British Journal of Nutrition (2016), **115**, 1273–1280
doi:10.1017/S0007114516000040

Improved interpretation of studies comparing methods of dietary assessment: combining equivalence testing with the limits of agreement

Marijka J. Batterham[1]*, Christel Van Loo[2], Karen E. Charlton[3], Dylan P. Cliff[2] and Anthony D. Okely[2]

[1]*National Institute for Applied Statistics Research Australia, University of Wollongong, Northfields Avenue, Wollongong, NSW 2522, Australia*
[2]*Faculty of Social Sciences, University of Wollongong, Wollongong, NSW 2522, Australia*
[3]*Faculty of Science, Medicine and Health, University of Wollongong, Wollongong, NSW 2522, Australia*

(Submitted 1 September 2015 – Final revision received 20 November 2015 – Accepted 30 December 2015 – First published online 16 February 2016)

Abstract

The aim of this study was to demonstrate the use of testing for equivalence in combination with the Bland and Altman method when assessing agreement between two dietary methods. A sample data set, with eighty subjects simulated from previously published studies, was used to compare a FFQ with three 24 h recalls (24HR) for assessing dietary I intake. The mean I intake using the FFQ was 126·51 (SD 54·06) μg and using the three 24HR was 124·23 (SD 48·62) μg. The bias was −2·28 (SD 43·93) μg with a 90 % CI 10·46, 5·89 μg. The limits of agreement (LOA) were −88·38, 83·82 μg. Four equivalence regions were compared. Using the conventional 10 % equivalence range, the methods are shown to be equivalent both by using the CI (−12·4, 12·4 μg) and the two one-sided tests approach (lower $t = -2{\cdot}99$ (79 df), $P = 0{\cdot}002$; upper $t = 2{\cdot}06$ (79 df), $P = 0{\cdot}021$). However, we make a case that clinical decision making should be used to set the equivalence limits, and for nutrients where there are potential issues with deficiency or toxicity stricter criteria may be needed. If the equivalence region is lowered to ±5 μg, or ±10 μg, these methods are no longer equivalent, and if a wider limit of ±15 μg is accepted they are again equivalent. Using equivalence testing, acceptable agreement must be assessed *a priori* and justified; this makes the process of defining agreement more transparent and results easier to interpret than relying on the LOA alone.

Key words: Equivalence: Agreement: Dietary assessment: Bland and Altman method

The Bland and Altman (BA) method[1] has been routinely used for assessing relative agreement between two dietary methods. The rationale for doing this, typically, is that, although the reference method – or gold standard – is deemed to be more accurate, it also has substantial participant burden to complete and resources to analyse. Often, the FFQ method is compared against food records, either weighed or unweighed, or against repeated 24-h dietary recalls. FFQ are easier to implement, less burdensome for participants to complete and less costly to analyse[2]. It is necessary to demonstrate that FFQ results are equivalent to a reference method before it can be used with confidence. Interpretation of the results of the BA method is straightforward when it is clear that the methods do not agree. In practice, this is defined by a large and statistically significant bias using a dependent samples test (paired *t* test or Wilcoxon's matched pairs test). However, difficulty arises in determining equivalence of two dietary methods when they are shown by the BA method to be in agreement. For example, a bias of 837 kJ with a limits of agreement (LOA) of −5192 to 6865 kJ was defined as 'reasonably acceptable' agreement in a study[3] comparing a FFQ with an estimated food diary. Likewise, compared with a 24 h recall (24HR), a FFQ was reported to have a bias of 1091 kJ with a limit of agreement of −2792 to 4974 kJ[4]. This was described as 'performing well' and was considered to have 'fair' or 'adequate' agreement despite the large and statistically significant bias and wide LOA. These two examples demonstrate a lack of consideration on what constitutes a clinically acceptable difference between dietary methods. The LOA in these studies encompass a range of intake between 7766 and 12 056 kJ, which is the magnitude of intake that represents the entire recommended daily intake for an adult (i.e. 8400–11 700 kJ[5,6]). This is clearly undesirable, yet appears to be the current practice in the published nutrition literature. As Bland and Altman[7] themselves stated, 'How far apart measurements can be without leading to problems will depend on the use to which the result is put, and is a question of clinical judgement. Statistical methods cannot answer such a question'.

The aim of this study was to consider how two methods can be demonstrated as being equivalent when the BA indicates agreement. In this study, we make a case for combining formal testing of equivalence with the BA method for assessing

Abbreviations: 24HR, 24 h recall; BA, Bland and Altman; LOA, limits of agreement.

* **Corresponding author:** M. J. Batterham, email marijka@uow.edu.au

agreement between methods. Performing a test of equivalence requires an *a priori* assessment of what constitutes a clinically acceptable difference between two methods. Therefore, we first considered how agreement is described in the nutrition literature for validation of FFQ using the BA method. Second, we compared the use of equivalence testing with the BA method for assessing agreement between two methods using an original data set. The emphasis of this study was to demonstrate the need to be able to accurately define what constitutes clinical agreement – before being able to interpret the level of agreement between these methods – and to encourage the use of both methods in validation studies.

Methods

To identify a sample of FFQ validation literature describing agreement using the BA method, a search of the database Web of Science (accessed 20 March 2015) was conducted. This search returned 24 847 citations for the initial Bland and Altman paper, of which 250 were identified under the sub-search for FFQ. We then selected the ten papers with the highest number of citations, available through our institutional subscriptions, which aimed to validate an FFQ using the BA method.

To demonstrate equivalence testing and compare this with the BA method, a data set consisting of a random sample of eighty participants was simulated based on a previously published analysis using the means of I intake assessed, using the average of three repeated 24-HR and a FFQ (3 × 24HR 118·88 (SD 48·95) μg, FFQ 120·19 (SD 55·98) μg and correlation 0·614; $P < 0{\cdot}001$)[8]. The data set was simulated using the matrix and drawnorm commands in STATA (version 12, STATA Inc.). Simulated data were chosen, instead of the actual data, in this example, to allow data sharing without any ethical considerations. In addition, the initial data set was right-skewed and transformed for analysis, and the simulated data were normally distributed to assist with interpretation.

The agreement of methods was interpreted using both a BA LOA and an equivalence approach. Both methods advocate acceptance on the basis of a clinical decision; however, in the case of the equivalence approach, this must be explicitly stated *a priori*[9].

The BA method[1] involves plotting the difference between the two methods against the average of the two methods and examining the mean bias, determining the 95 % CI of the bias and any trend in the bias. The precision of the limits is rarely considered in interpreting the BA plot. Interpreting the precision of the limits involves calculating and interpreting the 95 % CI of the upper and lower limits and is detailed with an example in the initial Bland and Altman paper[1]. Further reference to this on Martin Bland's website (https://www-users.york.ac.uk/~mb55/meas/sizemeth.htm, accessed 28 August 2015) demonstrates clearly the effect of sample size on these estimates, and emphasises that it is important not only to consider the width of the LOA but also the precision with which these have been estimated.

Equivalence testing was performed using the two one-sided test (TOST) procedures[10] and also by using the CI approach[11]. Both are valid approaches, and the use of one or the other depends on whether there is a preference for the use of a *P* value or CI. Equivalence testing is widely used in the pharmaceutical industry where a new drug, which may have fewer side-effects or be less costly to produce, is compared with the standard drug to determine whether the therapeutic effect is equivalent within a pre-defined range[12]. If differences in means (d) are considered using a paired t test, as in the traditional framework, the intention is to demonstrate that a new drug or method is different (generally with the aim of showing superiority). In this case, the null hypothesis states that there is no difference between the treatments, whereas the alternate hypothesis states that there is a difference. On the basis of this paradigm, established by Neyman and Pearson[13], it can only be demonstrated that $d \neq 0$, or that there is insufficient evidence to demonstrate that $d \neq 0$. What cannot be demonstrated is that $d = 0$ – that is, the null hypothesis cannot be proven. With a small sample size, it is difficult to show that $d \neq 0$ and an erroneous conclusion that there is no difference (type 2 error) may be made, particularly if the difference is small and the variance is large[14]. In this situation, we may conclude that the two methods agree as we do not have adequate power to demonstrate that the difference is statistically significant. Alternatively, for every d, there is a sample size where it can be demonstrated that $d \neq 0$, regardless of whether this difference has any practical meaning. In this situation, we may conclude that the methods do not agree when the difference between them is actually too small to have any clinical meaning. Thus, the statistical significance is unrelated to the practical or clinical significance. When demonstrating equivalence, these hypotheses are reversed such that the null states that there is a difference ($H_0: |d| \geq \Delta$, where d is the difference between the methods and Δ is the pre-specified equivalence interval) and the alternative hypothesis is that of no difference ($H_a: |d| < \Delta$)[15]. Equivalence trials require an *a priori* specification of an acceptable equivalence range. Determination of this range needs to be guided by clinical acceptability of the range of measures. Wellek[9] discusses arbitrary ranges when the equivalence range is unknown, and other arbitrary decisions such as ±10 % of the reference mean have been used in the literature on physical activity[16]. In general, this equivalence region is poorly defined. A review of 332 non-inferiority and equivalence pharmaceutical trials found that half of these considered 0·5 SD or less of the difference between treatments to be an 'irrelevant' difference[17]. Although TOST is not the most powerful equivalence test[9,18], its relative ease of use and interpretation[15] make it the preferred approach for nutritional applications.

Both the BA and equivalence approaches are most easily interpreted visually. In our analysis, we present the traditional BA plots with the equivalence intervals incorporated. The figures contain the equivalence interval, as well as the 90 % CI of the difference and the LOA. These figures can be plotted easily in most statistical packages or in Microsoft Excel. This approach is adapted from the one proposed in the SAS macro 'Concord', which presents a BA style plot, incorporating the equivalence interval and 90 % CI instead of the LOA[19], and we also present the results as confidence interval plots and in tabular form to show different options of presentation.

Given that we wished to provide practical guidelines on the conduct of equivalence tests, we considered their use in STATA

(version 12, StataCorp LP), SAS (version 9.3 SAS Inc.), SPSS (version 21, IBM Corporation) and R (version 3.2.1, www.cran.r-project.org[20]), and instructions on the use of each of these are considered in online Supplementary Appendix S1. In this example, we considered four regions of equivalence to demonstrate the proposed methodology and the differences between equivalence and non-equivalence. The four equivalence regions chosen for this example demonstrate how to interpret clear equivalence, non-equivalence and an intuitively ambiguous result.

Results

A summary of the ten validation studies identified from the literature review is provided in Table 1. Only three of the ten papers considered *a priori* what an acceptable difference between the methods would be, whereas none of the authors discussed what was considered an acceptable LOA. All the papers reported and discussed the correlation coefficient as a method of establishing validity, although two discussed the limitations of this approach. In most cases, the results were compared only with other literature and no clinically defined or practical implications of the LOA were discussed. Seven of the ten studies performed hypothesis testing (Wilcoxon, paired *t* test) to determine whether the mean difference between the methods was statistically significant.

Table 2 presents the results of the BA comparisons and the equivalence tests for the simulated data in tabular form. Fig. 1 presents the BA plots with the equivalence intervals and 90 % CI of the difference. Fig. 2 presents the CI plot. Fig. 2(a) shows a CI plot with the x-axis showing the difference between the two means, as is the traditional approach used for pharmaceutical trials. Fig. 2(b) shows a CI plot expressed relative to the mean intake of I using the 3 × 24HR; the two plots (Fig. 2(a) and (b)) are identical in interpretation, and in this case the methods are equivalent if the 90 % CI is contained within the pre-specified equivalence region. All equivalence methods show that the FFQ is only considered to be equivalent to the 3 × 24HR when the equivalence margin is set at 10 % of the mean of the 3 × 24HR (12·24 μg), or alternatively at 15 μg. These methods are not equivalent when the margin is set at 5 μg. The methods are also not equivalent when the margin is set at ±10 μg, because, although the mean difference meets the criteria on one side (the upper 90 % CI being 5·89, which is within the upper bound of 10 μg), the lower bound is outside the range (−10·49 < 10 μg) and both sides must be within the region to meet the assumption of equivalence. This is also reflected in the *P* values, both of which must be significant for equivalence to hold. Commands and outputs for the tests in SAS, R, STATA and SPSS are shown in online Supplementary Appendix S1. Fig. 3 shows the BA LOA plot with the 90 % CI of the mean bias used for the equivalence testing and the 95 % CI of the upper and lower LOA (numerically represented in Table 2).

Discussion

This study demonstrates the use of assessing equivalence in dietary studies that compare two methods for agreement. Equivalence is presented to be used in conjunction with the more commonly applied BA LOA method. The advantage of the equivalence method is that it requires the clinician to make an *a priori* assessment of what represents agreement, rather than accepting or rejecting the LOA determined in the BA analysis *a posteri*. The equivalence approach can be assessed using CI, either independently or in combination with a BA plot, or equivalence can be assessed in the traditional paradigm of *P* values using TOST.

Frequently, the agreement between two dietary methods is assessed using the BA analysis, and the decision whether or not to determine agreement is based on a dependent samples test (paired *t* test or Wilcoxon's matched pairs test). This approach was not advocated by BA and their initial paper that describes the method makes no reference to hypothesis testing regarding the bias. Rather, the initial manuscript by Bland & Altman[1] states 'How far apart the measurements can be without causing difficulties will be a question of judgement. Ideally, it should be defined in advance to help in the interpretation of the method comparison and to choose the sample size'.

Discussion to date on what constitutes a clinical LOA in the nutrition literature is limited. For example, among the studies reviewed in this analysis, many compared their LOA with other studies[21], but without discussion on whether this was acceptable in practice. In addition, when assessing agreement, the 95 % CI of the limits (as shown in Fig. 3) are rarely considered. These can be wide particularly for small data sets and should be reported, discussed and considered, particularly when estimating sample sizes, as advocated in the early BA literature. When only considering the LOA themselves, we may be prepared to accept that the measures agree; however, the interpretation of the 95 % CI of the LOA suggests that we could have an upper LOA as high as 117·96 μg or a lower LOA as low as −122·24 μg with repeated sampling.

Judging what is an acceptable equivalence between two methods is not a trivial procedure[17,30]. Even in the pharmaceutical domain, where equivalence tests are most often used, a systematic review found that only 134 of 314 studies provided a rationale for the difference used[17]. Given the number of agreement studies published in the field of nutrition, it is necessary to be able to determine the clinical rather than just the statistical interpretation of the results.

The question of what constitutes equivalence in the field of nutrition is complex. This may differ, depending on the nutrient being assessed and the population that is being studied. In the case of I, the estimated average requirement reported in the Australian Nutrient Reference Values is 100 μg/d for adults, with a recommended daily intake of 150 μg/d and an upper limit of 1100 μg/d[31]. Estimated average intakes in the Australian population based on the most recent (2011–2012) Australian nationally representative Health Survey were 191 μg in males and 152 μg in females[32]. Therefore, for the general population, a 10 % equivalence based on the mean of the reference food record appears reasonable. In populations where intakes may be inadequate (e.g. pregnant women)[33,34] and where the consequences of inadequacy have serious impacts on health outcomes, more stringent equivalence limits may be warranted.

Consideration of why it is important to state the acceptable LOA or equivalence a priori is warranted. Although there was a

Table 1. Summary of a highly cited sample of the literature assessing agreement of a FFQ with a reference method using the Bland and Altman (BA) method

Paper Comparator method	Sample size	Correlation performed to assess validity	Significance test of differences	BA plots presented	*A priori* assessment of acceptable agreement	Conclusion	Justification of conclusion	Number of citations*
Villegas *et al.*[21] FFQ, 24HR	195	Yes	Wilcoxon, all $P<0{\cdot}05$	Energy, protein, fat, carbohydrate, 8 nutrients not shown	100 % of ratio of FFQ/24HR, no LOA	'Good agreement'	Compared with other studies	64
Matthys *et al.*[22] FFQ/EFD	104	Yes	Wilcoxon	No BA plots, mean difference and standard deviation presented	Nil	'Acceptable for assessing population median intakes' for some food groups. LOA 'broad for all food groups'	Acceptability based on $P \geq 0{\cdot}05$ in the Wilcoxon tests	43
Cullen *et al.*[23] FFQ/24HR	83	Yes	Paired *t* tests	No BA plots, mean difference and standard deviation presented	Nil	BA plots done however not shown as they 'indicated no association between the difference and the mean of the two measures'. Validity for some nutrients, but not most food groups in adolescents based on *t* test.	Based on paired *t* tests	65
Hjartaker *et al.*[4] FFQ/24HR	238	Yes	Wilcoxon	Energy, fibre, alcohol, 20 nutrients not shown, however the three figures were shown to represent three trends observed in the nutrients assessed	Nil	LOA wide. Most means lower in FFQ. Three general trends in the bias over the range measured.	Compared with other studies. 'overall relative validity … comparable to that of FFQ used in other large cohorts often described as 'fair' or 'adquate'	48
Brantsaeter *et al.*[24] FFQ/WFD	119	Yes	Wilcoxon	Energy, fruit/vegetable/juice (g/d), twenty-four nutrients/vitamins/elements not shown as 'the plots were similar to the plot of energy intake'	Nil	'Bias was small, whereas the confidence limits were wide' '…produces realistic and relatively precise estimate of habitual intake of energy'	Based on small bias	85
Toft *et al.*[25] FFQ/DH	264	Yes, although limitations discussed	Wilcoxon by sex, 38/42 $P<0{\cdot}05$	SFA, energy, 19 nutrients not shown	Nil	'Acceptable agreement' with 'small differences'. SFA tendency for increased underreporting with increased intake	No definition of acceptable agreement given	35
Watson *et al.*[26] FFQ/AFR	113	Yes	No	β-Carotene, Ca, twenty nutrients not shown	Good (difference = 1 SD mean reference) Fairly good (difference = 2 SD mean reference) Poor (difference = 3SD mean reference)	'Not suitable for estimating absolute agreement'	Positive differences Wide limits of agreement, increasing difference with increasing bias	59
Zhang & Ho[27] FFQ/EFD	61	Yes	Wilcoxon, however results not reported	Energy, vitamin C, 12 nutrients not shown	100 % of ratio of FFQ/24 FR, no LOA	Differences stated no discussion on whether they agree with 100 %, LOA similar or narrower to other studies. State BA plots show no linear trend and there is 'reasonable comparative validity'	Compared with other studies	42
Ambrosini *et al.*[28] FFQ/EFD	785	Yes, although discussed limitations	No	Energy, carbohydrate, protein, fat, 18 nutrients not presented	Nil	'The majority of nutrients showed average agreement that was significantly different from 100 %' however the criteria for significance was not stated. 95 % CI are presented	'Most LOA ranged from 50 to 250 %' similar to other studies	39
Fernández-Ballart *et al.*[3] FFQ/EFD	158	Yes	No	Vegetable, meat/meat product, potato, legumes, energy, protein, thiamine, cobalamin presented as examples showing no trend or systematic trend over range of values, 26 nutrients/foods not shown	Nil	'Agreement between these two methods was also reasonably acceptable'. 'It was found to have acceptable levels of reproducibility and validity'	Compared to other studies, expected general overestimation by FFQ[29]	126

24HR, 24 h recall; LOA, limits of agreement; EFD, estimated food diary; WFD, weighed food diary; DH, diet history; AFR, assisted food record.
* Number of citations on Web of Science on 28 August 2015.

Table 2. Summary statistics, paired *t* test, Bland and Altman (BA) limits of agreement (LOA) and equivalence tests for assessing agreement between the 3×24HR and the FFQ

Method (*n* 80)	I (mean) (μg)	SD (μg)	Minimum (μg)		Maximum (μg)	
3×24HR	124·23	48·62	29·61		240·00	
FFQ	126·51	54·06	13·03		244·82	
	Paired *t* test					
Mean difference 3×24HR – FFQ (μg)	SD (μg)	SEM (μg)	95 % CI of the difference (μg)		*t* (df = 79)	*P*
−2·28	43·93	4·91	−12·06	7·49	−0·465	0·643
	BA LOA					
BA bias (μg)	SD (μg)	SEM (μg)	LOA (μg)		95 % CI of lower limit (μg)	95 % CI of upper limit (μg)
−2·28	43·93	4·91	−88·38	83·82	−122·24, −54·52	49·96, 117·96
	Paired equivalence test					
Mean difference 3×24HR – FFQ (μg)	SD (μg)	SEM (μg)	90 % CI of the difference (μg)		*t* (df = 79)	*P*
−2·28	43·93	4·91	−10·46	5·89		
Equivalence region −2·28 (SEM 5) μg I				*t* upper	0·55	0·291
−5 > −10·46, 5·89 > 5 (decision: not equivalent)				*t* lower	−1·48	0·071
Equivalence region −2·28 (SEM 10) μg I				*t* upper	1·57	0·060
−10 > −10·46, 5·89 > 10 (decision: not equivalent)				*t* lower	−2·50	0·007
Equivalence region −2·28 (SEM 15) μg I				*t* upper	2·59	0·006
−15 > −10·46, 5·89 > 15 (decision: equivalent)				*t* lower	−3·52	0·000
Equivalence region −2·28 (SEM 10) % (±12·4) μg I				*t* upper	2·06	0·021
−12·4 > −10·46, 5·89 > 12·4 (decision: equivalent)				*t* lower	−2·99	0·002

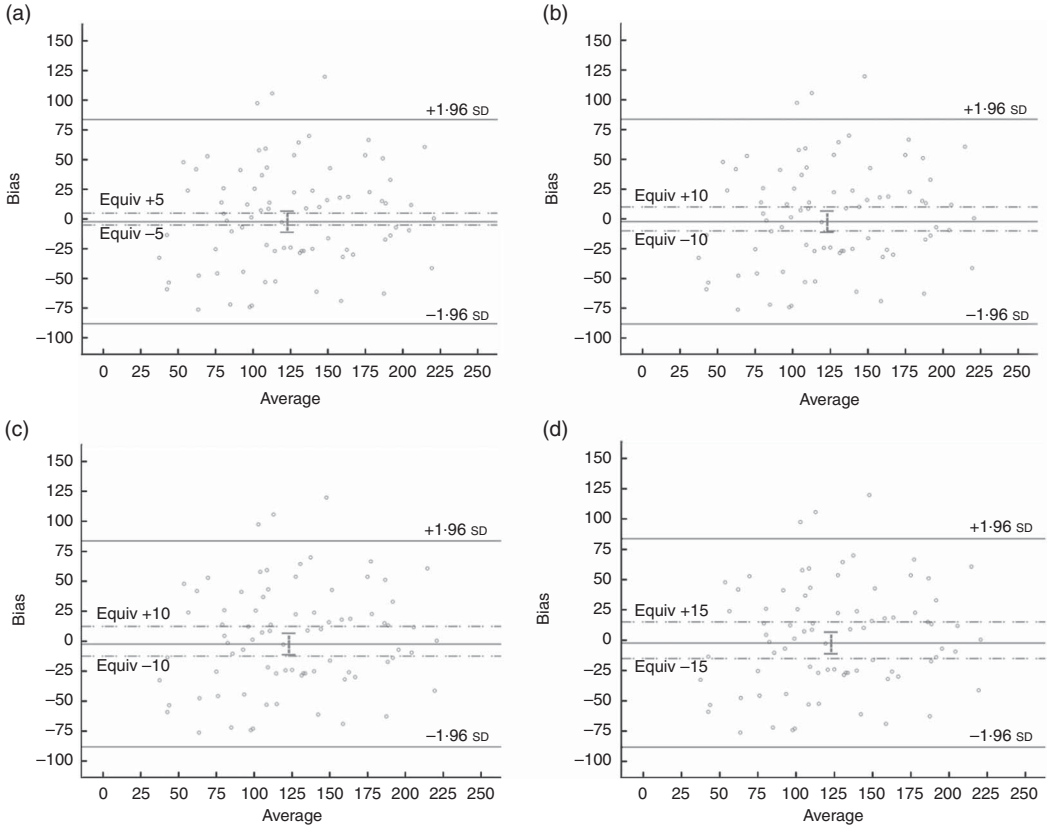

Fig. 1. Bland and Altman plots with superimposed equivalence intervals and the 90 % CI of the mean difference. (a) Equivalence ±5 μg I, (b) equivalence ±10 μg I, (c) equivalence ±10 % mean I 3×24HR and (d) equivalence ±15 μg I. 3×24HR, average of three 24 h recalls.

British Journal of Nutrition

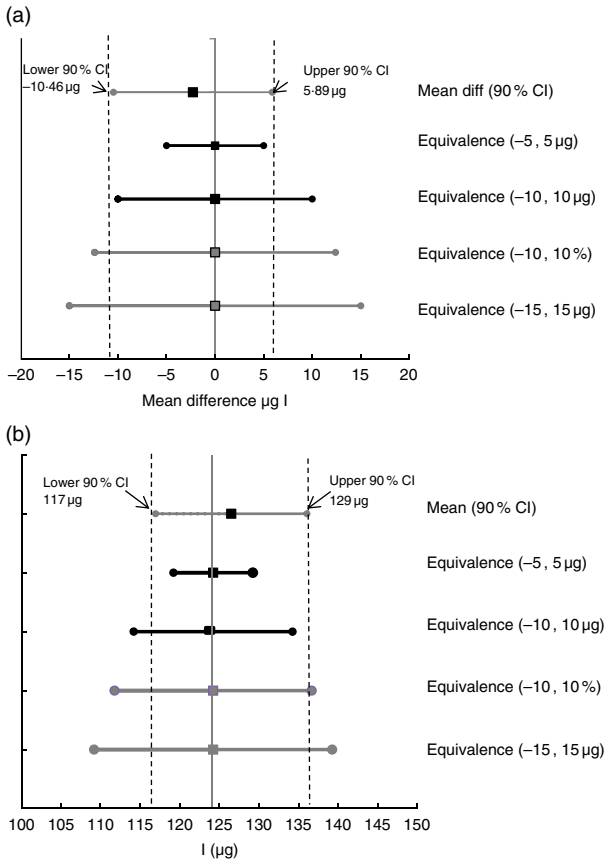

Fig. 2. CI plots. (a) CI plot using the mean difference between the 3×24 h recall (24HR) and FFQ and (b) CI plot using the mean I intake in the 3×24HR (124·23 µg). 3×24HR, average of three 24 h recalls.

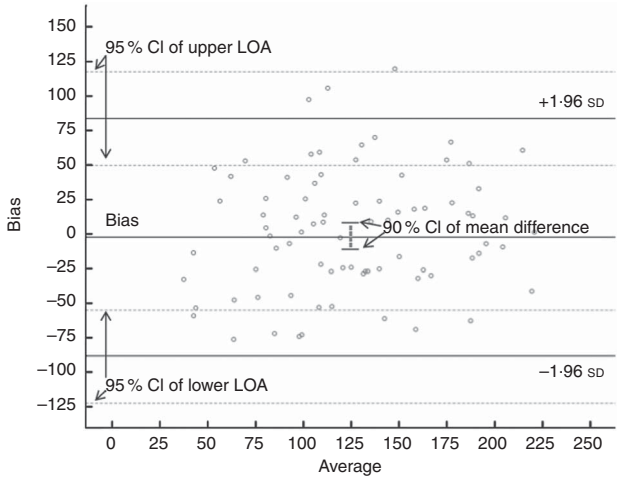

Fig. 3. 95 % CI of the upper and lower limits of agreement (LOA) for the mean bias in I intake (µg) between the 3×24HR and the FFQ.

large range of sample sizes in the studies presented here (n 61–785), these were selected as being the most cited, and dietary validation studies can be conducted with relatively small sample sizes (e.g. n 49[35]). This may lead to an erroneous acceptance of the null hypothesis due to limited power to detect a difference in the traditional hypothesis test (i.e. a type 2 error). In our particular example, the power to detect a mean difference of 2·28 (SD 43·93) µg with eighty subjects is only 0·084. In order for this difference (2·28 µg with an SD of 43·93 µg) to be statistically significant at an α of 0·05 with 80 % power, a sample size of 2112 would be required. As the sample size increases, the probability of rejecting the null hypothesis in the traditional null hypothesis testing framework increases, whereas smaller sample sizes result in the opposite trend[15].

In this example, we considered only the paired t test for equivalence as it is generally the case that two dietary assessment methods would be compared on the same subjects. Independent t test methods also exist for both normally distributed and non-normal data. It is often the case that dietary intake data are skewed as was the case with the initial data set on which the simulation used in the present analysis was based[8]. Non-normal data can be analysed using a similar approach for either paired or independent data based on the robust t test of Yuen[36] in the 'equivalence' package in R. Log transformation can also be considered. In this case, the interpretation relies on back-transformation, and the results represent the ratio of the two methods, generally expressed as a percentage with absolute equivalence being 100 %. SAS has a 'dist = lognormal' option in PROC TTEST where the TOST procedure is conducted, which will convert output and produce data based on the geometric (or back-transformed) mean. When back-transforming logarithmic data, a difference of ±10 % is approximately symmetrical, but wider limits will not be. For example, if the equivalence region chosen is ±20 %, this will correspond to a range of 80–125 % when the ratio is back-transformed. This relationship must be considered when setting equivalence limits with log-transformed data. Log transformations are commonly used in pharmaceutical equivalence testing and these concepts have been covered in the related literature[12]. The equivalence approach can also be applied to other hypothesis testing such as equivalence of slopes or trend[37]. In addition, multisample and multivariate tests have also been described but are beyond the scope of what is covered in this study.

This study was designed to assess methodological comparison studies based on agreement using an example based on our previous research. There are other methods for judging the usefulness of new dietary assessment tools, such as the method of triads, which we have used previously[8], or missclassification, but they are not discussed here. Lombard *et al.*[38] provided a recent review and recommendations on the use of other methods, specifically applicable to nutrient assessment. The comparison of 3×24HR with a FFQ outlined here is an example of an approach that can be applied not only to dietary methodology but also to other methods used in nutrition practice and research, which are commonly assessed for agreement using BA methodology. These include comparing resting energy expenditure prediction equations to indirect calorimetry[39–42], bioelectrical impedance analysers to dual energy x-ray absorptiometry for assessing body composition[43–46] and in validating physical activity assessment tools[47–49].

In summary, we have introduced an equivalence approach to be used in conjunction with the BA method in order to

encourage clinicians to establish up front what constitutes a clinically meaningful difference between the two methods being considered. This not only makes interpretation of the results of the study clear but also assists with assessing the necessary sample size in planning the study.

Acknowledgements

This research received no specific grant from any funding agency, commercial or not-for-profit sectors.

The research idea was conceived during discussions between M. J. B., C. V. L., D. P. C. and A. D. O. M. J. B. formalised the study idea, simulated and analysed the data and drafted the paper. C. V. L. assisted in study conception and comparing the analytical approaches and reviewing the manuscript. D. P. C. assisted in the study conception and reviewed the manuscript. K. E. C. assisted with the provision of the context and practical applications of the I example. A. D. O. provided constructive feedback on the study design and implications and reviewed the manuscript.

The authors declare that there are no conflicts of interest.

Supplementary material

For supplementary material/s referred to in this article, please visit http://dx.doi.org/doi:10.1017/S0007114516000040

References

1. Bland JM & Altman DG (1986) Statistical methods for assessing agreement between 2 methods of clinical measurement. *Lancet* **i**, 307–310.
2. Willett W (1998) *Nutritional Epidemiology*, 2nd ed. *Monographs in Epidemiology and Biostatistics*. New York, NY: Oxford University Press.
3. Fernández-Ballart JD, Piñol JL, Zazpe I, *et al.* (2010) Relative validity of a semi-quantitative food-frequency questionnaire in an elderly Mediterranean population of Spain. *Br J Nutr* **103**, 1808–1816.
4. Hjartaker A, Andersen LF & Lund E (2007) Comparison of diet measures from a food-frequency questionnaire with measures from repeated 24-hour dietary recalls. The Norwegian Women and Cancer Study. *Public Health Nutr* **10**, 1094–1103.
5. NSW Government (2015) 8700 find your ideal figure. www.8700.com.au (accessed November 2015).
6. European Food Information Council (2015) What is the recommended calorie intake for adults, children and toddlers? http://www.eufic.org/page/en/page/faq/faqid/recommended-calorie-intake-adults-children-toddlers/ (accessed November 2015).
7. Bland JM & Altman DG (1999) Measuring agreement in method comparison studies. *Stat Methods Med Res* **8**, 135–160.
8. Tan L-M, Charlton KE, Tan S-Y, *et al.* (2013) Validity and reproducibility of an iodine-specific food frequency questionnaire to estimate dietary iodine intake in older Australians. *Nutr Diet* **70**, 71–78.
9. Wellek S (2010) *Testing Statistical Hypothesis of Equivalence and Noninferiority*, 2nd ed. Boca Raton, FL: Chapman & Hall/CRC Press.
10. Schuirmann DJ (1987) A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *J Pharmacokinet Biopharm* **15**, 657–680.
11. Westlake WJ (1972) Use of confidence intervals in analysis of comparative bioavailability trials. *J Pharm Sci* **61**, 1340–1341.
12. Midha KK & McKay G (2009) Bioequivalence; its history, practice, and future. *AAPSJ* **11**, 664–670.
13. Neyman J & Pearson ES (1928) On the use and interpretation of certain test criteria for purposes of statistical inference: part 1. *Biometrika* **20**, 175–240.
14. Altman DG & Bland JM (1995) Statistics notes: absence of evidence is not evidence of absence. *BMJ* **311**, 485.
15. Hoenig JM & Heisey DM (2001) The abuse of power: the pervasive fallacy of power calculations for data analysis. *Am Stat* **55**, 19–24.
16. Kim Y, Crouter SE, Lee JM, *et al.* (2016) Comparisons of prediction equations for estimating energy expenditure in youth. *J Sci Med Sport* **19**, 35–40.
17. Lange S & Freitag G (2005) Choice of delta: requirements and reality – results of a systematic review. *Biom J* **47**, 12–27; discussion 99–107.
18. Berger RL & Hsu JC (1996) Bioequivalence trials, intersection-union tests and equivalence confidence sets. *Statist Sci* **11**, 283–319.
19. Groeneveld J (2011) Embedding equivalence t-test results in Bland Altman Plots visualising rater reliability. *Pharmaceutical Users Software Exchange*, PhUSE 2011, Brighton, UK, 9–12 October 2011. pp. SP06.
20. R Core Team (2013) *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing. http://wwwR-projectorg/ (accessed November 2015).
21. Villegas R, Yang G, Liu D, *et al.* (2007) Validity and reproducibility of the food-frequency questionnaire used in the Shanghai Men's Health study. *Br J Nutr* **97**, 993–1000.
22. Matthys C, Pynaert I, De Keyzer W, *et al.* (2007) Validity and reproducibility of an adolescent web-based food frequency questionnaire. *J Am Diet Assoc* **107**, 605–610.
23. Cullen KW, Watson K & Zakeri I (2008) Relative reliability and validity of the Block Kids Questionnaire among youth aged 10 to 17 years. *J Am Diet Assoc* **108**, 862–866.
24. Brantsaeter AL, Haugen M, Alexander J, *et al.* (2008) Validity of a new food frequency questionnaire for pregnant women in the Norwegian Mother and Child Cohort Study (MoBa). *Matern Child Nutr* **4**, 28–43.
25. Toft U, Kristoffersen L, Ladelund S, *et al.* (2008) Relative validity of a food frequency questionnaire used in the Inter99 study. *Eur J Clin Nutr* **62**, 1038–1046.
26. Watson JF, Collins CE, Sibbritt DW, *et al.* (2009) Reproducibility and comparative validity of a food frequency questionnaire for Australian children and adolescents. *Int J Behav Nutr Phys Act* **6**, 62.
27. Zhang C-X & Ho SC (2009) Validity and reproducibility of a food frequency questionnaire among Chinese women in Guangdong province. *Asia Pac J Clin Nutr* **18**, 240–250.
28. Ambrosini GL, de Klerk NH, O'Sullivan TA, *et al.* (2009) The reliability of a food frequency questionnaire for use among adolescents. *Eur J Clin Nutr* **63**, 1251–1259.
29. Subar AF, Thompson FE, Kipnis V, *et al.* (2001) Comparative validation of the Block, Willett, and National Cancer Institute food frequency questionnaires: the Eating at America's Table Study. *Am J Epidemiol* **154**, 1089–1099.
30. Senn SS (2008) *Statistical Issues in Drug Development*, 2nd ed. *Statistics in Practice*. Chichester: John Wiley & Sons Ltd.
31. National Health and Medical Research Council Australia & Ministry of Health New Zealand (2014) Nutrient reference

values for Australia and New Zealand. https://www.nrv.gov.au/contact (accessed June 2015).
32. Australian Bureau of Statistics (2014) Australian Health Survey: nutrition first results-food and nutrients, 2011–12. 4364.0.55.007. www.abs.gov.au (accessed November 2015).
33. Charlton KE, Yeatman H, Brock E, *et al.* (2013) Improved iodine status in pregnant women 3 years following mandatory iodine fortification of bread in Australia. *Prev Med* **57**, 26–30.
34. Hynes KL, Otahal P, Hay I, *et al.* (2013) Mild iodine deficiency during pregnancy is associated with reduced educational outcomes in the offspring: 9-year follow-up of the gestational iodine cohort. *J Clin Endocrinol Metab* **98**, 1954–1962.
35. Weir RR, Carson EL, Mulhern MS, *et al.* (2015) Validation of a food frequency questionnaire to determine vitamin D intakes using the method of triads. *J Hum Nutr Diet* (epublication ahead of print version 7 August 2015).
36. Yuen KK (1974) The two-sample trimmed t for unequal population variances. *Biometrika* **61**, 165–170.
37. Dixon PM & Pechmann JHK (2005) A statistical test to show negligible trend. *Ecology* **86**, 1751–1756.
38. Lombard MJ, Steyn NP, Charlton KE, *et al.* (2015) Application and interpretation of multiple statistical tests to evaluate validity of dietary intake assessment methods. *Nutr J* **14**, 40.
39. Reidlinger DP, Willis JM & Whelan K (2015) Resting metabolic rate and anthropometry in older people: a comparison of measured and calculated values. *J Hum Nutr Diet* **28**, 72–84.
40. Siervo M, Bertoli S, Battezzati A, *et al.* (2014) Accuracy of predictive equations for the measurement of resting energy expenditure in older subjects. *Clin Nutr* **33**, 613–619.
41. Vilar E, Machado A, Garrett A, *et al.* (2014) Disease-specific predictive formulas for energy expenditure in the dialysis population. *J Renal Nutr* **24**, 243–251.
42. Lazzer S, Patrizi A, De Coi A, *et al.* (2014) Prediction of basal metabolic rate in obese children and adolescents considering pubertal stages and anthropometric characteristics or body composition. *Eur J Clin Nutr* **68**, 695–699.
43. Pineau JC & Frey A (2014) Comparison of body composition measurement in highly trained athletes obtained by bioelectrical impedance analysis and dual-energy X-ray absorptiometry. *Sci Sports* **29**, 164–167.
44. Sillanpaa E, Cheng SL, Hakkinen K, *et al.* (2014) Body composition in 18- to 88-year-old adults – comparison of multifrequency bioimpedance and dual-energy X-ray absorptiometry. *Obesity (Silver Spring)* **22**, 101–109.
45. Wan CS, Ward LC, Halim J, *et al.* (2014) Bioelectrical impedance analysis to estimate body composition, and change in adiposity, in overweight and obese adolescents: comparison with dual-energy X-ray absorptiometry. *BMC Pediatr* **14**, 249.
46. Ziai S, Coriati A, Chabot K, *et al.* (2014) Agreement of bioelectric impedance analysis and dual-energy X-ray absorptiometry for body composition evaluation in adults with cystic fibrosis. *J Cyst Fibros* **13**, 585–588.
47. Oyeyemi AL, Umar M, Oguche F, *et al.* (2014) Accelerometer-determined physical activity and its comparison with the International Physical Activity Questionnaire in a sample of Nigerian adults. *PLOS ONE* **9**, e87233.
48. Aadland E & Ylvisaker E (2015) Reliability of the actigraph GT3X+ accelerometer in adults under free-living conditions. *PLOS ONE* **10**, e0134606.
49. Vanderloo LM, D'Alimonte NA, Proudfoot NA, *et al.* (2015) Comparing the actical and actigraph approach to measuring young children's physical activity levels and sedentary time. *Pediatr Exerc Sci* (epublication ahead of print version 17 July 2015).

British Journal of Nutrition