Data Variables Tool: Identifying and Collecting Data Variables

Currently, the collection of data variables as part of clinical research lacks uniformity, limiting the ability to capture results in a granular enough manner to accurately represent diverse populations and thus subsequently analyze within a study and compare across studies aggregate results, and assess heterogeneity of treatment effect across different subgroups. While all variables need not be collected for every research study, those that are dependent upon the nature and objectives of the research study should be collected using data standards that are as universal as possible (see *Achieving Diversity, Inclusion, and Equity in Clinical Research Guidance Document* Section 11.1, Data Variables and Collection, Background). The process by which data variables will be collected and the collection tool used to record data variables should be identified during study design and protocol development. This tool provides:

1) A framework to assist study designers in identifying relevant demographic/non-demographic data elements (Figures 1-3). The framework itself can be applied to any data element that will be collected as part of a research protocol.
2) A Data Collection Tool for baseline demographic variables (Figure 4). The Data Collection Tool serves as a template that sponsors and investigators can adapt and use when creating their own study specific data collection forms. The Data Collection Tool derives from previous work done by Clinical Data Interchange Standards Consortium (CDISC).[1]
3) An Aggregate Reporting Tool template (Figure 5) to be used for categorization and reporting of demographic information to regulatory authorities, oversight bodies and clinical trial registries.

Several important guiding features should be considered throughout this process:

- Data are most useful if collected at the most granular level. For example, when collecting age, a date of birth should be collected versus asking participants to categorize themselves into an age group (e.g., 20-29 years old, 30-39 years old). Data can be categorized and/or aggregated at the end of the study for different purposes, including regulatory submission or publication.
    - Some countries and regions limit the amount of personal data that may be collected. For example, in France there are limitations[2] to collecting date of birth due to privacy laws, in which case the data can be collected as year of birth (collected) and age (collected or derived).
- Demographic data variables should be self-reported. Self-report can mean that the participant completes a data collection form or that the researcher asks the participant a question and then records the answer that is given. Researchers should not assume answers regarding demographic information and should be trained on scripted, standardized methods for collection. Clear instructions in respectful, plain language should be provided to the participant.

[1] See online resources at: www.cdisc.org
[2] PHUSE Data Transparency Working Group – Recommendations for GDPR Compliancy: Version 1.0, 1-Apr-2020: https://www.phusewiki.org/docs/WorkingGroups/Deliverables/Recommendations%20for%20GDPR%20Compliancy-%20PHUSE%20Data%20Transparency%20Working%20Group.pdf [Accessed on 2020-06-10]

- Study designers should be sensitive to cultural distinctions in racial classificationsystems across different regions. For example, it is not allowed to collect "race" data in certain countries (see *Achieving Diversity, Inclusion, and Equity in Clinical Research Guidance Document* Section 11.1, Data Variables and Collection, Background), but is legally required in others.

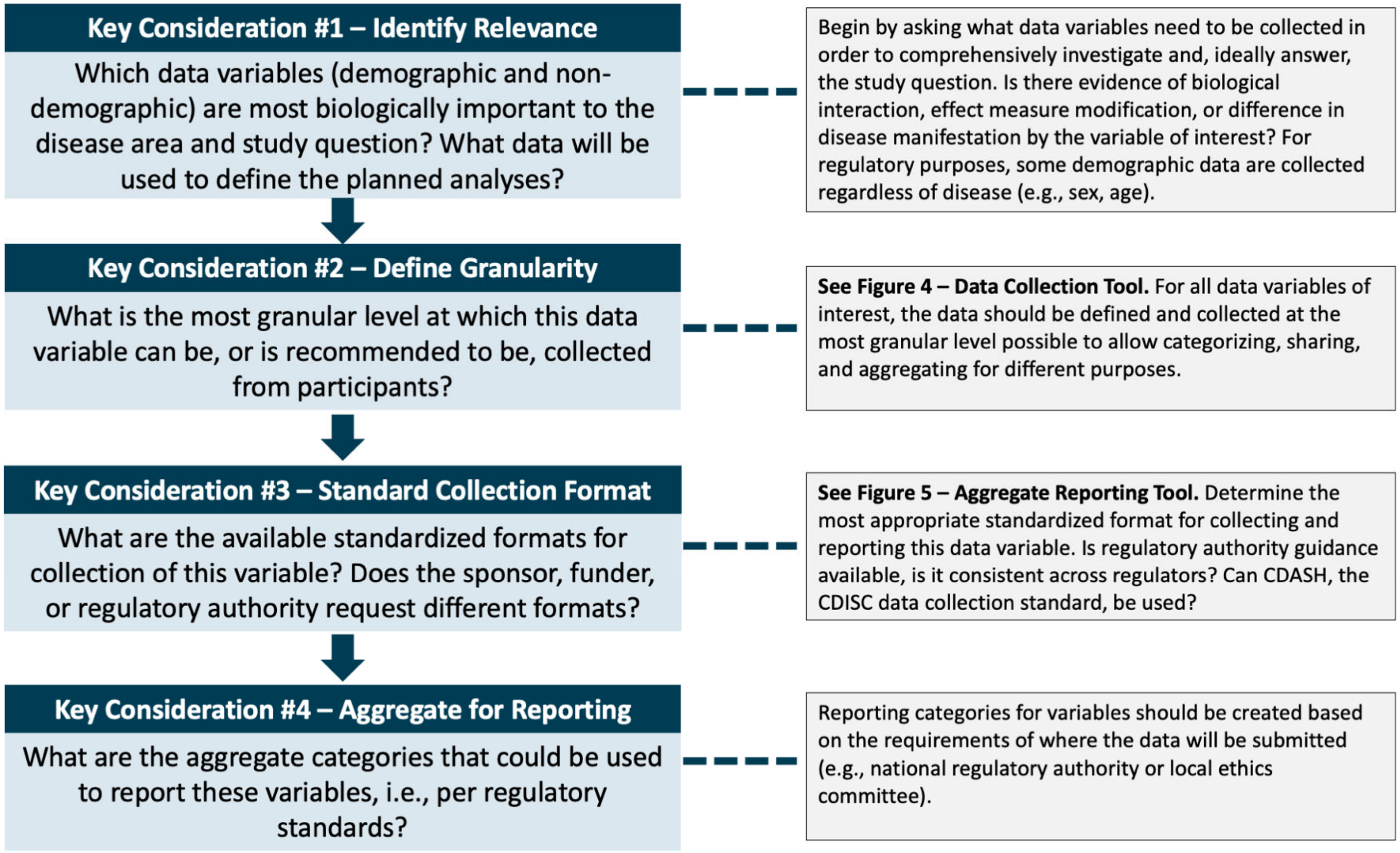
For additional information regarding demographic variables, please see Chapter 11 of the *Achieving Diversity, Inclusion, and Equity in Clinical Research* Guidance Document.

Ultimately, standardized data collection in a common electronic format would permit data to be structured in such a way that could be uploaded directly to regulatory authorities, oversight bodies (e.g., IRBs/RECs), data repositories, and clinical trial registries (e.g. ClinicalTrials.gov, EudraCT and other national registries). We recommend a similar, defined approach be utilized for every category of data and every datum element, with particular attention to whetherthere may be differences in diverse populations.

A four-stage approach to data collection:

Figure 1 annotates a four-stage approach to consider: (a) **which demographic and non- demographic variables** should be collected for a specific protocol; (b) the necessary level of **granularity of the data**; (c) the standardized collection method, tool, and format for **data collection**; and (d) approaches to data aggregation for **reporting**. This framework can be used to assist study designers in identifying relevant demographic and non-demographic data elements that will be collected as part of a research protocol. Two examples of applying this approach are given (Figure 2, 3). Figure 2 (race) is representative of a demographic variable that is well delineated in CDISC standards, while Figure 3 (gender) is an example of an element that is far more sensitive, inconsistent, and dependent on the protocol itself.

Figure 1: A four stage approach to data collection



Key Consideration #1 – Identify Relevance

Which data variables (demographic and non-demographic) are most biologically important to the disease area and study question? What data will be used to define the planned analyses?

Begin by asking what data variables need to be collected in order to comprehensively investigate and, ideally answer, the study question. Is there evidence of biological interaction, effect measure modification, or difference in disease manifestation by the variable of interest? For regulatory purposes, some demographic data are collected regardless of disease (e.g., sex, age).

Key Consideration #2 – Define Granularity

What is the most granular level at which this data variable can be, or is recommended to be, collected from participants?

See Figure 4 – Data Collection Tool. For all data variables of interest, the data should be defined and collected at the most granular level possible to allow categorizing, sharing, and aggregating for different purposes.

Key Consideration #3 – Standard Collection Format

What are the available standardized formats for collection of this variable? Does the sponsor, funder, or regulatory authority request different formats?

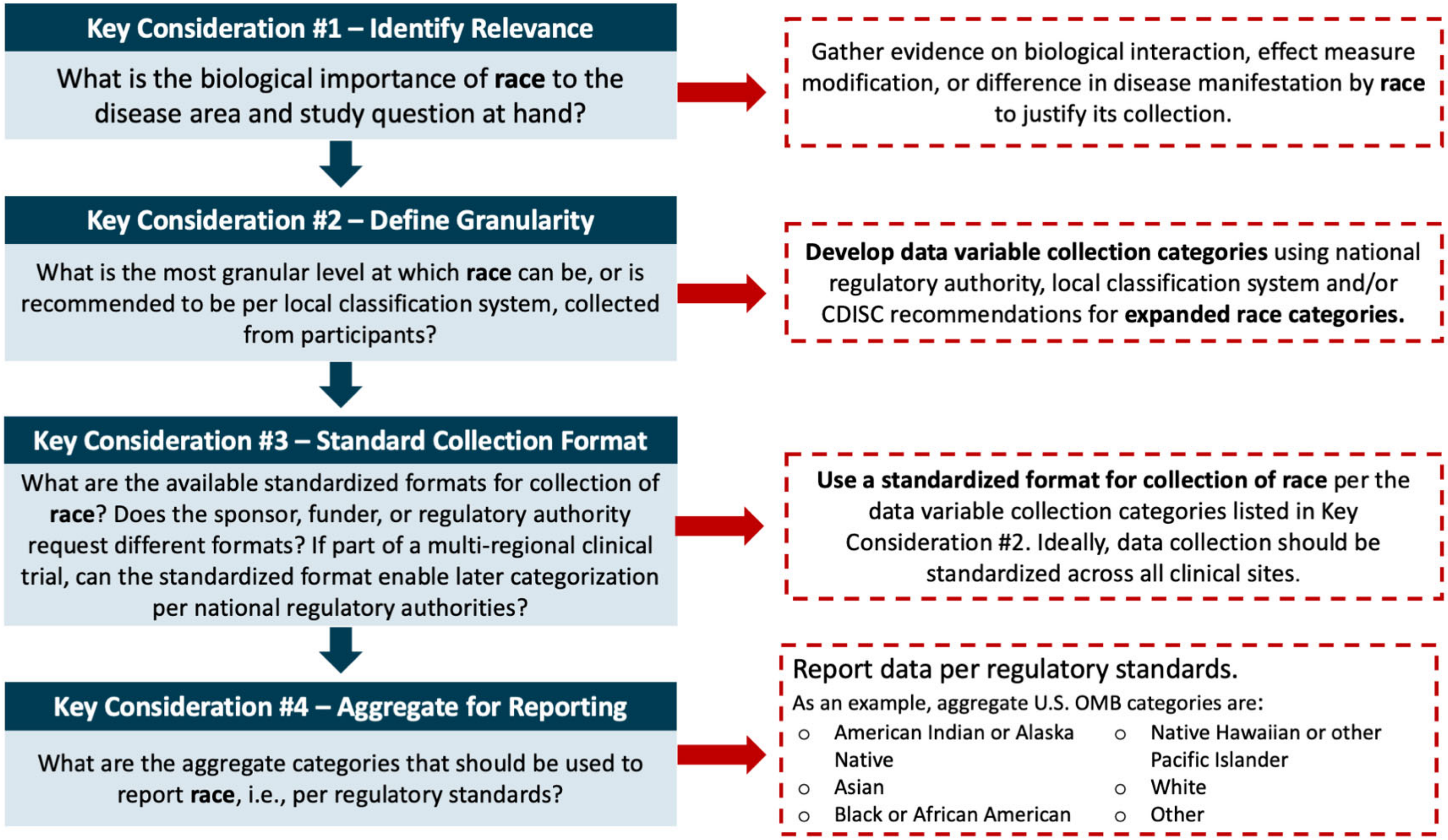
See Figure 5 – Aggregate Reporting Tool. Determine the most appropriate standardized format for collecting and reporting this data variable. Is regulatory authority guidance available, is it consistent across regulators? Can CDASH, the CDISC data collection standard, be used?

Key Consideration #4 – Aggregate for Reporting

What are the aggregate categories that could be used to report these variables, i.e., per regulatory standards?

Reporting categories for variables should be created based on the requirements of where the data will be submitted (e.g., national regulatory authority or local ethics committee).

Figure 2: Key considerations for **race** as a data element during protocol development and study design



Key Consideration #1 – Identify Relevance

What is the biological importance of **race** to the disease area and study question at hand?

Gather evidence on biological interaction, effect measure modification, or difference in disease manifestation by **race** to justify its collection.

Key Consideration #2 – Define Granularity

What is the most granular level at which **race** can be, or is recommended to be per local classification system, collected from participants?

Develop data variable collection categories using national regulatory authority, local classification system and/or CDISC recommendations for **expanded race categories.**

Key Consideration #3 – Standard Collection Format

What are the available standardized formats for collection of **race**? Does the sponsor, funder, or regulatory authority request different formats? If part of a multi-regional clinical trial, can the standardized format enable later categorization per national regulatory authorities?

Use a standardized format for collection of race per the data variable collection categories listed in Key Consideration #2. Ideally, data collection should be standardized across all clinical sites.

Key Consideration #4 – Aggregate for Reporting

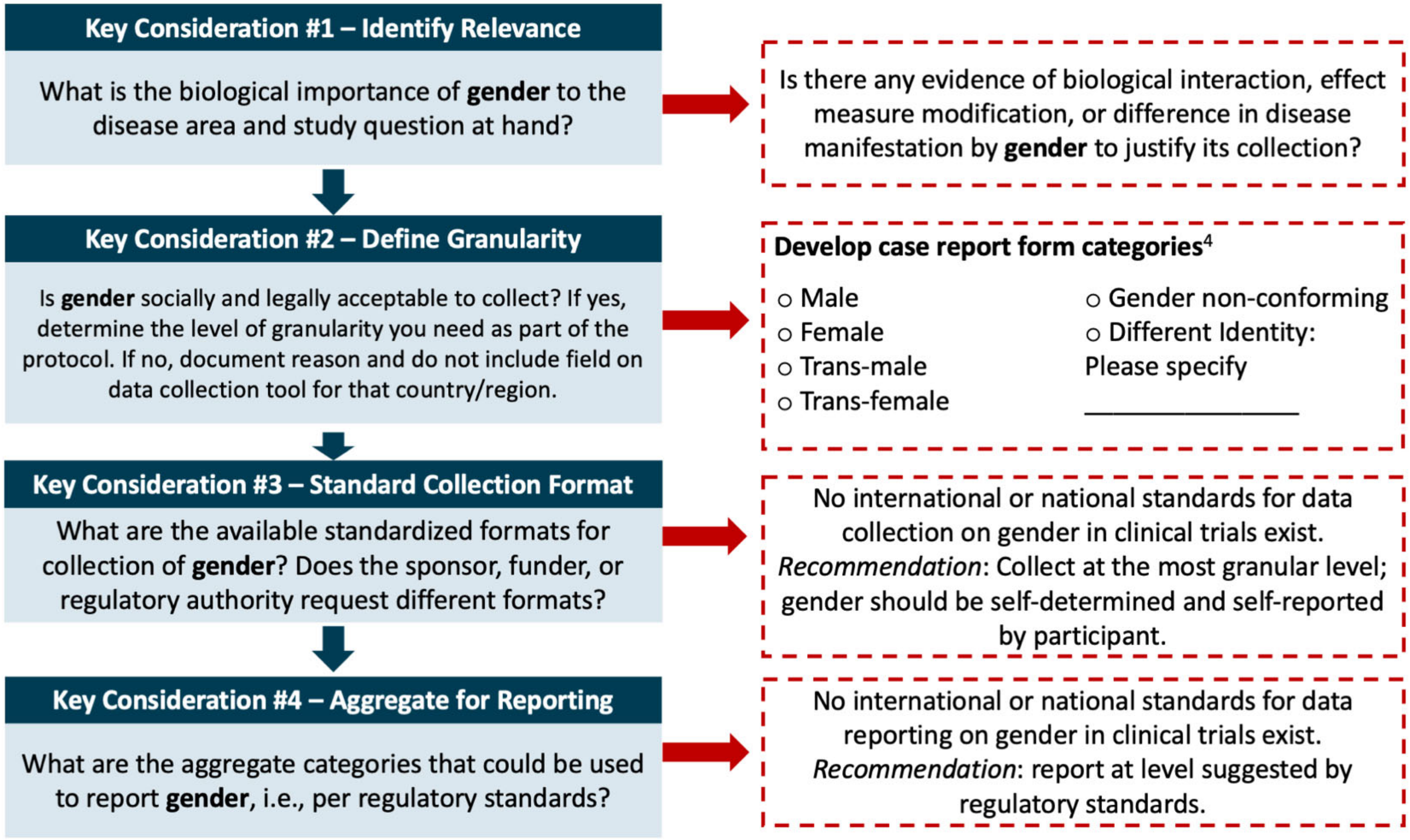
What are the aggregate categories that should be used to report **race**, i.e., per regulatory standards?

Report data per regulatory standards.

As an example, aggregate U.S. OMB categories are:

- American Indian or Alaska Native
- Asian
- Black or African American
- Native Hawaiian or other Pacific Islander
- White
- Other

Figure 3: Key considerations for **gender**[3] as a data element during protocol development and study design



Key Consideration #1 – Identify Relevance

What is the biological importance of **gender** to the disease area and study question at hand?

Is there any evidence of biological interaction, effect measure modification, or difference in disease manifestation by **gender** to justify its collection?

Key Consideration #2 – Define Granularity

Is **gender** socially and legally acceptable to collect? If yes, determine the level of granularity you need as part of the protocol. If no, document reason and do not include field on data collection tool for that country/region.

Develop case report form categories[4]

- ○ Male
- ○ Female
- ○ Trans-male
- ○ Trans-female
- ○ Gender non-conforming
- ○ Different Identity: Please specify ________________

Key Consideration #3 – Standard Collection Format

What are the available standardized formats for collection of **gender**? Does the sponsor, funder, or regulatory authority request different formats?

No international or national standards for data collection on gender in clinical trials exist. *Recommendation*: Collect at the most granular level; gender should be self-determined and self-reported by participant.

Key Consideration #4 – Aggregate for Reporting

What are the aggregate categories that could be used to report **gender**, i.e., per regulatory standards?

No international or national standards for data reporting on gender in clinical trials exist. *Recommendation*: report at level suggested by regulatory standards.

[3] Gender is defined as the socially constructed characteristics of women and men – such as norms, roles and relationships of and between groups of women and men. It varies from society to society and can be changed. World Health Organization. Glossary of terms and tools [Internet]. WHO. Available online: https://www.who.int/gender-equity-rights/knowledge/glossary/en/ (accessed May 07 2020).

[4] Bauer GR, Braimoh J, Scheim AI, Dharma C. Transgender-inclusive measures of sex/gender for population surveys: Mixed-methods evaluation and recommendations. PloS one. 2017 May 25;12(5):e0178043

The Data Collection Tool (Figure 4 below) serves as a template for study designers, including sponsors and investigators, to use when creating study specific demographic data collection forms. The demographic Data Collection Tool derives from previous work done by CDISC[5] unless noted otherwise. As noted above, data variables should be self-reported, meaning that the participant completes a data collection form or that the researcher asks the participant a specific, scripted question and then records the answer that is given. Clear instructions in plain language should be provided to the participant. Researchers should not assume answers regarding demographic information and should be trained on scripted, standardized methods for collection.

The format of this template should be modified as appropriate to the protocol. "Notes" are provided below the demographic variable fields to provide additional clarity in collecting and categorizing the variables.

Figure 4: Data collection tool for baseline demographic variables

Study ID: **Participant Study ID:** **Date of data collection:** (specify MM/DD/YYYY or DD/MM/YYYY)
AGE Instructions: Provide your date of birth to the best of your ability
Date of birth: (specify MM/DD/YYYY or DD/MM/YYYY)
Corresponding Age: (specify units: hours, days, months, years)
Note: • Collect age as a continuous variable, in order to summarize and/or report as required bythe regulatory authority. • Collect age in hours, days, months, years. Age may be grouped into categories to reflect important age-related distinctions or underlying biological differences. • If there are limitations to collecting date of birth (often related to national- or region-specific privacy laws), data can be collected as year of birth and corresponding age. Specify the Age Unit (e.g., years, months). • See Section *Achieving Diversity, Inclusion, and Equity in Clinical Research Guidance Document* 11.2.1 - 11.2.3 regarding data standards for specific age categories including neonates and the elderly.

[5] CDISC: Clinical Data Acquisition Standards Harmonization (CDASH): https://www.cdisc.org/standards/foundational/cdash

ETHNICITY

Instructions: Select one or more ethnicity that you most closely identify with at the high-level category or within the expanded categories. If you do not consider yourself "Hispanic or Latino," select "Not Hispanic or Latino."

Ethnicity	○ Hispanic or Latino	*Expanded Categories:* ○ Central American ○ Cuban ○ Cuban American ○ Latin American ○ Mexican ○ Mexican American ○ South American ○ Spanish
	○ Not Hispanic or Latino	
	○ Not Reported	

Note:

- Ethnicity terminology presented here is specific to U.S. During protocol development, a sponsor (or sponsor-investigator) should identify the classification system(s) for ethnicity, and/ornational origins where trial will be ongoing. Further, understand what is legally or socially acceptable to ask.
- In the U.S., questions regarding race and ethnicity should be asked in a standard order (e.g., questions about ethnicity precede race) with scripted questions. Individuals assigned to collect personal data should be cognizant of geographic variations and cultural sensitivities, asking questions that are locally respectful and internationally meaningful for the research.
- The Ethnicity, Expanded Categories code list is expanded based on CDISC user community requests. CDISC maintains one overall ethnicity code-list that is categorized as either "Hispanic or Latino" or "Not Hispanic or Latino." The code table is available for download from the CDISC.org terminology page here: https://www.cdisc.org/standards/terminology, login required.
- See Section 11.3 of the *Achieving Diversity, Inclusion, and Equity in Clinical Research* Guidance Document for reporting race and ethnicity to U.S. and ex-U.S. regions.

RACE

Instructions: Select one or more race that you most closely identify with at the high-level category or within the sub-category.

Race	○ American Indian or Alaska Native	*Expanded categories*: ○ Alaska Native ○ American Indian ○ Caribbean Indian ○ Central American Indian ○ Greenland Inuit ○ Nupiat Inuit ○ Siberian Eskimo ○ South American Indian ○ Yupik Eskimo
	○ Asian	*Expanded categories*: ○ Asian American ○ Asian Indian ○ Bangladesh ○ Bhutanese, Burmese ○ Malagasy ○ Malaysian ○ Maldivian ○ Mongolian ○ Nepalese

		○ Cambodian ○ Chinese ○ Filipino ○ Hmong ○ Indonesian ○ Iwo Jiman ○ Japanese ○ Korean ○ Laotian ○ Okinawn ○ Pakistani ○ Singaporean ○ Sri Lankan ○ Taiwanese ○ Thai ○ Vietnamese
	○ Black or African American	*Expanded categories*: ○ African ○ African American ○ African Caribbean ○ Bahamian ○ Barbadian ○ Black Central American ○ Black South American ○ Botswanan ○ Dominica Islander ○ Dominican ○ Ethiopian ○ Haitian ○ Jamaican ○ Liberian ○ Malagasy ○ Namibian ○ Nigerian ○ Trinidadian ○ West Indian ○ Zairean
	○ Native Hawaiian or Other Pacific Islander	*Expanded categories*: ○ Melanesian ○ Micronesian ○ Polynesian
	○ White	*Expanded categories*: ○ Arab ○ Eastern European ○ European ○ Mediterranean ○ Middle Eastern ○ North American ○ Northern European ○ Russian ○ Western European ○ White Caribbean ○ White Central American ○ White South American
	○ Other	*Expanded categories*: ○ Unknown ○ Not reported
Note: • Race terminology presented here is specific to U.S. During protocol development, identify the classification system(s) based on race and/or national origins where trial will be ongoing.Further, understand what is legally or socially acceptable to ask.		

<table>
<tr><td colspan="2">• In the U.S., questions regarding race and ethnicity should be asked in a standard order (e.g., questions about ethnicity precede race) with scripted questions. Individuals assigned to collect personal data should be cognizant of geographic variations and cultural sensitivities, asking questions that are locally respectful and internationally meaningful for the research.
• See Section 11.3 of the <i>Achieving Diversity, Inclusion, and Equity in Clinical Research</i> Guidance Document for regulatory guidance on reporting race and ethnicity to U.S. and ex-U.S. regions.</td></tr>
<tr><td colspan="2"><b>SEX</b>
Instructions: Select your biological sex at birth. Sex is defined as the different physiological and biological characteristics of males and females, such as reproductive organs, chromosomes, hormones, etc.[6]</td></tr>
<tr><td>SEX</td><td>○ Male
○ Female
○ Unknown or undifferentiated. Intersex is included in the term undifferentiated.</td></tr>
<tr><td colspan="2"><b>GENDER</b>
Instructions: Select the gender you most closely identify with. Gender is defined as the socially constructed characteristics of women and men – such as norms, roles and relationships of and between groups of women and men. It varies from society to society and can be changed.[7]</td></tr>
<tr><td>GENDER[8]</td><td>○ Male
○ Female
○ Trans-male
○ Trans-female
○ Gender non-conforming
○ Different Identity: Please specify___________
○ Chose to not answer the question</td></tr>
<tr><td colspan="2">Note:
• The collection of gender is sensitive. The individual collecting this information should be sensitive that this may make a participant uncomfortable and use scripted questions to ensure questions are asked in a respectful way.</td></tr>
</table>

Figure 5: **Aggregate reporting tool**

The Aggregate Reporting Tool is used to categorize and report demographic information to regulatory authorities, oversight bodies and clinical trial registries. The specific demographic variables listed and the individual categories reported should be developed according to regulatory standards to which data will be submitted or reported and should be identified during the development of the protocol and statistical analysis. The Aggregate Reporting Tool is populated by the more granular Data Collection Tool (Figure 4), therefore development of both tools prior to study conduct is important to ensure efficient collection and categorization of demographic data. The tool created below serves as an example for

[6] World Health Organization. Glossary of terms and tools. Accessible at https://www.who.int/gender-equity-rights/knowledge/glossary/en/.

[7] World Health Organization. Glossary of terms and tools. Accessible at https://www.who.int/gender-equity-rights/knowledge/glossary/en/.

[8] Adapted from: Bauer GR, Braimoh J, Scheim AI, Dharma C. Transgender-inclusive measures of sex/gender for population surveys: Mixed-methods evaluation and recommendations. PloS one. 2017 May 25;12(5):e0178043.

categorizing previously collected demographic data and includes the demographic categories described in Chapter 11, Data Variables and Collection, of the *Achieving Diversity, Inclusion, and Equity in Clinical Research* Guidance Document. The tool is currently designed for a global study enrolling participants over the age of 18 years old. It is designed according to U.S. regulatory standards. Additional categories can be included based on the specific protocol and study population (e.g. region of enrollment, language, etc.).

Study ID:

Baseline Demographics, Aggregated Data

Demographic Variables	**Treatment Group(s)**		**Control Group**	**Total**
	Group 1, N (%)	**Group 2, N (%)**	**N (%)**	**N**
Age				
>=18 - <65 years				
>=65 - <74 years				
>=75 - <84years				
>= 85 years				
Sex				
Male				
Female				
Unknown/Undifferentiated				
Gender				
Male Gender				
Female Gender				
Trans-Male				
Trans-Female				
Gender Nonconforming/ Unknown				
Ethnicity				
Hispanic or Latino				
Not Hispanic or Latino				
Not Reported				
Race				
White				
Black or African American				
Asian				
American Indian or Alaska Native				
Native Hawaiian or Other Pacific Islander				
Not reported/unknown				
Other/More than one				