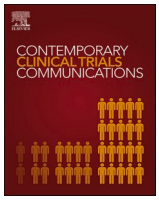
Contemporary Clinical Trials Communications 44 (2025) 101432

Contents lists available at ScienceDirect

Contemporary Clinical Trials Communications

journal homepage: www.elsevier.com/locate/conctc



Demographic clinical trial diversity assessment methods: Use of real-world data

Hua Chen [a], Nnadozie Emechebe [b], Sudeep Karve [b], Leon Raskin [b], Jailene Leal [b], Ning Cheng [b], Wendy Sebby [b], Kim Ribeiro [b], Samuel Crawford [b,*]

[a] Department of Pharmaceutical Health Outcomes and Policy, College of Pharmacy, University of Houston, USA
[b] AbbVie, Inc, North Chicago, IL, USA

ARTICLE INFO

Keywords:
Clinical trial diversity
Real-world data
RWD
Disease population
Health equity
Diversity plan
Benchmarking

ABSTRACT

Diversity in clinical trials is defined by the inclusion of clinical trial participants from various demographic groups that are representative of the broader population impacted by a disease state. Diversity in clinical trials is critical in identifying potential differences in safety and efficacy of treatments across races, ethnicities, ages, sexes, or other variables. In the United States, clinical trial diversity is often benchmarked against US Census data, which may limit the representativeness of patient demographics in clinical trials. Disease-specific, demographic estimates from real-world data (RWD) can facilitate benchmarking of clinical trials, support trial enrollment and the development of trial diversity plans. Notably, development and dissemination of these estimates from RWD can be challenging without a standardized process. To address this issue, we developed a new evaluation framework to assess patient demographics and characteristics within specific disease populations using RWD and disease population estimates.

Suitable databases were identified using predefined criteria such as accessibility to patient-level data, availability of all demographic variables of interest, sufficient sample size of the disease population, and availability of population weights to enhance generalizability. Concurrent data were gathered via targeted literature reviews for each disease condition. Together, this data was used to create disease-specific, demographic estimate profiles to inform diverse enrollment goals for prospective clinical trials. We present two examples of application of this framework to illustrate the results in the case of two disease states, rheumatoid arthritis and stroke.

1. Introduction

Diversity in clinical trials is defined as a sample of individuals that accurately reflect the demographic composition of those impacted by a disease. Clinical trial diversity is crucial to advancing health equity, as it provides the ability to ensure consistent treatment across a range of individuals, while potentially helping to identify differences in treatment response [1,2]. Clinical trial diversity can improve trust in reported trial outcomes, while broadening access to novel and potentially promising therapies [3–5].

Poor clinical trial representation can have significant implications for patient care, contributing to health disparities and unequal health outcomes, as well as limiting understanding of both risks and benefits for patients [3]. As the goal of a clinical trial is to determine safety and efficacy of treatments, it is critical to ensure that all groups benefit from advancements in medical research to the same extent, while identifying any differences that could potentially compromise either efficacy or safety. A lack of diversity in clinical trials that accurately reflects the population of a disease can result in uncertainties for underrepresented groups and could potentially lead to suboptimal care. Improving clinical trial representation to better reflect the real-world population of those dealing with a disease state can lead to more equitable access to effective treatment and help to develop a better understanding of how an intervention can work across diverse patient groups, informing better design of clinical trials [6,7].

Evaluation of diversity in clinical trials in recent studies have shown that trials are impacted by both underrepresentation and overrepresentation when comparisons are made to US Census data. Underrepresentation is often reported for racial or ethnic minority groups, as well as with age groups such as those 65 and older [8,9]. Conversely, overrepresentation can also lead to imbalances in trial diversity, which can happen when the biological sex of participants is skewed heavily in

* Corresponding author.
E-mail address: samuel.crawford@abbvie.com (S. Crawford).

https://doi.org/10.1016/j.conctc.2025.101432
Received 20 November 2024; Received in revised form 20 December 2024; Accepted 8 January 2025
Available online 10 January 2025

one direction or another, or when one racial or ethnic group has greater representation than what would align with a disease state [8,9]. Currently, there are limited resources available to provide an accurate reflection of diversity within a disease population. Use of standards such as the United States (US) national census to identify the diversity of the population provides a very broad, general picture without nuance to understand differences based on specific disease populations. Use of real-world data (RWD) and evidence gathered from existing studies to characterize disease-specific demographics provides greater understanding of who is impacted by a disease and can help to identify gaps in representativeness that may exist.

To address the need for disease-specific, demographic estimates to facilitate improved benchmarking of clinical trials, as well as support for trial enrollment and design of clinical trial diversity plans as outlined by the US Food and Drug Administration (FDA) [10,11], we sought to develop new ways of looking at diversity within specific disease populations using RWD and published literature. Here we report on a reproducible clinical trial diversity assessment framework and its utility to develop disease-specific, demographic estimates using multiple real-world datasets to inform diverse enrollment goals for prospective clinical trials. To illustrate the use of this framework, we identified two therapeutic areas, rheumatoid arthritis (RA) and stroke. These disease states were selected to provide a contrast in use of the framework for disease states with a narrow range of database and literature sources as is the case with RA, compared to stroke which provides a wider range of data sources to assess clinical trial diversity. These demographic estimates generated using this framework are more reflective of the diversity in specific disease populations compared to general census data and can inform the development of enrollment goals used for prospective clinical trials.

2. Methods

The trial diversity assessment framework was developed using large real-world population databases based in the US, along with published scientific literature, to assess disease-specific demographic characteristics.

2.1. Demographics

The key demographic variables of interest included age in years, sex (Female/Male), race (Asian, African American or Black, American Indian or Alaska native, Native Hawaiian or Other Pacific Islander, White), ethnicity (Hispanic/Non-Hispanic (NH)), race and ethnicity (NH Asian, NH Black, NH American Indian or Alaska native, Native Hawaiian or Other Pacific Islander, NH White, Hispanic), and US Census regions (Northeast, Midwest, West, South). Demographics of interest were selected based on those frequently collected in trials and highlighted in two FDA projects, Silver and Equity [12,13]. The FDA's Projects Equity and Silver are public health initiatives aimed at improving equitable access to medicines and increasing participation in clinical trials among historically underrepresented populations, including older adults. Initial analysis emphasized US demographic categories and estimates, with plans in future iterations to expand to other countries.

2.2. Database selection

For each disease of interest, several publicly and privately available databases were evaluated. Suitable databases were identified based on predefined criteria including measurement of the disease of interest, accessibility to patient-level data, availability of all demographic variables of interest, measurement of and sufficient sample size of the disease population in question, and availability of population weights to generate US representative estimates of civilian, non-institutionalized individuals (Table 1). Consequently, nationally representative survey datasets emerged as the preferred data source. When the disease of interest was unmeasured in these datasets, large nationwide disease registries, electronic health records (EHR), and administrative claims databases were preferred.

For each dataset under consideration, two reviewers independently evaluated suitability for a specific indication, assigning a score of 0–10, where 0 was defined as not suitable (ie, not representative and missing key demographic variables) and 10 was very suitable (ie, nationally representative with all key demographic variables) (Table S1). In the event of discordant scores, reviewers discussed evidence until a score was agreed upon. Reviewers then assigned the top two rated datasets as primary and secondary, allowing for two datasets selected per indication. Depending on rankings assigned to datasets, reviewers assigned analysis confidence grades as high, medium or low with a high confidence grade denoting better generalizability, better disease representation, and coverage of important sociodemographic characteristics.

2.3. Targeted literature reviews

Targeted literature reviews were conducted concurrently with the database evaluation by a separate team to determine published estimates on diseases of interest that could be used in lieu of databases with low confidence and to provide broader range of disease-specific population estimates (Fig. 1). Any published studies that utilized databases identified during database selection were excluded to avoid duplication of data.

Table 1
Hierarchy of real-world databases available at the time of analysis, including those used for evaluation in the current study.

Database Name	Data Type	Nationally Representative	Contains Key Variables	Ownership	Therapeutic Area	Selected for Current Study
National Health and Nutrition Examination Survey (NHANES)	National Survey	Yes	Yes	Public	Agnostic	No
National Health Interview Survey (NHIS)	National Survey	Yes	Yes	Public	Agnostic	Yes
Medical Expenditure Panel Survey (MEPS)	National Survey	Yes	Yes	Public	Agnostic	Yes
Behavioral Risk Factors Surveillance System (BRFSS)	National Survey	Yes	Yes	Public	Agnostic	Yes
National Survey of Children's Health (NSCH)	National Survey	Yes	Yes	Public	Agnostic	No
Surveillance, Epidemiology, End Results Program (SEER)	Disease Registries	No	Yes	Public	Oncology	No
ConcertAI	Electronic Health Records (EHR)	No	Yes	Private	Oncology	No
PurpleLab	Claims Database	No	Yes	Private	Agnostic	No
IQVIA	Claims Database	No	Yes	Private	Agnostic	No
Optum Market Clarity (MC)	Linked EHR and Claims	No	Yes	Private	Agnostic	No
Optum Clinformatics Data Mart (CDM)	Claims Database	No	Yes	Private	Agnostic	No

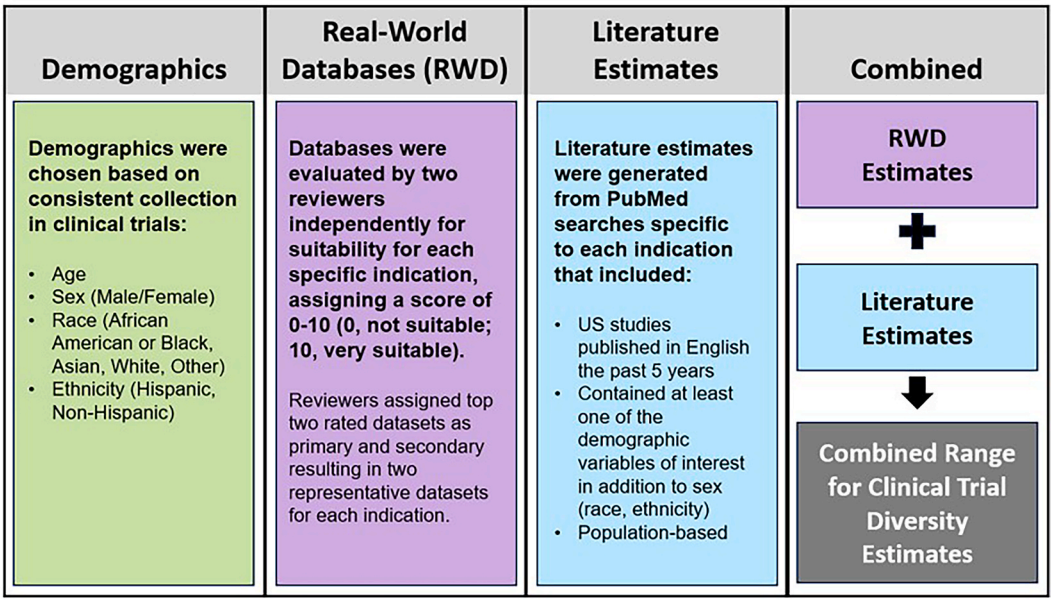


Fig. 1. Disease-specific profile generation method.

2.4. Statistical analysis

Following database selection, identification of disease cohorts was based on the source of the database. For national surveys, self-reported disease status of respondents was used, while ICD-10/ICD-0-3 coding systems were used for other real-world databases. Demographic characteristics were summarized using percentages for categorical variables and mean and standard deviation, and standard error for continuous variables. Standard error was computed for estimates generated using national survey data with a complex sampling design.

Only non-missing data for each variable were used in the analysis, thus, no imputation was done for missing data. Each dataset evaluated was examined over a three-year time horizon, such that the three most recent years of available data were evaluated.

Disease-specific profiles were created using demographic estimates generated from real-world databases and/or estimates from scientific literature. Once the final output was created, it was uploaded to the dashboard to create the disease-specific profiles. The dashboard was developed using the advanced analytics and data visualization software Spotfire (Tibco Software).

The trial diversity assessment framework can be applied to any disease of interest. To illustrate its utility, we provide examples for two disease states below, RA and stroke.

3. Results

3.1. Database selection

Four publicly available, national representative datasets were evaluated, namely: National Health and Nutrition Examination Survey (NHANES), Medical Expenditure Panel Survey (MEPS), National Health Interview Survey (NHIS), and Behavioral Risk Factors Surveillance System (BRFSS). Of these, three (stroke: NHIS and BRFSS; RA: MEPS) were utilized in generating the disease population estimates for both indications. Given the preference for these datasets, further evaluation of other real-world databases was not required. In some instances, NHANES was not chosen if other nationally representative data sources are available as race and ethnicity are reported jointly rather than separately.

MEPS was identified as a suitable dataset for RA because RA was measured in the database, population weights were available, all demographic characteristics were measured, had sufficient sample sizes (>50), and availability of patient-level data. Due to identification of a study from the literature review that utilized a large nationwide registry, one nationally representative dataset was deemed suitable for RA [14].

Database evaluation for stroke identified two publicly available databases, BRFSS and NHIS because both datasets met all the criteria. MEPS was not selected due to the heterogeneity in measurement of stroke where respondents were asked about a history of stroke or transient ischemic attack compared to BRFSS or NHIS where respondents were asked about a history of stroke, only. Furthermore, the condition code in MEPS included an ICD code for ischemic stroke, only (ICD10: I63).

Although both RA and stroke are measured in NHANES, the dataset was not chosen for due to lack of nationally representative data sources given that race and ethnicity are reported jointly rather than separately, as previously noted.

3.2. Targeted literature reviews

Targeted literature reviews were conducted for the two example disease states. For RA, the literature search initially yielded 8 records for full-text review, none of which met the criteria for inclusion outlined in Fig. 1; therefore, a subsequent search was extended by an additional 5 years. The 10-year search window yielded 98 records of which 97 were excluded due to duplication of data from registries, resulting in one paper that met inclusion criteria [14].

The targeted literature review for stroke resulted in a total of 73 records, of which 72 were excluded due to duplication of data from registries, resulting in 1 paper included in the literature estimates [15]. The search for stroke did not require an extension of the search years.

For both RA and stroke, data from RWD estimates were combined with the literature estimates to provide a final estimate of the disease population. Results of the implementation of the framework and

resulting estimates and comparison to US Census data can be found in Fig. 2.

4. Discussion

Through analysis of large, publicly and privately accessible datasets and targeted review of the literature, diversity in clinical trials can be examined at the disease level and can help to identify cases where groups may be underrepresented. Better representation has the potential to provide insight for diseases that impact some populations more than other groups, or groups where disease burden may have more severe effects. Here we provide methods to develop real-world disease estimates by combining RWD and literature estimates, providing a different perspective of the demographics impacted by the disease when compared to the US census. These differences are illustrated in Fig. 2, which provides a description of real-world estimates of RA and stroke compiled using the methods outlined here. These data reflect variation in disease estimates across indications, that often differ notably from demographic data obtained from the US census. For example, based on data gathered using the methods described here, the combined real-world estimate of individuals with RA that are Hispanic ranges from 5.3 to 15.1 %, but for stroke the estimate is 8.8–11.4 %, both of which differ from the US Census data reporting 19.1 % of the population is Hispanic. Estimates based on biological sex also differ when compared across indications, with estimates for women at 69.0–76.2 % for RA, which is higher than estimates for stroke (51.2–51.9 %) and higher than the US census data (50.4 %) (Fig. 2).

The current paper presents the results for two different disease states, RA which results in a wide range of estimates with lower precision and stroke which provides a narrow estimate range with higher precision, reflecting the generally low incidence of RA and higher incidence of stroke. Differences in the incidence of RA and stroke may explain the resulting lack of data for low-incidence indications like RA versus the vast amount of data for a higher-incidence indication like stroke. These examples also illustrate the results obtained using this method as well as the challenges that can arise with different disease states. For example, for RA, the targeted literature review did not return any results meeting the criteria over the last 5 years, requiring us to modify the search to extend to 10 years yielding one publication that met inclusion criteria. Though the literature review for stroke similarly found only one publication, it was published within the last 5 years.

While priority was given to data from national survey datasets, real-world databases like EHR and claims databases are valuable sources when generating demographic estimates for relatively rare diseases or diseases that require additional specificity for identification. Additionally, EHR was chosen over claims data due to availability of self-reported race vs derived race.

While the focus of this work was on US demographic estimates, future work will involve characterizing demographic estimates in populations outside the US.

We observed some data limitations in the development of this framework. Reporting of race can be inconsistent, with some real-world sources reporting race alone and others reporting both race and ethnicity together or as a non-informative “Other” posing challenges to harmonization of data. Similar challenges were observed with age, where medians or percentage of subjects 65 or older (population of interest in Project Silver) were challenging to implement across indications as some analytical platforms or published studies do not

	Rheumatoid Arthritis		Stroke		
Demographics	**RWD Estimates[a]**	**Literature Estimates[b]**	**RWD Estimates[c]**	**Literature Estimates[d]**	**Census[e,f]**
Sex, %					
Female	69.0	76.2	51.2-51.7	51.9	50.4
Male	31.0	23.2	48.3-48.8	49.1	49.6
Race, %					
White	76.3	90.5	71.5-76.5	75.5	75.5
Black	15.7	7.6	17.2-17.6	17.7	13.6
Asian	1.7	1.8	2.5-3.1	-	6.3
American Indian/ Alaskan Native	1.9	-	2.5-3.1	-	1.3
Native Hawaiian/Other Pacific Islander	1.6	-	0.4	-	0.3
Other	2.8	-	1.5-5.3	6.8	3.0
Ethnicity, %					
Hispanic	15.1	5.3	10.2-11.4	8.8	19.1
Non-Hispanic	84.9	94.7	88.6-89.8	91.2	80.9

RWD, Real-world database
[a]Based on disease specific population estimate assessed using the Medical Expenditure Panel Survey (MEPS), 2018 – 2020
[b]From Greenberg et al., 2013 (12). Due to a lack of literature meeting the search criteria in the initial 5-year search window, the literature search for rheumatoid arthritis was extended by an additional 5 years.
[c]Based on disease specific population estimates assessed using the National Health Interview Survey (NHIS) and the Behavioral Risk Factors Surveillance System (BRFSS) national surveys, 2018-2020.
[d]From Khan et al., 2022 (13)
[e]National data from United States Census as of April 1, 2020 (https://www.census.gov/quickfacts/fact/table/US/PST045222)
[f]Census data is provided for comparison only and is not a component of the disease-specific profile.

Fig. 2. Comparison of clinical estimates examples for rheumatoid arthritis and stroke alongside US census data.

consistently report these statistics, or report age in wide-ranging categories. Missing data on race and ethnicity were relatively higher in real-world databases like EHR compared to national survey databases, underscoring the importance of concerted efforts towards increased capture of race and ethnicity in these data sources. Last, disease populations of interest can differ in granularity such as specific biomarker groups or lines of therapy and the availability of databases with this level of granularity is limited.

5. Conclusions

Use of disease-specific population estimates generated from combining real-world sources and literature estimates may provide greater insight into the diversity across indications, informing development of clinical trials to better reflect specific disease populations and supporting the development of enrollment goals.

CRediT authorship contribution statement

Hua Chen: Writing – review & editing, Writing – original draft, Conceptualization. **Nnadozie Emechebe:** Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Formal analysis, Conceptualization. **Sudeep Karve:** Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Formal analysis, Conceptualization. **Leon Raskin:** Writing – review & editing, Writing – original draft, Validation, Methodology, Conceptualization. **Jailene Leal:** Writing – review & editing, Writing – original draft, Validation, Methodology, Conceptualization. **Ning Cheng:** Writing – review & editing, Writing – original draft, Validation, Methodology, Conceptualization. **Wendy Sebby:** Writing – review & editing, Writing – original draft, Validation, Methodology, Conceptualization. **Kim Ribeiro:** Writing – review & editing, Writing – original draft, Methodology. **Samuel Crawford:** Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Formal analysis, Conceptualization.

Funding

AbbVie funded this study and participated in the study design, research, analysis, data collection, interpretation of data, reviewing, and approval of the publication.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: **Chen** Member of the AbbVie Academic Steering Committee. **N Emechebe, S Karve, L Raskin, J Leal, N Cheng, W Sebby, K Ribeiro**, and **S Crawford** are full-time employees of AbbVie Inc. and may hold AbbVie stock and/or stock options.

Acknowledgements

The authors would like to thank Sean Sullivan, Jalpa Doshi, and Dima Qato for their valuable contributions to the development of this short communication. Medical writing support was provided by Elin M. Grissom, PhD of AbbVie.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.conctc.2025.101432.

References

[1] L.T. Clark, L. Watkins, I.L. Pina, M. Elmer, O. Akinboboye, M. Gorham, et al., Increasing diversity in clinical trials: overcoming critical barriers, Curr. Probl. Cardiol. 44 (5) (2019) 148–172.
[2] B. Johnson-Williams, D. Jean, Q. Liu, A. Ramamoorthy, The importance of diversity in clinical trials, Clin. Pharmacol. Ther. 113 (3) (2023) 486–488.
[3] A.A. Schwartz, M, A.A. Morris, S.D. Halpern, Why diverse clinical trial participation matters, N. Engl. J. Med. 388 (2023) 1252–1254.
[4] U. Peters, B. Turner, D. Alvarez, M. Murray, A. Sharma, S. Mohan, et al., Considerations for embedding inclusive research principles in the design and execution of clinical trials, Ther. Innov. Regul. Sci. 57 (2) (2023) 186–195.
[5] A. Corneli, E. Hanlen-Rosado, K. McKenna, R. Araojo, D. Corbett, K. Vasisht, et al., Enhancing diversity and inclusion in clinical trials, Clin. Pharmacol. Ther. 113 (3) (2023) 489–499.
[6] S. Dagenais, L. Russo, A. Madsen, J. Webster, L. Becnel, Use of real-world evidence to drive Drug development strategy and inform clinical trial design, Clin. Pharmacol. Ther. 111 (1) (2022) 77–89.
[7] Goldman DG, E.A.; del Rio, C. Lack of Diversity in Clinical Trials Costs Billions of Dollars. Incentives Can Spur Innovation2022 11/28/2023. Available from: https://healthpolicy.usc.edu/article/lack-of-diversity-in-clinical-trials-costs-billions-of-dollars-incentives-can-spur-innovation/.
[8] L.E. Flores, W.R. Frontera, M.P. Andrasik, C. Del Rio, A. Mondriguez-Gonzalez, S. A. Price, et al., Assessment of the inclusion of racial/ethnic minority, female, and older individuals in vaccine clinical trials, JAMA Netw. Open 4 (2) (2021) e2037640.
[9] A.H. Kaakour, H.U. Hua, A. Rachitskaya, Representation of race and ethnicity in randomized clinical trials of diabetic macular edema and retinal vein occlusion compared to 2010 US census data, JAMA Ophthalmol. 140 (11) (2022) 1096–1102.
[10] US Department of Health and Human Services UF, CDER, CBER, Enhancing the Diversity of Clinical Trial Populations — Eligibility Criteria, Enrollment Practices, and Trial Designs Guidance for Industry, fda.org, 2020 [Available from: https://www.fda.gov/media/127712/download.
[11] A. Ramamoorthy, R. Araojo, K.P. Vasisht, M. Fienkeng, D.J. Green, R. Madabushi, Promoting clinical trial diversity: a highlight of select us FDA initiatives, Clin. Pharmacol. Ther. 113 (3) (2023) 528–535.
[12] US Department of Health and Human Services UF, CDER, CBER, Project silver, improving the evidence base for the treatment of older adults with cancer: fda.org [Available from: https://www.fda.gov/about-fda/oncology-center-excellence/project-silver, 2021.
[13] US Department of Health and Human Services UF, CDER, CBER, Project equity, generating evidence for diverse populations in oncology: fda.org [Available from: https://www.fda.gov/about-fda/oncology-center-excellence/project-equity, 2023.
[14] J.D. Greenberg, T.M. Spruill, Y. Shan, G. Reed, J.M. Kremer, J. Potter, et al., Racial and ethnic disparities in disease activity in patients with rheumatoid arthritis, Am. J. Med. 126 (12) (2013) 1089–1098.
[15] M.Z. Khan, S. Zahid, A. Kichloo, S. Jamal, A.M.K. Minhas, M.U. Khan, et al., Gender, racial, ethnic, and socioeconomic disparities in palliative care encounters in ischemic stroke admissions, Cardiovasc. Revascularization Med. 35 (2022) 147–154.