Paper RE04

GSK Diversity Paper: The Challenges We Faced with Real World Data

Carletha Blanding, Upper Providence, PA, USA

Chris Leach, UK

Harrison James, UK

Abstract

In early 2021, Pfizer published a paper titled "Demographic Diversity of Participants in Pfizer Sponsored Clinical Trials in the United States", which issued a challenge to other Pharma companies to analyse and publish their data. This paper will document how GSK's Stats and Programming considered their approach and implemented the integration of 930 trials, conducted between 2002-2019. This was extended to include census and epidemiology data. The paper will cover how the data was combined, and the complex tables and figures used to summarize the data. In summary, it will also cover the challenges faced and lessons learned.

Purpose

There has been a recent industry wide push to enhance equity in clinical trials. The FDA has taken recent steps to increase racial and ethnic diversity in clinical trials by issuing a draft guidance for sponsor enrolment expectations, so that all races and ethnicities can benefit from biomedical research [1]. In response to this, the pharmaceutical industry is having discussions on how to properly address this inequity in clinical trial populations by beginning to delve into their past. Pfizer was the first pharmaceutical company to publish their diversity data in 2021 and summarised data for 213 clinical trials conducted between 2011 and 2020 encompassing over 103k subjects. Their analysis was selective, and more information can be obtained by referencing their paper [2]. Most importantly, this paper presented an industry wide challenge for other big pharma companies to release their diversity data, to start the conversation of how clinical trials can be more equitable. GSK heard the challenge and is looking to exceed the challenge by summarizing data for 495 US trials which included 108,261 participants and produced a series of complex tables and figures to summarize this data. In this paper we will discuss how we were able to summarize this data across company and therapeutic areas (TAs), how we evaluated GSK trial recruitment against US Census and epidemiologic data, some of the challenges we faced along the way, and how we incorporated modern ways of working with R output data.

Study Logistics

The initial challenge with a task such as this is the logistical planning stage. During this stage, consideration should be taken to understand things such as:

- Beginning with the end in mind (i.e., what is the team vision)
- Is there time available early on to develop mature ideas from ALL stakeholders before work begins?
- Has the project scope been clearly defined and how will the team be prepared to handle in-stream changes?
- What data sources are expected to be used in the analysis?
- What format will the data be coming in (i.e., SAS, CSV, TXT) and will it be CDISC compliant?
- What region specific encoding will the data use?
- Where will the data be stored so that the team may retrieve it?
- How will missing data be handled?

Additionally, for our GSK diversity paper, we wanted to include data from three sources: Pharma, Vaccines, & ViiV (HIV) studies. Since GSK is an aggregated entity, these sources used different systems and/or data structures. This required additional planning to ensure we would be able to integrate the data correctly.

It is best practice to map out answers to as many questions as possible early on so that the entirety of the task is visible and can be appropriately planned for. Figure 1 and Table 1 explain the final logistics of our study, which will be explained in detail.

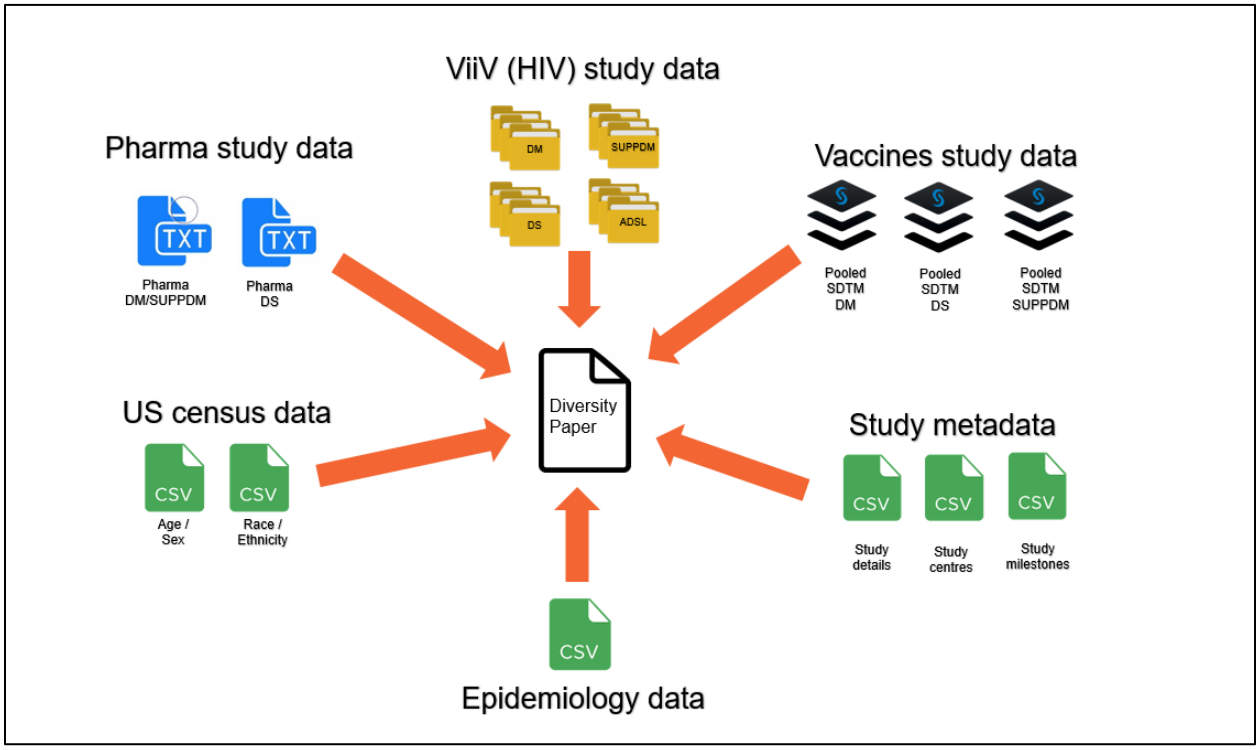


Figure 1: Study Logistics Roadmap

Pharma study data
Data Format: TXT & CSV Data Source: Single pooled DM and DS datasets
ViiV study data[1]
Data Format: SAS datasets Data Source: Programmer Specific Transfer Folder > Compound Folder > Study Folder > CDISC DM, SUPPDM, DS, ADSL datasets
Vaccines study data
Data Format: SAS datasets Data Source: Pooled DM and DS datasets
Study metadata
Data Format: CSV Data Source: Single files for study details, centres, and milestones
US Census Data
Data Format: CSV Data Source: Single files for race/ethnicity and sex/age group
Epidemiology Data
Data Format: CSV Data Source: Single file

Table 1: Data format and source information

[1] ViiV Healthcare is a specialist pharmaceutical company 100% dedicated to HIV medicines and research and focused on people living with HIV and AIDS, born out of a partnership between GSK and Pfizer in 2009, with Shionogi joining in 2012.

Study Data Logistics

As shown in Figure 1, the team was responsible for utilizing study data and metadata from many locations and in many different formats. The next step for the team, as seen in Figure 2, was to map how each data source was expected to be used:

- rawdata: contained original data sent to the programming team
- refdata: contained data converted to SAS datasets by the programming team
- SDTM: contained SDTM-like SAS datasets integrated by the programming team
- ADaM: contained ADaM-like SAS datasets created by the programming team

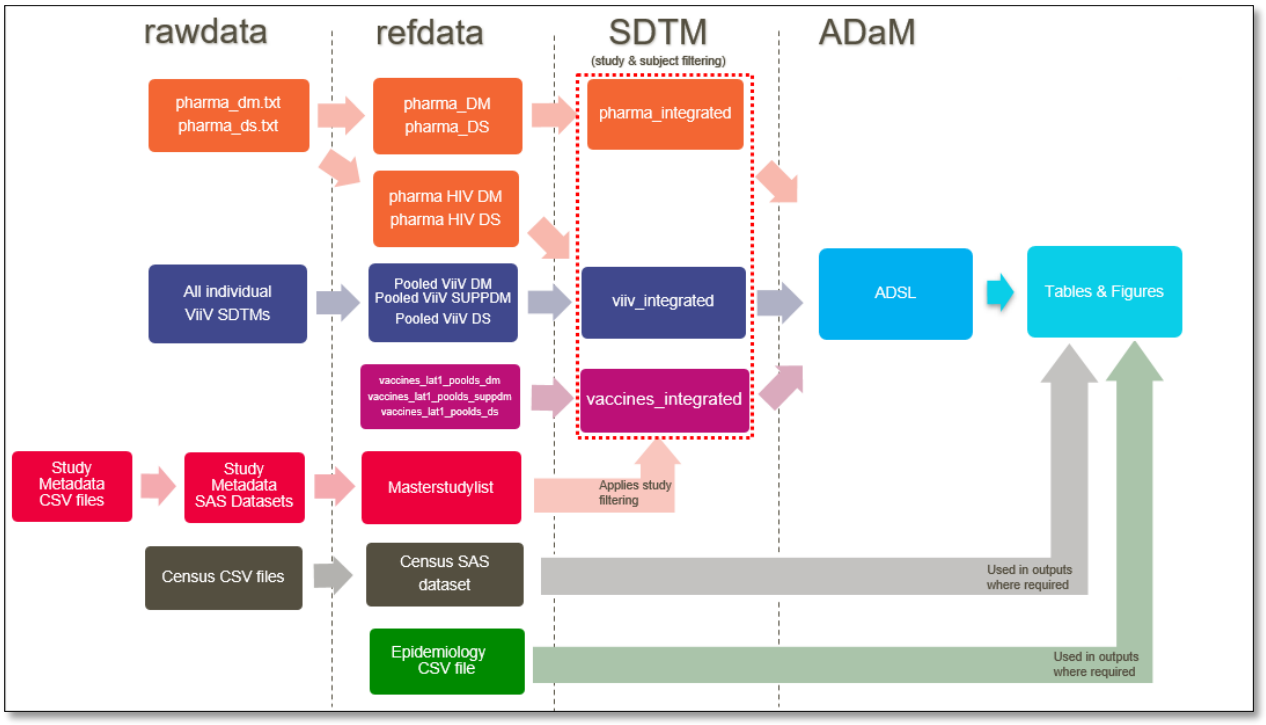


Figure 2: Study Data Logistics

Study Data Analysis

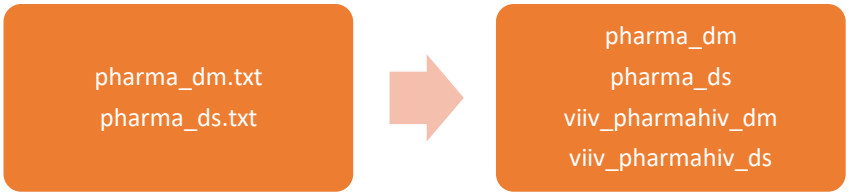
After receiving all the data from multiple sources, our next task was to collate it into a single dataset. This was done in stages, starting with converting the raw study data into individual datasets for Pharma, Vaccines and ViiV (HIV). These datasets were then ready to be combined into one final analysis dataset which would be used for all outputs. Using a single analysis dataset would mimic a typical study, meaning summaries could be created a lot simpler.

Reference Data

The study data provided included studies from 2002 to 2019, all of which used CDISC standards. Older studies, prior to CDISC being introduced, had been remapped to CDISC standards prior to producing the diversity paper. However, due to different systems being used, we received the data in different formats. These were converted to SAS datasets, where required, preparing them for integration.

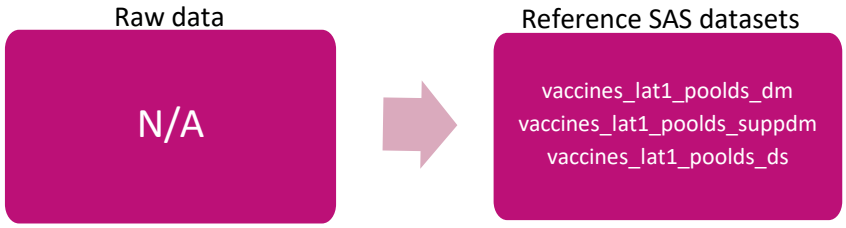
- **Pharma** – Two large text files were provided, one for demography data, and one for disposition data and these files were read into SAS. However, there was an added complication of HIV studies which occurred prior to ViiV studies being included in the Pharma data. So that we could include all HIV studies under one group, and to simplify later programming, all HIV studies were separated from non-HIV studies.

Raw data Reference SAS datasets

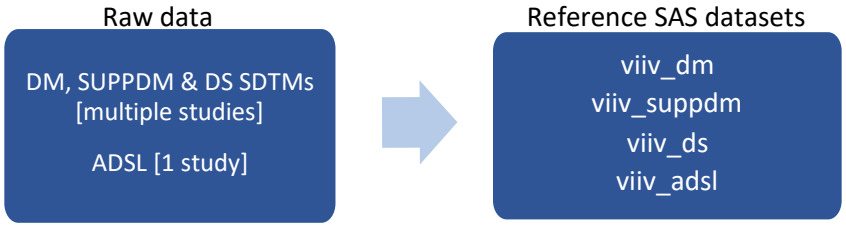


Extra caution should be taken when importing large text files. It may look as though the import was successful, but issues can exist further down in the file that can be easily overlooked. We independently validated all reference data imports, which allowed us to identify issues that would not have been visible with manual qc of such large files. We recommend performing independent QC for all large data imports.

- **Vaccines** – Data was provided as pooled SAS datasets, meaning no conversion was required at this stage.



- **ViiV (HIV)** – Multiple SDTMs for each ViiV study were provided. This meant that each individual study had to be given its own folder for compound and study. We created a program which recursively searched all the ViiV study folders and combined all the individual DM, SUPPDM and DS datasets into one dataset for each domain. One ADSL dataset was provided for one ViiV study, adding an extra complication. This ADSL dataset was moved to the reference datasets and was renamed at the same time.



We were provided with multiple CSV files that contained study metadata, study details, centre details and study milestones for GSK studies. In total, these files included the metadata for over 9000 studies. This metadata was read into SAS and was used to create a final reference dataset (i.e., “master study list”), that only included studies we should include in the diversity paper. The studies included met the following criteria:

- Pharma, Vaccines, and ViiV data
- GSK sponsored interventional trial data
- Trials that were completed and archived
- Non-extension studies

This final master study list contained the metadata for 3435 studies, and it was used to apply filtering criteria when combined with all the study data. Figure 3 shows a simple example of what the master study list contained.

studyid	indications	company	phase	study_type	therapy_area	siteid	country_name
200056	HIV Infections	ViiV	PHASE IIB	Intv-Clinical Trial	Infectious Diseases	207072	United States
200056	HIV Infections	ViiV	PHASE IIB	Intv-Clinical Trial	Infectious Diseases	209055	France
200056	HIV Infections	ViiV	PHASE IIB	Intv-Clinical Trial	Infectious Diseases	209057	France
200056	HIV Infections	ViiV	PHASE IIB	Intv-Clinical Trial	Infectious Diseases	211333	France
200096	Cardiovascular Disease	GSK Pharma	PHASE I	Intv-Clinical Trial	Cardiovascular Diseases	216919	India
200096	Cardiovascular Disease	GSK Pharma	PHASE I	Intv-Clinical Trial	Cardiovascular Diseases	217720	India
200147	Chickenpox	GSK Vaccines	PHASE IIIA	Intv-Clinical Trial	Vaccines Prophylactic	218098	United Kingdom
200147	Chickenpox	GSK Vaccines	PHASE IIIA	Intv-Clinical Trial	Vaccines Prophylactic	218104	United Kingdom
200147	Chickenpox	GSK Vaccines	PHASE IIIA	Intv-Clinical Trial	Vaccines Prophylactic	218146	Mexico

Figure 3: Example of master study list

In addition to incorporating study metadata, 2019 US census data [3] was also included as reference data, to be used in the tables and figures. Unfortunately, the census CSV files were in a format that could not be easily included in summaries, so the team encountered the following issues with the CSV files:

- There were two separate CSV files, one for race and ethnicity, and one for age and sex and they needed to be combined into one usable census file
- The files were not fully machine-readable due to including titles, blank cells, and subheadings
- The file included census data that spanned many years, and the team needed to retrieve only the year of interest

Below, Figure 4 and Figure 5, show an example of how each census CSV file was originally formatted.

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	table with row headers in column A and column headers in rows 3 through 4. (leading dots indicate sub-parts)												
2	Annual Estimates of the Resident Population by Sex, Race, and Hispanic Origin for the United States: April 1, 2010 to July 1, 2019												
3	Sex, Race, and Hispanic Origin	April 1, 2010	Population Estimate (as of July 1)										
4		Census	Estimates Base	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019
5	TOTAL POPULATION	308,745,538	308,758,105	309,321,666	311,556,874	313,830,990	315,993,715	318,301,008	320,635,163	322,941,311	324,985,539	326,687,501	328,239,523
6	.One Race:												
7	..White	241,937,061	241,945,431	242,235,328	243,292,145	244,354,778	245,308,889	246,350,641	247,382,690	248,413,058	249,270,773	249,961,025	250,522,190
8	..Black or African American	40,250,635	40,254,450	40,355,385	40,783,547	41,225,339	41,651,863	42,090,120	42,532,491	42,970,183	43,374,142	43,732,024	44,075,086
9	..American Indian and Alaska Native	3,739,506	3,739,607	3,752,274	3,801,451	3,853,222	3,902,766	3,953,354	4,004,358	4,054,740	4,101,605	4,145,811	4,188,092
10	..Asian	15,159,516	15,159,643	15,261,408	15,721,072	16,195,092	16,679,752	17,203,569	17,752,744	18,279,949	18,764,237	19,134,105	19,504,862
11	..Native Hawaiian and Other Pacific Islander	674,625	674,644	678,046	692,216	706,660	721,575	735,273	750,159	765,814	780,316	793,787	806,937
12	.Two or More Races	6,984,195	6,984,330	7,039,225	7,266,443	7,495,899	7,728,870	7,968,051	8,212,721	8,457,567	8,694,466	8,920,749	9,142,356

Figure 4: Snapshot of race and ethnicity census CSV file

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	table with row headers in column A and column headers in rows 3 through 5. (leading dots indicate sub-parts)												
2	Annual Estimates of the Resident Population for Selected Age Groups by Sex for the United States: April 1, 2010 to July 1, 2019												
3	Age	April 1, 2010						Population Estimate (as of July 1)					
4		Census			Estimates Base			2010			2011		
5		Both Sexes	Male	Female	Both Sexes	Male	Female	Both Sexes	Male	Female	Both Sexes	Male	Female
6	Total	308,745,538	151,781,326	156,964,212	308,758,105	151,788,777	156,969,328	309,321,666	152,074,758	157,246,908	311,556,874	153,200,058	158,356,816
7	.Under 5 years	20,201,362	10,319,427	9,881,935	20,201,426	10,319,452	9,881,974	20,188,815	10,312,617	9,876,198	20,123,103	10,279,719	9,843,384
8	.5 to 9 years	20,348,657	10,389,638	9,959,019	20,348,744	10,389,675	9,959,069	20,331,228	10,380,298	9,950,930	20,332,498	10,381,356	9,951,142
9	.10 to 14 years	20,677,194	10,579,862	10,097,332	20,677,377	10,579,975	10,097,402	20,680,606	10,578,206	10,102,400	20,712,798	10,586,779	10,126,019
10	.15 to 19 years	22,040,343	11,303,666	10,736,677	22,041,547	11,304,511	10,737,036	21,981,099	11,277,894	10,703,205	21,658,481	11,119,469	10,539,012
11	.20 to 24 years	21,585,999	11,014,176	10,571,823	21,587,573	11,015,251	10,572,322	21,701,298	11,072,082	10,629,216	22,159,198	11,303,591	10,855,607

Figure 5: Snapshot of sex and age census CSV file

These CSV files were combined, and the required 2019 US census data was extracted into a single SAS dataset. This dataset provides the expected populations for the US for each category of ethnicity, race, sex, and age group. These expected populations were used in tables and figures to analyse the

difference between the census data and subjects in GSK studies. This can be seen in Figure 6. The 2019 census was used due to it being the latest year that had complete data available.

	Category	Value	Count	Total Population	Percentage
1	Ethnicity	HISPANIC	60572237	328239523	18.453669578
2	Ethnicity	NOT HISPANIC	267667286	328239523	81.546330422
3	Ethnicity	NOT HISPANIC - AMERICAN INDIAN AND ALASKA NATIVE	2434908	328239523	0.7418082922
4	Ethnicity	NOT HISPANIC - ASIAN	18905879	328239523	5.7597814021
5	Ethnicity	NOT HISPANIC - BLACK OR AFRICAN AMERICAN	41147488	328239523	12.535811539
6	Ethnicity	NOT HISPANIC - NATIVE HAWAIIAN AND OTHER PACIFIC ISLANDER	595908	328239523	0.1815466933
7	Ethnicity	NOT HISPANIC - TWO OR MORE RACES	7273281	328239523	2.2158455915
8	Ethnicity	NOT HISPANIC - WHITE	197309822	328239523	60.111536903
9	Race	AMERICAN INDIAN AND ALASKA NATIVE	4188092	328239523	1.2759255685
10	Race	ASIAN	19504862	328239523	5.9422649112
11	Race	BLACK OR AFRICAN AMERICAN	44075086	328239523	13.427720586
12	Race	NATIVE HAWAIIAN AND OTHER PACIFIC ISLANDER	806937	328239523	0.2458378542
13	Race	TWO OR MORE RACES	9142356	328239523	2.7852697068
14	Race	WHITE	250522190	328239523	76.322981374
15	Sex	M	161657324	328239523	49.249804692
16	Sex	F	166582199	328239523	50.750195308
17	Age	UNDER 18 YEARS	73039150	328239523	22.251784103
18	Age	UNDER 5 YEARS	19576683	328239523	5.9641455791
19	Age	5 TO 13 YEARS	36829704	328239523	11.22037458
20	Age	14 TO 17 YEARS	16632763	328239523	5.0672639443
21	Age	18 TO 64 YEARS	201142110	328239523	61.2790648
22	Age	18 TO 24 YEARS	30219206	328239523	9.2064495231
23	Age	25 TO 44 YEARS	87599465	328239523	26.687665215
24	Age	45 TO 64 YEARS	83323439	328239523	25.384950063
25	Age	65 YEARS AND OVER	54058263	328239523	16.469151096

Figure 6: Reference SAS dataset of 2019 census data

In addition to the census data, epidemiology data was also needed for the real-world comparison summary figures. This data was sent to the team in a simple CSV file format and easily stored with the other reference datasets.

Indication	Race_ethnicity	Percent
Therapeutic Area 1	White	72.85%
Therapeutic Area 1	Black or African American	17.02%
Therapeutic Area 1	Asian	2.97%
Therapeutic Area 1	American Indian or Alaskan Native	1.37%
Therapeutic Area 1	Native Hawaiian or Other Pacific Islander	-
Therapeutic Area 1	Mixed Race	5.80%
Therapeutic Area 1	Hispanic or Latino	14.37%
Therapeutic Area 2	White	85.66%
Therapeutic Area 2	Black or African American	7.10%
Therapeutic Area 2	Asian	2.14%
Therapeutic Area 2	American Indian or Alaskan Native	-
Therapeutic Area 2	Native Hawaiian or Other Pacific Islander	-
Therapeutic Area 2	Mixed Race	5.10%
Therapeutic Area 2	Hispanic or Latino	6.45%

Table 2: Example of epidemiology CSV file

Please note, for the purpose of this paper we have removed the TA names from the epidemiology data. We did not want to reveal the analysis results of the GSK diversity paper. Once the GSK diversity paper is published you will be able to see the analysis and results which are discussed throughout this paper. All epidemiology data was collected from documentation available on the Centers for Disease Control and Prevention (CDC) website [4].

Integrated Data (SDTM)

Creating collated, TA specific, SAS reference datasets to be used as a source for the integrated datasets, meant integration was a lot simpler. We then used the master study list dataset to subsequently filter the data, keeping only studies in Pharma, Vaccines, and ViiV, that were also included in the list. We also applied additional subject level filtering at this stage to keep subject data that met pre-defined criteria, which included:

- Excluding studies that did not collect both race and ethnicity data
- Excluding subjects who did not pass screening (Disposition (DS) reference data was used to apply this filtering)
- Excluding subjects who were not assigned treatment

The master study list was also used to retain study information so that we could easily group subjects in tables and figures. The following data was merged to the subject records, by study id and site id (see Figure 3 for variables):

- Country
- Therapeutic area category (Pharma, Vaccines, or ViiV)
- Therapeutic area
- Indication
- Study phase

To gain efficiencies while working with large datasets, a separate SAS program was created to handle each TA category independently. This allowed programs to run much quicker and permitted the dataset specific downstream tasks to proceed without interfering with the other TAs. This allowed the team to troubleshoot and investigate without having to rerun the entire integration when an issue was found. Figure 7 shows the process of integrating the master study list with the reference study data, along with the number of subjects we had before and after integration.

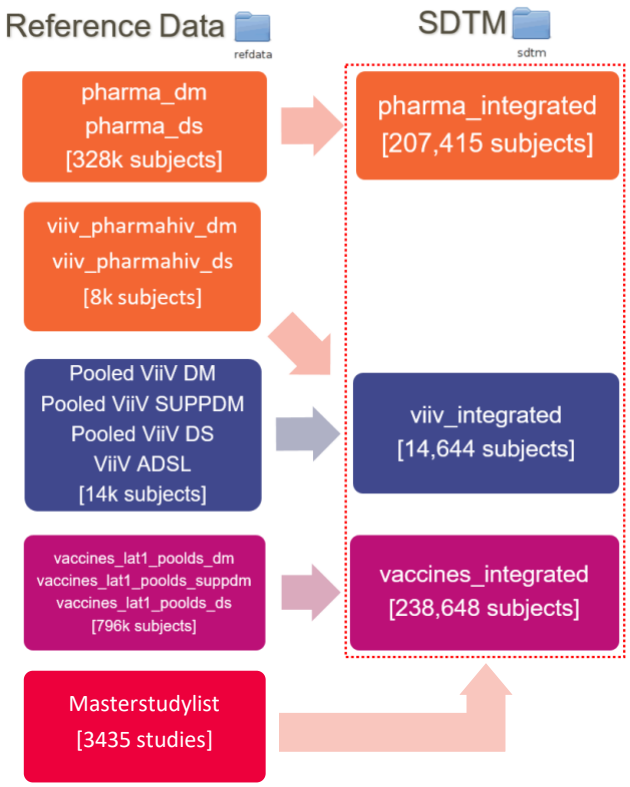


Figure 7: Flow diagram showing integration of study data with master study list

We refer to these datasets as SDTM, as they were the source of our final analysis dataset (ADSL). Clearly these datasets are not SDTM domains and do not follow SDTM standards, meaning no CDISC requirements were necessary. This meant we did not create dataset metadata or define-XML, and CDISC validation was not required.

The integration was independently validated by comparing the integrated datasets created by the statistical programmers against the data scientist's R output CSV files. The data scientist was also producing the same tables and figures we were creating. To add an additional step of validation, we agreed to compare our data at this stage, with data created in R. This additional validation step meant we could resolve data issues much earlier in the pipeline, which accelerated the review process of the tables and figures considerably. Some example of issues that were found are explained in the Major Impacts to Programming using Real-World Data section.

It is important to perform this validation as early as possible, so that you have time to discuss the differences and resolve issues on both sides before you are too far into the analysis. Working with so much data, across many years, means it will be highly unlikely that there is an exact match immediately. Most of our mismatches resulted in updates to our filtering criteria to clarify things and make them more robust. In one example, we removed extension studies by using the filter, extension_study = "No". After further review, it was apparent that some studies had this information set to missing, instead of "No". Therefore, we agreed to update the filtering to, extension_study ≠ "Yes".

Final Analysis Data (ADSL)

To facilitate the production of tables and figures a single dataset was created which contained all the required demographic data. On a CDISC study this would be the ADSL dataset, so we decided to use the

same naming convention. Figure 8 shows our process of combining the integrated datasets, along with the number of subjects we had before and after.

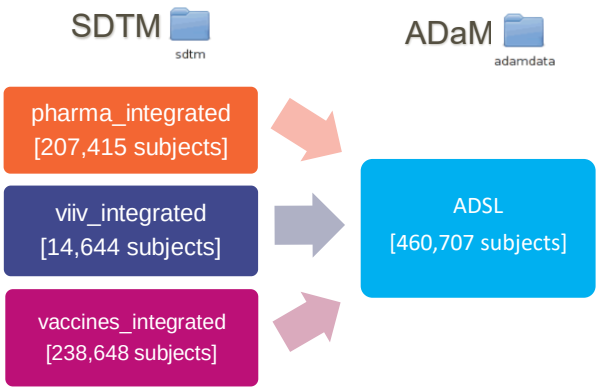


Figure 8: Flow diagram showing creation of final analysis dataset

All three integrated datasets were combined, and some additional processing was performed. This included:

- Deriving analysis age variable (AAGE) – this was derived for two reasons:
    1. Any ages recorded using units other than "YEARS", were recalculated in years.
    2. Some studies had AGE missing, even though birth date (BRTHDTC) and reference dates (RFSTDTC, RFICDTC) were provided. In these cases, age was derived using the standard age formula. Reference start date (RFSTDTC) was used if available, otherwise informed consent date (RFICDTC) was used. Age remained missing if any of the reference dates were missing. Any partial birth dates were imputed using the following rules:
        - Year recorded – assigned YYYY-06-30.
        - Year and month recorded – assigned YYYY-MM-15.
- Assigned age groups (AGEGR1) – age groups were assigned based on their derived analysis age (AAGE). Any subjects who could not be assigned AAGE, were assigned an age group of "NA".
- Assigned numeric sorting variables (e.g., RACEN, AGEGR1N, etc.) – used to sort summaries.
- Assigned a flag for "Non-Hispanic White" subjects based on RACE and ETHNICITY – subjects who had RACE = "WHITE" and ETHNICITY = "HISPANIC OR LATINO" were flagged. This was included to simplify the summary of this race-ethnicity combination against the census values.
- Updated TA for "Infectious Diseases" – all ViiV studies had TA assigned as "Infectious Diseases". The team also encountered an issue with additional Infectious Diseases data being included in our Pharma data source. It was necessary to move the HIV studies from the Pharma data into the ViiV data during the integration (SDTM) stage, so that we could easily relabel these TAs based on the data source. This would avoid any confusion in the outputs. Table 3 shows how we relabelled these TAs.

Group	Data Source	Original Therapeutic Area	New Therapeutic Area
Pharma	pharma_integrated.sas7bdat	Infectious Diseases	Infectious Diseases (Non-HIV)
ViiV	viiv_integrated.sas7bdat	Infectious Diseases	Infectious Diseases (HIV)

Table 3: TA label updates for Pharma and ViiV data

As mentioned previously, the naming convention “ADSL” was used for this final dataset, but we have not followed any strict CDISC standards. No define-XML was produced for the purpose of this paper and Pinnacle21 was not used to validate the final dataset or metadata.

Major Impacts to Programming using Real-World Data

While we were reviewing the integrated data, we noticed some irregularities. These were checked programmatically and after further investigation, it was decided that additional mapping had to be performed during the integration stage. We would recommend an independent validation program is created so that counts for each variable can be assessed. Counting non-preferred values such as “UNKNOWN”, “OTHER”, “NA”, etc., by subgroup and/or study, will indicate if there are any patterns emerging in the data.

Example #1

It was discovered that the race category “OTHER”, had much higher counts in vaccine studies than any other TA category. Looking into the pooled vaccines DM data, it was clear that the race and ethnicity data was not recorded in the same way as Pharma and ViiV. In many cases RACE was set to “OTHER”, and RACEOTH in SUPPDM contained the subject’s race information. We agreed to remap race using the RACEOTH, where possible. However, since RACEOTH is a free text field on a CRF, the values recorded had many variations. We had to consider different formats, abbreviated text, and typos. To perform this mapping, we used regular expressions to look for specific patterns, meaning we could identify mixed race subjects easily. We also looked for specific values for single race categories. For example, some subjects had RACEOTH recorded as standard code list values, so these cases were mapped directly to RACE. Anything that could not be clearly identified, was left as “OTHER”. Tables 4 shows an example of some of these free text values, and what they were remapped to.

	DM.RACE	SUPPDM.RACEOTH	Mapped Race
1	OTHER	BLACK AND WHITE	MIXED RACE
2	OTHER	AFRICAN AM/WHITE	MIXED RACE
3	OTHER	1/2BLACK,1/2WHITE	MIXED RACE
4	OTHER	CAUCASIAN/BLACK	MIXED RACE
5	OTHER	BLACK,CAUCASIAN	MIXED RACE
6	OTHER	MULTI RACIAL /MOM WHITE /DAD BLACK	MIXED RACE
7	OTHER	WHITE AND BLACK	MIXED RACE
8	OTHER	AFRICAN AMERICAN/CAUCASIAN	MIXED RACE
9	OTHER	MOTHER - AFRICAN AMERICAN,FATHER - WHITE	MIXED RACE
10	OTHER	BIRACIAL; WHITE AND AFRICAN AMERICAN	MIXED RACE
11	OTHER	MIXED: AFRICAN AMERICAN / CAUCASIAN	MIXED RACE
12	OTHER	AFRICAN AMERICAN & CAUCASIAN EUROP.HERIT	MIXED RACE
13	OTHER	MOM WHITE/DAD BLACK	MIXED RACE
14	OTHER	BLACK/BIRACIAL	MIXED RACE
15	OTHER	HALF WHITE HALF BLACK	MIXED RACE
16	OTHER	MIXED AFRICAN AMERICAN AND CAUCASION	MIXED RACE
17	OTHER	AFRICAN AMERICAN AND WHITE	MIXED RACE
18	OTHER	AFRICAN AMERICAN/CAUCASIAN	MIXED RACE
19	OTHER	BI-RACIAL/WHITE-CAUC & AFRICAN AMERICAN	MIXED RACE

20	OTHER	50%BLACK AND 50%CAUC	MIXED RACE
21	OTHER	AMERICAN	OTHER
22	OTHER	ASIAN	ASIAN
23	OTHER	UNKNOWN	NA
24	OTHER	EAST ASIAN/SOUTH EAST ASIAN	ASIAN
25	OTHER	ASIAN-EAST ASIAN/ASIAN-SOUTH EAST ASIAN	ASIAN
26	OTHER	DECLINE TO STATE	NA
27	OTHER	WHITE	WHITE

Table 4: Mapping of RACEOTH to RACE

Furthermore, some vaccine studies had subjects that had race recorded as "OTHER" and were missing ethnicity. In some cases, the RACEOTH value was recorded as "HISPANIC". For these exact cases we assigned their ethnicity, but we did not have enough information to remap their race. Table 5 shows how race and ethnicity were mapped in these cases.

	DM.RACE	DM.ETHNICITY	SUPPDM.RACEOTH	Mapped Race	Mapped Ethnicity
1	OTHER	null	HISPANIC	OTHER	HISPANIC OR LATINO

Table 5: Mapping of ethnicity based on RACEOTH

Example #2

There were very few cases of sex category "UNKNOWN", except for one study which had a very large percentage of subjects with unknown sex. These "UNKNOWN" values were mapped to "NA" for summary purposes. However, Table 6 shows that out of 42 subjects, 14 of the subjects (33%) had sex unknown.

Sex	Total (N=42)
n	42
F	20 (48%)
M	8 (19%)
NA	14 (33%)

Table 6: Sex summary for one irregular study

This was significantly different from all other studies. In fact, excluding this study, there was only one subject across all the US studies that had sex recorded as "UNKNOWN". Clearly there was something different occurring in this study. After further review, it was discovered that the study was an obstetric labour study, which included the unborn foetus as a subject. After the baby was born, they had a new record assigned, where full demographic data was recorded. An example of this can be seen in Table 7.

SUBJECT	USUBJID	BRTHDTC	AGE	AGEU	SEX	RACE	ETHNIC
Mother	111111-000001-M	1994	21	YEARS	F	WHITE	NOT HISPANIC OR LATINO
Foetus	111111-000001-F	null	null	null	U	null	null
New-born	111111-000001-N	2015-06-01	60	DAYS	M	WHITE	NOT HISPANIC OR LATINO

Table 7: Example of how data was recorded for obstetric labour study

This meant the child subject was recorded twice. We removed subjects with the SUBJID suffix "-F", and this eliminated the issue. This issue, which was not obvious when all the data was combined, indicated that extra attention should be given to paediatric studies.

Summary Tables

Multiple tables were created to summarise the combined ADSL data. This included simple summaries of study and subject counts, along with complex summaries of the demographic characteristics.

Summary of Studies

We summarised the number of studies and subjects per TA category, for Pharma, Vaccines, and ViiV studies. Two tables were produced, one summary for global studies, and one summary which only included US studies. Figure 9 shows how many studies and subjects were included when we filtered the data to only include US studies (i.e., every study in ADSL where COUNTRY = "USA").

	Number of Subjects [1]	% of Subjects	Number of Clinical Trials [1]	% of Clinical Trials
Pharma Therapy Area				
Autoimmune	0	0.00%	0	0.00%
Cardiovascular Diseases	9188	15.30%	52	14.57%
Dermatology	53	0.09%	4	1.12%
Gastrointestinal Diseases	245	0.41%	6	1.68%
Infectious Diseases (Non-HIV)	3161	5.26%	33	9.24%
Inflammation	534	0.89%	8	2.24%
Metabolic Diseases	4322	7.20%	43	12.04%
Musculoskeletal Diseases	567	0.94%	10	2.80%
Neurosciences	5687	9.47%	45	12.61%
Oncology	418	0.70%	14	3.92%
Ophthalmology	225	0.37%	2	0.56%
Other	28	0.05%	1	0.28%
Respiratory	34533	57.52%	121	33.89%
Urology	1077	1.79%	18	5.04%
Pharma Overall	60038	100.00%	357	100.00%
ViiV Therapy Area				
Infectious Diseases (HIV)	5720	100.00%	93	100.00%
Vaccines Therapy Area				
Vaccines Prophylactic	42448	99.87%	40	88.89%
Vaccines Therapeutic	55	0.13%	5	11.11%
Vaccines Overall	42503	100.00%	45	100.00%
Overall Therapy Area				
Overall	108261	100.00%	495	100.00%

[1] The number of subjects and number of clinical trials are based on the total number of subjects and clinical trials within each therapeutic area category.

Figure 9: Summary of US studies

Some important points to consider when creating these summaries:

- The percentages should be calculated based on the total counts within each of the TA categories.
- Review the formatting of the percentages when the table is drafted. Due to working with so much data, we had to display our percentages to 2 decimal places to avoid percentages being incorrectly displayed as "0.0%".
- Make your program as efficient as possible, to decrease run times when working with large data. E.g., in these tables we did not need any demographic variables, therefore it was more efficient to drop these variables prior to sorting and counting.

Summary of Demographic Characteristics

The demography data was summarised two ways, one table summarised demography by subjects and the other table by studies. Both tables calculated a summary for each TA, including a total summary for Pharma and Vaccines categories. ViiV did not need a total summary as all studies in this category were HIV.

Summary of US Subjects

This summary was created in a similar fashion to a normal study. We calculated summary statistics for age, and we calculated frequency counts for age group, sex, race, and ethnicity. Figure 10 shows our results for Pharma overall.

	Pharma (N=60038)
Sex	
n	60038
F	28387 (47%)
(95% Confidence Interval)	(46.9, 47.7)
M	31650 (53%)
(95% Confidence Interval)	(52.3, 53.1)
NA	1 (<1%)
(95% Confidence Interval)	(0.0, 0.0)
Age Group (YEARS)	
n	60038
Under 5	441 (<1%)
(95% Confidence Interval)	(0.7, 0.8)
5-17	6543 (11%)
(95% Confidence Interval)	(10.7, 11.1)
18-64	37832 (63%)
(95% Confidence Interval)	(62.6, 63.4)
65 and Older	14433 (24%)
(95% Confidence Interval)	(23.7, 24.4)
Age NA	789 (1%)
(95% Confidence Interval)	(1.2, 1.4)

Figure 10: Summary of sex and age groups for US Pharma studies

The 95% confidence intervals (CIs) are based on the Wilson method [5]. These can be obtained using the freq procedure and using the following options in the tables statement: binomial (wilson level=X), alpha=0.5, where X is each value of the categorical variable. I.e., for age group CIs, we needed to call proc freq 5 times. Each time, replacing the level with the next age group category, including the "Age NA" category.

Summary of US Studies

The other demography table summarised race and ethnicity by study. This table calculated how many studies had each race/ethnicity category for at least one subject on the study, and how many studies met the 2019 US census for that category. E.g., if we consider the race "Asian", Figure 11 indicates that out of 45 vaccine studies, 40 studies had at least one subject with the race Asian. However, only 7 of the vaccine studies had a percentage of subjects greater than the census percentage for that race, which in this case is 5.9% for Asian subjects. The census percentages in the summary table were based on the 2019 US census percentages extracted from the reference census dataset, Figure 6.

	Vaccines (N=45)	
Ethnicity		
n	45	
Hispanic or Latino	41	(91%)
Trials at or Above 18.5% US Census	16	(36%)
Not Hispanic or Latino	43	(96%)
Trials at or Above 81.5% US Census	27	(60%)
Non Hispanic White	41	(91%)
Trials at or Above 60.1% US Census	21	(47%)
High Level Race		
n	45	
American Indian or Alaskan Native	27	(60%)
Trials at or Above 1.3% US Census	5	(11%)
Asian	40	(89%)
Trials at or Above 5.9% US Census	7	(16%)
Black or African American	41	(91%)
Trials at or Above 13.4% US Census	14	(31%)
Native Hawaiian or Other Pacific Islander	23	(51%)
Trials at or Above 0.2% US Census	14	(31%)
White	43	(96%)
Trials at or Above 76.3% US Census	19	(42%)
More Than One Race	28	(62%)
Trials at or Above 2.8% US Census	13	(29%)

Figure 11: Summary of race and ethnicity for US Vaccines studies

Summary Figures

Six figures were requested for the publication. These figures use various methods to compare the demographic makeup of the clinical trials to those found in the census and epidemiology data. The request involved replicating images of displays that were provided as examples of what was desired. SAS Graph Template Language was used to provide the options and control necessary to replicate the complex and non-standard nature of these displays.

Summary Figure 1 – Stacked Bar Chart of US Demographic Characteristics by Subjects and US Census

Figure 12 presents a cluster grouped bar chart for sex and age groups compared to the US census.

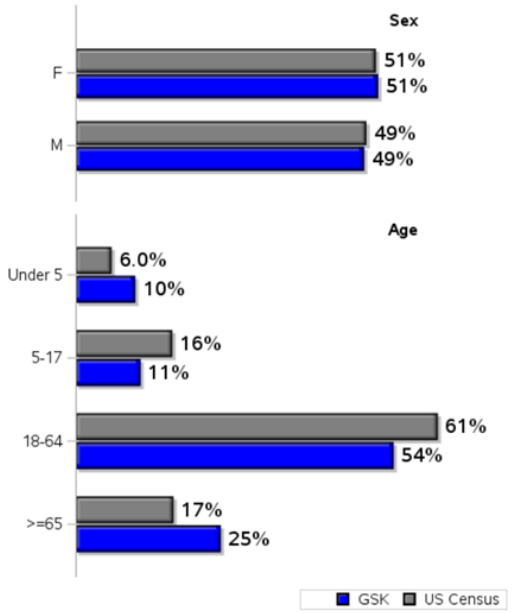


Figure 12: Rows of cluster grouped bar charts for demographics (race, ethnicity, sex, age).

The percentage of subjects of all the clinical trials was determined for each demographic group using proc freq. The results were merged with census data and a lattice layout with 5 rows and one column was created containing horizontal bar charts with demographic categories on the vertical X axis and percent of the total population as the bar length. The data was grouped according to the source using a clustered group display and the bars were labelled with the percentages using a bar label format of percent6.1. See Appendix 1 for full template.

There was minor adjustment to the 2019 US census data required, prior to reading it into the figure. The census data provided two age groups for people aged 5 to 13 and 5 to 17. However, we wanted to summarise subjects aged 5 to 17 years old together. This meant we needed to combine the counts in Figure 12 for these two age groups, in the figure program. We combined the data in the figure program, instead of modifying the census reference dataset, so that traceability was not affected.

Summary Figure 2 – Histogram and Cumulative Distribution Curve of Trial Makeup

Figure 13 presents a histogram, overlayed with a cumulative distribution curve, for subjects who had ethnicity recorded as "Hispanic or Latino".

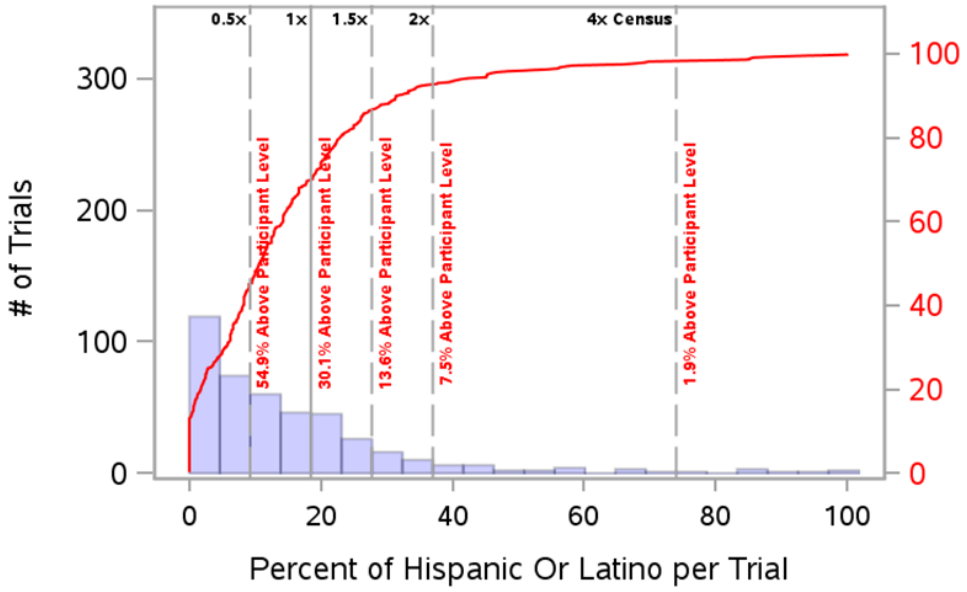


Figure 13: Histogram for each category of counts of trials by total percent with reference lines for multiples of census with an overlayed distribution curve showing Cumulative % of Trials at or Below Participant Level

This figure was generated using a lattice layout with 4 rows and 2 columns to create a plot for each of eight race categories containing a series plot of cumulative distribution (red line) overlayed over a histogram of study counts. The percentage of subjects was determined for each race category using the proc freq and summarized at the study level.

The bin sizes (bar widths) used in each histogram were assigned programmatically based on the census data. For the race categories White & non-Hispanic White these were set as 1/16th of the census values. For all other categories they were set as 1/4th of the census value. The histogram bar heights indicate

the number of studies in each bin. The first bar is the number of studies where the proportion if Hispanic or Latino participants was 0-1/4th of the census value (0-4.75%).

To handle a few cases of extremely skewed or biphasic distribution the histogram X axis was cut off at the minimum of the highest percentage found in any study or 8 times the census value. Since the US census value for American Indian or Alaskan native was 1.3, the X axis was cut off at 10.4 even though there was at least one study where the population was 100%.

For each possible percent value (X) found in the data the cumulative distribution was calculated as 100 * the number of studies where the percentage of study subjects were in the race category <=X% / total number of studies. In the plot above the red line crosses the 1x Census line at 69.9 indicating that in 69.9% of all the studies, the proportion of study participants that were Hispanic or Latino was less than 19%.

Vertical reference lines were added at X values programmatically determined by the census data. For the race categories White & non-Hispanic White the reference line values were derived for 0.25, 0.5, 0.75, 1.25, and 1.5 times the census level. For all other race categories, the reference line values were derived for 0.5, 1, 1.5, 2, and 4 times the census level.

The reference lines were labelled with text indicating the percentage of all studies found in bins above the reference line. This was calculated as 100 * the number of studies where the percentage of study subjects in the race category was greater than the multiple of the census value / total number of studies. To generate these labels, an annotate statement was used to call %sgtext. The labels were positioned based on the values derived for the reference lines, and a rotate option allowed the text to be rotated 90°. A macro variable assigned a value derived from the data was used to produce the percentage value in the label's text. See Appendix 2 for full template.

Summary Figure 3 – Heatmap of Percent of Census by Therapeutic Area Category and Trial Phase

Figure 14 presents a heatmap of percentages of studies achieving census level. The figure summarises race by therapeutic aera category and study phases.

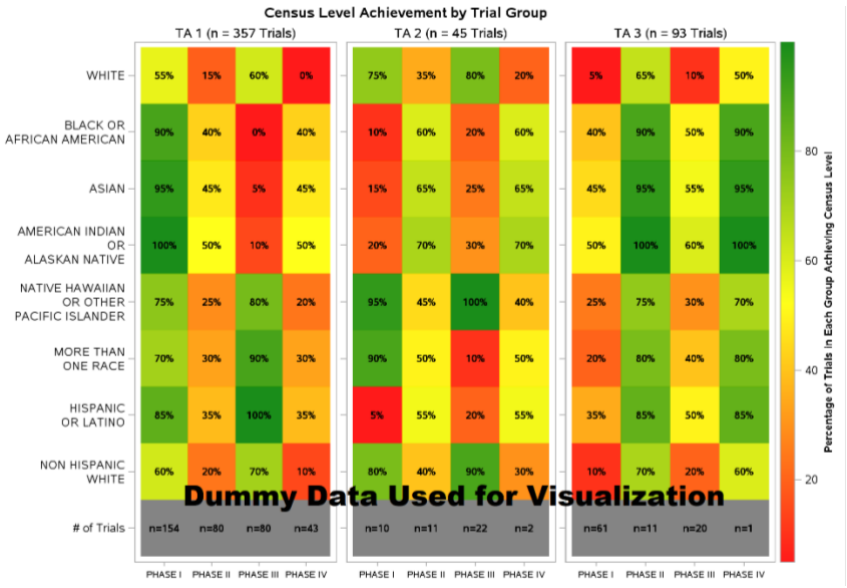


Figure 14: Heatmap of % of studies achieving census level for each race/ethnicity category by TA category and trial phase (Numbers and colours are produced with dummy data)

For each study the percentage of study participants in each category was calculated. This was compared to the census and each study was assigned a flag for each category. A flag value of "Y" was assigned if study population percentage was greater or equal to that of the census. For each category the percent of studies with "Y" values was summarized for each group within each TA category.

A lattice layout plot was created with 1 row and 3 columns, one column per TA category. A common Y axis indicated the race/ethnicity category being summarised or the total number of trials included in each group. Each column has a scatterplot of a text version of the percentage value overlayed on a heatmap (heatmapparm) using the percentage value as the colorresponse and a colormodel=(red yellow green) option.

For the "# of Trials" category the variable for the text version of the percentage value was assigned "n=X" where X is the total number of studies in that group. This appears as a scatterplot marker, but because the actual percentage value is missing no heatmap colour is assigned. See Appendix 3 for full template.

Summary Figure 4 – Number of Trials by Percent of Trial Makeup

Figure 15 presents a bar chart of the number of studies by proportion of the census level. The figure includes subjects who had ethnicity recorded as "Hispanic or Latino" and subjects who were recorded as "Non-Hispanic White".

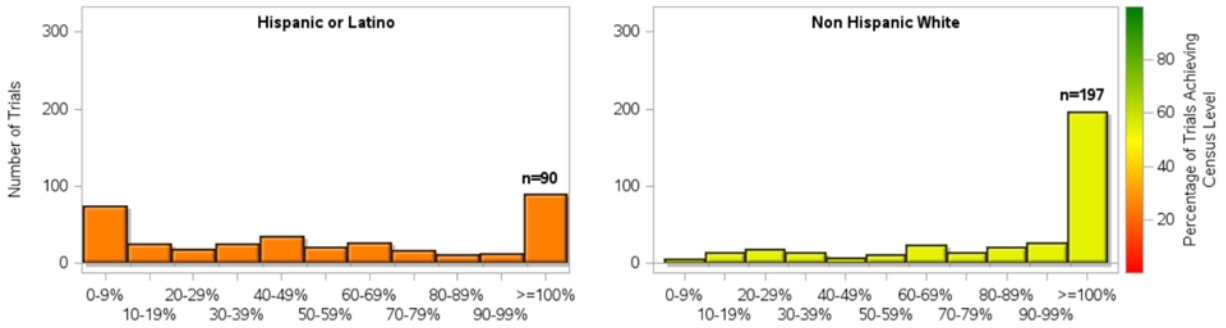


Figure 15: Plot for each TA containing a bar chart for each race/ethnicity category

For each study the percentage of study participants in each category was calculated. This value was then compared to the census to calculate the percent of census value. For each race category, each study was assigned to a group based on the percentage of census proportion represented. A study where the proportion of subjects was >=0% and <10% was assigned to the 0-9% group. If the proportion was >=10% and <20% of the census level it was assigned to the 10-19% group. Any study with equal or higher representation than found in the census was assigned to the >=100% group. The number of studies assigned to each group determines the height of the bar for that group.

A lattice layout with 4 rows and 2 columns was used to create plots for each category. Each plot was assigned the same axes attributes to allow easy comparison of groups. Bar charts were created using the number of studies as bar height and the discrete x axis ticks corresponding to each bar. Each tick was assigned a label indicating the proportion of census contained withing the bar.

The percentage of all studies achieving census level was calculated using Pct = 100 * (count where % >=100 / total number of studies). This was used to assign an overall colour to all bars in each plot using a colormap assigned by a rangeattrmap statement with range 0-100 and a rangecolormodel option.

A scatter plot was used to overlay text to label the >=100% bar with the number of studies achieving census level. See Appendix 4 for full template.

Summary Figure 5 – Mean and Median Additional Subjects Needed per Trial to Achieve Census Level Representation

Figure 16 presents a bar chart showing mean and median number of subject required to meet census levels for race and ethnicity. This figure only includes one therapeutic area category.

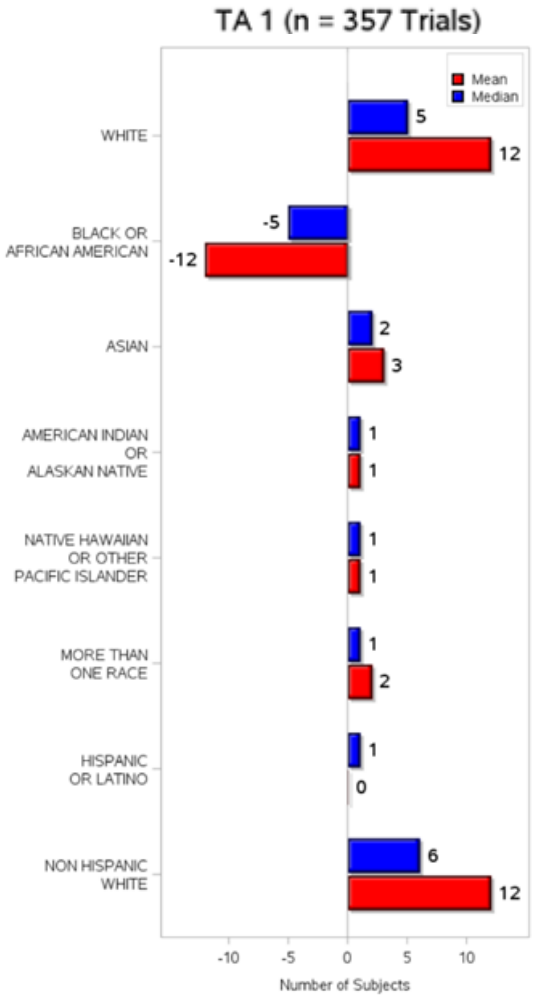


Figure 16: Rows of bar charts showing the number of subjects needed per study to achieve census level representation, cluster grouped by statistic

For each study the number of subjects required to match census proportion was calculated for each race category. The actual number of subjects in each category was subtracted from the expected value to determine the additional number of subjects that would have been necessary to add to each study to achieve a proportion that matches the census. This difference was rounded up to the nearest integer.

For each therapeutic area category, the mean and median of these values were reported using horizontal bar plots for each race category using cluster grouping by statistic. See Appendix 5 for full template.

Summary Figure 6 – Stacked Bar Chart of Ethnic Proportions of US GSK Trials, Epidemiology Data, and US Census

Figure 17 presents a cluster grouped bar chart for some race values compared to the census and epidemiology data.

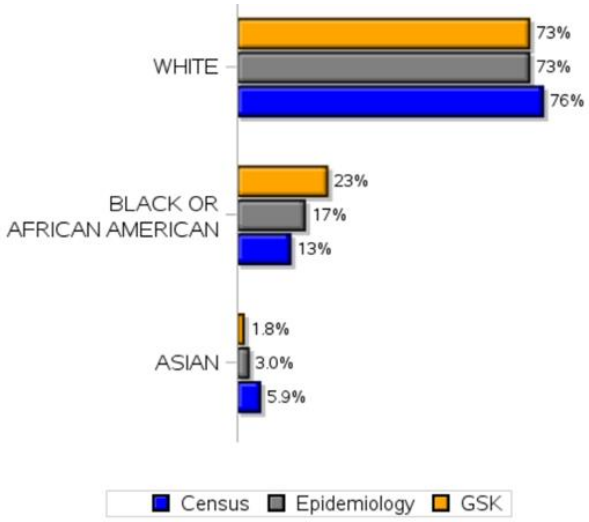


Figure 17: Rows of cluster grouped bar charts for race categories comparing trial demographics to census and epidemiology

A horizontal bar chart was produced for Census, Epidemiology, and GSK clinical trials with race categories on the vertical X axis and percent of the total population as the bar length. The data was grouped according to the source using a clustered group display. The bars were labelled with the percentages using a bar label format of percent6.1. See Appendix 6 for full template.

References

[1] Diversity Plans to Improve Enrollment of Participants from Underrepresented Racial and Ethnic Populations in Clinical Trials Guidance for Industry, FDA Draft Guidance

[2] Demographic diversity of participants in Pfizer sponsored clinical trials in the United States, Melinda Rottas, Peter Thadeio, Rachel Simons, Raven Houck, David Gruben, David Keller, David Scholfield, Koshika Soma, Brian Corrigan, Annette Schettino, Patrick J. McCann, Marie-Pierre Hellio, Kannan Natarajan, Rob Goodwin, Judy Sewards, Peter Honig, Rod MacKenzie

[3] US Census Bureau Population Division, Annual Estimates of the Resident Population by Sex, Race, and Hispanic Origin for the United States: April 1, 2010 to July 1, 2019 (NC-EST2019-SR11H), Release Date: June 2020. (Accessed 12 Jul 2021).

[4] Centers for Disease Control and Prevention

[5] Wilson, E. B. (1927). "Probable Inference, the Law of Succession, and Statistical Inference." Journal of the American Statistical Association 22:209–212.

Acknowledgements

The authors would like to thank the GSK Cell and Gene Therapy biostatistics, clinical development, data management, and programming teams for their support, review, and encouragement. With special thanks to Ashwin Mathews, Michel Reid, and Danae Williams. We would also like to thank Adrienne Bonwick & Rabia Ahmad for their oversight and guidance.

Contact Information

Carletha Blanding
GlaxoSmithKline
Email carletha.r.blanding@gsk.com

Chris Leach
GlaxoSmithKline
Email chris.x.leach@gsk.com

Harrison James
GlaxoSmithKline
Email harrison.x.james@gsk.com

Appendices

Appendix 1 – Template for Summary Figure 1

```
proc template;
   define style graphtmpl;
      parent=styles.statistical;
      class GraphFonts /
         'GraphDataFont'=("<sans-serif>, <MTsans-serif>",9pt)
         'GraphUnicodeFont'=("<MTsans-serif-unicode>",10pt)
         'GraphValueFont'=("<sans-serif>, <MTsans-serif>",8pt)
         'GraphLabelFont'=("<sans-serif>, <MTsans-serif>",8pt)
         'GraphFootnoteFont'=("<sans-serif>, <MTsans-serif>",8pt)
         'GraphTitleFont'=("<sans-serif>, <MTsans-serif>",10pt,bold)
         'GraphAnnoFont'=("<sans-serif>, <MTsans-serif>",8pt);
      class GraphData1 /  color=blue         contrastcolor=blue;
      class GraphData2 /  color=gray         contrastcolor=gray;
   end;
run;

proc template;
   define statgraph splot;
   begingraph / designwidth=10in designheight=13in border=no;
     layout lattice / rows=5 columns=1
                      rowweights=(0.37 0.09 0.09 .15 0.3)
                      rowgutter=10;
   sidebar / align=top ;
      entry "Stacked Bar Chart of US Demographic Characteristics by Subjects
             and US Census" / textattrs=GraphTitleText pad=(bottom=5px);
   endsidebar;
   layout overlay / walldisplay=(fill)
      xaxisopts=(display=none type=linear linearopts=(viewmin=0 viewmax=1.4))
      yaxisopts=(label=" " reverse=1 offsetmin=0.15);
      entry "Race" / textattrs=(weight=bold) valign=top;
      barchart x=cat1 y=pct1 / orient=horizontal     group=bar1
                               groupdisplay=cluster  clusterwidth=.7
                               barwidth=.9           dataskin=crisp
                               barlabel=true         barlabelfitpolicy=none
                               barlabelformat=percent6.1;
   endlayout;
   layout overlay / walldisplay=(fill)
      xaxisopts=(display=none type=linear linearopts=(viewmin=0 viewmax=1.1))
      yaxisopts=(label=" " reverse=1 offsetmin=0.2);
      entry "Ethnicity" / textattrs=(weight=bold) valign=top;
      barchart x=cat2 y=pct2 / orient=horizontal     group=bar2
                               groupdisplay=cluster  clusterwidth=.7
                               barwidth=.9           dataskin=crisp
                               barlabel=true         barlabelfitpolicy=none
                               barlabelformat=percent6.1;
   endlayout;
   layout overlay / walldisplay=(fill)
      xaxisopts=(display=none type=linear linearopts=(viewmin=0 viewmax=1.1))
      yaxisopts=(label=" " reverse=1 offsetmin=0.2);
      entry "Race/Ethnicity Combined" /  textattrs=(weight=bold) valign=top;

      barchart x=cat3 y=pct3 / orient=horizontal     group=bar3
```

```
                              groupdisplay=cluster  clusterwidth=.7
                              barwidth=.9           dataskin=crisp
                              barlabel=true         barlabelfitpolicy=none
                              barlabelformat=percent6.1;
   endlayout;
   layout overlay / walldisplay=(fill)
      xaxisopts=(display=none type=linear linearopts=(viewmin=0 viewmax=1.1))
      yaxisopts=(label=" " reverse=1 offsetmin=0.35);
      entry "Sex" / textattrs=(weight=bold) valign=top;
      barchart x=cat4 y=pct4 / orient=horizontal     group=bar4
                              groupdisplay=cluster  clusterwidth=.7
                              barwidth=.9           dataskin=crisp
                              barlabel=true         barlabelfitpolicy=none
                              barlabelformat=percent6.1;
   endlayout;
   layout overlay / walldisplay=(fill)
      xaxisopts=(display=none type=linear linearopts=(viewmin=0 viewmax=1.1))
      yaxisopts=(label=" " reverse=1 offsetmin=0.17);
      entry "Age" / textattrs=(weight=bold) valign=top;
      barchart x=cat5 y=pct5 / orient=horizontal group=bar5
                              groupdisplay=cluster  clusterwidth=.7
                              barwidth=.9           dataskin=crisp
                              barlabel=true         barlabelfitpolicy=none
                              barlabelformat=percent6.1;
                              groupdisplay=cluster  clusterwidth=.7
                              name='bar'
      DiscreteLegend "bar" /title=" " EXCLUDE=(" " ".") across=2
                             location=outside halign=center valign=bottom;
   endlayout;
   endlayout;
   endgraph;
   end;
run;
```

Appendix 2 – Template for Summary Figure 2

```
%sganno;

data anno;
   length id $4;
   retain id 'grp1';
   keep id;
   set figure;

   id='grp1';
   if grp1=1 and cl1^='' and pct1<100 then do;
      %sgtext(x1=clx1, y1=160,     label=cl1, x1space="datavalue",
              y1space="datavalue", width=50,  anchor="top",  textcolor="red",
              textweight="bold",   rotate=90, textsize=4);
   end;

   id='grp2';
   if grp2=2 and cl2^='' then do;
      %sgtext(x1=clx2, y1=160,     label=cl2,  x1space="datavalue",
              y1space="datavalue", width=50,   anchor="top", textcolor="red",
              textweight="bold",   rotate=90,  textsize=5);
   end;

   id='grp3';
   if grp3=3 and cl3^='' then do;
      %sgtext(x1=clx3, y1=160,     label=cl3,  x1space="datavalue",
              y1space="datavalue", width=50,   anchor="top", textcolor="red",
              textweight="bold",   rotate=90,  textsize=5);
   end;

   id='grp4';
   if grp4=4 and cl4^='' then do;
      %sgtext(x1=clx4, y1=160,     label=cl4,  x1space="datavalue",
              y1space="datavalue", width=50,   anchor="top", textcolor="red",
              textweight="bold",   rotate=90,  textsize=5);
   end;

   id='grp5';
   if grp5=5 and cl5^='' then do;
      %sgtext(x1=clx5, y1=160,     label=cl5,  x1space="datavalue",
              y1space="datavalue", width=50,   anchor="top", textcolor="red",
              textweight="bold",   rotate=90,  textsize=5);
   end;

   id='grp6';
   if grp6=6 and cl6^='' then do;
      %sgtext(x1=clx6, y1=160,     label=cl6,  x1space="datavalue",
              y1space="datavalue", width=50,   anchor="top", textcolor="red",
              textweight="bold",   rotate=90,  textsize=5);
   end;

   id='grp7';
   if grp7=7 and cl7^='' then do;
      %sgtext(x1=clx7, y1=160,     label=cl7,  x1space="datavalue",
```

```
            y1space="datavalue", width=50,   anchor="top", textcolor="red",
            textweight="bold",   rotate=90,  textsize=5);
   end;

   id='grp8';
   if grp8=8 and cl8^='' then do;
      %sgtext(x1=clx8, y1=160,     label=cl8,  x1space="datavalue",
            y1space="datavalue", width=50,   anchor="top", textcolor="red",
            textweight="bold",   rotate=90,  textsize=5);
   end;

run;

proc template;
   define style graphtmpl;
      parent=styles.statistical;
      class GraphFonts /
         'GraphDataFont'=("<sans-serif>, <MTsans-serif>",9pt)
         'GraphUnicodeFont'=("<MTsans-serif-unicode>",10pt)
         'GraphValueFont'=("<sans-serif>, <MTsans-serif>",8pt)
         'GraphLabelFont'=("<sans-serif>, <MTsans-serif>",8pt)
         'GraphFootnoteFont'=("<sans-serif>, <MTsans-serif>",8pt)
         'GraphTitleFont'=("<sans-serif>, <MTsans-serif>",10pt,bold)
         'GraphAnnoFont'=("<sans-serif>, <MTsans-serif>",8pt);
   end;
run;

proc template;
   define statgraph splot;

   begingraph / designwidth=11in designheight=13in border=no;

   layout lattice / rows=4 columns=2
                    rowweights=(.25 .25 .25 .25) columnweights=(0.50 0.50);

   entrytitle "Number and Cumulative Percentage US Clinical Trials vs
               Percentage of Subjects Per Trial By Race and Ethnicity"
             / textattrs= (size=9 weight=bold);

   layout overlay /
      xaxisopts=(offsetmin=0.05  offsetmax=0.05  tickvalueattrs=(size=7pt)
                 label="Percent of &cat1. per Trial"  labelattrs=(size=8)
                 type=linear linearopts= (tickvaluefitpolicy=none
                 viewmin=0 viewmax=&xmax1.))

      yaxisopts=(label="# of Trials" type=linear
                 linearopts=(viewmin=0 viewmax=&vmax
                 tickvaluesequence=(start=0 end=&ymax increment=&yinc)))

      y2axisopts=(offsetmax=0.1 label=" " labelattrs= (size=6 color=red)
                  tickvalueattrs=(color=red));

      histogram pct1 / primary=true endlabels=true scale=count binstart=0
                       binwidth=&wbin1. xvalues=leftpoints
                       fillattrs=(color=bluegray) datatransparency=.8;

      seriesplot y=cd1 x=pct1 /
```

```
            yaxis=y2 lineattrs=(thickness=1 pattern=solid color=red);

      annotate / id="grp1";

      referenceline x=eval(0.25 * &cen1.) /
         lineattrs=(thickness=1 pattern=mediumdash) clip=true
         curvelabel="0.25x"     curvelabelattrs=(size=4 weight=bold);

      referenceline x=eval(0.5 * &cen1.) /
         lineattrs=(thickness=1 pattern=mediumdash) clip=true
         curvelabel="0.5x"      curvelabelattrs=(size=4 weight=bold);

      referenceline x=eval(0.75 * &cen1.) /
         lineattrs=(thickness=1 pattern=mediumdash) clip=true
         curvelabel="0.75x"     curvelabelattrs=(size=4 weight=bold);

      referenceline x=&cen1. /
         lineattrs=(thickness=1 pattern=solid) clip=true
         curvelabel="1x Census" curvelabelattrs=(size=4 weight=bold);

      referenceline x=eval(1.25 * &cen1.) /
         lineattrs=(thickness=1 pattern=mediumdash) clip=true
         curvelabel="1.25x"     curvelabelattrs=(size=4 weight=bold);

      referenceline x=eval(1.5 * &cen1.) /
         lineattrs=(thickness=1 pattern=mediumdash) clip=true
         curvelabel="1.5x"      curvelabelattrs=(size=4 weight=bold);

endlayout;

layout overlay /
   xaxisopts=(offsetmin=0.05 offsetmax=0.05 tickvalueattrs=(size=7pt)
              label="Percent of &cat2. per Trial" labelattrs= (size=8)
              type=linear linearopts= (tickvaluefitpolicy=none
              viewmin=0 viewmax=&xmax2.))

   yaxisopts=(label=" " type=linear  linearopts=(viewmin=0 viewmax=&vmax
              tickvaluesequence=(start=0 end=&ymax increment=&yinc)))

   y2axisopts=(offsetmax=0.1 label="Cumulative % of Trials at or
              (*ESC*){unicode '000a'x} Below Participant Level"
              labelattrs= (size=6 color=red) tickvalueattrs=(color=red));

   histogram pct2 / primary=true endlabels=true scale=count binstart=0
                    binwidth=&bin2. xvalues=leftpoints
                    fillattrs=(color=bluegray) datatransparency=.8;

   seriesplot y=cd2 x=pct2 /
              yaxis=y2 lineattrs=(thickness=1 pattern=solid color=red) ;

      annotate / id="grp2";

      referenceline x=eval(0.5 * &cen2.) /
         lineattrs=(thickness=1 pattern=mediumdash) clip=true
         curvelabel="0.5x"      curvelabelattrs=(size=4 weight=bold);

      referenceline x=&cen2. /
```

```
         lineattrs=(thickness=1 pattern=solid) clip=true
         curvelabel="1x"        curvelabelattrs=(size=4 weight=bold);

      referenceline x=eval(1.5 * &cen2.) /
         lineattrs=(thickness=1 pattern=mediumdash) clip=true
         curvelabel="1.5x"      curvelabelattrs=(size=4 weight=bold);

      referenceline x=eval(2 * &cen2.) /
         lineattrs=(thickness=1 pattern=mediumdash) clip=true
         curvelabel="2x"        curvelabelattrs=(size=4 weight=bold);

      referenceline x=eval(4 * &cen2.) /
         lineattrs=(thickness=1 pattern=mediumdash) clip=true
         curvelabel="4x Census" curvelabelattrs=(size=4 weight=bold) ;

endlayout;

layout overlay /
   xaxisopts=(offsetmin=0.05 offsetmax=0.05 tickvalueattrs=(size=7pt)
              label="Percent of &cat3. per Trial" labelattrs= (size=8)
              type=linear linearopts=(tickvaluefitpolicy=none
              viewmin=0 viewmax=&xmax3.))

   yaxisopts=(label="# of Trials" type=linear
              linearopts=(viewmin=0 viewmax=&vmax
              tickvaluesequence=(start=0 end=&ymax increment=&yinc)))

   y2axisopts=(offsetmax=0.1 label=" "
               labelattrs= (size=6 color=red) tickvalueattrs=(color=red));

   histogram pct3 / primary=true endlabels=true scale=count binstart=0
                    binwidth=&bin3. xvalues=leftpoints
                    fillattrs=(color=bluegray) datatransparency=.8;

   seriesplot y=cd3 x=pct3 /
              yaxis=y2 lineattrs=(thickness=1 pattern=solid color=red);

      annotate / id="grp3";

      referenceline x=eval(0.5 * &cen3.) /
         lineattrs=(thickness=1 pattern=mediumdash clip=true
         curvelabel="0.5x"      curvelabelattrs=(size=4 weight=bold) ;

      referenceline x=&cen3. /
         lineattrs=(thickness=1 pattern=solid)clip=true
         curvelabel="1x"        curvelabelattrs=(size=4 weight=bold) ;

      referenceline x=eval(1.5 * &cen3.) /
         lineattrs=(thickness=1 pattern=mediumdash) clip=true
         curvelabel="1.5x"      curvelabelattrs=(size=4 weight=bold) ;

      referenceline x=eval(2 * &cen3.) /
         lineattrs=(thickness=1 pattern=mediumdash) clip=true
         curvelabel="2x"        curvelabelattrs=(size=4 weight=bold) ;

      referenceline x=eval(4 * &cen3.) /
         lineattrs=(thickness=1 pattern=mediumdash) clip=true
```

```
         curvelabel="4x Census" curvelabelattrs=(size=4 weight=bold) ;

endlayout;

layout overlay /
   xaxisopts=(offsetmin=0.05 offsetmax=0.05 tickvalueattrs=(size=7pt)
              label="Percent of &cat4. per Trial" labelattrs= (size=8)
              type=linear linearopts=(tickvaluefitpolicy=none
              viewmin=0 viewmax=&xmax4.))

   yaxisopts=(label=" " type=linear
              linearopts=(viewmin=0 viewmax=&vmax
              tickvaluesequence=(start=0 end=&ymax increment=&yinc)))

   y2axisopts=(offsetmax=0.1 label="Cumulative % of Trials at or
              (*ESC*){unicode '000a'x} Below Participant Level"
              labelattrs= (size=6 color=red) tickvalueattrs=(color=red));

   histogram pct4 / primary=true endlabels=true scale=count binstart=0
                    binwidth=&bin4. xvalues=leftpoints
                    fillattrs=(color=bluegray) datatransparency=.8;

   seriesplot y=cd4 x=pct4 /
              yaxis=y2 lineattrs=(thickness=1 pattern=solid color=red) ;

      annotate / id="grp4";

      referenceline x=eval(0.5 * &cen4.) /
         lineattrs=(thickness=1 pattern=mediumdash clip=true
         curvelabel="0.5x"      curvelabelattrs=(size=4 weight=bold) ;

      referenceline x=&cen4. /
         lineattrs=(thickness=1 pattern=solid) clip=true
         curvelabel="1x"        curvelabelattrs=(size=4 weight=bold) ;

      referenceline x=eval(1.5 * &cen4.) /
         lineattrs=(thickness=1 pattern=mediumdash) clip=true
         curvelabel="1.5x"      curvelabelattrs=(size=4 weight=bold) ;

      referenceline x=eval(2 * &cen4.) /
         lineattrs=(thickness=1 pattern=mediumdash) clip=true
         curvelabel="2x"        curvelabelattrs=(size=4 weight=bold) ;

      referenceline x=eval(4 * &cen4.) /
         lineattrs=(thickness=1 pattern=mediumdash) clip=true
         curvelabel="4x Census" curvelabelattrs=(size=4 weight=bold) ;

endlayout;

layout overlay /
   xaxisopts=(offsetmin=0.05 offsetmax=0.05 tickvalueattrs=(size=7pt)
              label="Percent of &cat5. per Trial" labelattrs= (size=8)
              type=linear linearopts=(tickvaluefitpolicy=none
              viewmin=0 viewmax=&xmax5.))

   yaxisopts=(label="# of Trials" type=linear
              linearopts=(viewmin=0 viewmax=&vmax
```

```
            tickvaluesequence=(start=0 end=&ymax increment=&yinc)))

   y2axisopts=(offsetmax=0.1 label=" "
               labelattrs= (size=6 color=red) tickvalueattrs=(color=red));

   histogram pct5 / primary=true endlabels=true scale=count binstart=0
                    binwidth=&bin5. xvalues=leftpoints
                    fillattrs=(color=bluegray) datatransparency=.8;

   seriesplot y=cd5 x=pct5 /
              yaxis=y2 lineattrs=(thickness=1 pattern=solid color=red) ;

      annotate / id="grp5";

      referenceline x=eval(0.5 * &cen5.) /
         lineattrs=(thickness=1 pattern=mediumdash) clip=true
         curvelabel="0.5x"      curvelabelattrs=(size=4 weight=bold) ;

      referenceline x=&cen5. /
         lineattrs=(thickness=1 pattern=solid) clip=true
         curvelabel="1x"        curvelabelattrs=(size=4 weight=bold) ;

      referenceline x=eval(1.5 * &cen5.) /
         lineattrs=(thickness=1 pattern=mediumdash) clip=true
         curvelabel="1.5x"      curvelabelattrs=(size=4 weight=bold) ;

      referenceline x=eval(2 * &cen5.) /
         lineattrs=(thickness=1 pattern=mediumdash) clip=true
         curvelabel="2x"        curvelabelattrs=(size=4 weight=bold) ;

      referenceline x=eval(4 * &cen5.) /
         lineattrs=(thickness=1 pattern=mediumdash) clip=true
         curvelabel="4x Census" curvelabelattrs=(size=4 weight=bold) ;

endlayout;

layout overlay /
   xaxisopts=(offsetmin=0.05 offsetmax=0.05 tickvalueattrs=(size=7pt)
              label="Percent of &cat6. per Trial" labelattrs= (size=8)
              type=linear linearopts=(tickvaluefitpolicy=none
              viewmin=0 viewmax=&xmax6.))

   yaxisopts=(label=" " type=linear
              linearopts=(viewmin=0 viewmax=&vmax
              tickvaluesequence=(start=0 end=&ymax increment=&yinc)))

   y2axisopts=(offsetmax=0.1 label="Cumulative % of Trials at or
              (*ESC*){unicode '000a'x} Below Participant Level"
              labelattrs= (size=6 color=red) tickvalueattrs=(color=red));

   histogram pct6 / primary=true endlabels=true scale=count binstart=0
                    binwidth=&bin6. xvalues=leftpoints
                    fillattrs=(color=bluegray) datatransparency=.8;

   seriesplot y=cd6 x=pct6 /
              yaxis=y2 lineattrs=(thickness=1 pattern=solid color=red) ;
```

```
      annotate / id="grp6";

      referenceline x=eval(0.5 * &cen6.) /
         lineattrs=(thickness=1 pattern=mediumdash)clip=true
         curvelabel="0.5x"      curvelabelattrs=(size=4 weight=bold) ;

      referenceline x=&cen6. /
         lineattrs=(thickness=1 pattern=solid) clip=true
         curvelabel="1x"        curvelabelattrs=(size=4 weight=bold) ;

      referenceline x=eval(1.5 * &cen6.) /
         lineattrs=(thickness=1 pattern=mediumdash) clip=true
         curvelabel="1.5x"      curvelabelattrs=(size=4 weight=bold) ;

      referenceline x=eval(2 * &cen6.) /
         lineattrs=(thickness=1 pattern=mediumdash) clip=true
         curvelabel="2x"        curvelabelattrs=(size=4 weight=bold) ;

      referenceline x=eval(4 * &cen6.) /
         lineattrs=(thickness=1 pattern=mediumdash) clip=true
         curvelabel="4x Census" curvelabelattrs=(size=4 weight=bold) ;

endlayout;

layout overlay /
   xaxisopts=(offsetmin=0.05 offsetmax=0.05 tickvalueattrs=(size=7pt)
              label="Percent of &cat7. per Trial" labelattrs= (size=8)
              type=linear linearopts=(tickvaluefitpolicy=none
              viewmin=0 viewmax=&xmax7.))

   yaxisopts=(label="# of Trials " type=linear
              linearopts=(viewmin=0 viewmax=&vmax
              tickvaluesequence=(start=0 end=&ymax increment=&yinc)))

   y2axisopts=(offsetmax=0.1 label=" "
               labelattrs= (size=6 color=red) tickvalueattrs=(color=red));

   histogram pct7 / primary=true endlabels=true scale=count binstart=0
                    binwidth=&bin7. xvalues=leftpoints
                    fillattrs=(color=bluegray) datatransparency=.8;

   seriesplot y=cd7 x=pct7 /
              yaxis=y2 lineattrs=(thickness=1 pattern=solid color=red) ;

      annotate / id="grp7";

      referenceline x=eval(0.5 * &cen7.) /
         lineattrs=(thickness=1 pattern=mediumdash) clip=true
         curvelabel="0.5x"      curvelabelattrs=(size=4 weight=bold) ;

      referenceline x=&cen7. /
         lineattrs=(thickness=1 pattern=solid) clip=true
         curvelabel="1x"        curvelabelattrs=(size=4 weight=bold) ;

      referenceline x=eval(1.5 * &cen7.) /
         lineattrs=(thickness=1 pattern=mediumdash) clip=true
         curvelabel="1.5x"      curvelabelattrs=(size=4 weight=bold) ;
```

```
      referenceline x=eval(2 * &cen7.) /
         lineattrs=(thickness=1 pattern=mediumdash) clip=true
         curvelabel="2x"        curvelabelattrs=(size=4 weight=bold) ;

      referenceline x=eval(4 * &cen7.) /
         lineattrs=(thickness=1 pattern=mediumdash) clip=true
         curvelabel="4x Census" curvelabelattrs=(size=4 weight=bold) ;

endlayout;

layout overlay /
   xaxisopts=(offsetmin=0.05 offsetmax=0.05 tickvalueattrs=(size=7pt)
              label="Percent of &cat8. per Trial" labelattrs= (size=8)
              type=linear linearopts=(tickvaluefitpolicy=none
              viewmin=0 viewmax=&xmax8.))

   yaxisopts=(label=" " type=linear
              linearopts=(viewmin=0 viewmax=&vmax
              tickvaluesequence=(start=0 end=&ymax increment=&yinc)))

   y2axisopts=(offsetmax=0.1 label="Cumulative % of Trials at or
              (*ESC*){unicode '000a'x} Below Participant Level"
              labelattrs= (size=6 color=red) tickvalueattrs=(color=red));

   histogram pct8 / primary=true endlabels=true scale=count binstart=0
                    binwidth=&wbin8. xvalues=leftpoints
                    fillattrs=(color=bluegray ) datatransparency=.8;

   seriesplot y=cd8 x=pct8 /
              yaxis=y2 lineattrs=(thickness=1 pattern=solid color=red) ;

      annotate / id="grp8";

      referenceline x=eval(0.25 * &cen8.) /
         lineattrs=(thickness=1 pattern=mediumdash) clip=true
         curvelabel="0.25x"     curvelabelattrs=(size=4 weight=bold) ;

      referenceline x=eval(0.5 * &cen8.) /
         lineattrs=(thickness=1 pattern=mediumdash) clip=true
         curvelabel="0.5x"      curvelabelattrs=(size=4 weight=bold) ;

      referenceline x=eval(0.75 * &cen8.) /
         lineattrs=(thickness=1 pattern=mediumdash) clip=true
         curvelabel="0.75x"     curvelabelattrs=(size=4 weight=bold) ;

      referenceline x=&cen8. /
         lineattrs=(thickness=1 pattern=solid) clip=true
         curvelabel="1x Census" curvelabelattrs=(size=4 weight=bold) ;

      referenceline x=eval(1.25 * &cen8.) /
         lineattrs=(thickness=1 pattern=mediumdash)clip=true
         curvelabel="1.25x"     curvelabelattrs=(size=4 weight=bold) ;

      referenceline x=eval(1.5 * &cen8.) /
         lineattrs=(thickness=1 pattern=mediumdash) clip=true
         curvelabel="1.5x"      curvelabelattrs=(size=4 weight=bold) ;
```

```
   endlayout;

   endlayout;
   endgraph;
   end;
run;
```

Appendix 3 – Template for Summary Figure 3

```
proc template;
   define style graphtmpl;
      parent=styles.statistical;
      class GraphFonts /
         'GraphDataFont'=("<sans-serif>, <MTsans-serif>",9pt)
         'GraphUnicodeFont'=("<MTsans-serif-unicode>",10pt)
         'GraphValueFont'=("<sans-serif>, <MTsans-serif>",8pt)
         'GraphLabelFont'=("<sans-serif>, <MTsans-serif>",8pt)
         'GraphFootnoteFont'=("<sans-serif>, <MTsans-serif>",8pt)
         'GraphTitleFont'=("<sans-serif>, <MTsans-serif>",10pt,bold)
         'GraphAnnoFont'=("<sans-serif>, <MTsans-serif>",8pt);
   end;
run;


proc template;
   define statgraph splot;

   begingraph / designwidth=13in designheight=9in border=no;

   layout lattice / rows=1 columns=3 columnweights=(0.43 0.25 0.32) ;

   sidebar / align=top;
      entry "Census Level Achievement by Trial Phase" /
         textattrs=GraphTitleText pad=(bottom=5px);
   endsidebar;

   layout overlay /
      xaxisopts=(label=" ")

      yaxisopts=(label=" " reverse=1 type=discrete tickvalueattrs=(size=9)
                 discreteopts=(tickvaluefitpolicy=split
                 tickvaluesplitchar='@'))

      x2axisopts=(display=(line label) label="&g1 (n = &n1 Trials)") ;

      heatmapparm y=y1 x=x1  colorresponse=val1 /
         colormodel=(red yellow green) datatransparency=.1;

      scatterplot y=y1 x=x1 / xaxis=x2 markercharacter=lab1 labelstrip=true
                              markercharacterattrs=(size=8 weight=bold );
   endlayout;

   layout overlay /
      xaxisopts=(label=" ")

      yaxisopts=(display=(line ticks) reverse=1)

      x2axisopts=(display=(line label) label="&g2 (n = &n2 Trials)") ;

      heatmapparm y=y2 x=x2 colorresponse=val2 /
         colormodel=(red yellow green) datatransparency=.1;

      scatterplot y=y2 x=x2 / xaxis=x2 markercharacter=lab2 labelstrip=true
```

```
                              markercharacterattrs=(size=8 weight=bold );
   endlayout;

   layout overlay /
      xaxisopts=(label=" ")

      yaxisopts=(display=(line ticks)  reverse=1)

      x2axisopts=(display=(line label) label="&g3 (n = &n3 Trials)" );

      heatmapparm y=y3 x=x3 colorresponse=val3 / name='heatmap'
         colormodel=(red yellow green) datatransparency=.1;

      scatterplot y=y3 x=x3 / xaxis=x2 markercharacter=lab3 labelstrip=true
                              markercharacterattrs=(size=8 weight=bold );

      continuouslegend "heatmap" /
         title="Percentage of Trials in Each Phase Achieving Census Level"
         location=outside halign=right valign=center
         titleattrs=(size=8 weight=bold ) ;

   endlayout;

   endlayout;
   endgraph;
   end;
run;
```

Appendix 4 – Template for Summary Figure 4

```
proc template;
   define style graphtmpl;
      parent=styles.statistical;
      class GraphFonts /
         'GraphDataFont'=("<sans-serif>, <MTsans-serif>",9pt)
         'GraphUnicodeFont'=("<MTsans-serif-unicode>",10pt)
         'GraphValueFont'=("<sans-serif>, <MTsans-serif>",8pt)
         'GraphLabelFont'=("<sans-serif>, <MTsans-serif>",8pt)
         'GraphFootnoteFont'=("<sans-serif>, <MTsans-serif>",5pt)
         'GraphTitleFont'=("<sans-serif>, <MTsans-serif>",9pt,bold)
         'GraphAnnoFont'=("<sans-serif>, <MTsans-serif>",8pt);
   end;
run;

%let tickval='0-9%' '10-19%' '20-29%' '30-39%' '40-49%' '50-59%' '60-69%'
'70-79%' '80-89%' '90-99%' '>=100%';

proc template;
   define statgraph splot;

   begingraph / designwidth=11in designheight=13in border=no;

   rangeattrmap name="ColorMap";
      range 0-100 / rangecolormodel=(red yellow green);
   endrangeattrmap;

   rangeattrvar var=pct
                attrmap="ColorMap"
                attrvar=pcol;

   layout lattice / rows=4 columns=2 rowweights=(.25 .25 .25 .25)
                                     columnweights=(0.47 0.53) ;

      entrytitle "&title.";

   layout overlay /
      xaxisopts=(offsetmin=0.05 offsetmax=0.05 tickvalueattrs=(size=8pt)
                 label=" " type=discrete
                 discreteopts= (tickvaluefitpolicy=stagger
                 tickvaluelist=(&tickval.)))

      yaxisopts=(label="Number of Trials"
                 linearopts=(viewmin=0 viewmax=%eval(&ymax. + &laboff.)
                 tickvaluesequence=(start=0 end=&ymax increment=&yinc)))

      x2axisopts=(display=(line) ) ;

         entry "White" /
            textattrs=(weight=bold) valign=top;

      barchartparm y=y1 x=x1 / barwidth=1 dataskin=crisp colorresponse=pcol;

      scatterplot y=ylab1 x=x1 /
                  xaxis=x2 markercharacter=lab1 labelstrip=true
```

```
               markercharacterattrs=(size=8 weight=bold);

endlayout;

layout overlay /
   xaxisopts=(offsetmin=0.05 offsetmax=0.05 tickvalueattrs=(size=8pt)
              label=" " type=discrete
              discreteopts=(tickvaluefitpolicy=stagger
              tickvaluelist=(&tickval.)))

   yaxisopts=(display=(line ticks) label=" "
              linearopts=(viewmin=0 viewmax=%eval(&ymax. + &laboff.)
              tickvaluesequence=(start=0 end=&ymax increment=&yinc)))

   x2axisopts=(display=(line) ) ;

      entry "Black or African American" /
         textattrs=(weight=bold) valign=top;


   barchartparm y=y2 x=x2 / barwidth=1 dataskin=crisp colorresponse=pcol;

   scatterplot y=ylab2 x=x2 /
               xaxis=x2 markercharacter=lab2 labelstrip=true
               markercharacterattrs=(size=8 weight=bold );

endlayout;

layout overlay /
   xaxisopts=(offsetmin=0.05 offsetmax=0.05 tickvalueattrs=(size=8pt)
              label=" " type=discrete
              discreteopts=(tickvaluefitpolicy=stagger
              tickvaluelist=(&tickval.) ))

   yaxisopts=(label="Number of Trials"
              linearopts=(viewmin=0 viewmax=%eval(&ymax. + &laboff.)
              tickvaluesequence=(start=0 end=&ymax increment=&yinc)))

   x2axisopts=(display=(line) ) ;

      entry "Asian" /
         textattrs=(weight=bold) valign=top;

   barchartparm y=y3 x=x3 / barwidth=1 dataskin=crisp colorresponse=pcol;

   scatterplot y=ylab3 x=x3 /
               xaxis=x2 markercharacter=lab3 labelstrip=true
               markercharacterattrs=(size=8 weight=bold );

endlayout;

layout overlay /
   xaxisopts=(offsetmin=0.05 offsetmax=0.05 tickvalueattrs=(size=8pt)
              label=" " type=discrete
              discreteopts=(tickvaluefitpolicy=stagger
              tickvaluelist=(&tickval.) ))
```

```
   yaxisopts=(display=(line ticks) label=" "
              linearopts=(viewmin=0 viewmax=%eval(&ymax. + &laboff.)
              tickvaluesequence=(start=0 end=&ymax increment=&yinc)))

   x2axisopts=(display=(line) ) ;

      entry "American Indian or Alaskan Native" /
         textattrs=(weight=bold) valign=top;

   barchartparm y=y4 x=x4 / barwidth=1 dataskin=crisp colorresponse=pcol;

   scatterplot y=ylab4 x=x4 /
               xaxis=x2 markercharacter=lab4 labelstrip=true
               markercharacterattrs=(size=8 weight=bold );

endlayout;

layout overlay /
   xaxisopts=(offsetmin=0.05 offsetmax=0.05 tickvalueattrs=(size=8pt)
              label=" " type=discrete
              discreteopts=(tickvaluefitpolicy=stagger
              tickvaluelist=(&tickval.) ))

   yaxisopts=(label="Number of Trials"
              linearopts=(viewmin=0 viewmax=%eval(&ymax. + &laboff.)
              tickvaluesequence=(start=0 end=&ymax increment=&yinc)))

   x2axisopts=(display=(line) ) ;

      entry "Native Hawaiian or Other Pacific Islander" /
         textattrs=(weight=bold) valign=top;

   barchartparm y=y5 x=x5 / barwidth=1 dataskin=crisp colorresponse=pcol;

   scatterplot y=ylab5 x=x5 /
               xaxis=x2 markercharacter=lab5 labelstrip=true
               markercharacterattrs=(size=8 weight=bold );

endlayout;

layout overlay /
   xaxisopts=(offsetmin=0.05 offsetmax=0.05 tickvalueattrs=(size=8pt)
              label=" " type=discrete
              discreteopts=(tickvaluefitpolicy=stagger
              tickvaluelist=(&tickval.) ))

   yaxisopts=(display=(line ticks) label=" "
              linearopts=(viewmin=0 viewmax=%eval(&ymax. + &laboff.)
              tickvaluesequence=(start=0 end=&ymax increment=&yinc)))

   x2axisopts=(display=(line) ) ;

      entry "More Than One Race" /
         textattrs=(weight=bold) valign=top;

   barchartparm y=y6 x=x6 / barwidth=1 dataskin=crisp colorresponse=pcol;
```

```
   scatterplot y=ylab6 x=x6 /
               xaxis=x2 markercharacter=lab6 labelstrip=true
               markercharacterattrs=(size=8 weight=bold );

endlayout;

layout overlay /
   xaxisopts=(offsetmin=0.05 offsetmax=0.05 tickvalueattrs=(size=8pt)
              label=" " type=discrete
              discreteopts=(tickvaluefitpolicy=stagger
              tickvaluelist=(&tickval.) ))

   yaxisopts=(label="Number of Trials"
              linearopts=(viewmin=0 viewmax=%eval(&ymax. + &laboff.)
              tickvaluesequence=(start=0 end=&ymax increment=&yinc)))

   x2axisopts=(display=(line) ) ;

      entry "Hispanic or Latino" /
         textattrs=(weight=bold) valign=top;

   barchartparm y=y7 x=x7 / barwidth=1 dataskin=crisp colorresponse=pcol;

   scatterplot y=ylab7 x=x7 /
               xaxis=x2 markercharacter=lab7 labelstrip=true
               markercharacterattrs=(size=8 weight=bold );

endlayout;

layout overlay /
   xaxisopts=(offsetmin=0.05 offsetmax=0.05 tickvalueattrs=(size=8pt)
              label=" " type=discrete
              discreteopts=(tickvaluefitpolicy=stagger
              tickvaluelist=(&tickval.) ))

   yaxisopts=(label=" "
              linearopts=(viewmin=0 viewmax=%eval(&ymax. + &laboff.)
              tickvaluesequence=(start=0 end=&ymax increment=&yinc)))

   x2axisopts=(display=(line) ) ;

      entry "Non Hispanic White" /
         textattrs=(weight=bold) valign=top;

   barchartparm y=y8 x=x8 / barwidth=1 dataskin=crisp colorresponse=pcol
                            name="heatmap";

   scatterplot y=ylab8 x=x8 /
               xaxis=x2 markercharacter=lab8 labelstrip=true
               markercharacterattrs=(size=8 weight=bold );

   continuouslegend "heatmap" /
      title="Percentage of Trials Achieving (*ESC*)
            {unicode '000a'x} Census Level"
      location=outside halign=right valign=center titleattrs=(size=8 );

endlayout;
```

```
   endlayout
   endgraph;
   end;
run;
```

Appendix 5 – Template for Summary Figure 5

```
proc template;
   define style graphtmpl;
      parent=styles.statistical;
      class GraphFonts /
         'GraphDataFont'=("<sans-serif>, <MTsans-serif>",9pt)
         'GraphUnicodeFont'=("<MTsans-serif-unicode>",10pt)
         'GraphValueFont'=("<sans-serif>, <MTsans-serif>",8pt)
         'GraphLabelFont'=("<sans-serif>, <MTsans-serif>",8pt)
         'GraphFootnoteFont'=("<sans-serif>, <MTsans-serif>",8pt)
         'GraphTitleFont'=("<sans-serif>, <MTsans-serif>",10pt,bold)
         'GraphAnnoFont'=("<sans-serif>, <MTsans-serif>",8pt);
      class GraphData1 /  color=red    contrastcolor=red;
      class GraphData2 /  color=blue   contrastcolor=blue;
   end;
run;

proc template;
   define statgraph splot;

   begingraph / designwidth=13in designheight=10in border=no  ;
     layout lattice / rows=1 columns=3 columnweights=(0.40 0.30 0.30) ;

   sidebar / align=top ;
      entry "Number of Subjects needed to Achieve Census Level" /
      textattrs=GraphTitleText ;
   endsidebar;

   layout overlay /
      xaxisopts=(label="Number of Subjects"  type=linear )

      yaxisopts=(label=" " reverse=1 type=discrete
                 discreteopts=(tickvaluefitpolicy=split
                 tickvaluesplitchar='@') offsetmin=0.1);

         entry "&g1" /
            textattrs=(weight=bold size=12) valign=top opaque=true;

      barchart x=y1 y=val1 /
         orient=horizontal group=stat1 groupdisplay=cluster clusterwidth=.7
         barwidth=.9 dataskin=crisp
         barlabel=true barlabelfitpolicy=none;

   endlayout;

   layout overlay /
      xaxisopts=(label="Number of Subjects"  type=linear)

      yaxisopts=(display=(line ticks)  reverse=1 offsetmin=0.1);

         entry "&g2" /
            textattrs=(weight=bold size=12) valign=top opaque=true;

      barchart x=y2 y=val2 /
```

```
      orient=horizontal group=stat2 groupdisplay=cluster clusterwidth=.7
      barwidth=.9 dataskin=crisp
      barlabel=true barlabelfitpolicy=none;

   endlayout;

   layout overlay /
      xaxisopts=(label="Number of Subjects"  type=linear)

      yaxisopts=(display=(line ticks)  reverse=1 offsetmin=0.1);

         entry "&g3" /
            textattrs=(weight=bold size=12) valign=top opaque=true;

      barchart x=y3 y=val3 /
         orient=horizontal group=stat3 groupdisplay=cluster clusterwidth=.7
         barwidth=.9 dataskin=crisp
         barlabel=true barlabelfitpolicy=none name="bar";

      DiscreteLegend "bar" /title=" " EXCLUDE=(" " ".") across=1
                            location=inside halign=right valign=top;

   endlayout;

   endlayout;
   endgraph;
   end;
run;
```

Appendix 6 – Template for Summary Figure 6

```
proc template;
   define style graphtmpl;
      parent=styles.statistical;
      class GraphFonts /
         'GraphDataFont'=("<sans-serif>, <MTsans-serif>",9pt)
         'GraphUnicodeFont'=("<MTsans-serif-unicode>",10pt)
         'GraphValueFont'=("<sans-serif>, <MTsans-serif>",8pt)
         'GraphLabelFont'=("<sans-serif>, <MTsans-serif>",8pt)
         'GraphFootnoteFont'=("<sans-serif>, <MTsans-serif>",8pt)
         'GraphTitleFont'=("<sans-serif>, <MTsans-serif>",10pt,bold)
         'GraphAnnoFont'=("<sans-serif>, <MTsans-serif>",8pt);
      class GraphData1 /  color=blue         contrastcolor=blue;
      class GraphData2 /  color=gray         contrastcolor=gray;
      class GraphData3 /  color=orange       contrastcolor=orange;
   end;
run;

proc template;
   define statgraph splot;

   begingraph / designwidth=12in designheight=8in border=no;

   layout lattice / rows=1 rowdatarange=union columns=4
                    columnweights=(0.40 0.2 0.2 0.2) ;

   sidebar / align=top ;
      entry "Bar Chart of US Demographic Characteristics by Subjects,
             Epidemiology, and US Census" /
         textattrs=GraphTitleText pad=(bottom=5px);
   endsidebar;

   layout overlay / walldisplay=(fill)

      xaxisopts=(display=none label=" " type=linear
                 linearopts=(viewmin=0 viewmax=1.05))

      yaxisopts=(label=" " reverse=1 type=discrete
                 discreteopts=(tickvaluefitpolicy=split
                 tickvaluesplitchar='@') offsetmin=0.1  );

         entry "&ecat1." / textattrs=(weight=bold) valign=top;

      barchart x=cat1 y=pct1 / name='bar'
               orient=horizontal   group=bar1    groupdisplay=cluster
               clusterwidth=.7     barwidth=.9   dataskin=crisp
               barlabel=true       barlabelfitpolicy=none
               barlabelattrs=(size=6)   barlabelformat=percent6.1;

      DiscreteLegend "bar" /title=" " EXCLUDE=(" " ".")
                            displayclipped=true across=3
                            location=outside halign=center valign=bottom;

   endlayout;
```

```
   layout overlay / walldisplay=(fill)

      xaxisopts=(display=none label=" " type=linear
                 linearopts=(viewmin=0 viewmax=1.05))

      yaxisopts=(display=none reverse=1 offsetmin=0.1);

         entry "&ecat2." / textattrs=(weight=bold) valign=top;

      barchart x=cat2 y=pct2 /
               orient=horizontal   group=bar2    groupdisplay=cluster
               clusterwidth=.7     barwidth=.9   dataskin=crisp
               barlabel=true       barlabelfitpolicy=none
               barlabelattrs=(size=6)   barlabelformat=percent6.1;
   endlayout;

   layout overlay / walldisplay=(fill)

      xaxisopts=(display=none label=" " type=linear
                 linearopts=(viewmin=0 viewmax=1.05))

      yaxisopts=(display=none reverse=1 offsetmin=0.1);

         entry "&ecat3." / textattrs=(weight=bold) valign=top;

      barchart x=cat3 y=pct3 /
               orient=horizontal  group=bar3    groupdisplay=cluster
               clusterwidth=.7    barwidth=.9   dataskin=crisp
               barlabel=true      barlabelfitpolicy=none
               barlabelattrs=(size=6)  barlabelformat=percent6.1;
   endlayout;

   layout overlay / walldisplay=(fill)

      xaxisopts=(display=none  label=" " type=linear
                 linearopts viewmin=0 viewmax=1.05))

      yaxisopts=(display=none reverse=1 offsetmin=0.1);

         entry "&ecat4." / textattrs=(weight=bold) valign=top;

      barchart x=cat4 y=pct4 /
               orient=horizontal   group=bar4    groupdisplay=cluster
               clusterwidth=.7     barwidth=.9   dataskin=crisp
               barlabel=true       barlabelfitpolicy=none
               barlabelattrs=(size=6)   barlabelformat=percent6.1;

   endlayout;

   endlayout;
   endgraph;
   end;
run;
```