Database | Credentialed Access

C-REACT: Contextualized Race and Ethnicity Annotations for Clinical Text

Oliver Bear Don't Walk IV , Adrienne Pichon , Harry Reyes Nieva , Tony Sun , Jaan Lı , Joshua Winston Joseph , Sivan Kinberg , Lauren R Richter , Salvatore Crusco , Kyle Kulas , Shaan Ahmed , Daniel Snyder , Ashkon Rahbari , Benjamin Ranard , Pallavi Juneja , Dina Demner-Fushman , Noemie Elhadad

Published: Oct. 21, 2024. Version: 1.0.0

When using this resource, please cite: (show more options)
Bear Don't Walk IV, O., Pichon, A., Reyes Nieva, H., Sun, T., Lı, J., Joseph, J. W., Kinberg, S., Richter, L. R., Crusco, S., Kulas, K., Ahmed, S., Snyder, D., Rahbari, A., Ranard, B., Juneja, P., Demner-Fushman, D., & Elhadad, N. (2024). C-REACT: Contextualized Race and Ethnicity Annotations for Clinical Text (version 1.0.0). *PhysioNet*. https://doi.org/10.13026/z8tq-v658.

Please include the standard citation for PhysioNet: (show more options)
Goldberger, A., Amaral, L., Glass, L., Hausdorff, J., Ivanov, P. C., Mark, R., ... & Stanley, H. E. (2000). PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. Circulation [Online]. 101 (23), pp. e215–e220.

Abstract

The Contextualized Race and Ethnicity Annotations for Clinical Text (C-REACT) dataset is a large publicly available corpus of sentences from clinical notes manually annotated for information related to race and ethnicity (RE). The corpus presented here contains 17,281 sentences drawn from 12,000 patients and their clinical notes at the Beth Israel Deaconess Medical Center critical care units between 2001 and 2012. This corpus contains two sets of reference standard annotations for RE data. The first set contains granular RE-information such as patient country of origin and spoken language. The second set of annotations contains RE labels manually assigned by clinicians. This corpus is intended to improve understanding about granular information related to RE contained within the clinical note and how this information might be used to infer RE.

Background

Accurate and complete race and ethnicity (RE) data underpins many areas of clinical informatics, such as disease risk estimation [1–3], quality and performance metrics assessment [4], and health disparities identification [5–8]. Within the electronic health record (EHR), easily accessible RE information is stored in structured data sources. For many patients, however, structured RE data are often missing, inadequate, or inaccurate [9–16]. Additionally, common RE categories such as the Office of Management and Budget's (OMB) five race categories (American Indian or Alaska Native, Black or African American, Asian, Native Hawaiian or Other Pacific Islander, white) and two ethnic categories (Hispanic or Latino versus Not Hispanic or Latino) may lack the granularity needed to convey important subgroup differences [13, 17] and are thus inadequate. Patients with missing RE data are made less visible and cannot be fully included in and benefit from research. When identified in a careful and thoughtful manner, granular and accurate RE data can support health equity through diversifying representation in research.

Clinical notes often contain rich and nuanced information such as immigration status [18], country of origin [19], and spoken language [20] that clinical informaticians can potentially leverage to infer RE. While manually extracting RE data is costly and time-consuming, natural language processing (NLP) models can be trained to perform this task in an automated fashion using publicly available reference standard RE annotations. Ideally, these datasets provide rigorous annotation guidelines to facilitate transparency and offer sufficiently granular information related to RE to support a variety of use cases.

In this work, we present a detailed description of the RE annotation process and provide two sets of reference standard annotations on the same corpus of sentences. A total of 17,281 sentences for 12,000 patients were drawn from clinical notes in the MIMIC-III dataset [21], extracted from Beth Israel Deaconess Medical Center critical care units between 2001 and 2012. The first set contains granular RE-information such as patient country of origin and spoken language, annotated at the span level. The second set of annotations contains sentence-level RE labels manually assigned by clinicians. The purpose of this corpus is to foster utilization of granular information within the clinical note related to RE and improve understanding of how such information might be used to infer RE via NLP.

Methods

In this section, we describe our annotation and validation approach to independently annotate span level information related to RE (indicators) and then annotate sentences from clinical notes for RE assignments.

Data and Pre-Processing

We used NLTK [22] to extract sentences from discharge summaries (n=59,652) for patients (n=41,127) from version 1.4 of the MIMIC-III dataset [21] (version 1.4). Heuristics were developed to handle medical lists such as medication and condition lists. We identified 794,841 sentences that may contain discussions of a patient's RE by selecting for sentences with patient-related demographic keywords (e.g., "male", "female", "patient") and section headings (e.g., "Past Medical History", "PMH", "Social History", "SHX"). We then marked and prioritized sentences with RE keywords (e.g., "Black", "AA", "Native American", "Latino", "Spanish") for annotation. Of the 794,841 sentences, we sampled 17,281 sentences for two independent annotation processes (RE indicators and labels). These sentences were drawn from 13,507 notes. While we prioritized sentences with RE, we also sampled sentences without RE keywords. This corpus of 17,281 sentences was used for both annotation processes and is thus referred to as the central corpus.

Explicit Indicator Annotations

All 17,281 sentences from the central corpus were annotated for explicit indicators by non-clinicians (n=4). Four indicator categories were chosen to capture explicit spans of text that could potentially describe patient RE: 1) discussions of patient country/nation of origin or geographic ancestry (country); 2) discussions of primary, preferred, or spoken language (language); 3) direct discussions of race (race); and 4) direct discussions of ethnicity (ethnicity). Examples of the indicator categories used can be found in Table 1.

Table 1: Race and Ethnicity indicator descriptions and examples.

	Indicator Category	**Description**	**Examples**
Indirect	Country	Discussions of country/nation of origin or geographic ancestry.	"Pt is recently immigrated from **France**" "Pt is a 23 yo **Korean** female"
	Language	Discussions of primary, preferred, or spoken language.	"Pt required a **Spanish** interpreter" "Pt speaks fluent **Russian**"
Direct	Race	Direct mentions of racial categories from the U.S. census.	"Pt is an elderly **American Indian** man" "Pt is 42 yo **AA** female"
	Ethnicity	Direct mentions of ethnic categories from the U.S. census.	"Pt is 23 yo **Latino**" "Pt's mother is **Latina**"

Annotating was performed using the software Prodigy[23]. All sentences were double-coded until sufficient agreement was reached (macro F1 > 0.85) and annotation guidelines were iteratively developed to support reproducibility and transparency. Given that this task is at the span level and the number of negative cases is unknown, we measured inter-annotator agreement using macro F1 instead of Cohen's kappa[24, 25]. We aggregated F1-scores across indicator categories using macro F1 instead of micro F1 as we did not want the score dominated by the majority class. Annotation guidelines are available in Supplementary File 1.

Ten clinicians with a medical degree and at least one year of completed post-graduate residency training performed the RE category annotations. We presented each clinician with two subsets of sentences for independent annotation such that all sentences were annotated by at least one clinician. The first subset contained 5,834 sentences and was annotated by all clinicians (shared annotation subset). The second subset contained the remaining 11,447 sentences (single annotation subset). The single annotation subset was split evenly among all clinician annotators. The shared annotation subset was chosen such that all 4,834 sentences with at least one positive indicator annotation were used along with 1,000 sentences without a positive indicator (one sentence was later corrected as containing a positive indicator; the new C-REACT dataset reflects this change). The single annotation subset contained sentences that were not assigned any positive indicator annotations (five sentences were later corrected as containing positive indicators; the new C-REACT dataset reflects this change). Annotation guidelines for RE labeling can be found in Supplementary File 2.

Clinicians were asked to assign each sentence race and ethnicity categories based on the five race (Native American or Alaska Native, Black or African American (Black/AA), Asian, Native Hawaiian or Other Pacific Islander, and white) and two ethnicity (Hispanic/Latino/Latina and non-Hispanic/non-Latino/non-Latina) categories defined by federal standards [17]. Two additional categories were also used for race and ethnicity labeling. The category “Not Covered” indicates that a race or ethnicity label should be assigned to a sentence, but that the other categories provided were not sufficient. The category “No Information Indicated” is the only negative category and was used when a sentence did not contain sufficient information to assign a race or ethnicity label. It is important to note that these two categories were used independently for labeling both race and ethnicity.

We measured clinician agreement using Cohen’s Kappa on the shared annotation subset by comparing each clinician to the majority vote annotations of the remaining nine clinicians. We define the majority vote annotation by assigning an RE label to a sentence if a majority of clinicians (>=5) agreed on each label over all categories. Clinicians had moderate to almost perfect agreement for Black/AA, Asian, white, and Latino RE categories [26]. For all categories with 300 or more sentences assigned, half of the annotators had almost perfect (0.81-1.00) agreement and all annotators had substantial (0.61-0.80) agreement [26]. Generally, clinicians had lower agreement for the non-Latino and “No Information Indicated” ethnicity categories. Given the overall high agreement between clinicians, we resolved differences using the majority vote label between all clinicians.

Data Description

We provide slightly processed indicator and majority vote RE assignments and raw RE assignment data for all 17,281 sentences in the central corpus. All sentences can be linked together using a sentence identifier (sentence_id or ID in the raw RE assignment files). The remainder of this section describes the directory structure and files.

In the `collected_data` subdirectory, the `indicators_df.jsonl` (17,281 sentences) file contains output from the annotation software Prodigy (spans) as well as the associated sentences (text) and tokens (tokens). The file also contains information from the MIMIC NOTEEVENTS file such as identifiers for visits (visit_id), patients (patient_id), and notes (note_id).

In the `collected_data` subdirectory, the `all_re_assignments_df.jsonl` (17,281 sentences) file contains sentence (text) and one column for each race and ethnicity category used in this work. The prefix ‘RACE’ is used for racial categories while the prefix ‘ETH’ is used for ethnicity categories. Finally, sentence identifiers (sentence_id) and information on whether a sentence was part of the shared or single annotation subset (shared_subset, single_subset). Shared annotation subset sentences are assigned majority vote RE categories.

In the `raw_race_ethnicity_assignment_annotations` subdirectory, we also provide each annotator’s raw RE assignment files (all of which follow the same structure). Drawn from the shared subset, there are six xlsx files with the prefix ‘all_clinician_sentences_’ numbered 0-5. A subdirectory in this folder contains the sentences from the single annotation subset assigned to the annotator in a single xlsx file. All xlsx files follow the same format, and contain columns for all RE categories described previously, in addition to columns for the sentence identifier (ID) and sentence text (Sentence). Assigned columns are marked with any character, most often an ‘x’, and blank cells (including white space) do not contain an assignment.

Usage Notes

This data is meant as an extension of the RE data provided in MIMIC-III by providing sentence-level RE assignments as well as span-level RE indicators for 17,281 sentences. We support reuse and transparency by providing annotation guidelines and raw RE assignments. Future researchers are welcome to use these assignments to supplement missing RE data in MIMIC-III. We acknowledge that all researchers may not agree with the assumptions made during the assignment phase. As such, we provide span-level information on RE indicators. Given that sentences with positive RE labels were overwhelmingly likely to contain a positive indicator span, RE indicators can be used to identify sentences to assign different RE labels than those given by our annotators. Additionally, RE indicators can provide more granular information on patient race, ethnicity, country of origin, and spoken language.

Several limitations of this data exist. First, the RE categories and indicators used here align with RE constructs that are U.S.-centric in nature. While we attempt to address this by explicitly stating our assumptions through annotation guidelines and by providing RE indicators, other researchers might employ different definitions of race or ethnicity that are not fully encompassed by the indicators defined here. Secondly, the single annotation subset contains RE assignments for sentences without positive RE indicator spans. Some of these sentences contain discussions of food, occupation, immigration status, and wars/conflicts. We do not consider these discussions as indicative of race or ethnicity, but nonetheless we have chosen not to change the RE assignments in order to leave use of these assignments up to future researchers. Finally, the main limitation of C-REACT dataset is that all information recorded is from the subjective lens of the care provider who originally wrote the clinical notes and we cannot guarantee that the RE information recorded was patient reported.

The co-authors OBDW and NE are the main stewards of the C-REACT dataset and are responsible for its maintenance and update. For additional details and more information on the validation of the C-REACT dataset, please see our related publication in *Scientific Data* titled "Contextualized race and ethnicity annotations for clinical text from MIMIC-III" [27].

Release Notes

Version 1.0.0: First version of C-REACT.

Ethics

The C-REACT dataset was created with the intent to inform future research about granular information related to RE contained within the clinical note and how this information might be used to infer RE through NLP. While we believe that this information is much more impactful at the granular level, we understand that broader RE categories are important for many areas of informatics research. Whatever level of granularity is used, future researchers should be intentional and transparent about their definitions for race and ethnicity when using this dataset. Finally, while leveraging NLP trained on datasets like C-REACT to identify a patient’s RE information can be useful, this information should be used with the caveat that self-reported RE data is more faithful to a patient’s identity. Researchers should carefully consider when it is appropriate to use extracted over self-reported RE data.

Acknowledgements

We would like to acknowledge Adler Perotte and Karthik Natarajan for their guidance, feedback, and support during this project. We would also like to acknowledge Alistair Johnson and Travis Goodwin for their help with navigating the publication process. This work was supported by grants from the National Library of Medicine (OBDW, HRN, TS, AP: T15LM007079), Agency for Healthcare Research and Quality (BR: T32HS026121), Computational and Data Science Fellowship from the Association for Computing Machinery Special Interest Group in High Performance Computing (HRN), and the Intramural Research Program of the National Library of Medicine and National Institutes of Health (OBDW, DDF).

Conflicts of Interest

The author(s) have no conflicts of interest to declare.

References

1. Stevens LA, Coresh J, Greene T, Levey AS (2006) Assessing Kidney Function — Measured and Estimated Glomerular Filtration Rate. N Engl J Med 354:2473–2483
2. Levey AS, Stevens LA, Schmid CH, et al (2009) A New Equation to Estimate Glomerular Filtration Rate. Ann Intern Med 150:604–612
3. Gail MH, Brinton LA, Byar DP, Corle DK, Green SB, Schairer C, Mulvihill JJ (1989) Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. J Natl Cancer Inst 81:1879–1886
4. Blumenthal D, Tavenner M (2010) The “meaningful use” regulation for electronic health records. N Engl J Med 363:501–504
5. LaVeist TA, Gaskin D, Richard P (2011) Estimating the Economic Burden of Racial Health Inequalities in the United States. Int J Health Serv 41:231–238
6. Dorsey R, Graham G, Glied S, Meyers D, Clancy C, Koh H (2014) Implementing health reform: improved data collection and the monitoring of health disparities. Annu Rev Public Health 35:123–138
7. Douglas MD, Dawes DE, Holden KB, Mack D (2015) Missed Policy Opportunities to Advance Health Equity by Recording Demographic Data in Electronic Health Records. Am J Public Health 105:S380–S388
8. Kressin NR (2015) Race/Ethnicity Identification: Vital for Disparities Research, Quality Improvement, and Much More Than “Meets the Eye.” Medical Care 53:663–665
9. Polubriaginof FCG, Ryan P, Salmasian H, Shapiro AW, Perotte A, Safford MM, Hripcsak G, Smith S, Tatonetti NP, Vawdrey DK (2019) Challenges with quality of race and ethnicity data in observational databases. J Am Med Inform Assoc 26:730–736
10. Chakkalakal RJ, Green JC, Krumholz HM, Nallamothu BK (2015) Standardized data collection practices and the racial/ethnic distribution of hospitalized patients. Med Care 53:666–672
11. Kressin NR, Chang B-H, Hendricks A, Kazis LE (2003) Agreement between administrative data and patients’ self-reports of race/ethnicity. Am J Public Health 93:1734–1739
12. Institute of Medicine (US) Subcommittee on Standardized Collection of Race/Ethnicity Data for Healthcare Quality Improvement (2009) Race, Ethnicity, and Language Data: Standardization for Health Care Quality Improvement. National Academies Press (US), Washington (DC)
13. Nelson A (2002) Unequal treatment: confronting racial and ethnic disparities in health care. J Natl Med Assoc 94:666–668
14. Hasnain-Wynia R, Pierce D, Pittman MA (2004) Who, When, and How: The Current State of Race, Ethnicity, and Primary Language Data Collection in Hospitals. New York: Commonwealth Fund 42
15. Magaña López M, Bevans M, Wehrlen L, Yang L, Wallen GR (2016) Discrepancies in Race and Ethnicity Documentation: a Potential Barrier in Identifying Racial and Ethnic Disparities. J Racial Ethn Health Disparities. https://doi.org/10.1007/s40615-016-0283-3
16. Zingmond DS, Parikh P, Louie R, Lichtensztajn DY, Ponce N, Hasnain-Wynia R, Gomez SL (2015) Improving Hospital Reporting of Patient Race and Ethnicity—Approaches to Data Auditing. Health Serv Res 50:1372–1389

17. Revisions to the Standards for the Classification of Federal Data on Race and Ethnicity. In: The White House. https://obamawhitehouse.archives.gov/node/15626. Accessed 19 Dec 2021
18. Ross J, Hanna DB, Felsen UR, Cunningham CO, Patel VV (2017) Emerging from the database shadows: characterizing undocumented immigrants in a large cohort of HIV-infected persons. AIDS Care 29:1491–1498
19. Farber-Eger E, Goodloe R, Boston J, Bush WS, Crawford DC (2017) Extracting Country-of-Origin from Electronic Health Records for Gene- Environment Studies as Part of the Epidemiologic Architecture for Genes Linked to Environment (EAGLE) Study. AMIA Jt Summits Transl Sci Proc 2017:50–57
20. Hatef E, Rouhizadeh M, Tia I, Lasser E, Hill-Briggs F, Marsteller J, Kharrazi H (2019) Assessing the Availability of Data on Social and Behavioral Determinants in Structured and Unstructured Electronic Health Records: A Retrospective Analysis of a Multilevel Health Care System. JMIR Med Inform 7:e13802
21. Johnson AEW, Pollard TJ, Shen L, Lehman L-WH, Feng M, Ghassemi M, Moody B, Szolovits P, Celi LA, Mark RG (2016) MIMIC-III, a freely accessible critical care database. Sci Data 3:160035
22. Bird S, Klein E, Loper E (2009) Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit, 1st edition. O'Reilly Media, Beijing ; Cambridge Mass.
23. Montani I, Honnibal H (2018) Prodigy: A new annotation tool for radically efficient machine teaching. Artificial Intelligence to appear:
24. Hripcsak G, Rothschild AS (2005) Agreement, the F-Measure, and Reliability in Information Retrieval. J Am Med Inform Assoc 12:296–298
25. Deleger L, Li Q, Lingren T, Kaiser M, Molnar K, Stoutenborough L, Kouril M, Marsolo K, Solti I (2012) Building Gold Standard Corpora for Medical Natural Language Processing Tasks. AMIA Annu Symp Proc 2012:144–153
26. McHugh ML (2012) Interrater reliability: the kappa statistic. Biochem Med (Zagreb) 22:276–282
27. Bear Don't Walk OJ, Pichon A, Reyes Nieva H, et al (2024) Contextualized race and ethnicity annotations for clinical text from MIMIC-III. Sci Data.

Contents

Parent Projects

C-REACT: Contextualized Race and Ethnicity Annotations for Clinical Text was derived from:

- MIMIC-III Clinical Database v1.4

Please cite them when using this project.

Share

Access

Access Policy:
Only credentialed users who sign the DUA can access the files.

License (for files):
PhysioNet Credentialed Health Data License 1.5.0

Data Use Agreement:
PhysioNet Credentialed Health Data Use Agreement 1.5.0

Required training:
CITI Data or Specimens Only Research

Discovery

DOI (version 1.0.0):
https://doi.org/10.13026/z8tq-v658

DOI (latest version):
https://doi.org/10.13026/t9ka-6k29

Topics:
clinical notes | patient country information | race and ethnicity | patient language information

Corresponding Author

You must be logged in to view the contact information.

Files

This is a restricted-access resource. To access the files, you must fulfill all of the following requirements:

- be a credentialed user
- complete required training:
  - CITI Data or Specimens Only Research
    You may submit your training here.
- sign the data use agreement for the project

Supported by the National Institute of Biomedical Imaging and Bioengineering (NIBIB) under NIH grant number R01EB030362.

Back to top