They Call Me Whello Yello: Revisiting the SEER Race and Nationality Descriptions

Francis P. Boscoe, Laura E. Soloway, New York State Cancer Registry

Background

The Division of Vital Statistics of the National Center for Health Statistics and the Census Bureau each maintain a list of race recodes for write-in responses to the race question. For example, if "Italian" is written in, this is recoded as white. Appendix D of the SEER Program Coding and Staging Manual consists of an approximate union of these two lists. It is to be used "when race is not stated but other information is provided in the medical record", either in the birthplace field or in a text field. Some birthplace codes are also governed by the NAACCR Asian/Pacific Islander Identification Algorithm (NAPIIA).

Appendix D embeds a number of obsolete, obscure, non-specific, overly-specific, and undefinable terms. For example, "Bavarian" appears, but not "Saxon" or any other German region; "Muganda" but not any of the other dozens of ethnic groups from Uganda. A Google search of the terms "Whello" and "Yello" locates only Appendix D and other similarly-derived lists used in public health data systems - the terms have no independent meaning. In a 2002 paper Laws and Hecksher concluded that terms such as these were generally harmless because they are never actually applied, but that is still no reason to maintain them indefinitely.

Methods

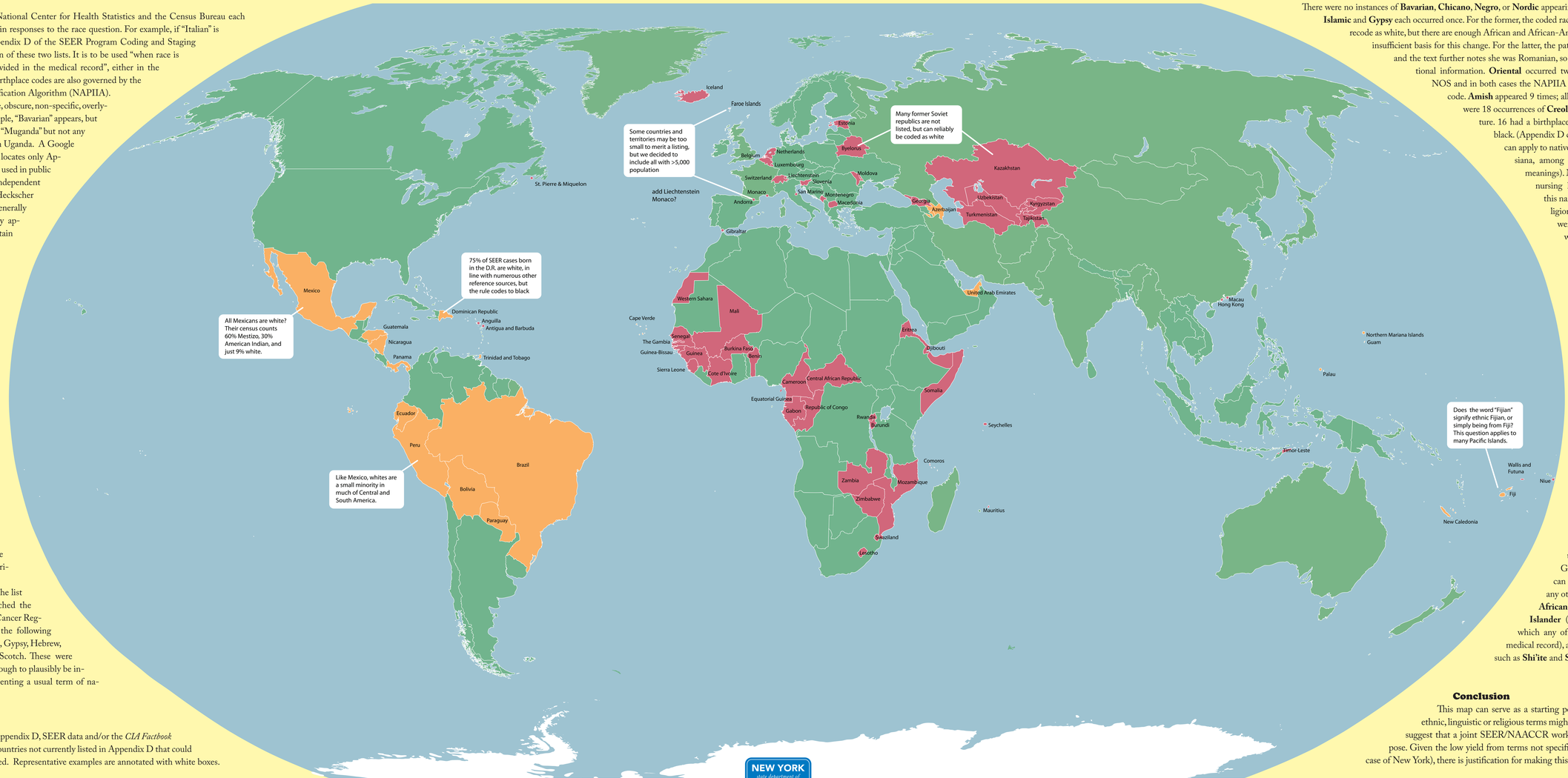
We assessed the validity of Appendix D by comparing it with race/ethnicity distributions for each country given in the current *CIA Factbook* and with race-birthplace cross tabulations from the SEER public-use file (1973-2008). While they measure different things -- the *CIA Factbook* reflects current race distributions within each country; the SEER data reflect the race distributions of pre-1960s emigrants who later developed cancer -- they were generally in agreement.

In order to assign a race based on national origin, we required a single-race percentage above 85% in both the SEER data and the *CIA Factbook*, after disregarding any mixed or multiple-race categories. Lest this seem an overly strict criterion, we note that the percentage of whites in the United States is 80%, and no one would seriously propose coding all Americans of unknown race to white.

We also identified additional entries on the list of questionable utility. Finally, we searched the source text fields of the New York State Cancer Registry (NYSCR) for any appearance of the following terms: Amish, Bavarian, Chicano, Creole, Gypsy, Hebrew, Islamic, Negro, Nordic, Oriental and Scotch. These were chosen because they seemed common enough to plausibly be included in a case report while not representing a usual term of national origin.

Results

Countries with disagreements between Appendix D, SEER data and/or the *CIA Factbook* are shaded orange in the map at right. Countries not currently listed in Appendix D that could potentially be added to it are shaded in red. Representative examples are annotated with white boxes.



There were no instances of **Bavarian**, **Chicano**, **Negro**, or **Nordic** appearing in any text field in the NYSCR. **Islamic** and **Gypsy** each occurred once. For the former, the coded race was "other". The rule directs us to recode as white, but there are enough African and African-American Muslims that this seems an insufficient basis for this change. For the latter, the patient was already reported as white, and the text further notes she was Romanian, so the term Gypsy provided no additional information. **Oriental** occurred twice; both were reported as Asian NOS and in both cases the NAPIIA algorithm provides a more specific code. **Amish** appeared 9 times; all were also reported as white. There were 18 occurrences of **Creole**, all referring to language, not culture. 16 had a birthplace of Haiti and all were reported as black. (Appendix D codes Creole as white, and the term can apply to natives of West Africa, Belize, and Louisiana, among other places, with very different meanings). Most references to **Hebrew** were to nursing homes and a hospital containing this name. Of 14 references to Hebrew religion and 3 to Hebrew language, all were reported as white. Finally, there were over 300 occurrences of Scotch but all referred to alcohol consumption and none to ethnicity.

In sum, these nine terms did not yield a single improvement in race coding in the NYSCR.

In addition to **Whello** and **Yello**, other examples of obsolete and/or undefinable terms included **Brava/Bravo**, **Bilalian**, **Celebesian**, **Ceram**, **Ebian**, **Hamitic**, **Marshenese**, **Morena**, and **Nigritian**. Other problematic terms included **Basque**, **Chechnyan**, **Assyrian**, and **Nubian** (it is unclear why these but not other far more common ethnic terms would be listed), **Cayenne**, **Nassau**, and **Santo Domingo** (presumably referring to natives of the capitals of French Guiana, Bahamas, and the Dominican Republic, but this is not done for any other countries), **Other Arab**, **Other African**, **Other Asian**, and **Other Pacific Islander** (it is hard to imagine a situation in which any of these phrases would appear in a medical record), and further religious designmations such as **Shi'ite** and **Sunni**.

Conclusion

This map can serve as a starting point for revisiting which national, ethnic, linguistic or religious terms might be used to assign a race code. We suggest that a joint SEER/NAACCR work group be convened for this purpose. Given the low yield from terms not specific to a national origin (zero in the case of New York), there is justification for making this list more exclusive than inclusive.

NEW YORK state department of HEALTH