Lab on a Chip

Dynamic Article Links

Cite this: *Lab Chip*, 2012, **12**, 4884–4893

www.rsc.org/loc

PAPER

Prediction and control of 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] Rami El Assal,[b] Ali Acikgenc,[b] Chung-an Max Wu[b] and Utkan Demirci*[bc]

Received 12th October 2011, Accepted 30th May 2012
DOI: 10.1039/c2lc40523g

Manipulation and encapsulation of cells in microdroplets has found many applications in various fields such as clinical diagnostics, pharmaceutical research, and regenerative medicine. The control over the number of cells in individual droplets is important especially for microfluidic and bioprinting applications. There is a growing need for modeling approaches that enable control over a number of cells within individual droplets. In this study, we developed statistical models based on negative binomial regression to determine the dependence of number of cells per droplet on three main factors: cell concentration in the ejection fluid, droplet size, and cell size. These models were based on experimental data obtained by using a microdroplet generator, where the presented statistical models estimated the number of cells encapsulated in droplets. We also propose a stochastic model for the total volume of cells per droplet. The statistical and stochastic models introduced in this study are adaptable to various cell types and cell encapsulation technologies such as microfluidic and acoustic methods that require reliable control over number of cells per droplet provided that settings and interaction of the variables is similar.

1. Introduction

Microscale droplets (microdroplets) have widespread applications in various areas, such as inkjet printing,[1] colloidal research,[2] biology,[3,4] and medicine.[5] Recently, cell encapsulation in microdroplets has found new fields of application including microfluidics,[6,7] cryobiology,[3,8–11] clinical diagnostics,[12] cell patterning,[3,13–16] tissue engineering,[13,17] high throughput drug studies for cancer,[15] stem cells,[18,19] and pharmaceutical research.[20] These applications require control over the number of cells encapsulated within individual droplets. For example, individual cells can be encapsulated in microscale droplets as single cell bioreactors[13] to rapidly detect concentrations of secreted molecules. However, a stochastic model for predicting the number of cells in microdroplets with the current encapsulation methods has not been developed.

There has been a growing interest in cell encapsulation in nano- and micro-scale droplets for biological and genetic analysis,[21–25] in which the control over the number of cells in a droplet and cell-to-cell distances are essential.[26] On the other hand, in a bottom-up tissue engineering approach, cell-encapsulating hydrogels are used as building blocks, where the number of cells per building block determines the overall cell density in the resulting constructs.[14,27,28] Cell density and cell-to-cell distance are critical, which affect the structural and functional properties of the engineered tissues.[29] These applications all require encapsulation of a few cells in a small volume of fluids or microdroplets with highly controllable density for consistent and repeatable results.

There are currently several cell encapsulation techniques at the microscale, such as pneumatic valve-based bioprinting,[14,30] acoustic technologies,[3,13] inkjet bioprinting,[31,32] laser bioprinting,[33] microfluidic based cell manipulation,[34–37] and encapsulation methods.[38–40] All these techniques aim to manipulate cells in microscale volumes, and control cell density and cell-to-cell distance. However, the variability in number of cells per droplet due to the stochastic nature of cell encapsulation is a barrier for effective use of these techniques.

Previously, the number of cells per droplet and the dependence on the concentration in a suspension and droplet size were reported.[13] The data were fit to a Poisson distribution to estimate the probability of the number of cells per droplet using a

[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. E-mail: udemirci@rics.bwh.harvard.edu*
† Electronic Supplementary Information (ESI) available. See DOI: 10.1039/c2lc40523g
‡ The authors contributed equally to this work.
§ Feng Xu is currently with the Key Laboratory of Biomedical Information Engineering of Ministry of Education, School of Life Science and Technology, Xi'an Jiaotong University, Xi'an 71049, P.R. China and Biomedical Engineering and Biomechanics Center, Xi'an Jiaotong University, Xi'an 710049, P.R. China

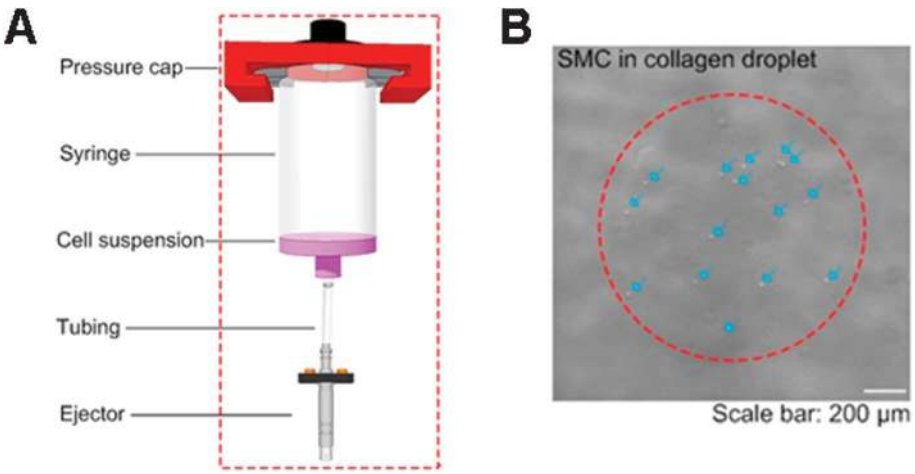

Fig. 1 **(A)** Cell encapsulation and microdroplet generation system. **(B)** A typical printed cell-encapsulating collagen droplet. Arrows represent the smooth muscle cells (SMC) encapsulated in a droplet.

simplified model that is not statistical or stochastic. That is, the model did not determine the statistical relationship between the variables explicitly. In this work, we perform this important aspect of cell encapsulation, where we also include cell radius as a predictor in one of the three models. We developed statistical models to determine the relationship between the number of cells per droplet (denoted N_{CPD} henceforth), and the following parameters: **(i)** cell concentration in the ejection fluid, **(ii)** droplet size, and **(iii)** cell size in terms of radius. The models can also be used to predict and control N_{CPD}. Furthermore, we develop stochastic models for total volume of cells per droplet based on the above statistical models, hence the three parameters considered. To the best of our knowledge this is attempted for the first time for cell encapsulating droplets. We considered the ranges of these parameters as follows: the cell concentration in the ejection fluid (1, 2, 4, 8, and 16 million cells per milliliter (mil ml^{-1})), cell radius (3–16 μm) and droplet radius (300–700 μm). We developed a statistical model of N_{CPD} by negative binomial regression as a function of these parameters using generalized linear modeling techniques appropriate for count data (*i.e.*, data that provide the numbers or counts of particles or units in a fixed region or time) and stochastic modeling of the number and volume of cells as a form of negative binomial process. The novelty here is the statistical modeling of number of cells per droplet (as the response variable), based on the other variables (size of droplets and cell concentration) as predictor variables. The developed models can be used for reliable predictions and to improve the control over cell encapsulation in droplets, offering theoretical and experimental insights into the involved mechanisms.

2. Materials and methods

2.1. Cell-encapsulation in microdroplets

In this study, we used a commercially available droplet ejector to generate cell encapsulating microdroplets (solenoid microvalve ejector, model G100-150300, TechElan, Mountainside, NJ). We have used this ejector to generate droplets encapsulating various cell types (*e.g.*, smooth muscle cells (SMCs), embryonic stem cells, and cancer cells) with high cell viability,[14,15,17,19,30] (Fig. 1). The cell encapsulation systems involve the model formation of a breaking droplet that encapsulates cells within the contents. In this system, N_{CPD} was controlled by changing the droplet radius or cell concentration in the ejection fluid (*i.e.*, cell suspension in a syringe before ejection). In this study, we used data based on smooth muscle cell (SMC) printing. We used primary bladder SMCs from a Sprague Dawley rat. Most cell types are spherical when suspended in solution. We perform ejection when the cells are in suspension. The cell suspension was mixed by manual pipetting before printing and the printing process took less than 1 min, which prevented cell settling in the reservoir. During printing, a droplet was taken from the ejection fluid from the bottom of the syringe *via* tubing that connected the syringe to the ejector (Fig. 1). Images of printed cells in droplets were taken using a bright-field microscope (Nikon TE2000). Number of cells in printed droplets was counted manually in the 4× images obtained. Cell radii were measured, assuming a spherical geometry. Since it was challenging to measure droplet radius in a three-dimensional (3D) spherical shape before landing on the substrate surface, we used the two-dimensional (2D) radius of the droplets on the surface (*i.e.* droplet spread radius), which correlate to the 3D droplet radius. As the cell concentration in the ejection fluid, we used 1, 2, 4, 8, and 16 mil ml^{-1}. We collected data from 178 droplets at five cell concentrations (see Table 1). There are several other factors affecting N_{CPD}, which were fixed in our experiments. These factors are: **(i)** dispensing pressure in ejection reservoir (5 psi), **(ii)** cell type (rat SMCs), **(iii)** fluid viscosity (0.2% collagen), and **(iv)** droplet ejection rate (10 droplets per second). We provided the list of variables and abbreviations used in the article in Table 2, and the ranges of these variables in Table 3.

Table 1 Number of droplets for each cell concentration

Cell concentration (mil ml^{-1})	1	2	4	8	16
Number of droplets	23	55	30	30	37

Table 2 Notation (variable names and abbreviations) and its description for statistical modeling of number of cells per droplet

Notation	Description
N_{CPD}	Number of cells per droplet
X_{DR}	Droplet radius (μm)
X_{CR}	Cell radius (μm)
X_{CC}	Cell concentration (million cells per milliliter)
X_{RR}	Radius ratio
GLM	Generalized linear models
V_T	Total volume of the cells in a droplet (μm^3)
V_i	Volume of the cell i in a droplet (μm^3)
α	Level of significance for hypothesis testing
λ	Poisson rate parameter
μ and σ^2	Mean and variance
n	Number of observations
Q_1, Q_3	First and third quartile values

2.2. Stochastic and statistical modeling of number of cells per droplet

We hypothesized that the number of cells per droplet (N_{CPD}) highly depends on the droplet radius, cell radius, and cell concentration in the suspension. To test this hypothesis, we developed mathematical models to understand the stochastic processes for N_{CPD} and the total cell volume (per droplet). We also assessed empirically these models by fitting them to the experimental data. For count data, usually the relationship between the mean and the variance is determined as $\mathrm{Var}(Y_i) = \tau\mu_i$, where Y_i is the count variable with mean μ_i and τ is the dispersion parameter. Depending on the values of τ, two sets of models are used. If τ equals one (*i.e.,* not significantly different than one), a Poisson regression model (a generalized linear model (GLM) model) with logarithm function as the canonical link function and Poisson distributed errors[41] was fit to the data. When τ is significantly different than one, other GLMs such as the negative binomial model are more appropriate.[42] Our data are consistent with the underlying assumptions for a GLM model.

In the models, we used N_{CPD} as the response (or dependent) variable and the other variables (see Table 3) as the predictor (or independent) variables in the GLM procedures. We applied a model selection procedure to obtain a concise and descriptive model (with the least number of variables possible, but with a high explanatory power). We started with a model containing all the variables (called ''full model'') with some non-linear terms that were added to reflect the significant relationship between N_{CPD} and the predictor variables. The full model is then reduced using a stepwise backward elimination procedure together with Akaike Information Criteria (AIC),[43] *i.e.* some insignificant variables were removed until each of the remaining variables has a significant effect on the N_{CPD} at $\alpha = 0.05$ level.

The underlying assumptions, model selection procedure, and some of the discussion on the model diagnostics for each model that we consider are detailed in the ESI† for brevity in presentation; as they are also peripheral for the main message and results of this article.

Table 3 Variables used in this study, their values and ranges

Variable	Values and ranges
Cell concentration	1, 2, 4, 8, and 16 million cells per ml
Cell radius	3–16 μm
Droplet radius	300–700 μm

3. Results and discussion

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

The summary statistics (such as mean, median and first quartile) of the variables of droplet radius, cell concentration, and N_{CPD} are summarized in Table 4 and the corresponding histograms are plotted in Fig. 2. The histograms indicate a mild leftward skew for droplet radii and severe rightward skew for N_{CPD} whose mean, 63.63, is much larger than its median, 21, while the rightward skew is reduced for $\log(N_{CPD})$. In particular, the standard deviations for N_{CPD}, droplet radius, cell concentration, and cell radius are 81.87, 91.99, 5.52, and 2.05 (Table 4). So among these variables, the variation of cell radius is much smaller compared to those of N_{CPD}, cell concentration and droplet radius in our setup. Here one might be misled by comparing the ranges (maximum minus minimum) of these variables which is not a robust measure of spread. Hence, we first model N_{CPD} as a function of only cell concentration (X_{CC}) and droplet radius (X_{DR}) without considering the influence of cell radius (X_{CR}). Our experimental data (Table 4) show that the variance of N_{CPD} is significantly larger than its mean: $\mathrm{Var}(N_{CPD}) = 6702.70$ and $\mathrm{Mean}(N_{CPD}) = 63.63$ with $p < 0.0001$ based on Dean's P_B test for overdispersion[44]). This indicates that negative binomial regression is more appropriate for our data compared to the more common Poisson regression.

We start with the negative binomial GLM which models logarithm of N_{CPD} as a function of droplet radius and cell concentration and obtain the following model:

$$\log(E(N_{CPD})) = -11.0022 - 0.0247 \times X_{DR} + 0.1890 \times X_{CC} + 1.1528 \times \sqrt{X_{DR}} \quad (1)$$

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

Model D-C2:

$$E(N_{CPD}) = \left(1.6663 \times 10^{-5}\right) \times 0.9756^{X_{DR}} \times 1.2081^{X_{CC}} \times 3.1672^{\sqrt{X_{DR}}} \quad (2)$$

Observe that the expected value of N_{CPD} increases as the droplet radius or cell concentration increases. For example, the expected $\log(N_{CPD})$ increase is 0.1890 for a one-unit increase in

Table 4 Summary statistics of the variables: droplet radius, cell concentration, N_{CPD} for models that ignore the cell radius (top three rows) and the cell radius (bottom row). The abbreviations are as in Table 2

	n	mean	SD	min	Q_1	median	Q_3	max
N_{CPD}	178	63.63	81.87	2.00	13.00	21.00	83.750	301.00
X_{DR}	178	508.70	91.99	318.52	428.10	516.7	583.30	709.3
X_{CC}	178	6.23	5.52	1.00	2.00	4.00	8.00	16.00
X_{CR}	10 150	6.24	2.05	3.00	4.99	5.61	7.00	15.96

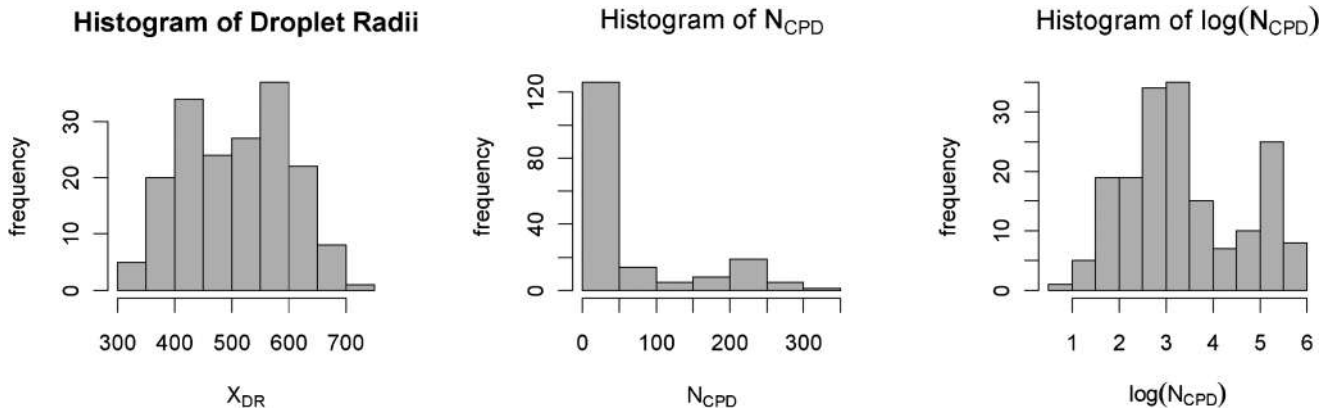

Fig. 2 Histograms of droplet radii (left), N_{CPD} values (middle) and logarithm of N_{CPD} values (right).

cell concentration (*i.e.,* if the cell concentration increases by 1 mil ml^{-1}). 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 effects of the cell concentration and droplet radius are both strong in estimating N_{CPD}, but that of the droplet radius is much greater.

Based on the diagnostic plots in Fig. 3, we observe that model assumptions are satisfied for Model D-C2. Hence, when the cell radius is fixed or its variation is negligible compared to that of the other variables (*i.e.,* when the variance of the cell radius is much smaller compared to those of other variables), Model D-C2 can be used to estimate the N_{CPD} values for a given droplet radius and a cell concentration (within the variable ranges given in Table 3). For example, with a droplet radius of 500 μm and cell concentration of 5 mil ml^{-1}, we estimate the expected N_{CPD} to be

$$E(N_{CPD}) = \left(1.6663 \times 10^{-5}\right) \times 0.9756^{500} \times 1.2081^{5} \times 3.1672^{\sqrt{500}} = 29$$

The notation of the models is summarized in Table 5.

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

Unlike Model D-C2 (Table 5), at this stage of analysis, we consider the cell radius (X_{CR}) as a potentially important factor in explaining or modeling the N_{CPD} by incorporating cell radius into the modeling procedure. That is, the response variable of interest (N_{CPD}) is modeled as a function of independent (predictor) variables, *i.e.* cell concentration (X_{CC}), droplet radius (X_{DR}), and cell radius (X_{CR}). 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 9539 sets of X_{DR}, X_{CR}, N_{CPD} and X_{CC} values from 148 droplets at five cell concentrations. 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. This indicated that negative binomial regression is more appropriate.

We implement the negative binomial GLM that models the logarithm of N_{CPD} as a function of droplet radius, cell radius, and cell concentration together with non-linear terms. By our model selection procedure, the model is reduced to one that only contains X_{DR} and X_{CC} as predictors. That is, in the presence of droplet radius and cell concentration, cell radius has no significant contribution to modeling N_{CPD}. However, this does not necessarily mean that X_{CR} has no impact in the modeling of N_{CPD}. In particular, if the cell radius is used as the only predictor variable in modeling the response variable N_{CPD}, then it is significant. We construct models at each cell concentration value treating cell concentration as a qualitative factor (*i.e.,* **Model D-C3**). This is justifiable, because in practice, usually an experimenter takes X_{CC} to be any one of the 5 values specified. When many replications are taken at few levels of a numerical variable, it is a common practice to also treat this numerical variable as a categorical variable which sometimes provides a better fit of the model to the data at hand. With such modeling, we observe that the cell radius is significant at some cell concentration levels (1 and 8 mil ml^{-1}), but not at other levels (2, 4, 16 mil ml^{-1}). For example, for X_{CC} = 1 mil ml^{-1}, we have

$$\log\left(E(N_{CPD})\right) = 22.2990 + 0.0366X_{DR} - 1.5999\sqrt{X_{DR}} - 1.7804\sqrt{X_{CR}} + 0.3071X_{CR} \quad (3)$$

When these coefficients are translated into multiplicative effects in the predicted N_{CPD} count, we get

$$E(N_{CPD}) = 4.8345 \times 10^{9} \times 1.0372^{X_{DR}} \times 0.2019^{\sqrt{X_{DR}}} \times 0.1686^{\sqrt{X_{CR}}} \times 1.3594^{X_{CR}} \quad (4)$$

See Table 6 for the explicit forms of **Model D-C3** for each cell concentration. Notice that the dependence of N_{CPD} on X_{CR} and X_{CD} is different at each X_{CC}. Observe that N_{CPD} increases with increasing droplet radius, while N_{CPD} tends to decrease with increasing cell radius. For example, at X_{CC} = 1 mil ml^{-1}, for a one-unit increase in cell radius from, say 5 to 6 μm (*i.e.,* if cell radius value increases by 1 μm at 5 μm), the decrease in log of expected N_{CPD} is 0.0729 or the expected N_{CPD} decrease is by a factor of exp(−0.0729) = 0.9297, when X_{DR} is held constant.

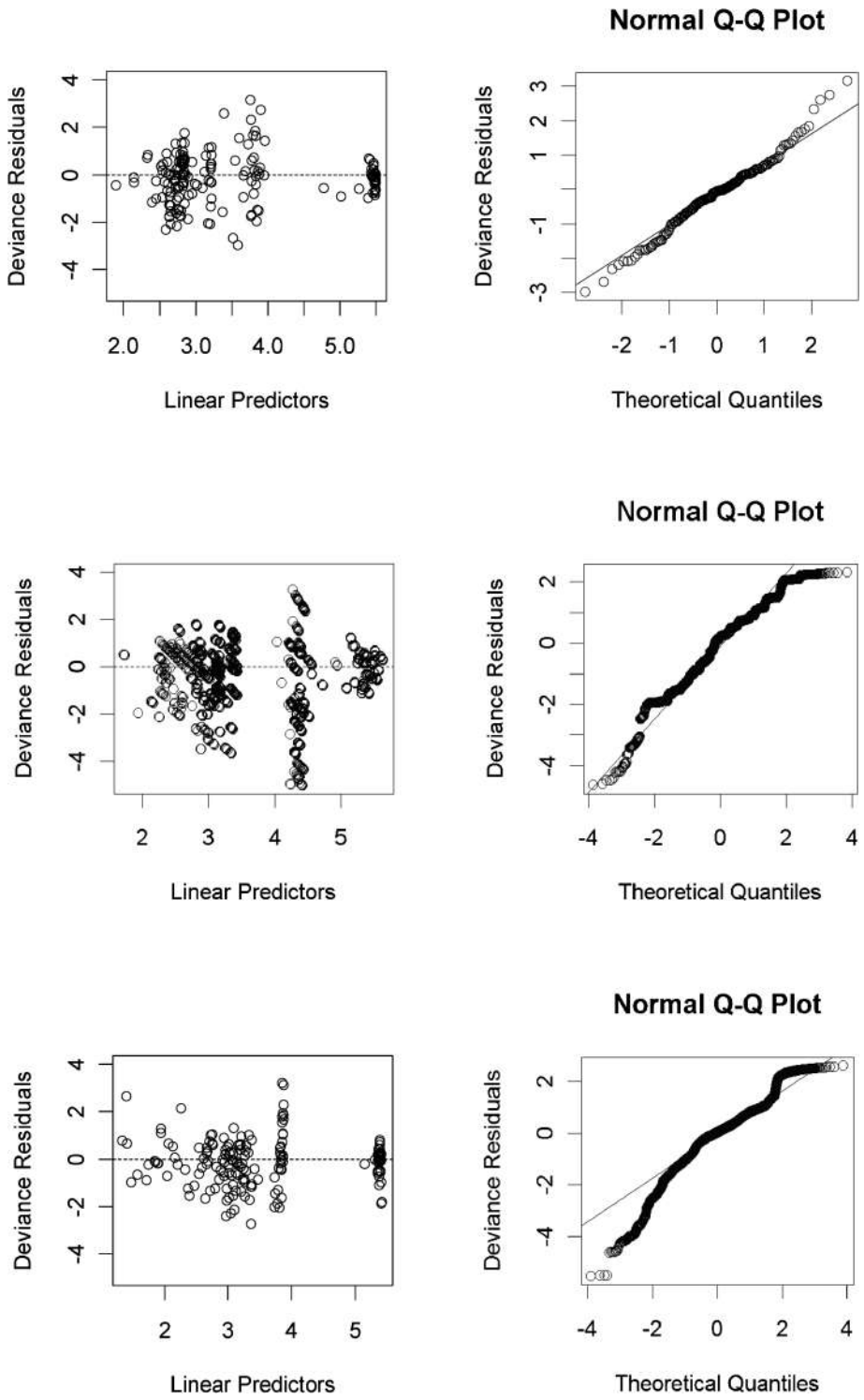

Fig. 3 Diagnostic plots for Model D-C2 (top row), Model D-C3 (middle row) and Model R2-C2 (bottom row). The deviance residuals *versus* predicted values (left) and the normal *QQ*-plot for deviance residuals *versus* theoretical quantiles, where the straight line passes through the first and third quartiles (right).

Furthermore, the droplet radius has a stronger influence on N_{CPD} compared to the cell radius.

Based on the diagnostic plots presented in Fig. 3, we observe that model assumptions are valid in this case. Hence, when the cell radius is considered, **Model D-C3** is a good alternative to estimate the N_{CPD} values for a given droplet radius and cell radius, at the cell concentrations tested in this study (1, 2, 4, 8, and 16 mil ml^{-1}). That is, if one wants to use these cell concentration values in a cell encapsulation experiment, Model D-C3 can be employed. For example, for a droplet radius of 500 μm, a cell concentration of 1 mil ml^{-1} and a cell radius of 15 μm, we estimate the expected N_{CPD} to be

Table 5 Abbreviations used for the notation of models are indicated by the bold face and capitalization in the description column

Models[a]	Description
Model D-C2	Modeling N_{CPD} as a function of **D**roplet radius and **C**ell **C**oncentration
Model D-C3	Modeling N_{CPD} as a function of **D**roplet radius, **C**ell **C**oncentration, and **C**ell radius
Model R2-C2	Modeling N_{CPD} as a function of the ratio of droplet **R**adius to cell **R**adius and **C**ell **C**oncentration

[a] The MS Excel and R code of the three statistical models (*i.e.* D-C2, C-D3, and R2-C2) are provided as a supplement.†

$$E(N_{CPD}) = 4.8345 \times 10^9 \times 1.0372^{500} \times 0.2019^{\sqrt{500}} \times 0.1686^{\sqrt{15}} \times 1.3594^{15} = 12$$

On the other hand, Model D-C2 is also applicable to any cell concentration value within 1–16 mil ml^{-1} range. However, for concentration values other than 1, 2, 4, 8, and 16 mil ml^{-1}, one can also estimate N_{CPD} values with linear interpolation. For example, at X_{DR} = 500 μm and X_{CR} = 10 μm, Model D-C3 estimates N_{CPD} value to be 22 for X_{CC} = 4 mil ml^{-1} and 105 for X_{CC} = 8 mil/ml. Then, at the same X_{DR} = 500 μm and X_{CR} = 10 μm values, for X_{CC} = 5 mil ml^{-1}, by linear interpolation, we obtain $N_{CPD} \approx 22 + \left(\frac{5-4}{8-4}\right)(105-22) = 42.75$.

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

We modelled N_{CPD} as a function of X_{CC} and ratio of droplet radius to cell radius for each cell, called the radius ratio and denoted X_{RR}. Negative binomial regression is more appropriate here, since Var(N_{CPD}) = 6702.70 is significantly larger than the mean: Mean(N_{CPD}) = 63.63, $p < 0.0001$ based on Dean's P_B test for overdispersion. We used cell droplet data on 171 droplets and 10,226 cells.

We implemented the negative binomial GLM with X_{RR} and X_{CC} as predictors and obtain the following reduced model

$$\log(E(N_{CPD})) = 2.0485 + 0.0021 \times X_{RR} + 0.3546 \times X_{CC} - 0.0098 \times X_{CC}^2 \quad (5)$$

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

Model R2-C2:

$$E(N_{CPD}) = 7.7564 \times 1.0021^{X_{RR}} \times 1.4256^{X_{CC}} \times 0.9902^{X_{CC}^2} \quad (6)$$

Note that N_{CPD} tends to increase as X_{CC} or X_{RR} increases. That is, when the cell concentration increases, it is more likely to have more cells per droplet. Similarly, when the ratio of droplet radius to cell radius increases, the droplet volume tends to be much larger than the cell volumes, so it is more likely to encapsulate more cells in such droplets. For example, the expected log(N_{CPD}) increase is 0.0021 for a one-unit increase in radius ratio (*i.e.,* if the radius ratio increases by 1). That is, a one-unit increase in radius ratio causes the expected N_{CPD} to increase by a factor of exp(0.0021) = 1.0021, holding X_{CC} constant. 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 that of other variables), Model R2-C2 can be used to estimate the N_{CPD} values for a given radius ratio and a cell concentration within the range of the variables. The ranges for the droplet radii and cell radii in Table 3 yield the range for X_{RR} to be 18.75–233.33. For example, with the radius ratio being 500 μm/10 μm = 50 and the cell concentration being 1 mil ml^{-1}, we estimate the expected N_{CPD} to be

$$E(N_{CPD}) = 7.7564 \times 1.0021^{50} \times 1.4256^{1} \times 0.9902^{1} = 12.$$

Furthermore, the model diagnostic plots in Fig. 3 suggest that although the model assumptions seem to be not severely violated, the quantile-quantile (*QQ*)-plot suggests more severe non-normality compared to other models. Besides, the plot of the deviances indicates a worse fit compared to other models (see Fig. 3).

3.4. Comparison and discussion of the models for N_{CPD}

The models D-C2, D-C3, and R2-C2 have AIC values of 105,376.6, 100,407.1, and 103,697.8 respectively. Hence D-C3 with the smallest AIC value provides the best fit to the available data. Further, comparing the above three models, we find that the inclusion of the cell radius and treating cell concentration as a categorical variable in Model D-C3 provides a significant improvement over Model D-C2 (likelihood ratio χ^2 = 5001.4, df = 16, $p < 0.0001$). However, the effect of the cell radius is not as strong as the other variables in the modeling of N_{CPD}. For example, for a five-fold increase in cell radius, which is roughly the ratio of the largest cell radius to the smallest cell radius in our data, the expected value of N_{CPD} decreases by a factor of 1.0470

Table 6 Model D-C3: The negative binomial model for N_{CPD} at each cell concentration.

Cell concentration (mil ml^{-1})	The model
X_{CC} = 1	$E(N_{CPD}) = 4.8345 \times 10^9 \times 1.0372^{X_{DR}} \times 0.2019^{\sqrt{X_{DR}}} \times 0.1686^{\sqrt{X_{CR}}} \times 1.3594^{X_{CR}}$
X_{CC} = 2	$E(N_{CPD}) = (8.2390 \times 10^{-23}) \times 0.9081^{X_{DR}} \times 96.8620^{\sqrt{X_{DR}}} \times 0.8627^{\sqrt{X_{CR}}}$
X_{CC} = 4	$E(N_{CPD}) = (3.8970 \times 10^{-9}) \times 0.9588^{X_{DR}} \times 6.8313^{\sqrt{X_{DR}}} \times 0.1740^{\sqrt{X_{CR}}}$
X_{CC} = 8	$E(N_{CPD}) = (2.7180 \times 10^{2}) \times 1.0070^{X_{DR}} \times 0.8088^{\sqrt{X_{DR}}} \times 1.1030^{\sqrt{X_{CR}}}$
X_{CC} = 16	$E(N_{CPD}) = 1.4094 \times 0.9967^{X_{DR}} \times 1.3299^{\sqrt{X_{DR}}} \times 1.1143^{\sqrt{X_{CR}}}$

at cell concentration X_{CC} = 1 mil ml^{-1} and increases by a factor of 1.2489 at cell concentration X_{CC} = 8 mil ml^{-1}. Therefore, we can conclude that the influence of cell radius is statistically significant in modeling N_{CPD}, but its practical significance is only moderate. Hence, for practical purposes, Model D-C2 is better along the lines of principle of parsimony, *i.e.,* simple yet explanatory for estimating N_{CPD}.

Comparing Model R2-C2 with Model D-C2, we observe that using the radius ratio instead of droplet radius does not significantly improve the model performance in the sense that the fit of Model R2-C2 is not better than that of Model D-C2. In fact, Model D-C2 is better in explaining the variation in N_{CPD} compared to Model R2-C2 (the likelihood ratio χ^2 = 1678.8, df = 0, $p < 0.0001$).

On the other hand, comparing Model R2-C2 to Model D-C3, we see that Model D-C3 is significantly better in explaining the variation in N_{CPD} compared to Model R2-C2 (the likelihood ratio χ^2 = 3322.6, df = 16, $p < 0.0001$). That is, the raw radius values for droplets and cells are better for explaining the variation in N_{CPD} compared to the radius ratios. Therefore, if the cell radius is fixed or its variation is negligible, Model D-C2 can be applied; otherwise Model D-C3 should be applied. These models explain the encapsulation process that determines the number of cells per droplet. Further, we can perform predictions to control the conditions that will yield designed cell encapsulation performance (*i.e.,* N_{CPD} with high probability).

Furthermore, if the cell concentration and droplet radius are fixed, the dependence of N_{CPD} on cell radius can be determined more precisely. The cell radii can be measured by imaging cells in suspension. It takes time for the cells to attach to a surface and spread after ejection. We eject the cells in suspension form; hence, their spread sizes on the surface do not come into play. Some cells may seem larger in cultures, since they spread, however, their sizes range within the tens of microns when they are suspended and become spherical. The cell concentration can be precisely controlled, and the cell radius is dependent on the cell type. The droplet radius is measured after the droplet lands on the substrate. However, this does not mean that we cannot control the droplet radius. The droplet size is controlled by the valve-opening duration as described in the methods. The longer the microvalve stays open the larger the droplet that is ejected. Hence, to make use of the presented models in this work in estimation or prediction, we also determine the conditions of the experimental settings to achieve specific droplet radius values on the substrate. Therefore, for a given cell type, models can determine the required cell concentrations and droplet radius to achieve a predetermined number of cells per droplet with a high probability. Additionally, for given cell concentration and droplet radius values, models can estimate the expected number of cells per droplet with a high probability.

See also Table 7 for estimated values of N_{CPD} for X_{DR} = 500 μm, and X_{CR} = 15 μm for each of 1, 2, 4, 8, and 16 mil ml^{-1} X_{CC} values. For small X_{CC} values (*i.e.,* for X_{CC} =1 and 2 mil ml^{-1}), the models agree with the prediction of N_{CPD} values. However, for larger X_{CC} values, Model D-C3 is more reliable. For higher concentration values Models D-C2 and R2-C2 seem to be over-averaging, and, hence, underestimating the N_{CPD} values. For concentration values other than 1, 2, 4, 8, and 16 mil ml^{-1}, one can perform a linear interpolation based on Model D-C3 as described at the end of Section 3.2.

Table 7 Estimated N_{CPD} values based on the models D-C2, D-C3, and R2-C2 for X_{DR} = 500 μm and X_{CR} = 15 μm

Cell concentration (million cells ml^{-1})	1	2	4	8	16
D-C2	14	16	24	51	233
D-C3	12	14	24	113	241
R2-C2	12	16	29	76	195

4. Stochastic modeling of number and volume of cells per droplet

4.1. N_{CPD} modeled as a negative binomial process

In Sections 3.1–3.3, we have shown that the volume of the droplet, *i.e.* droplet radius and cell concentration are the main factors to determine the N_{CPD} values. For a 3D region R with a certain volume in the ejection fluid, the number of cells in R denoted N(R), can be modeled as a negative binomial process with the following probability distribution function (pdf):

$$P(N(\mathrm{R})=k)=\frac{\Gamma(k+r)}{k!\Gamma(r)}\left(1-\frac{\lambda V(\mathrm{R})}{\lambda V(\mathrm{R})+r}\right)^{r}\left(\frac{\lambda V(\mathrm{R})}{\lambda V(\mathrm{R})+r}\right)^{k} \quad (7)$$

for k = 0, 1, 2,... and where λ is the rate parameter with its unit chosen to be mil ml^{-1} (so $\lambda = X_{CC}$), and V(R) is the volume of the region R in ml, and $r=\frac{\lambda V(\mathrm{R})}{\tau-1}$ Var(N(R)) = τ × Mean(N(R)). Since the ejection fluid is assumed to be homogenized and droplets are taken from the fluid so that a droplet represents a region with volume V(D). In particular, the number of cells per droplet, N_{CPD}, has the distribution as in eqn (7) with V(R) being replaced by V(D).

4.2. Total volume of cells per droplet modeled as a compound spatial inhomogeneous negative binomial process

In the cell encapsulation procedure, the cell radius may vary in a range even for a given cell type. Furthermore, N_{CPD} is also related with the cell radius (or volumes). Therefore, a more complex model that incorporates the randomness in the cell radius in addition to the negative binomial property of the number of cells is the compound homogeneous negative binomial process. Given a droplet, let V_T(D) be the total volume of the cells at a droplet and V_i is the volume of cell i in the droplet for i = 1, 2,…,N(D) where N(D) = N_{CPD} in the droplet. Then

$$V_{\mathrm{T}}(\mathrm{D})=\sum_{i=1}^{N(\mathrm{D})} V_i \quad (8)$$

where N(D) is the homogeneous negative binomial process described in Section 3.5.1. In particular, using eqn (8), we get with Model D-C2:

$$V_{\mathrm{T}}(X_{\mathrm{DR}},X_{\mathrm{CC}})=\sum_{i=1}^{N(X_{\mathrm{DR}},X_{\mathrm{CC}})} V_i \quad (9)$$

with Model D-C3:

$$V_{\mathrm{T}}(X_{\mathrm{DR}},X_{\mathrm{CC}},X_{\mathrm{CR}})=\sum_{i=1}^{N(X_{\mathrm{DR}},X_{\mathrm{CC}},X_{\mathrm{CR}})} V_i \quad (10)$$

with Model R2-C2:

$$V_{\mathrm{T}}(X_{\mathrm{RR}},X_{\mathrm{CC}})=\sum_{i=1}^{N(X_{\mathrm{RR}},X_{\mathrm{CC}})} V_i \quad (11)$$

In the model in eqn (10), N(D) directly depends on the cell radius, so it is a dependent-type compound negative binomial

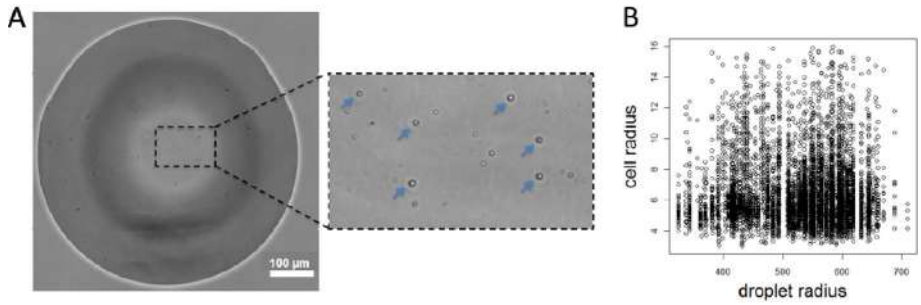

Fig. 4 Cells in a typical printed droplet and quantification of their radii. **(A)** An image of cells encapsulated in a printed droplet. Arrowheads in the inset indicate some of the cells encapsulated in a droplet. **(B)** A scatter plot of cell radius versus droplet radius values.

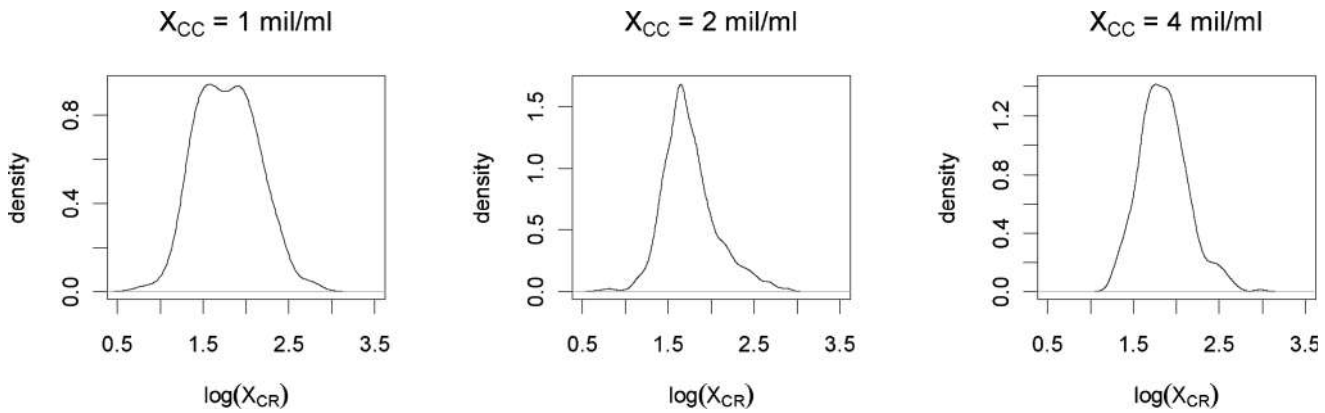

Fig. 5 Kernel density estimates of the log of the cell radii for cell concentrations 1 mil ml^{-1}, 2 mil ml^{-1}, and 4 mil ml^{-1}.

process. In the model of eqn (9), $N(\mathrm{D})$ only depends on the droplet radius and the cell concentration. There is a positive relationship with a small slope between the cell radius and the droplet radius (see Fig. 4 (right)). So, the model in eqn (9) can be assumed to be a compound negative binomial process.

What remains is the distribution of the volume, V_i, of the cells that can be determined by measuring the cell radii and assuming a spherical cell geometry, *i.e.,* $V = \frac{4}{3}\pi X_{\mathrm{CR}}^3$. Hence, it suffices to determine the distribution of the cell radii, X_{CR}. 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 Fig. 5. 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. Therefore, the cell radii (pooled together in the aggregate data) as a mixture of log-normal distributions has the pdf $f_x(x) = \sum_{i=1}^{n} a_i f_{Y_i}(x)$ where $n = 5$, i stands for cell concentration 2^{i-1} mil ml^{-1} for $i = 1, 2,\ldots,5$, and a_i is the proportion of cells from concentration i, and $f_{Y_i}(x)$ is the pdf of cell radii at concentration i.

Table 8 Mean and standard deviation values for the mixed log-normal distribution in Section 4.2

Cell concentration (million cells ml^{-1})	1	2	4	8	16
μ_i	6.29	6.07	6.65	5.95	5.83
σ_i	2.38	2.22	2.05	1.69	1.93

That is, $\log(Y_i) \sim N(\mu_i, \sigma_i)$ (*i.e.* $\log(\mathrm{Y}_i)$ is distributed as normal distribution with mean μ_i and standard deviation σ_i). In particular, we presented the distribution parameters for each concentration in Table 8.

5. Conclusions

Our statistical models in Section 3 estimate the relationship between number and volume of cells per droplet (N_{CPD}) and other variables including cell concentration (X_{CC}), droplet radius (X_{DR}), and cell radius (X_{CR}) using negative binomial regression. Considering the nature of the relationships and the structure of the statistical models, we conclude that the influential factors that affect N_{CPD} are the cell concentration, droplet radius, and cell radius, with the former two having greater influence. Furthermore, if the cell concentration is fixed, more subtle relationships are observed between N_{CPD} values *versus* droplet and cell radii (see Model D-C3). On the other hand, our stochastic models in Section 4 incorporate the statistical models in Section 3 to describe the total volume of cells per droplet as a compound spatial inhomogeneous negative binomial process.

In conclusion, we have developed three statistical models, namely, Models D-C2, D-C3, and R2-C2. The models are more appropriate under different conditions (cell concentration, droplet radius, and cell radius) to optimize N_{CPD}, *i.e.,* estimate the optimal conditions to encapsulate a desired number of cells within a nanoliter droplet volume. For example, if one wants to estimate N_{CPD} at the specific cell concentration values in Table 3, Model D-C3 is the most appropriate choice, while if one wants

to estimate for any cell concentration value within a vicinity of 1–3 mil ml^{-1} (*i.e.*, around small cell concentration values relative to the ones considered), Model D-C2 could be employed. Considering Models D-C2, D-C3, and R2-C2 (Table 5) based on our experimental data, we conclude that each of the three variables (*e.g.*, cell concentration, droplet radius, and cell radius) can be estimated for a specific goal when the other two are given. For example, for a given droplet and cell radii, the cell concentration can be estimated to achieve a specific N_{CPD} value. Thus, one can design the conditions for a desired N_{CPD}, based on the models introduced here; or in a given setup, one can estimate the number of cells per droplet reliably. In particular, at the specific cell concentration values in Table 3, one might employ Model D-C3 for estimation or prediction purposes. On the other hand, for smaller cell concentration values (*i.e.,* between 1–3 mil ml^{-1}), one might employ Model D-C2 as well. For larger cell concentration values (*i.e.*, 3–16 mil ml^{-1}), we recommend the linear interpolation based on Model D-C3 (see end of Section 3.2). Model R2-C2 is fit mostly for comparative purposes, and found to perform less efficiently than the other two models in the sense that the goodness of fit for the other models are better than Model R2-C2.

The models introduced in this paper are applicable to different cell types, other encapsulation media and cell encapsulation technologies, which require a reliable control over the number of cells per microdroplet. The models are usable when the experimental setup is replicated in the current form. Additionally, the models developed in this study and steps taken to validate the experimental and modeling results are applicable broadly to other cell encapsulation systems, since similar parameters such as cell concentration and droplet size that are analyzed here apply. For example, if the setup, *e.g.* dispenser does not seriously confound the relationship between number of cells per droplet and the other variables, then the models are applicable in that setting as well. Otherwise, the models are instructive in forming models of dispenser families that affect the droplet formation or number of cells per droplet that are substantially different than our setup. In addition, we expect viscosity to affect the overall system when ejecting different solutions. Since the droplet is generated under constant pressure, the viscosity will affect the droplet size with keeping all the other conditions the same. However, we do not expect an effect of viscosity on the number of cells per droplet if the droplet sizes are the same. The statistical and stochastic models introduced in this study are adaptable to various cell types and cell encapsulation technologies such as microfluidic and acoustic methods that require reliable control over the number of cells per droplet provided that setting or interaction of the variables is similar to ours. Here, by adaptability we mean that certain parameters are common to all cell encapsulation systems, *e.g.*, cell concentration and droplet size. A few restrictions of the model are provided in Section S1 of the ESI † as the underlying assumptions for the models. Hence, the models developed in this study can be used to provide reliable predictions and to improve the control over cell encapsulation in droplets for a wide range of applications in biomedicine and biomedical research.

Acknowledgements

This work was performed at the Demirci Bio-Acoustic MEMS in Medicine (BAMM) Labs at the HST-BWH Center for Bioengineering, Harvard Medical School. This work was supported by the W. H. Coulter Foundation Young Investigator Award, the Center for Integration of Medicine and Innovative Technology under U.S. Army Medical Research Acquisition Activity Cooperative Agreements DAMD17-02-2-0006, W81XWH-07-2-0011, W81XWH-09-2-0001, NIH R21-AI087107, and NIH R21-HL095960. In addition, this research is made possible by a research grant that was awarded and administered by the U.S. Army Medical Research & Materiel Command (USAMRMC) and the Telemedicine & Advanced Technology Research Center (TATRC), at Fort Detrick, MD. The information contained herein does not necessarily reflect the position or policy of the Government, and no official endorsement should be inferred. This material is based in part upon work supported by the National Science Foundation under NSF CAREER Award Number 1150733. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. We would like to thank Vition Mbrica for contributing in this study as a high school student under the Student Success Job Program at Brigham and Women's Hospital.

References

1 H. J. Jeong, W. R. Hwang, C. Kim and S. J. Kim, *J. Mater. Process. Technol.*, 2010, **210**, 297–305.
2 R. D. Deegan, O. Bakajin, T. F. Dupont, G. Huber, S. R. Nagel and T. A. Witten, *Nature*, 1997, **389**, 827–829.
3 U. Demirci and G. Montesano, *Lab Chip*, 2007, **7**, 1428–1433.
4 M. E. Leunissen, A. van Blaaderen, A. D. Hollingsworth, M. T. Sullivan and P. M. Chaikin, *Proc. Natl. Acad. Sci. U. S. A.*, 2007, **104**, 2585–2590.
5 A. V. Reis, M. R. Guilherme, L. H. C. Mattoso, A. F. Rubira, E. B. Tambourgi and E. C. Muniz, *Pharm. Res.*, 2009, **26**, 438–444.
6 S. Koster, F. E. Angile, H. Duan, J. J. Agresti, A. Wintner, C. Schmitz, A. C. Rowat, C. A. Merten, D. Pisignano, A. D. Griffiths and D. A. Weitz, *Lab Chip*, 2008, **8**, 1110–1115.
7 D. Di Carlo, D. Irimia, R. G. Tompkins and M. Toner, *Proc. Natl. Acad. Sci. U. S. A.*, 2007, **104**, 18892–18897.
8 F. Xu, S. Moon, X. Zhang, L. Shao, Y. S. Song and U. Demirci, *Philos. Trans. R. Soc. London, Ser. A*, 2010, **368**, 561–583.
9 Y. S. Song, D. Adler, F. Xu, E. Kayaalp, A. Nureddin, R. M. Anchan, R. L. Maas and U. Demirci, *Proc. Natl. Acad. Sci. U. S. A.*, 2010, **107**, 4596–4600.
10 A. Sakai and F. Engelmann, *Cryoletters*, 2007, **28**, 151–172.
11 J. Samot, S. Moon, L. Shao, X. Zhang, F. Xu, Y. Song, H. O. Keles, L. Matloff, J. Markel and U. Demirci, *PLoS One*, 2011, **6**, e17530.
12 S. Y. Teh, R. Lin, L. H. Hung and A. P. Lee, *Lab Chip*, 2008, **8**, 198–220.
13 U. Demirci and G. Montesano, *Lab Chip*, 2007, **7**, 1139–1145.
14 F. Xu, S. Moon, A. E. Emre, E. S. Turali, Y. S. Song, A. Hacking, J. Nagatomi and U. Demirci, *Biofabrication*, 2010, **3**(3), 034101.
15 F. Xu, J. Celli, I. Rizvi, S. Moon, T. Hasan and U. Demirci, *Biotechnol. J.*, 2011, **6**, 204–212.
16 E. Tumarkin, L. Tzadu, E. Csaszar, M. Seo, H. Zhang, A. Lee, R. Peerani, K. Purpura, P. W. Zandstra and E. Kumacheva, *Integr. Biol.*, 2011, **3**, 653–662.
17 F. Xu, B. Sridharan, S. Q. Wang, N. G. Durmus, U. A. Gurkan and U. Demirci, *PLoS ONE*, 2011, **6**(4), e19344.
18 S. Moon, Y. G. Kim, L. Dong, M. Lombardi, E. Haeggstrom, R. V. Jensen, L. L. Hsiao and U. Demirci, *PLoS One*, 2011, **6**, e17455.
19 F. Xu, B. Sridharan, S. Q. Wang, A. U. Gurkan, B. Syverud and U. Demirci, *Biomicrofluidics*, 2011, **5**(2), 22207.
20 M. A. Kohanski, D. J. Dwyer, J. Wierzbowski, G. Cottarel and J. J. Collins, *Cell*, 2008, **135**, 679–690.
21 P. Kumaresan, C. J. Yang, S. A. Cronier, R. G. Blazej and R. A. Mathies, *Anal. Chem.*, 2008, **80**, 3522–3529.

22 K. L. Roach, K. R. King, K. Uygun, S. C. Hand, I. S. Kohane, M. L. Yarmush and M. Toner, *Cryobiology*, 2009, **58**, 315–321.
23 P. S. Dittrich and A. Manz, *Nat. Rev. Drug Discovery*, 2006, **5**, 210–218.
24 J. C. Baret, Y. Beck, I. Billas-Massobrio, D. Moras and A. D. Griffiths, *Chem. Biol.*, 2010, **17**, 528–536.
25 J. Clausell-Tormos, D. Lieber, J. C. Baret, A. El-Harrak, O. J. Miller, L. Frenz, J. Blouwolff, K. J. Humphry, S. Koster, H. Duan, C. Holtze, D. A. Weitz, A. D. Griffiths and C. A. Merten, *Chem. Biol.*, 2008, **15**, 427–437.
26 H. A. Kenny, T. Krausz, S. D. Yamada and E. Lengyel, *Int. J. Cancer*, 2007, **121**, 1463–1472.
27 E. A. Roth, T. Xu, M. Das, C. Gregory, J. J. Hickman and T. Boland, *Biomaterials*, 2004, **25**, 3707–3715.
28 F. Xu, C. M. Wu, V. Rengarajan, T. D. Finley, H. O. Keles, Y. Sung, B. Li, U. A. Gurkan, and U. Demirci, *Adv. Mater.*, 2011, **23**(37), 4254–4260.
29 H. Geckil, F. Xu, X. Zhang, S. Moon and U. Demirci, *Nanomedicine*, 2010, **5**, 469–484.
30 S. Moon, S. K. Hasan, Y. S. Song, F. Xu, H. O. Keles, F. Manzur, S. Mikkilineni, J. W. Hong, J. Nagatomi, E. Haeggstrom, A. Khademhosseini and U. Demirci, *Tissue Eng., Part C*, 2010, **16**, 157–166.
31 B. R. Ringeisen, C. M. Othon, J. A. Barron, D. Young and B. J. Spargo, *Biotechnol. J.*, 2006, **1**, 930–948.
32 T. Boland, T. Xu, B. Damon and X. Cui, *Biotechnol. J.*, 2006, **1**, 910–917.
33 R. K. Pirlo, D. M. D. Dean, D. R. Knapp and B. Z. Gao, *Biotechnol. J.*, 2006, **1**, 1007–1013.
34 Y. S. Song, R. L. Lin, G. Montesano, N. G. Durmus, G. Lee, S.-S. Yoo, E. Kayaalp, E. Haeggstrom, A. Khademhosseini and U. Demirci, *Anal. Bioanal. Chem.*, 2009, **395**, 185–193.
35 Y. S. Song, R. L. Lin, G. Montesano, N. G. Durmus, G. Lee, S. S. Yoo, E. Kayaalp, E. Haeggstrom, A. Khademhosseini and U. Demirci, *Anal. Bioanal. Chem.*, 2009, **395**, 185–193.
36 Y. S. Song, S. Moon, L. Hulli, S. K. Hasan, E. Kayaalp and U. Demirci, *Lab Chip*, 2009, **9**, 1874–1881.
37 Y. G. Kim, S. Moon, D. R. Kuritzkes and U. Demirci, *Biosens. Bioelectron.*, 2009, **25**, 253–258.
38 M. Chabert and J. L. Viovy, *Proc. Natl. Acad. Sci. U. S. A.*, 2008, **105**, 3191–3196.
39 C. H. Choi, J. H. Jung, Y. W. Rhee, D. P. Kim, S. E. Shim and C. S. Lee, *Biomed. Microdevices*, 2007, **9**, 855–862.
40 A. Kumachev, J. Greener, E. Tumarkin, E. Eiser, P. W. Zandstra and E. Kumacheva, *Biomaterials*, 2011, **32**, 1477–1483.
41 A. C. Cameron and P. K. Trivedi, *Regression analysis of count data*, Cambridge University Press, Cambridge; New York, 1998.
42 P. McCullagh and J. Nelder, *Generalized Linear Models*, Chapman and Hall, London, 1989.
43 K. P. Burnham and D. Anderson, *Model Selection and Multi-Model Inference*, Springer, New York, 2003.
44 C. B. Dean, *J. Amer. Statist. Assoc.*, 1992, **87**, 451–457.