February 20, 2007

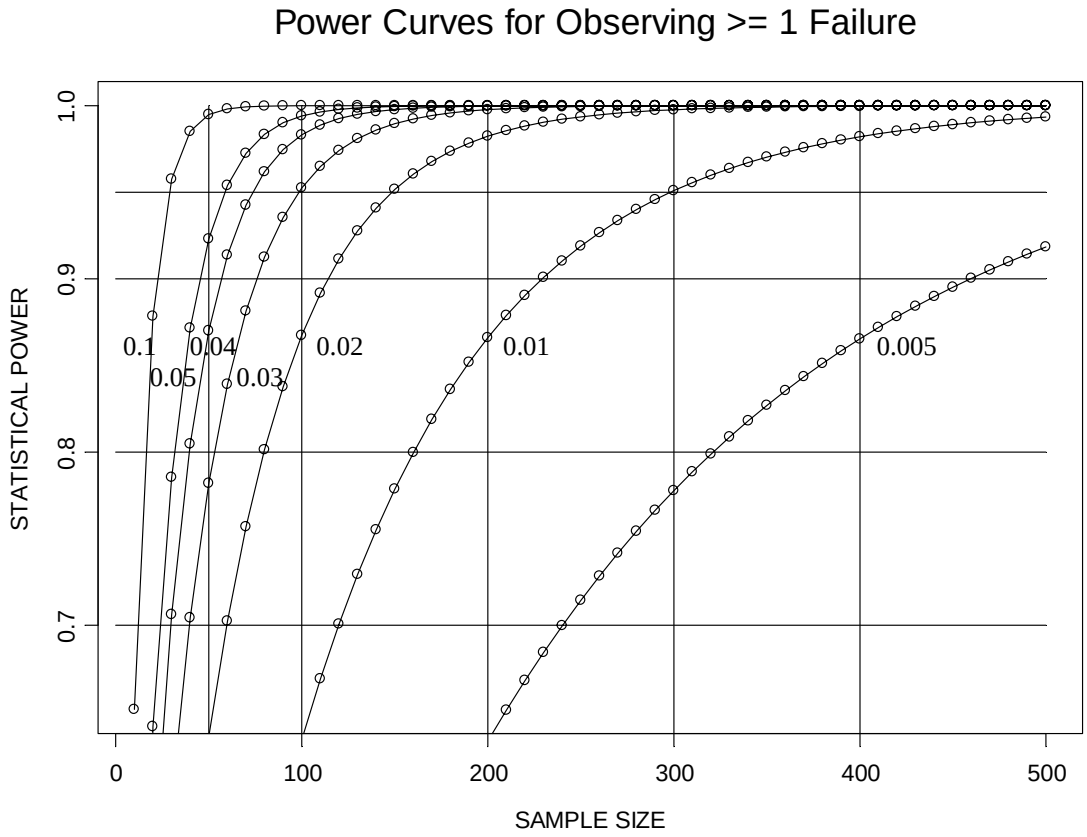
Sampling Plan and Statistical Considerations for the Long Term Field Evaluation Program

1. Criteria for Evaluating Results and Determining Sample Size Requirements:

For purposes of this analysis, a set of criteria, e.g. maximum CO_2 no greater than 4%, minimum O_2 level of at least 15%, and a capacity of at least 60 minutes, define acceptable limits. Based on that definition, the data are treated as yes/no measurements, with the goal of observing 0% failures. Power estimates for detecting at least one failure are then calculated under different assumptions about the true underlying failure rate (for a given population of units). Although we can never be certain that a sample, or subset of the population, will include a failure if the true rate is extremely low, the subsequent power estimates provide reasonable guidelines for specifying a sufficiently large number of units to achieve a high degree of confidence for detecting at least one failure under unacceptable conditions. To define unacceptable conditions, different assumptions are made about the underlying failure rate.

2. General Method for Estimating a Sufficient Sample Size:

The following section outlines a methodology for determining a required sample size. Later sections will address the incorporation of that sample size requirement into an overall sampling plan. To determine a sufficient sample size for this goal, we can evaluate the statistical power, or probability of detecting one or more failures under different assumptions and using different sample sizes. In other words, we are attempting find the minimum number of units that we need to have a high degree of confidence that we will observe one or more failures given some true failure rate for the underlying population. This leads to the following results in Figure 1. (See explanation on the following page.)



2. a. Interpreting the Power Curves in Figure 1:

The y-axis of Figure 1 gives the probability that we will observe at least 1 failure (i.e. power) for a given sample size (labeled on the x-axis). Each curve represents a different assumption about the underlying failure rate of the given population of units; curves are drawn for failure rates of 10%, 5%, 4%, 3%, 2%, 1% and 0.5%. Horizontal lines are drawn to mark specific cases that give 70, 80, 90 and 95% power. The vertical lines mark sample sizes of 50, 100, 200, 300 and 400 (n = 500 is given by the far right side of the graph). Open circles on each curve are drawn at intervals of n = 10.

To interpret the resulting graph, take the following steps.

1) make an assumption about the underlying failure rate; find the corresponding curve
2) decide on an acceptable power; follow the curve up until it reaches that power on the y-axis (i.e. intersects the horizontal line corresponding to the desired power)
3) trace that point down to the x-axis

For instance, say we want to find a sufficient sample for achieving a 90% probability that we will observe at least one failure if the true failure rate is 1%. First, find the curve labeled as 0.01. Then follow that curve to the horizontal line corresponding to 0.9 on the y-axis. The curve intersects that horizontal line somewhere between n = 200 and n =

300. Since it intersects the curve at approximately the third open circle past n = 200, the corresponding sample size is approximately n = 230. In fact, n = 229 gives just under 90% power (89.9894%); therefore, if we assume the true failure rate is 1%, we should use n = 230 to have 90% for observing a failure.

Another approach to using the power curves in Figure 1 is to fix the sample size and then assess the power for a given assumption about the true failure rate, or assess what assumption we need to make for a given power. For instance, fixing the sample size at n = 100 corresponds to the vertical line at n = 100. Following that line up from the x-axis gives the following results. A sample size of n = 100 gives 1) approximately 65% power for detecting a failure rate of 1%, 2) approximately 87% power for detecting a failure rate of 2%, 3) just over 95% power for detecting a failure rate of 3%, and 4) very high power (approximately 98-100%) for detecting failure rates of 4% or more. Any curves for failure rates of under 1% will not show up on the current graph (where the y-axis starts at about 65% power), implying that n = 100 gives a relatively low power for detecting failure rates below 1%.

3. Overall Strategy for a Sampling Plan:

Utilizing the ability to enumerate the entire population of units in mining environments (through MSHA listings), the overall sampling strategy will be outlined as follows.

1) Determine the overall sample size to be collected annually; denote that sample size as N
2) collect an inventory from each mine of all respirator units; concatenate those lists into a single master list of all mining respirator units
3) sort that master list by approved respirator model; create a separate list for each approved model
4) within each of the k approved models (currently $k = 4$), randomly sort the corresponding list of respirator units and select the first n units, where $n = N/k$
5) evaluate the resulting ability (e.g. statistical power) to detect failures (as defined in Section 1) over one year under different assumptions, and outline a strategy for increasing statistical power and/or detecting failures within specific subgroups (with a set power) by accumulating data over multiple years

3. a. Eligibility Criteria, Lost to Follow-up, and Maintaining a Random Sample:

Once the k randomly ordered lists are specified, the units (with corresponding serial numbers) will then be collected from the mines. However, it is suspected that some appreciable percentage of the units will not be available for collection (e.g. due to inability to find a given serial number). In addition, a unit must meet the pre-specified guidelines for usability (e.g. lacking visible damage and achieving other such criteria) to be eligible for the sampling process (since the objective of the LTFE program is to evaluate only the units meeting such criteria); units not meeting those criteria are not eligible for data collection, although all such cases will be carefully documented. In either case, when a selected unit is not collected, the next unit on the random list (under that approved model) then becomes part of the sample, thus maintaining the randomness of the sampling plan. Variation from this random mechanism, although potentially aiding

feasibility of the data collection effort, will lead to an unquantifiable bias in the resulting measurements.

3. b. Primary Year 1 Goal:

The goal for year 1 is to detect one or more failures for any approved respirator model that has a sufficiently high true failure rate (of 2-3% or higher).

Table 1 gives the probability of observing one or more failures for a given approved model under different assumptions about the underlying failure rate, the number of approved models, and the overall sample size. (Note: these results are calculated in the same manner as in Figure 1; results are just presented differently for this specific goal.) Results illustrate that an overall sample size of 400, and the current total of 4 approved models, gives a high probability (over 95%) of detecting at least one failure (for a given approved model) if the true failure rate is 3% or higher; using an overall sample size of $N = 300$ gives slightly lower ability to detect failures, as we have at least 95% power for true failure rates of 4% or higher. If we add one approved model, the power for a 3% or 4% failure rate drops to about 91% (from 95%). We also have relatively high power (78% or more) to observe at least one failure for a true failure rate of 2% with the current total of 4 approved models. However, the power drops substantially for a true failure rate of 1%. We can therefore define a 'sufficiently high' true failure rate as 2-3% or more with the given sample sizes.

Table 1. Statistical power for detecting failures within a given approved model (with total sample sizes of $N = 300$ and $N = 400$)

Overall	True	Statistical Power by Number of Approved Units			
Sample Size	Failure Rate	3	4 (current #)	5	6
400	0.10	> 99.9%	> 99.9%	>99.9%	99.9%
	0.05	99.9%	99.4%	98.3%	96.6%
	0.04	99.6%	98.3%	96.2%	93.2%
	0.03	98.3%	95.2%	91.3%	86.6%
	0.02	93.2%	86.7%	80.1%	73.6%
	0.01	73.7%	63.4%	55.2%	48.5%
300	0.10	> 99.9%	> 99.9%	99.8%	99.5%
	0.05	99.4%	97.9%	95.4%	92.3%
	0.04	98.3%	95.3%	91.4%	87.0%
	0.03	95.2%	89.8%	83.9%	78.2%
	0.02	86.7%	78.0%	70.2%	63.6%
	0.01	63.4%	52.9%	45.3%	39.5%

3. c. Primary Year 2 Goal:

The goal for year 2 is to detect one or more failures for any approved respirator model that has a true failure rate of 1 to 2% or higher.

Results from Table 2 illustrate that an overall sample size of 400 per year after two years, and the current number of approved models, gives a high probability (over 95%) of detecting at least one failure (for a given approved model) if the true failure rate is 1.5% or higher; using an overall sample size of 300 per year gives slightly lower ability to detect failures, as we have at least 95% power for true failure rates of 2% or higher. If we add one approved model, the power for a 1.5% or 2% failure rate drops to about 91% (from 95%). We also have relatively high power (78% or more) to observe at least one failure for a true failure rate of 1% with the current total of 4 approved units. However, the power drops substantially for a true failure rate below 1%.

Table 2. Statistical power for detecting failures within a given approved model (with total sample sizes of $N = 600$ and $N = 800$)

Overall	True	Statistical Power by Number of Approved Units			
Sample Size	Failure Rate	3	4 (current #)	5	6
800	0.02	99.5%	98.2%	96.1%	93.2%
	0.015	98.2%	95.1%	91.1%	86.6%
	0.01	93.1%	86.6%	80.0%	73.7%
	0.005	73.6%	63.3%	55.2%	48.7%
	0.001	23.4%	18.1%	14.8%	12.5%
600	0.02	98.2%	95.2%	91.1%	86.7%
	0.015	95.1%	89.6%	83.7%	77.9%
	0.01	86.6%	77.9%	70.1%	63.4%
	0.005	63.3%	52.9%	45.2%	39.4%
	0.001	18.1%	13.9%	11.3%	9.5%

3. d. Additional Goals for Years 2-5:

An additional goal for subsequent years is to detect one or more failures for any specific subgroup of units with sufficiently high failure rates.

The previously-specified analyses all focus on achieving sufficient statistical power across either the entire population of units, or across a given approved model. However, it may also be of interest to consider a more select subgroup of units and evaluate the power to detect at least one failure under different assumptions about the underlying failure rate within that subgroup. These subgroups might include, but are not limited to the deployment type (e.g. carried, stored, warehoused), age of the units, and height or size of the mine.

The specific details for this additional goal depend on the distribution of units across the different subgroups. For instance, for a variable, or model characteristic, with two

subgroups that are evenly distributed across the population (e.g. if 50% of units are 'newer' and 50% are 'older' units), the sample sizes and power estimates in Table 1 will apply to each subgroup of units (within the each approved model) after two years of sampling. In contrast, for a subgroup which only accounts for 10% of units (e.g. for smaller sized mines) it will take 10 years of sampling to accumulate the sample sizes and power estimates in Table 1 for making a specific statement about units within that subgroup.

In addition to assessing the number of years needed to apply estimates in Tables 1 and 2, we can also assess the ability to detect a failure within a given subgroup under the more liberal assumption of a true failure rate of 10%. For instance, as quantified in Figure 1, a sample size of only 30 gives over 95% power to detect at least one failure (when the true failure rate is 10%). Thus, a subgroup which accounts for just under one-third of the units (within a given approved model), and has a true failure rate of 10%, will lead to at least one failure with over 95% probability after only one year. Statistical power for other such scenarios can be specified using the graphs in Figure 1.

4. Other Statistical Analyses:

All available information on respirator units will be collected and (assuming a given variable can be consistently collected across most units) considered in the subsequent analysis. This information will include, but is not limited to, deployment type, age of the unit, and mine height and size. As a first step, a descriptive analysis will be conducted to tabulate the number of failures, and subsequent rates across different subgroups, the relative frequencies of the different subgroups, and summaries of any continuous measurements. Descriptive statistics will be calculated on both the level of the respirator unit (e.g. ages of the units and deployment types) and on the level of the mine (e.g. frequencies of different mine heights and sizes). Information on units failing initial inspection, or otherwise not included in the sample data, will also be included in the analysis. Failure rates will be calculated both in terms of failures resulting from testing of a unit, and failure to meet usability criteria (subsequently leading to exclusions from the sample data). As failure data accumulates over subsequent years of the program, Poisson regression models will be fit to evaluate the significance and effect of different variables on failure rates.

Although the primary analysis will focus on yes/no definitions of a failure, the main parameters (e.g. O_2, CO_2, and capacity time) will also be analyzed in terms of the continuous measurements. Using summary statistics, graphical plots, and appropriate non-linear regression models of those data, trends over time will be analyzed both separately and jointly for those parameters. The above-mentioned variables will again be considered as possible predictors in the model. Hierarchical models may also be implemented to simultaneously consider both mine-level and (respirator) unit-level variables.

5. Strengths and Limitations:

The above-outlined sampling plan presents a significant improvement over current data collection and surveillance of respirator performance, and features several significant strengths in terms of the study design. First, enumeration of the complete respirator population in mines allows for completely random selection (within a given approved model), which is rarely achieved in any epidemiologic or other statistical studies. Second, the fixed configuration of a given approved respirator model allows for accumulation of data over multiple years to build statistical significance and subsequent ability to make more specific conclusions from the data. Finally, although it is expected that a substantial proportion of the units will be lost to follow-up or deemed ineligible for testing, the use of the MSHA list allows for documentation of such events, and calculation of the resulting lost-to-follow-up rates, and, even with these types of missing data, maintenance of completely random data collection, and subsequent minimization of any resulting bias.

Despite these significant strengths, there are some potential limitations to the LTFE program. First, the feasibility of collecting a completely random sample (within each approved model) may be challenging due to the geographic spread of the resulting list of units. Further, collection of units with specific serial numbers may prove difficult, especially in larger mines. In the event that targeted units are not collected, the sampling plan allows for extension of the sampling list according to the previously determined randomization. However, this may impose additional practical hurdles, as collection of additional units may be required from geographically diverse locations. Several steps may be taken to reduce this burden, such as initially over-sampling (again according to the random list) to avoid the practical constraints of having to return to (potentially) geographically diverse sites. The timeline needed to achieve the stated levels of statistical significance might also need to be extended if additional data collection (to make up for units lost-to-follow-up) is not possible. Alternatively, the sampling plan might be revised (in the future) to allow for some type of clustered, or multi-staged sampling across a more narrow range of mines, thus sacrificing randomization, and potentially some measure of representativeness, for increased feasibility.

Although the objective of the LTFE program is to test and evaluate respirators which are deemed usable, it is also necessary to consider the implications of excluding respirator units that do not meet such criteria. If for instance, handling of the unit leads to decreased usability, and that handling is associated with some other factor, such as mine size for instance, that factor will tend to be included in the data at a lower frequency. We might therefore think of the analysis as biased or at least less efficient, especially if any such factor is less prevalent to begin with. Such logic might then lead to a different sampling plan which stratifies on any such factor, thus guaranteeing a given relative frequency for that factor. Despite these concerns, it is important to note that the LTFE program is focused on assessment of respirators which are selected to be usable by the miner. Therefore, these concerns are not of primary importance for this study. Other future studies might address the outcome of usability and associated prevalence's more directly. Overall, the current sampling plan provides the optimal approach for the given objectives of the LTFE program.