Appraisal of within- and between-laboratory reproducibility of non-radioisotopic local lymph node assay using flow cytometry, LLNA: BrdU-FCM: Comparison of OECD TG429 performance standard and statistical evaluation

Hyeri Yang a, Jiyhe Na a, Won-Hee Jang a, Mi-Sook Jung a, Jun-Young Jeon a, Yong Heo d, Kyung-Wook Yeo d, Ji-Hoon Jo d, Kyung-Min Lim a,*, Seungjin Bae a, **

a College of Pharmacy, Ewha Womans University, Seoul 120-808, Republic of Korea
b Medical Beauty Research Division, Amorepacific Corp. R&D Center, Republic of Korea
c Pharmacology Efficacy Team, Biotoxtech Co., Ltd., Republic of Korea
d Department of Occupational Health, College of Natural Sciences, Catholic University of Daegu, Daegu, 712-702, Republic of Korea

HIGHLIGHTS

• Here, we assessed within- and between-laboratory reproducibility of LLNA:BrdU-FCM, a non-radioisotopic analog of LLNA.
• We compared the criteria given by OECD TG429 and formal statistical methodologies.
• We found that reproducibility may be assessed more rigorously through the application of statistical methods.

ARTICLE INFO

Article history:
Received 31 December 2014
Received in revised form 15 February 2015
Accepted 18 February 2015
Available online 27 February 2015

Keywords:
Local lymph node assay
LLNA:BrdU-FCM
Within and between laboratory reproducibility
OECD TG performance standard
Descriptive and inferential statistics

ABSTRACT

Mouse local lymph node assay (LLNA, OECD TG429) is an alternative test replacing conventional guinea pig tests (OECD TG406) for the skin sensitization test but the use of a radioisotopic agent, 3H-thymidine, deters its active dissemination. New non-radioisotopic LLNA, LLNA:BrdU-FCM employs a non-radioisotopic analog, 5-bromo-2’-deoxyuridine (BrdU) and flow cytometry. For an analogous method, OECD TG429 performance standard (PS) advises that two reference compounds be tested repeatedly and ECT(threshold) values obtained must fall within acceptable ranges to prove within- and between-laboratory reproducibility. However, this criteria is somewhat arbitrary and sample size of ECT is less than 5, raising concerns about insufficient reliability. Here, we explored various statistical methods to evaluate the reproducibility of LLNA:BrdU-FCM with stimulation index (SI), the raw data for ECT calculation, produced from 3 laboratories. Descriptive statistics and graphical representation of SI alone could illustrate the within- and between-laboratory reproducibility. Inferential statistics employing parametric and non-parametric methods drew similar conclusion. While all labs passed within- and between-laboratory reproducibility criteria given by OECD TG429 PS based on ECT values, statistical evaluation based on SI values showed that only two labs succeeded in achieving within-laboratory reproducibility. For those two labs that satisfied the within-lab reproducibility, between-laboratory reproducibility could be also attained based on inferential as well as descriptive statistics.

© 2015 Elsevier Ireland Ltd. All rights reserved.

1. Introduction

Murine local lymph node assay (LLNA) is an OECD-endorsed alternative test method to evaluate the skin sensitization (OECD, 2010a), replacing conventional guinea pig tests (OECD, 1992).
Although LLNA is still “in vivo” and should be an interim measure to be replaced by true in vitro alternatives in the future, in vitro alternatives currently available or to be available in the foreseeable future as exemplified by hCLAT, direct peptide reactivity assay (DPRA) or Keratinosens®5, are not expected to replace LLNA for a substantial period of time since they do not overcome the intrinsic limitation, i.e., will not be able to transcend the realm of hazard identification. Ultimately, skin sensitization test method must be able to produce relative potency data that can contribute to the risk assessment/management (Kimber et al., 2001). In this context, LLNA has an unmatched merit in its capability to produce dose-response relationship data that can provide invaluable information for risk assessment.

Original LLNA method employs the radioisotopic 3H-thymidine to quantitate proliferating lymph node cells upon exposure to test substances. Due to the in vivo use of a radioisotope with a long physical half-life, however, its wide diffusion is hampered since many countries are implementing strict regulations regarding the disposal of radioactive wastes. To circumvent this issue, newly developed LLNA:Brdu-FCM uses a non-radioisotopic thymidine analog, 5-bromo-2′-deoxyuridine (Brdu) and detects the Brdu-incorporated lymph node cells through antibody-assisted flow cytometric method (Jung et al., 2012, 2010).

As with the case of original vs. generic drugs, OECD test guideline recommends that a new analogue or me-too test method should be evaluated for its equivalence to the original test method in reproducibility and predictive capacity based on the pre-determined criteria described in performance standard (PS) as appended to the original test guideline). To evaluate the within-laboratory reproducibility of a new test method to LLNA, OECD TG429 PS advises that one reference (positive) compound (hexylcinnamaldehyde, HCA) shall be tested repeatedly four times and ECT values, an estimate of the test material concentration required to produce a stimulation index (SI) of threshold or cutoff for determination of sensitizers, must fall within pre-determined acceptable range (5–20%). To evaluate the between-laboratory reproducibility, two positive reference compounds (HCA and 2,4-dinitrochlorobenzene, DNCB) shall be tested by 3 independent laboratories (that attained within laboratory reproducibility) and ECT values obtained must fall within acceptable range (5–20% for HCA and 0.025–0.01% for DNCB). Up to now two analogous LLNA methods, LLNA:DA (OECD, 2010a) and LLNA:Brdu-ELISA (OECD, 2010b), have been approved by OECD after demonstrating that they satisfy the criteria provided by PS. However, this criteria is somewhat arbitrary and the sample size of ECT is less than 5, raising some concerns about insufficient reliability and statistical power.

Previously, the consistency in ECT as expressed in coefficients of variation has been used as an index for the estimation of within- or between-laboratory reproducibility (Dean et al., 2001). Actually, ECT is a figure summarizing a line constructed by the regression of 3 concentration-SI points (Loveless et al., 1996). Each concentration-SI point is again obtained from 4 to 5 animals, reflecting that an ECT value represents at least 12 SI values. Considering that one point of a concurrent positive control (generally, 25% HCA) is included with 3 concentration points of test article, numerous SI values are generated during each LLNA trial. Actually, SI is an important index by providing a cut-off for the classification of sensitizer (Basketter et al., 1999) and Ehling et al., tried to provide a statistical rationale for determining sensitizers in LLNA employing SI values (Ehling et al., 2005a,b).

Several studies attempted to demonstrate the reproducibility of LLNA or LLNA:Brdu-ELISA through the graphical representation of mean and standard deviation (SD) or 95% confidence interval of SI for each concentration point (Kojima et al., 2011b; Omori et al., 2008; Omori and Sozu, 2007). To obtain representative SI values produced from multiple laboratories, the variance component, $r^2$, which is commonly used in meta-analysis and estimated based on random-effect model for the log-transformed SI, was used to appraise the between-laboratory variation. (Kojima et al., 2011; Omori et al., 2008). Hanke et al. (2001) estimated the within- and between-laboratory reproducibility (or reliability) by calculating the consistency statistics (k and h, respectively) with SI. Kimber et al. (1991) examined between-laboratory reproducibility by applying analysis of covariance with the dose-dependent SI values produced from 4 laboratories (Kimber et al., 1991). Another important and commonly used methodology for estimating reproducibility is to evaluate the consistency in binary decision (non-sensitizer vs. sensitizer) between trials in single laboratory or trials of multiple laboratories (Idehara et al., 2008; Kimber et al., 1998; Scholes et al., 1992) which can be further analyzed by kappa-statistics (Viera and Garrett, 2005).

Unlike inferential statistics which tests the null hypothesis based on the level of significance and the p-value thus rejects or fails to reject the null hypothesis, descriptive statistics only “describe” the magnitude of reproducibility. Incidentally, SI values for a concurrent positive control have limited sample size (N=4–5 in one LLNA trial), common sense indicates that normality assumption may fail; yet considering that baseline characteristics of experimental studies are relatively homogenous (Hothorn, 2014; Na et al., 2014) as contrasted with those of clinical studies, small sample sizes of SI do not necessarily imply that the classification of sensitizer based on SI may be statistically flawed (Basketter et al., 2009) or that parametric method should not be used. However, unlike clinical studies which emphasize the importance of sufficiently large sample sizes and the use of the appropriate statistical methods (Pagano and Gauvreau, 2000), few experimental studies have fully discussed the appropriate application of the statistical methods (Na et al., 2014).

In this study, using the data produced by 3 laboratories, we first investigated the within- and between-laboratory reproducibility of LLNA:Brdu-FCM based on ECT values of HCA and DNCB, according to criteria given by PS of OECD TG429. To further examine the reproducibility, we analyzed the SI data employing descriptive statistics and inferential statistics. Briefly, the mean and SD of all SI values was graphically presented, and the reproducibility in SI values of a concurrent positive control, HCA 25% (Dearman et al., 2001) was investigated based on inferential statistics. We employed both parametric (one-way ANOVA and student t-test) and non-parametric (Kruskal–Wallis and Wilcoxon rank sum test) methods to evaluate the within- and between-laboratory reproducibility along with examining assumptions behind parametric approach (such as test of normality and equal variance assumption), and results obtained from both approaches were compared and discussed.

2. Materials and methods

2.1. Chemicals and reagents

2.4-Dinitrochlorobenzene (DNCB), hexylcinnamaldehyde (HCA) and 5-bromo-2′-deoxyuridine (Brdu) were obtained from Sigma–Aldrich (San Diego, CA, USA). DNCB and HCA were dissolved in acetone:olive oil (AOO; 4:1). Brdu was dissolved in phosphate-buffered saline (PBS) at a concentration of 20 mg/mL.

2.2. LLNA:Brdu-FCM

Both the animal care and study protocol employed were in accordance with Institutional Animal Care and Use Committee (IACUC) of each participating laboratory. LLNA:Brdu-FCM assays were conducted according to previous reports (Jung et al., 2012;
Female Balb/c mice (7–8 weeks old, body weight 18–22 g) were purchased from OrientBio (Seoul, Korea), and were used in all experiments. Groups of mice (N = 4 or 5 per a group) were treated with 25 μl of the test substances (HCA and DNCB) in AOO or AOO alone on the back of both ears daily for 3 consecutive days (days 1–3). On day 5, mice were intraperitoneally injected with BrdU (Sigma–Aldrich). On day 6, animals were humanely sacrificed through CO2 asphyxiation and auricular lymph nodes were isolated, weighed and undergone lymphocyte preparation. Lymph node cells (LNCs) were prepared from lymph node by disintegration through 70 μm mesh (BD Biosciences, Franklin Lakes, NJ) in 1 mL PBS. The LNCs were counted in a hemocytometer after staining with trypan blue. The LNCs (2 × 10⁶ cells/mL) were washed once by centrifugation (300 × g) for 5 min with PBS and re-suspended for fixation and permeabilization step, according to the instruction manual of BrdU Flow kits (BD Pharmingen™, Franklin Lakes, NJ). Then LNCs were permeabilized using Cytoperm plus buffer, which contains 10% DMSO. After DNA was denatured by incubating for 1 h with DNase, LNCs were washed, and incubated with FITC conjugated anti-BrdU antibody for 20 min at RT in the dark at a dilution of 1:50. Cells were washed once more and then re-suspended in 20 μL 7-AAD solution to label DNA. Ten thousand 7-AAD expressing cells were gated, and the number of the cells expressing BrdU was analyzed using BD FACScaliburTM system.

Stimulation index (SI) value was obtained by calculating the ratio of the mean number of LNCs with incorporated BrdU from mice in each of the test substance dose groups to the mean number of LNCs with incorporated BrdU from mice in the vehicle control group.

2.3. Assessment of within- and between-lab reproducibility

2.3.1. Production of data for the assessment of within- and between-lab reproducibility

Three laboratories including a lead lab (AmorePacific R&D center) and two participating labs, (Biotoxtech Co. and Catholic University of Daegu) assessed the skin sensitization potentials of known skin sensitizers, HCA and DNCB according to the protocol developed by the lead lab. For the assessment of within-laboratory reproducibility, HCA was assessed 4 times and for the between-laboratory reproducibility, DNCB was assessed 3 times.

2.3.2. Analysis of within- and between-lab reproducibility based on performance standard of OECD test guideline 429

Firstly, we examined the within- and between-lab reproducibility based on EC1 value according to OECD TG429. The EC1 value is defined as the estimated concentration that yields an SI value of predetermined threshold. Among known methods of estimating EC1 values from the SI values, we used the linear interpolation method (Basketter et al., 1999). Briefly, EC1 value was obtained from an ordinary least square equation based on SI values, which matches an x (concentration) where the y (SI) equals threshold. Acceptable within-lab reproducibility was being recognized when EC1 values of HCA generated from 4 repetitions fall between 5 and 20%. Between-lab reproducibility was accepted based on dual criteria, i.e., EC1 values of HCA must be between 5 and 20% and those of DNCB, between 0.025 and 0.01%.

2.3.3. Descriptive statistics for SI values

Stimulation index, the primary raw data for estimation of EC1 and classification of sensitizer was visually presented as mean ± standard deviation (SD), scatterplots and box-whisker plots.

2.3.4. Inferential statistics with EC1 values and SI values of a concurrent positive control, HCA 25%

The SI values and EC1 values of the positive reference compounds were investigated based on inferential statistics. We employed both parametric (one-way ANOVA and student t-test) and non-parametric (Kruskal–Wallis and Wilcoxon rank sum test) methods to evaluate the within- and between-laboratory reproducibility, and results obtained from both approaches were compared and discussed. Normality assumption was tested based on Shapiro–Wilks test and Kolmogorov–Smirnov test for one-way ANOVA and student t-test (Shapiro et al., 1968), and homoscedasticity (equal variance) assumption was examined for one-way ANOVA based on Bartlett’s test or Levene’s test (Brown and Forsythe, 1974). When relevant assumptions regarding parametric approach fail and the conclusions from parametric/non-parametric approach diverge, then non-parametric approach was more heavily considered. If the results of multiple comparison tests were

Table 1

<table>
<thead>
<tr>
<th></th>
<th>Shown in %</th>
<th>EC2.5</th>
<th>EC2.6</th>
<th>EC2.7</th>
</tr>
</thead>
<tbody>
<tr>
<td>HCA</td>
<td></td>
<td>7.011</td>
<td>7.366</td>
<td>7.722</td>
</tr>
<tr>
<td>DNCB</td>
<td></td>
<td>0.075</td>
<td>0.077</td>
<td>0.078</td>
</tr>
<tr>
<td>Variance (SD)</td>
<td></td>
<td>0.00014</td>
<td>0.00015</td>
<td>0.00016</td>
</tr>
<tr>
<td><strong>DNCB</strong></td>
<td></td>
<td>1.0638</td>
<td>1.1670</td>
<td>1.3162</td>
</tr>
</tbody>
</table>
significantly different \((p < 0.05)\), then parametric Tukey method for one-way ANOVA and non-parametric Dwass–Steel–Critchlow–Fligner method was applied \((\text{Critchlow and Fligner, 1991})\) as post hoc analyses.

2.4. Data analysis and statistics

Data is presented as mean \(\pm SD\) unless described otherwise. All statistical analyses were performed using SAS version 9.3 \((\text{SAS Inc., Cary, NC, USA})\) and \(p < 0.05\) was considered significant.

3. Results

3.1. Determination of optimal ECt value and evaluation of reproducibility based OECD Test Guideline 429

In order to assess the within- and between-lab reproducibility, skin sensitization potential of HCA was evaluated four times and DNCB, three times by 3 labs using LLNA:BrdU-FCM. Three concentrations of HCA \((5, 10\) and 25%) and DNCB \((0.05, 0.1\) and 0.25%) were tested. To obtain an optimal threshold for the determination of EC\(_t\), EC\(_t\) was calculated as the threshold varies from 2 to 4 \((\text{Fig. 1})\). Of those EC\(_t\) values, EC2.5, EC2.6, and EC2.7 for HCA and DNCB fell within acceptable ranges described in OECD TG429 \((5\%\) to 20\% for HCA and 0.025 to 0.1\% for DNCB). The optimal threshold could be selected based on clustering of the EC\(_t\) values near the reference values, i.e., if mean of EC\(_t\) values for HCA and DNCB was close to 10\% and 0.05\%, respectively, and the variances were small, then the corresponding EC\(_t\) could be considered optimal. Based on the mean values, EC2.7 were favored for HCA and EC2.5 for DNCB, but the smaller variance was observed in EC2.5 than EC2.7 for both HCA and DNCB (Table. 1), suggesting that EC2.5 could be considered to be optimal for LLNA:BrdU-FCM. All EC2.5 values for HCA and DNCB fell within the respective acceptable ranges stated in OECD TG429 PS, indicating that LLNA:BrdU-FCM was acceptable in terms of within- and between-lab reproducibility.

3.2. Descriptive statistics for SI values

An EC\(_t\) value summarizes a set of SI values consisting of three concentration points \((\text{Gerberick et al., 2007})\), thus a lot of information regarding the underlying dispersion of SI values is lost. Therefore, an EC\(_t\) value may not be a proper indicator to evaluate the within- or between-lab reproducibility. Visual presentation of SI values may give a more detailed and concrete idea on the variance and reproducibility of LLNA:BrdU-FCM. Fig. 2 shows the scatterplot with mean \(\pm SD\) of SI values obtained from 3 labs for HCA 5, 10 and 25\%. With the plot, substantial level of variance could be more clearly seen \((\text{Fig. 2})\), especially in lab 3.
Fig. 3 represents the scatterplot with mean ± SD of SI values obtained from 3 labs for DNCB 0.05, 0.1 and 0.25%, and PC (25% HCA), and again substantial variation was observed in lab 3.

Fig. 4 shows the SI values for PC obtained during the assessment of within-lab reproducibility in a scatterplot with mean ± SD and a box-whisker plot. Overall, within-lab reproducibility of SI values of PC obtained during 4 trials appeared appropriate except for lab 3 which showed incongruence in value ranges from trial to trial. Also, although SI values of PC are supposed to be higher than the threshold, 2.5, a value obtained at 2nd trial in lab 3 were less than 2.5 (Fig. 4). All SI values obtained during the tests of HCA and DNCB from respective laboratory (N = 33–35) were shown in a scatterplot with mean ± SD and a box-whisker plot in Fig. 5 for the presentation of between-lab reproducibility which showed a sign of difference between lab 1 and lab 3 and one value was below the cut-off, 2.5.

3.3. Inferential statistics for within- and between-lab reproducibility of SI values of PC

To further analyze the reproducibility in the data, inferential statistics was conducted with the establishment of null hypothesis stating that the mean SI values within – or between-lab were equivalent. The hypothesis of within-lab reproducibility in 4 trials was tested based on parametric (one-way ANOVA) as well as non-parametric (Kruskal–Wallis) approach. Equal variance of SI values of the PC for one-way ANOVA were confirmed for all trials from three labs based on both Levene’s test and Bartlett’s test (p > 0.05). The normality assumption regarding within-lab reproducibility was confirmed in lab 1 and lab 3 for both Shapiro–Wilk test and Komogorov–Smimov test, yet the first trial of lab 2 failed based on both tests (p = 0.0309 for Shapiro–Wilk and p = 0.0359 for Komogorov–Smimov test).

Table 2 illustrates that within-lab reproducibility (namely, equivalence in mean SI values of PC) was observed in lab 1 and lab 2 based on both parametric (p = 0.9255 and 0.9040, respectively) and non-parametric (p = 0.9722 and 0.9709) approach, yet significant difference between within-lab trials was observed in lab 3 based on both parametric (p = 0.0004) and non-parametric (p = 0.0020) approach. Post hoc analysis of lab 3 (both parametric and nonparametric) indicated that trial 1 was significantly different from trial 2 or 3; values in trial 2 were clearly lower than other values, not just within lab 3 but compared with values in other labs. Of those two labs (lab 1 and lab 2) whose within-lab reproducibility was attained, the between-lab reproducibility was investigated with all SI obtained (N = 33–35) and the robustness of the results were examined by comparing results of the parametric test (student t-test) and non-parametric test (Wilcoxon rank sum test) as shown in Table 3. Firstly, the assumptions for parametric test were examined which showed that the normality assumption was violated for lab 2 (p = 0.001 for Shapiro–Wilk test and p = 0.01 for Komogorov–Smimov test) but it was not violated for lab 1 (p = 0.0814 for Shapiro–Wilk test and p = 0.15 for Komogorov–Smimov test). Between-lab reproducibility for lab 1 and lab 2 was observed based on both parametric (t-test, p = 0.138) and non-parametric (Wilcoxon rank sum test, p = 0.100), again illustrated consistency.

4. Discussion

With the increasing awareness on animal welfare, many alternative approaches to conventional animal test methods are

---

Fig. 4. SI values of a concurrent positive control (HCA 25%) obtained during within-lab reproducibility. SI values of a concurrent positive control (HCA 25%) obtained during within-lab in [a] a scatterplot with mean ± SD, N = 4–5 and in (b) box-whisker plot. Upper or lower whisker extends to the highest or lowest data value within the upper (Q3 + 1.5 (Q3–Q1)) or lower limit (Q1–1.5 (Q3–Q1)).
Table 2
Summary and statistical analysis of within-lab reproducibility in SI values of HCA 25% obtained during within-lab reproducibility.

<table>
<thead>
<tr>
<th>Lab</th>
<th>Trial</th>
<th>Mean</th>
<th>SD</th>
<th>N</th>
<th>Median (Min. Max.)</th>
<th>Significance (II value)</th>
<th>ANOVA</th>
<th>Kruskal-Wallis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lab 1</td>
<td>1</td>
<td>10.3</td>
<td>3.5</td>
<td>5</td>
<td>10.1(6.3, 15.6)</td>
<td>0.9255</td>
<td>0.9722</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>9.2</td>
<td>2.6</td>
<td>5</td>
<td>7.8(6.8, 12.7)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>10.3</td>
<td>3.5</td>
<td>5</td>
<td>10.2(9.6, 22.8)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>9.6</td>
<td>3.0</td>
<td>5</td>
<td>8.1(7.0, 13.3)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lab 2</td>
<td>1</td>
<td>6.8</td>
<td>2.5</td>
<td>5</td>
<td>5.1(5.0, 9.9)</td>
<td>0.9040</td>
<td>0.9709</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>8.1</td>
<td>3.3</td>
<td>4</td>
<td>7.8(4.3, 12.7)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>7.9</td>
<td>2.5</td>
<td>5</td>
<td>8.5(3.8, 10.7)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>7.6</td>
<td>3.2</td>
<td>5</td>
<td>8.4(2.7, 11.1)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lab 3</td>
<td>1</td>
<td>9.7</td>
<td>2.1</td>
<td>5</td>
<td>9.1(7.2, 12.6)</td>
<td>0.0004*</td>
<td>0.0020*</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>4.7</td>
<td>1.7</td>
<td>5</td>
<td>4.8(2.3, 6.6)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>5.4</td>
<td>1.1</td>
<td>5</td>
<td>5.6(3.7, 6.6)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>7.1</td>
<td>0.7</td>
<td>5</td>
<td>6.8(6.4, 7.8)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*p < 0.05.

(a) Grouped by Tukey’s post hoc analysis.
(b) Grouped by Tukey’s post hoc analysis.
(c) Grouped by Dwass, Steel, Critchlow-Fligner method.
(d) Grouped by Dwass, Steel, Critchlow-Fligner method.

Table 3
Overall summary and statistical analysis of between-lab reproducibility in SI values of a concurrent positive control (HCA 25%).

<table>
<thead>
<tr>
<th>Lab</th>
<th>MEAN</th>
<th>SD</th>
<th>N</th>
<th>Median (Min. Max.)</th>
<th>Significance (p value)</th>
<th>Comparison</th>
<th>Wilcoxon rank sum test</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>t-test</td>
<td></td>
</tr>
<tr>
<td>Lab 1</td>
<td>9.7</td>
<td>4.1</td>
<td>33</td>
<td>9.2(3.2, 11.8)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lab 2</td>
<td>8.2</td>
<td>3.7</td>
<td>34</td>
<td>8.3(2.7, 24.3)</td>
<td></td>
<td>0.138</td>
<td>0.1004</td>
</tr>
<tr>
<td>Lab 3</td>
<td>6.7</td>
<td>3.8</td>
<td>35</td>
<td>6.4(2.3, 22.0)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Fig. 5. Between-lab reproducibility of SI values of a concurrent positive control (HCA 25%).
All SI values of a concurrent positive control (HCA 25%) obtained during the evaluation of within- and between-laboratories reproducibility in (a) a scatterplot with mean ±SD, N = 33–35 and in (b) a box-whisker plot. *Outlier: an unusually large or small observation. Values beyond the whiskers are outliers.

being newly introduced. Newly developed test methods shall be
evaluated to demonstrate reproducibility and relevance of the
test results they produce. For the assessment of reproducibility of
test results, diverse statistical methodologies have been applied as
could be categorized into descriptive statistics, correlation analysis
and hypothesis-driven inferential statistics to examine the
 equivalence in the data produced (Avram et al., 1985). Through
graphically representing the scatter of data, or estimating the
mean or median value with variances or ranges, descriptive
statistics can give an overall picture of the reproducibility or
variance of the data produced (Martinez et al., 2006; Pfannen-
becker et al., 2013; Sheskin, 2003). Correlation analysis of the data
(Vian et al., 1995) or consistency in the decision (Nichols et al.,
2004; Viera and Garrett, 2005) between trial to trial or between lab
to lab can also explain the reproducibility of the test results. Finally,
inferential statistics can provide a decision whether the test results
are reproducible or not through testing the null hypothesis by
means of giving significance level (Ehling et al., 2005b).
Incidentally, these statistical methods sometimes demand a
substantial quantity of data for ensuring relevant analysis which
may overwhelm the original efforts to reduce the sacrifice of
animals. Therefore, flexible standards of reproducibility shall be
applied to an individual test method based on its purpose and
applicability domain.
Analogous, or me-too test methods capitalizing the principle or
basis of an original test significantly contribute to the propagation
of alternative methods, ultimately reducing the sacrifice of animals. OECD test guideline performance standard (PS) provide the basis for the determination of reliability and accuracy of those
similar test methods in comparison with the original method.
Standards for accuracy and reliability that the proposed test
method should meet or exceed are established such that minimum
amount of time, resource and animals is demanded. However, this
minimum level may be somewhat arbitrary and questions have
been raised about ‘the standards’ of PS such as the selection of
reference compounds for predictive ability assessment (Basketter
et al., 2012). And it is necessary to evaluate the criteria of the
reproducibility more thoroughly in comparison with the conven-
tional statistical methodology.

Our study demonstrates that although all three labs met the
criteria provided by PS of OECD TG429 in terms of the within
and between-lab reproducibility, the statistical analyses provided
different results. For example, the descriptive statistics of SI values
raised concerns regarding the within-lab reproducibility of lab 3,
and the inferential statistics showed that the SI values of a
concurrent positive control (PC) from lab 3 were significantly
different, regardless of the types of statistical methodologies
we applied, suggesting that the within-lab reproducibility in terms
of SI of PC may not be guaranteed for lab 3. Since PS of OECD
TG429 was based on EC5 whereas our statistical analyses used SI
values of PC, it is not surprising that the results from the two
studies diverged.

The results of inferential statistics were consistent with those of
descriptive statistics; descriptive statistics illustrated that SI values
of lab 3 were significantly different, since the mean and the median
(Fig. 4) value of 1st trial of lab 3 were higher than the upper limits
of the rest of the trials in lab 3. Inferential statistics showed that the
within-lab reproducibility of lab 3 failed based on parametric (one-
way ANOVA) and non-parametric (Kruskal–Wallis) test, and the post hoc analyses demonstrated that the 1st trial of lab 3 were significantly different from the rest of the trials in lab 3 (both Tukey and Dwass–Steel–Critchlow–Fligner test), which is consistent with the results shown in descriptive statistics. Also, the mean and the median of SI values in lab 3 were generally lower than lab 1 or lab 2 (Table 2), suggesting that including lab 3 would compromise the between-lab reproducibility regardless of the within-lab reproducibility in lab 3.

An interesting point to note is that lab 3 passed both within–and between-lab reproducibility based on OECD TG429 PS. Our analyses suggest that passing the standards of OECD guidelines might not necessarily indicate that reproducibility has been guaranteed; given that OECD guidelines are based on ECT values, which summarize a set of SI values consisting of three concentration points, ECT value might underestimate the dispersion of the distribution, which could influence the evaluation of the reproducibility. The criteria for the within- and between-lab reproducibility is somewhat arbitrary and wide to provide a minimum cut-off point for reproducibility, yet more rigorous standard with clear justification may enhance within- and between-lab reproducibility.

Although many studies have tried to provide statistical grounds for the classification of sensitizers in LLNA (Ehling et al., 2005b; Hothorn and Hasler, 2008; Hothorn and Vohr, 2010), few studies have introduced inferential statistics to the evaluation of reproducibility of LLNA. Our study attempted to compare the parametric and non-parametric approaches in evaluating within– and between-laboratory reproducibility, and the assumptions behind the test statistics (parametric approach) were systematically tested. An interesting point to observe is that the result of parametric test and non-parametric test were consistent; however, given that the normality assumption failed only in the 1st trial of HCA assay of lab 2 and the homoscedasticity assumption was not violated in all three labs, and the ANOVA is usually sensitive to equal variance assumption but less so with normality assumption (Schmider et al., 2010), our result should be interpreted with caution and further analyses should be conducted before generalizing our results. However, given that the test animals are usually tightly controlled thus the baseline characteristics of them not as heterogeneous as clinical studies, the use of parametric approach should not be entirely avoided, given that the assumptions regarding the parametric approach be tested. Also, since the power of normality tests were usually low for the studies with small sample size, which is common for experimental studies, the type I error is likely to be escalated, indicating that the null hypothesis (normality assumption) is more likely to be violated although the null hypothesis is actually true (Razali and Wah, 2011), thus the results should be interpreted with caution.

Even though the appraisal of reproducibility by PS of OECD TG429 and the application of proper statistical methods has diverged, it is beyond the scope of the current study, to conclude which criteria should be preferred to estimate the reproducibility of LLNA. ECT values obtained all fell within the acceptable range as stated in the PS of OECD TG429 and statistical analysis of ECT values failed to show significant difference between within-lab trials or between-lab trials (data not shown) indicating that three labs produced reproducible ECT values. However, since the statistical evaluation of the reproducibility of SI have failed, within- and between-lab reproducibility of LLNA:BrDU–FCM in producing consistent SI is not established. We believe that further studies should be preceded before examining which outcomes should be more critically referred for estimating the within- and between-lab reproducibility, which warrants further studies in the future.

Author’s contributions


Conflict of interest

The authors declare that there are no conflicts of interest.

Transparency document

The Transparency document associated with this article can be found in the online version.

Acknowledgements

This research was supported by Grants (13172MFD5987) from the Ministry of Food & Drug Safety of Korea.

References


