MRMC Sample Size Program

User Guide

Stephen L. Hillis
(steve-hillis@uiowa.edu)

Kevin S. Berbaum
(kevin-berbaum@uiowa.edu)

1. Introduction

The programs *MRMC sample size input1.sas* and *MRMC sample size input2.sas* are SAS programs that perform sample size computations for ROC studies where each reader reads all cases under each treatment (or modality). These programs compute the sample size needed to detect a specified difference in the AUC (or other ROC parameter such as sensitivity for a fixed specificity) between two modalities when using the analysis method proposed by Dorfman, Berbaum, and Metz [1], which we refer to as the DBM method. These program are based on the updated DBM method, which differs from the original DBM method [1] in the following two ways: it uses less model reduction, as proposed by Hillis, Obuchowski, Schartz, & Berbaum [2] and validated in simulations by Hillis & Berbaum [3], and it incorporates the new denominator degrees of freedom (ddf_H) proposed by Hillis [4].

The programs compute sample sizes under three different inference situations: (1) both readers and cases are random; (2) readers are fixed and cases are random; and (3) readers are random and cases are fixed. Corresponding results generalize to (1) the reader and case populations for which the study reader and cases are representative; (2) the case population when read by the particular readers in the study; and (3) the reader population when reading the particular cases used in the study.

There are two versions of the program to accommodate two different types of input. The input for *MRMC sample size program input1.sas* consists of the mean squares and reader and case sample sizes from a previous DBM analysis for a similar study. The input for *MRMC sample size program input2* consists of variance component estimates that typically would be obtained from a previous DBM analysis or from pooling estimates across several previous DBM analyses.

2. Running the program

Data must be inputted at the beginning of the program before running it. The program statements shown below, which are the same for both versions, consist of `%let`

statements that create macro variables and assign values to them. For example, `%let min_cases = 20` defines *`min_cases`* to be a macro variable having a value of `20`. This has the effect of replacing *`&min_cases`* throughout the program by `20` when you run it. Similarly, `%let AUCdiff = .03, .05` has the effect of replacing *`%AUCdiff`* by `.03, .05`. The statements can be modified in the obvious way; for instance, we could write

```
%let AUCdiff = .03, .05, .10, .20.
```

```
**NOTE:  set the following parameters as desired**;

%let min_cases = 20;  **minimum number of cases to consider**;
%let max_cases = 2000;   **maximum number of cases to consider*;
%let power = .80 ;  **desired minimum power**;
%let AUCdiff = .03, .05;  **AUC values**;
%let readers = 3 to 15;  **number of readers*;
%let alpha = .05;  **significance level**;
```

The above statements instruct the program to computes the power for effect sizes of AUC = .03 and .05 with readers varying between 3 and 15 and cases varying between 20 and 2000. The program begins by setting AUC = .03 and readers = 3 and computing the power for cases = 20, 21, 22, ..., 2000. As soon as the power exceeds .80, then the case do-loop ends and the case sample size is outputted. If .8 power is not obtained for 2000 readers, then a missing value is outputted. The program then starts over, computing power for 4 readers and cases = 20, 21, 22, ..., 2000, etc. This process continues through 15 readers.

The statements below, from *MRMC sample size input1.sas*, show where you input the pilot data mean squares and reader and case sample sizes.

```
data pilotdata;  **input pilot study statistics in this data step**;
  input dataset $ r_star c_star mstr mstc mstrc ;

   **dataset is a label identifying the data set in the output,
     r_star = number of readers, c_star = number of cases,
     mstr = MS(treat x reader),
     mstc = MS(treat x case),
     mstrc = MS(treat x reader x case)**;

  **now compute the variance components**;
  var_trc = max(mstrc, 0);
  var_tr = (mstr - mstrc)/c_star; *use unbiased estimate*;
  var_tc = max( (mstc - mstrc)/r_star, 0);

  cards; **input data here**
VanDyke 5 114 0.11027549 0.15011443 0.06825495
Franken 4 100 0.00778009 0.07807153 0.08364310
```

Note that the data are inputted only in the last two lines. For example,

```
Franken 4 100 0.00778009 0.07807153 0.08364310
```

shows that for the "Franken" data set there are 4 readers, 100 cases, and *mstr* = 0.00778009, *mstc* = 0.07807153, and *mstrc* = 0.09364310, where *mstr*, *mstc*, and *mstrc* are the treatment-by-reader, treatment-by-case, and treatment-by-reader-by-case mean squares, respectively.

In *MRMC sample size input2sas* you enter the variance components, instead of the mean squares, in the following statements:

```
data var_components; **input pilot study statistics in this data
step**;
  input dataset $ var_tr var_tc var_trc;

   **dataset is a label identifying the data set in the output,
     var_tr = treatment x reader variance component,
     var_tc = treatment x case variance component.
     var_trc = treatment x reader x case variance component;

  cards; **input data here**
VanDyke  0.000368640 0.016371 0.068255
Franken  -.000758630 0.000000 .083643
```

Note that the data are inputted only in the last two lines. For example,

```
Franken  -.000758630 0.000000 .083643
```

shows that for the "Franken" data set *var_tr* = -.00075860, *var_tc* = 0.000000, and *var_trc* = .083643, where *var_tr*, *var_tc*, and *var_trc* are the treatment-by-reader, treatment-by-case, and treatment-by-reader-by-case variance components. Although theoretically a variance component cannot be negative, we note that the unbiased variance component *estimates* can be negative, as in this example. This point is discussed further in the Section 4.

To be consistent we have used the same input data for our examples as used by Hillis and Berbaum [5]. These input values were obtained use a DBM analysis where the outcome was the maximum likelihood estimate of the AUC, assuming a binormal model. Note that the input values would be different if another estimation method, such as PROPROC or the trapezoidal method, had been used.

3. Output

Below I show the output from running *MRMC sample size input1.sas*.

Beginning lines of program:

```
**NOTE:  set the following parameters as desired**;

%let min_cases = 20;  **minimum number of cases to consider**;
%let max_cases = 2000;   **maximum number of cases to consider*;
%let power = .80 ;  **desired minimum power**;
%let AUCdiff = .03, .05;  **AUC values**;
```

```
%let readers = 3 to 15;  **number of readers*;
%let alpha = .05;  **significance level**;

**NOTE:  input the pilot data in the next data set**;

data pilotdata;  **input pilot study statistics in this data step**;
  input dataset $ r_star c_star mstr mstc mstrc ;

   **dataset is a label identifying the data set in the output,
     r_star = number of readers, c_star = number of cases,
     mstr = MS(treat x reader),
     mstc = MS(treat x case),
     mstrc = MS(treat x reader x case)**;

  **now compute the variance components**;
  var_trc = max(mstrc, 0);
  var_tr = (mstr - mstrc)/c_star; *use unbiased estimate*;
  var_tc = max( (mstc - mstrc)/r_star, 0);

  cards; **input data here**
VanDyke 5 114 0.11027549 0.15011443 0.06825495
Franken 4 100 0.00778009 0.07807153 0.08364310
```

Output:

inputted data and computed variance components 1

Obs	dataset	r_star	c_star	mstr	mstc	mstrc	var_trc	var_tr	var_tc
1	VanDyke	5	114	0.11028	0.15011	0.068255	0.068255	0.000368601	0.016372
2	Franken	4	100	0.00778	0.07807	0.083643	0.083643	-.000758630	0.000000

Reader and sample combinations needed for .80 power 2

NOTE: cases = . means that .80 power not available for cases <= 2000

Obs	dataset	AUCdiff	readers	c_both_ random	c_cases_ random	c_readers_ random
1	Franken	0.03	3	105	489	105
2	Franken	0.03	4	98	367	98
3	**Franken**	**0.03**	**5**	**92**	**294**	**92**
4	Franken	0.03	6	86	246	86
5	Franken	0.03	7	81	211	81
6	Franken	0.03	8	77	185	77
7	Franken	0.03	9	73	165	73
8	Franken	0.03	10	69	148	69
9	Franken	0.03	11	66	135	66
10	Franken	0.03	12	63	124	63
11	Franken	0.03	13	61	115	61
12	Franken	0.03	14	58	107	58
13	Franken	0.03	15	56	100	56
14	Franken	0.05	3	96	177	96
15	Franken	0.05	4	81	134	81
16	**Franken**	**0.05**	**5**	**70**	**107**	**70**
17	Franken	0.05	6	62	90	62

18	Franken	0.05	7	55	77	55
19	Franken	0.05	8	50	68	50
20	Franken	0.05	9	46	61	46
21	Franken	0.05	10	42	55	42
22	Franken	0.05	11	39	50	39
23	Franken	0.05	12	36	46	36
24	Franken	0.05	13	34	43	34
25	Franken	0.05	14	32	40	32
26	Franken	0.05	15	30	37	30
27	VanDyke	0.03	3	.	685	.
28	VanDyke	0.03	4	.	586	.
29	**VanDyke**	**0.03**	**5**	**.**	**526**	**.**
30	VanDyke	0.03	6	.	486	.
31	VanDyke	0.03	7	.	458	.
32	VanDyke	0.03	8	.	437	.
33	VanDyke	0.03	9	.	420	.
34	VanDyke	0.03	10	1890	407	803
35	VanDyke	0.03	11	1277	396	477
36	VanDyke	0.03	12	1010	387	340
37	VanDyke	0.03	13	859	380	264
38	VanDyke	0.03	14	761	373	216
39	VanDyke	0.03	15	693	367	183
40	VanDyke	0.05	3	.	248	.
41	VanDyke	0.05	4	.	212	.
42	**VanDyke**	**0.05**	**5**	**833**	**191**	**933**
43	VanDyke	0.05	6	400	177	286
44	VanDyke	0.05	7	298	166	170
45	VanDyke	0.05	8	249	159	121
46	VanDyke	0.05	9	221	153	94
47	VanDyke	0.05	10	202	148	77
48	VanDyke	0.05	11	189	144	65
49	VanDyke	0.05	12	178	141	57
50	VanDyke	0.05	13	170	138	50
51	VanDyke	0.05	14	164	136	45
52	VanDyke	0.05	15	159	134	41

In the above output the variables *c_both_random, c_cases_random,* and *c_readers_random* give the number of cases needed to achieve at least .8 power under these three inference situations, respectively: (a) random readers and random cases (b) fixed readers and random cases, and (c) random readers and fixed cases.

Consider the following lines taken from the above output:

Obs	dataset	AUCdiff	readers	c_both_ random	c_cases_ random	c_readers_ random
3	**Franken**	**0.03**	**5**	**92**	**294**	**92**
29	**VanDyke**	**0.03**	**5**	**.**	**526**	**.**
16	**Franken**	**0.05**	**5**	**70**	**107**	**70**
42	**VanDyke**	**0.05**	**5**	**833**	**191**	**933**

We see that using 5 readers we need 92 cases for the Franken study to obtain .8 power to detect a .03 AUC difference with alpha = .05, treating readers and cases as random,. For the VanDyke study the missing value for *c_both_random* indicates that .8 power is not

achieved with 2000 cases. For fixed readers and random cases, we need 294 and 526 cases for the Franken and VanDyke studies, respectively.

Similarly, we see that to obtain .8 power to detect a .05 AUC difference with alpha = .05 using 5 readers and treating readers and cases as random, we need 70 and 833 cases for the Franken and VanDyke studies, respectively. For fixed readers and random cases, we need 107 and 191 cases, respectively, for the Franken and VanDyke studies

We note that the number of cases (92) needed for the Franken study with 3 readers and AUC = .03 when both readers and cases are random is considerably less than the number needed (294) when only cases are random. If the variance components were all known, this would not be possible. It happens here because we are using *estimates* of the variance components rather than the true unknown parameter values; in particular, it happens because the treatment-by-reader variance component estimate is negative (*var_tr* = -.00075860), as shown on page 1 of the output:

inputted data and computed variance components 1

Obs	dataset	r_star	c_star	mstr	mstc	mstrc	var_trc	var_tr	var_tc
1	VanDyke	5	114	0.11028	0.15011	0.068255	0.068255	0.000368601	0.016372
2	Franken	4	100	0.00778	0.07807	0.083643	0.083643	**-.000758630**	0.000000

In this situation a conservative approach is to rerun the program with the treatment-by-reader variance component estimate set to zero. That is, use the variance component estimates in the above output as input for *MRMC sample size input2.sas*, but change the negative treatment-by-reader variance component estimate to zero, as shown below:

```
data var_components; **input pilot study statistics in this data
step**;
  input dataset $ var_tr var_tc var_trc;

   **dataset is a label identifying the data set in the output,
     var_tr = treatment x reader variance component,
     var_tc = treatment x case variance component.
     var_trc = treatment x reader x case variance component;

  cards; **input data here**
Franken  0 0.000000 .083643
```

Selected output:

Obs	dataset	AUCdiff	readers	c_both_ random	c_cases_ random	c_readers_ random
3	**Franken**	**0.03**	**5**	**526**	**294**	**526**
16	**Franken**	**0.05**	**5**	**190**	**107**	**190**

Now we see that the number of cases for random readers and random cases has increased substantially. Alternatively, in this situation you may want to consider pooling information from several similar studies, resulting in more precise variance component estimates and hence more precise sample size estimates. See Section 4 for details.

4. Details

4.1 The DBM Model: random readers and random cases

For the DBM method, AUC (or other ROC parameter) pseudovalues are computed using the Quenouille-Tukey jackknife separately for each reader-modality combination as described in Dorfman et al [1]. Let Y_{ijk} denote the AUC pseudovalue for modality *i*, reader *j*, and case *k*; by definition $Y_{ijk} = c\hat{\theta}_{ij} - (c-1)\hat{\theta}_{ij(k)}$, where $\hat{\theta}_{ij}$ denotes the AUC estimate based on all of the data for the *i*th modality and *j*th reader, and $\hat{\theta}_{ij(k)}$ denotes the AUC estimate based on the same data but with data for the *k*th case removed. Using the Y_{ijk} as the responses, we then test for a modality effect using a fully crossed three-factor ANOVA, with modality treated as a fixed factor and reader and case as random factors. The jackknife estimate of the AUC for the *i*th modality and *k*th reader, denoted by $\widehat{\mathrm{AUC}}_{ij}$, is equal to the mean of the corresponding pseudovalues; that is, $\widehat{\mathrm{AUC}}_{ij} = \bar{Y}_{ij.}$, where $\bar{Y}_{ij.} = \frac{1}{c}\sum_{i=1}^{t}\sum_{j=1}^{r} Y_{ijk}$. Here the subscript replaced by a dot indicates that values under the bar are averaged across the missing subscript.

The analysis model is given by

$$Y_{ijk} = \mu + \tau_i + R_j + C_k + (\tau R)_{ij} + (\tau C)_{ik} + (RC)_{jk} + (\tau RC)_{ijk} + \varepsilon_{ijk}, \qquad (1)$$

$i = 1,\ldots,t,\ j = 1,\ldots,r, k = 1,\ldots,c,$ where τ_i denotes the fixed effect of modality (or treatment) *i*, R_j denotes the random effect of reader *j*, C_k denotes the random effect of case *k*, the multiple symbols in parentheses denote interactions, and ε_{ijk} is the error term. Main fixed effects are denoted with a Greek letter and random effects with a capital English letter. The interaction terms are all random effects. The random effects are assumed to be mutually independent and normally distributed with zero means and variances $\sigma_R^2, \sigma_C^2, \sigma_{\tau R}^2, \sigma_{\tau C}^2, \sigma_{RC}^2, \sigma_{\tau RC}^2$, and σ_ε^2, where the subscript indicates the corresponding random effect. Since there are no replications, $\sigma_{\tau RC}^2$ and σ_ε^2 are inseparable, and hence we define $\sigma^2 = \sigma_{\tau RC}^2 + \sigma_\varepsilon^2$.

Let MS(T), MS(T*R), MS(T*C), and MS(T*R*C) denote the means squares corresponding to the treatment, treatment × reader, treatment × case, and treatment × reader × case effects, respectively. The updated DBM F statistic, as discussed in References [2-4], for testing for a treatment effect is given by

$$F = \frac{\text{MS(T)}}{\text{MS(T*R)} + \max[\text{MS(T*C)} - \text{MS(T*R*C)}, 0]} \tag{2}$$

The null hypothesis is rejected if $F > F_{\alpha;t-1,\text{ddf}_\text{H}}$, where α is the significance level and

$$\text{ddf}_\text{H} = \frac{\left[\text{MS(T*R)} + \max[\text{MS(T*C)} - \text{MS(T*R*C)}, 0]\right]^2}{\frac{\text{MS(T*R)}^2}{(t-1)(r-1)}}.$$

Unbiased estimates of the variance components σ^2, $\sigma^2_{\tau R}$, and $\sigma^2_{\tau C}$ are given by the following ANOVA estimates:

$$\hat{\sigma}^2 = \text{MS(T*R*C)},$$

$$\hat{\sigma}^2_{\tau R} = \frac{\text{MS(T*R)} - \text{MS(T*R*C)}}{c},$$

and

$$\hat{\sigma}^2_{\tau C} = \frac{\text{MS(T*C)} - \text{MS(T*R*C)}}{r}. \tag{3}$$

It follows that the F statistic and ddf_H can be written as

$$F = \frac{\text{MS(T)}}{c\hat{\sigma}^2_{\tau R} + \hat{\sigma}^2 + \max[r\hat{\sigma}^2_{\tau C}, 0]}$$

$$\text{ddf}_\text{H} = \frac{\left\{c\hat{\sigma}^2_{\tau R} + \hat{\sigma}^2 + \max[r\hat{\sigma}^2_{\tau C}, 0]\right\}^2}{\frac{\left[c\hat{\sigma}^2_{\tau R} + \hat{\sigma}^2\right]^2}{(t-1)(r-1)}} \tag{4}$$

The power for comparing two treatments is approximated by

$$\text{power} \approx \Pr\left(F_{1,\text{ddf}_\text{H};\hat{\Delta}} > F_{1-\alpha;1,\text{ddf}_\text{H}}\right) \tag{5}$$

where the noncentrality parameter estimate is given by

$$\hat{\Delta} = \frac{\left[\text{AUC}_1 - \text{AUC}_2\right]^2}{\frac{2}{rc}\left(c\hat{\sigma}^2_{\tau R} + \hat{\sigma}^2 + \max\left[r\hat{\sigma}^2_{\tau C}, 0\right]\right)}. \tag{6}$$

In *MRMC sample size input1.sas* the power is computed by computing the variance component estimates (3) from the inputted mean squares and reader and case sample sizes for the pilot data analysis, followed by computing equations (4), (6), and (5) for various values of *r* and *c*. *MRMC sample size input2.sas* uses the same algorithm but does not compute the variance components since they are inputted.

This method of computing the power is the same as used by Hillis and Berbaum [5], except for two aspects: (a) we use the new denominator degrees of freedom, ddf_H, as suggested by Hillis [4]; and (b) we use the unbiased variance component estimates (3) in equations (4) and (6). In contrast, Hillis and Berbaum set the treatment-by-reader variance component estimate to zero when it is negative. However, by doing so the estimate of the noncentrality parameter is positively biased.

As seen in the Franken example, use of a negative treatment-by-reader variance component estimate can result in the number of cases needed for random readers and random cases being considerably less than the number needed when only cases are random. Again, if the variance components were all known, this would not be possible, but it can happen because we are using *estimates* of the variance components rather than the true unknown parameter values. We note that when treatment-by-reader variance component is small, then it is natural that the estimate will be negative with high probability since it is an unbiased estimate. However, our estimate of the variance of the estimator will always be positive. As discussed previously, in this situation a conservative approach is to rerun the program with the treatment-by-reader variance component estimate set to zero. That is, use the variance component estimates in the above output as input for *MRMC sample size input2.sas*, but change the negative treatment-by-reader variance component estimate to zero, as was previously shown. Alternatively, in this situation you may want to consider averaging variance componets across from several similar studies, resulting in more precise variance component estimates and hence more precise sample size estimates.

4.2 The DBM Model: fixed readers and random cases

The model is the same as in Section 4.1, except that now $\sigma^2_{\tau R} = \sigma^2_R = 0$. The appropriate test statistic, as discussed in References [2,4], for testing for a treatment effect is given by

$$F = \frac{\text{MS(T)}}{\text{MS(T*C)}} \tag{7}$$

The null hypothesis is rejected if $F > F_{\alpha;t-1,(t-1)(c-1)}$, where α is the significance level.

Using an approach similar to that used in the previous section, the power for comparing two treatments is approximated by

$$\text{power} \approx \Pr\left(F_{1,c\text{-}1;\hat{\Delta}} > F_{1-\alpha;1,c-1}\right) \tag{8}$$

where the noncentrality parameter estimate is given by

$$\hat{\Delta} = \frac{\left[\mathrm{AUC}_1 - \mathrm{AUC}_2\right]^2}{\frac{2}{rc}\left(r\hat{\sigma}_{\tau C}^2 + \hat{\sigma}^2\right)}. \tag{9}$$

In *MRMC sample size input1.sas* the power is computed by computing the variance component estimates (3) from the inputted mean squares and reader and case sample sizes for the pilot data analysis, followed by computing equations (8) and (9) for various values of *r* and *c*. *MRMC sample size input2.sas* uses the same algorithm but does not compute the variance components since they are inputted.

We note that Hillis and Berbaum [5] do not discuss power for the fixed readers and random cases situation.

4.3 The DBM Model: random readers and fixed cases

The model is the same as in Section 4.1, except that now $\sigma_{\tau C}^2 = \sigma_C^2 = 0$. The appropriate test statistic, as discussed in References [2,4], for testing for a treatment effect is given by

$$F = \frac{\mathrm{MS(T)}}{\mathrm{MS(T*R)}} \tag{10}$$

The null hypothesis is rejected if $F > F_{\alpha;t-1,(t-1)(r-1)}$, where α is the significance level.

Using an approach similar to that used in the previous section, the power for comparing two treatments is approximated by

$$\text{power} \approx \Pr\left(F_{1,r-1;\hat{\Delta}} > F_{1-\alpha;1,r-1}\right) \tag{11}$$

where the noncentrality parameter estimate is given by

$$\hat{\Delta} = \frac{\left[\mathrm{AUC}_1 - \mathrm{AUC}_2\right]^2}{\frac{2}{rc}\left(c\hat{\sigma}_{\tau R}^2 + \hat{\sigma}^2\right)}. \tag{12}$$

In *MRMC sample size input1.sas* the power is computed by computing the variance component estimates (3) from the inputted mean squares and reader and case sample sizes for the pilot data analysis, followed by computing equations (11) and (12) for various values of *r* and *c*. *MRMC sample size input2.sas* uses the same algorithm but does not compute the variance components since they are inputted.

We note that Hillis and Berbaum [5] did not discuss power for the random readers and fixed cases situation.

4.4 Theoretical justification

Throughout we have been computing the power as though the DBM model is acceptable. However, we have pointed out [2,4] that it is not an acceptable model because the pseudovalues have no intrinsic interpretation, and furthermore, they are not independent or normally distributed. At the same time, though, we have also pointed out

[2,4] that it serves as a "working" model that can be justified by the conceptually and theoretically acceptable model proposed by Obuchowski and Rockette [6] which provides identical hypotheses tests and confidence intervals. We are presently writing a paper that shows how all of these sample size formulas can be derived from the Obuchowski and Rockette model.

References

1. Dorfman DD, Berbaum KS, Metz CE. Receiver operating characteristic rating analysis: generalization to the population of readers and patients with the jackknife method. Investigative Radiology 1992; 27: 723-731.
2. Hillis SL, Obuchowski NA, Schartz KM, Berbaum KS. A comparison of the Dorfman-Berbaum-Metz and Obuchowski-Rockette Methods for receiver operating characteristic (ROC) data. Statistics in Medicine 2005;24:1579-1607.
3. Hillis SL, Berbaum KS. Monte Carlo validation of the Dorfman-Berbaum-Metz method using normalized pseudovalues and less data-based model simplification. Academic Radiology 2005;12:1534-1542.
4. Hillis SL. A comparison of denominator degrees of freedom methods for multiple observer ROC analysis. Statistics in Medicine 2006; in press.
5. Hillis SL, Berbaum KS. Power estimation for the Dorfman-Berbaum-Metz method. Academic Radiology 2004;11:1260-1273.
6. Obuchowski NA, Rockette HE. Hypothesis testing of the diagnostic accuracy for multiple diagnostic tests: an ANOVA approach with dependent observations. Communications in Statistics: Simulation and Computation 1995; 24:285-308.