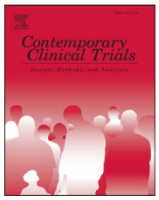
Contemporary Clinical Trials 121 (2022) 106890

Contents lists available at ScienceDirect

Contemporary Clinical Trials

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



Short Communication

Defining demographic cohorts in clinical trial populations using large electronic health records databases

Stephen J. Peroutka

PPD, a part of Thermo Fisher Scientific, 929 North Front Street, Wilmington, NC 28401-3331, United States of America

ARTICLE INFO

Keywords:
FDA
Demographics
Trials
Model
Gender
Race

ABSTRACT

The Food and Drug Administration (FDA) has stressed the need to ensure that clinical trial study populations accurately reflect the patients likely to use the product, if approved. However, the FDA has not provided specific guidance on how clinically relevant demographic characteristics might be defined. Therefore, the present study was designed to develop a framework that could be used to rapidly identify population demographics for any medical condition. Then, these real-world data were used as the basis to calculate acceptable demographic parameters (with 95% confidence intervals) for clinical trial populations. Data on Alzheimer's Disease were used as an example of the proposed approach.

1. Introduction

Recent Food and Drug Administration (FDA) Guidance has stressed the importance of clinical trial study populations that accurately reflect the patients likely to take the drug in the United States (US), if approved. Moreover, FDA now recommends that sponsors should include a plan for inclusion of clinically relevant populations during the development process. The plan should be submitted, and discussed with the FDA, prior to the onset of their pivotal trials [1].

However, the FDA did not provide specific guidance on how clinically relevant trial population characteristics might be defined and noted that limited data might be available for certain medical conditions. Ideally, demographic data from all patients diagnosed with a given disease in the US could be used. Unfortunately, these data rarely exist for most clinical disorders. Alternatively, data from epidemiological studies could be used to model trial populations, but such data are rare or non-existent for most disorders. US Census demographic data could be used, but there are many clinical disorders that vary in prevalence between different demographic groups.

Over the past few years, electronic health record databases have been created from large subsets of the population. Demographic data, based on International Classification of Diseases (ICD) codes for medical conditions, can now be assessed rapidly on millions of US healthcare consumers. Moreover, these real-world data provide an opportunity to define the demographic characteristics of patients who are likely to use new treatments, as per FDA recommendations.

Therefore, the present study was designed to assess the demographic characteristics of a large subpopulation of individuals, diagnosed with Alzheimer's Disease (AD), in the US population. The demographic data were then used to propose statistical parameters around each selected demographic variable. The objective was to develop a data driven approach that allows for the calculation of recommended cohort sample sizes, for any demographic variable. This approach may provide an acceptable framework to ensure the inclusion of clinically relevant subpopulations in clinical trials.

2. Methods

The TriNetX Analytics Network [2] database contains electronic health records from >130 million individuals from >70 Health Care Organizations across the US. The payers include commercial, Medicaid, Medicare and VA providers. TriNetX, LLC is compliant with the Health Insurance Portability and Accountability Act (HIPPA), the US federal law which protects the privacy and security of health care data, and any additional data privacy regulations applicable to the contributing HCO. TriNetX is certified to the ISO 27001:2013 standard and maintains an Information Security Management System (ISMS) to ensure the protection of the health care data it has access to and to meet the requirements of the HIPAA Security Rule. Any data displayed on the TriNetX Platform in aggregate form, or any patient level data provided in a data set generated by the TriNetX Platform, only contains de-identified data as per the de-identification standard defined in Section §164.514(a) of the

Abbreviations: FDA, Food and Drug Administration.

E-mail address: Stephen.Peroutka@ppd.com.

https://doi.org/10.1016/j.cct.2022.106890
Received 18 July 2022; Received in revised form 18 August 2022; Accepted 18 August 2022
Available online 24 August 2022

HIPAA Privacy Rule. The process by which the data are de-identified is attested to through a formal determination by a qualified expert as defined in Section §164.514(b)(1) of the HIPAA Privacy Rule.

Because this study used only de-identified patient records and did not involve the collection, use, or transmittal of individually identifiable data, this study was exempted from Institutional Review Board approval.

A database query was generated for individuals who had received an ICD-10 G.30 diagnosis of AD and had obtained healthcare within the past 5 years in the US (as of March 30, 2022). Demographic data were obtained and then used to define the characteristics of this "representative" sample of the AD population. Specifically, the gender and race demographic parameters in the TriNetX dataset were used to define the "expected" parameters for hypothetical sample sizes. Binomial distribution was used to calculate the 95% Confidence Interval (CI) for each demographic cohort.

3. Results

3.1. TriNetX AD population

As of March 30, 2022, there were 266,640 individuals in the TriNetX database who had an ICD-10 G.30 diagnosis of AD within the previous 5 years and had received healthcare services. The demographic characteristics of this US only cohort included average age (82 ± 8 years), gender (63% female, 37% male), race (White = 72%, Black or African American = 11%, Asian = 2%, Other or unknown = 15%) and ethnicity data (Not Hispanic or Latino = 73%, Unknown = 22%, Hispanic or Latino = 5%) (Table 1).

3.2. Gender cohort models

For each hypothetical sample size, the TriNetX data were used to estimate the "expected" proportion of female and male cohorts (Table 2). For example,an AD trial consisting of 100 subjects could be considered "representative" of the >266,000 patients in the TriNetX database if it included 54 to 72 females (i.e., values that would fall within the 95% CI). As a result, enrollment of 28 to 46 males would be needed to also be within the 95% CI (Table 2A).

As sample size increases (Table 2B and C), the "representative" range of acceptable cohort numbers decreases in terms of allowable percentages. For example, in a 1000 CE subject trial, gender cohorts that are representative of the TriNetX database population would include 600 to 660 females and balanced by 340 to 400 males in order to be within the 95% CI. At least 60% of subjects should be female and 34% male to be considered "representative" of the AD gender distribution in the TriNetX database, based on the 95% CI (Table 2C).

Table 1
Demographic characteristics of ICD-10 G.30 Alzheimer's disease patients in the TriNetX database (as of March 30, 2022).

Number of patients identified	266,640
Age (mean ± Standard Deviation)	82 ± 8
Gender (%)	
Female	63
Male	37
Race (%)	
White	72
Black or African American	11
Asian	2
Other, Mixed or Unknown	15
Ethnicity (%)	
Not Hispanic or Latino	73
Unknown	22
Hispanic or Latino	5

Table 2
Sample size and gender cohort models.

Demographic cohorts	Expected # of cohort subjects	95% Confidence interval
(A) Sample Size = 100		
Female	63	54–72
Male	37	28–46
(B) Sample Size = 300		
Female	189	173–205
Male	111	95–127
(C) Sample Size = 1000		
Female	630	600–660
Male	370	340–400

3.3. Race cohort models

A similar approach was used to model representative trial cohorts based on race. For example, White subjects comprised 72% of the TriNetX AD dataset (Table 3A) while all other races comprised 28% of the dataset. Binomial proportion analysis showed that an AD trial population of 100 subjects that included 63 to 81 White subjects would be within the 95% CI for the sample size. Inclusion of 28 to 46 non-White subjects would also be within the 95% CI (Table 3A) for this sample size. In the same 100 subject trial, enrollment of 6–17 Black or African American subjects would be within the 95% CI (Table 3A). Table 3 also provides recommended cohort sizes for 300 and 1000 subject AD trials, based on the 95% CI.

4. Conclusions

The major finding of the present study is that data from a large electronic medical record databases can be used to rapidly identify demographic parameters that can guide the enrollment of clinical trial populations that are representative of patients likely to use new products. Although only gender and race data were analyzed in this report, this method can be applied to any binomial demographic parameter. These "expected" data can then be used to guide the appropriate selection of demographic cohorts for clinical trial populations, as recommended by the FDA.

Traditionally, epidemiological studies, disease specific databases and patient registries have been used to describe disease demographics. For example, the Einstein Aging Study [3] used systematic recruiting methods to enroll 1,944 individuals from Bronx County, NY, between 1993 and 2004. Detailed cognitive assessments were made at baseline and at 12-month intervals thereafter. The study identified 102

Table 3
Sample size and race cohort models.

Demographic cohorts	Expected # of cohort subjects	95% Confidence interval
(A) Sample Size = 100		
White	72	63–81
Black or African American	11	5–17
Asian	2	0–5
Other/Mixed/Unknown	15	8–22
(B) Sample Size = 300		
White	216	201–231
Black or African American	33	22–44
Asian	6	1–11
Other/Mixed/Unknown	45	33–57
(C) Sample Size = 1000		
White	720	692–748
Black or African American	110	91–129
Asian	20	11–29
Other/Mixed/Unknown	150	128–172

individuals with AD. Although this study represents an outstanding example of epidemiological research, it should be noted that after more than a decade of work, a relatively small number of AD patients were identified in a small geographic area. Similarly, the Framingham Study enrolled 5,205 individuals, 60 years of age or older, over a 40-year period [4]. This study identified 264 cases of AD. Thus, epidemiological studies in AD are relatively limited in size and geographic reach, despite their excellent long-term longitudinal data. In addition, epidemiological studies may become outdated due to the evolution of diagnostic criteria and standard of care over time.

In comparison to epidemiological research, electronic healthcare databases allow for real-world evidence that can be generated within minutes. For example, the OneFlorida Data Trust contains information on 17.2 M individuals, obtained from 10 healthcare systems within the state [5]. The database contained information on 100,033 individuals with AD, as of 2022 [6]. However, this excellent data repository is geographically limited to ~5% of individuals the US.

By contrast, the TriNetX Analytics Network contains health records on >130 M individuals from >70 healthcare organizations across the US [2]. The data in this, and other health record databases, do not contain the level of detail nor the diagnostic rigor of traditional epidemiological studies, yet they do provide real world evidence that may be more indicative of patient populations that are likely to use new therapies.

Although this study has focused on gender and race, there are many other demographic characteristics that could be evaluated using the same modeling method. For example, age, ethnicity, educational levels, biomarkers, comorbities and concomitant medication use are additional demographic characteristics that can be used to determine the similarities between trial populations and the general population. In the future, it is likely that the FDA will provide additional guidance on how to select the key demographic variables for given medical conditions. In addition, guidance will be needed as to the level of similarity that would be considered acceptable (e.g., values that fall within the 80%, 95% or 99% CIs?).

In conclusion, optimal clinical drug development requires that potential therapeutics should be studied in trial populations that resemble the real-world populations who are likely to use new products. The FDA has provided strong recent guidance to highlight this important scientific tenet. Given the number and variety of clinical indications, it is unlikely that traditional epidemiological studies could address the need to model trial populations in all areas of medical research. By contrast, large electronic health record databases can provide for a rapid method to describe and define demographic parameters that can then be used to enroll subject populations that approximate real-world clinical populations.

Declaration of Competing Interest

The author declares that he has no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

I thank Drs. Alberto Lledo, Rose Blackbourne, Rodrigo Garcia, Jonca Bull and Carol Olson for their thoughtful comments on the manuscript.

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

[1] Food and Drug Administration, Enhancing the Diversity of Clinical Trial Populations—Eligibility Criteria, Enrollment Practices, and Trial Designs Guidance for Industry, Available at, https://www.fda.gov/media/127712/download, 2022 (Accessed 2022-03-30).

[2] TriNetX Website, Available at, https://trinetx.com/life-sciences/, 2022 (Accessed 2022-03-30).

[3] M.J. Katz, R.B. Lipton, C.B. Hall, M.E. Zimmerman, A.E. Sanders, J. Verghese, D. W. Dickson, C.A. Derby, Age-specific and sex-specific prevalence and incidence of mild cognitive impairment, dementia, and Alzheimer dementia in blacks and whites: a report from the Einstein Aging Study, Alzheimer Dis. Assoc. Disord. 26 (4) (2012 Oct-Dec) 335–343, https://doi.org/10.1097/WAD.0b013e31823dbcfc. PMID: 22156756; PMCID: PMC3334445.

[4] C.L. Satizabal, A.S. Beiser, V. Chouraki, G. Chêne, C. Dufouil, S. Seshadri, Incidence of dementia over three decades in the Framingham heart study, N. Engl. J. Med. 374 (6) (2016 Feb 11) 523–532, https://doi.org/10.1056/NEJMoa1504327. PMID: 26863354; PMCID: PMC4943081.

[5] W.R. Hogan, E.A. Shenkman, T. Robinson, O. Carasquillo, P.S. Robinson, R. Z. Essner, J. Bian, G. Lipori, C. Harle, T. Magoc, L. Manini, T. Mendoza, S. White, A. Loiacono, J. Hall, D. Nelson, The OneFlorida data trust: a centralized, translational research data infrastructure of statewide scope, J. Am. Med. Inform. Assoc. 29 (4) (2022 Mar 15) 686–693, https://doi.org/10.1093/jamia/ocab221. PMID: 34664656; PMCID: PMC8922180.

[6] A.H. Miller, D.E. Marra, Y. Wu, J. Bian, E.A. Shenkman, D.M. Maraganore, G. E. Smith, Characterizing dementia prevalence in the State of Florida: An electronic health record study, Suppl. Public Health 17 (S10) (31 December 2021), https://doi.org/10.1002/alz.052364.