Item S1

# Supplementary Description of Statistical Methods

## Beta-Binomial Model for %GSG

### Data

After the needle core biopsy sections were scanned and the glomerular profiles and area of cortex were traced, the data for kidney biopsy $i$ ($i = 1, \ldots, n$) consisted of

$$\begin{aligned} a_i &= \text{ area of cortex in mm}^2 \\ n_i &= \text{ total number of glomeruli} \\ g_i &= \text{ number of GSG} \end{aligned}$$

### Model

*Note: Much of our description of the model is based on the documentation of the* `betabinomial` *function in the* `VGAM` *package (see reference below).*

Our beta-binomial model assumed that

$$g_i \sim \text{binomial}(n_i, p_i) \tag{1}$$

where $\sim$ signifies "is distributed as", and binomial$(n, p)$ refers to the binomial distribution with number of trials $n$ and success probability $p$. Here, $p_i$ is the patient-specific probability that a glomerulus is globally sclerotic, i.e. $\frac{1}{100}$(patient %GSG). The model further assumed that

$$p_i \sim \text{beta}(\mu_i, \psi_i) \tag{2}$$

where $\mu_i$ ($0 < \mu_i < 1$) is the mean of the beta distribution, and $\psi_i$ ($0 < \psi_i < 1$) is a variability parameter. The beta distribution is more commonly described in terms of shape parameters $\alpha$ and $\beta$. The mean-variability parameterization relates to the shape parameterization as follows:

$$\mu = \frac{\alpha}{\alpha + \beta} \tag{3}$$

$$\psi = \frac{1}{1 + \alpha + \beta} \tag{4}$$

The variance of the beta distribution is directly related to the parameter $\psi$ (hence our term "variability parameter") but depends on $\mu$ also:

$$\text{Var}(p_i) = \mu_i(1 - \mu_i)\psi_i \tag{5}$$

The mean and variability of the beta distribution were both allowed to depend on the amount of biopsied cortex, measured by either $n_i$ or $a_i$, in a manner described by logit link equations (below, replace $n_i$ with $a_i$ for the model based on cortex area):

$$\log\left(\frac{\mu_i}{1-\mu_i}\right) = \gamma_{\mu 0} + \gamma_{\mu 1} n_i \tag{6}$$

$$\log\left(\frac{\psi_i}{1-\psi_i}\right) = \gamma_{\psi 0} + \gamma_{\psi 1} n_i \tag{7}$$

### Parameter Estimates

The model was fit via maximum likelihood. The parameter estimates of interest were $\gamma_{\mu 1}$ and $\gamma_{\psi 1}$, exponentiated to yield estimates of

$$e^{\gamma_{\mu 1}} = \text{multiplicative change in } \frac{\mu}{1-\mu} \text{ associated with 1 unit increase in } n \text{ (or } a\text{)} \tag{8}$$

$$e^{\gamma_{\psi 1}} = \text{multiplicative change in } \frac{\psi}{1-\psi} \text{ associated with 1 unit increase in } n \text{ (or } a\text{)} \tag{9}$$

Since both $\mu$ and $\psi$ were estimated to be small ($\mu < 0.06$, $\psi < 0.12$) within the range of our data, we used the approximation

$$\frac{x}{1-x} \approx x \text{ for small } x \tag{10}$$

and report $100 \times (e^{\gamma_{\mu 1}} - 1)$ as the percent change in $\mu$ associated with each 1 unit increase in $n$ or $a$. Strictly speaking, it is the percent change in $\frac{\mu}{1-\mu}$. Likewise for our reported estimates relating to $\psi$.

### Implementation in R

We fit the beta-binomial model using the `vglm` function in the R package `VGAM`. The R code below demonstrates how the model was fit with continuous-valued number of glomeruli (in standard deviations) as the predictor.

```
library(VGAM)
head(biopsies)
```

```
##   n_gsg n_nsg n_glom_sd cort_area_sd
## 1     1    20 1.9105717     2.019344
## 2     1     8 0.8188165     1.368577
## 3     1    11 1.0917553     1.907890
## 4     0    12 1.0917553     1.687694
## 5     1    11 1.0917553     1.173255
## 6    11    22 3.0023270     2.241369
```

```
fit <- vglm(cbind(n_gsg, n_nsg) ~ n_glom_sd,
            # The "zero = NULL" argument specifies that neither
            # parameter should be intercept-only; both mu and psi
            # should depend on the predictor.
            family = betabinomial(zero = NULL),
            data = biopsies,
            # Using biopsies with >1 glomeruli avoids a
            # warning about extremely small working weights
            # in the model-fitting procedure.
            subset = n_gsg + n_nsg > 1)
# In the parameter names, ":1" refers to mu, ":2" to psi.
exp(coef(fit))
```

```
## (Intercept):1 (Intercept):2   n_glom_sd:1   n_glom_sd:2
##    0.04849442    0.06538981    0.84721524    0.81192099
```

```
100 * (exp(coef(fit)) - 1)
```

```
## (Intercept):1 (Intercept):2   n_glom_sd:1   n_glom_sd:2
##     -95.15056     -93.46102     -15.27848     -18.80790
```

```
100 * (exp(confint(fit)) - 1)
```

```
##                     2.5 %     97.5 %
## (Intercept):1 -95.75838 -94.455631
## (Intercept):2 -95.53338 -90.427147
## n_glom_sd:1   -20.37274  -9.858304
## n_glom_sd:2   -31.13928  -4.268256
```

### References


- Thomas W. Yee (2018). VGAM: Vector Generalized Linear and Additive Models. R package version 1.0-6. URL https://CRAN.R-project.org/package=VGAM

## Detection of 10% GSG

For the comparison of biopsies with 1–9 glomeruli to biopsies with 10 or more glomeruli in terms of their utility for detecting clinically significant %GSG, we estimated the distribution of patient %GSG associated with each group of biopsies using beta-binomial models. One intercept-only model was fit to the biopsies with 1–9 glomeruli and another to the biopsies with 10 or more glomeruli, producing estimates of $\mu$ and $\psi$ for each group. The probabilities

$$\text{PPV} = \Pr(\text{patient \%GSG} \geq 10\% \mid \text{biopsy \%GSG} \geq 10\%) \tag{11}$$

$$\text{NPV} = \Pr(\text{patient \%GSG} < 10\% \mid \text{biopsy \%GSG} < 10\%) \tag{12}$$

$$\text{Sensitivity} = \Pr(\text{biopsy \%GSG} \geq 10\% \mid \text{patient \%GSG} \geq 10\%) \tag{13}$$

$$\text{Specificity} = \Pr(\text{biopsy \%GSG} < 10\% \mid \text{patient \%GSG} < 10\%) \tag{14}$$

were calculated using simulation. Simulated numbers of biopsy glomeruli ($n$) were sampled with replacement from the numbers in our data, then true patient proportion GSG values ($p$) were sampled from beta distributions with $\mu$ and $\psi$ set equal to the values estimated above for 1–9 glomeruli and for 10 or more glomeruli. GSG counts ($g$) were sampled from binomial$(n, p)$ distributions using the values of $n$ and $p$ from the previous steps. The probabilities were estimated from $10^6$ simulated kidney/biopsy samples for each biopsy size group. PPV, for example, was calculated as the proportion of cases with $p \geq 0.1$ among all those with $g/n \geq 0.1$.