On Jones et al.'s method for assessing limits of agreement with the mean for multiple observers

Heidi S. Christensen[1,2,3], Jens Borgbjerg[4], Lars Børty[2] and Martin Bøgsted[1,2,3*]

[1]Department of Clinical Medicine, Aalborg University, Aalborg, Denmark

[2]Department of Haematology, Aalborg University Hospital, Aalborg, Denmark

[3]Clinical Cancer Research Center, Aalborg University Hospital, Aalborg, Denmark

[4]Department of Radiology, Aarhus University Hospital, Aarhus, Denmark

[*]Corresponding author, email: martin.boegsted@rn.dk

Abstract

Background To assess the agreement of continuous measurements between a number of observers, Jones et al. introduced limits of agreement with the mean (LOAM) for multiple observers, representing how much an individual observer can deviate from the mean measurement of all observers. Besides the graphical visualisation of LOAM, suggested by Jones et al., it is desirable to supply LOAM with confidence intervals and to extend the method to the case of multiple measurements per observer.

Methods We reformulate LOAM under the assumption the measurements follow an additive two-way random effects model. Assuming this model, we provide estimates and confidence intervals for the proposed LOAM. Further, this approach is easily extended to the case of multiple measurements per observer.

Results The proposed method is applied on two data sets to illustrate its use. Specifically, we consider agreement between measurements regarding tumour size and aortic diameter. For the latter study, three measurement methods are considered.

Conclusions The proposed LOAM and the associated confidence intervals are useful for assessing agreement between continuous measurements.

Keywords Accuracy, limits of agreement with the mean, continuous measurements

1 Background

Clinical decisions regarding diagnosis or treatment are often based on one or more measured quantities such as blood pressure, tumour size, or the diameter of an aorta. To understand the limitations of using such measurements in clinical practice, it is important to quantify how much the measurements may vary.

For almost three decades, Bland-Altman plots have been the standard method for graphical assessment of agreement between continuous measurements made by two observers or methods on a number of subjects [1]. In particular, Bland-Altman plots are often used to assess how well a new measurement method compares to a current golden standard method. However, if the goal is to assess the variability of measurements made by different observers it is preferable to consider more than two observers.

This prompted Jones et al. to suggest an extension of Bland-Altman's graphical method for assessing *limits of agreement between two observers* to the *limits of agreement with the mean (LOAM) for multiple observers* [2]. Jones et al.'s LOAM have the advantage that they quantify agreement between measurements on the same scale as the measurements themselves, in contrast to the intra-class correlation (ICC) that has no unit of measure and always takes value between 0 and 1.

In more detail, consider a study where a continuous quantity is observed on a subjects by b observers (or methods). We let y_{ij} denote an observation from a random variable Y_{ij}, which models the measurement performed on the *i*'th subject by the *j*'th observer for $i = 1, \dots, a$ and $j = 1, \dots, b$. Assuming no preferred observer, Jones et al. suggested to assess the agreement between measurements made by different observers by investigating how much the measurements vary around the subject-specific average [2]. More formally, they were interested in how much the differences $D_{ij} = Y_{ij} - \bar{Y}_{i\cdot}$ are likely to vary, where $\bar{Y}_{i\cdot}$ denotes the average measurement for subject i across the b observers. For visualising the data, Jones et al. propose to consider a plot of the observed differences $d_{ij} = y_{ij} - \bar{y}_{i\cdot}$ against the observed subject-specific average $\bar{y}_{i\cdot}$. We will refer to this as an *agreement plot*. For an example of an agreement plot see Figure 1 below. In the special case of two observers, i.e., b = 2, the agreement plot corresponds to the scatter plot of $\left(\bar{y}_{i\cdot}, 0.5(y_{i1} - y_{i2})\right)$ for $i = 1, \dots, a$, which again corresponds to a scaled Bland-Altman plot. An agreement plot can for example help to detect whether the spread of the differences is associated

to the size of the measurements, or, at least when a and b are not too large, whether some observers tend to always make large, small, or more varying measurements.

Further, Jones et al. equipped the agreement plot with horizontal lines representing the estimated 95% LOAM, which are given by $\pm 1.96s$, where s is the estimate of the residual standard deviation in a two-way analysis of variance (ANOVA) including subject and observer effect. Thus, any possible variation due to observer is not included in s and therefore disregarded in Jones et al.'s LOAM. Jones et al. suggest investigating whether there is a significant systematic observer variation using the two-way ANOVA. However, in the (unrealistic) case of no systematic observer variation, the suggested 95% LOAM lines are biased and inefficiently estimated, as it would be custom to refit the ANOVA model without the adjustment for observer variation and adjust the degrees of freedom for s, accordingly. Further, no alternative is provided for incorporating a non-negligible observer variation into the LOAM.

In conclusion, although the method has gained an increasing interest over the years, Jones et al. did not provide a way to: 1) assess the variation of the LOAM estimate, 2) integrate systematic differences between the observers, and 3) extend the method to multiple observations per observer.

In this paper, we suggest formalising Jones et al.'s approach under a simple two-way random effects model, which allows us to formulate a coherent statistical inference procedure for the LOAM. In addition, we provide not only an implementation in the statistical programming software R, but also simple formulae which can be implemented in, e.g., statistical programming languages, Excel, or automatic web-modules for data collection.

2 Methods

2.1 A revised version of the limits of agreement with the mean

As an alternative, we propose to derive LOAM assuming a statistical model for the measurements. Assuming a model provides a theoretical framework in which the LOAM can be constructed in a transparent way and furthermore enables us to supply estimates and confidence intervals for the LOAM.

2.1.1 Statistical model

In the following we assume the measurements follow a two-way random effects model given by

$$Y_{ij} = \mu + A_i + B_j + E_{ij}, \quad (1)$$

where μ describes the overall mean, and A_i, B_j, and E_{ij} are independent random variables following zero-mean normal distributions with variances σ_A^2 , σ_B^2 , and σ_E^2, respectively. Under this model, measurements made by different observers on different subjects are uncorrelated, while measurements made by different observers, but on the same subject, have covariance σ_A^2, and measurements made by the same observer for different subjects are assumed to have covariance σ_B^2. Thus, the model accounts for correlation among measurements made by the same observer or on the same subject. Note that the measurements are assumed to be homoscedastic with variance $\sigma_A^2 + \sigma_B^2 + \sigma_E^2$. That is, the variance is split into three components: the inter-subject, inter-observer, and residual variance. We follow here the tradition of some authors and call interchangeably the residual variance for intra-observer variance.

2.1.2 Proposed limits of agreement with the mean

Under the two-way random effects model stated in Eq. (1), the difference to the mean, D_{ij}, is normally distributed with mean zero and variance $(\sigma_B^2 + \sigma_E^2)(b-1)/b$. Thus, under this model we expect 95% of the differences to be within the limits

$$\pm 1.96 \sqrt{\frac{b-1}{b}(\sigma_B^2 + \sigma_E^2)}. \quad (2)$$

We therefore propose the above as the 95 % LOAM.

To estimate σ_B^2 and σ_E^2 under the suggested two-way random effects model, we use the unbiased and consistent ANOVA estimates (see, e.g., Chapter 4 of Searle et al. [3]), given by

$$\hat{\sigma}_B^2 = \frac{MSB - MSE}{a}, \qquad \hat{\sigma}_E^2 = MSE, \quad (3)$$

where $MSB = SSB/\nu_B$ and $MSE = SSE/\nu_E$ with $SSB = a \sum_{j=1}^{b} (\bar{y}_{\cdot j} - \bar{y}_{\cdot\cdot})^2$ and $SSE = \sum_{i=1}^{a} \sum_{j=1}^{b} (y_{ij} - \bar{y}_{i\cdot} - \bar{y}_{\cdot j} + \bar{y}_{\cdot\cdot})^2$ denoting the sums of squares for the observer and residual term, and $\nu_B = b - 1$ and $\nu_E = (a-1)(b-1)$. Using these estimates of σ_B^2 and σ_E^2 , we obtain the following estimate of the 95% LOAM:

$$\pm 1.96 \sqrt{\frac{SSB + SSE}{N}}, \tag{4}$$

where $N = ab$ is the total number of measurements. For comparison Jones et al.'s estimate of the LOAM is given by $\pm 1.96\ \hat{\sigma}_E$, where variation due to observers is not included.

2.1.3 Confidence intervals

Instead of simply reporting the estimated LOAM given by Eq. (4), it is more informative to report confidence intervals. However, as the distribution of the LOAM is quite complicated, we only supply approximate confidence intervals.

A symmetric confidence interval can be constructed using the asymptotic normality results outlined by Graybill and Wang [4] and the statistical delta method. This gives the approximate 95% confidence interval with endpoints

$$1.96 \sqrt{\frac{SSB + SSE}{N}} \pm 1.96^2 \sqrt{\frac{SSB^2/\nu_B + SSE^2/\nu_E}{2N(SSB + SSE)}} \tag{5}$$

for the upper 95% LOAM. Flipping the sign of the endpoints provides the corresponding confidence interval for the lower 95% LOAM. Simulations under the two-way random effects model from Eq. (1) indicate that the coverage probability for the symmetric confidence interval can be quite far away from 95% even with a reasonable high number of measurements. In particular, the confidence interval tends to be too narrow to obtain the desired coverage probability when the number of observers is small or moderate. In that case, we recommend using the following asymmetric confidence interval instead.

First, an approximate 95% confidence interval can be obtained for $\sigma_B^2 + \sigma_E^2$ using Eq. (2.2) in Graybill and Wang [4]. Next, transforming this in accordance with Eq. (2) we get an approximate confidence interval for the upper 95% LOAM given by

$$(1.96\sqrt{(SSB + SSE - L)/N}\,, 1.96\sqrt{(SSB + SSE + H)/N}\,), \tag{6}$$

where

$$L = \sqrt{l_B^2 SSB^2 + l_E^2 SSE^2}, \qquad H = \sqrt{h_B^2 SSB^2 + h_E^2 SSE^2}$$

with $l_x = 1 - 1/F_{0.975;\,\nu_x,\infty}$ and $h_x = 1/F_{0.025;\,\nu_x,\infty} - 1$ for $x = B$ and $x = E$ (see Graybill and Wang for other choices of l_x and h_x [4]). Here $F_{\alpha;m,n}$ is the α-quantile for the F-distribution with m numerator and n

denominator degrees of freedom. A confidence interval for the lower 95% LOAM is simply obtained by flipping the sign of Eq. (6).

Results from a small simulation study, [see Additional file 1], indicate that the "sufficient" number of observers depends on the inter-observer and residual variation. However, it seems that 30-40 observers in general is enough to obtain an actual coverage probability of around 90%-95%.

2.1.4 Sample size calculations

Assume we have a fixed number of subjects a we want to include in a future study to assess agreement between measurements. Then we may want to determine the necessary number of observers b to obtain a certain half-width M of the confidence interval in Eq. (5), such that the confidence interval is the estimated LOAM $\pm M$. This requires initial estimates of σ_B^2 and σ_E^2, say $\hat{\sigma}_{B,0}^2$ and $\hat{\sigma}_{E,0}^2$, which can be obtained from, e.g., a pilot study. Then b can be estimated by

$$\frac{1.96^4}{2aM^2}\ \frac{(a-1)\left(\hat{\sigma}_{B,0}^2\right)^2+\left(\hat{\sigma}_{B,0}^2+\hat{\sigma}_{E,0}^2\right)^2}{\hat{\sigma}_{B,0}^2+\hat{\sigma}_{E,0}^2}.$$

Caution should be taken here, as the symmetric confidence interval tends to be artificially narrow as mentioned in Section 2.1.3. It would be preferable to estimate b using the asymmetric confidence interval in Eq. (6) instead, as this in general has a coverage probability closer to the desired 95%. However, as the dependency on b is more complicated we cannot obtain an estimate of b on closed form and numerical approximation is needed. To keep it simple, b may be estimated using the above formula, whereupon the width of the resulting asymmetric confidence interval is investigated for that specific choice of b.

2.1.5 Inference on the variance components

In order to assess the extent of the inter-observer and intra-observer variations, we suggest to consider a 95% confidence interval for σ_B and σ_E, respectively.

If the ANOVA estimate $\hat{\sigma}_B^2 > 0$, we simply estimate σ_B by $\hat{\sigma}_B = \sqrt{\hat{\sigma}_B^2}$. Using the statistical delta method we obtain the following approximate 95% confidence interval for σ_B:

$$\hat{\sigma}_B \pm \frac{1.96}{a\hat{\sigma}_B}\sqrt{\frac{(a\hat{\sigma}_B^2+\hat{\sigma}_E^2)^2}{2\nu_B}+\frac{(\hat{\sigma}_E^2)^2}{2\nu_E}}\ . \qquad (7)$$

It is well known that $\hat{\sigma}_B^2$ can be negative due to negative correlation between observers or simple random variation of the estimator. As we anticipate negative correlation between observers is unrealistic, it is tempting to suggest setting $\hat{\sigma}_B^2$ to zero, which, however, introduces bias in the estimation. We therefore suggest to report the negative estimates, and recommend the researcher to comment of the possibility of negatively correlated measurements, and if that does not seem realistic, to assess whether the confidence intervals are too wide to provide any clinical meaningful conclusion, or if more observers should be included.

As the distribution of $\hat{\sigma}_E^2$ is known on closed form, an exact asymmetric 95% confidence interval can easily be constructed for σ_E, see e.g. Chapter 4 of Searle et al. [3]. Alternatively, a symmetric but only approximate 95% confidence interval can be obtained for σ_E using the delta method:

$$\hat{\sigma}_E \pm 1.96 \frac{\hat{\sigma}_E}{\sqrt{2\nu_E}}. \tag{8}$$

2.1.6 Performing an agreement analysis

To investigate agreement between observers, we propose first to make the agreement plot with the estimate and confidence interval for the 95% LOAM from Sections 2.1.2 - 2.1.3, and to calculate the sample means and standard deviations for the measurements grouped by observer or subject. Inspection of the agreement plot and the sample means grouped by observer can be used to reveal whether any observers tend to make unusual large or small measurements. Further, the agreement plot and the grouped standard deviations can be used to check whether the assumption of homoscedasticity of the random model is fulfilled. If the model seems reasonable, we report the estimate and confidence interval for the LOAM along with the associated confidence intervals. Next, we may compare the order of magnitude of $\hat{\sigma}_B^2$ with $\hat{\sigma}_E^2$ to investigate how much of the variation is due to different observers. Further, we calculate the confidence interval of σ_B and σ_E. If clinicians deem the observer variation to be negligible, the observer effect could in principle be removed from the random model, entailing that the LOAM and the associated estimate and confidence intervals should be adjusted accordingly [see Additional file 2].

The agreement analysis may be supplemented with an estimate and confidence interval for the ICC. Various forms of ICCs are listed in McGraw and Wong for a range of models [5]. The two-way random effects model proposed in this paper corresponds to case 2A in McGraw and Wong, with subject as row

effect and observer as column effect, and ICC(A, 1) can then be used to assess absolute agreement of the measurements [5].

Based on the estimate and confidence interval for the LOAM (and possibly the ICC), it is up to clinicians to decide whether the agreement between measurements is satisfactory.

2.2 Multiple measurements on each subject per observer

The proposed LOAM and their estimates and confidence intervals can easily be extended to the case where each observer performs multiple measurements on every subject. If each observer performs c measurements on each subject, then the two-way random effects model is extended to:

$$Y_{ijk} = \mu + A_i + B_j + E_{ijk},$$

where Y_{ijk} is the k'th measurement performed by the j'th observer on the i'th subject for $i = 1, \dots, a$, $j = 1, \dots, b$, and $k = 1, \dots, c$.

Mimicking the arguments for the single measurement case, but now considering the differences $D_{ijk} = Y_{ijk} - \bar{Y}_{i\cdot\cdot}$, we propose the following 95% LOAM:

$$\pm 1.96 \sqrt{\frac{b-1}{b}\sigma_B^2 + \frac{bc-1}{bc}\sigma_E^2}.$$

Again σ_B^2 and σ_E^2 are estimated by the ANOVA estimates (see, e.g., Chapter 4 of Searle et al. [3]), which are given by

$$\hat{\sigma}_B^2 = \frac{MSB - MSE}{ac}, \qquad \hat{\sigma}_E^2 = MSE,$$

where now $MSB = SSB/\nu_B$ and $MSE = SSE/\nu_E$ with $SSB = ac\sum_{j=1}^{b}\left(\bar{y}_{\cdot j\cdot} - \bar{y}_{\cdots}\right)^2$, $SSE = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c}\left(y_{ijk} - \bar{y}_{i\cdot\cdot} - \bar{y}_{\cdot j\cdot} - \bar{y}_{\cdots}\right)^2$, $\nu_B = b - 1$, and $\nu_E = abc - a - b + 1$.

Note that the overall, subject-specific, and observer-specific averages ($\bar{y}_{\cdots}$, $\bar{y}_{i\cdot\cdot}$, and $\bar{y}_{\cdot j\cdot}$) are now also averaging across the multiple measurement index. With these definitions of SSB, SSE, ν_B, and ν_E and with $N = abc$, the LOAM estimate and confidence intervals still have the form given by Eq. (4)-(6).

Further, confidence intervals for σ_B and σ_E are obtained by Eq. (7)-(8), except that a is replaced with ac and $\hat{\sigma}_B^2, \hat{\sigma}_E^2, \nu_B$, and ν_E are defined as above.

Note that all formulas for the multiple measurement case reduce to those for the single measurement case, when $c = 1$.

2.3 Data and software

The statistical programming language R, version 3.6.1 [7], was used to analyse the data in the paper. An R-package, R-scripts, and the aortic data for the LOAM calculations in the present paper can be obtained from the GitHub repository: https://github.com/HaemAalborg/loamr.

3 Results

Example 1. In a study $b = 5$ thoracic radiologists measured $a = 40$ lung tumours (6). This study was also used as an example in Jones et al. (2). Table 1 shows the sample mean and standard deviation of the measurements for each radiologist, and Figure 1 displays the agreement plot. Estimates and confidence intervals of the 95% LOAM, ICC, σ_B, and σ_E are listed in Table 2. Neither the agreement plot nor the grouped sample mean indicate any observer systematically making unusual small or large measurements. Further, there is no indication of heteroscedasticity in relation to change in observer or to the size of the tumour.

The estimated 95% LOAM are ±1.1 centimetres. Note that the asymmetric confidence interval is much wider than the symmetric. As mentioned in Section 2.1.3, the symmetric confidence interval may be artificially narrow. The inter-observer standard deviation estimate is 0.29 cm with a confidence interval from 0.07 cm to 0.50 cm. In comparison, the intra-observer standard deviation estimate is 0.58. Jones et al. concluded that the LOAM are $\pm 1.96\,\hat{\sigma}_E = 1.1$ cm. Although, significant, this indicates the inter-observer variation in this study contributes with a negligible source of variation to the LOAM, which supports the practice where lung nodule measurement is performed by different radiologists.

Radiologist	*Mean*	*SD*
1	3.9	1.6
2	3.7	1.5
3	4.4	1.6

4	4.4	1.6
5	4.1	1.6

Table 1. Sample mean and standard deviation for each radiologist's tumour measurements.

<Figure 1 here>

Figure 1. Agreement plot for tumour size measurements in centimetres with the proposed 95% LOAM (dashed line) and associated 95% confidence interval (shading). In figure A, the symmetric confidence interval and in figure B, the asymmetric confidence interval.

LOAM (CI)	***ICC (CI)***	$\widehat{\boldsymbol{\sigma}}_{\boldsymbol{B}}$ ***(CI)***	$\widehat{\boldsymbol{\sigma}}_{\boldsymbol{E}}$ ***(CI)***
1.1 (0.94, 1.33)	0.84 (0.74, 0.90)	0.29 (0.07, 0.50)	0.58 (0.51, 0.64)

Table 2. Estimates and confidence intervals of the upper 95% LOAM, ICC, σ_B, and σ_E. Here the asymmetric confidence interval for the LOAM is given.

Example 2. Borgbjerg et al. consider three methods (OTO, LTL, and ITI) for assessing the maximum antero-posterior abdominal aortic diameter [6]. A total of b = 12 radiologists measured the aortic diameter $c = 2$ times on a = 50 still abdominal aortic images to assess which of the three methods were most reliable.

Using the methods described in Section 2.2 for multiple measurements, we calculate estimates and confidence intervals for the LOAM, σ_B, and σ_E (see Table 3) and make an agreement plot (see Figure 2). The observer variation constitutes a large part of the total variation and should not be excluded. The LTL method have the largest estimated LOAM, meaning that measurements made by this method tend to vary more. However, the wide confidence intervals for the LOAM indicate that more observers may be needed to assess this properly. We found significantly less intra-observer variation for the LTL and ITI compared to the OTO method. This finding is in line with the conclusion by Borgbjerg et al. which

suggests that it is advantageous to employ either the ITI or LTL method when repeated measurements are performed by the same observer [6].

Method	*LOAM (CI)*	$\widehat{\sigma}_B$ *(CI)*	$\widehat{\sigma}_E$ *(CI)*
OTO	3.15 *(2.75, 4.31)*	1.14 (0.65, 1.63)	1.20 (1.15, 1.25)
LTL	3.38 *(2.77, 5.06)*	1.46 (0.84, 2.07)	1.04 (0.99, 1.08)
ITI	2.88 *(2.37, 4.29)*	1.23 (0.71, 1.75)	0.90 (0.86, 0.93)

Table 3. Estimates and 95% confidence intervals for the upper 95% LOAM, ICC, σ_B, and σ_E for the aortic diameter measurements. Here the asymmetric confidence interval is considered for the LOAM.

<Figure 2 here>

Figure 2. Agreement plots for each of the three methods used to measure the aortic diameter along with the estimate and the asymmetric confidence interval for the 95% LOAM.

4 Discussion

Our results show it is possible to formulate measures for the agreement between multiple observers, equip them with confidence intervals, and extend them to multiple observations per observer, thereby providing a natural extension of Bland-Altman's graphical method.

In the study we have chosen to formulate a simple two-way random effects model, with additive observer and subject effects. However, in several cases the observers can react differently upon varying subjects. For single measurements this interaction effect is confounded with the residual error, but for multiple measurements this effect could in principle be modelled and estimated. However, we have in this work chosen not to walk down this alley in order to keep the paper focussed on a simple, yet useful extension of Bland-Altman's graphical method.

The limits proposed by Jones et al. do not account for any systematic observer variation, and no clear argumentation for choosing these exact limits are given. Further, Jones et al. only provide an estimate of the limits and no confidence interval, which would be useful for assessing the variation of the

limits. Inspired by the extensive literature (see e.g. McGraw and Wong (5)) on random effects models for observer agreement data we suggested to formalise Jones et al.'s approach under a simple two-way random effects model, which allows us to formulate a coherent statistical inference procedure for the LOAM.

5 Conclusions

We believe to have provided an easily accessible and useful statistical toolbox for researchers involved in assessing agreement between methods or individuals performing clinical measurements.

List of abbreviations

LOAM	Limits of agreement with the mean
ICC	Inter-class correlation
ANOVA	Analysis of variance

References

1. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet. 1986;327:307–10.

2. Jones M, Dobson A, O'brian S. A graphical method for assessing agreement with the mean between multiple observers using continuous measures. Int J Epidemiol. 2011;40:1308–13.

3. Searle SR, Casella G, McCulloch CE. Variance Components. Hoboken: John Wiley & Sons, Inc.; 1992.

4. Graybill FA, Wang C-M. Confidence intervals on nonnegative linear combinations of variances. J Am Stat Assoc. 1980;75:869–73.

5. McGraw KO, Wong SP. Forming inferences about some intraclass correlation coefficients. Psychol Methods. 1996;1:30–46.

6. Borgbjerg J, Bøgsted M, Lindholt JS, Behr-Rasmussen C, Hørlyck A, Frøkjær JB. Superior reproducibility

of the leading to leading edge and inner to inner edge methods in the ultrasound assessment of maximum abdominal aortic diameter. Eur J Vasc Endovasc Surg. 2018;55:206–13.

7. R Core Team. R: A Language and Environment for Statistical Computing. 2019. https://www.r-project.org/.

8. Erasmus JJ, Gladish GW, Broemeling L, Sabloff BS, Truong MT, Herbst RS, et al. Interobserver and intraobserver variability in measurement of non-small-cell carcinoma lung lesions: Implications for assessment of tumor response. J Clin Oncol. 2003.

Declarations

Ethics approval and consent to participate

Regarding the ethics approval and consent to participate, we refer to the statements in the original papers by Erasmus et al. [8] for the tumour sizes data and Borgbjerg et al. [6] for the abdominal aortic diameter measurement data.

Consent for publication

Not applicable

Availability of data and materials

The dataset on abdominal aortic diameter measurements supporting the conclusions of this article is available in the *loamr* repository: https://github.com/HaemAalborg/loamr. The dataset on tumour sizes is not publicly available but is available from the corresponding author of the original paper on request [8].

Competing interests

Not applicable

Funding

Not applicable

Authors' contributions

MB and JB designed the study. MB and HSC did the statistical modelling and analyzed the data. HSC wrote the first version of the manuscript. LB produced figures and organized data and scripts into an R package. All authors read and approved the final manuscript.

Acknowledgements

Not applicable

Additional material

The following additional pdf files are provided:

Additional file 1: Coverage probabilities from a small simulation study

Additional file 2: Formulae after removing the observer effect