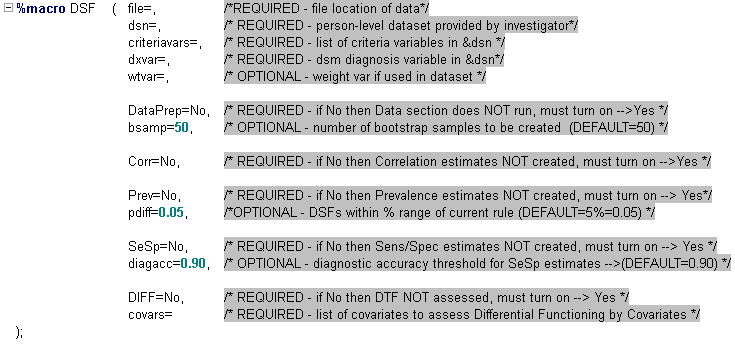
**Introduction**

Lengthy criteria sets can be problematic and provide the primary motivation to identify short-forms. This paper explores how custom SAS macro, ***%DSFv2***, can be used to systematically identify subsets and associated cut-offs that yield diagnoses as similar as possible to using the original full-length criteria set. Using DSM-5 alcohol use disorder (AUD) as our example with 11 diagnostic criteria and the current rule of 2+ criteria endorsed for an AUD diagnosis, the macro identifies diagnostic short-forms by: (1) maximizing the association between the sum scores of all 11 criteria with newly constructed subscales from subsets of criteria, (2) optimizing the similarity of AUD prevalence between the current DSM-5 rule and newly constructed diagnostic short-forms, (3) maximizing sensitivity and specificity of the short-forms against the current DSM-5 rule, and (4) minimizing differences in the accuracy of the short-form across chosen covariates. At the end of the macro, a final list of optimal diagnostic short-form alternatives are provided with comparative information for decision making.

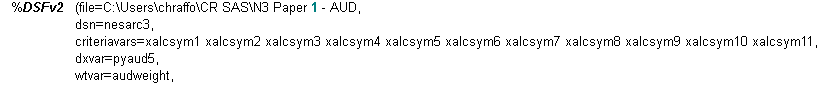
**Macro Call**



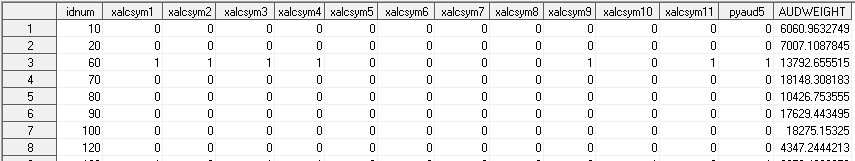
***[file=, dsn=, criteriavars=, dxvar=]***

The first 4 macro calls (file=, dsn=, criteriavars=, dxvar=) are all REQUIRED and the first two identify the file location on your computer (file=) where the users dataset is found (dsn=). Then, you must list out the names of the criteria with a single space in between (criteriavars=var1 var2) and identify the name of the current diagnostic rule (dxvar=). If the data has a weight variable then you can identify the name of the var (wtvar=) but is OPTIONAL.

The example below provides the file location of the nesarc3.sas7bdat dataset with the list of 11 diagnostic criteria used for the diagnosis of AUD. Also, the variable name for the current rule to diagnosis AUD (dxvar=pyaud5) and the weight var.



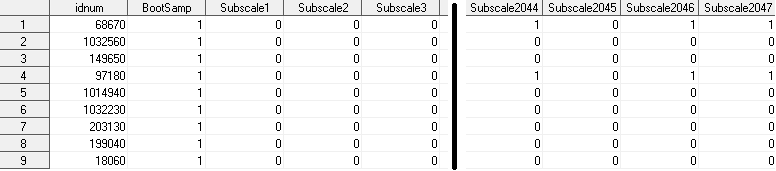
A snapshot example of the person-level dataset is below:



***[DataPrep=No, bsamp=50]***

The rest of the macro call mimic the steps as outlined in the introduction (i.e., correlation step, prevalence step). The first macro call (DataPrep=No) needs to be switch on (DataPrep=Yes) to create bootstrap data off of the dataset (&dsn) provided by the user. The default has 50 bootstrap samples (bsamp=50) but can be adjusted depending on the user.

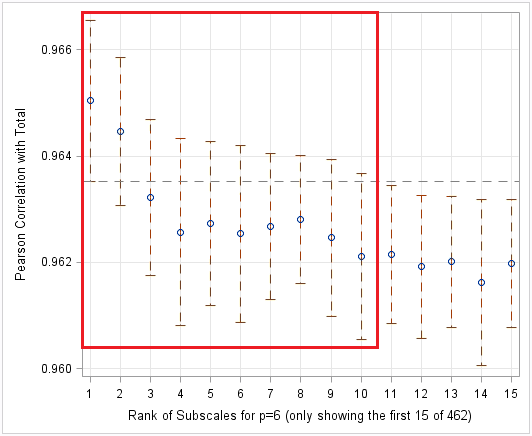
When there are 11 criteria, there are 2,047 possible subscales that can be constructed of size p=1 to 11. Below provides an example of the bootstrapped data (BootSamp=1 to 50) with variables for sum count of criteria (Subscale) for each of the 2,047 possible subsets.



***[Corr=No]***

The next macro call (Corr=No) needs to be switched on (Corr=Yes) for the first screening step of the algorithm to run. This uses the bootstrapped data created previously and creates all possible combinations of subsets of criteria, obtains sum scores for each subset to create subscales and correlates each subscale with the current diagnosis (as specified with dxvar=). Once all estimates of correlation are calculated, 95% confidence intervals are estimated and for each size of subscale (p=1, 2, ... total number of criteria), the top ranked subscale and any not significantly different are selected to move to the next step.

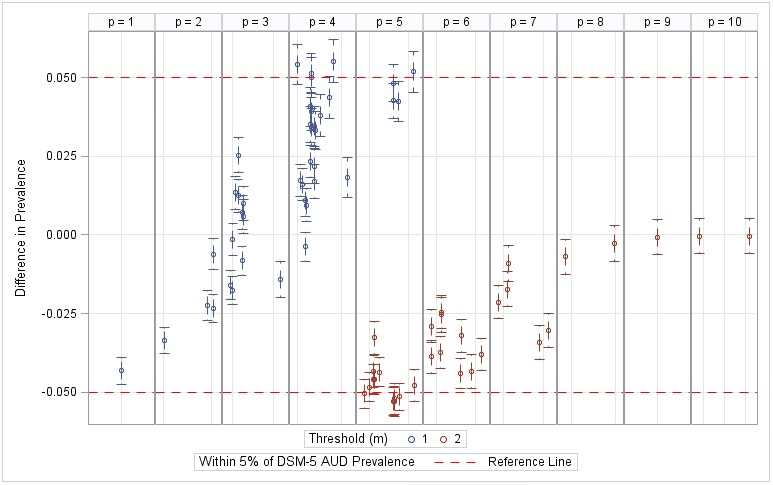
Below provides an example of the selection of subscales with six criteria included (p=6). This step takes the top subscale for size p=6 and any not significantly different. Among the 462 possible subscales of this size, this step takes the 10 top performing subscales to move to the next step.



***[Prev=No, pdiff=0.05]***

This macro call is the second screening step in the algorithm. It takes the optimal subscales selected in the previous section and incorporates the threshold component for diagnosing disorder. Specifically, for each size of the subscale (p), we let the threshold (m) vary across all possible sizes. If there are four criteria included in a subscale (p=4), then four diagnostic short-forms can be constructed where the threshold ranges from m=1 to 4. For the first diagnostic short-form, only 1 criterion needs to be endorsed for a diagnosis made but for the last one, all four criteria must be endorsed for a diagnosis. This is extrapolated to all subscales. Once all possible diagnostic short-forms are constructed, the prevalence of disorder is estimated in the sample and those that are within a percentage of the current rule prevalence are selected to move to the next step. The percentage difference is set at 5% (default) but can be adjusted depending on the user.

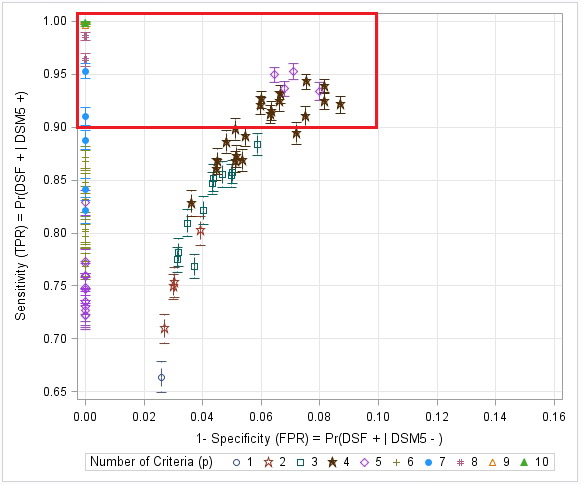
The figure below provides the difference in prevalence between DSM-5 AUD and diagnostic short-forms where values greater than zero indicate DSF prevalence is higher than DSM-5 AUD prevalence and values less than zero indicate DSF prevalence is lower than DSM-5 AUD. Those falling with 5% of the DSM-5 AUD prevalence (inside the red reference lines) move to the next step.

******

***[SeSp=No, diagacc=0.90]***

This is the second to last screening step and uses the current rule (dxvar=) to estimate sensitivity and specificity for each diagnostic short-form. The user can set the threshold for the diagnostic short-forms that make it to the next step (diagacc=) which is currently set to only take those that have Sensitivity/Specificity estimates (and 95% confidence intervals) ≥ 0.90.

The figure below provides the sensitivity vs. 1-specificity and 95% confidence intervals around sensitivity among the diagnostic short-forms that made it to this screening step. The ones that fall above the threshold set (diagacc=0.90) move to the last step. In this example, the red reference line indicates the diagnostic short-forms that have sensitivity and specificity estimates ≥ 0.90.

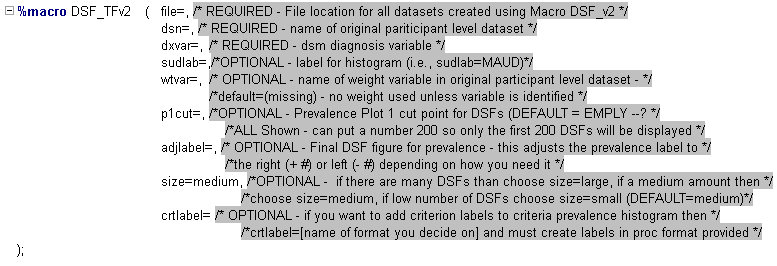
******

***[DIFF=No, covars=]***

The last screening step, requires the user to impute the variable names of the covariates used to assess differential test functioning (DTF) for each of the remaining diagnostic short-forms. That is, it should not exhibit systematic differences in the way it measures the target construct in different subgroups than what is currently being captured by the full-length rule. In our setting, we are concerned with reducing the systematic differences with the entire diagnostic short-form rather than with each item within the short-form.

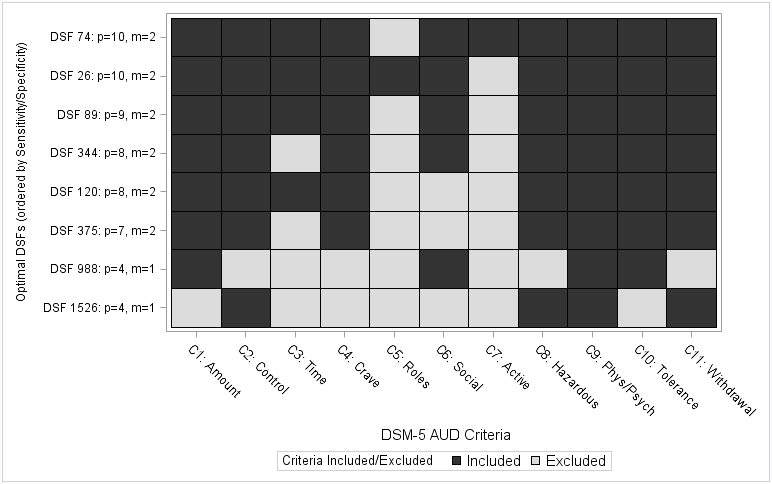
In our example, we assess whether the diagnostic short-forms that made it through all the previous steps have any differential functioning by the covariates listed (covars=). However, currently the macro can only handle binary covariates.

At the end of the algorithm, a final list of optimal diagnostic short-forms are provided. Each step in the process creates a new dataset saved to your file location which are called upon in the tables and figures macro (***%DSF\_TFv2***) to create the output for the user. See example below:



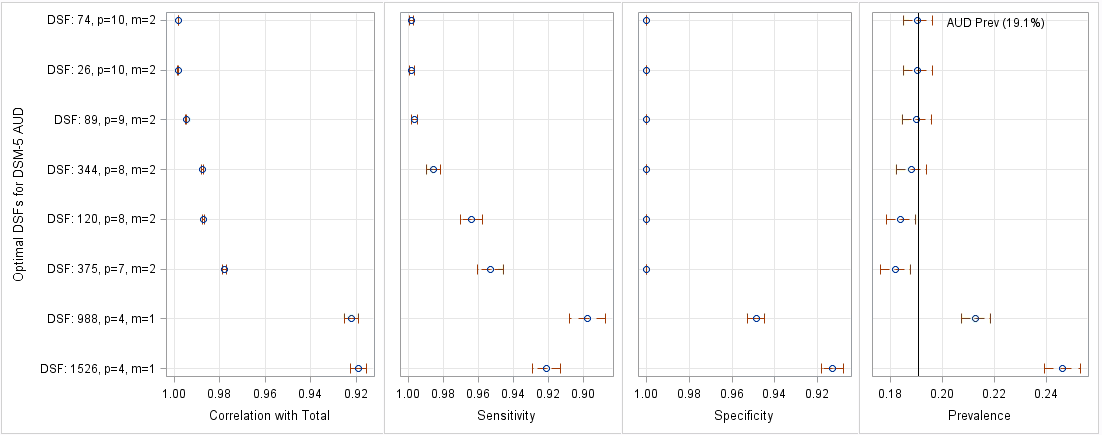
**Example of Tables & Figures produced by macro:**

Criteria Included/Excluded in the 8 Optimal Diagnostic Short-forms for DSM-5 AUD\* - Step 8



Descriptive Statistics on Optimal Diagnostic Short-Forms for DSM-5 AUD

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **DSF label** |  |  | **Correlation with Total** | **Prevalence1** | **Sensitivity** | **Specificity** |
| 74 | 10 | 2 | 0.998 (0.998, 0.998) | 0.191 (0.185, 0.196) | 0.998 (0.997, 0.999) | 1.000 (1.000, 1.000) |
| 26 | 10 | 2 | 0.998 (0.998, 0.998) | 0.191 (0.185, 0.196) | 0.998 (0.997, 1.000) | 1.000 (1.000, 1.000) |
| 89 | 9 | 2 | 0.995 (0.994, 0.995) | 0.190 (0.185, 0.196) | 0.996 (0.995, 0.998) | 1.000 (1.000, 1.000) |
| 344 | 8 | 2 | 0.988 (0.987, 0.988) | 0.188 (0.183, 0.194) | 0.986 (0.982, 0.990) | 1.000 (1.000, 1.000) |
| 120 | 8 | 2 | 0.987 (0.987, 0.988) | 0.184 (0.178, 0.190) | 0.964 (0.957, 0.970) | 1.000 (1.000, 1.000) |
| 375 | 7 | 2 | 0.978 (0.977, 0.979) | 0.182 (0.176, 0.188) | 0.953 (0.946, 0.960) | 1.000 (1.000, 1.000) |
| 988 | 4 | 1 | 0.922 (0.919, 0.925) | 0.213 (0.207, 0.218) | 0.898 (0.887, 0.908) | 0.949 (0.945, 0.953) |
| 1526 | 4 | 1 | 0.919 (0.916, 0.922) | 0.246 (0.239, 0.253) | 0.921 (0.913, 0.929) | 0.913 (0.908, 0.918) |

Descriptive Statistics on Optimal Diagnostic Short-Forms for DSM-5 AUD****

**Information**

Developer: Cheri Raffo

Location: Brooklyn, NY

Date: March 2018

There are two macros available for download:

1. ***%DSF\_v2***
   1. This macro runs through each step as described in this article and provides a final dataset with the optimal diagnostic short-forms for the specified data set.
2. ***%DSFTF\_v2***
   1. This macro uses the datasets created in the previous macro to output tables and figures used to investigate findings

Email [cheriraffo@gmail.com](mailto:cheriraffo@gmail.com) for downloadable versions.