[redacted] FDA Clinical Study Protocol – Sample Size Justification
[redacted], on behalf of [redacted]
[redacted] (Revised)

We powered the study based on the primary endpoint, the final sensitivity and specificity of the [redacted] as compared to the [redacted] in detecting clinically relevant breast lumps. To establish non-inferiority of the [redacted], the target outcomes are at least 80% of the sensitivity and specificity of the [redacted]. Initial [redacted] studies reported a sensitivity range of 84-87% and a specificity range of 80-94% ([redacted]). Accordingly with our non-inferiority targets, the minimum desired sensitivity and specificity of the [redacted] are 67% and 64%, respectively. Additionally, based on the upper end of the ranges reported for the [redacted], the sensitivity and specificity of the [redacted] could be as high as 70 and 75%, respectively, while achieving the target outcomes.

Our sample cohort will consist of adult women in the Philadelphia metropolitan area, including both symptomatic and asymptomatic participants. We expect the prevalence of clinically relevant breast lumps to be **40%** within symptomatic women and **10%** within asymptomatic women. We currently plan to recruit an equal balance of symptomatic and asymptomatic participants. Therefore, we anticipate that the cohort prevalence of clinically relevant breast lumps will be between **15 - 30%**, though it will depend on the relative proportion of recruited symptomatic and asymptomatic women.

Our primary endpoint consists of two hypotheses involving the sensitivity and specificity of the [redacted]. For hypothesis testing of sensitivity, we will perform a two-sided binomial test of the proportion of correctly identified breast lumps (see Appendix for details). Likewise for specificity, we will perform a two-sided binomial test of the proportion of correctly identified negative findings. We will analyze results at the breast examination level (i.e., two cases per woman) using a false positive rate of 5%. In addition to the primary endpoint, we will also compare the sensitivity and specificity of the [redacted] to those of the clinical breast exam, which we expect to perform comparably to the [redacted] based on recent data ([redacted]).

The following table presents the required number of women for varying scenarios with cohort prevalence ranging from 15 to 30%, desired sensitivity ranging from 67 to 70%, and desired specificity ranging from 64 to 75%. In all scenarios we set desired power at 80%. For each scenario, we reported the required number of positive cases and the required number of negative cases based on power calculations for sensitivity and specificity, separately. Then, using cohort prevalence, we computed the required number of total cases based on the test that yielded the larger sample size in each scenario. For example, the number of positive cases required to detect a sensitivity of 67% at 80% power is 64, and the number of negative cases required to detect a specificity of 64% at 80% power is 98. Assuming a cohort prevalence of 30%, the minimum sample size required for testing sensitivity is 214 (e.g., 64 / 0.3), and the minimum sample size required for testing

specificity is 140 (e.g., 90 / 0.7). We take the larger of these two numbers and arrive at a total required sample size of 214 breast exams, or 107 women.

Table 1: Required number of positive, negative, and total cases for varying scenarios

Cohort Prevalence: 30%					
Sensitivity	**Specificity**	**$N_{positive}$**	**$N_{negative}$**	**N_{total}**	**N_{women}**
67	64	64	98	214	**107**
70	75	49	29	164	**82**
Cohort Prevalence: 20%					
Sensitivity	**Specificity**	**$N_{positive}$**	**$N_{negative}$**	**N_{total}**	**N_{women}**
67	64	64	98	320	**160**
70	75	49	29	245	**123**
Cohort Prevalence: 15%					
Sensitivity	**Specificity**	**$N_{positive}$**	**$N_{negative}$**	**N_{total}**	**N_{women}**
67	64	64	98	427	**214**
70	75	49	29	327	**164**

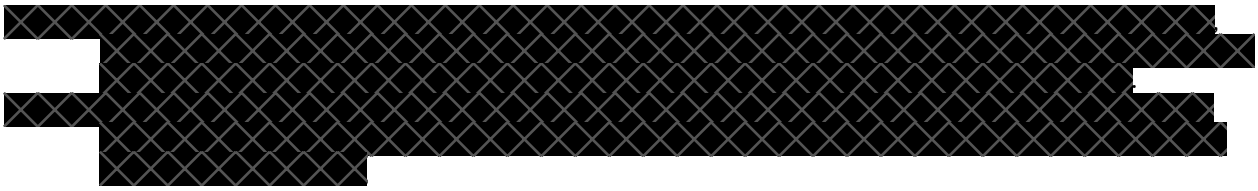
Caption: Cohort prevalence of clinically relevant breast lumps was assumed to vary between 15 and 30%. Minimum desired sensitivity of the [redacted] varied from 67 to 70%; minimum desired specificity varied from 64 to 75%. In all scenarios 80% power was required with a type I error rate of 5%. $N_{positive}$ is the number of required positive cases confirmed by Mammogram or similar gold standards; $N_{negative}$ is the number of required negative cases. Cohort prevalence was used to determine the number of total cases to detect the desired sensitivity and specificity, separately, and the larger of these two estimates was reported as N_{total}. N_{women} is the number of participants required for the study.

Note, all power calculations were performed in the R statistical programming language (version 4.3.2) using the *pwr* package (Champely, et al. 2020).

Conclusions from the Sample Size Justification

- Our base case scenario assumes a cohort prevalence of 30% and a sensitivity/specificity of 67/64%. Under this scenario, **107** women are required to detect the desired sensitivity and specificity levels with 80% power.
- Our most conservative scenario assumes a cohort prevalence of 15% and sensitivity/specificity of 67/64%. Under this scenario, **214** women are required to detect the desired sensitivity and specificity levels with 80% power.
- Using the base case scenario, we believe our study will be adequately powered with **200** women to establish non-inferiority of the [redacted] versus the [redacted]. We can afford a dropout rate of 10% as well as an additional 10% simple cyst exclusion rate, while still retaining data from 162 participants for analysis.

REFERENCES



APPENDIX – Formulas and Code

The binomial test is used to test a hypothesis about the probability of a particular outcome. In the case of **sensitivity**, the outcome is the correct identification of a clinically relevant breast lump. We begin with the conventional null hypothesis,

$$H_0: \pi = 0.5$$

The two-sided p-value is computed from the binomial distribution probability mass function,

$$p = \sum_{i \in \mathrm{I}} \Pr\,(X = i) = \sum_{i \in \mathrm{I}} \binom{n}{i} \pi_0^{i(1-\pi_0)^{n-i}}$$

Where $\mathrm{I} = \{i: \Pr\,(X = i) \leq \Pr\,(X = k)\}$ denotes the set of all events as or more extreme than the observed event, $X = k$.

The following R code was used to perform the power calculations using a one-sample proportion test from the *pwr* package.

```
library(pwr)

# target sensitivity & specificity
TARGET.SENS <- 0.8 * 0.84
TARGET.SPEC <- 0.8 * 0.80

# target prevalence
PREVALENCE <- 0.30

# calculate sample size needed for given prevalence
h.sens <- ES.h(0.5, TARGET.SENS)
pwr.sens <- pwr.p.test(h=h.sens, n=NULL, sig.level=0.05, power=0.8)
plot(pwr.sens)
message("Number of positives needed: ", ceiling(pwr.sens$n))
n.sens <- ceiling(ceiling(pwr.sens$n) / PREVALENCE)

h.spec <- ES.h(0.5, TARGET.SPEC)
pwr.spec <- pwr.p.test(h=h.spec, n=NULL, sig.level=0.05, power=0.8)
plot(pwr.spec)
message("Number of negatives needed: ", ceiling(pwr.spec$n))
n.spec <- ceiling(ceiling(pwr.spec$n) / (1- PREVALENCE))

n.total <- max(n.sens, n.spec)
message("Total needed with current prevalence: ", n.total)
message("Total women needed: ", ceiling(n.total/2))
```