Albert Mehl, Sven Reich, Florian Beuer, Jan-Frederik Güth

Accuracy, trueness, and precision – a guideline for the evaluation of these basic values in digital dentistry

Abstract

An increasing number of accuracy studies on 3D digitizing systems, especially intraoral scanning devices, are being published in scientific and educational journals. The methods, measurement values, and statistical parameters of these studies vary. Certain inconsistencies exist, which lead to difficulty in terms of interpretation and sometimes even questionable conclusions being drawn. These issues make it almost impossible to compare the results of such studies. One aspect inherent in this is the mutable use of basic terms describing the quality of measurement outcomes. A clear definition of such terms and clear instructions as to their respective calculation processes is essential for communication among scientists as well as for reporting measurement results to the dental community. Therefore, the aim of the present guideline is to provide a clear definition of the accuracy, trueness, and precision as the basic terms in the context of digital dentistry. The survey for this guideline included the application of ISO Norms and their expansion to special aspects concerning 3D data acquisition and, in particular, surface meshes. Additionally, the literature was screened to collect approaches, which can be seen as useful for dealing with these terms when performing different kinds of studies.

Keywords: *intraoral scanning, accuracy, precision, trueness, ISO standard, 3D evaluation*

Fundamental aspects

According to the description in ISO 5725-1[1], **accuracy** consists of precision and trueness. The term 'accuracy' in general involves a combination of all possible random components and a common systematic error or bias component during the measurement process (paragraph 3.6, ISO 5725-1[1] and ISO 3534-1[2]). A precondition for applying this ISO Norm is that the measurement method yields measurement values *on a continuous scale* and gives a *single value* (= one-dimensional) as the test result (paragraph 1.2, ISO 5725-1[1]).

In this context, **trueness** refers to the closeness of agreement between the arithmetic mean of a large number of test results and the true or accepted reference value (paragraph 0.1, ISO 5725-1[1]) (NB: the reference value itself is also one-dimensional). This trueness value displays, in general, the *systematic* errors. The measure of trueness is usually expressed in terms of **bias** (paragraph 3.7, ISO 5725-1[1]). In this sense, bias is defined as the *difference* between the expectation of the test results (arithmetic mean) and the true or accepted reference value (paragraph 3.8, ISO 5725-1[1] and ISO 3534-1[2]). In particular, ISO 20896-1[3] for accuracy testing of intraoral scanning systems uses the term 'bias' as a measure for trueness.

Precision refers to the closeness of agreement between independent test results obtained under stipulated conditions (paragraphs 0.1 and 3.12, ISO 5725-1[1]). Precision describes the *statistical* or *random* errors under repeated measurements (paragraph 0.2, ISO 5725-1[1]). Precision in this sense is also the general term for variability between repeated measurements (paragraph 0.4, ISO 5725-1[1]). According to the factors of variability, which are held constant or which can vary, repeatability and reproducibility are two special subgroups of precision (paragraph 0.4, ISO 5725-1[1]).

Precision is normally expressed in terms of standard deviations (SDs) (paragraph 0.4, ISO 5725-1[1]). In this recommendation, 'normally' expresses that the SD can also be replaced by other measures that are more suitable to describe the random errors in a specific test. Another important aspect to emphasize is that precision depends only on the distribution of random errors and does not relate to the true value or the specified value (paragraph 3.12, ISO 5725-1[1]). In particular, it explicitly does not mean that the SD of the trueness is the precision. This is sometimes used in a misleading manner.

The abovementioned terms and expressions can roughly be summarized as shown in Figure 1. Accuracy, trueness, and precision are *qualitative* performance characteristics, expressing the closeness of agreement between a measurement result and the value of the measurand. A quantitative estimate of the accuracy of a result is essential to define the degree of confidence that can be placed in it as well as the reliability of the decisions based on such a result[4]. This par-

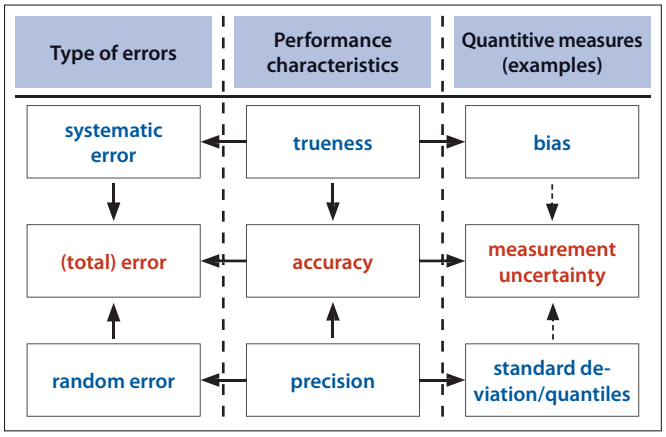

Fig 1 Basic concepts for the definition of terms used in the accuracy assessments (modified according to Menditto et al, 2006[4]).

ameter is the measurement uncertainty, which describes "the dispersion of the values that could reasonably be attributed to the measurand," often expressed as a standard deviation or as an interval including a larger fraction of such values (expanded uncertainty, quantiles)[4]. Guidance has been provided to use both the information provided by repeatability/reproducibility (or, in general, precision) and trueness (bias) estimates for the evaluation of the uncertainty of measurement (ISO 21748[5]). The dashed arrows in Figure 1 take into account the ongoing debate on the contribution of both bias and precision components to measurement uncertainty[4].

Standard deviation, standard error, and root mean square error

In statistics, the term 'error' means, in general, a deviation from a known or true value. The **root mean square error** (RMSE), in this sense, quantifies the average dispersion of a set of observations from a *known* value[6]. RMSE is calculated as:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2}$$

where μ is the known, correct value; n is the number of observations of μ; and x_i is one of a set of the n observations. As RMSE is a dispersion around a true value, it is a possible measure of accuracy[7]. However, it should not be mixed up with bias, as bias is a difference and therefore a linear error, whereas RMSE describes quadratic errors.

The **standard deviation** σ quantifies the average dispersion of a set of observations from an estimate of the set's mean value, and σ is calculated as:

$$\sigma = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\hat{\mu})^2}$$

where $\hat{\mu}$ is the *estimate* (an average of some kind) of the correct value computed from the observation set, n is the number of observations, and x_i is one of a set of the n observations[7]. Important to notice here is that the true value of μ is not known, and $\hat{\mu}$ must be calculated from:

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

According to the above definitions, this SD can be used as a measure for precision, where in a strict sense precision does not have to relate to the true value (paragraph 3.12, ISO 5725-1[1]). (Comment: in a similar way to the mean, the SD above is an estimate of the real SD, which should be denoted as $\hat{\sigma}$ in an exact sense. The problem with the definition of $\hat{\sigma}$ is that this value is not an expectationally correct parameter, which, in contrast, is the case for the variance $\hat{\sigma}^2$. This, and the usual notation, favors the use of σ instead of $\hat{\sigma}$).

Despite the similar equations, the SD and the RMSE can differ remarkably. For example, in case of a high bias, $\hat{\mu}$ is closer to the x_i than μ and, therefore, RMSE may be much higher than the SD; vice versa, a small σ does not imply a small RMSE.

The parameters SD and RMSE are related to each single test value x_i and describe how scattered or dispersed the single (!) measurement data are. However, if one wanted to indicate the uncertainty of such parameters themselves, which are calculated from a *set of* values x_i, the concept of **standard error** (or its synonym according to ISO 20896-1[3], **standard uncertainty**) is used. For example, if the uncertainty of the estimate of the mean measurement x_i is of interest, the standard error σ_{SE} of the mean is quoted[8]. This error correlates with the SD and is given by the following relation:

$$\sigma_{SE}(\hat{\mu}) = \sqrt{\frac{1}{n\,(n-1)}\sum_{i=1}^{n}(x_i-\hat{\mu})^2} = \frac{\sigma}{\sqrt{n}}$$

For a large sample, this standard error is also closely related to the **confidence interval** (CI) used in statistical testing: the 95% CI is obtained as the values of $1.96 \times \sigma_{SE}(\hat{\mu})$ either side of the mean[8]. In an analogous manner, the standard error σ_{SE} of the SD σ $(\sigma_{SE}(\sigma))$ and the standard error σ_{SE} of the RMSE $(\sigma_{SE}(RMSE))$ can be calculated.

Extension to non-normal distributions

If the underlying distribution of the test values x_i is not a normal distribution, the SD, RMSE, and standard errors can be used only with caution. In such situations, other parameters may describe the dispersion and errors of the test values. Possible parameters are median and quantiles, replacing the respective parameters mean and SD.

Before determining median and quantiles, a preprocessing step is necessary: all the test values x_i must be sorted according to their increasing value. This process results in a histogram, displaying the distribution of the values x_i (see also Figs 8 and 9). The median is now defined as the value in the histogram, where half of the values x_i lie below and half lie above this value. The median is also identical to the 50% quantile. The p%-quantile in general is defined as the value, below which p% of all values x_i in the histogram lie (this definition of quantiles here corresponds to the usage in the statistical program R; in other definitions this quantile usage here is equivalent to the percentile definiton).

The interval of $\pm\sigma$ around the mean $\hat{\mu}$ represents approximately 68.3% of the test values x_i ($\pm 2\sigma$ corresponds to 95.4%; $\pm 3\sigma$ corresponds to 99.7%). This percentage follows directly from the normal distribution and is therefore only correct if the test values x_i are distributed according to a normal distribution. As an alternative to that, for example, the 20%-quantile (q20) and the 80%-quantile (q80) can be determined. The range of q80 to q20 (q80–q20) covers *per definitionem* 60% of the test values x_i and is therefore comparable to $\pm\sigma$. The advantage of using the quantiles is:

a. Each arbitrary level of safety can be determined, ie, (q90, q10) covering 80% of the test values, (q95, q05) covering 90% of the test values, etc.
b. This quantity can be applied independently of the underlying distribution of the test values.
c. Even in case of a normal distribution (or approximately a normal distribution), small errors on the outer ends of the normal distribution have a critical influence on the concept of SD and on calculating the percentage of covering data. With quantiles this is more stable.

For example, if the aim is to investigate the quality of an intraoral scanner at the preparation margin with respect to the fit of the restoration, a measure that covers only 68% and excludes 32% of the measured points may not be very meaningful. In this case, for higher safety, the measure can more easily and more stably be described by quantiles, eg, the (q95, q05) range, covering 90% of all values and neglecting only 10%. Altogether, this is the reason why the measure with quantiles is increasingly used in metrology and quality assurance.

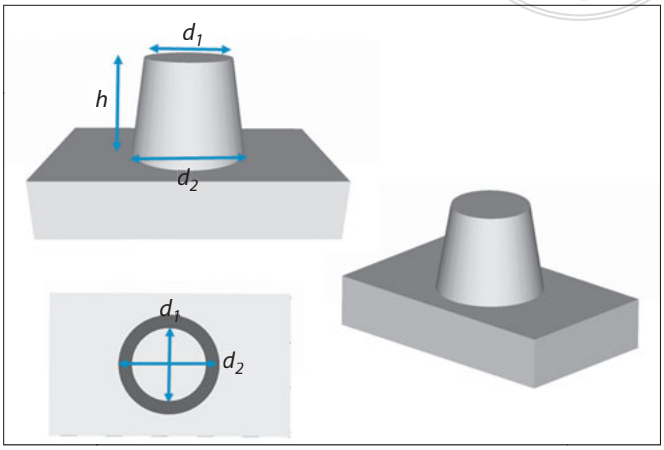

Fig 2 Example of a reference object with some known dimensions of interest: this standardized crown preparation is designed according to one reference object, as suggested by ISO 20896-1[3].

Calculation of accuracy parameters in case of one-dimensional measures

For intraoral scanning devices, the ISO Norm 20896-1[3] defines the relevant methods for assessing the accuracy. Three different reference objects are suggested: a schematic crown preparation, a schematic inlay preparation ('negative crown'), and a full jaw model with four spheres in the region of the first premolars and the second molars. For all reference objects, accuracy is assessed by measuring some known distances (eg, d_1, d_2, and h in case of a crown preparation similar to Fig 2). In ISO 20896-1[3], these known distances are defined by the term **dimensions of interest**, and are described as the distance between features of a test object that are required to be both measured independently as a reference or true value and estimated by the digital impression device[3]. The ISO Norm does explicitly rely only on in vitro measurements. In order to assess the accuracy under in vivo conditions, some research groups have extended the ISO methods in such a way that measurements can be performed in a clinically relevant intraoral situation (Fig 3)[9-12].

What all these methods have in common is that some known reference values (dimensions of interest) are compared with the measured test values, *and* the evaluation of the accuracy is performed for each known reference value

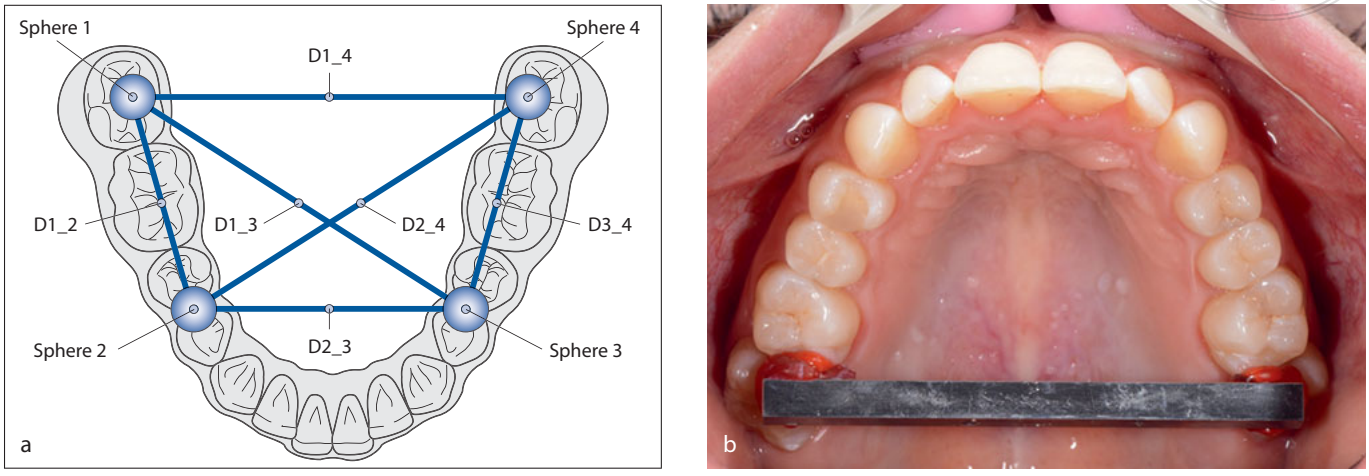

Fig 3 Examples of possible methods that transfer distance measurements, as suggested for in vitro investigations by ISO 20896-1[3], into an intraoral clinical situation (according to Schmidt et al[11] [a] and Keul and Güth[9] [b]).

separately. Therefore, for each evaluation, a single test value is assessed (and it is obvious that this value can be assumed as continuous), so that all the above-described preconditions apply for using the appropriate definitions from ISO 5725-1[1]. In ISO 20896-1[3] for intraoral scanning devices, these concepts are specified as follows (to describe the process more concisely, here – as an example – the description relies on one dimension of interest $d = d_1$; however, it can also be transferred to all other dimension of interest values $d = d_2$, $d = h$, etc [see Figs 2 and 3])[3]:

1. Reference measurement of dimension of interest

a. The dimension of interest d shall be determined by an **independent**, calibrated measurement, which in general is performed by a reference measurement device of high quality. The value obtained shall be considered the true value for the dimension of interest d_R. If the value comes from k repeated measurements, the true value is calculated as the mean of these measurements $d_R = \frac{1}{k} \sum_{j=1}^{k} d_{R,j}$.
b. The precision of this determination of k repeated measurements is to be expressed as the **standard error** (standard uncertainty) $\sigma_{SE}(d_R)$.
c. The standard error (standard uncertainty) in the reference value d_R shall be no greater than **one-fifth** (ie, 0.2 times) of the accuracy expected, required or claimed for the digitizing device.

2. Assessment of accuracy

a. Repeated measurements (n) of the **scanning device** under investigation is performed for the dimension of interest d. The mean $\hat{d}$ and SD σ is then evaluated from the n observations d_i (i = 1, … n).
b. Expression of bias: Bias Δd in the dimension of interest is expressed as the difference between the mean value obtained according to point 2a (above) and the corresponding reference value according to point 1a (above):
$\Delta d = \hat{d} - d_R$
Additionally, suitable statistics shall test the hypothesis that the bias is zero, and the statistical significance shall be noted.
c. Expression of precision: The precision is expressed as the standard deviation σ:
$\sigma = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (d_i - \hat{d})^2}$
d. Expression of accuracy: The accuracy of the measurement of a dimension of interest d shall be assessed as the combination of the trueness and the precision. The accuracy shall be evaluated as the greater of 1) the standard error (standard uncertainty) in the reference value, and 2) the RMSE between the measured values and the reference value, ie, by evaluating the statistic:
$s = max\left(\sqrt{\frac{1}{n} \sum_{i=1}^{n} (d_i - d_R)^2}; \sigma_{SE}(d_R)\right) = max\,(RMSE\,(d); \sigma_{SE}(d_R))$

After performing these steps, the following results should be summarized in a table for each dimension of interest (Table 1)[3]: the mean value $\hat{d}$, the precision expressed as the

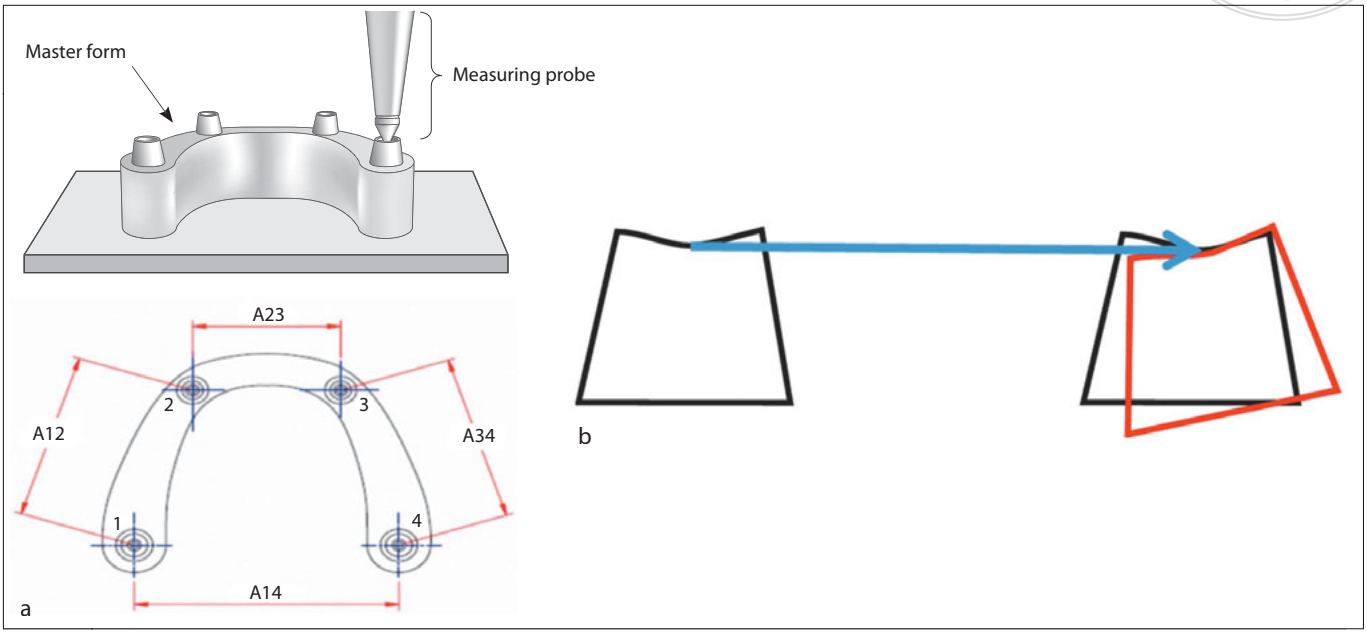

Fig 4 Reference measurements with single reference values – in this case, four distances (a): Deformations of the impressions and local inaccuracies cannot be revealed (b).

Table 1 Example of a table listing the measurement results as recommended by the ISO 20896-1[3]

Dimension of interest	Mean value $\bar{d}$	Precision σ	True value d_R	Standard error of d_R: $\sigma_{SE}(d_R)$	Bias (= trueness) Δd	Accuracy s
1.						
2.						
3.						

SD σ, the corresponding true value d_R, the standard error in the true value $\sigma_{SE}(d_R)$, the bias Δd (if the bias is assessed as non-zero, the level of significance shall be cited), and the accuracy s. It is important to mention that the RMSE measure is used for accuracy, whereas the SD σ is used for precision. The precision does not therefore rely on the true values (therefore not on trueness or bias), whereas the accuracy (defined as above) tries to combine trueness and precision into one measure.

Extending the ISO Norm to multidimensional measures

The ISO and the other abovementioned methods of accuracy evaluations focus on single dimensions of interest, ie, one-dimensional data such as one distance or one angle. These concepts are, of course, very reliable and are used in a broad bandwidth of quality assurance processes. They are also very useful to assess the quality of digital impression devices. However, intraoral scanning devices deliver thousands up to millions of surface points, resulting in a detailed and highly resolved surface representation necessary for different kinds of treatments or for diagnostic purposes. Therefore, the information in a single surface scan of an intraoral scanning device is high and is not fully described by a few dimensions of interest (see Figs 2 and 3). One problem can be seen in Figure 4: Measuring only linear distances does not give information about possible distortions, which can happen within an impression. Another important aspect is that, in case of linear measurements, geometric objects with symmetric shapes and surfaces without struc-

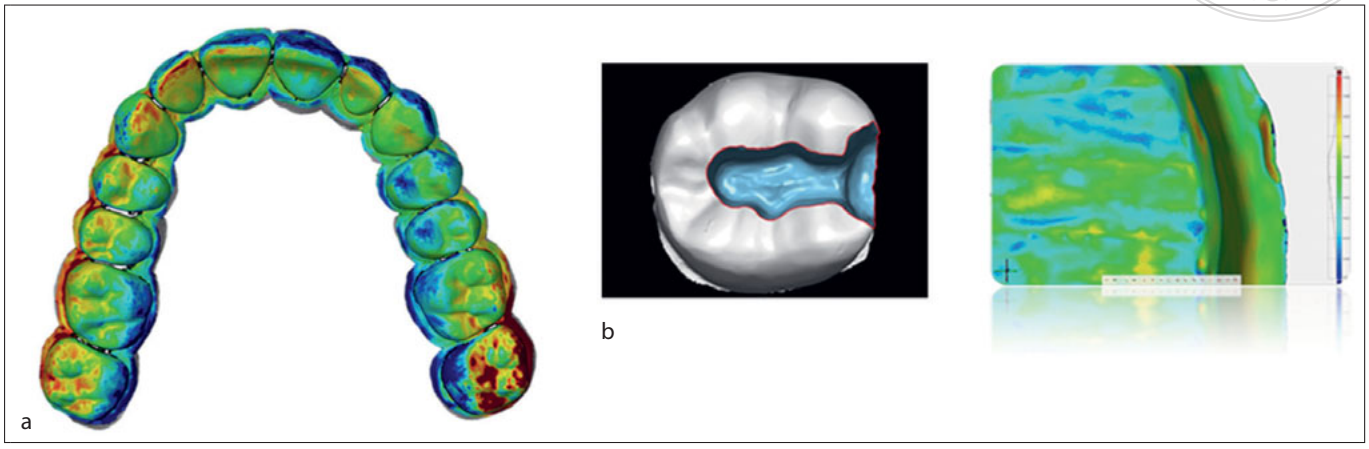

Fig 5 (a) Superimposition of scans and calculation of a high-dense pointwise difference image. (b) Surface scans for evaluating the preparation margins with high detail resolution.

tures are used. Although symmetric shapes are advantageous for mathematical procedures to detect linear one-dimensional measures, strictly symmetric objects are not quite suitable for algorithms that apply superimposing processes, both during accuracy evaluation and during data acquisition, such as with intraoral scanners. Therefore, more and more clinically realistic models are used, and from that a high proportion of accuracy studies nowadays focus on surface comparison methods (Fig 5)[13,14].

In contrast to the single value processes before, in case of surface comparison methods there is no specific reference dimension of interest that can be evaluated separately; instead, the reference surface consists of a combination of single reference vertices and each single distance value in the color-coded images in Figure 5 corresponds to one single reference dimension of interest. In general, this results in a large number of distance values (equivalently to a multidimensional measure) that are calculated from each surface point (vertex) of one scan to the surface or vertices of another scan. The task is to define a process that comes close to the definitions above, as there is no ISO Norm to date that deals with this situation. Therefore, the qualitative performance definitions of accuracy, trueness, and precision in their original sense must be applied (see Fig 1), and suitable qualitative measures have to be deduced from that. To repeat: Accuracy consists of precision and trueness (ISO 5725-1[1]). Precision describes how close repeated measurements are to each other. The higher the precision, the more predictable the measurement. Trueness describes how far the measurement deviates from the actual dimensions of the measured object. A high trueness delivers a result that is close or equal to the actual dimensions of the measured object[13].

In this sense, the following stepwise processes for assessing trueness and precision can be recommended:

1. Reference measurement of the physical object surface

a. The object shall be measured by an **independent**, calibrated measurement device of a high quality. The data set of the 3D object surface obtained shall be considered the *true reference surface*. (Comment: If k repeated measurements are performed to determine the precision of the reference measurement device, the true surface cannot, in general, be calculated by averaging the single surfaces. Instead, a single representative out of the k measurements should be selected as the *true reference surface* [see, as an example, Reich et al[14]]).
b. If the precision of this determination is not known, k repeated measurements should be made to determine this precision, analogous to what is described below in 2c.
c. The error (precision) in the reference values of the dimensions of interest shall be no greater than **one-fifth** (ie, 0.2 times) of the trueness and/or the precision expected, required or claimed for the digitizing device. (Comment: The threshold of one-fifth comes from ISO 20896-1[3]. Especially in case of multidimensional measurements, it is sometimes difficult to fulfill this threshold. Therefore, it should be interpreted here as a recommendation; nevertheless, the precision of the independent, calibrated reference measuring device should be noted).

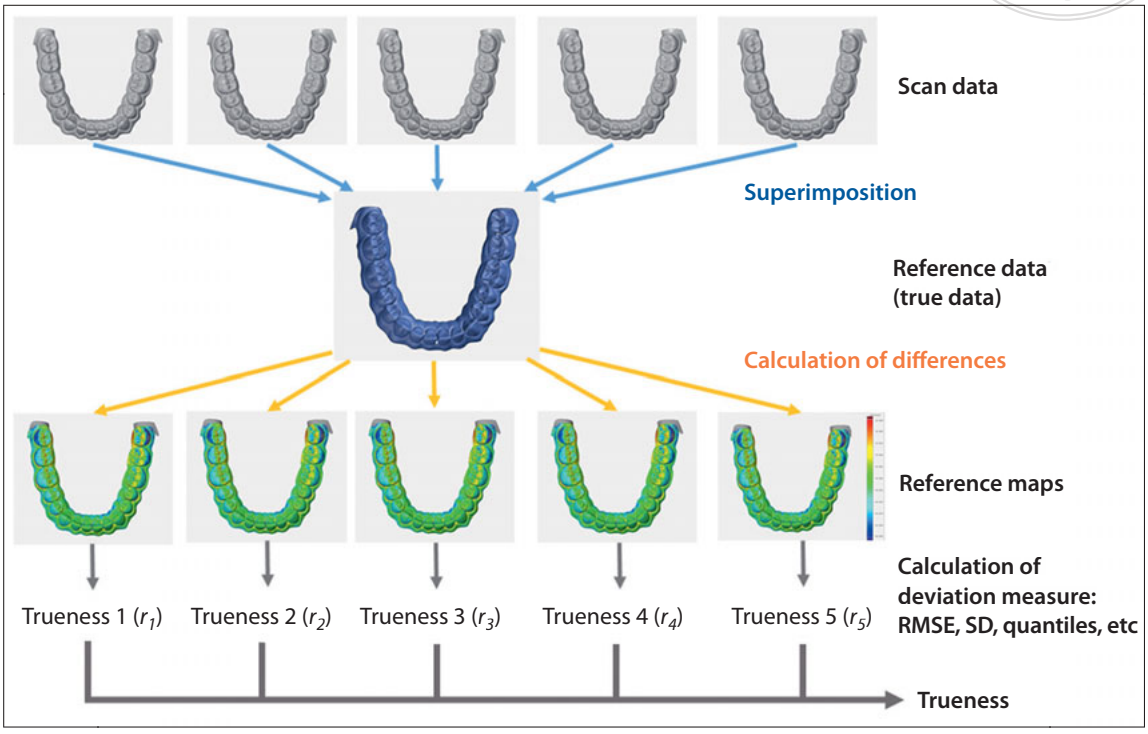

Fig 6
Procedure of determination of the trueness: Each scan is superimposed with the reference data set (eg, with an ICP [iterative closest point] or other general best fit algorithm). The difference map is then calculated for each scan and the measure of trueness (r_i) is determined.

2. Assessment of accuracy

a. Repeated surface measurements (n) of the scanning device under investigation are performed for the object surface, and the n respective 3D data sets are stored.
b. Expression of trueness (see also Fig 6): For trueness, each of the n 3D data sets is superimposed with the true reference surface, and a quantitative measure r_i (i = 1, …, n) for the 'closeness' for each superimposition is calculated. The trueness is then given by:
$tr = \hat{r} \pm \sigma(r_i)$ or $tr = median\ (r_i)\ [IQR(r_i)]$
c. Expression of precision (see also Fig 7): For precision, the n scans are superimposed pairwise, which results in $k = \binom{n}{2}$ pairwise superimpositions, and a quantitative measure m_i (i = 1, …, k) for the 'closeness' for each superimposition is calculated. The precision is then given in an analogous manner by:
$pr = \hat{m} \pm \sigma(m_i)$ or $pr = median\ (m_i)\ [IQR(m_i)]$

Here, the trueness *tr* and precision *pr* are expressed with an additional standard deviation (or IQR [interquartile range] in case of non-normal distribution) to provide an estimate for the distribution of the respective values. For all the above-mentioned superimposition processes, it is important to use a quantitative meaningful measure r_i or m_i that represents the inaccuracy for *one* arbitrary surface point, representative of the entire surface region under consideration. To calculate the deviation to the true surface or to another surface for each surface point, several parameters are useful: these are usually the distance from a vertex of one surface to the nearest vertex of the other surface (point–point), the distance from a vertex of one surface to the nearest surface point (point-surface) or the distance from a vertex of one surface in the direction of the surface normal until a hit with the other surface occurs (point–normal direction). The distances can be signed or absolute values. By calculating these distance parameters for each surface vertex, a distribution as shown in Figure 8a and Figure 9b will result.

This distribution provides all the information relevant for a meaningful quantitative measure to assess the agreement between two surfaces. Such a measure can be one of the following descriptors: RMSE (assuming a normal distribution around 0, used mainly for trueness), σ (used mainly for precision), quantiles or absolute mean (ie, the mean value of the absolute values of all distances; see Figs 8 and 9). For a complete description of the evaluation procedure in a study, therefore, the following parameters and information should be specified:

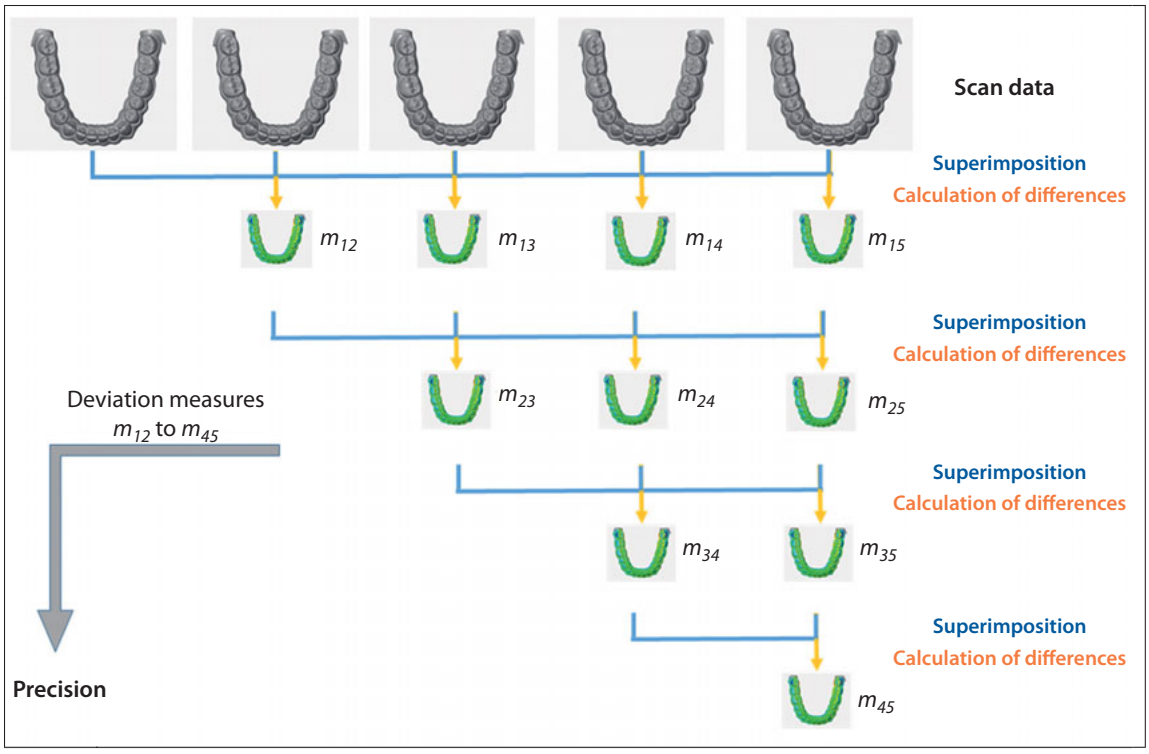

Fig 7 Procedure of determination of the precision: Each scan is superimposed with each other scan (eg, with an ICP or other general best fit algorithm). The difference maps are then calculated and the respective measure of deviation m_{ij} between the two scans *(i and j)* is determined for each combination.

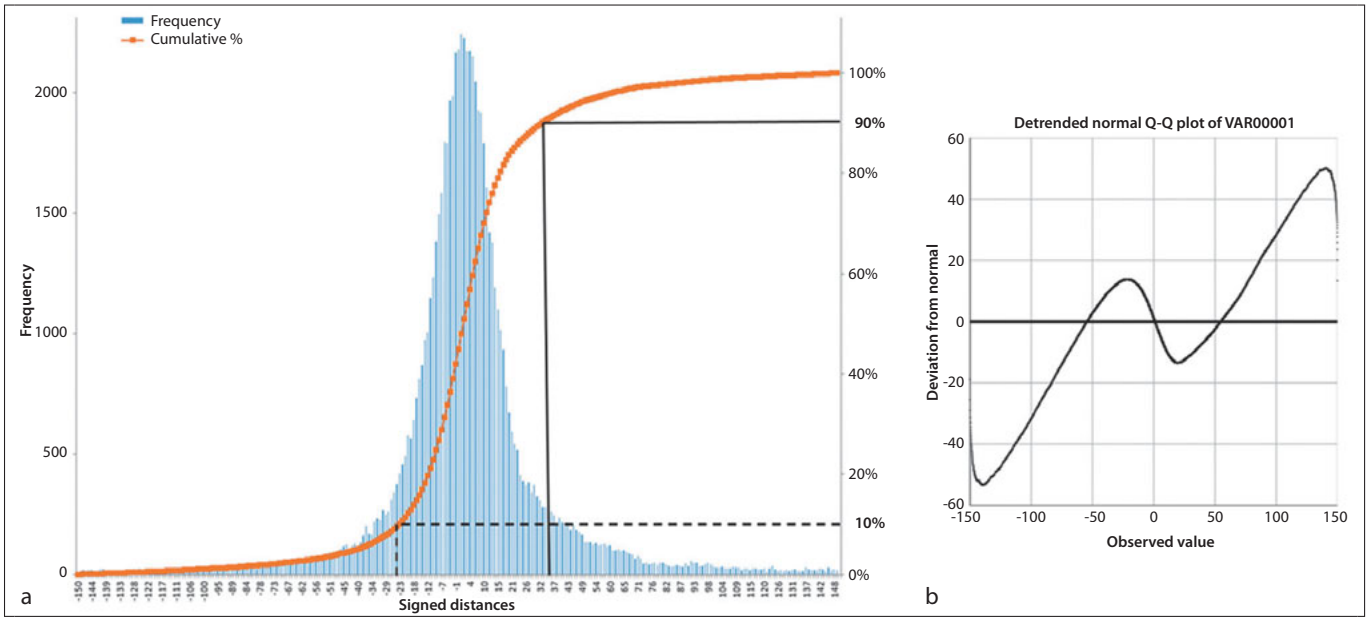

Fig 8 (a) Histogram of distance value distribution (signed distances) after superimposition of two scanned surfaces (eg, superimposition from Fig 5, with around 75,000 surface points and distance values, respectively). The 10%-quantiles (q10) and the 90%-quantiles (q90) are shown. The standard deviation (SD) calculated for these distributions is $\sigma = 37\ \mu m$. (b) The Q-Q plot clearly shows that the distribution of the distance values does not follow a normal distribution and the quantiles seem to be the more reliable measures.

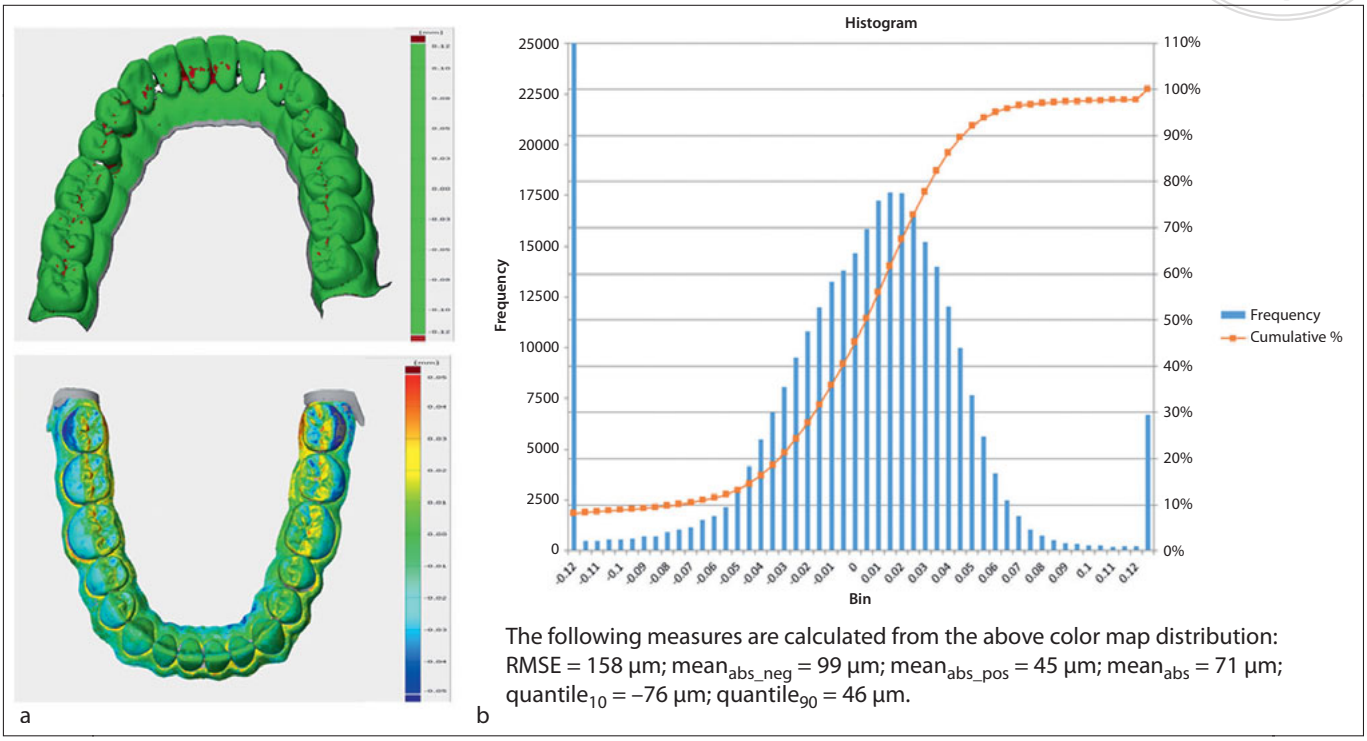

Fig 9 An example of a difference image and the respective distance value distribution for 311,000 surface points. (a) Bottom: color map with range from -100 to 100 µm. Top: Only distance values (in red) are shown, which are < -120 and > 120 µm. Such areas are mostly located in critical regions such as interproximal surfaces or at the margins of the scan. These outlier values are represented in the histogram in 'b' in the first and last frequency bar (< -120 and > 120 µm). (b) Some measures can be calculated from the histogram distribution. With this typical situation, it can be seen that the outliers have a strong influence on the RMSE and on the mean absolute values. Therefore, in most cases, it is recommended to rely on quantile measures.

- The method of distance measurement (point-point, point–surface, point–normal direction, etc).
- The quantitative measure that describes the distance distribution (RMSE, σ, quantiles, absolute mean, absolute negative mean, absolute positive mean, etc).
- The upper and lower limits, which define the range of the distance values used for the histogram or, equivalently, for the calculation of the quantitative measures. These limits are set automatically or manually in most programs to exclude outliers; however, this may have a strong influence on the quantitative measure.

Discussion

Trueness and precision for systems and processes in digital dentistry are determined in a confusing variety, and often not in accordance with metrological and ISO standards. The present article provides a guideline for the determination of trueness and precision. In case of linear (one-dimensional) distance measurements, various ISO standards provide a clear framework for the evaluation processes, which are described and developed here in detail. In case of surface comparison methods (multidimensional measures), no ISO standard is actually applicable. For that situation, an evaluation process is suggested, which extrapolates the concepts for one-dimensional cases to multidimensional measures and which also relies on the fundamental definitions of trueness and precision. This guideline can help to standardize accuracy studies in digital dentistry and to deliver enhanced comparability of results. This article focuses on the correct procedure for the determination of trueness and precision. In general, the preprocessing steps such as data acquisition from standardized models, data representation, and computing algorithms also have an influence on the respective accuracy outcome and should be carried out with care. Surface comparison methods are especially critical with respect to the superimposition method used, eg,

which best fit method, which outlier quantile, which distance measure or which region is used. The same is true for the calculation of the distance map, in which, for example, the two or more superimposed images should be cut to the same region of interest. This could even be accomplished automatically by a preregistration of all data files to be compared, so that they are aligned in the same coordinate system. Thereafter, all files can be cut identically.

In general, the range of the color map (eg, see Figs 5 and 6) should be chosen in accordance with the relevant hypothesis of a given investigation, for example, if the accuracy of full-jaw arch impressions is investigated, the range of interest should be defined somewhere around ± 50 or ± 100 µm. This allows the visual comparison of different systems/groups, if all the color maps are fixed to these limits. However, if the differences for only one group are under investigation, the color maps favorably coincide with the range or quantiles used as a measure for the precision or trueness, eg, in case of 90%-quantile to 10%-quantile, the upper and lower limit of the color bar should also be set to the 90%-quantile (upper limit) and to the 10%-quantile (lower limit). In some cases, this will lead to an asymmetric setting, but this can be handled in most of the programs used for the alignment calculation (additionally, they keep the 0 value green, and scale the positive and negative color ranges according to the positive and negative quantiles). If there is only the possibility of a symmetric adjustment, then it is recommended to take the min (abs[90%quantile], abs[10%quantile]) as the upper and lower limits of the color range.

The distribution of the distance values in the 3D-evaluation process can be described by RMSE, σ or quantiles. Other measures also exist and are used in some cases. One of them is the amount of the object surface, which lies over and under the reference surface. The drawback of this measure is twofold: first, on morphological complex surfaces it may be difficult to calculate the exact surface (especially if there are holes, etc, which then has a great impact on the measure); second, and more importantly, the amount of area below and above gives no hint of the real extent of the deviation in a specific region: a deviation of only a few microns will account for the same impact as a deviation of a few hundred microns. This gives rise to many problems. Further measures are the average positive and negative mean (or, in addition, the 'average absolute mean'); in principle, these values may also be used. However, if we assume a normal or an approximatively normal distribution, these values correspond to approximately $0.8 \cdot \sigma$ (0.8 SD), therefore leaving out around 25% of points with a lower deviation and 75% of points with a higher deviation (it also corresponds approximately to the 25%- and the 75%-quantiles). The risk here is that these values underestimate errors, which may happen on morphological constellations relevant to the dental tasks at hand. Therefore, it is strongly recommended to use a more restrictive measure, which includes 60% (80%-/20%-quantile), 66% (SD) or, even better, 80% (90%-/10%-quantile) and more deviations (= surface points) of the surface.

Finally, it is also worth mentioning that data acquisition with intraoral scanners and the superimposition of very smooth geometric or symmetric surfaces cannot be performed in a precise manner by most programs and methods. Therefore, accuracy evaluations using surface comparisons should be performed with special care and under full control over the evaluation software and program used.

Conclusions

- Accuracy comprises precision and trueness.
- For single distance measurements or single dimensions of interest (one-dimensional measures), the evaluation process of accuracy, precision, and trueness (bias) can be fully based on ISO standards.
- In case of surface comparison methods (multidimensional measures), the evaluation process of precision and trueness can be extrapolated from the concepts for one-dimensional cases by applying the qualitative definitions of trueness and precision.
- In case of surface comparison methods, a meaningful quantitative measure has to be defined that describes the distance distribution (RMSE, σ, quantiles, absolute mean, etc).
- The quantitative measure should be based on whether or not there is a normal distribution of the distance values.

Acknowledgments

The authors would like to thank the DDM (Dynamic Digital Model) Working Group and Dr B. Reiss for initializing the idea for this guideline and giving the space for profound discussions.

Disclaimer

The authors declare no conflicts of interest and received no funding in relation to this study.

References

1. ISO 5725-1. Accuracy (trueness and precision) of measurement methods and results -- Part 1: General principles and definitions. International Organization for Standardization, Beuth Verlag Berlin, 1994.
2. ISO 3534-1. Statistics — Vocabulary and symbols — Part 1: General statistical terms and terms used in probability. International Organization for Standardization, Beuth Verlag, Berlin, 2006.
3. ISO 20896-1:2019(E). Dentistry — Digital impression devices — Part 1: Methods for assessing accuracy. ISO copyright office, Geneva, Switzerland, 2019.
4. Menditto A, Patriarca M, Magnusson B. Understanding the meaning of accuracy, trueness and precision. Accredit Qual Assur 2006;12:45-47.
5. ISO 21748:2017. Guidance for the use of repeatability, reproducibility and trueness estimates in measurement uncertainty evaluation. ISO copyright office, Geneva, Switzerland, 2017.
6. Ghilani CD. Adjustment Computations: Spatial Data Analysis, ed 5. Hoboken, New Jersey: Wiley, 2010:672.
7. Meyer ThH. Root mean square error compared to, and contrasted with, standard deviation. Surveying and Land Information Science 2012;72:107-108.
8. Altman DG, Bland JM. Statistics notes - standard deviations and standard errors. BMJ 2005;331:903.
9. Keul C, Güth J-F. Accuracy of full-arch digital impressions: an in vitro and in vivo comparison. Clin Oral Investig 2019;24:735–745.
10. Kuhr F, Schmidt A, Rehmann P, Wöstmann B. A new method for assessing the accuracy of full arch impressions in patients. J Dent 2016;55:68–74.
11. Schmidt A, Klussmann L, Wöstmann B, Schlenz MA. Accuracy of digital and conventional full-arch impressions in patients: an update. J Clin Med 2020;9: 688, doi:10.3390/jcm9030688.
12. Wesemann Ch, Kienbaum H, Thun M, Spies BC, Beuer F, Bumann A. Does Ambient Light Affect the Accuracy and Scanning Time of Intraoral Scans? J Prosthet Dent 2020; doi: 10.1016/j.prosdent.2020.03.021.
13. Ender, A, Mehl A. Accuracy of complete-arch dental impressions: a new method of measuring trueness and precision. J Prosthet Dent 213;109: 121-8.
14. Reich S, Yatmaz B, Raith S. Do "cut out-rescan" procedures have an impact on the accuracy of intraoral digital scans? J Prosthet Dent 2020; doi: 10.1016/j.prosdent.2019.11.018.

Genauigkeit, Richtigkeit und Präzision – Leitfaden zur Ermittlung dieser Messgrößen in der digitalen Zahnmedizin

Zusammenfassung

In wissenschaftlichen Fachzeitschriften wird eine zunehmend hohe Anzahl an Genauigkeitsstudien zu digitalen 3-D-Systemen, insbesondere zu intraoralen Scannern und Fräsmaschinen, veröffentlicht. Die Methoden, die Messgrößen und die statistischen Parameter unterscheiden sich dabei zum Teil sehr deutlich, was zu einer schwierigen Interpretation, manchmal sogar zu fraglichen Schlussfolgerungen, und damit zu einer geringen Vergleichbarkeit der Ergebnisse solcher Studien führt. Ein Aspekt, der in diesem Zusammenhang besonders beachtet werden muss, ist die richtige Verwendung der Begriffe Genauigkeit, Richtigkeit und Präzision. Eine eindeutige Definition dieser Begriffe und klare Anweisungen zu ihrer jeweiligen Ermittlung sind für die Kommunikation unter Wissenschaftlern sowie für die Weitergabe von Messergebnissen an die zahnmedizinische Fachwelt unerlässlich. Ziel dieser Publikation ist es daher, einen Leitfaden für die Grundbegriffe Genauigkeit, Richtigkeit und Präzision im Kontext der digitalen Zahnmedizin zu geben. Grundlage für den vorliegenden Leitfaden war die Anwendung der einschlägigen ISO-Normen und deren Erweiterung auf spezielle Aspekte in Bezug zur 3-D-Datenerfassung, insbesondere auf die 3-D-Oberflächendaten. Zusätzlich erfolgte eine Literatursuche, um verschiedene weitere Verfahren zu berücksichtigen, die für die Ermittlung dieser Messgrößen bezugnehmend auf spezielle Fragestellungen als empfehlenswert angesehen werden können.

Indizes: *Intraorales Scannen, Genauigkeit, Präzision, Richtigkeit, ISO-Norm, 3-D-Auswertung*

Albert Mehl

Albert Mehl, Prof Dr med dent Dr rer biol hum
Division of Computerized Restorative Dentistry, Clinic of Conservative and Preventive Dentistry, Center of Dental Medicine, University of Zurich, Zurich, Switzerland

Sven Reich, Prof Dr med dent
Subject and Research Area of Computerized Dentistry at the Department of Prosthodontics and Biomaterials, RWTH University Hospital, Aachen, Germany

Florian Beuer, Prof Dr med dent
Department of Prosthodontics, Geriatric Dentistry and Craniomandibular Disorders, University Charité Berlin, Germany

Jan-Frederik Güth, Prof Dr med dent
Professor, Department of Prosthetic Dentistry, University Hospital, Ludwig-Maximilians-University Munich, Munich, Germany

Address Prof Dr Dr Albert Mehl, Division of Computerized Restorative Dentistry, Clinic of Conservative and Preventive Dentistry, Center of Dental Medicine, University of Zurich, Plattenstrasse 11, Zurich CH-8032, Switzerland, Email: albert.mehl@zzm.uzh.ch