

International Journal of Population Data Science

Journal Website: www.ijpds.org



Understanding the quality of ethnicity data recorded in health-related administrative data sources compared with Census 2021 in England

Cameron Razieh^{1,2}

¹Office for National Statistics, Newport, United Kingdom

²University of Leicester, Leicester, United Kingdom

Objectives

Electronic health records (EHR) are increasingly used to investigate health inequalities across ethnic groups. Whilst there are some studies showing that the recording of ethnicity in EHR is imperfect, there is no robust evidence on the accuracy between the ethnicity information recorded in various real-world sources and census data.

Conclusions

Certain ethnic categories across all data sources have high discordance with census ethnic categories. These differences may lead to biased estimates of differences in health outcomes between ethnic groups, a critical data point used when making health policy and planning decisions.

Methods

We linked primary and secondary care NHS England data sources with Census 2021 data and compared individual-level agreement of ethnicity recording in General Practice Extraction Service (GPES) Data for Pandemic Planning and Research (GDPPR), Hospital Episode Statistics (HES), Ethnic Category Information Asset (ECIA) and Talking Therapies for anxiety and depression (TT) with ethnicity reported in the census. Census ethnicity is self-reported and, therefore, regarded as the most reliable population-level source of ethnicity recording. We further assessed the impact of multiple approaches to assigning a person an ethnic category. The number of people that could be linked to census from ECIA, GDPPR, HES and TT were 47.4m, 43.5m, 47.8m and 6.3m, respectively.

Results

Across all four data sources, the White British category had the highest level of agreement with census ($\geq 96\%$), followed by the Bangladeshi category ($\geq 93\%$). Levels of agreement for Pakistani, Indian, and Chinese categories were $\geq 87\%$, $\geq 83\%$ and $\geq 80\%$ across all sources. Agreement was lower for Mixed ($\leq 75\%$) and Other ($\leq 71\%$) categories across all data sources. The categories with the lowest agreement were Gypsy or Irish Traveller category ($\leq 6\%$), Other Black ($\leq 19\%$) and Any Other Ethnic Group ($\leq 25\%$). Different methodologies to assign ethnic category resulted in different coverage between data sources, particularly in HES, with increased coverage seen in HES when reallocating Not Stated and Not Known categories.

