Race and ethnicity variables are commonly used in research using administrative claims data sets. Recently updated guidelines on the use of race and ethnicity in medical journals specify that the methods section of articles “should include an explanation of who identified participant race and ethnicity and the source of the classification used.” Administrative claims data sets, consisting of insurance claims and/or electronic health record data, are a frequent and increasingly used source for medical research. A MEDLINE search of titles and abstracts from cohort and case-control studies using large, national-level administrative claims data sets shown in the Table returns more than 10,000 articles in the past 5 years.

Analyses in administrative claims data sets frequently include race and ethnicity, when available, in analytic models as covariables (eg, confounders) or as part of the primary research question (eg, comparing outcomes by race). However, the articles reporting such analyses do not always include an explanation of how and by whom participant race and ethnicity were determined.

Table. US National Administrative Data Race and Ethnicity Derivation

<table>
<thead>
<tr>
<th>National claims data set</th>
<th>Race and ethnicity data source</th>
<th>Validation status</th>
<th>Percentage missing race and ethnicity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anthem Digital Data Sandbox</td>
<td>Does not contain race or ethnicity data</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>IBM MarketScan Commercial Database</td>
<td>Does not contain race or ethnicity data</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>Medicare</td>
<td>(1) Common Medicare Environment* (from Social Security)</td>
<td>Validation published</td>
<td>0.3% Missing or unknown from Enrollment Database* (2005)</td>
</tr>
<tr>
<td></td>
<td>(2) RTP algorithm*</td>
<td>NA</td>
<td></td>
</tr>
</tbody>
</table>
| Medicaid                         | Patient (self-reported) with missing data addressed using language, geography, and matching with other data sources* | Validations not published* | As of 2019:
  • <10%: 17 States
  • 10%-20%: 13 States
  • 20%-50%: 15 States and Washington, DC
  • >50%: 5 States |
| Optum Clininformatics Data Mart  | Algorithm based on geography and name                                                           | Unvalidated       | 26% (2000-2016)                       |
| TriNetX                          | Electronic health record data                                                                  | Unvalidated       | Unknown                               |
| Veterans Health Administration   | Patient (self-reported)                                                                       | NA                | 8% (Fiscal year 2018)                 |

Abbreviations: NA, not applicable; RTP, Research Triangle Institute.

* Common Medicare Environment is the source for Medicare beneficiary enrollment and demographic data and includes information from the Enrollment Database, which was used for race and ethnicity prior to 2017. This race and ethnicity variable is based on race and ethnicity data from the Social Security Administration’s Master Beneficiary Records obtained from birth certificates along with form SS-5 (Application for a Social Security Card). From 1935 to 1980, the SS-5 form only contained 3 race codes (Black, White, and other), but beginning in 1994, SS-5 forms used expanded race and ethnicity codes.

b The RTI imputation algorithm is used to improve the accuracy of the Enrollment Database race and ethnicity code by improving the sensitivity for Hispanic individuals and Asian/Pacific Islander individuals using language preference, beneficiary language, beneficiary race source code, state, and Hispanic and Asian/Pacific Islander surname lists developed by the US Census Bureau.

c Self-reported race and ethnicity information can be requested but is not standard in data sets provided to researchers.

d Medicaid race and ethnicity information is collected on a state-by-state basis and varies. Most states collect voluntary information on self-reported race and ethnicity, and states often address missing or incomplete data using language, geography, or matching with other state data or other administrative or claims data.

e States internally validate race and ethnicity, but these validations are not available in the literature to our knowledge.

Open Access. This is an open access article distributed under the terms of the CC-BY License.
or include a discussion of the implications of not having race and ethnicity variables. Neither do the articles always state whether methods to determine race and ethnicity have been validated or discuss how the quality of race and ethnicity variables may influence the study interpretation.

We have summarized the collection and derivation methods for race and ethnicity variables in large US national administrative claims data sets commonly used for research (Table). Race and ethnicity variables in some of these data sets are fully or partially derived from algorithms based on patient name and demographics. Medicare data have 2 race and ethnicity variables. One variable is based on Social Security data. The other variable was created to improve Social Security–based classification, particularly in Asian/Pacific Islander and Hispanic individuals. The other variable uses a validated algorithm, including language, state, and last name. This variable was created to improve Social Security–based classification, particularly in Asian/Pacific Islander individuals and Hispanic individuals. This approach results in disparate accuracy and reliability across racial and ethnic groups with kappa coefficients (κ, a measure of interrater reliability) ranging from 0.44 to 0.96 when compared with self-report (κ > 0.81, which indicates excellent agreement). ⁷

Among commercial administrative data sets, Optum Clinformatics Data Mart reports race and ethnicity using an unvalidated proprietary algorithm derived using member geographic location and name (email communication, Optum Life Science Department Data, August 13, 2021). TriNetX uses an unvalidated electronic health record–based approach that reports race and ethnicity information provided by participating health care organizations (email communication, US Healthcare Organization Partnerships, TriNetX, August 25, 2021). IBM MarketScan and Anthem Digital Data Sandbox report no race or ethnicity data (email communication, Anthem Inc, February 18, 2022). Overall, there are substantial potential issues with the use and interpretation of race and ethnicity variables in national administrative claims data sets used for research.

Race and ethnicity classifications do not have a well-defined biological correlation, and their terminology and classification can be problematic. ¹ The use and interpretation of race and ethnicity are further complicated in large administrative claims data sets by variables with imperfect and frequently unknown accuracy and reliability. Even when data on race and ethnicity are self-reported, the categories available to choose from in the self-report, the frequency that data are missing or incomplete, and the possibility that the likelihood of missing data is associated with racial or ethnic group identity could introduce meaningful biases.

For researchers reporting studies using race and ethnicity data in large administrative claims data sets, we propose the following principles: methods sections should clearly state how race and ethnicity variables were derived, including specific algorithm input parameters; methods sections should include a statement specifying whether the approach used to classify race and ethnicity has been validated and published, and if so, the accuracy and reliability of this method in the racial and ethnic groups examined; and authors should discuss the implications of the accuracy and reliability of the methods used to classify race and ethnicity on the interpretation of the reported results.

In contrast to government-led administrative claims data sets, more recently available commercial insurance data sets frequently do not report race and ethnicity data or use methods other than self-report without published validation studies. Administrative claims data sets, particularly commercial insurance data sets marketed for research, should present the following information: report race and ethnicity data if they currently do not; provide a clear and detailed explanation of the methods used to derive race and ethnicity variables; publish validation studies of their race and ethnicity variables if not based solely on self-report; and actively move to race and ethnicity variables obtained via self-report.

The utility of methods relying on a person's name, language, and geographic location will require ongoing assessment with increasing diversity in the US. Most counties have an increasing diversity index (likelihood that 2 people chosen at random will be from different racial and ethnic groups). ⁸ Increases in racial and ethnic diversity within geographic areas could present a challenge to algorithms using geography. Even self-report has limitations: the US Department of Health and
Human Services data collection standards categorize people of Middle Eastern descent as White, despite evidence that people of Middle Eastern descent may not be perceived or perceive themselves to be White.\(^9\)

Administrative claims data sets are a valuable research resource. Massive sample sizes and rapidly updated data sets can facilitate the timely investigation of key research questions, but there is a need to actively evaluate and enhance their quality for research. Improving race and ethnicity reporting practices using these data sets will increase awareness and transparency regarding the utility and limitations of these variables in research publications. The US Department of Health and Human Services provides standards for race and ethnicity questions that are required for reporting by federal agencies or the recipients of federal funds,\(^10\) which could be adopted by commercial data sets.

Ongoing evaluation of existing race and ethnicity data collection standards will improve the quality of data available, though it will also require resources to support the implementation of improved standards. Looking to the future, more uniform use and enforcement of existing standards for the definition, collection, derivation, validation, and use of race and ethnicity variables in administrative claims data sets are needed, particularly in commercial insurance data sets.

**ARTICLE INFORMATION**

Published: July 1, 2022. doi:10.1001/jamahealthforum.2022.1812

Open Access: This is an open access article distributed under the terms of the CC-BY License. © 2022 Nead KT et al. JAMA Health Forum.

**Corresponding Author:** Mackenzie R. Wehner, MD, MPhil, Department of Health Services Research, Department of Dermatology, University of Texas MD Anderson Cancer Center, 1400 Pressler St. Unit 1444, Houston, TX 77030 (mwehner@mdanderson.org).

**Author Affiliations:** Department of Epidemiology, The University of Texas MD Anderson Cancer Center, Houston (Nead); Department of Radiation Oncology, The University of Texas MD Anderson Cancer Center, Houston (Nead); Department of Health Services Research, The University of Texas MD Anderson Cancer Center, Houston (Hinkston, Wehner); Department of Dermatology, The University of Texas MD Anderson Cancer Center, Houston (Wehner).

**Conflict of Interest Disclosures:** None reported.

**REFERENCES**

