Evaluating Agreement Metrics Used in Analytical Validation: A Simulation Study for a Digital Measure of Step Count

Gerasimos Dumi[1], Paolo Eusebi[2], Dara O'Neill[3], Aleksandra Sjöström-Bujacz[4]
[1]IQVIA, Athens, Greece; [2]IQVIA, Milan, Italy; [3]IQVIA, Barcelona, Spain; [4]IQVIA, Stockholm, Sweden

Background

Analytical validation requires researchers to ensure agreement between the digital measure and a **gold standard** reference measure

Digital measure (e.g., step count)

Establish agreement

Gold standard measure (e.g., manual step count)

Traditional agreement metrics include[1,2]:

A. Primary metrics

- Mean absolute percent error
- Concordance correlation coefficient (CCC) based on linear mixed model (LMM)

B. Supplementary metrics

- Pearson's correlation
- T-test

However, these metrics are designed for continuous data, which may not be appropriate for count-based parameters

So, we propose an **alternative metric**[3]:

- CCC based on Poisson generalized linear mixed model (GLMM))

Objectives

- A simulation study was conducted with the following objectives:
- **Objective 1:** Evaluate the performance of CCC based on a Poisson[3] GLMM versus a LMM[2] as an agreement metric for count data
- **Objective 2:** Compare CCC with alternative approaches originally designed for continuous data, such as Pearson's correlation and t-test, and with commonly used metric of mean absolute percentage error (MAPE) to assess inference consistency under different levels of agreement (true CCC<0.7 and true CCC≥0.7)

Methods

- **Data-generating mechanism:** A Poisson GLMM (assuming this is a more representative distribution than normal distribution for step count) with random subject and fixed measurement method effects was used to simulate step counts (500 datasets) on selected parameters (see **Table 1**). The selected CCC values were closely aligned with observed values from previous analytical validation studies for step counts (e.g., 0.991[4], 0.837[5], 0.750[5], 0.620[6], 0.570[6])

Table 1: Selected parameters of the GLMM model

Intercept (b_0)	Systematic bias (b_1)	Sample size (n)	Agreement (true CCC)
8.9	0.001	50, 200,1000	0.594, 0.687, 0.785, 0.879, 0.999

Abbreviations: CCC: Concordance correlation coefficient; GLMM: Generalized linear mixed model.

b_0 is the intercept of the linear predictor: the average number of steps (in log scale) based on the manual count; b_1 is the average difference (in log scale) between the manual step count and the step count based on the algorithm under evaluation.

The between-measurement method variability was fixed to 0.0000007; The true CCC was calculated based on different values of the between-subject variability: 0.700, 0.031, 0.022, 0.017, 0.014

The main difference between CCC_{GLMM} and CCC_{LMM} is that the former is mean dependent and requires the variance terms to be exponentiated in order the CCC to refer to the original scale (given that the GLMM Poisson model is on log scale).

- **Estimates of interest:** In each simulated dataset, the following estimates were derived:
  - CCC, which represents the agreement between the step counts produced by the algorithm vs the manual step count
  - Pearson's correlation, *p*-value$_{t\text{-test}}$ and MAPE which are used to assess inference consistency with CCC
- **Method:** Each simulated dataset was analyzed in two ways: via LMM and GLMM
- **Objective 1:** The performance of **CCC_{GLMM}** and **CCC_{LMM}** was assessed using absolute bias and mean square error
- **Objective 2:** Consistency in inference across the metrics was assessed based on the percentage of times (denoted as P1) that each metric provided supporting evidence of agreement (CCC≥0.7[7], correlation≥0.7[8], *p*-value$_{t\text{-test}}$ >0.05, MAPE<0.05[9]) across 500 datasets

Results

Objective 1

Figure 1: Absolute bias for CCC_{GLMM} and CCC_{LMM} under different scenarios

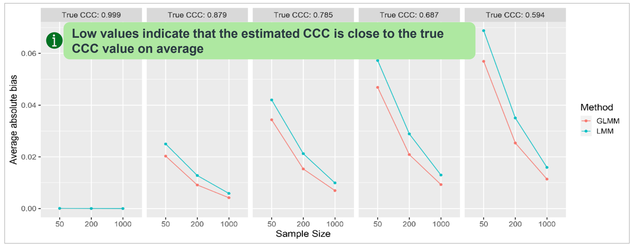
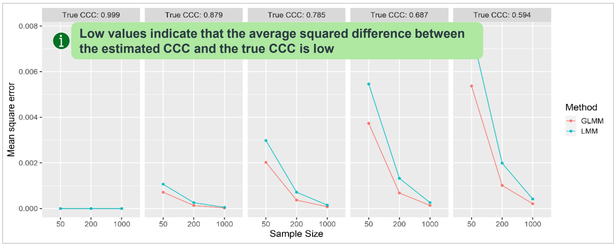


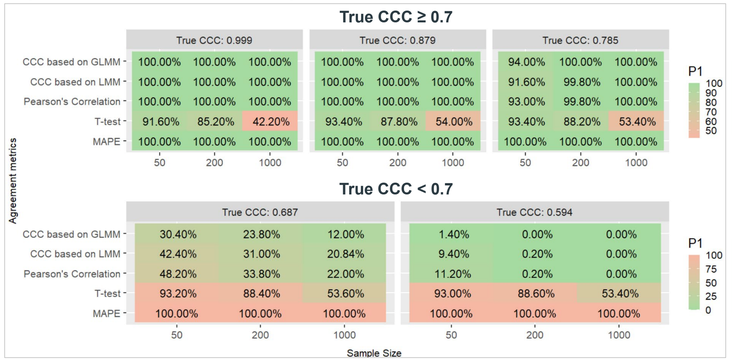
Figure 2: Mean square error for CCC_{GLMM} and CCC_{LMM} under different scenarios



- CCC_{GLMM} performed better (i.e., lower absolute bias and mean square error) than CCC_{LMM}, especially when true CCC<0.7 (see **Figures 1-2**)
- As the true CCC decreases, the impact of sample size becomes more evident for both CCC_{GLMM} and CCC_{LMM} (see **Figures 1-2**)

Objective 2

Figure 3: Heat map for P1 for each agreement metric under different scenarios



High P1 expected

Low P1 expected

- When true CCC>0.7, Pearson's correlation and MAPE provided supportive evidence of agreement with the gold standard measure in most datasets, while t-test provided variable results (see **Figure 3**)
- As expected, when true CCC<0.7, all the metrics less frequently provided supportive evidence of agreement with the gold standard measure (except for MAPE and t-test; see **Figure 3**)
- For n=50 and 200 and true CCC=0.687, CCC_{GLMM} provided notably less frequent supportive evidence of agreement with the gold standard measure than CCC_{LMM} (see **Figure 3**)

Conclusions and limitations

- **Conclusions: CCC_{GLMM}** is an appropriate metric for count-based parameters, accounting for the subject effect (unlike MAPE) and measurement method effect. It should be complemented with other agreement metrics (e.g., MAPE) and preferred over **CCC_{LMM}**, unless expected agreement is high (e.g., CCC>0.9) or data are near normal. Pearson correlation or t-test should be used only as a supplementary assessment for comparing the two measures.
- **Limitations:** The results of the current simulation study were based on Poisson GLMM, which gave an advantage to the **CCC_{GLMM}** estimator. Other data generating mechanisms could be also considered for future work.

References

[1]Welk, G. J., Bai, Y., Lee, J. M., Godino, J. O. B., Saint-Maurice, P. F., & Carr, L. (2019). Standardizing analytic methods and reporting in activity monitor validation studies. Medicine and science in sports and exercise, 51(8), 1767.
[2]Lawrence, I., & Lin, K. (1989). A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, 255-268.
[3]Carrasco, J. L. (2010) A generalized concordance correlation coefficient based on the variance components generalized linear mixed models for overdispersed count data. *Biometrics*, 66(3), 897-904;
[4]Rigot, S. K., Maronati, R., Lettenberger, A., O'Brien, M. K., Alamdari, K., Hoppe-Ludwig, S., ... & Jayaraman, A. (2024). Validation of proprietary and novel step-counting algorithms for individuals ambulating with a lower limb prosthesis. *Archives of physical medicine and rehabilitation*, 105(3), 546-557.
[5]Li, Z., Feng, W., Zhou, L., & Gong, S. (2024). Accuracy of wrist-worn activity trackers for measuring steps in patients after major abdominal surgery: A validation study. *Digital Health*, 10, 20552076241297036.
[6]Wahl, Y., Düking, P., Droszez, A., Wahl, P., & Mester, J. (2017). Criterion-validity of commercially available physical activity tracker to estimate step count, covered distance and energy expenditure during sports conditions. *Frontiers in physiology*, 8, 725.
[7]Hahn, E. A., Cella, D., Chassany, O., Fairclough, D. L., Wong, G. Y., Hays, R. D., & the Clinical Significance Consensus Meeting Group (2007). Precision of Health-Related Quality-of-Life Data Compared With Other Clinical Measures. *Mayo Clinic Proceedings*, 82(10), 1244–1254.
[8]Mukaka, M. M. (2012) A guide to appropriate use of correlation coefficient in medical research. *Malawi Medical Journal*, *24*(3), 69-71.
[9]Lewis, C. D. (1982). Industrial and business forecasting methods: A practical guide to exponential smoothing and curve fitting, London: Butterworth Scientific.