Supplementary Information:

Prediction and Control of the Number of Cells in Microdroplets by Stochastic Modeling

Elvan Ceyhan[a,*], Feng Xu[b,*], Umut Atakan Gurkan[b,*], Ahmet Emrehan Emre[b], Emine Sumeyra Turali[b], Chung-an Max Wu[b], Utkan Demirci[b,c,#]

[a] *Department of Mathematics, College of Sciences, Koç University, Istanbul, Turkey*

[b] *Demirci Bio-Acoustic-MEMS in Medicine (BAMM) Laboratory, Center for Biomedical Engineering, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA*

[c] *Harvard-MIT Health Sciences and Technology, Cambridge, MA, USA.*

[*] *The authors contributed equally to this work*

[#] *Corresponding author:* udemirci@rics.bwh.harvard.edu

S.1. Generalized Linear Models

Generalized linear models (GLM) are a useful generalization of ordinary least squares regression. Poisson regression models are a special type of GLMs used for count data with logarithm as the canonical link function and Poisson distributed errors. The Poisson regression model attributes to a response variable Y a Poisson distribution whose expected value depends on a predictor variable X in the following fashion:

$$\log(E(Y)) = \alpha + \beta X. \tag{S1}$$

If Y are independent, then α and β can be estimated numerically with maximum likelihood estimation methods. Poisson regression is appropriate when the dependent variable, Y, is count or rate data, where count is the number of events for a specified time or region and the rate is a count of events occurring in a particular unit of time or space. The variance of Y is proportional to its mean, that is, $\mathrm{Var}(Y) = \tau\mu$ where μ is the mean of Y and the dispersion parameter τ usually equals one. When it is not, the model is often described as overdispersed Poisson or quasi-Poisson model. Other GLMs such as the negative binomial model may function better in such cases[1].

The underlying assumptions for a GLM model are[2]: (i) (probabilistic) independence of the n observations in the data set, (ii) correct specification of the variance function, (iii) correct specification of τ in $\mathrm{Var}(Y_i) = \tau\mu_i$, (iv) correct specification of the link function, (v) correct form for the predictor variables, X, in Eq. (S1), (vi) lack of undue influence of individual observations on the fit (*i.e.*, lack of outliers)[3]. In our data set it can be assumed that there is independence between observations from different droplets. In other words, the measurements in one droplet do not affect the measurements in the other droplets. The only measurements repeated within each droplet are the cell radius values and radius of one cell has a small influence on the radius of other cells. That is, if a cell is much larger than others, then its existence in a droplet may impede the presence of other cells in that droplet. However, in our experimental setup, the cell radii are comparable (3-16 μm, **Fig. S4**) and much smaller than the droplet radii (300-700 μm). We require variance homogeneity of the residuals (hence the data, more specifically number of cell per droplet values) for the range of variables in the data set in assumption (ii). The overdispersion parameter τ in assumption (iii) is one for Poisson and binomial data and needs to be checked to choose the appropriate GLM modeling (in our case, τ is much larger than one, suggesting the use of negative binomial model). The appropriate link function in assumption (iv) is the log link function in Poisson or negative binomial GLMs. We specified the explanatory variables as in their original form; in our data, there was no need for a transformation. Finally, to satisfy assumption (vi) we specified and excluded the outlier values (as 18 out of 166 droplets had extremely unusual cell count values, they were deemed as outliers).

S.2. Technical Details of the Models in the Article

In this study, we developed statistical and stochastic models to predict and control the number of cells per droplet (denoted N_{CPD}, henceforth) based on three factors: **(i)** the cell concentration in the ejection fluid, **(ii)** droplet radius, and **(iii)** cell radius. We modeled N_{CPD} values based on the cell concentration in the encapsulation fluid (1, 2, 4, 8, and 16 million cells per milliliter (mil/ml)), cell radius (3-16 μm) and droplet radius (300-700 μm). For cell concentration in the ejection fluid, we used five levels: 1, 2, 4, 8, and 16 mil/ml. We collected cell droplet data on 178 droplets at various concentration levels (see **Table S1**). We present the pairwise scatter plots of these three variables in Figure S1. We used the statistical software, R version 2.12.0 (http://www.r-project.org/), for our statistical computations and tests. We compared the variance and the mean of N_{CPD} values to determine which model would be better in each case.

We also provide the scatter plot of each pair of the three variables, namely, droplet radius, cell concentration, and N_{CPD} in **Figure S1**. We observe no relation between the droplet radius and the cell concentration. However N_{CPD} is positively correlated with the droplet radius and cell concentration. More specifically, we observed a positive linear relationship between droplet radius and number of cells per droplet and there is a quadratic or cubic relationship between N_{CPD} and cell concentration (**Fig. S1**).

S.2.1. Modeling N_{CPD} as a function of cell concentration and droplet radius (Model D-C2)

Among the variables, the variation of cell radius is much smaller compared to those of cell concentration and droplet radius. Hence, we first model N_{CPD} as a function of only cell concentration (X_{CC}) and droplet radius (X_{DR}) and ignore the cell radius (X_{CR}) for the time being. Since our experimental data indicates that the variance of N_{CPD} is significantly larger than its mean ($\mathrm{Var}(N_{CPD}) =$ 6734.12 and $\mathrm{Mean}(N_{CPD}) = 63.82$ with $p<.0001$ based on Dean's P_B test for overdispersion[4], indicating that negative binomial regression is more appropriate (than Poisson regression) for our data.

We start with the following quadratic model in negative binomial GLM:

$$\log(E(N_{CPD})) = \alpha + \beta_1 X_{DR} + \beta_2 X_{CC} + \beta_3\sqrt{X_{DR}} + \beta_4 X_{DR}^2 + \beta_5 X_{CC}^2 + \beta_6 (X_{DR} \times X_{CC}) + \beta_7 (X_{DR} \times X_{CC})^2 \qquad \text{(S2)}$$

where α is the intercept, β_i are the coefficients of the variables in the model. That is, we are modeling logarithm of N_{CPD} as a function of droplet radius and cell concentration together with some non-linear terms in our model. By our model selection procedure, the model is reduced to:

$$\log(E(N_{CPD})) = \alpha + \beta_1 X_{DR} + \beta_2 X_{CC} + \beta_3\sqrt{X_{DR}}. \qquad \text{(S3)}$$

Hence, only the variables X_{DR} and X_{CC} have significant contribution in explaining the variation in $\log(N_{CPD})$, but the nonlinear variables (except $\sqrt{X_{DR}}$) do not. See **Table S2** for the summary of coefficients (estimates, standard errors, z -scores, and p -values).

When we compare our model with the *null model* of no predictors (also known as *intercept-only model*, since it looks like $\log(E(N_{CPD})) = \alpha$) by likelihood ratio test, we observe that our model explains the variation in N_{CPD} significantly better than this intercept-only model. Hence the negative binomial regression model predicting N_{CPD} from droplet radius and cell concentration is statistically significant (χ^2 =333.21, df = 4, p < 0.0001) compared to the null model. Our model in Eq. (S3) with the estimated coefficients is as follows.

$$\log(E(N_{CPD})) = -11.0022 - 0.0247\times X_{DR} + 0.1890\times X_{CC} + 1.1528\times\sqrt{X_{DR}}. \qquad \text{(S4)}$$

Since the model is log linear, we can translate these coefficients into multiplicative effects in the predicted N_{CPD} as

$$\textbf{Model D-C2:}\quad E(N_{CPD}) = \left(1.6663\times10^{-5}\right)\times0.9756^{X_{DR}}\times1.2081^{X_{CC}}\times3.1672^{\sqrt{X_{DR}}}. \qquad \text{(S5)}$$

Observe that, the expected value of N_{CPD} increases as droplet radius or cell concentration increases. For example, the expected $\log(N_{CPD})$ increase is 0.1890 for a one-unit increase in cell concentration. That is, a one-unit increase in cell concentration causes the expected N_{CPD} to increase by a factor of $\exp(0.1890) = 1.2081$, holding X_{DR} constant. Notice also that the influence of the cell concentration and droplet radius are both strong in estimating N_{CPD} (but droplet radius is much stronger).

We also performed diagnostic checks for the validity of the model assumptions (for **Model D-C2**) by examining the deviance results (deviance is the difference in observed N_{CPD} value from the N_{CPD} value predicted by the model), **Figure S2**. First plot in **Figure S2** shows the deviance residuals of the current data from our fitted model, *i.e.*, it is a visual inspection of the goodness of fit of our model. If the assumptions of the model are met, the residuals should be scattered around 0 and should not have a trend in width of the band for the range of predicted values. The second plot in **Figure S2** is the normal QQ-plot of the deviance residuals, which checks for the normality (*i.e.*, Gaussian distribution) of the residuals. When the plot is close to $y = x$ line, the residuals are deemed approximately Gaussian (which is the case in our plot). Finally, the third plot in **Figure S2** presents the observed values against the predicted values. For a good model, this plot should be close to $y = x$ also. We observed that model assumptions are satisfied in this case. Hence, when the cell radius is fixed or its variation is negligible compared to the variation in the other variables (i.e., when the variance of cell radius is much smaller compared to the variances of other variables), **Model D-C2** seems to be a reliable tool to predict or estimate the N_{CPD} values for a given droplet radius and a cell concentration (within and around the ranges of the droplet radius and cell concentration values (see **Table 2** in the main text)).

S.2.2. Modeling N_{CPD} as a function of cell concentration, droplet radius, and cell radius (Model D-C3)

Unlike Model D-C2, now we incorporate cell radius into our modeling procedure. That is, the response variable of interest, N_{CPD}, is modeled with a negative binomial regression as a function of independent (i.e., predictor) variables, i.e., cell concentration, droplet radius, and cell radius. We treat each cell related data as a single data point, so for the cells in each droplet, N_{CPD} values are replicated, as well as X_{CC} and X_{DR} values. Hence, we have 11229 many of each of X_{DR}, X_{CR}, N_{CPD} and X_{CC} values. That is, we have cell droplet data on 11229 cells that are dispersed on 178 droplets at 5 concentration levels. However, our experimental data showed that the variance of N_{CPD} is significantly larger than its mean ($Var(N_{CPD}) = 8211.90$ and $mean(N_{CPD}) =$ 168.74 with $p < 0.0001$ based on Dean's P_B test for overdispersion), indicating that negative binomial regression is more appropriate. Notice that at this stage of analysis, we consider the cell radius (X_{CR}) as a potentially important factor in explaining or modeling the N_{CPD}, as cells have variable radius.

We start with the following model in negative binomial GLM:

$$\begin{aligned}\log(E(N_{CPD})) = \alpha + \beta_1 X_{DR} + \beta_2 X_{CR} + \beta_3 X_{CC} + \beta_4 \sqrt{X_{DR}} + \beta_5 X_{CC}^2 + \beta_6 (X_{DR} \times X_{CR}) + \\ \beta_7 (X_{DR} \times X_{CC}) + \beta_8 (X_{CR} \times X_{CC}) + \beta_9 (X_{DR} \times X_{CR} \times X_{CC}).\end{aligned} \tag{S6}$$

That is, we are modeling $\log(N_{CPD})$ as a function of droplet radius, cell radius, and cell concentration, and we consider some non-linear terms in our model. By our model selection procedure, the model was reduced to:

$$\log(E(N_{CPD})) = \alpha + \beta_1 X_{DR} + \beta_3 X_{CC} + \beta_4 \sqrt{X_{DR}} \tag{S7}$$

Notice that in the presence of droplet radius and cell concentration, cell radius has no additional significant contribution to the model, *i.e.*, after droplet radius and cell concentration are accounted for, cell radius only explains an insignificant part of the variation in N_{CPD}. That is, X_{CR} does not belong to the model in Eq. (S7). We compare the significant improvement of our model in Eq. (S7) with the intercept-only (i.e., the null) model, namely $\log(N_{CPD}) = \alpha + error$. The negative binomial regression model in Eq. (S7) predicting N_{CPD} was statistically significant (χ^2 =19552, df =3, p < 0.0001). On the other hand, if droplet radius and cell concentration were ignored, *i.e.*, only the cell radius (and square root of it) is used as the predictor variable, we get the following model:

$$\log(E(N_{CPD})) = \alpha + \beta_1 X_{CR} + \beta_2 \sqrt{X_{CR}}\,. \tag{S8}$$

But, when models in Equations (S8) and (S7) are compared, Eq. (S7) is better in explaining the variation in N_{CPD} (Akaike Information Criterion (AIC) for Eq. (S7) is 117297 while AIC for Eq. (S8) is 123634). However, given a specific cell concentration value, cell radius may be a significant variable in explaining N_{CPD}. In particular, when cell concentration levels are treated as a categorical factor (rather than a quantitative variable), we obtain the following model.

$$\begin{aligned}\log(E(N_{CPD})) = \alpha + \alpha_1 I(X_{CC}=2) + \alpha_2 I(X_{CC}=4) + \alpha_3 I(X_{CC}=8) + \alpha_4 I(X_{CC}=16) + \\ \beta_1 X_{DR} + \beta_2\sqrt{X_{DR}} + \beta_3\sqrt{X_{CR}} + \beta_4 I(X_{CC}=1)X_{CR} + \beta_5 I(X_{CC}=8)X_{CR} + \\ \beta_6 I(X_{CC}=2)\sqrt{X_{CR}} + \beta_7 I(X_{CC}=4)\sqrt{X_{CR}} + \beta_8 I(X_{CC}=8)\sqrt{X_{CR}} + \\ \beta_9 I(X_{CC}=16)\sqrt{X_{CR}} + \beta_{10} I(X_{CC}=2)X_{DR} + \beta_{11} I(X_{CC}=4)X_{DR} + \\ \beta_{12} I(X_{CC}=8)X_{DR} + \beta_{13} I(X_{CC}=16)X_{DR} + \beta_{14} I(X_{CC}=2)\sqrt{X_{DR}} + \\ \beta_{15} I(X_{CC}=4)\sqrt{X_{DR}} + \beta_{16} I(X_{CC}=8)\sqrt{X_{DR}} + \beta_{17} I(X_{CC}=16)\sqrt{X_{DR}}\end{aligned} \qquad \text{(S9)}$$

where $I(X_{CC}=k)$ is the indicator function which is 1, if $X_{CC}=k$ and 0 otherwise for $k=1,2,4,8,16$ mil/ml. The model in Eq. (S9) has a different form for each cell concentration level. In particular, we have

for $X_{CC}=1$ mil/ml, $\log(E(N_{CPD})) = \alpha + \beta_1 X_{DR} + \beta_2\sqrt{X_{DR}} + \beta_3\sqrt{X_{CR}} + \beta_4 X_{CR}$

for $X_{CC}=2$ mil/ml,

$$\log(E(N_{CPD})) = \alpha + \alpha_1 + (\beta_1+\beta_{10})X_{DR} + (\beta_2+\beta_{14})\sqrt{X_{DR}} + (\beta_3+\beta_6)\sqrt{X_{CR}}$$

for $X_{CC}=4$ mil/ml,

$$\log(E(N_{CPD})) = \alpha + \alpha_2 + (\beta_1+\beta_{11})X_{DR} + (\beta_2+\beta_{15})\sqrt{X_{DR}} + (\beta_3+\beta_7)\sqrt{X_{CR}}$$

for $X_{CC}=8$ mil/ml,

$$\log(E(N_{CPD})) = \alpha + \alpha_3 + (\beta_1+\beta_{12})X_{DR} + (\beta_2+\beta_{16})\sqrt{X_{DR}} + (\beta_3+\beta_8)\sqrt{X_{CR}} \quad \text{and}$$

for $X_{CC}=16$ mil/ml,

$$\log(E(N_{CPD})) = \alpha + \alpha_4 + (\beta_1+\beta_{13})X_{DR} + (\beta_2+\beta_{17})\sqrt{X_{DR}} + (\beta_3+\beta_9)\sqrt{X_{CR}}$$

Compared to the null model, namely $\log(E(N_{CPD})) = \alpha$, the model in Eq. (S9) is statistically significant (χ^2 =26504, df =18, p < 0.0001) predicting N_{CPD}. The model in Eq. (S9) with the estimated coefficients is

$$\begin{aligned}\log(E(N_{CPD})) = {} & 22.2990 - 73.1496 I(X_{CC}=2) - 41.6621 I(X_{CC}=4) - 16.6940 I(X_{CC}=8) - \\ & 21.9559 I(X_{CC}=16) + 0.0366 X_{DR} - 1.5999\sqrt{X_{DR}} - 1.7804\sqrt{X_{CR}} + 0.3071 I(X_{CC}=1) X_{CR} + \\ & 1.6326 I(X_{CC}=2)\sqrt{X_{CR}} + 2.3339 I(X_{CC}=4)\sqrt{X_{CR}} + 1.8783 I(X_{CC}=8)\sqrt{X_{CR}} + \\ & 1.8885 I(X_{CC}=16)\sqrt{X_{CR}} - 0.1329 I(X_{CC}=2) X_{DR} - 0.0785 I(X_{CC}=4) X_{DR} - \\ & 0.0296 I(X_{CC}=8) X_{DR} - 0.0398 I(X_{CC}=16) X_{DR} + 6.1733 I(X_{CC}=2)\sqrt{X_{DR}} + \\ & 3.5215 I(X_{CC}=4)\sqrt{X_{DR}} + 1.3878 I(X_{CC}=8)\sqrt{X_{DR}} + 1.8851 I(X_{CC}=16)\sqrt{X_{DR}}\end{aligned} \tag{S10}$$

Then, for $X_{CC}=1$ mil/ml,

$$\log(E(N_{CPD})) = 22.2990 + 0.0366 X_{DR} - 1.5999\sqrt{X_{DR}} + 0.3071 X_{CR} - 1.7804\sqrt{X_{CR}}$$

for $X_{CC}=2$ mil/ml, $\log(E(N_{CPD})) = -50.8506 - 0.0963 X_{DR} + 4.5734\sqrt{X_{DR}} - 0.1478\sqrt{X_{CR}}$

for $X_{CC}=4$ mil/ml, $\log(E(N_{CPD})) = -19.3631 - 0.0419 X_{DR} + 1.9216\sqrt{X_{DR}} + 0.5535\sqrt{X_{CR}}$

for $X_{CC}=8$ mil/ml, $\log(E(N_{CPD})) = 5.6050 + 0.0070 X_{DR} - 0.2121\sqrt{X_{DR}} + 0.0979\sqrt{X_{CR}}$

and

for $X_{CC}=16$ mil/ml, $\log(E(N_{CPD})) = 0.3431 - 0.0032 X_{DR} + 0.2852\sqrt{X_{DR}} + 0.1081\sqrt{X_{CR}}$

Since the model is log linear, we can translate these coefficients into multiplicative effects in the predicted N_{CPD} as

Model D-C3:

$$\begin{aligned}E(N_{CPD}) = {} & 4.8345\times10^{9} \times \left(1.7042\times10^{-32}\right)^{I(X_{CC}=2)} \times \left(8.0608\times10^{-19}\right)^{I(X_{CC}=4)} \times \left(5.6220\times10^{-8}\right)^{I(X_{CC}=8)} \times \\ & \left(2.9152\times10^{-10}\right)^{I(X_{CC}=16)} \times 1.0372^{X_{DR}} \times 0.2019^{\sqrt{X_{DR}}} \times 0.1686^{\sqrt{X_{CR}}} \times 1.3594^{I(X_{CC}=1)X_{CR}} \times \\ & 5.1169^{I(X_{CC}=2)\sqrt{X_{CR}}} \times 1.0318^{I(X_{CC}=4)\sqrt{X_{CR}}} \times 6.5421^{I(X_{CC}=8)\sqrt{X_{CR}}} \times 6.6093^{I(X_{CC}=16)\sqrt{X_{CR}}} \times \\ & 0.8755^{I(X_{CC}=2)X_{DR}} \times 0.9244^{I(X_{CC}=4)X_{DR}} \times 0.9709^{I(X_{CC}=8)X_{DR}} \times 0.9610^{I(X_{CC}=16)X_{DR}} \times \\ & 479.7524^{I(X_{CC}=2)\sqrt{X_{DR}}} \times 33.8350^{I(X_{CC}=4)\sqrt{X_{DR}}} \times 4.0060^{I(X_{CC}=8)\sqrt{X_{DR}}} \times 6.5870^{I(X_{CC}=16)\sqrt{X_{DR}}}.\end{aligned} \tag{S11}$$

Then, for $X_{CC}=1$ mil/ml,

$$E(N_{CPD}) = 4.8345\times10^{9} \times 1.0372^{X_{DR}} \times 0.2019^{\sqrt{X_{DR}}} \times 1.3594^{X_{CR}} \times 0.1686^{\sqrt{X_{CR}}}.$$

for $X_{CC}=2$ mil/ml, $E(N_{CPD}) = 8.2390\times10^{-23} \times 0.9081^{X_{DR}} \times 96.8620^{\sqrt{X_{DR}}} \times 0.8627^{\sqrt{X_{CR}}}.$

for $X_{CC} = 4$ mil/ml, $E(N_{CPD}) = 3.8970 \times 10^{-9} \times 0.9588^{X_{DR}} \times 6.8313^{\sqrt{X_{DR}}} \times 0.1740^{\sqrt{X_{CR}}}$.

for $X_{CC} = 8$ mil/ml, $E(N_{CPD}) = 271.7956 \times 1.0070^{X_{DR}} \times 0.8088^{\sqrt{X_{DR}}} \times 1.1030^{\sqrt{X_{CR}}}$ and

for $X_{CC} = 16$ mil/ml, $E(N_{CPD}) = 1.4094 \times 0.9967^{X_{DR}} \times 1.3299^{\sqrt{X_{DR}}} \times 1.1143^{\sqrt{X_{CR}}}$.

Notice that the dependence of N_{CPD} on X_{CR} and X_{DR} is different at each X_{CC} level. Observe that N_{CPD} increases as droplet radius increases, and N_{CPD} tends to decrease as cell radius increases. Furthermore, the droplet radius has stronger influence on N_{CPD} compared to cell radius and the droplet and cell radii have influence on N_{CPD} in reverse directions.

We also performed diagnostic checks for the validity of the model assumptions (for **Model D-C3**) by examining the deviance results (see **Fig. S3**). We observe that model assumptions are satisfied in this case. Hence, when the cell radius is assumed to vary, **Model D-C3** seems to be a reliable tool to predict or estimate the N_{CPD} values for a given droplet radius, cell radius, at a specific cell concentration value (within and around the ranges of droplet radius and cell radius values at the cell concentration values 1, 2, 4, 8, and 16 mil/ml only).

S.2.3. Modeling N_{CPD} as a function of cell concentration and the ratio of droplet radius to cell radius (Model R2-C2)

Previously, it was demonstrated that N_{CPD} is mostly determined by the cell concentration in the ejection fluid and the ratio of droplet radius to cell radius in acoustic picoliter droplets[5]. We model this relationship using GLM methods. The size ratio could be viewed as the volume ratio (*i.e.*, ratio of the total volume of the droplet to the total volume of cells encapsulated in the droplet, X_{VR}) assuming the droplets and cells are spherical. Let $X_{DR}^{(k)}$ be the droplet radius for k^{th} droplet and $X_{CR}^{k,i}$ cell radius for i^{th} cell in droplet k and n_k be the number of cells in k^{th} droplet. Hence, the volume ratio, X_{VR}, is $\left(X_{DR}^{(k)}\right)^3 \Big/ \sum_{i=1}^{n_k} \left(X_{CR}^{k,i}\right)^3$ in each droplet or $\left(X_{DR}^{(k)}\right)^3 \Big/ \left(X_{CR}^{k,i}\right)^3$ or simply radius ratio as $X_{RR} = X_{DR}^{(k)} \big/ X_{CR}^{k,i}$ for each cell in a droplet. Analysis of the histograms in **Fig. S6** suggests that X_{RR} is more appropriate for our analysis. Here, recall that X_{DR} is not the radius of the actual three dimensional (3D) droplet, but the radius on the substrate that we could measure.

We consider the ratio of droplet radius to cell radius ratio for each cell, i.e., X_{RR}. We model N_{CPD} as a function of X_{RR} and X_{CC}. Similar to the previous cases, negative binomial regression is more appropriate, since $Var(N_{CPD}) = 6846.88$ is significantly larger than the mean $mean(N_{CPD}) = 65.89$ ($p < .0001$ based on Dean's P_B test for overdispersion). We have cell droplet data on 171 droplets and 10226 cells (radius values were not available for 1079 of the cells, hence removed from the analysis) at some concentration levels. The response variable is N_{CPD} and the predictor variables are X_{RR} and X_{CC}.

We started with the following quadratic model in negative binomial GLM:

$$\log(E(N_{CPD})) = \alpha + \beta_1 X_{RR} + \beta_2 X_{CC} + \beta_3 \sqrt{X_{RR}} + \beta_4 X_{RR}^2 + \beta_5 X_{CC}^2 + \beta_6 X_{RR} \times X_{CC} + \beta_7 (X_{RR} \times X_{CC})^2. \quad \text{(S12)}$$

After model selection procedure, the model was reduced to:

$$\log(E(N_{CPD})) = \alpha + \beta_1 X_{RR} + \beta_2 X_{CC} + \beta_5 X_{CC}^2. \quad \text{(S13)}$$

When we compare our model with the initial model of no predictors (*i.e.*, with the null model) by likelihood ratio test, we observe that our model explains the variation in N_{CPD} significantly better than the intercept only (i.e., null) model. Hence the negative binomial regression model predicting N_{CPD} from X_{RR} and X_{CC} is statistically significant ($\chi^2 = 19175$, df = 3, $p < 0.0001$).

The model with the estimated coefficients is

$$\log(E(N_{CPD})) = 2.0485 + 0.0021 \times X_{RR} + 0.3546 \times X_{CC} - 0.0098 \times X_{CC}^2. \quad \text{(S14)}$$

The coefficients of the log linear model can be translated into multiplicative effects in the predicted count as

Model R2-C2:

$$E(N_{CPD}) = 7.7564 \times 1.0021^{X_{RR}} \times 1.4256^{X_{CC}} \times 0.9902^{X_{CC}^2}. \quad \text{(S15)}$$

Notice that, N_{CPD} value tends to increase as X_{CC} increases and likewise the same holds for X_{RR} (to a weaker extent). Furthermore, the model diagnostic plots are presented in **Figure S5**, where we observe that although the model assumptions are be not severely violated, the QQ-plot suggests more severe non-normality compared to other models and the predicted-observed N_{CPD} plot indicates a worse fit compared to other models (see **Figures S2 and S3**).

S.3. Poisson Process

Poisson process is a stochastic counting process as a function of time $N(t)$ or space $N(R)$, which represents the number of events since time $t = 0$ or in region R, respectively. The number of events between time a and time b (or in region R) is given as $N(b) - N(a)$ (or $N(R)$) and has a Poisson distribution. The Poisson process has several general characteristics[6]: (i) orderliness which means that events do not occur simultaneously at the same time or spatial location; (ii) "no memory" which means the number of arrivals or events occurring in any bounded interval of time after time t is independent of the number of arrivals occurring before time t. Similarly, the number of events over two disjoint regions is independent.

A homogeneous Poisson process, $N(t)$, is characterized by a rate parameter λ, which denotes the number of occurrences of the event per unit time, such that the number of events in time interval $(a,b]$ follows a Poisson distribution with parameter $\lambda(b-a)$. Similarly, a homogeneous spatial Poisson process $N(R)$ over region R implies that the number of events over R follows a Poisson distribution with mean $\lambda \times \mathrm{volume}(R)$. If the rate parameter is a function of time or space, then the rate function is generalized to account for such spatiotemporal dependence[6].

S.4. Negative Binomial Process

The negative binomial distribution is a discrete probability distribution of the number of successes in a sequence of Bernoulli trials before a fixed number r of failures. For example, if one tosses a coin repeatedly until the third time “head” appears, then the probability distribution of the number of “tails” that had appeared will be negative binomial. Suppose in each Bernoulli trial the probability of success is p and of failure is $(1-p)$. Then the random number of successes we have seen, X, will have the negative binomial distribution: $X \sim NB(r,p)$. The probability mass function of the negative binomial distribution is

$$P(X=k)=\binom{k+r-1}{r-1}(1-p)^r p^k \text{ for } k=0,1,2,\ldots \tag{S16}$$

It is possible to extend the definition of the negative binomial distribution to the case of real-valued r with the probability mass function:

$$P(X = k) = \frac{\Gamma(k + r)}{k! \cdot \Gamma(r)} (1 - p)^r p^k \text{ for } k = 0, 1, 2, \ldots \tag{S17}$$

where r is a real, positive number and $\Gamma(r)$ is the gamma function. The negative binomial distribution, especially in its alternative parameterization described in Eq. (S17), can be used as an alternative to the Poisson distribution. It is especially useful for discrete data over an unbounded positive range whose sample variance exceeds its sample mean. In that case, the observations are over-dispersed with respect to the Poisson model. If a Poisson distribution is used to model such data, the model mean and variance are assumed to be equal. Since the negative binomial distribution has one more parameter than the Poisson, the second parameter can be used to adjust the variance independently of the mean [1].

S.5. Compound Spatial Inhomogeneous Poisson Process

A compound Poisson process with rate $\lambda > 0$ and jump size distribution G is a continuous-time stochastic process $\{\mathrm{Y(t)}, \mathrm{t} > 0\}$ which is given by $\mathrm{Y(t)} = \sum_{\mathrm{i=1}}^{\mathrm{N(t)}} \mathrm{S_i}$. $\{\mathrm{N(t)}, \mathrm{t} > 0\}$ is a Poisson process with rate λ and $\{\mathrm{S_i}, \mathrm{t} > 0\}$ are independent and identically distributed random variables with distribution function G, which are also independent of the Poisson process [6]. As before, when the variance is much larger than the mean, the Poisson process can be approximated by a negative binomial distribution.

To model the volumes of cells in the bioprinted droplets, we have performed a pilot study with 75 cells. A sample picture for the cells is presented in **Figure S4** (left). The kernel density plot for the volumes is presented in **Figure S7**. In the pilot study of 75 cells, we have measured the cell diameters. We also plot the kernel density estimate for the diameters in **Figure S7**. Cell radii have a distribution that resembles a normal distribution, which was further confirmed by the Lilliefor's test of normality ($p = 0.6681$)[7]. Cell volumes are approximately normal with a more emphasized right skew. However Lilliefor's test of normality yields $p = 0.0972$, hence it is close but not a significant evidence for non-normality. But when the entire data is considered (with around 10000 cells), we have evidence that cell radii is significantly non-normal ($p < .0001$) and in fact they have a log-normal distribution. We present the kernel density estimates of the cell radii for the cell concentration values (only 1, 2, and 4 mil ml^{-1} are presented) in **Figure S7**. The figures for the other concentrations are similar, hence are not presented here. These figures support the claim that cell radii are log-normal with different parameters at each cell concentration. The distribution of the logarithm of cell radius in our data can be modeled as a mixture of normal distributions. The probability density function of a log-normal distribution is:

$$f_x(x;\mu,\sigma) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{(\ln x-\mu)^2}{2\sigma^2}}, x > 0 \qquad \text{(S18)}$$

where μ and σ are the mean and standard deviation of the variable's natural logarithm (by definition, the variable's logarithm is normally distributed).

If X is a log-normally distributed variable, its expected value (mean), variance, and standard deviation are

$$E[X] = e^{\mu+0.5\sigma^2}, \quad Var[X] = \left(e^{\sigma^2}-1\right)e^{2\mu+\sigma^2}, \quad SD[X] = \sqrt{Var[X]} \qquad \text{(S19)}$$

S.6. N_{CPD} Modeled as a Negative Binomial Process

In this work, we count the N_{CPD} values, which depend on the volume of the droplet, *i.e.*, the droplet radius. Moreover, in our statistical modeling in **Section S2**, cell concentration and droplet radius were found to be the main factor in determining the N_{CPD} values. In the ejection fluid, we could model the number of cells, $N(R)$, in a given 3D region R as a spatial Poisson process. However, in **Models D-C2**, **D-C3**, and **R2-C2**, N_{CPD} is modeled using negative binomial regression, since variance of the dependent variable $Var(N_{CPD}) = \tau \times mean(N_{CPD})$ where the dispersion parameter τ is significantly larger than one (if it were not, we could have used Poisson regression). Hence, here we have an "overdispersed" Poisson random variable.

Considering the general characteristics of the negative binomial process, and our experimental setup, we can model the N_{CPD} by a homogeneous negative binomial process. Because the cell concentration is homogenized in the fluid by constantly stirring during the ejection process, the generalized rate function will mostly depend on the concentration of the cells X_{CC}: $\lambda = g(X_{CC})$. The form of this dependence is crucial, but for practical purposes, it is reasonable to assume g is the identity function, $g(x) = x$, provided that we have the same units for λ and X_{CC}.

A homogeneous Poisson process, $N(R)$, is characterized by a rate parameter λ, which denotes the number of occurrences of the event per unit volume or area, such that the number of events in region R follows a Poisson distribution with mean $\lambda \times V(R)$. Hence if the Poisson distribution were appropriate, the number of cells in a given 3D region R in would have the following pdf:

$$P(N(R) = k) = \frac{e^{-\lambda V(R)} (\lambda V(R))^k}{k!} \text{ for } k = 0,1,2,\ldots n \tag{S20}$$

where λ is the rate parameter (with its unit chosen to be mil/ml) and $V(R)$ is the volume of the region R in ml. Since the negative binomial distribution is more appropriate, the pdf of the number of cells in region R has the following pdf:

$$P(N(R) = k) = \frac{\Gamma(k+r)}{k!\Gamma(r)} \left(1 - \frac{\lambda V(R)}{\lambda V(R) + r}\right)^r \left(\frac{\lambda V(R)}{\lambda V(R) + r}\right)^k \text{ for } k = 0,1,2,.. \tag{S21}$$

where $r = \dfrac{\lambda V(R)}{\tau - 1}$ and $Var(N(R)) = \tau \times mean(N(R))$. Since the ejection fluid is homogenized and droplets are taken from the fluid so that a droplet represents a region with volume of the droplet, $V(D)$. In particular, N_{CPD} has a negative binomial distribution as in Equation (S21) above with $V(R)$ being replaced by $V(D)$.

Tables

Table S1. Number of droplets for each cell concentration

Cell concentration (mil/ml)	1	2	4	8	16
Number of droplets	23	55	30	30	37

Table S2: The summary table for Model D-C2 in the logarithmic form given in Eq. (S5).

Coefficients	Estimate	Std. Error	z score	p-value
Intercept	-10.9075	4.7268	-2.308	0.0210
X_{DR}	-0.0244	0.0096	-2.548	0.0108
X_{CC}	0.1890	0.0069	27.429	$<.0001$
$\sqrt{X_{DR}}$	1.1432	0.4274	2.675	0.0075

Table S3: The summary table for Model D-C3 in the logarithmic form given in Eq. (S11). Notice that some of the variables are omitted in the table for brevity.

Coefficients	Estimate	Std. Error	z score	p-value
Intercept	22.30	4.199	5.311	$<.0001$
$I(X_{CC}=2)$	-73.152	5.194	-14.084	$<.0001$
$I(X_{CC}=4)$	-41.66	4.671	-8.918	$<.0001$
$I(X_{CC}=8)$	-16.69	4.648	-3.591	0.0003
$I(X_{CC}=16)$	-21.96	4.235	-5.185	$<.0001$
X_{DR}	0.0366	.0083	4.381	$<.0001$
$\sqrt{X_{DR}}$	-1.5999	.3734	-4.285	$<.0001$
$I(X_{CC}=1)\times X_{CR}$	.3071	.1177	2.608	0.0091
…	…	…	…	…
$I(X_{CC}=4)\times X_{DR}$	-.0785	.0093	-8.406	$<.0001$
$I(X_{CC}=8)\times X_{DR}$	-.0296	.0096	-3.074	0.0021
$I(X_{CC}=16)\times X_{DR}$	-.0398	.0084	-4.734	$<.0001$
$I(X_{CC}=2)\times\sqrt{X_{DR}}$	6.173	.4516	13.671	$<.0001$
$I(X_{CC}=4)\times\sqrt{X_{DR}}$	3.521	.4161	8.463	$<.0001$
$I(X_{CC}=8)\times\sqrt{X_{DR}}$	1.388	.4222	3.287	0.0010
$I(X_{CC}=16)\times\sqrt{X_{DR}}$	1.885	.3763	5.009	$<.0001$

Table S4: The summary table of the logarithmic version of Model R2-C2 in Eq. (S15).

Coefficients	Estimate	Std. Error	z score	p-value
Intercept	2.0485	0.0231	88.79	$<.0001$
X_{RR}	0.0021	0.0001	16.75	$<.0001$
X_{CC}	0.3545	0.0045	78.18	$<.0001$
X_{CC}^2	-0.0098	0.0002	-43.20	$<.0001$

Figures

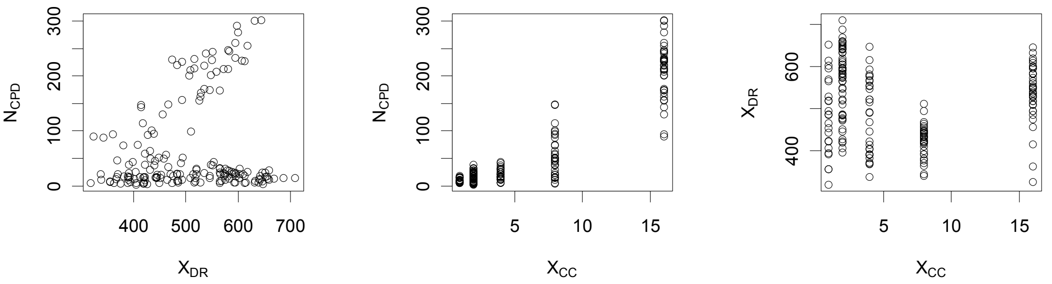

Figure S1. Pairs plot (*i.e.*, pairwise scatter plots for each pair of variables) of N_{CPD}, X_{DR} and X_{CC}.

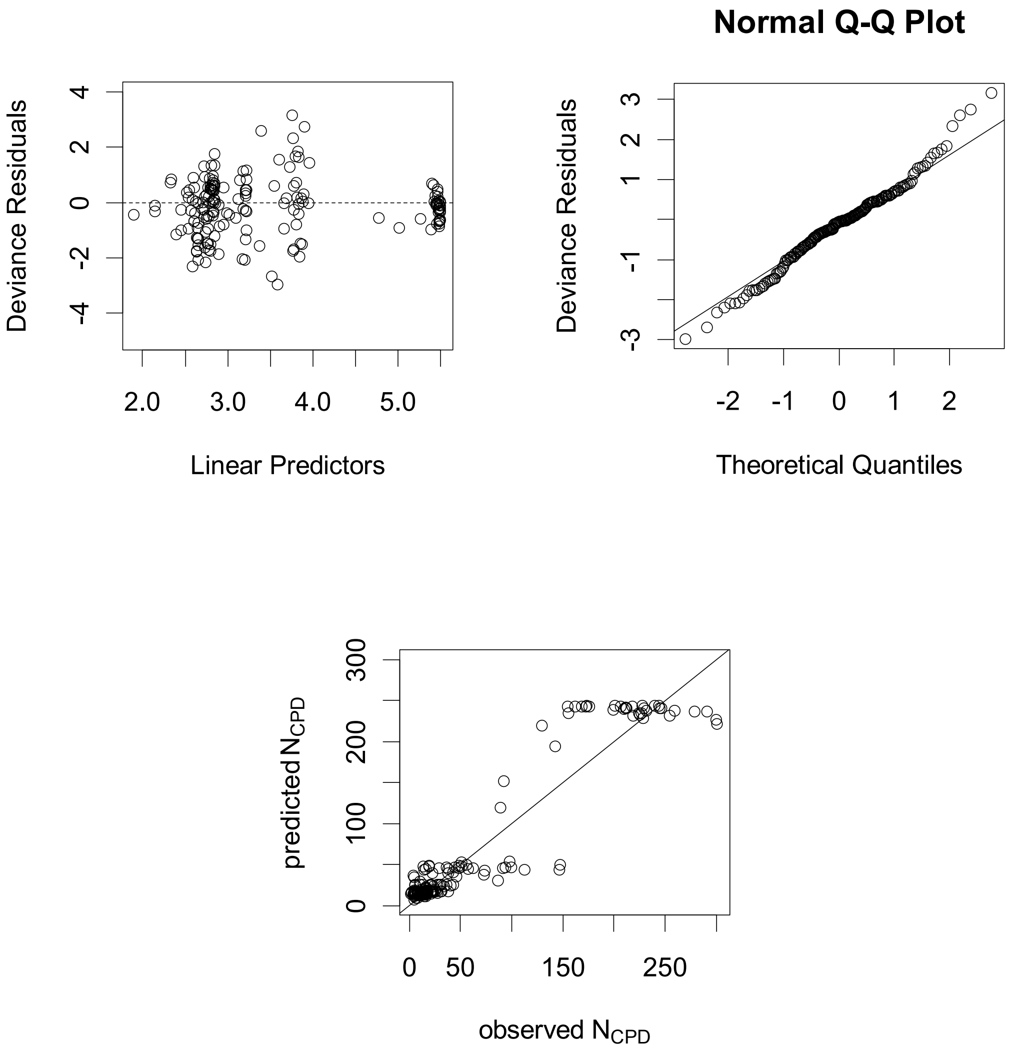

Figure S2. Diagnostic plots for Model D-C2. Plotted left is the deviance residuals versus predicted values and right is the normal QQ-plot for deviance residuals versus theoretical quantiles where the straight line passes through the first and third quartiles.

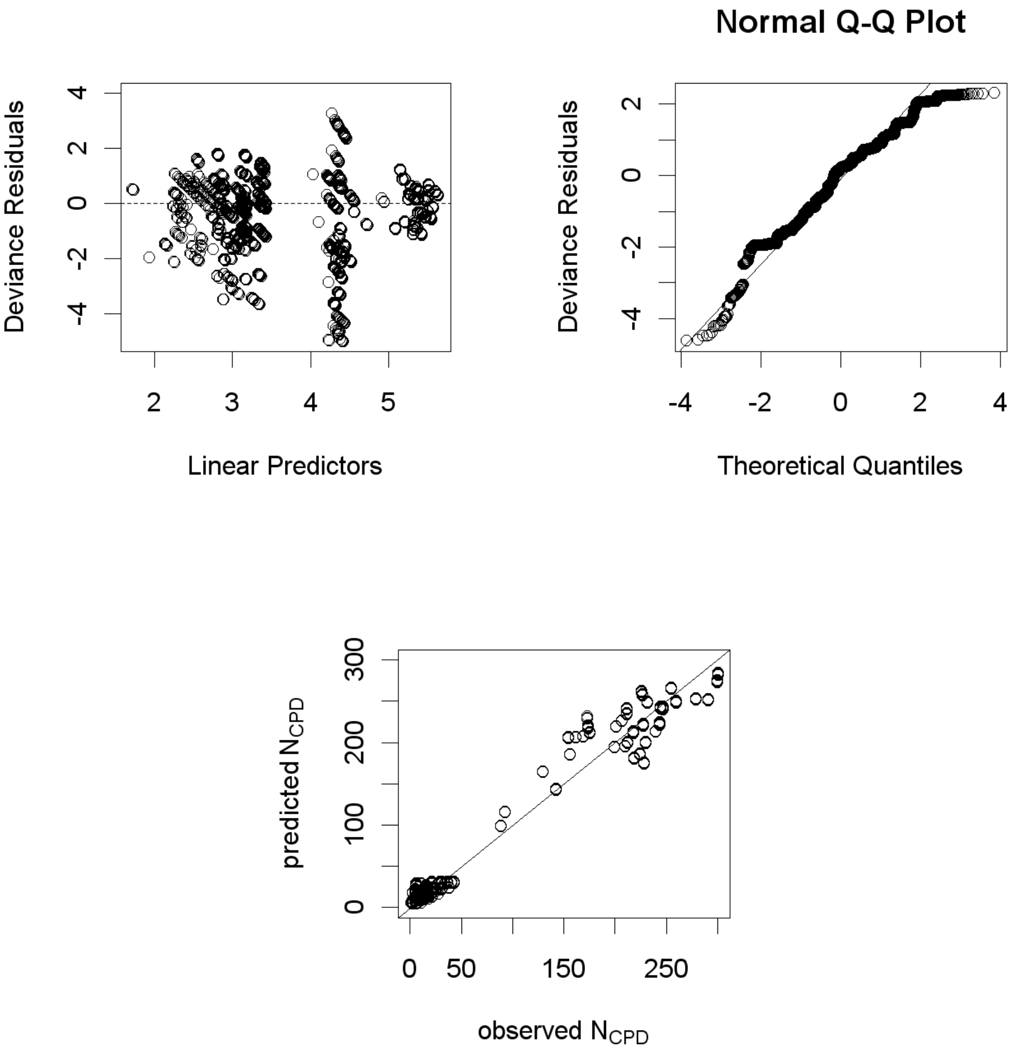

Figure S3. Diagnostic plots for Model D-C3. The description of the plots is as in Figure S2.

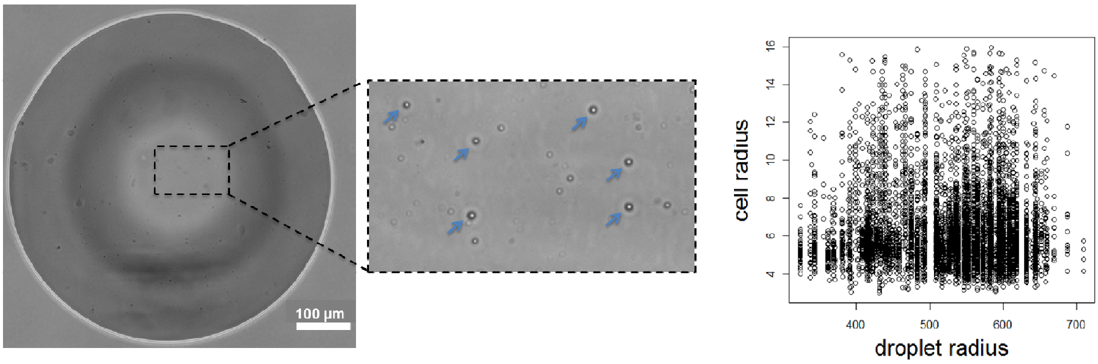

Figure S4. A sample figure for cell pictures taken to measure the cell radii (left). A scatter plot of cell radius versus droplet radius values (right).

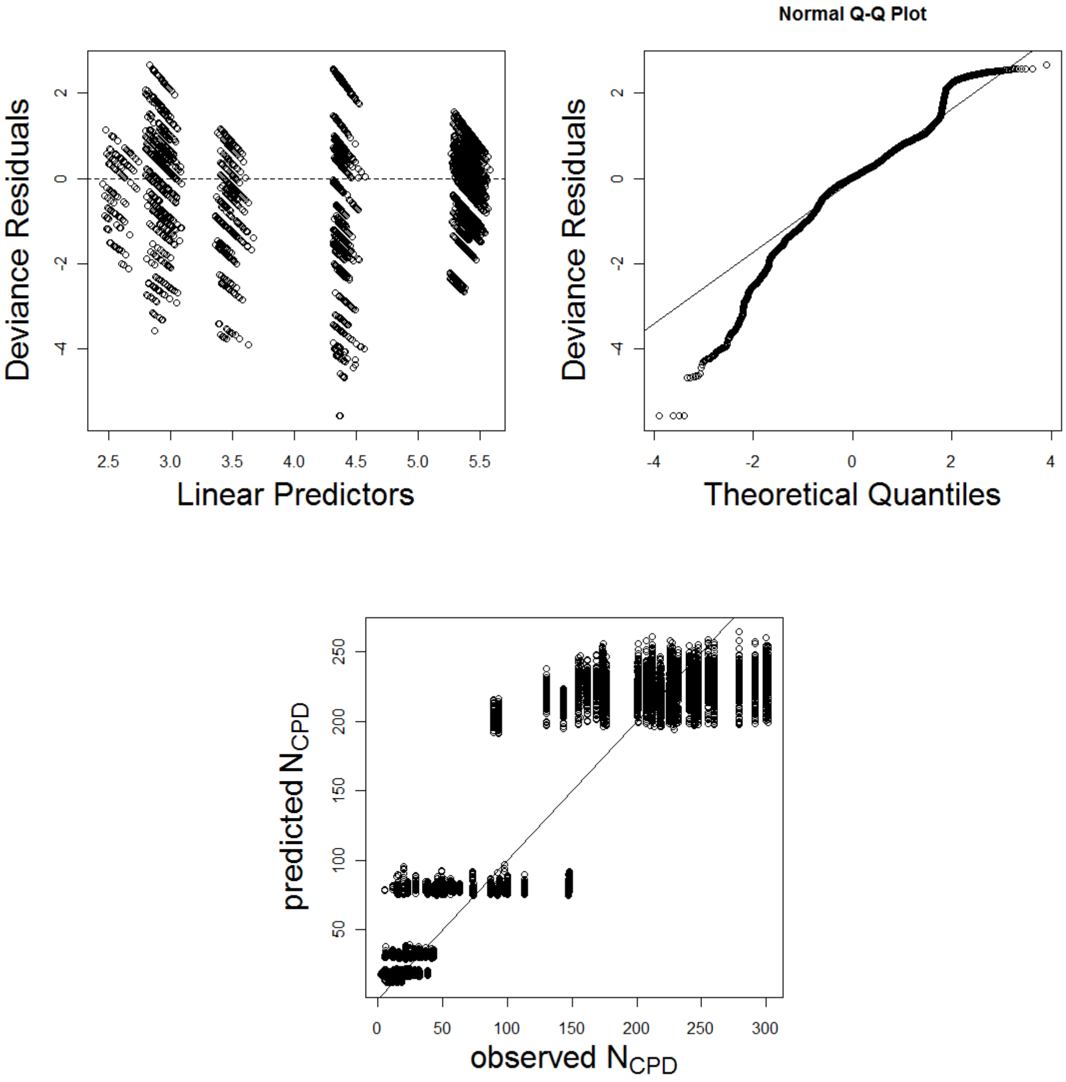

Figure S5. Diagnostic plots for Model R2-C2. The description of the plots is as in Figure S2.

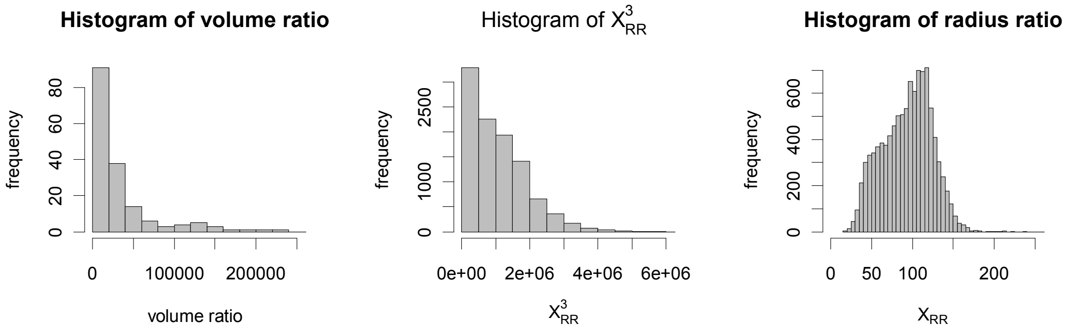

Figure S6. Histograms of the volume ratios (a), cube of radius ratios (b), and radius ratios (c).

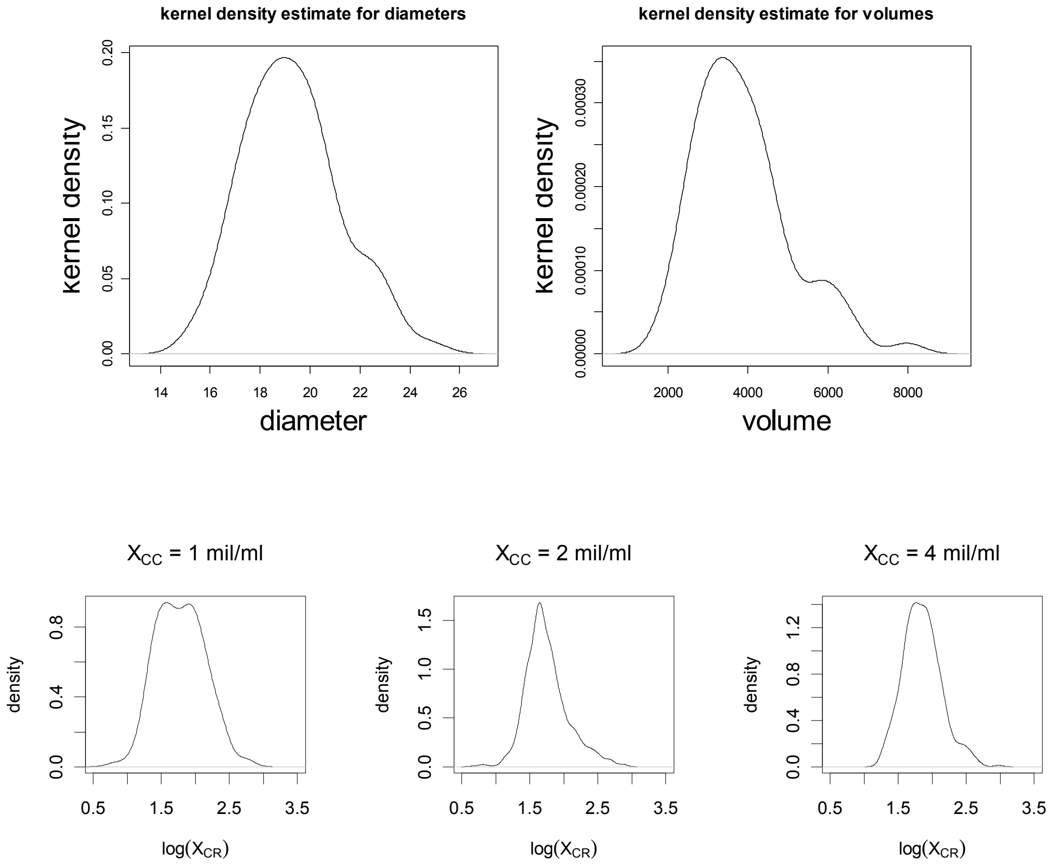

Figure S7. Kernel density estimates for the cell diameters (top left) and volumes (top right) based on the pilot study with 75 cells. Kernel density estimates of the log of the cell radii for all cells at cell concentrations 1 mil/ml, 2 mil/ml, and 4 mil/ml, respectively (bottom row).

References

1. P. McCullagh and J. Nelder, *Generalized Linear Models*, Chapman and Hall, London, 1989.
2. N. E. Breslow, *Statistica Applicata*, 1996, **4**, 23-41.
3. N. E. Breslow, Statistica Applicata, 1996, 8, 23-41.
4. C. B. Dean, J. Amer. Statist. Assoc., 1992, 87, 451-457.
5. U. Demirci and G. Montesano, *Lab Chip*, 2007, **7**, 1139-1145.
6. D. L. Snyder and M. I. Miller, *Random Point Processes in Time and Space*, Springer-Verlag, New York, 1991.
7. H. C. Thode, *Testing for Normality*, Marcel Dekker, New York, 2002.