and positive predictive value in the case of a binary classifier, and difference scores and estimates of bias in the case of quantitative outcomes.

Of secondary importance for regulatory clearance are the aspects of repeatability and reproducibility – how consistent the provided software outcome is, whether binary or quantitative, across different operational characteristics. Demonstrating repeatability and reproducibility is often necessary because of the human component involved in the image analysis and preparation for usage of the CDS software. These human steps can include:

- Manual exclusion of images that have severe artifacts or excess noise,
- Modification or correction of the “mask” that the software overlays on an image to identify anatomical characteristics of interest.

Therefore, operator training and skill level influence the accuracy of measurements. Two additional variables that influence accuracy are the different brands of image acquisition machinery as well as the severity of disease represented by the image.

This paper will address the aspects of powering and analyzing repeatability and reproducibility experiments for purposes of regulatory submission; as well as illustrate methods to overcome, where possible, the common challenge in which there is only one replicate per combination of experimental factors.

CASE STUDY

For purposes of illustrating the methods, consider an example based on a real, investigational, AI-based Clinical Decision Support Tool. This tool, call it MyCardioView, uses coronary CT angiography (CCTA) scans and quantifies – by measurement – a functional characteristic at several anatomical locations on the image. Measurements can range from 0 to 1.0. For simplicity, we will refer to it as the HRT score. In clinical practice, HRT values ≤ 0.8 are evidence for the presence of disease while values > 0.8 indicate the absence of disease; making the value of 0.8 a medical decision point.

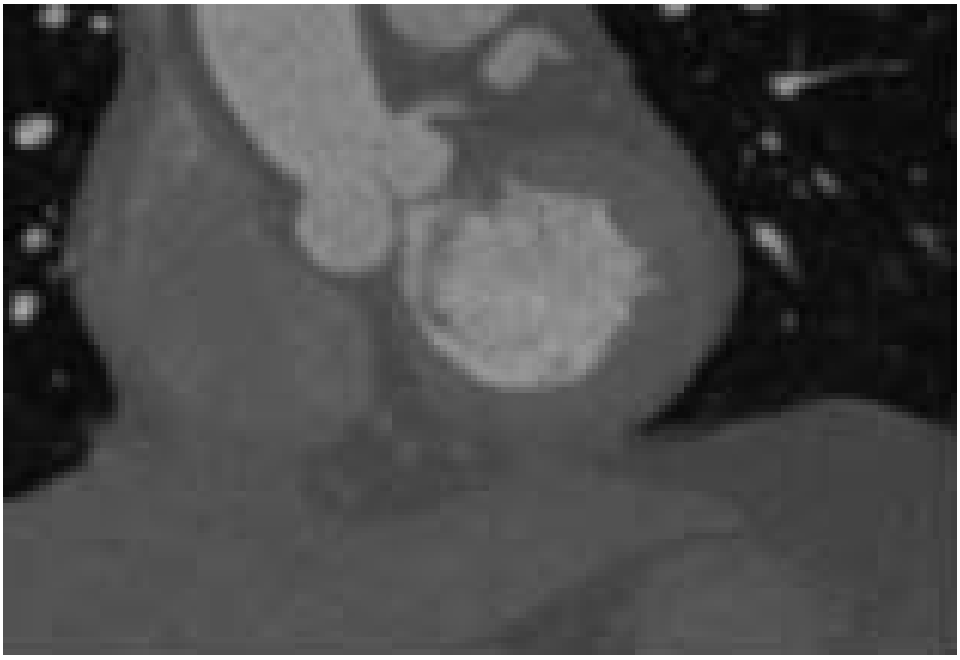

Figure 1. Coronary CT Angiogram (CCTA) (use permission provided by owner)

Measurement of HRT in clinical practice requires an invasive diagnostic procedure. Comparing the HRT score obtained by MyCardioView to the score obtained by the invasive procedure during clinical trials demonstrated the accuracy of MyCardioView.

In addition to the demonstration of accuracy, it is also necessary for regulatory clearance to demonstrate the repeatability and reproducibility of the measurement over factors that influence measurement accuracy. As the software requires annotating the image by the analyst, analyst skill level as well as analyst consistency over time were factors considered for demonstration of repeatability and reproducibility.

The Repeatability and Reproducibility experiment conducted to explore the contributions of these two factors used a set of 20 CCTA scans representing a wide range of HRT values. It is in the interest of the tool developer to demonstrate to the regulatory agency that there is consistency of measurement across different users, and with the same user over time, as long as users are adequately trained in the operation of the software. Therefore, the developer created a rigorous and documented training protocol for all potential users. In the case of MyCardioView, the term “analysts” only apply to those who are successful in that training program. Three trained image analysts used the software to compute HRT values for each of the 20 images on different days, with a period of 14 days between experiment days. All three analysts used the same set of 20 images, creating a data layout like this:

	Analyst 1			Analyst 2			Analyst 3		
CaseID	**Day 1**	**Day 2**	**Day 3**	**Day 1**	**Day 2**	**Day 3**	**Day 1**	**Day 2**	**Day 3**
1	0.78	0.81	0.82	0.77	0.79	0.76	0.79	0.80	0.81
2	0.81	0.89	0.89	0.86	0.86	0.84	0.86	0.86	0.86
3	0.86	0.86	0.84	0.83	0.86	0.85	0.85	0.84	0.87
4	0.77	0.78	0.76	0.73	0.76	0.77	0.77	0.78	0.78
5	0.69	0.68	0.68	0.69	0.71	0.71	0.69	0.69	0.68
6	0.80	0.78	0.77	0.80	0.79	0.80	0.80	0.80	0.80
7	0.89	0.90	0.90	0.90	0.91	0.91	0.91	0.90	0.91
8	0.78	0.78	0.76	0.76	0.76	0.75	0.77	0.76	0.75
9	0.81	0.82	0.81	0.83	0.83	0.83	0.85	0.84	0.84
10	0.82	0.80	0.78	0.83	0.81	0.76	0.81	0.79	0.78
11	0.86	0.86	0.86	0.86	0.87	0.85	0.86	0.85	0.86
12	0.82	0.80	0.83	0.81	0.85	0.82	0.82	0.84	0.83
13	0.93	0.90	0.92	0.94	0.90	0.90	0.91	0.93	0.87
14	0.84	0.83	0.84	0.84	0.85	0.85	0.84	0.85	0.84
15	0.89	0.90	0.89	0.91	0.90	0.90	0.90	0.90	0.90
16	0.76	0.71	0.73	0.76	0.74	0.78	0.70	0.73	0.76
17	0.83	0.84	0.84	0.84	0.84	0.85	0.85	0.85	0.83
18	0.77	0.76	0.79	0.79	0.80	0.79	0.78	0.78	0.81
19	0.79	0.82	0.82	0.80	0.80	0.79	0.79	0.78	0.79
20	0.72	0.69	0.72	0.68	0.68	0.68	0.66	0.69	0.69

Table 1. Sample HRT Values Calculated from 20 Images by Three Analysts Over Three Days using MyCardioView

EXPLORATORY STATISTICS AND GRAPHICS

The goal of repeatability and reproducibility experiments is to identify whether suspected sources of variation in the outcome are within acceptable limits to use the product for its intended medical purpose. A typical descriptive statistic in quantifying the amount of variation in any one factor is the coefficient of variation, often expressed as %CV, where:

$$\%CV_{factor} = 100 * SD_{factor} / \text{mean}$$

For example, in this experiment, for each case (image), we could calculate a %CV due to Analyst by collapsing the data to one mean HRT value per Analyst.

$$\%CV_{Analyst,case\ i} = 100 * SD_{Analyst} / \text{mean}_i$$

Then we obtain $SD_{Analyst}$ by taking the SD of the three analyst values and divide the result by the mean HRT score for the case and multiply by 100.

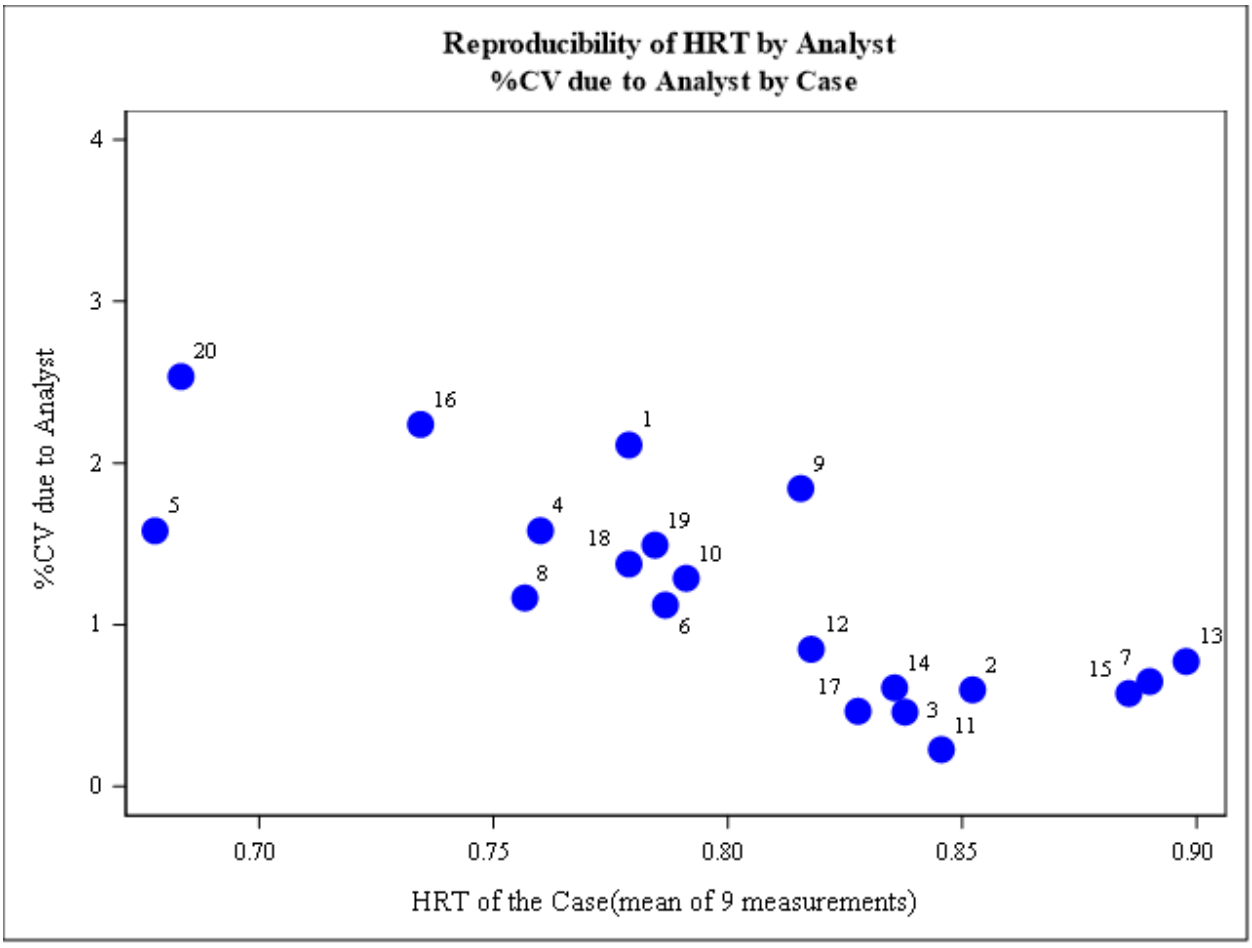

Figure 2. Descriptive Estimates of Reproducibility of HRT Due to Analyst

In this example, each case (image) has a %CV due to Analyst of less than 3%, so there is little analyst-to-analyst variability. In general, the higher the HRT value, the lower the %CV due to Analyst. This is the expected outcome since higher HRT values represent less disease and reduces the Analyst skill factor. These same steps can also create a data set for the Reproducibility by Day.

ESTIMATING VARIANCE COMPONENTS

Although helpful in understanding sources of variation and their patterns, it's not enough to describe variation. To obtain regulatory clearance to market this Clinical Decision Support Tool, we need to make inferences from the sample to the population of all future uses of this tool and compare those inferences to pre-specified acceptance criteria. Therefore, we need statistical models that *simultaneously* estimate the contributions to variance from Case, Analyst, and Day. These models are termed Variance Components Models or Random Effects Models.

First, it is necessary to get the data into a tall-skinny format so that there is one record per combination of Case, Analyst, and Day. For simplicity, I have represented the HRT value by the variable VAL. The code in PROC MIXED looks like:

```
ods output covparms=VarEst;
ods output SolutionF=SolutionF ;

PROC MIXED DATA=TALL COVTEST CL ;
    CLASS CASEID ANALYST DAY ;
    MODEL VAL= /SOLUTION CL ;
    RANDOM INTERCEPT ANALYST /SUBJECT= CASEID ;
RUN;
ods output close ;
```

Note the following about this model:

1. Since no variable appears to the right of the equal sign in the MODEL statement, all effects are RANDOM. In other words, our inference isn't about these specific 3 analysts, and these specific 3 days, but rather *populations* of analysts, and days, from which we have sampled.
2. Since there is only 1 CCTA image per combination of case, image, and analyst, there is no replication, from which we could estimate a true error term. Therefore, one of the model factors must be contaminated with the error term. We choose to represent the Day term as being equivalent to the error term.
3. The VAREST dataset contains the variance estimates and their 95% confidence intervals.
4. The SolutionF dataset contains the grand mean HRT value as the estimate for the INTERCEPT.

To construct the %CV due to each cause, we merge the VarEst and SolutionF datasets.

```
data Process_Overall ;
 set VarEst ;


          SD_Resid = sqrt(Estimate) ;
          SD_Lower = sqrt(Lower) ;
          SD_Upper = sqrt(Upper) ;

run ;
proc sql;
     create table wide_overall as
     select a.CovParm, a.SD_Resid, a.SD_Lower, a.SD_Upper,
           b.Estimate as GrandMean,
                100 * ( a.SD_Resid/b.Estimate) as Pct_CV_point ,
                100 * ( a.SD_Lower/b.Estimate) as Pct_CV_lower ,
```

```
            100 * ( a.SD_Upper/b.Estimate) as Pct_CV_upper

CASE
      WHEN CovParm="Intercept" then 'CASE'
      WHEN CovParm="ANALYST" then 'ANALYST'
      WHEN CovParm="Residual" then 'DAY'
      ELSE ''
END
      as SOURCE

from Process_Overall as a,
       SolutionF as b
;
quit ;
```

Table 2 presents the estimated %CVs for each factor resulting from the model. For this particular product, a %CV of 6.0% was chosen as the acceptance criterion, i.e., the upper bound of the 95% confidence interval for each %CV under investigation was required to be less than 6.0% to be sufficiently reproducible. In this case, both Day and Analyst pass the reproducibility criterion.

Factor	Estimated %CV	Lower 95% ci	Upper 95% ci	Acceptance Criterion	PASS/ FAIL
Case	7.69	5.83	11.27	n/a	n/a
Analyst	0.61	0.32	3.66	6.0	PASS
Day	1.85	1.64	2.11	6.0	PASS

Table 2. Estimated %CV by Factor in HRT Measurement

ESTIMATING VARIANCE COMPONENTS FOR EACH CASE

Guidance CLSI EP05-A3(CLSI 2014) provides detailed coverage of the design and analysis of experiments to characterize Repeatability and Reproducibility of a quantitative diagnostic as well as detailed definitions for Repeatability and Reproducibility. To paraphrase the guidance,

Repeatability is measurement precision under the condition of replicate measurements within a short period of time, with the replicate measurements made using the same operator, location, and measuring equipment.

Reproducibility is measurement precision comparing measurements taken under conditions allowed to change, (e.g., different operators, laboratories, runs, lots of material, equipment) and others taken under unchanged conditions.

Specifically, using the nomenclature of EP05-A3(Section 4.6.2),

$$\text{Standard deviation due to Repeatability} = \sqrt{VarComp_{error}}$$

and

$$\text{Standard deviation due to Reproducibility} = \sqrt{VarComp_{Analyst} + VarComp_{Day} + VarComp_{error}},$$

where VarComp is the estimated variance component for the model term in question from the 2-way MIXED model obtained using the ODS Table name COVPARMS.

In the case of our experiment that has no replication, there is no Repeatability component and the Day effect is contaminated with Error, so the Reproducibility simplifies to:

$$\text{Standard deviation due to Reproducibility} = \sqrt{VarComp_{Analyst} + VarComp_{Day}}$$

and

$$\%\text{CV due to Reproducibility} = 100 \times \frac{\sqrt{VarComp_{Analyst} + VarComp_{Day}}}{Grand\ Mean\ HRT}$$

Since the CCTA images selected for this experiment deliberately represented a wide range of disease severity seen in clinical practice, it makes sense to examine reproducibility one case at a time. Cases with large reproducibility CV may indicate particularly difficult cases for the algorithm.

The model is similar to the overall model:

```
ods output covparms=VarEst_by_Case ;
ods output Type3=ANOVA_by_Case;

PROC MIXED DATA=tall COVTEST CL NOINFO NOBOUND method=type3 ;
    BY caseid ;
    CLASS ANALYST DAY ;
    MODEL VAL= /SOLUTION CL ;
    RANDOM ANALYST;
RUN;
ods output close ;
```

We can compute

$$VarComp_{Reproducibility} = VarComp_{Analyst} + VarComp_{Day},$$

provided that both components are non-negative. A negative variance component is set to zero for purposes of this summation.

Table 3, below, presents a sampling of the %CVs for Analyst, Day, and Reproducibility by Case. (Space limitations prevents display of all results.)

CaseID	Mean DVFFR	Source of Variation	Variance Component Estimate*	Estimated %CV
1	0.792	Analyst	0.00019	1.72
		Day	0.00026	2.02
		Reproducibility (Analyst + Day)	0.00044	2.65
2	0.859	Analyst	-0.00023	.
		Day	0.00076	3.20
		Reproducibility (Analyst + Day)	0.00076	3.20
…	…	…	…	…
20	0.690	Analyst	0.00023	2.21
		Day	0.00020	2.05
		Reproducibility (Analyst + Day)	0.00043	3.02

Table 3. %CV by Case as Estimated from Mixed Model

REDUCING A QUANTITATIVE OUTCOME TO A BINARY CLASSIFIER

In clinical practice, an HRT value ≤ 0.80 may suggest presence of a disease state requiring an interventional procedure. Therefore, it is necessary to know whether the software's binary classification of a CCTA case (Positive/Negative for disease state requiring intervention) is consistent across days and

analysts. To do this, reduce the quantitative HRT value to a binary category (≤ 0.80, > 0.80), where values ≤ 0.80 are considered “Positive” and values > 0.80 are considered negative.

For each case, there are nine opportunities (three analysts over three days) for binary classification of the image. A measure of binary reproducibility, then, is the proportion of those nine opportunities in which the image was classified as Positive or Negative. Images classified as Positive in 100% of the experimental conditions are evidence of high reproducibility, as are images classified as Negative in 100% of the experimental conditions. On the other hand, images for which classifications are 50% Positive and 50% as Negative are indicative of either low reproducibility or underlying HRT values that are very close to the cutoff value.

For the data in our experiment, the proportions classified as Positive and Negative for each case can summarize nicely with a stacked bar graph. After creating a dummy variable DISEASE that takes on values of Positive or Negative corresponding to HRT≤ 0.80 and HRT> 0.80, respectively, the following SAS code produces the stacked bar graph:

```
proc freq data=tall noprint;
    by caseid;
    tables DISEASE / out=FreqOut;
run;

ods rtf file="&path.\Classification_by_Case.rtf" bodytitle notoc_data ;


proc sgplot data=FreqOut;
    vbar caseid / response=Percent group=disease groupdisplay=stack;
    xaxis discreteorder=data;
    yaxis grid values=(0 to 100 by 10) label="Percentage of Total within
Case";
run;
ods rtf close ;
ods listing;
```

Figure 3, below, displays the stacked bar graph showing, for each case, the proportion of experimental conditions that produced a Positive or a Negative classification. In this example, Cases 1, 10, and 19 showed evidence of inconsistency with Percent Negative ranging from 44% to 56%. Upon closer inspection, this is expected because these cases had average HRT values of 0.792, 0.798, and 0.798 – very close to the medical decision cut point.

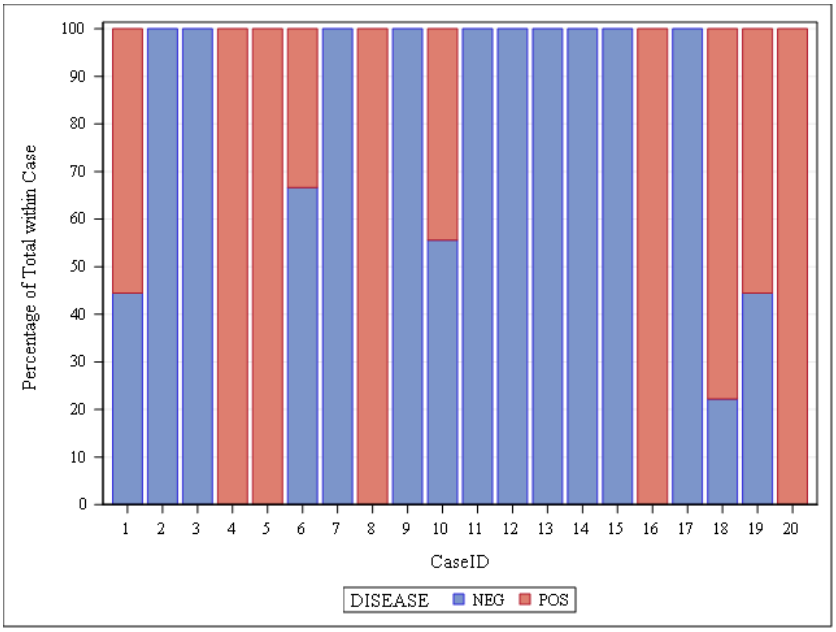

Figure 3. Percentage of HRT Reads Classified Positive or Negative for Disease State Requiring Intervention

CONCLUSION

The rapid and increasing development of AI-based quantitative Clinical Decision Support Tools in medical diagnostics will drive a need for clinical validation studies for regulatory clearance. These studies will require demonstration of not only sufficient accuracy, but sufficient repeatability and reproducibility to warrant regulatory approval. The methods covered in this paper illustrate exploratory, descriptive, and modelling statistical techniques to analyze repeatability and reproducibility and compare them to pre-determined acceptance criteria. Furthermore, when decisions for medical intervention revolve around a cut point or threshold value of a quantitative indicator, these techniques are useful for exploring repeatability or reproducibility on a case-by-case basis. Additionally, these techniques can also apply to situations in which the Repeatability and Reproducibility experiment has only one replicate per combination of experimental factors – a common limitation of this type of Clinical Decision Support Tool.

REFERENCES

CLSI. 2014. “Evaluation of Precision of Quantitative Measurement Procedures; Approved Guideline – Third Edition”. *CLSI document EP05-A3.* Clinical and Laboratory Standards Institute, 34:13

ACKNOWLEDGMENTS

The author is grateful to Jesse A. Canchola for advice on modelling and Lorrie Nicoles for editing and manuscript preparation.

CONTACT INFORMATION

The author appreciates all comments and questions, feel free to contact the author at:

Douglas A. Milikien
Accudata Solutions, Inc.
dougm@accudatasolutions.com
www.accudatasolutions.com