

Background: The U.S. Food and Drug Administration (FDA) has a large repository of clinical trial data from regulatory submissions of new drug applications. However, these data exists in disparate formats, with data structure, terminology, and data elements unique to each application. As the FDA increasingly looks for safety signals across applications, overcoming the challenge of assembling data - not originally designed to go together - becomes more important.

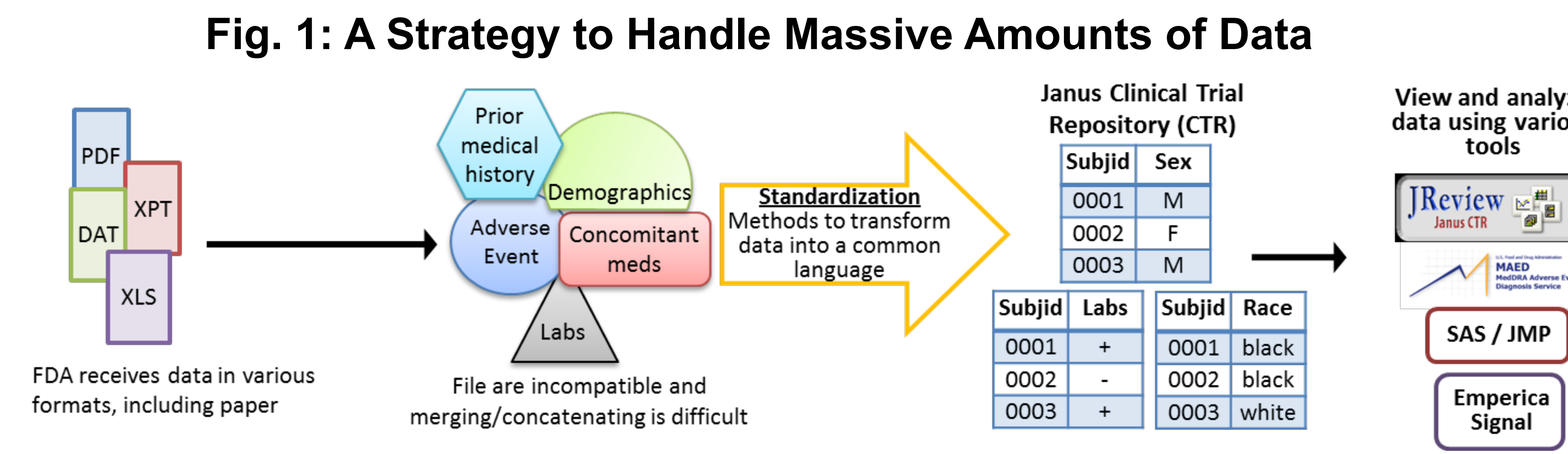
The Office of Translational Sciences in the Center for Drug Evaluation and Research (CDER) at FDA recently converted data from 51 pivotal clinical diabetes trials submitted in support of new diabetes drug applications into a standardized format to facilitate patient-centered outcomes research (PCOR). This initiative was funded through the American Recovery and Re-Investment Act (ARRA) to test the usability of converted and stacked data from multiple applications. Data structure was harmonized to a single standard (CDISC, SDTM 3.1.2) to enable an aggregated trial database. We describe the technical challenges and benefits of aggregating data from 6 clinical trials conducted for a diabetes drug that belong to a single drug class.

Methods: We used a subset of 6 of the 51 trials to describe the demographic characteristics of adults patients enrolled in diabetes trials for a single drug class (Thiazolidinedione), stratified by geographic study region (U.S., U.S. and non-U.S., and Non-U.S. studies). Using a scientific workstation with sufficient computing capability, we created SAS code (v9.3), to join Analysis Data Model (ADaM subject-level, ADSL and laboratory, ADLB) datasets within an individual trial submission. This resulted in an aggregated dataset with a structure of one record per observation; with the subject (USUBJID) as the unit of observation. We describe the demographic characteristics of patients enrolled in these 6 clinical trials.

Results: We performed a demographic subset analysis and examined baseline laboratory test by study region. Using the aggregated ADaM database (N=9002), we found that 62% of study participants were enrolled in Non-US studies and, on average, were older (Mean ± SD; 61.8y ± 7.8) than their US and US and non-US counterparts. Across all 3 regions, a majority of study participants were male and classified as White race. Limited data was collected on the enrollment of American Indian/ Alaskan Natives and Hispanic individuals. On average, participants in the US study sites reported a higher baseline BMI (Mean ± SD; 32.8 kg/m2 ± 5.4) relative to their counterparts.

Table 1. Demographic Characteristics of Pioglitazone Studies by Region

	US Studies (N=2853)	US and non-US Studies (N=547)	Non-US Studies (N=5602)
Age (Mean ± SD)	55.6 ± 10.3	59.8 ± 9.3	61.8 ± 7.8
Sex, N (%)			
Male	1658 (58.1)	369 (67.5)	3474 (62.0)
Female	1195 (41.9)	178 (32.5)	1783 (31.8)
Missing			345 (6.2)
Study Population, N (%)			
Study 341	763 (26.7)	--	--
Study 342	883 (31.0)	--	--
Study 343	745 (26.1)	--	--
Study 518	462 (16.2)	--	--
Study 516	--	547 (100.0)	--
Study 444	--	--	5602 (100.0)
Race, N(%)			
American Indian/ Alaskan Natives	15 (0.5)	5 (0.9)	--
Asian	333 (11.7)	28 (5.1)	52 (0.9)
Black	582 (20.4)	58 (10.6)	17 (0.3)
Other	44 (1.5)	--	5 (0.1)
White	1802 (63.2)	322 (58.9)	5164 (92.2)
Missing	77 (2.7)	134 (24.5)	364 (6.5)
Hispanic Ethnicity, N (%)			
Yes	76 (2.7)	134 (24.5)	--
No	418 (14.7)	413 (75.5)	--
Missing	1550 (82.7)	--	--
Baseline BMI (Mean ± SD)*	32.8 ± 5.4	32.0 ± 5.2	30.9 ± 4.8



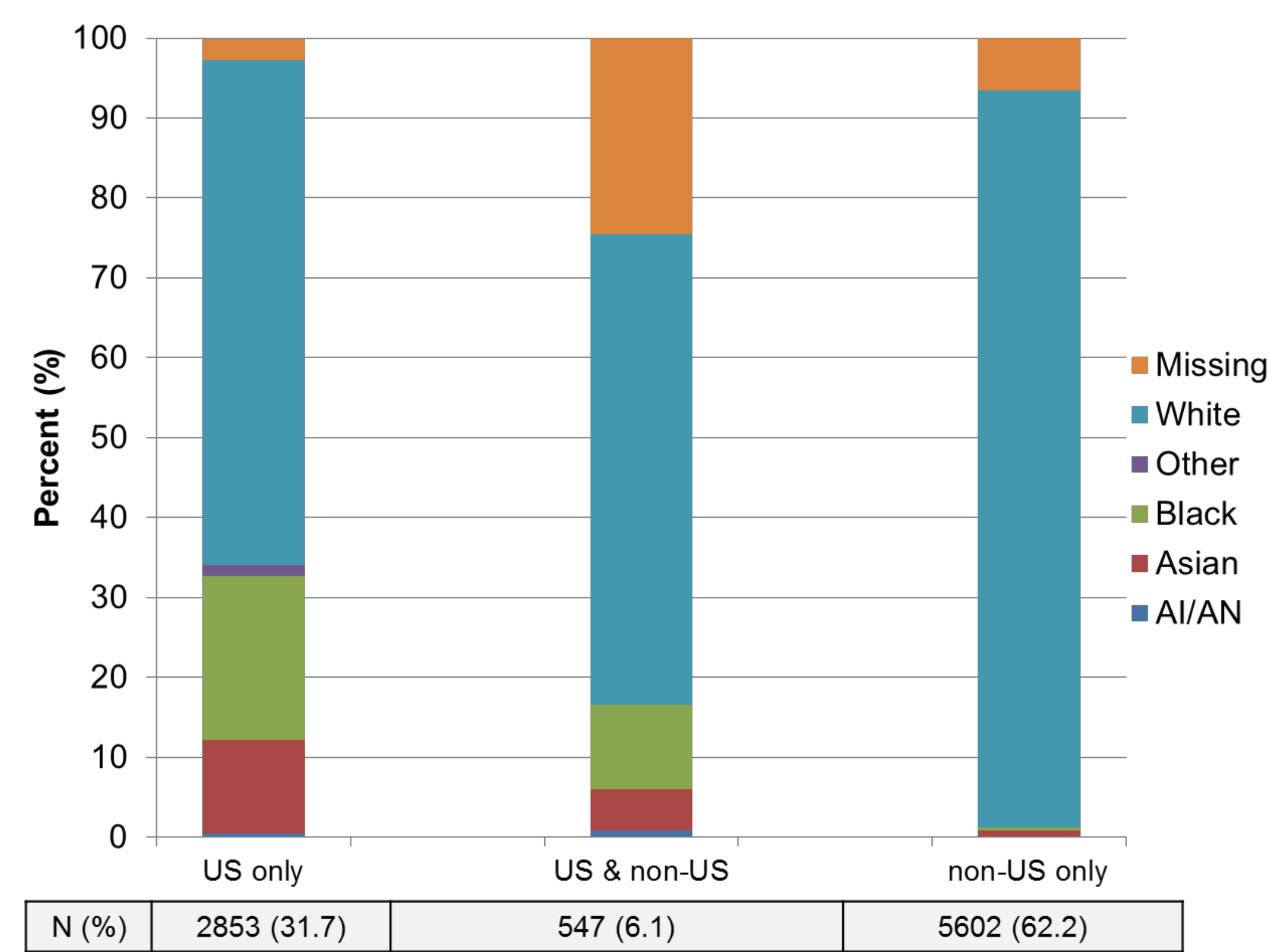
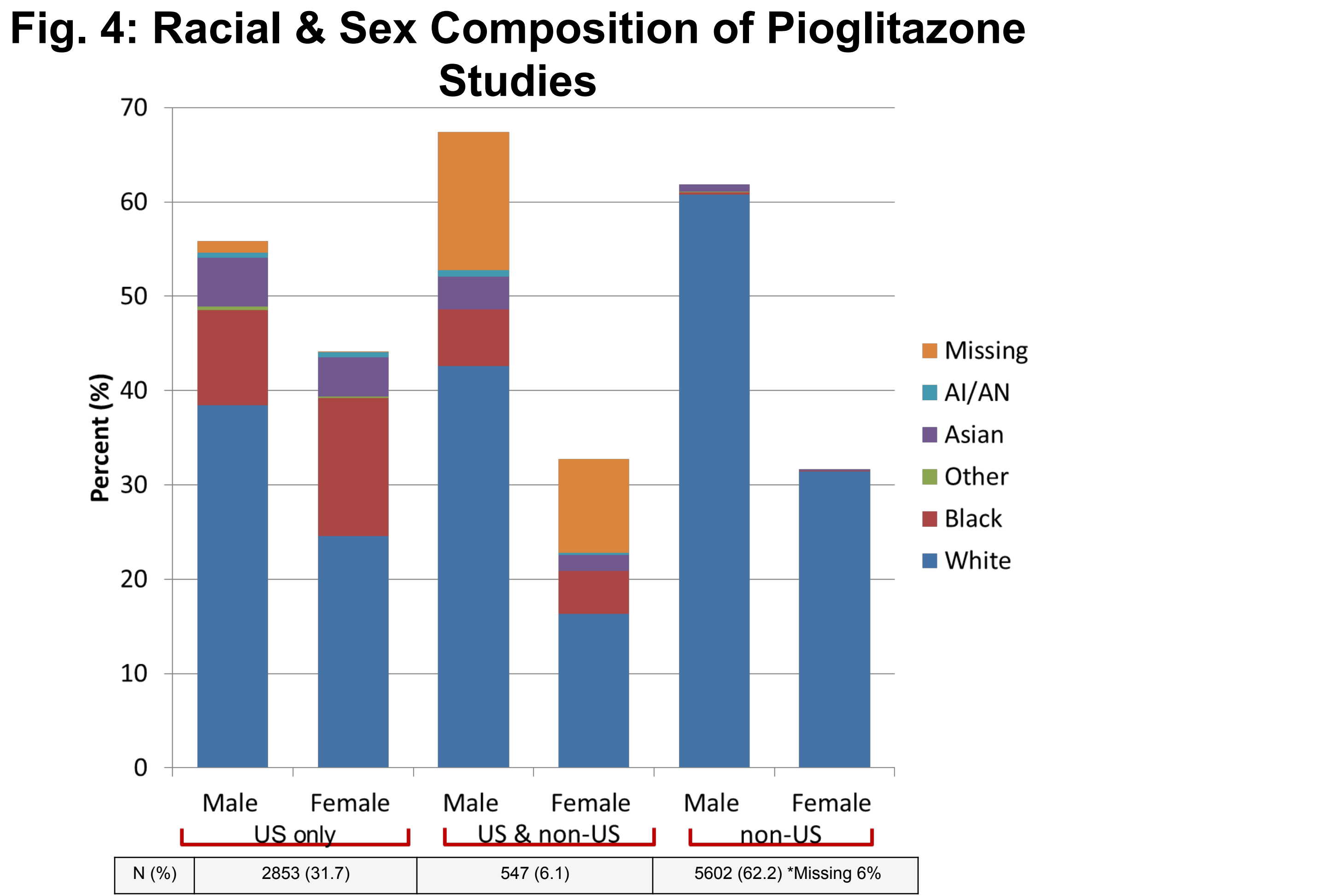
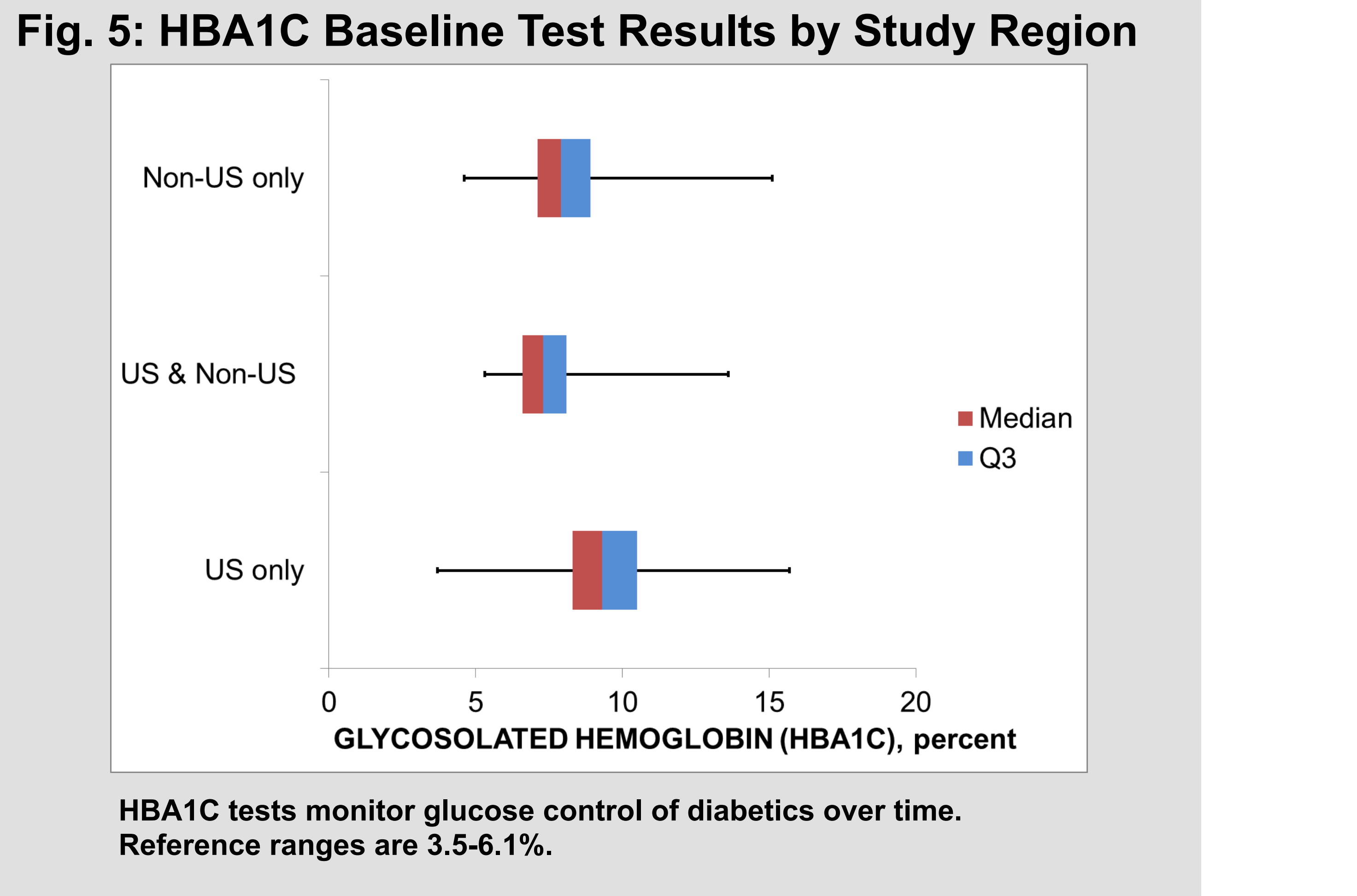
FDA receives a vast amount of data in disparate formats, including paper/pdf. Because these files are not systematically coded or even compatible, merging or concatenating them poses challenges for data users. By applying standardized methods to convert data into a common language, information can be entered and stored into Janus Clinical Trial Repository (CTR). CTR is a data warehouse that transforms data into a usable format recognized by various statistical software programs and tools, such as JReview, MAED, SAS, JMP, etc. Having the data in a standardized format provides a systematic way to address important scientific and regulatory questions.

Fig. 2: Approaches for Using Standardized Data

DOMAINS									
DM					AE		LB		
ID	SITE	ARM	AGE	SEX	AE1	AE2	LAB TEST	RESULT	LAB TEST UNIT
001	USA	PIOG	44	M	1	0	HBA1C	7	%
002	USA	PLACEBO	37	F	0	0	HBA1C	7.5	%
003	USA	PIOG	35	F	1	1	HBA1C	8	%
004	USA	PLACEBO	40	M	0	1	HBA1C	10	%
005	USA	PLACEBO	42	F	0	1	HBA1C	9	%

Highlighted in yellow are domains. Domains are conserved two letter naming conventions used to identify a SDTM formatted data sets. To create the dataset shown, data from the grey columns would come from DM domain, data in the blue-colored columns would come from the adverse events, and green from LB.

Fig. 3: Racial Inclusion by Region of Pioglitazone Studies (N=9002)



Lessons Learned:

- Able to join multiple SDTM domains however, ADaM datasets facilitate cross study analyses
- Secondary use of these data should be well planned
- Prior to conducting analyses, clearly define research inquiry and strategy

Conclusions:

- Analysis validates successful data integration
- Proof of concept for aggregated clinical trial datasets that lend themselves to analytic re-purposing
- Populations in these 6 diabetes trials do not reflect the full range of patients with the disease in the US populations
 - women and racial /ethnic minorities are under-represented but are likely to be treated with these products.
- To ensure safety and efficacy of marketed therapies, strategies should be identified to increase participation of these subgroups in clinical research

Acknowledgements: Dr. Shondelle Wilson-Frederick would like to thank the FDA Commissioner's Fellowship and the FDA Office of Minority Health for supporting this research