Exploring the Feasibility of Using Real-World Data from a Large Clinical Data Research Network to Simulate Clinical Trials of Alzheimer's Disease

Zhaoyi Chen, PhD[1*], Hansi Zhang, MS[1*], Yi Guo, PhD[1], Thomas J George Jr, MD, FCAP[1], Mattia Prosperi, PhD[1], William R. Hogan, MD, MS[1], Zhe He, PhD[2], Elizabeth A Shenkman, PhD[1], Jiang Bian, PhD[1†]

[1]University of Florida, Gainesville, Florida; [2]Florida State University, Tallahassee, Florida

Abstract

Clinical trials are essential but often have high financial costs and long execution time. Trial simulation using real world data (RWD) could potentially provide insights on a treatment's efficacy and safety before running a large-scale trial. In this work, we explored the feasibility of using RWD from a large clinical data research network to simulate a randomized controlled trial of Alzheimer's disease considering two different scenarios: an one-arm simulation of the standard-of-care control arm; and a two-arm simulation comparing treatment safety between the intervention and control arms with proper patient matching algorithms. We followed original trial's design and addressed some key questions, including how to translate trial criteria to database queries and establish measures of safety (i.e., serious adverse events) from RWD. Our simulation generated results comparable to the original trial, but also exposed gaps in both trial simulation methodology and the generalizability issue of clinical trials.

Introduction

Clinical trials, especially randomized controlled trials (RCTs), are critical in the drug discovery and development process to assess how the treatment being developed will interact with the human body.[1] While the controlled conditions of clinical trials can reduce bias and ensure the internal validity of the study results, they also come with the drawbacks of high financial costs and long execution time.[2] For example, the total cost of developing an Alzheimer's disease (AD) drug were estimated at $5.6 billion with a timeline of 13 years from preclinical studies to FDA approval,[3] even though no effective drugs have been developed for either AD treatment or prevention thus far. In general, the median cost of pivotal trials (i.e., Phase III trials) for new therapeutic agents approved by the U.S. Food and Drug Administration (FDA) is estimated at $19.0 million;[4] while Phase III AD trials typically cost $1.79 billion and take on average more than 4 years to complete.[3] Strategies that can accelerate the drug development process and reduce costs will not only be of interest to pharmaceutical companies but also ultimately benefit the patients.

In the last decade, there has been an increased uptake of electronic health record (EHR) systems in the United States (US). These technological advances and policy changes in the US have created a fertile ground with increasing opportunity to use EHR data to improve current methods of clinical evidence generation. The FDA coined the term real-world data (RWD) refereeing to data collected from sources outside of conventional research settings, including EHRs, administrative claims, and billing data among others.[5,6] RWD have emerged as an important data source reflecting real-world clinical environment of where the treatments are actually used, including patient demographics, comorbidities, adherence, disease status, treatment outcomes, and concurrent treatments that are tracked in detail and longitudinally. The opportunity to design studies simulating clinical trials using RWD could (1) provide insights on a treatment's efficacy and safety before running a large-scale RCT and help to decide whether a trial would be beneficial and yield a high return on investment, and (2) replace standard-of-care control arms to reduce trial costs.

The design of pivotal Phase III trials typically includes intervention arms using the new chemical entities and control arms with the current standard of care for the disease. The control arms are often repeated across different trials, have rigid eligibility criteria, and often incur high costs to enroll sufficient samples. Simulating contemporaneous external control arms using RWD collected from routine care could potentially reduce these costs. In addition, clinical trials are often conducted under rigorously controlled conditions to assure their internal validity. In clinical trials, the target population (TP) is the patients to whom the trial results are intended to be applied. The study population (SP, also called trial-eligible population) is the set of patients defined by the trial's eligibility criteria. To ensure patient safety and demonstrate efficacy, a trial's criteria are often restrictive, leading to an SP that is a constrained subset of the TP. Low trial generalizability has been widely documented across different clinical areas[7] including AD and dementia[8,9] when applying the findings from trials to the broad TP. Therefore, real-world evidence (RWE) is needed to reflect

[*] Zhaoyi Chen, PhD and Hansi Zhang, MS contributed equally, co-first authors
[†] Corresponding: Jiang Bian, PhD; bianjiang@ufl.edu

the population who would most likely use the treatment. A simulation-based study that simulates the logic flow of a clinical trial using RWD can be used to estimate the safety and efficacy of the treatment in a broad, diverse, and more general patient population, thus, generating RWE. In addition, by simulating a trial, the on treatment-outcome effects observed from observational RWD could potentially illuminate causal relationships.

The concept of trial simulation has been explored before.[10–13] For example, Danaei *et al.* conducted a comparative effectiveness research (CER) study using EHR data from United Kingdom (UK) by emulating a hypothetical RCT to estimate the effect of statins for primary prevention of coronary heart disease.[12] Like many other emulation studies,[14–16] this is essentially a retrospective cohort study, where the authors followed a clinical trial design to identify unbiased initiation of exposures and eventually to reach an unbiased estimation of the casual relationship. On the other hand, studies simulating external control arms aim to identify eligible individuals who are on standard-of-care treatments for the specific disease from patient databases, as comparators to treatment arms in RCTs. For example, Carrigan *et al.* used Flatiron, a US-based oncology EHR database, to assess how closely results from advanced non-small cell lung cancer (aNSCLC) RCTs could be approximated, by substituting EHR-based external control arms as the comparator.[13] Their findings showed promising results as the hazard ratio estimates of overall survival aligned closely with those from the corresponding RCT (i.e., a Pearson correlation coefficient of 0.86). These prior studies showed great potential for trial simulations and asking causal questions using observational RWD data.[17]

In this study, we aimed to explore the feasibility of using RWD from the OneFlorida Clinical Research Consortium—one of the large clinical data research networks funded by the Patient-Centered Outcomes Research Institute (PCORI) contributing to the national Patient-Centered Clinical Research Network (PCORnet)—to simulate a real-world AD RCT as a use case. We attempted to address a number of key barriers that have not been well-explored in previous studies working with RWD, for example, the lack of discussions on how to translate eligibility criteria to database queries and the difficulties in establishing the outcome measures (e.g., serious adverse events [SAEs]). We will consider two main scenarios: (1) a one-arm simulation: simulating a standard-of-care arm that can serve as an external control arm; and (2) a two-arm simulation: simulating both intervention and control arms with proper patient matching algorithms for comparative effectiveness analysis.

Methods

Identification of Alzheimer's disease (AD) clinical trials for simulation

To identify AD clinical trials for simulation, we searched all clinical trials on ClinicalTrials.gov using a group of AD-related keywords (e.g., "*Alzheimer's disease*", "*Dementia of AD type*"). ClinicalTrials.gov[18] is a registry of clinical research studies maintained by the U.S. National Library of Medicine. It enables researchers to find clinical trials by disease category with different searching filters (e.g., study phase, recruitment status, etc.) to refine the search results. The trial summaries are semi-structured in ClinicalTrials.gov: study descriptors (e.g., study phase, intervention type, and locations) and study results (e.g., baseline characteristics of participants, serious adverse events [SAEs]) are stored in structured fields, whereas eligibility criteria are largely free text. In this analysis, we focused on Phase 3 or 4 drug development RCTs on AD. To compare the study results with our simulated trials regarding drug safety (e.g., SAEs) and participants' characteristics (e.g., age, gender, and race), we only included studies that have already been completed and the study results were already published on ClincalTrials.gov. 44 trials were identified. Next, we attempted to search for studies that have study protocols available on ClinicalTrials.gov as the trial protocol documents the objectives, design, methodology, statistical considerations, and other aspects related to the organization of clinical trials.[19] Out of the initial 44 trials, only 6 published their study protocols. Nevertheless, they were all excluded for various reasons (e.g., primarily focused on other diseases, small sample sizes in the target patient database—OneFlorida—as the drug being developed is still in early stage and not widely used in real-world settings). Thus, we opted to looking for the most frequently used AD drugs and their corresponding RCTs that have generated publications with sufficient details on the study design including the eligibility criteria used to select subjects, treatment protocols of the subjects, and assessments of treatment efficacy and safety (i.e., definitions of adverse events [AEs]).

Target trial characteristics

We identified donepezil as the most widely tested AD drug and selected trial NCT00478205.[20] This study is a Phase III double-blind, double-dummy, parallel-group comparison of 23 mg donepezil sustained release (SR) with the 10 mg donepezil immediate release (IR) formulation (currently marketed standard-of-care) in patients with moderate to severe Alzheimer's disease. Patients who have been taking 10 mg IR (or a bioequivalent generic) for at least 3 months prior to screening were recruited. The study consisted of 24 weeks of daily administration of study medication, with clinic visits at screening, baseline, 3 weeks (safety only), 6 weeks, 12 weeks, 18 weeks, and 24 weeks or early termination. Patients received either 10 mg donepezil IR in combination with the placebo corresponding to 23 mg

donepezil SR, or 23 mg donepezil SR in combination with the placebo corresponding to 10 mg donepezil IR. A total of 400 patients and 800 patients were needed for the 10mg arm and 23mg arm, respectively, with a total of 471 and 963 patients enrolled eventually, respectively. The study was conducted at approximately 200 global sites (Asia, Oceania, Europe, India, Israel, North America, South Africa, and South America). In our simulation, we followed the detailed study procedures outlined in their published article[21] to formulate our simulation protocol, including the treatment regimen, population eligibility, and follow-up assessments for SAEs. **Table 1** describes how the original trial design was followed in our simulation.

Table 1. Overall study design of the simulated trial in comparison with the original trial.

Component	Target trial (NCT00478205)	Simulated trial
Aim	Assess the safety and effectiveness of 23mg compared to 10 mg	Assess whether the simulated trial can generate similar results to the "real" trial
Eligibility	36 eligibility criteria	25 are computable or partially computable
Treatment strategies	Randomized allocation of 23mg :10 mg ratio is 2:1	Scenarios 1 and 2 are two-arm simulations. Propensity score matching was performed on baseline covariates: sex, race, age, and Charlson comorbidity index (CCI). Scenario 3 is a one-arm simulation of the 10 mg control arm only. The same sample size as calculated in the original trial, random sampling and proportional sampling were used.
Sampling strategies	N/A	Bootstrap with replacement was repeated 1,000 times to randomly generate the sample population, and mean value and 95% confidence interval were reported.
Follow-up	The outcomes were measured from the first dose to 24 weeks after the first dose.	
Outcome	SAE and cognition function measures	SAE
Statistical analysis	Compare the average number of SAEs per patient, and the SAE rates (i.e., how many patients have SAE).	
SAE: Serious Adverse Events		

Real-world patient data (RWD) from the OneFlorida network

Our OneFlorida data contain robust longitudinal and linked patient-level RWD of ~15 million (>50%) Floridians, including data from Medicaid & Medicare claims, cancer registry, vital statistics, and EHRs from its clinical partners. As one of the PCORI-funded clinical research networks in the national PCORnet, OneFlorida includes 12 healthcare organizations that provide care through 4,100 physicians, 914 clinical practices, and 22 hospitals, covering all 67 Florida counties. The OneFlorida data is a HIPAA limited data set (i.e., dates are not shifted; and 9 digit zip codes of patients' residences are available) that contains detailed patient characteristics and clinical variables, including demographics, encounters, diagnoses, procedures, vitals, medications, and labs.[22] **Table 2** shows the demographics of AD patients in OneFlorida. Note that even though clinical notes are potentially available through OneFlorida, we focused on the structured data immediately available to us formatted according to the PCORnet common data model (PCORnet CDM) version 5.1.[23]

Table 2. Alzheimer's disease (AD) patients in OneFlorida; N = 101,904 (100%), Q3 2019

Sex	
Male	31,680 (31.1%)
Female	69,032 (67.7%)
Unknown	1,192 (1.2%)
Race/Ethnicity	
NHW	50,067 (49.1%)
NHB	12,451 (12.2%)
Hispanics	25,237 (25.7%)
Other	1,314 (1.3%)
Unknown	11,835 (11.8%)
Age	
< 65	6,422 (6.3%)
65 – 74	15,184 (14.9%)
75 – 84	36,407 (35.7%)
≥ 85	43,890 (43.1%)
Unknown	1 (0.0%)
NHW = Non-Hispanic White; **NHB** = Non-Hispanic Black;	

Cohort identification: target population, study population, and trial not eligible population

We identified three populations: the target population (TP), the study population (SP), and the trial not eligible population (NEP) for the selected trial following the process shown in **Figure 1**. Starting with the overall OneFlorida AD population, we defined the target population as patients who (1) had the disease of interest (i.e., AD), and (2) had used the study drug (i.e., donepezil) for a specific time period according to the study protocol. The criteria to define the TP were extracted from both the study description on ClinicalTrials.gov and study-related publications. We then identified the study population (i.e., patients who met both the TP criteria and the trial eligibility criteria) and

trial not eligible population (i.e., patients who meet the TP criteria but do not meet the trial eligibility criteria) by applying clinical trial eligibility criteria to the TP. To do so, for the selected trial, we analyzed its eligibility criteria and determined the computability of each criterion. A criterion is computable when its required data elements are available and clearly defined in the target patient database (i.e., the OneFlorida data in our study). Then, we manually translated computable criteria into database queries against the OneFlorida database. Note that not all eligibility criteria are queryable against the patient database as the needed data elements may not exist in the target database (e.g., "*A cranial image is required, with no evidence of focal brain disease that would account for dementia.*" and "*Written informed consent*."). We assumed that all patients met the non-computable criteria, which is a limitation of our study.

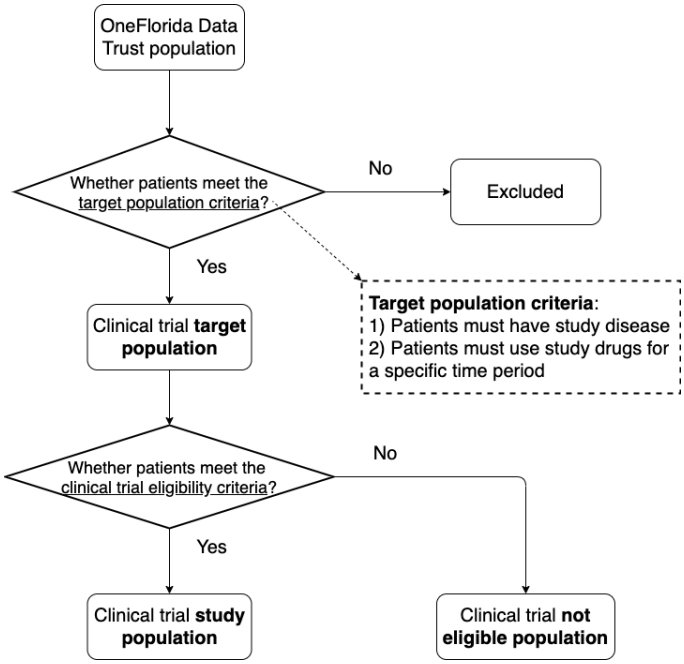

Figure 1. The cohort identification process for the target, study, and trial not eligible populations.

Definition and identification of serious adverse events (SAE) from EHRs

The drug efficacy of AD is often measured based on neuropsychological tests (e.g., Mini-Mental State Examination) that may only exist in unstructured EHR data (i.e., clinical notes). The target trial used Severe Impairment Battery (SIB) and the Clinician's Interview-Based Impression of Change Plus Caregiver Input scale (CIBIC+; global function rating) to assess the efficacy of donepezil. Because these data are not in the OneFlorida structured data, we instead focused our simulation on drug safety in terms of the occurrences of SAEs. To define an SAE, we first followed the FDA[24] definition of SAEs and the Common Terminology Criteria for Adverse Events (CTCAE) version 5 and then consulted with clinical experts. The final guideline for identifying SAEs for the target trial NCT00478205 is as below:

- ***Step 1***. For each reported AE in the trial result on ClinicalTrails.gov, map the AE term to CTCAE .
- ***Step 2***. For those mapped terms, use the CTCAE manual to identify the severity grading scale.
- ***Step 3***. To identify SAE, if an AE happened during the SAE selection window (i.e., 24 weeks after study injection and within 30 days after study end), and is graded with grade 4 (life-threatening or hospitalization) and 5 (death), we consider the AE as an SAE. If death happened with 30 days after an AE, we consider the AE as an SAE. Note that if the AE condition is a chronic disease and happened before the study, we did not consider it as an SAE.

Simulation protocol

Table 1 shows our design of the simulated trial corresponding to the original target trial. Based on the calculation from the original trial[21], a sample size of 400 and 800 were needed for the 10mg and 23mg arms, respectively. We did not find a sufficient number of patients who took 23mg donepezil in our OneFlorida data. Thus, we decided to explore two different scenarios for the two-arm simulation, and one scenario for the one-arm simulation. In scenarios 1 and 2, we aimed to simulate both 10mg and 23 mg arms with different ratios for the number of subjects between the

two arms. We used propensity score matching on baseline covariates: sex, race, age, and Charlson Comorbidity Index (CCI) to simulate the randomization of the original trial. In scenario 1, we set the 23mg to 10mg number of subjects ratio as 1:1, while in scenario 2, the ratio was set to 1:3. Because of the limited number of individuals who took the 23mg form, we can only increase the number of subjects in the 10mg arm. In scenario 3, we only simulated the control arm using standard therapy (i.e., the 10 mg arm of the original trial), where we have a sufficiently large sample size from the OneFlorida data. Thus, we designed our sampling strategy based on the sample size calculated in the original trial (N=400). For all scenarios, bootstrap sampling with replacement was repeated for 1,000 times to generate the sample populations, and the mean value and 95% confidence interval of each bootstrap sample were used to generate the overall estimates. We focused on comparing the average number of SAE per patient, and the overall SAE rates (how many patients have SAE) in our simulation.

Results

Computability of eligibility criteria in the original trial (i.e., NCT00478205)

In total, there are 36 eligibility criteria in trial NCT00478205 based on ClinicalTrials.gov, where 17 are inclusion criteria and 19 are exclusion criteria. However, not all eligibility criteria are computable against our OneFlorida patient database. The reasons are summarized in **Table 3.**

Table 3. Reasons for eligibility criteria that are either partially computable or not computable for NCT00478205.

Computability	Reasons	Examples
Not computable (N = 11)	Data elements needed for the criterion are not present in the OneFlorida data. (Inclusion： N = 3; Exclusion: N = 3)	e.g., "*A cranial image is required, with no evidence of focal brain disease that would account for dementia."*
	Patients need caregiver support. (Inclusion： N = 1; Exclusion: N = 1)	*e.g., "The patient must have a relative/caregiver who supervises the regular taking of the drug at the correct dose and is alert for possible side effects, unless the patient's legal guardian takes on this task."*
	The criterion asked for subjective information from patients. (Inclusion： N = 0; Exclusion: N = 1)	*e.g., "Patients who are unwilling or unable to fulfill the requirements of the study."*
	Data elements have granularity issues. (Inclusion： N = 0; Exclusion: N = 1)	*e.g., "Known hypersensitivity to acetylcholinesterase inhibitors or memantine."*
	Requires information about another clinical trial. (Inclusion： N = 0; Exclusion: N = 1)	*e.g., "Involvement in any other investigational drug clinical trial during the preceding 3 months, or likely involvement in any other such trial during the course of this study."*
Partially computable (N = 7)	Data elements of interest are not clearly defined. (Inclusion： N = 4; Exclusion: N = 1)	*e.g., "The patient must meet certain psychometric test criteria related to the degree of impairment of cognitive functioning."*
	Requires physicians' subjective judgment. (Inclusion： N = 1; Exclusion: N = 1)	*e.g., "Clinical laboratory values must be within normal limits or, if abnormal, must be judged not clinically significant by the investigator."*

Characteristics of the target, study, and trial not eligible populations from OneFlorida

Overall, a total of 90 and 2048 TP patients were identified in OneFlorida for the 23 mg arm and 10 mg arm, respectively. Among them, 38 and 782 met the eligibility criteria of the original target RCT for the two arms, respectively. ***Table 4*** shows the demographic characteristics and SAE statistics of the original trial population as well as the target population (TP), study population (SP), and trial not eligible population (NEP) from OneFlorida.

Table 4. Population characteristics and SAE statistics of the target trial vs. TP, SP, and NEP from OneFlorida.

	23mg Arm				10mg Arm			
	Original Trial[a]	**Overall TP[b]**	**Overall SP[c]**	**Overall NEP[d]**	**Original Trial**	**Overall TP[*]**	**Overall SP[#]**	**Overall NEP[&]**
# of Subject	963	90	38	52	471	2,048	782	1,266

Age Mean (SD)	73.9 (8.53)	74.3 (9.01)	73.3 (9.01)	81.6 (12.37)	73.8 (8.56)	73.4 (11.0)	74.2 (9.67)	77.1 (11.8)
Gender								
Male	356 (37.0%)	24 (26.7%)	10 (26.3%)	14 (26.9%)	177 (37.6%)	727 (35.5%)	234 (29.9%)	493 (38.9%)
Female	607 (63.0%)	66 (73.3%)	28 (73.7%)	38 (73.1%)	294 (62.4%)	1321 (64.5%)	548 (70.1%)	773 (61.1%)
Race[e]								
White	708 (73.5%)	63 (70.0%)	25 (65.8%)	38 (73.1%)	346 (73.5%)	829 (40.5%)	280 (35.8%)	549 (43.4%)
Asian/Pacific	161 (16.7%)	0 (0%)	0 (0%)	0 (0%)	87 (18.5%)	22 (1.1%)	11 (1.4%)	11 (0.9%)
Hispanic	67 (7.0%)	15 (16.7%)	4 (10.5%)	11 (21.2%)	26 (5.5%)	440 (21.5%)	192 (24.6%)	248 (19.6%)
Black	22 (2.3%)	11 (12.2%)	4 (10.5%)	7 (13.5%)	9 (1.9%)	380 (18.6%)	126 (16.1%)	254 (20.1%)
Other	5 (0.5%)	1 (1.1%)	5 (13.2%)	7 (13.5%)	3 (0.6%)	377 (18.4%)	173 (22.1%)	204 (16.1%)
CCI[f]	N/A	1.54	1.32	1.53	N/A	2.36	1.64	2.36
Mean SAE[g]	0.15	1.89	0.92	2.60	0.14	1.68	0.64	2.59
# of patients with ⩾ 1 SAE	45 (9.6%)	20 (22.2%)	4 (10.5%)	16 (30.7%)	80 (8.3%)	573 (28.0%)	121 (15.5%)	452 (35.7%)

[a]Reported in the original trial on ClinicalTrials.gov
[b]TP: Target population—patients who (1) had the disease of interest (i.e., AD), and (2) had used the study drug (i.e. donepezil) for a specific time period according to the study protocol.
[c]SP: Study population—patients in the TP who met the computable eligibility criteria of the original trial.
[d]NEP: Trial not eligible population—patients in the TP who did NOT meet the eligibility criteria of the original trial.
[e]The original trial reported Hispanic as a race, thus, we followed the same convention to make sure the results are comparable even though race and ethnicity are two different fields in OneFlorida.
[f]Charlson Comorbidity Index.
[g]Mean SAE: average number of SAEs per patient.

For demographic characteristics, relative to the target RCT population, we observed a large difference in race in our OneFlorida population (all p-values of race group comparison were smaller than 0.05). OneFlorida had more Hispanics (10.5% - 24.6% vs. 5.5% - 7%) and Blacks (10.5% - 20.1% vs. 1.9% - 2.3%), but less Whites (35.8% - 73.1% vs. 73.5% - 73.5%) or Asian/Pacific islanders (0% - 1.4% vs. 16.7% - 18.5%). The age distributions were similar across all populations. For clinical variables, we calculated the Charlson Comorbidity Index (CCI) of the various populations from OneFlorida. Smaller CCIs were observed in the SP compared with the TP for both arms ($p<0.05$), and a smaller CCI was observed in the 23mg arm compared with the 10mg arm ($p<0.05$). Our primary outcomes of interest in this analysis were SAEs. Thus, we calculated the mean SAE (i.e., the average number of SAEs per patient) and the number of patients who had more than 1 SAE during the study period. For both 23mg and 10mg arms, the mean SAE and the number of patients with SAEs were the largest in the TP, followed by the SP, and then the original trial. Consistent with the original trial, populations derived from the OneFlorida data in the 23mg arm have higher numbers of mean SAE and more patients with SAE compared with the 10mg arm.

Two-arm trial simulation results

Because of the limited number of eligible patients who took 23mg donepezil in OneFlorida data, we were unable to sample the 23mg arm with the original trial's sample size. Thus, we used all 38 OneFlorida subjects for the 23mg arm simulation and did not run the bootstrap sampling for this arm. But the matched 10mg arm was very stable in the matching variables (age, gender, race, and CCI) across the 1000 bootstrap samples. **Table 5** shows our two-arm simulation results, where we show the average and 95% confidence interval (CI) of all variables for the 10mg simulation arms. In both scenarios 1 and 2, the mean SAE and SAE rates were higher in the 23mg arm than in the 10mg arm, which is consistent with the original trial. However, the variance for both SAE outcomes for the 10mg arm are higher in scenario 1 than in scenario 2. This is understandable as the sample size for 10mg arm in scenario 2 is much bigger (i.e., in scenario 1, we set the 23mg to 10mg number of subjects ratio as 1:1, while in scenario 2, the

ratio was set to 1:3). Further, in scenario 2, the 10mg arm does have a higher SAE rate comparing to the original trial and scenario 1. Because of the sample size difference, estimates from scenario 2 should be more reliable.

Table 5. Results for the two-arm simulation for both scenarios 1 and 2.

	Target RCT		Scenario 1		Scenario 2	
	23 mg	**10mg**	**23mg**	**10mg**	**23mg**	**10mg**
# of Subjects	963	471	38	38	38	114
Age (mean)	73.9	73.8	73.3	72.7±0.1	73.3	71.8±0.1
Gender						
Male	37.0%	37.6%	26.3%	31.6%±0.1 %	26.3%	34.2%±0.1 %
Female	63.0%	62.4%	73.7%	68.4% ±0.1%	73.7%	65.8% ±0.1%
Race						
White	73.5%	73.5%	65.8%	63.2% ±0.1%	65.8%	64.9% ±0.1%
Asian/Pacific	16.7%	18.5%	0.0%	0.0%	0.0%	0.0%
Hispanic	7.0%	5.5%	10.5%	7.9%±0.1%	10.5%	9.6%±0.1%
Black	2.3%	1.9%	10.5%	7.9%±0.1%	10.5%	6.1%±0.1%
Other	0.5%	0.6%	13.2%	9.6%±0.1%	13.2%	8.0%±0.1%
Charlson Comorbidity Index (mean)	N/A	N/A	1.32	0.97±0.1	1.32	1.11±0.1
Mean SAE	0.15	0.14	0.92	0.118 ±0.121	0.92	0.547 ±0.016
SAE Rate[a]	9.6%	8.3%	10.5%	5.3% ±3.4%	10.5%	**15.8% ±0.3%**

[a]Percentage of patients with ⩾ 1 SAE.

External standard-of-care control arm (i.e., one-arm) simulation results

Our third scenario was to simulate an external control arm of the original trial (i.e., the 10mg stand-of-care arm). **Table 6** displays the results for this scenario. Two different sampling approaches were used: (1) random sampling, and (2) proportional sampling controlling for race distribution. When using the random sampling approach, compared with the control arm in the original trial, higher mean SAE and SAE rates were observed, in addition to discrepancies in demographic variables. When using proportional sampling, the results were closer and more consistent with the original trial. Notably, the SAE rates in the simulated control were similar to the SAE rates from the original control (8.9% vs. 8.3%), and a z-score test for population proportion had a p-value of 0.75, suggesting there were no significant differences between the two SAE rates.

Table 6. One-arm simulation results for the external control arm (i.e., 10mg arm).

	Original Control	Simulated Control	
		Random Sampling	**Proportional Sampling**
# of Subjects	471	400	400
Age	73.8	74.1±0.1	73.8±0.1
Gender			
Male	37.6%	29.9%±0.1%	32.7%±0.1%
Female	62.4%	70.1%±0.1%	67.3%±0.1%
Race			
White	73.5%	35.8%±0.1%	73.5%
Asian/Pacific	18.5%	1.4%±0.1%	18.5%
Hispanic	5.5%	24.5%±0.1%	5.5%
Black	1.9%	16.1%±0.1%	1.9%
Other	0.6%	22.1%±0.1	0.6%
Mean SAE	0.14	0.643±0.005	0.448±0.007
SAE Rates	8.3%	15.5%±0.1%	8.9%±0.1%

Discussions and conclusion

In this work, we simulated an AD RCT utilizing the rich RWD from OneFlorida—a large clinical data research network, considering three different simulation scenarios. In the two scenarios of two-arm simulation, we showed that randomization in the original trial might be achieved by controlling for baseline characteristics using propensity score matching. However, the outcomes measured in the simulated trial were different from the original trial for various reasons (e.g., sample size issue, conducted in research settings vs. in real-world clinical settings). In the one-arm simulation scenario, we attempted to simulate an external control arm for the original trial. We demonstrated that we could achieve a similar estimate of SAE rates as the original trial when proportional sampling was used to control for race distribution, and the statistics of the simulated control arm were very stable across all bootstrap simulation runs. Overall, our findings suggested that our trial simulation framework has some potential to mimic an original clinical trial; and the outcome estimates (i.e., SAEs) are stable and reliable, especially when simulating the standard-of-care control arm. However, there are still gaps, especially data gaps, that led to the differences between our simulation results and the original trial results when considering the two-arm simulation scenarios. Future studies are warranted to identify strategies to fill these gaps.

While simulating the original AD trial followed the study protocol in **Table 1**, we found it is difficult to replicate every single eligibility criterion in the original trial since not all of them are computable. Out of the original 36 eligibility criteria, only 25 of them were computable or partially computable. Since these criteria were used to weed out patients who are unlikely to complete the protocol (e.g., due to safety concerns), ignoring some of the criteria (not computable eligibility criteria) could potentially explain some of the increases either in the mean SAE or SAE rates. For example, we were unable to query for OneFlorida patients who met the inclusion criterion, "*Clinical laboratory values must be within normal limits or, if abnormal, must be judged not clinically significant by the investigator*", because the criterion was vague and did not define what abnormal clinical laboratory values are. Also, we found some of the SAEs (e.g., abnormal behavior, presyncope) reported in the trial's results cannot be mapped to any AE terms in CTCAE, and the definitions of AEs in the original trial were unavailable, which increased the difficulty of accurately accounting for all SAEs. Further, even though trials' SAEs reported in ClinicalTrials.gov largely follow the Medical Dictionary for Regulatory Activities Terminology (MedDRA), not all reported SAEs were correctly defined in the trial results. For example, we found "*Back pain*" and "*Fall*" were defined as SAEs in the original AD trial we modeled. However, in CTCAE, there is no corresponding category 4 or 5 definition for them. More effort is needed to consistently model SAEs reported in clinical trials.

Our findings are consistent with previous literature on clinical trial generalizability.[25–28] More SAEs were observed in real-world settings. In our data, the overall number of patients who had SAEs and the average number of SAEs per patient were the highest in the target population (i.e., patients who took the medication for the target disease), which is the population who actually used the medication in real-world situation. Compared with reports from the original trial, the mean SAE and SAE rates were also higher in the study population—the population who used the medication and also met the original trial's eligibility criteria. Some of the differences may be due to the incomputable eligibility criteria and SAE types that we did not account for, but it is also possible that the original trial samples did not adequately reflect the TP and thus there might be treatment effect heterogeneity across patient subgroups, not captured by the original trial. In the two-arm simulations, large variances were observed, especially when the matched sample size was small. This may also indicate the heterogeneous treatment effects of donepezil when applied to different patient subgroups in real-world settings.

Our study demonstrated the feasibility of trial simulation using RWD, especially when simulating external standard-of-care control arms. Our one-arm simulation provided stable and robust estimates and sufficient sample sizes to compare with the original trial's control arm. The SAE rates observed in the simulated control arm with proportional sampling were very close to what was reported in the original trial. The mean SAE per patient, however, was larger in the simulated control arms, which suggested that, in a real-world setting, the patients who experienced SAEs tend to have more occurrences of SAEs. On the other hand, the two-arm simulation, although it provided insights, was not entirely successful. Although the randomization process was effectively simulated by using propensity score matching, the outcome measures were very different from the original trial. The reasons for the differences could be multi-fold (e.g., research setting vs. real-world clinical setting, difference in sample size, overly restrictive eligibility criteria that limits the generalizability of the original trial), but cannot be explored due to limited data reported by the original trial (i.e., no patient-level data is available).

There are some other limitations in this study. First, we only looked at one original trial for one medication (i.e. donepezil). Simulations on different drugs and diseases may have different results. Future studies could build on the

findings from our work and apply the same methodology to simulating other clinical trials. Second, the population who took the 23mg form in our data is very small (even though the overall OneFlorida population is large with more than 15 million patients), where we only identified 38 patients who took the 23mg donepezil and met the eligibility criteria of the original trial. The 23mg donepezil form was approved by the FDA in 2012, so it is still a relatively new drug on the market, and following its approval, the clinical utility of the 23mg form was called into question because of its limited efficacy and higher rates of adverse events.[29] The current practice of using the donepezil 23mg form is reserved for AD patients who have been on stable donepezil 10 mg form for at least 3–6 months with no significant improvement,[31,32] which limited its use in real-world clinical practice. Third, when comparing the study population of the simulated control arm to the study population in the original trial, although it is possible to match the distribution of demographic variables, we were unable to match the baseline health status (e.g., CCI) between our simulation and the original trial, because the health status of the patients enrolled in the original trial was not reported. The potential difference in baseline health status in the study populations may explain some of the differences in the number of SAEs detected. Strategies that can help make the simulated control arm more comparable to the original trials are needed. For example, backward estimation of CCI in the trial population based on the exclusion criteria and average CCI in the general population may provide insights on the baseline health status of the trial population. Of course, it would be beneficial if future clinical trials do actually report participants' overall baseline health status (e.g., CCI). Finally, because of data limitations, we were not able to assess the efficacy of AD treatment. Nevertheless, neuropsychological tests (e.g., Mini-Mental State Examination and Severe Impairment Battery) may exist in clinical narratives. Future studies that explore the use of advanced natural language processing (NLP) methods to extract useful variables from clinical notes will be important. Further, variables extracted from clinical notes with NLP could also be used to render some of the incomputable eligibility criteria computable. Thus, systematic efforts are necessary to explore the benefits of NLP.

Acknowledgment

This work was supported in part by NIH grants R21AG061431 and UL1TR001427. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

References

1. Aronson JK. What is a clinical trial? [Internet]. John Wiley & Sons, Ltd; 2004. Available from: http://doi.wiley.com/10.1111/j.1365-2125.2004.02184.x
2. Sertkaya A, Wong HH, Jessup A, Beleche T. Key cost drivers of pharmaceutical clinical trials in the United States. Clin Trials. 2016;
3. Scott TJ, O'Connor AC, Link AN, Beaulieu TJ. Economic analysis of opportunities to accelerate Alzheimer's disease research and development. Ann N Y Acad Sci. 2014 Apr;1313:17–34. PMCID: PMC4285871
4. Moore TJ, Zhang H, Anderson G, Alexander GC. Estimated Costs of Pivotal Trials for Novel Therapeutic Agents Approved by the US Food and Drug Administration, 2015-2016. JAMA Intern Med. 2018;
5. Sherman RE, Anderson SA, Dal Pan GJ, Gray GW, Gross T, Hunter NL, LaVange L, Marinac-Dabic D, Marks PW, Robb MA, Shuren J, Temple R, Woodcock J, Yue LQ, Califf RM. Real-world evidence - What is it and what can it tell us? N Engl J Med. Massachussetts Medical Society; 2016 Dec;375(23):2293–2297. PMID: 27959688
6. Real-World Evidence | FDA [Internet]. [cited 2020 Feb 20]. Available from: https://www.fda.gov/science-research/science-and-research-special-topics/real-world-evidence
7. Rothwell PM. External validity of randomised controlled trials: "to whom do the results of this trial apply?" Lancet Lond Engl. 2005 Jan 1;365(9453):82–93. PMID: 15639683
8. Banzi R, Camaioni P, Tettamanti M, Bertele' V, Lucca U. Older patients are still under-represented in clinical trials of Alzheimer's disease. Alzheimers Res Ther. 2016 12;8:32. PMCID: PMC4982205
9. Leinonen A, Koponen M, Hartikainen S. Systematic Review: Representativeness of Participants in RCTs of Acetylcholinesterase Inhibitors. PloS One. 2015;10(5):e0124500. PMCID: PMC4416896
10. Zhang Y, Young JG, Thamer M, Hernan MA. Comparing the Effectiveness of Dynamic Treatment Strategies Using Electronic Health Records: An Application of the Parametric g-Formula to Anemia Management Strategies. Health Serv Res. United States; 2018 Jun;53(3):1900–1918. PMID: 28560811
11. Dickerman BA, García-Albéniz X, Logan RW, Denaxas S, Hernán MA. Avoidable flaws in observational analyses: an application to statins and cancer. Nat Med. 2019;
12. Danaei G, Rodríguez LAG, Cantero OF, Logan R, Hernán MA. Observational data for comparative effectiveness research: An emulation of randomised trials of statins and primary prevention of coronary heart disease. Stat Methods Med Res. 2013;

13. Carrigan G, Whipple S, Capra WB, Taylor MD, Brown JS, Lu M, Arnieri B, Copping R, Rothman KJ. Using Electronic Health Records to Derive Control Arms for Early Phase Single-Arm Lung Cancer Trials: Proof-of-Concept in Randomized Controlled Trials. Clin Pharmacol Ther. 2020;
14. Garcia-Albeniz X, Hsu J, Hernan MA. The value of explicitly emulating a target trial when using real world evidence: an application to colorectal cancer screening. Eur J Epidemiol. 2017 Jun;32(6):495–500. PMID: 28748498
15. Admon AJ, Donnelly JP, Casey JD, Janz DR, Russell DW, Joffe AM, Vonderhaar DJ, Dischert KM, Stempek SB, Dargin JM, Rice TW, Iwashyna TJ, Semler MW. Emulating a Novel Clinical Trial Using Existing Observational Data. Predicting Results of the PreVent Study. Ann Am Thorac Soc. American Thoracic Society - AJRCCM; 2019 Apr 30;16(8):998–1007.
16. Hernán MA, Robins JM. Using Big Data to Emulate a Target Trial When a Randomized Trial Is Not Available. Am J Epidemiol. 2016 Apr 15;183(8):758–764. PMCID: PMC4832051
17. Brown SM, Paine R. Asking Causal Questions of Observational Data: The Quest Continues. Ann Am Thorac Soc. 2019 Aug;16(8):977–979. PMID: 31368800
18. Home - ClinicalTrials.gov [Internet]. [cited 2020 Mar 6]. Available from: https://clinicaltrials.gov/
19. Cipriani A, Barbui C. What is a clinical trial protocol? Epidemiol Psichiatr Soc. 2010 Apr;19(2):116–117. PMID: 20815294
20. Comparison of 23 mg Donepezil Sustained Release (SR) to 10 mg Donepezil Immediate Release (IR) in Patients With Moderate to Severe Alzheimer's Disease - Full Text View - ClinicalTrials.gov [Internet]. [cited 2020 Mar 6]. Available from: https://clinicaltrials.gov/ct2/show/NCT00478205
21. Farlow MR, Salloway S, Tariot PN, Yardley J, Moline ML, Wang Q, Brand-Schieber E, Zou H, Hsu T, Satlin A. Effectiveness and tolerability of high-dose (23 mg/d) versus standard-dose (10 mg/d) donepezil in moderate to severe Alzheimer's disease: A 24-week, randomized, double-blind study. Clin Ther. 2010 Jul;32(7):1234–1251. PMCID: PMC3068609
22. Shenkman E, Hurt M, Hogan W, Carrasquillo O, Smith S, Brickman A, Nelson D. OneFlorida Clinical Research Consortium: Linking a Clinical and Translational Science Institute With a Community-Based Distributive Medical Education Model. Acad Med J Assoc Am Med Coll. 2018;93(3):451–455. PMCID: PMC5839715
23. PCORnet. PCORnet Common Data Model v5.1 Specification (12 Sep 2019) [Internet]. 2019 [cited 2020 Mar 9]. Available from: https://pcornet.org/data-driven-common-model/
24. CFR - Code of Federal Regulations Title 21 [Internet]. [cited 2020 Mar 6]. Available from: https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfcfr/CFRSearch.cfm?fr=314.80
25. Stuart EA, Bradshaw CP, Leaf PJ. Assessing the Generalizability of Randomized Trial Results to Target Populations. Prev Sci Off J Soc Prev Res. 2015 Apr;16(3):475–485. PMCID: PMC4359056
26. He Z, Gonzalez-Izquierdo A, Denaxas S, Sura A, Guo Y, Hogan WR, Shenkman E, Bian J. Comparing and Contrasting A Priori and A Posteriori Generalizability Assessment of Clinical Trials on Type 2 Diabetes Mellitus. AMIA Annu Symp Proc AMIA Symp. 2017; PMID: 29854151
27. Shrimanker R, Beasley R, Kearns C. Letting the right one in: evaluating the generalisability of clinical trials. Eur Respir J [Internet]. European Respiratory Society; 2018 Dec 1 [cited 2020 Mar 8];52(6). Available from: https://erj.ersjournals.com/content/52/6/1802218 PMID: 30545963
28. Susukida R, Crum RM, Ebnesajjad C, Stuart E, Mojtabai R. Generalizability of findings from randomized controlled trials: Application to the National Institute of Drug Abuse Clinical Trials Network. Addiction [Internet]. Wiley-Blackwell; 2017 [cited 2020 Mar 8]; Available from: https://jhu.pure.elsevier.com/en/publications/generalizability-of-findings-from-randomized-controlled-trials-ap PMID: 28191694
29. Schwartz LM, Woloshin S. How the FDA forgot the evidence: the case of donepezil 23 mg. BMJ [Internet]. British Medical Journal Publishing Group; 2012 Mar 22 [cited 2020 Mar 8];344. Available from: https://www.bmj.com/content/344/bmj.e1086 PMID: 22442352
30. Deardorff WJ, Feen E, Grossberg GT. The Use of Cholinesterase Inhibitors Across All Stages of Alzheimer's Disease. Drugs Aging. 2015 Jul;32(7):537–547. PMID: 26033268
31. Cummings JL, Geldmacher D, Farlow M, Sabbagh M, Christensen D, Betz P. High-dose donepezil (23 mg/day) for the treatment of moderate and severe Alzheimer's disease: drug profile and clinical guidelines. CNS Neurosci Ther. 2013 May;19(5):294–301. PMCID: PMC6493345