 doi:10.20944/preprints202204.0104.v1

Statistical modelling and experimental design for the validation of droplet digital PCR methods

Steffen Uhlig[1], Christopher Weidner[2] and Bertrand Colson[1]

[1] QuoData GmbH, Prellerstr. 14, 01309 Dresden, Germany
uhlig@quodata.de

[2] Department Method Standardisation, Reference Laboratories, Resistance to Antibiotics, Federal Office of Consumer Protection and Food Safety (BVL), P.O. Box 110260, 10832 Berlin, Germany

Abstract

For the in-house validation of a droplet digital PCR method, a factorial experimental design was implemented. This design serves different purposes. On the one hand, it is an efficient design in relation to the workload involved in achieving a desirable level of reliability of variance estimates. On the other hand, it allows a partitioning of total variance into different components, thus providing information regarding the dominant sources of random variation. The statistical modelling reflects the actual measurement mechanism, establishing relationships between nominal target DNA copies per well, the range of variation of copy numbers per droplet, probability of detection values, and estimated numbers of copies.

Keywords

Method validation, droplet digital PCR, orthogonal factorial design, variance components, Poisson assumption, cloglog model, target DNA copies per droplet, Monte Carlo, prediction interval

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 April 2022 doi:10.20944/preprints202204.0104.v1

Introduction

The in-house reproducibility precision reflects random variation inside a given laboratory. Having identified the influence factors which account for this random variation, a factorial design can be implemented. In connection with digital PCR, an emerging technology for sequence specific detection and quantification of nucleic acids [1, 2, 3], factors such as the following may have an effect on test results: technician, PCR system, fluorescence marker, PCR-enzyme-mix, use of restriction enzyme, etc. Each factor is realized across different factor levels. For example, if the design requires 2 different technicians, then the factor *technician* has 2 levels. In a factorial design, different factor levels are combined to form measurement conditions called *settings*. If there are q factors, each with 2 levels, there are a total of 2^q possible settings. Thus, if there are more than 3 factors, it may be necessary to reduce the number of settings. This can be achieved via orthogonal designs, ensuring that it remains possible to distinguish the main effects. A general introduction to orthogonal designs can be found in [8]. The use of orthogonal designs in connection with method validation studies is described in [4, 5].

Statistical modelling

The statistical model was developed in such a way as to reflect the actual measurement mechanism of droplet digital PCR (ddPCR). It is based on models described in [6, 7]. The quantitative test result is calculated on the basis of individual *binary* results for between 10 000 and 20 000 droplets. The term binary is used here to refer to detection/non-detection of target DNA copies.

The fundamental statistical assumption is that, for a given number of DNA copies x per well, the number of copies inside a given droplet follows a Poisson distribution with parameter λ_x (average number of copies per droplet). Accordingly, the average number of copies per droplet is estimated as follows:

$$\hat{\lambda}_x = -\ln\left(N_0/N_a\right) \qquad \text{Equation 1}$$

where N_0 denotes the number of droplets with no detected copies and N_a denotes the number of “accepted” droplets. It is this estimate $\hat{\lambda}_x$ from which the final quantitative test result is obtained (via the droplet volume).

Assuming that every DNA copy is detected, the POD (probability of detection, i.e. the probability that an individual droplet is positive) for a given x is

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 April 2022 doi:10.20944/preprints202204.0104.v1

$$POD(x) = 1 - \exp(-\lambda_x). \qquad \text{Equation 2}$$

This model is refined by introducing the sensitivity parameter a $(0 \leq a \leq 1)$:

$$POD(x) = 1 - \exp(-a \cdot \lambda_x). \qquad \text{Equation 3}$$

As can be seen, the POD increases with a. The value $a = 0$ corresponds to $POD = 0$, no matter how large x is. The value $a = 1$ corresponds to $POD = 1 - \exp(-\lambda_x)$, i.e. the method functions perfectly in the sense that all copies are indeed detected.

Since the sensitivity may itself depend on x, the model is further refined as follows:

$$POD(x) = 1 - \exp\left(-a \cdot {\lambda_x}^b\right). \qquad \text{Equation 4}$$

The new parameter b allows the sensitivity to depend on x via the new parameter $a_b(x)$:

$$a_b(x) = a \cdot {\lambda_x}^{b-1}. \qquad \text{Equation 5}$$

Ideally, b is equal to 1 and the sensitivity does not depend on x.

Rearranging Equation 4 and taking the logarithm (twice), we obtain

$$\ln\left(-\ln\left(1 - \mathrm{POD}(x)\right)\right) = \ln a + b \cdot \ln \lambda_x. \qquad \text{Equation 6}$$

In view of the linear structure on the right-hand side of Equation 6, the parameter b is called the slope parameter.

It should be noted that, in Equation 6, λ_x denotes the *nominal* number of copies per droplet, rather than the estimate $\hat{\lambda}_x$ (from which the quantitative test result is calculated). The nominal value, λ_x is obtained via the nominal copy number value per well (or per unit volume) by simple conversion, taking into account the droplet volume and any other relevant quantities (e.g. relative proportions of sample DNA and master-mix in each well). The quality of the fit is enhanced by taking the relative quantity $\ln\frac{\lambda_{max}}{\lambda_x}$ rather than $\ln \lambda_x$ as the nominal value on the right-hand side of Equation 6. In this relative quantity, λ_{max} denotes the (nominal) number of copies per droplet for the *highest* concentration level in the validation study. It

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 April 2022 doi:10.20944/preprints202204.0104.v1

should be noted that $\ln\frac{\lambda_{max}}{\lambda_x}$ is proportional to the number of amplification cycles. Indeed, since we have $\ln\frac{\lambda_{max}}{\lambda_x} = \ln\lambda_{max} - \ln\lambda_x$, the lower $\ln\lambda_x$ is, the greater $\ln\frac{\lambda_{max}}{\lambda_x}$. Conversely, when $\ln\lambda_x = \ln\lambda_{max}$, we have $\ln\frac{\lambda_{max}}{\lambda_x} = 0$. With this new nominal value, the two regression parameters are now denoted $\tilde{a}$ and $\tilde{b}$.

The factors are taken into consideration as follows:

$$\ln\left(-\ln\left(1-\mathrm{POD}_i(x)\right)\right) = \ln\tilde{a} + \tilde{b}\cdot\frac{\lambda_{max}}{\lambda_x} + \eta_i + \theta_i\cdot\ln\frac{\lambda_{max}}{\lambda_x} \qquad \text{Equation 7}$$

In Equation 7, the index i corresponds to the number of the run during the droplet digital PCR measurement series. The constant effect η_i is the sum of all the constant factorial effects $\eta_i = \gamma_{11}\cdot z_{i11} + \gamma_{12}\cdot z_{i12} + \cdots + \gamma_{q1}\cdot z_{iq1} + \gamma_{q2}\cdot z_{iq2}$, where γ_{kl} denotes the effect of factor k ($k = 1, \ldots, q$) for factor level l and z_{ikl} denote the corresponding design matrix entry (0 or 1).

The *proportional* run effect θ_i has the same structure.

Accordingly, in this model the variation in $\ln\left(-\ln\left(1-\mathrm{POD}_i(x)\right)\right)$ in accounted for by a constant component $Var(\ln\tilde{a} + \eta_i)$ and a proportional component $\left(\ln\frac{\lambda_{max}}{\lambda_x}\right)^2\cdot\mathrm{Var}(\tilde{b} + \theta_i)$. The factorial effects are modelled as fixed effects but interpreted as random effects. In order words, for each factor $k = 1, \ldots, q$, a variance component σ_k^2 is calculated from the difference between the $\ln\hat{\lambda}_x$ values at the two factor levels. Furthermore, estimates for run and repeatability variance components can be obtained. The in-house reproducibility variance σ_{iR}^2 is calculated as

$$\sigma_{iR}^2 = \sigma_{run}^2 + \sigma_r^2 + \sigma_1^2 + \cdots + \sigma_q^2. \qquad \text{Equation 8}$$

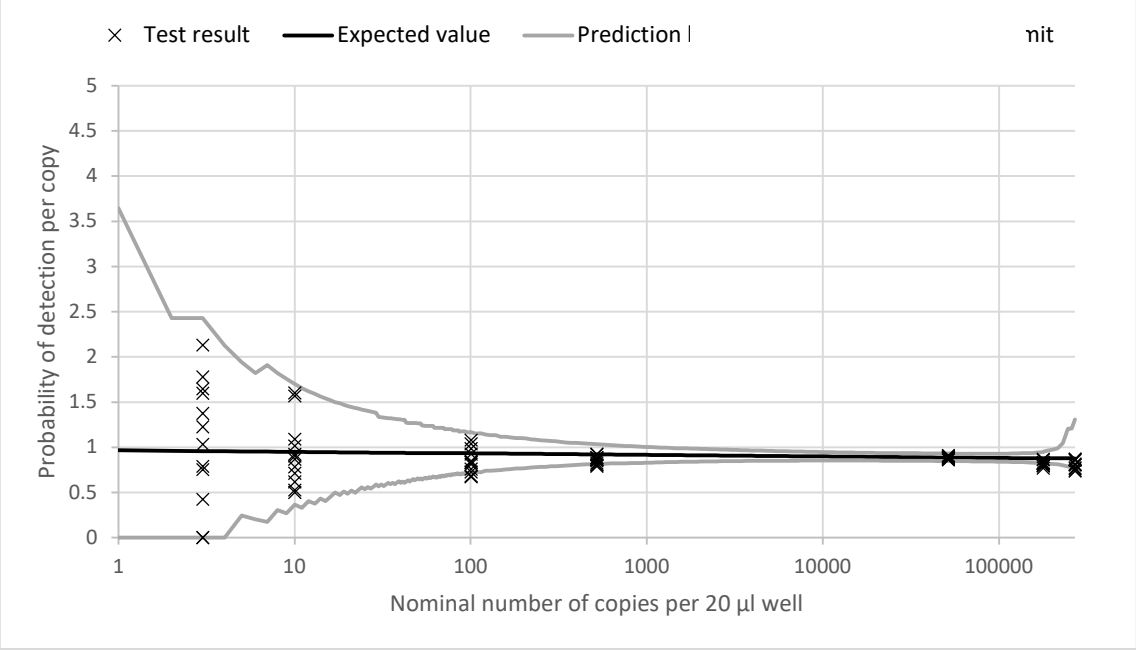
It should be noted that σ_{iR}^2 depends on the nominal number of DNA copies.

While σ_{iR}^2 is a measure of the random variation of the natural logarithm of the number of copies per droplet, a more relevant question is how to characterize the random variation of the quantitative ddPCR test results expressed as copies per unit volume. The latter can be calculated on the basis of σ_{iR}^2 via a Monte Carlo approach, see [9]. Separately for each nominal number of copies, σ_{iR} can be used

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 April 2022 doi:10.20944/preprints202204.0104.v1

to calculate a prediction interval around $\ln \tilde{a} + \tilde{b} \cdot \frac{\lambda_{max}}{\lambda_x}$. A large number N_{sim} (e.g. $N_{sim} = 10^6$) of random normally distributed values from this prediction interval can then be generated. Each such value y can then be converted to a probability $p = 1 - \exp(-\exp(y))$ that a droplet is positive. A random binomially-distributed vector (whose length is e.g. the mean number of acceptable droplets) can then be transformed to "ones" and "zeros" via p, and then converted into DNA copies per well via the basic Poisson assumption as described above (see Equation 1). The random variation of the quantitative ddPCR test results can then be characterized via the distribution of these N_{sim} copies per well values. This distribution reflects both the random variation due to the factors and repeatability effects (corresponding to the technical uncertainty in ISO 19036 [10]) and the random variation due to Poisson and binomial effects (corresponding to the distribution uncertainty in ISO 19036 [10]). At low copy numbers, the latter component can become very large, as seen in the following figure.

Figure 1: Test and predicted results for the ddPCR measurements. The test results are best displayed after division by the nominal copy numbers, yielding probability of detection per copy values. For each nominal copy number, the expected value is the mean value as predicted by the model. The prediction range reflects all the sources of variation (i.e. in-house reproducibility variation).



References

[1] Vogelstein B, Kinzler KW (1999) Digital PCR. Proceedings of the National Academy of Sciences of the United States of America, 96(16), 9236-9241.

[2] Demeke T, Dobnik D (2018) Critical assessment of digital PCR for the detection and quantification of genetically modified organisms. Anal Bioanal Chem 410, 4039-4050 https://doi.org/10.1007/s00216-018-1010-1

[3] Huggett JF, Cowen S, Foy CA (2015) Considerations for digital PCR as an accurate molecular diagnostic tool. Clin Chem. 2015 Jan, 61(1), 79-88. doi: 10.1373/clinchem.2014.221366

[4] Jülicher B, Gowik P, Uhlig S (1998) Assessment of detection methods in trace analysis by means of a statistically based in-house validation concept. Analyst, Vol.123 (173-179).

[5] Jülicher B, Gowik P, Uhlig S (1999) A top-down in-house validation based approach for the investigation of the measurement uncertainty using factional factorial experiments. Analyst, Vol.124 (537-545).

[6] Uhlig et al. (2015) Validation of qualitative PCR methods on the basis of mathematical-statistical modelling of the probability of detection. Accred Qual Assur 20, 75–83. https://doi.org/10.1007/s00769-015-1112-9

[7] Uhlig S, Gowik P (2018) Efficient estimation of the limit of detection and the relative limit of detection along with their reproducibility in the validation of qualitative microbiological methods by means of generalized linear mixed models. J Consum Prot Food Saf 13, 79–87. https://doi.org/10.1007/s00003-017-1130-0

[8] Tamhane AC (2009) Statistical Analysis of Designed Experiments: Theory and Applications, John Wiley & Sons, Hoboken, New Jersey

[9] JCGM 101:2008 Evaluation of measurement data – Supplement 1 to the “Guide to the expression of uncertainty in measurement” – Propagation of distributions using a Monte Carlo method

[10] ISO 19036:2020, Microbiology of the food chain — Estimation of measurement uncertainty for quantitative determinations