Paper SM03

# Group, Collapse, Repeat: A Dynamic SAS Macro for Stratification Pooling in Oncology Efficacy Analysis

Venkatesh Burla, Merck & Co., Inc., Rahway, NJ, USA
Lulu Wang, Merck & Co., Inc., Rahway, NJ, USA

## ABSTRACT

Stratified randomization is a common design employed in clinical trials, used to ensure balance across key prognostic factors such as disease risk categories and geographic region. These stratification variables are often carried forward into the efficacy analysis of endpoints like progression-free survival (PFS), overall survival (OS), and objective response rate (ORR). However, sparse events count within certain strata can compromise model stability and interpretability. To address this challenge, we developed a dynamic and reusable SAS macro that evaluates events that counts across predefined strata and applies a data-driven pooling strategy when minimum thresholds are not met. The collapsing sequence is customizable and typically proceeds by removing region, combining IMDC (International Metastatic Renal Cell Carcinoma Database Consortium) categories, and collapsing sarcomatoid status. As efficacy endpoints are usually derived post-ADSL, a standalone stratification dataset (ADSTRATA) is generated. This paper describes the implementation and use of this macro in an RCC (Renal Cell Carcinoma) study and its broader applicability across clinical trials.

## INTRODUCTION

Stratified randomization is a common technique in clinical trials to ensure treatment assignment is balanced across important prognostic factors, such as disease stage, region or biomarker status. These stratification factors are pre-specified in the protocol and used during randomization to mitigate the impact of confounding variables on treatment comparisons.

In time-to-event and response-based efficacy analysis such as progression-free survival (PFS), overall survival (OS), or objective response rate (ORR), it is often necessary to conduct stratified analysis (e.g., using stratified log-rank tests or CMH methods). However, the true values needed for these stratification factors-particularly when pooling strata based on actual event counts – may only become available after deriving endpoint-specific ADaM datasets like ADTTE (for PFS/OS) or ADRS (for ORR).

Therefore, it is not feasible to derive these dynamically pooled stratification variables in ADSL, since endpoint event flags and response assessments do not exist at that stage. To address this, a dedicated dataset, ADSTRATA, is created to capture only the pooled stratification variables used in the final efficacy analysis, based on verified values derived from downstream ADaM datasets such as ADTTE and ADRS. This enables traceability, flexibility, and dynamic derivation of strata based on actual analysis populations and observed events, ensuring that statistical analysis reflects protocol-defined stratification while aligning with the realities of the observed data.

## POOLED STRATIFICATION ANALYSIS DATASET (ADSTRATA)

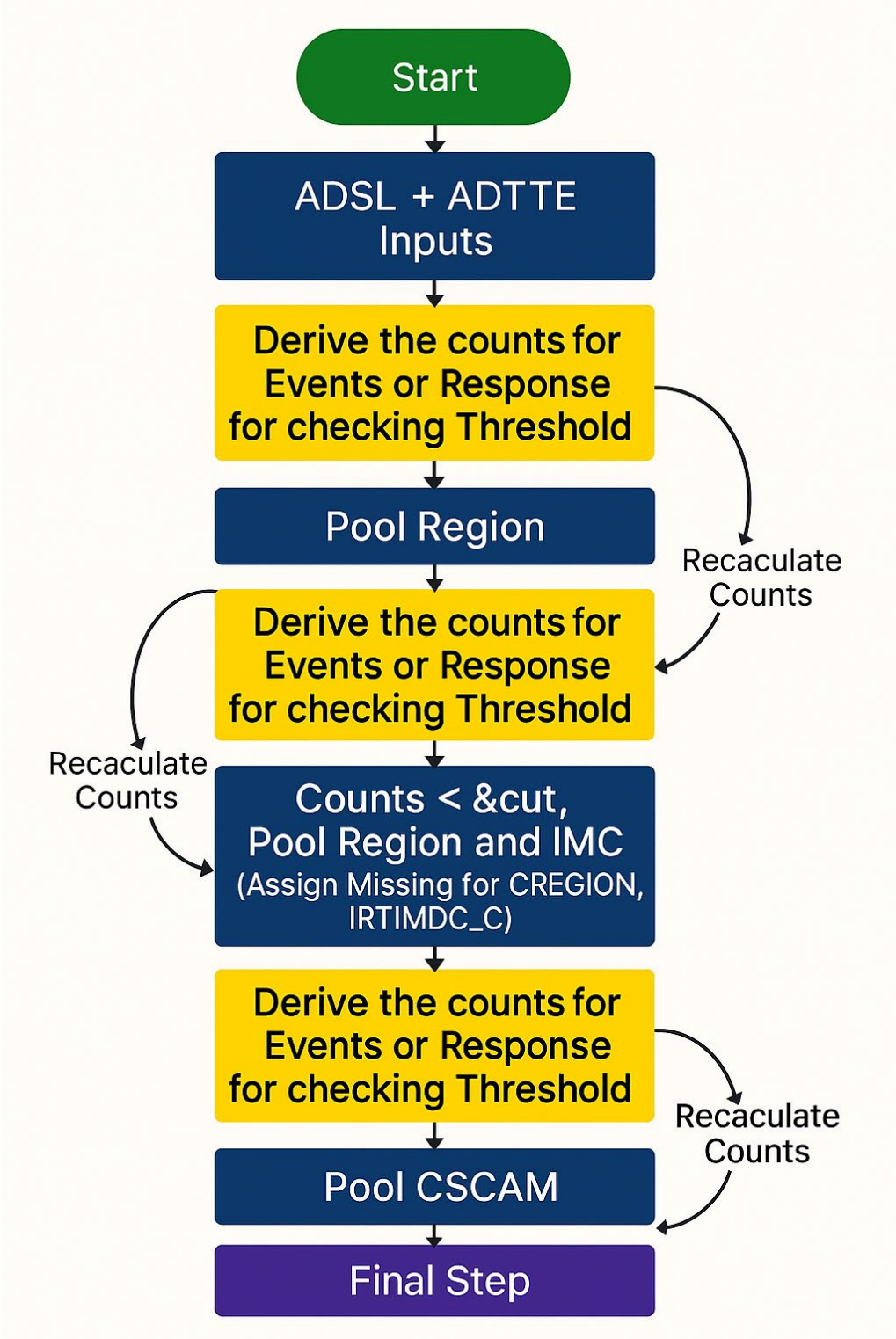
The ADSTRATA dataset contains pooled stratification variables used for efficacy analysis across endpoints such as progression-free survival (PFS), Overall Survival (OS), and Objective Response Rate (ORR). Stratification factors – Sarcomatoid – sarcomatoid status (csam), IMDC risk group (IRTIMDC), and region (CREGION) are initially derived from ADSL and dynamically evaluated based on the number of observed events in ADTTE. When event counts within a stratum fall below a predefined threshold (e.g., fewer than 5 events), strata are sequentially pooled in a predefined order: region, IMDC category, and then sarcomatoid status. The dataset includes both descriptive (STRATPx) and numeric (STRATAPxN) variables representing the final pooled strata.

This dataset is created using a single macro (%ADSTRATA) that calls two utility macros - %get_group_sheet and %my_pooling.

**%GET_GROUP_SHEET:**

This macro extracts the stratification factors and corresponding event counts by combining information from the ADSL and ADTTE datasets. For time-to-event endpoints such as PFSIRC, and OS, the macro filters for non-censored subjects (cnsr=0) to ensure that strata are evaluated based on the observed events only. For response-based endpoints such as ORR, the macro summarizes counts directly from ADSL. The output dataset, a "&group_sheet" provides a stratification-level summary of events across combinations of CSCAM, IRTIMDC, and CREGION. The sample (draft) datasets illustrating the counts before and after pooling are provided in **Appedix I** for reference and the SAS code in **Appendix II** provides more information.

**%MY_POOLING:**
This macro performs the core logic of the sequential pooling and creation of the final analysis stratification variables. It takes the event summary dataset ("&group_sheet") as the input and iteratively evaluates whether the number of events in each stratum falls below a user-defined threshold (&cut). When low-count strata are detected, the macro combines the stratification factors following a defined sequence – first by collapsing by Region (CREGION), then IMDC risk group (IRTIMDC), and finally Sarcomatoid status (CSCAM). Each iteration reassesses the strata after pooling until all strata meet the stability criterion. The detailed steps of this process are described below.

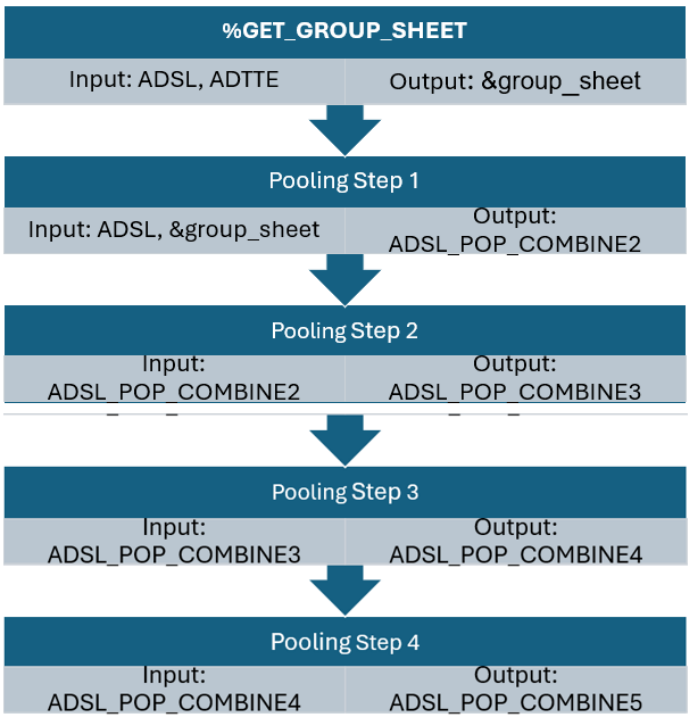


**STEP 1:**
In this step, the program joins the subject-level population dataset **(ADSL_POP)** with the previously generated event summary (GROUP_SHEET1) to bring the event counts for each stratification combination. It then flags strata where the number of events falls below the predefined threshold (&cut.). Next, it calculates the total number of flagged strata within each Sarcomatoid (CSCAM) and IMDC combination. If any low-count strata are detected, the corresponding Region (CREGION) variable is assigned to be missing, effectively pooling across regions. The code snippet in **Appendix III** provides more information.

**STEP 2:**
In this step, strata are further collapsed based on IMDC risk groups once region pooling is complete. The code first merges the previously pooled dataset in step 1 (ADSL_POP_COMBINE2) with the updated event summary (GROUP_SHEET2) to reassess event counts after region pooling. It flags any strata where event counts remain below the specified threshold (&cut) and creates a combined IMDC category (Favorable/Intermediate) to merge those two risk groups when needed. Next, it evaluates the number of flagged strata within each CSCAM and IMDC combination. If low-count strata persist, both Region (CREGION) and IMDC (IRTIMDC) values are updated to reflect the pooled categories. The code snippet in **Appendix IV** provides more information.

**STEP 3:**
In this step, If there are still small strata across treatment arms after previous pooling steps, it completely drops the IMDC grouping to get the adequate event counts. The code merges the latest subject-level dataset (ADSL_POP_COMBINE3) with the updated event summary (GROUP_SHEET3) to reassess the event and flag strata that still below the threshold (&cut.). If present, then both CREGION and IRTIMDC variables are reset to missing. The code snippet in **Appendix V** provides more information

**STEP 4:**
In this step, if there are still small strata across treatment arms even after region and IMDC-based pooling, the final pooling step completely removes theCSCAM grouping. The code begins by merging the intermediate dataset (ADSL_POP_COMBINE4) with the event summary (GROUP_SHEET4) to recheck event counts at the CSCAM-IMDC-CREGION group. Strata below the defined threshold (&cut) are flagged. Next, cumulative flags are computed within each group. Finally, if any low low-count strata remain, the code sets REGION, IRTIMDC, CSCAM to missing effectively collapsing all strata into a single combined group. The code snippet in **Appendix VI** provides more information



**FINAL STEP – CONSTRUCTING THE POOLING STATIFICATION VARIABLE**

In the final step, the pooled stratification variables are formatted into a single, descriptive character variable that represents the final analysis strata variables. The data step creates a new variable (&pooling_var = strata1) that concatenates the contributing stratification factors – Sarcomatoid status (CSCAM), IMDC risk group (IRTIMDC), and Region (CREGION) into a single variable. The code snippet in **Appendix VII** provides more information

For each subject, the macro dynamically builds this pooled description using the CATX function, including only non-missing components. For example, if a subject belongs to Sarcomatoid: No, IMDC: Favorable/Intermediate, and Region: Rest of the World, the final pooled variable will read: The **Appendix VIII** provides the final dataset snippet.

**"Sarcomatoid: No, IMDC: Favorable/Intermediate, Region: Rest of the World"**

The macro also supports multiple efficacy endpoints (PFSIRC, OS, and ORR) and population subsets (e.g., Arm A+C, Arm B+C, and specific IMDC subgroups). Separate thresholds (e.g., cut=5 for time-to-event endpoints, cut=10 for binary endpoints) are applied dynamically depending on the endpoint type.

In studies with multiple treatment arms, the macro is designed to dynamically accommodate pairwise and overall comparisons within a single execution. For Instance, in a three-arm study set up (Arm A, Arm B, and Arm C), the macro automatically handles two pairwise comparisons – Arm A vs Arm B and Arm A vs Arm C – along with the overall population analysis. Depending on the comparison type, the macro generates a corresponding set of stratification variables: for example, six pooled stratification variables for a pairwise comparison and nine for the overall analysis. This dynamic structure allows the same macro to be applied consistently across multiple efficacy endpoints and comparison groups without additional changes.

Another important functionality of the macro is its automated export process. Once each endpoint's analysis is complete, the macro exports the results into Excel workbooks, where each worksheet corresponds to a specific endpoint-comparator combination. For example:

- orr0trtavstrtb
- orr0trtavstrtc
- pfs0trtavstrtb
- pfs0trtavstrtb

The export worksheet names are dynamically generated based on endpoint, treatment comparator values.

## CONCLUSION

The development of the ADSTRATA dataset and its supporting pooling macro provides a standardized and automated approach to handling stratification factors in efficacy analysis. By dynamically evaluating event counts and applying pooling rules across endpoints like PFS, OS, and ORR, the macro ensures statistical robustness while maintaining traceability. This approach eliminates repetitive manual programming, reduces potential for error, and enhances consistency across studies. Overall, the ADSTRATA macro supports efficient implementation of stratified analysis in Oncology trials and can be easily adapted for use in other therapeutic areas.

## REFERENCES

References go at the end of your paper. This section is not required.


## ACKNOWLEDGMENTS

The author would like to thank Amy, Sai Govind Chenna, Himanshu Patel, Dileep Penmetsa, and Li Liu for reviewing the paper and providing great suggestions.


## CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:


Venkatesh Burla
Merck & Co., Inc., Rahway, NJ, USA
215-631-5295
Email: venkatesh.burla@merck.com
Lulu Wang
Merck & Co., Inc., Rahway, NJ, USA
267-305-6745
Email: lulu.wang2@merck.com

**APPENDIX I:**

**Event counts before pooling for Event based endpoints (PFS/OS):**

| CSAM | IRTIMDC | CREGION | EVENT_COUNT |
|---|---|---|---|
| No | Favorable | North America/Western Europe | 12 |
| No | Favorable | Rest of the World | 3 |
| No | Intermediate | North America/Western Europe | 18 |
| No | Intermediate | Rest of the World | 10 |
| No | Poor | North America/Western Europe | 6 |
| No | Poor | Rest of the World | 2 |
| Yes | Favorable | North America/Western Europe | 8 |
| Yes | Favorable | Rest of the World | 4 |
| Yes | Intermediate | North America/Western Europe | 20 |
| Yes | Intermediate | Rest of the World | 15 |
| Yes | Poor | North America/Western Europe | 7 |
| Yes | Poor | Rest of the World | 5 |

**Event counts after pooling for Event based endpoints (PFS/OS):**

| CSAM | IRTIMDC | CREGION | EVENT_COUNT |
|---|---|---|---|
| No | Favorable | | 15 |
| No | Intermediate | North America/Western Europe | 18 |
| No | Intermediate | Rest of the World | 10 |
| No | Poor | | 8 |
| Yes | Favorable | | 12 |
| Yes | Intermediate | North America/Western Europe | 20 |
| Yes | Intermediate | Rest of the World | 15 |
| Yes | Poor | | 12 |

**Event counts before pooling for Response based endpoints (ORR):**

| CSAM | IRTIMDC | CREGION | EVENT_COUNT |
|---|---|---|---|
| No | Favorable | North America/Western Europe | 9 |
| No | Favorable | Rest of the World | 7 |
| No | Intermediate | North America/Western Europe | 25 |
| No | Intermediate | Rest of the World | 18 |
| No | Poor | North America/Western Europe | 6 |
| No | Poor | Rest of the World | 5 |
| Yes | Favorable | North America/Western Europe | 8 |
| Yes | Favorable | Rest of the World | 6 |
| Yes | Intermediate | North America/Western Europe | 20 |
| Yes | Intermediate | Rest of the World | 14 |
| Yes | Poor | North America/Western Europe | 7 |
| Yes | Poor | Rest of the World | 4 |

**Event counts after pooling for Response based endpoints (ORR):**

| CSAM | IRTIMDC | CREGION | EVENT_COUNT |
|---|---|---|---|
| No | Favorable | | 16 |
| No | Intermediate | North America/Western Europe | 25 |
| No | Intermediate | Rest of the World | 18 |
| No | Poor | | 11 |
| Yes | Favorable | | 14 |
| Yes | Intermediate | North America/Western Europe | 20 |
| Yes | Intermediate | Rest of the World | 14 |
| Yes | Poor | | 11 |

**APPENDIX II:**

| SAS SCRIPT | FUNCTION |
|---|---|
| %if &_paramcd = "PFSIRC" \| &_paramcd = "OS"<br>%then %do;<br>proc sql;<br>create table _merged as<br>select a.CNSR, b.CSCAM, b.IRTIMDC, b.CREGION<br>from &_adtte as a<br>left join &_adsl as b<br>on a.usubjid = b.usubjid<br>where a.paramcd = &_paramcd & CNSR = **0**;<br>quit;<br><br>proc sql;<br>create table &group_sheet as<br>select CSCAM, IRTIMDC, CREGION,<br>count(*) as events_count<br>from _merged<br>group by CSCAM, IRTIMDC, CREGION;<br>quit;<br>%end; | To check the number of events for PFS, OS by CSCAM, IRTIMDC, and CREGION |
| %else %do;<br>proc sql;<br>create table &group_sheet as<br>select CSCAM, IRTIMDC, CREGION,<br>count(*) as events_count<br>from &_adsl<br>group by CSCAM, IRTIMDC, CREGION;<br>quit;<br>%end; | To check the number of events for other than PFS, OS by CSCAM, IRTIMDC, and CREGION |

**APPENDIX III:**

| SAS SCRIPT | FUNCTION |
|---|---|
| proc sql;<br>create table adsl_pop_combine as<br>select usubjid, a.cscam, a.irtimdc, a.cregion,<br>events_count,<br>case<br>when events_count < &cut then 1<br>else 0<br>end as flag<br>from &adsl_pop as a<br>left join group_sheet1 as b<br>on a.CSCAM = b.CSCAM<br>and a.IRTIMDC = b.IRTIMDC<br>and a.CREGION = b.CREGION;<br>quit; | Flag the events counts less than the required threshold value by required stratification variables.<br><br>1: Data with events less than threshold<br>0: Data greater than events greater than threshold |
| proc sql;<br>create table adsl_pop_combine2 as<br>select *, sum(flag) as flag_c<br>from adsl_pop_combine<br>group by CSCAM, IRTIMDC;<br>quit; | Checking the data by CSCAM, and IRTIMDC, If the value of flag_c is not 0, then CREGION will be combined |
| proc sql;<br>update adsl_pop_combine2 | Assigning the missing value for Region (CREGION) for those data with events less than threshold |

| | |
|---|---|
| set cregion = ""<br>where flag_c >= 1;<br>quit; | |

**APPENDIX IV:**

| SAS SCRIPT | FUNCTION |
|---|---|
| proc sql;<br>create table adsl_pop_combine as<br>select usubjid, a.cscam, a.irtimdc, a.cregion,<br>b.events_count,<br>case<br>when b.events_count < &cut then **1**<br>else **0**<br>end as flag,<br>case<br>when a.irtimdc = "Intermediate" \| a.irtimdc = "Favorable" then "Favorable/Intermediate"<br>else a.irtimdc<br>end as irtimdc_c<br>from adsl_pop_combine2 as a<br>left join group_sheet2 as b<br>on a.CSCAM = b.CSCAM<br>and a.IRTIMDC = b.IRTIMDC<br>and a.CREGION = b.CREGION;<br>quit; | If there are still small strata (<5 OS/PFS events across 2 arms, <10 participants for ORR across 2 arms), completely drop the factor of Regions of the World and combine and Favorable with Intermediate.<br><br>Flag the events counts less than the required threshold value by required stratification variables.<br><br>1: Data with events less than threshold<br>0: Data greater than events greater than threshold |
| proc sql;<br>create table adsl_pop_combine3 as<br>select *, sum(flag) as flag_c<br>from adsl_pop_combine<br>group by CSCAM, irtimdc_c;<br>quit; | if flag_c is not 0, then CREGION will be dropped and IRTIMDC will be combined |
| proc sql;<br>update adsl_pop_combine3<br>set cregion = "", irtimdc = irtimdc_c<br>where flag_c >= **1**;<br>quit; | Assigning the missing value for Region (CREGION) for those data with events less than threshold and combining the values of IRTIMDC (Fav/Inter) |

**APPENDIX V:**

| SAS SCRIPT | FUNCTION |
|---|---|
| proc sql;<br>create table adsl_pop_combine as<br>select usubjid, a.cscam, a.irtimdc, a.cregion,<br>b.events_count,<br>case<br>when b.events_count < &cut then 1<br>else 0<br>end as flag,<br>"" as irtimdc_c<br>from adsl_pop_combine3 as a<br>left join group_sheet3 as b<br>on a.CSCAM = b.CSCAM<br>and a.IRTIMDC = b.IRTIMDC<br>and a.CREGION = b.CREGION;<br>quit; | If there are still small strata (<5 OS/PFS events across 2 arms, <10 participants for ORR across 2 arms), completely drop the IRTIMDC<br><br>Flag the events counts less than the required threshold value by required stratification variables.<br><br>1: Data with events less than threshold<br>0: Data greater than events greater than threshold |
| proc sql;<br>create table adsl_pop_combine4 as<br>select *, sum(flag) as flag_c | if flag_c is not 0, then CREGION and irtimdc_c will be dropped |

| | |
|---|---|
| from adsl_pop_combine<br>group by CSCAM, irtimdc_c;<br>quit; | |
| proc sql;<br>update adsl_pop_combine4<br>set cregion = "", irtimdc = irtimdc_c<br>where flag_c >= 1;<br>quit; | Assigning the missing value for CREGION and assigning the value of irtimdc_c (missing) to the IRTIMDC |

**APPENDIX VI:**

| SAS SCRIPT | FUNCTION |
|---|---|
| proc sql;<br>create table adsl_pop_combine as<br>select usubjid, a.cscam, a.irtimdc, a.cregion,<br>b.events_count,<br>case<br>when b.events_count < &cut then 1<br>else 0<br>end as flag,<br>"" as irtimdc_c,<br>"" as cscam_c<br>from adsl_pop_combine4 as a<br>left join group_sheet4 as b<br>on a.CSCAM = b.CSCAM<br>and a.IRTIMDC = b.IRTIMDC<br>and a.CREGION = b.CREGION;<br>quit; | If there are still small strata (<5 OS/PFS events across 2 arms, <10 participants for ORR across 2 arms), completely drop CSCAM<br><br>Flag the events counts less than the required threshold value by required stratification variables.<br><br>1: Data with events less than threshold<br>0: Data greater than events greater than threshold |
| proc sql;<br>create table adsl_pop_combine5 as<br>select *, sum(flag) as flag_c<br>from adsl_pop_combine<br>group by cscam_c, irtimdc_c;<br>quit; | if flag_c is not 0, then drop the CSCAM |
| proc sql;<br>update adsl_pop_combine5<br>set cregion = "", irtimdc = irtimdc_c, cscam = cscam_c<br>where flag_c >= 1;<br>quit; | Assigning the missing value for CREGION, set IRTIMDC_C (missing) to IRTIMDC, and CSCAM = CSCAM_C (Missing) |

**Appendix VII**

| SAS SCRIPT | FUNCTION |
|---|---|
| data strchar;<br>set adsl_pop_combine5;<br>length &pooling_var $**200**;<br>&pooling_var = '';<br>if CSCAM ne "" then &pooling_var = catx(', ',<br>&pooling_var, 'Sarcomatoid: ' \|\| strip(CSCAM));<br>if IRTIMDC ne "" then &pooling_var = catx(', ',<br>&pooling_var, 'IMDC: ' \|\| strip(IRTIMDC));<br>if CREGION ne "" then &pooling_var = catx(', ',<br>&pooling_var, 'Region: ' \|\| strip(CREGION));<br>keep usubjid &pooling_var CSCAM IRTIMDC<br>CREGION;<br>run; | In the final step, creating the pooling variable |

| | |
|---|---|

**Appendix VIII**

| USUBJID | CSCAM | IRTIMDC | CREGION | STRATP1 | STRATP2 | STRATP3 |
|---|---|---|---|---|---|---|
| 1001 | No | Favorable | North America/Western Europe | Sarcomatoid: No, IMDC: Favorable, Region: Combined | Sarcomatoid: No, IMDC: Favorable | Sarcomatoid: No |
| 1002 | Yes | Intermediate | Rest of the World | Sarcomatoid: Yes, IMDC: Intermediate, Region: Combined | Sarcomatoid: Yes, IMDC: Intermediate | Sarcomatoid: Yes |
| 1003 | No | Poor | North America/Western Europe | Sarcomatoid: No, IMDC: Poor, Region: Combined | Sarcomatoid: No, IMDC: Poor | Sarcomatoid: No |
| 1004 | Yes | Favorable | North America/Western Europe | Sarcomatoid: Yes, IMDC: Favorable/Intermediate | Sarcomatoid: Yes, IMDC: Favorable/Intermediate | Sarcomatoid: Yes |
| 1005 | No | Poor | Rest of the World | Sarcomatoid: No, IMDC: Poor, Region: Combined | Sarcomatoid: No, IMDC: Poor | Sarcomatoid: No |