Trials

METHODOLOGY

Open Access

Estimates of intra-cluster correlation coefficients from 2018 USA Medicare data to inform the design of cluster randomized trials in Alzheimer's and related dementias

Yongdong Ouyang[1,2*], Fan Li[3,4], Xiaojuan Li[5], Julie Bynum[6], Vincent Mor[7] and Monica Taljaard[8,9*]

Abstract

Background Cluster randomized trials (CRTs) are increasingly important for evaluating interventions embedded in health care systems. An essential parameter in sample size calculation to detect both overall and heterogeneous treatment effects for CRTs is the intra-cluster correlation coefficient (ICC) of both outcome and covariates of interest. However, obtaining advance estimates for the ICC can be challenging. When trial outcomes will be obtained from routinely collected data sources, there is an opportunity to obtain reliable ICC estimates in advance of the trial. Using USA national Medicare data, we estimated ICCs for a range of outcomes to inform the design of CRTs for people living with Alzheimer's and related dementias (ADRD).

Method Data from 2018 Medicare Fee-for-Service beneficiaries, specifically, 1,898,812 individuals (≥ 65 years) with diagnosis of ADRD within 3436 hospital service areas (treated as clusters) and 306 hospital referral regions (treated as fixed strata), were used to calculate unadjusted and adjusted ICC estimates for three outcomes: death, any hospitalizations, and any emergency department (ED) visits and three covariates: age, race and sex. We present both overall and stratum-specific ICC estimates. We illustrate their use in sample size calculations for overall treatment effects as well as detecting treatment effect heterogeneity.

Results The unadjusted overall ICCs for death, hospitalizations, and ED visits were 0.001, 0.010, and 0.017 respectively. Stratum-specific ICCs varied widely across the 306 HRRs: median 0.001, 0.010 and 0.025 for death, hospitalizations, and ED visits respectively and 0.007, 0.001, and 0.080 for age, sex and race. An interactive R Shiny app is provided that allows users to retrieve estimates overlayed on a map of the USA.

Conclusions We presented both adjusted and unadjusted ICCs for outcomes as well as unadjusted ICCs for covariates of potential interest from population-level data in the USA and demonstrated how the estimates may be used in sample size calculations for CRTs in ADRD.

Keywords Intra-cluster correlation coefficient, Cluster randomized trials, Interactive R shiny app, Treatment effect heterogeneity, Sample size calculation, Power

*Correspondence:
Yongdong Ouyang
yongdongouyang12@gmail.com
Monica Taljaard
mtaljaard@ohri.ca
Full list of author information is available at the end of the article

Introduction

Alzheimer's disease and related dementias (ADRD) are crippling illnesses that affect older persons' memory, mental processes, and functioning. For those who have Alzheimer's and their families, these diseases can have severe repercussions [1]. In 2021, approximately 6.2 million Americans aged 65 years or older were living with dementia; about two-thirds of Americans with ADRD are women [2]. This number could grow to 13.8 million by 2060 if there are no medical breakthroughs to prevent, slow, or cure the disease [2].

Randomized controlled trials are considered the gold standard for evaluating effectiveness of interventions. They allow researchers to provide high-quality evidence for identifying new potential treatments and interventions to slow down the progression of the disease, improve symptoms, or even prevent the disease altogether. In 2019, the National Institute on Aging funded the Imbedded Pragmatic Alzheimer's Disease and AD-Related Dementias Clinical Trials (IMPACT) Collaboratory to support pragmatic clinical trials of non-pharmacological interventions for dementia patients and their care partners that are integrated into health care systems. The cluster randomized trial (CRT) is an essential design for evaluating such interventions [3]. In a CRT, the units of randomization are not individuals but clusters (i.e., groups) of individuals where the clusters may constitute, for example, entire geographic regions, communities, hospitals, or clinics. Interventions may be delivered at the cluster level and/or directly to individual study subjects, but importantly, all participants within the same cluster receive the same intervention.

A key characteristic of a CRT includes that outcomes measured on multiple participants from the same cluster tend to be correlated, for example, due to shared influences associated with the cluster environment. The degree of correlation is typically measured by the intracluster correlation coefficient (ICC) [4]. The ICC is defined as the ratio of the between-cluster variance to the total variance in the outcome and can be interpreted as the average correlation between any two individuals from the same cluster. The ICC is a crucial parameter in both the design and analysis of CRTs. At the design stage, an estimate of the ICC is needed for sample size and power calculations [5–10]. The simplest way to account for the ICC in the sample size calculation is to multiply the sample size required under individual randomization by a design effect, defined as $DE = 1 + (n - 1)ICC$, where n is the average cluster size. Failing to account for the ICC or using an under-estimate may result in a smaller sample size than necessary, thereby increasing the risk of a type II error (failing to detect a treatment effect when it exists) [11, 12]. On the other hand, using an over-estimate may result in a larger sample size than necessary and may have implications for feasibility of the planned trial.

Designing a CRT with adequate power to detect heterogeneous treatment effects is of increasing interest [13]. Heterogeneous treatment effects, also known as "difference in difference estimates" refers to differential treatment effects across subgroups defined by covariates, for example, sex and ethnicity. These effects are often estimated via subgroup analyses or adding a treatment-by-covariate interaction term into the analysis. Sample size and power calculation methods for detecting heterogeneous treatment effects in CRTs have been developed and require estimates of adjusted ICCs and unadjusted ICCs for covariates of interest [14].

The CONSORT statement requires investigators reporting the results of a CRT to included estimated ICCs in their study as such estimates can be helpful to inform the design of future CRTs in similar populations [15]. However, it is often challenging for investigators to identify a plausible estimate for their planned trial at the design stage of a new trial. Several methodological reviews of CRTs have shown that authors seldom adhere to the CONSORT requirement to report ICCs [16]. CONSORT also does not require authors to report ICC estimates for covariates. One potential solution is to access historical data, e.g., from routinely collected data sources and estimate the ICC. Ouyang et al. [17] described and summarized methods to estimate ICCs when historical data are available. In several clinical areas, investigators have published databases of ICCs [18–20]. For some of these databases, investigators have identified "patterns" of ICCs to provide useful rules of thumb for the likely range of ICC values [20]. However, to our knowledge, such databases are not available in the field of ADRD research. Further, investigators may not have access or the necessary resources to allow them to obtain historical data specific to this disease area.

In this manuscript, we aimed to use the 2018 USA Medicare Fee-for-Service Beneficiaries database to provide ranges of ICCs for commonly used outcomes of interest in pragmatic trials in ADRD (death, hospitalizations, and emergency department [ED] visits) and covariates for potential treatment effect heterogeneity assessment (age, sex, and race) [14, 21–24]. In section "Methods", we describe the data source and the population included. We also describe the methods we used to calculate both adjusted and unadjusted ICCs for the variables mentioned above. Section "Results" presents the estimated ICCs and describes an interactive app that allows researchers to obtain relevant ICC values by region for their planned trial and to explore patterns of variation in the estimated ICC values. Section "Results" also includes examples to illustrate how to use these

estimates in informing the design of a hypothetical CRT to detect an overall treatment effect as well as differential treatment effects across subgroups of interest. Finally, we discuss the implications of our findings as well as some limitations in section "Discussion".

Methods

Data source and study population

We analyzed routinely collected data obtained from 2018 Medicare Fee-for-Service Beneficiaries. Medicare is a national health insurance program in the USA for people 65 and older, people under 65 with certain disabilities or people with end-stage renal disease. The program provides coverage through two payment mechanisms: either fee-for-service or managed care. For our purposes, beneficiaries were included if they were 65 years of age or older on January 1, 2018, resided in the USA, were continuously enrolled in Medicare under the fee-for-service mechanism for the entire year or until death. At the time of data preparation, the majority of older adults in the USA were managed as fee-for-service. We included only beneficiaries with diagnosed ADRD identified using a validated claims-based algorithm [25].

We obtained patient demographic characteristics (age, sex, race classified using the Research Triangle Institute race code), beneficiary ZIP code, and outcome information (death, hospitalization, and ED visit for any cause) from the administrative data. Age, sex, race, ZIP code and date of death were obtained from the Medicare Beneficiary Summary File. Medicare Provider Analysis and Review (MEDPAR) and Carrier (Physician/Supplier Part B claims) files were used to obtain hospitalizations and ED visit counts. A crosswalk from the Dartmouth Atlas was used to assign beneficiaries to their Hospital Referral Region (HRR) and Hospital Service Area (HSA) based on beneficiary ZIP code [26]. The HSAs were considered clusters in our analysis and HRRs as fixed strata. Note that the concept of the HSA is well-established in the USA, defined as geographic areas around hospitals where individual Medicare beneficiaries tend to be admitted [27]. HSAs are made up of ZIP code region groups based on tertiary medical care referral trends, defined by assigning HSAs to the region with the highest number of major cardiovascular operations performed, with minor adjustments. Thus, HSAs are larger geographic areas in which most Medicare beneficiaries secure specialty care [28]. Each HSA belongs to one HRR.

This study received expedited review approval from the Institutional Review Board of the University of Michigan.

Analytical dataset creation and descriptive analysis

The analytical dataset was constructed first by identifying all eligible beneficiaries and the presence of an ADRD diagnosis in 2018. For each individual, demographic characteristics and the occurrence of outcomes (any hospitalization, any ED visit or death in 2018) were identified. Each beneficiary was followed for one calendar year to assess outcomes. Individual-level patient records were clustered by HSA, with each HSA belonging to one HRR (treated as fixed strata in our analysis). The resultant dataset contained all Medicare Fee-For-Service beneficiaries in the USA in 2018 with a diagnosis of ADRD, across all 3436 HSAs and 306 HRRs across the USA.

To describe the analytic dataset, we present the mean (SD), median (inter-quartile range, IQR) and range of the population size across all HRRs and HSAs. Then, we summarize the outcomes and demographic characteristics (i.e., prevalence of death, hospitalization, and ED visits, as well as average age, proportion of males, proportion of whites) overall as well as by HRR. We show the variation in HRR-specific prevalence by presenting the mean (SD), median (IQR and range) across regions.

Estimation of intra-cluster correlation coefficients

We calculated ICCs for three binary outcomes (death, any hospitalization, any ED visit without hospital admission) and three covariates (age [continuous], sex [male vs female], and race [white vs non-white]). For each variable, we report four types of ICC estimates: (1) unadjusted overall ICC across all HRRs, (2) unadjusted HRR-specific ICC, (3) age-, sex-, and race-adjusted overall ICC across all regions, and (4) age-, sex-, and race-adjusted HRR-specific ICCs.

We chose the ANOVA estimator to obtain the unadjusted ICC estimates for our binary outcomes as the ANOVA estimator is commonly used for binary outcomes and conveniently yields estimates on the proportions scale, which is the scale needed for sample size calculation for binary outcomes. In particular, to obtain the overall unadjusted ICCs for binary variables across all HRRs, we used the stratified ANOVA estimator, with regions treated as fixed strata [29, 30]:

$$\widehat{\rho} = \frac{MSG - MSE}{MSG + (n_A - 1)MSE}$$

where

$$n_A = \left[N - \sum_{i=1}^{m} \sum_{j=1}^{k_i} \frac{n_{ij}}{N_i} \right] / \sum_{i=1}^{m} (k_i - 1)$$

$$MSG = \frac{\sum_{i=1}^{m} \sum_{j=1}^{k_i} n_{ij} (\overline{Y}_{ij.} - \overline{Y}_{i..})^2}{\sum_{i=1}^{m} k_i - m}$$

$$MSE = \frac{\sum_{i=1}^{m}\sum_{j=1}^{k_i}\sum_{l=1}^{n_{ij}}(Y_{ijk} - \overline{Y}_{ij.})^2}{N - \sum_{i=1}^{m} k_i}$$

In these formulae, $\widehat{\rho}$ is the ICC estimator, i indexes HRRs, j indexes HSAs, and l indexes individuals. We have m regions (in our case, m = 306), k_i service areas in the i th region, and n_{ij} individuals in the i th region and j th service area. Let Y_{ijl} be the binary response of the l th individual in the i th region and j th service area. Let $\overline{Y}_{ij.}$ Denote the mean response across individuals in the i th region and j th service area, and $\overline{Y}_{i..}$ the mean response for all individuals the i th region (or in the case of binary variables, the proportion). $N_i = \sum_{j=1}^{k_i} n_{ij}$ is the total number of individuals included in the i th region, and $N = \sum_{i=1}^{m} N_i$ is the total sample size. To calculate HRR-specific ICCs, we applied the standard ANOVA estimator within each region [29].

As the ANOVA estimator does not permit covariate adjustment, the adjusted ICC estimates for all three outcomes were calculated using a model-based generalized estimating equation (GEE) approach with an exchangeable correlation structure; GEE was chosen as it yields correlation estimates on the natural scale of the outcome [31, 32]. We fitted the model for each outcome adjusting for three demographic covariates (age, sex and race) for each HSA; for the overall ICC, we also included HRR as a fixed covariate (stratum indicator). For each variable, we presented the overall ICC estimates and the distribution of the 306 HRR-specific ICCs using mean, median, IQR, and boxplots. We did not report the individual HRR-adjusted ICCs due to the small number of HSAs in some regions.

The unadjusted ICCs for age, a continuous variable, were calculated using a linear mixed-effect model, treating HSA as a clustering variable. When obtaining overall ICCs, we added HRRs as a fixed covariate. Because negative ICCs are typically regarded as improbable in the context of CRTs, we set negative ICCs to zero [3]. All the analyses were conducted using SAS System for Unix.

Results

Cohort characteristics

A descriptive summary of the analytical dataset is presented in Table 1. We identified 1,898,812 Medicare beneficiaries with ADRD diagnosis (35.5% male; 81.2% white, mean [SD] age, 83.0 [0.9] years). In the year of 2018, 51.4% of these individuals had at least one hospitalization, 37.5% had at least one ED visit, and 23.9% died. These beneficiaries resided across 306 HRRs, with the average number of individuals per region ranging from 698 to 41,234. There was a total of 3436 HSA, averaging about 11 HSAs (range: 1 to 76) per region. The average number of individuals per HSA was 731 (range 66 to 12,168). Outcome prevalence varied across regions. The mean prevalence of death, any hospitalizations and any ED visits 23.9, 51.4 and 37.5% and ranged across regions from 17.2 to 34.4, 26.2 to 64.6% and 12.1 to 49.6%, respectively.

Intra-cluster correlation coefficients

Table 2 presents the overall unadjusted and adjusted ICCs for the three outcome variables and three covariates. The unadjusted overall ICCs for the outcomes were relatively small: death, 0.0008; hospitalization, 0.0095; and ED visits 0.0174. After adjusting for covariates,

Table 1 Information of 2018 Medicare Fee-for-Service beneficiaries who were 65 years or older with Alzheimer's disease and related dementias diagnosis

Characteristics of HRRs and HSAs	**Mean (SD)**	**Median (Q_1, Q_3)**	**Range (Min, Max)**	
HRR population size	6205 (6360)	4087 (2329, 7387)	(698, 41234)	
Number of HSAs per HRR	11 (10)	8 (5, 13)	(1, 76)	
HSA population size	731 (871)	516 (326, 869)	(66, 12168)	
Characteristics of Beneficiaries	**Overall**	**HRR-specific**		
Potential Trial Covariates of interests	***Mean (SD) or Prevalence (%)***	***Mean (SD) or Prevalence (%)***	***Median [Q_1, Q_3]***	***Range across HRRs (Min, Max)***
Age	83.0 (0.9)	83.0 (0.9)	83.0 (82.4, 83.7)	(80.9, 85.2)
Sex (% Male)	35.5	35.7	35.4 (34.4, 36.7)	(31.2, 42.2)
Race (% White)	81.2	84.4	88.8 (78.2, 94.4)	(16.3, 98.6)
Potential Trial Outcomes of interest	***Prevalence (%)***	***Mean (SD)***	***Median [Q_1, Q_3]***	***Range across HRRs (Min, Max)***
Death (%)	23.9%	24.4 (1.8)	24.4 (23.3, 25.4)	(17.2, 34.4)
Hospitalizations (%)	51.4%	49.8 (6.6)	50.7 (46.6, 54.0)	(26.2, 64.6)
At least one ED visit (%)	37.5%	36.5 (6.6)	37.8 (33.6, 40.6)	(12.1, 49.6)

HRR hospital referral region, *HSA* hospital service area

Table 2 Estimated ICCs for three outcomes (death, hospitalization and ED visit) and three covariates (age, sex and race) across all hospital referral regions

Outcome	Unadjusted		Adjusted[a]
	Overall[b]	Median (Q_1, Q_3)[b]	Overall
Death	0.001	0.001 [0, 0.002]	0.001
Hospitalizations	0.010	0.010 [0.003, 0.023]	0.006
Emergency department visits	0.017	0.025 [0.006, 0.052]	0.012
Covariates			
Age	0.007	0 [0, 0][c]	-
Sex	0.001	0.001 [0, 0.002]	-
Race	0.080	0.032 [0.011, 0.077]	-

[a] ICCs were adjusted for age, sex and race

[b] Treating hospital referral regions as fixed strata. The Median and IQR were obtained from the 306 hospital referral region-specific ICCs. The hospital referral region-specific ICCs are available in the R Shiny app

[c] The estimates were zeros when keeping three digits

ICCs for these outcomes slightly decreased to 0.0007, 0.0062, 0.0118, respectively. The distributions of HRR-specific unadjusted ICCs for the three outcome variables are visualized using boxplots in Fig. 1. The median and interquartile range (IQR) of the ICCs for the three outcome variables across the 306 HRRs were 0.001 (IQR 0, 0.002), 0.010 (IQR 0.003, 0.023) and 0.025 (IQR 0.006, 0.052), respectively. For covariates, the overall ICCs were 0.0070 for age, 0.0008 for sex, and 0.0801 for race. It is worth noting that ICCs for race were the largest among the covariates considered. HRR-specific ICCs of age were all close to zero. The HRR-specific ICCs varied moderately for both sex and race. Of most importance is that the outcome ICCs varied considerably across HRRs, suggesting that sample size requirements may vary considerably, depending on where the future trial is taking place. For example, consider a trial being planned with a primary outcome of hospitalizations. If the trial were to take place in HRR 339, the anticipated ICC would be 0.01 with an average cluster size of 300, yielding a design effect of 4, whereas if the trial were to take place in HRR 442, the anticipated ICC would be 0.11 with an average cluster size of 326 (similar to HRR 339), yielding a design effect of 37.

Interactive R Shiny app

To help researchers visualize and efficiently obtain relevant information to inform the design of their trials, we created an R Shiny app (https://douyang.shinyapps.io/adrdicc/). The R Shiny app contains two major features on separate tabs. The "Crude ICC" tab provides a geographical map of unadjusted ICCs for all 306 HRRs (see Fig. 2 for a screenshot). Users can hover over the cursor to a specific region to view the unadjusted ICCs. Users can also switch to different outcomes and view the distribution of ICCs. Note that adjusted ICCs are not provided as these estimates are potentially biased when the number of clusters is small.

Focusing on the HRR-specific information, users can move to the "HRR comparison" tab (Fig. 3). This section allows users to select multiple HRRs, which will be particularly useful when planning a trial across multiple

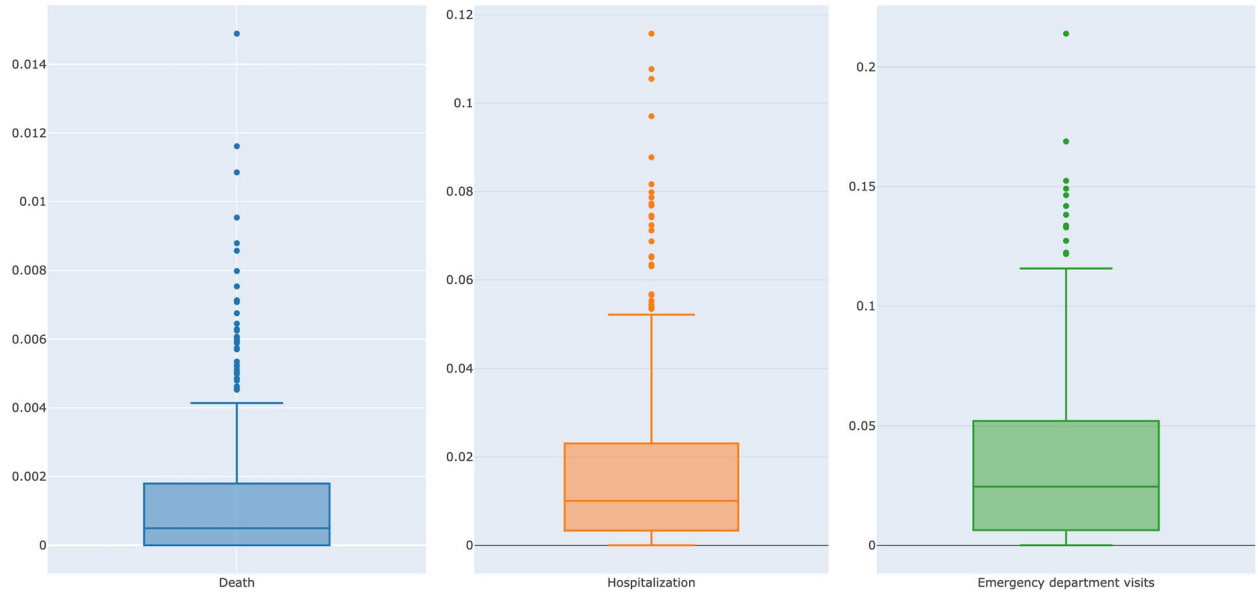

Fig. 1 Boxplots of health referral region-specific unadjusted ICC estimates for three outcomes

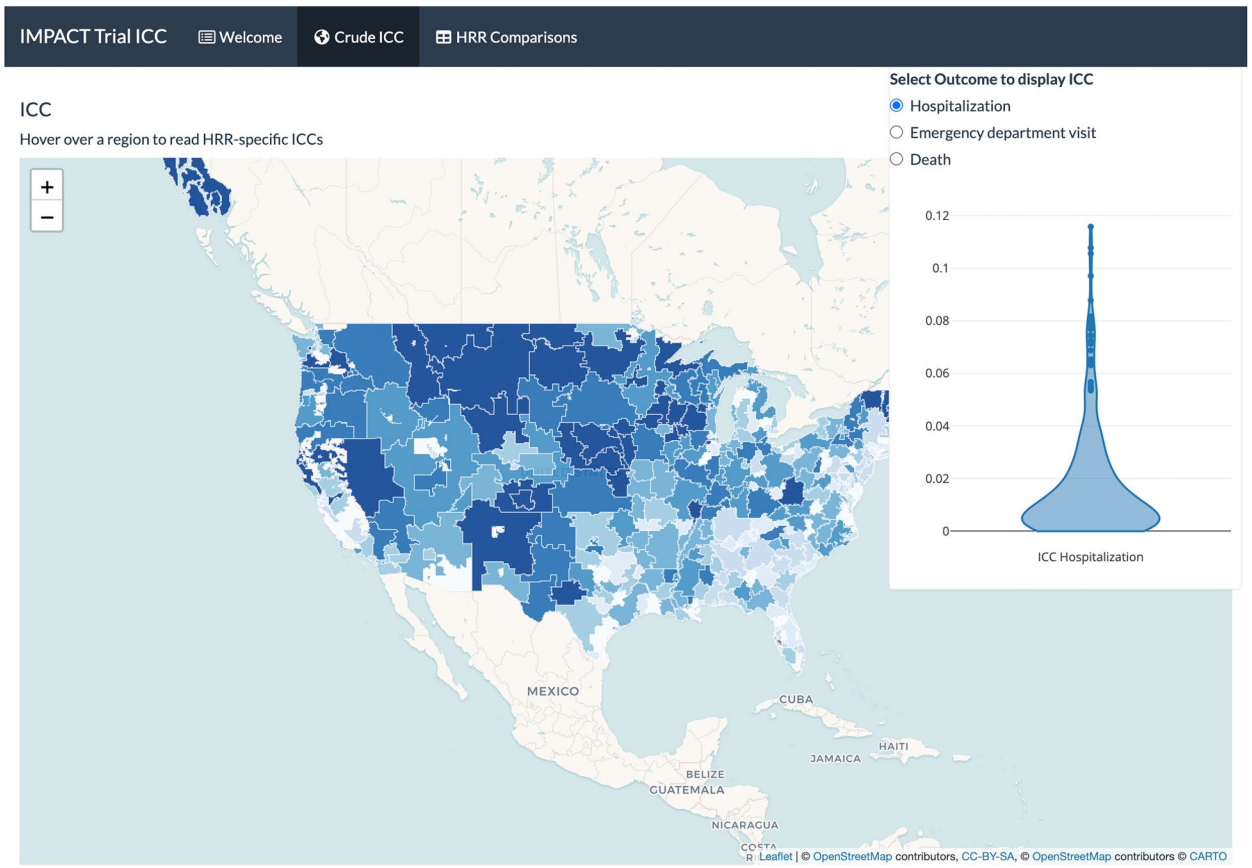

Fig. 2 A screenshot of the R Shiny app that displays health referral region-specific ICC estimates in an interactive map

regions. Users can select specific regions (top left panel), which will display relevant information (e.g., prevalence, ICC of outcomes, covariate information) in the table on the bottom panel, and the summary statistics (e.g., quartiles and mean) on the top right panel above the table.

Application

In this section, we illustrate how researchers can use the R Shiny app and ICC estimates to inform sample size calculations for a planned trial. Consider a hypothetical scenario, motivated by a recently published trial to evaluate the effectiveness of the high-dose trivalent influenza vaccine on hospital admissions among nursing home residents in the USA [33]. In this trial, 823 Medicare-certified nursing homes, located within 50 miles of a Centers for Disease Control and Prevention influenza reporting city, were randomized to either high-dose or standard-dose (standard of care) vaccine arms. The final analytical sample included individuals who were 65 or older and had been long-stay residents (staying in the facility for 90 days or more). The risks of respiratory-related hospital admissions were found to be 13% lower in facilities where residents were offered high-dose influenza vaccines compared to in facilities where residents were offered standard-dose vaccines.

Example 1: Powering a trial for detecting an overall treatment effect

Consider a planned trial for a similar intervention among individuals 65 years or older with ADRD in the USA, i.e., a similar target population as included in our Medicare data. The trial is to take place in 10 HRRs in New York state (Albany, Binghamton, Bronx, Buffalo, Elmira, East Long Island, New York, Rochester, Syracuse, and White Plains). There are 120 available HSAs (clusters) for the trial. Each HSA will be randomized to either intervention or control arm. The investigator wants to detect a minimum clinically important difference of 5% in the risk of any hospitalization (primary outcome), and determine the required sample size (number of residents) to reach 80% power with a two-sided test at a 5% significance level.

Using our R Shiny app, the baseline prevalence of hospital admissions among the 10 HRRs is approximately 53% and the estimated ICC for the outcome of hospitalization is 0.01. The required number of individuals per

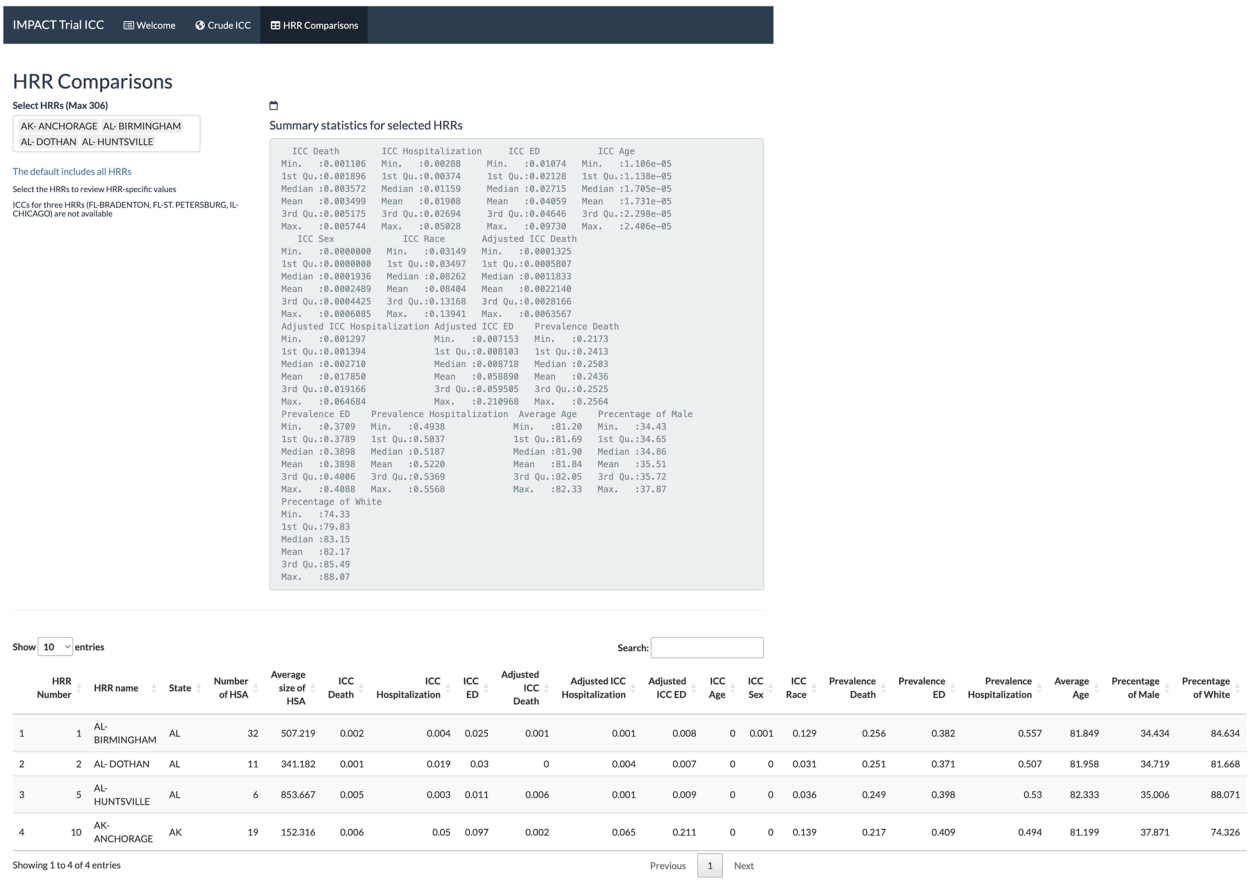

Fig. 3 A screenshot of the R Shiny app to display selected health referral region information and summaries. Legend: Summary statistics provided in the app include, by hospital referral region (HRR): Number of hospital service areas (HSA), average size of hospital service area, ICC estimates, prevalence estimates, and descriptive summaries of the covariates

cluster for a parallel arm CRT, given the available number of clusters (120) can then be determined. For example, by using the Shiny CRT calculator [7], an average of 35 individuals per HSA is required to achieve 80% power. As a sensitivity analysis, the required sample sizes under a range of ICC values can be determined, for example, by using lower and upper quartiles. Among the 10 HRRs, the 25th and 75th percentiles of ICCs were 0.003 and 0.016 respectively which yield required sample sizes ranging from 29 to 45 individuals per HSA.

Example 2: Powering a trial for detecting heterogeneous treatment effects

Consider the same scenario as example 1, and now suppose the planned trial has a secondary objective to examine the differential impact of the high-dose influenza vaccine on reducing hospitalizations among white vs non-white individuals. Using the sample size calculation formula to detect heterogeneous treatment effects developed by Yang et al. [14], we require the following information: (1) the adjusted ICC of hospitalization, (2) the prevalence of hospitalization (the variance then can be derived by $P(1-P)$, where P is the prevalence), (3) the ICC of race (white vs. non-white), and (4) the prevalence of the binary race variable.

Following example 1, we know the prevalence of hospitalization is 53%. We obtain the adjusted outcome ICC at 0.01. The variance (calculated from the prevalence of white residents) and ICC of race can both be obtained from the Shiny app at 0.16 and 0.09, respectively. With 35 residents per HSA (according to the sample size calculation in example 1), we would be able to detect a treatment-by-race interaction effect size of 11% (i.e., an 11% difference in the effect of the high-dose vaccine on white versus non-white individuals) with 80% power.

Discussion

In this study, we used data from Medicare Fee-for-Service beneficiaries who were 65 years old or older with a diagnosis of ADRD in 2018 to estimate the ICCs for three relevant binary outcomes for a future planned CRT: death, any hospitalization, and any ED visits, and three

covariates that are potentially relevant when interest lies in assessing treatment effect heterogeneity (age, sex, and race). With the ICC estimates, we presented examples to calculate sample sizes for detecting both overall and heterogeneous treatment effects.

Hospitalization was the most common among the three outcomes we investigated. Our analysis suggested that the overall ICCs for all three outcomes were generally small (but, as the design effect is a function of both the ICC and the cluster size, clustering can still have a substantial impact on the required sample size for a future CRT). The ICCs for death were similar across HRRs partially because the prevalence of death was similar across HRRs. The ICCs for ED visits had the largest variation. Among the three covariates, age and sex had small overall ICCs, while race had a much larger overall ICC. The ICCs and their variation were also large for race, which indicates individuals with the same race may have more strongly correlated responses, and this should be considered during the trial design. Compared to the overall unadjusted ICCs of the three outcomes, ICCs decreased after adjusting for age-, sex-, and race, suggesting that covariate adjustment may help to reduce between-cluster variation [34]. If the prespecified primary analysis will adjust for covariates, investigators may use the adjusted outcome ICCs for powering the trial, although unadjusted estimates may be preferred as they yield more conservative sample size estimates. The adjusted outcome ICCs may also be used to power the trial for detecting treatment effect heterogeneity [14, 22, 35].

We created an R Shiny app that allows users to visualize and retrieve estimated overall and HRR-specific ICCs on an interactive map and explore spatial patterns in variation of these estimates. When a trial is being planned across multiple HRRs, this web-based app can also summarize the anticipated prevalence, plausible range of ICCs, and covariate ICC values for the selected HRRs.

We have presented examples to illustrate how researchers can determine sample sizes needed to detect overall and heterogeneous treatment effects in a CRT using these estimates. Given the same sample size, we can detect 5% differences in the overall treatment effect, but 11% differences in heterogeneous treatment effects (i.e., difference-in-difference estimates). This means that the conventional sample size calculation for detecting overall treatment effects may not be sufficient to yield the same power for detecting heterogeneous treatment effects. As a result, a separate calculation is required.

A somewhat surprising finding was that in some cases, ICCs did not decrease after covariate adjustment. Although one would expect ICCs to decrease after covariate adjustment, especially when covariates are strongly predictive of the outcome, Murray and Blistein [34] found that ICCs may not always decrease after covariate adjustment. This issue warrants further investigation, but we suspect that it may occur when (1) the number of clusters is small, (2) covariates are only weakly correlated with the outcome, or (3) outcomes are binary.

To our best knowledge, this is one of the first studies to empirically estimate ICCs using a dataset focusing on populations with ADRD, which is highly relevant given the substantial efforts in building national infrastructure for conducting embedded pragmatic clinical trials for the ADRD population in the USA [36]. Similar studies have been done using datasets in other clinical areas [18–20]. ICC estimates for planning CRTs are generally imprecise. When empirical data are not available, researchers often have to rely on crude rules of thumb; when empirical data are available, a large number of clusters and individuals are usually desirable to obtain a more precise estimate. The main strength of our study is the use of a population-level data source that allowed us to analyze data for the entire population of potentially eligible clusters and patients. We included 3436 clusters (HSAs), with large numbers of people in each cluster. Usually, access to Medicare dataset is restricted, and researchers may not be able to obtain relevant values to inform their trial designs on similar populations. Our results included a comprehensive list of estimates for both overall and HRR-specific prevalence and ICCs for the entire population. Researchers planning a CRT often have to rely on published ICC estimates from other studies and necessarily have to extrapolate to their target population. The extent to which our ICC estimates apply to other clinical areas and other jurisdictions is unknown, but we encourage investigators to use judgment, supported by the reported demographic characteristics, cluster size information and prevalence estimates to make conclusions about applicability of these overall or region-specific ICC estimates to their planned trial.

There are some limitations of our work. First, the number of HSA in each HRR ranges from 1 to 76, and over 60% of HRRs have fewer than 10 HSAs. Therefore, some HRR-specific ICCs were not estimable or may be inaccurate due to a small number of clusters [37]. Interpreting ICCs derived from a very small number of clusters is challenging but sometimes unavoidable in CRTs [38]. Second, we used data from 2018 which was the most recent data available to us; however, these are population-level data and we do not expect that major changes could have occurred at the population-level in a small number of years. Third, the available analytical dataset is entirely cross-sectional (diagnosis and utilization were determined in the same year). However, this was considered an acceptable limitation as dementia does not have an abrupt onset, and people typically receive a diagnosis

at a late stage in their illness. A claim on a particular date is therefore an indication that the person has the disease and likely has had for at least a year prior (more likely longer). Fourth, we estimated unadjusted overall and stratified ICCs using the ANOVA approach. There is no consensus on the optimal definition and best estimation method for ICC parameters with binary outcomes [29, 39]. Several studies have presented different methods to estimate ICCs for binary outcomes [40–42], and recent studies have compared different methods [29, 39, 42]. Wu et al. [35] compared the performance of different ICC estimators for binary outcomes including GEE, ANOVA, Fleiss-Cuzick, and Pearson and found that the ANOVA estimator is generally acceptable, except in cases where the outcome probability varies across different trial arms. Ridout and Firth [27] evaluated the ANOVA estimator using simulation and found this was among the estimators with the lowest MSE. However, none of the methods dominates others in term of performance. In our analysis, we chose the ANOVA estimator for reporting the unadjusted ICCs of binary variables and GEE for reporting the adjusted ICCs because they are commonly used methods in the literature, and the resulting ICC estimates can be interpreted on the natural scale of the outcome.

Conclusion

In summary, planning a CRT to detect either overall or heterogeneous treatment effects requires accurate estimates of ICCs for both outcome and covariates of interest. We present estimated ICCs for the outcomes of death, any hospitalizations and any emergency department visits as well as for three covariates (age, sex, and race). We find that estimates vary widely across the USA and thus, sample size calculation for planned future CRTs randomizing HSA with one of these outcomes as primary outcome should ideally use region-specific estimates. We describe sample size parameters for three potential trial outcomes and provide an interactive R Shiny app that can help investigators obtain relevant parameters for these outcomes to inform the design of future CRTs in ADRD.

Abbreviations

CRT	Cluster randomized trial
ICC	Intracluster correlation coefficient
ADRD	Alzheimer's and related dementias
HRR	Health referral regions
ED	Emergency department
HSA	Health service area

Acknowledgements

We obtained Hospital Referral Region Geographical Boundary File from Dartmouth Atlas Data website to create our Rshiny app. Dartmouth Atlas Data website was funded by the Robert Wood Johnson Foundation, The Dartmouth Clinical and Translational Science Institute, under award number UL1TR001086 from the National Center for Advancing Translational Sciences (NCATS) of the National Institutes of Health (NIH), and in part, by the National Institute of Aging, under award number U01 AG046830.

Authors' contributions

Y.O led the writing of this manuscript. F.L and M.T. provided valuable inputs and contributed significantly to the revision of this manuscript. All authors contributed to the writing and revision of this manuscript.

Funding

M.T., X.L., J.B., V.M., and F.L. are supported by the National Institute of Aging (NIA) of the National Institutes of Health (NIH) under Award Number U54AG063546, which funds NIA Imbedded Pragmatic Alzheimer's Disease and AD-Related Dementias Clinical Trials Collaboratory (NIA IMPACT Collaboratory). F.L. is also supported by a Patient-Centered Outcomes Research Institute Award (PCORI Award ME-2020C3-21072). Y.O. received the Canadian Institutes of Health Research Health System Impact Postdoctoral Fellowship. X.L. is supported by Career Development Award K01AG073651 from the NIA of the National Institutes of Health. The statements presented in this article are solely the responsibility of the authors and do not necessarily represent the official views of NIH, PCORI, or its Board of Governors or Methodology Committee.

Availability of data and materials

Data from Medicare is not available due to privacy and legislation.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Child Health Evaluative Sciences, The Hospital for Sick Children, 686 Bay Street, Toronto, ON, Canada. [2]Dalla Lana School of Public Health, University of Toronto, 155 College St, Toronto, ON, Canada. [3]Department of Biostatistics, Yale School of Public Health, New Haven, CT, USA. [4]Center for Methods in Implementation and Prevention Science, Yale School of Public Health, New Haven, CT, USA. [5]Department of Population Medicine, Harvard Medical School and Harvard Pilgrim Health Care Institute, Boston, MA, USA. [6]Division of Geriatric and Palliative Medicine, Department of Internal Medicine, University of Michigan Medical School, Ann Arbor, MI, USA. [7]Center for Gerontology and Healthcare Research, School of Public Health, Brown University, Providence, RI, USA. [8]Clinical Epidemiology Program, Ottawa Hospital Research Institute, 1053 Carling Ave, Ottawa, ON, Canada. [9]School of Epidemiology and Public Health, University of Ottawa, 600 Peter Morand Crescent, Ottawa, ON, Canada.

Received: 13 March 2024 Accepted: 16 August 2024
Published online: 30 October 2024

References

1. Alzheimer's Disease and Related Dementias. National Institute on Aging, https://www.nia.nih.gov/health/alzheimers. Accessed 9 Nov 2022.
2. 2021 Alzheimer's disease facts and figures. Alzheimers Dement. 2021;17(3):327–406. https://doi.org/10.1002/alz.12328.
3. Donner A. Design and analysis of cluster randomization trials in health research. London: Arnold; 2000.
4. Donner A, Koval JJ. Design considerations in the estimation of intraclass correlation. Ann Hum Genet. 1982;46:271–7.
5. Ouyang Y, Xu L, Karim ME, et al. CRTpowerdist: an R package to calculate attained power and construct the power distribution for cross-sectional stepped-wedge and parallel cluster randomized trials. Comput Methods Programs Biomed. 2021;208:106255.
6. Chen J, Zhou X, Li F, et al. swdpwr: a SAS macro and an R package for power calculations in stepped wedge cluster randomized trials. Comput Methods Programs Biomed. 2022;213:106522.

7. Hemming K, Kasza J, Hooper R, et al. A tutorial on sample size calculation for multiple-period cluster randomized parallel, cross-over and stepped-wedge trials using the Shiny CRT Calculator. Int J Epidemiol. https://doi.org/10.1093/ije/dyz237. Epub ahead of print 22 February 2020.
8. Zhang Y, Preisser JS, Li F, et al. GEEMAEE: A SAS macro for the analysis of correlated outcomes based on GEE and finite-sample adjustments with application to cluster randomized trials. Comput Methods Programs Biomed. 2023;230:107362.
9. Ouyang Y, Li F, Preisser JS, Taljaard M. Sample size calculators for planning stepped-wedge cluster randomized trials: a review and comparison. Int J Epidemiol. 2022;51(6):2000–13. https://doi.org/10.1093/ije/dyac123.
10. Zhan D, Xu L, Ouyang Y, et al. Methods for dealing with unequal cluster sizes in cluster randomized trials: a scoping review. PLoS One. 2021;16:e0255389.
11. Ouyang Y, Kulkarni MA, Protopopoff N, et al. Accounting for complex intracluster correlations in longitudinal cluster randomized trials: a case study in malaria vector control. BMC Med Res Methodol. 2023;23:64.
12. Kasza J, Bowden R, Ouyang Y, Taljaard M, Forbes AB. Does it decay? Obtaining decaying correlation parameter values from previously analysed cluster randomised trials. Stat Methods Med Res. 2023;32(11):2123–34. https://doi.org/10.1177/09622802231194753.
13. Starks MA, Sanders GD, Coeytaux RR, et al. Assessing heterogeneity of treatment effect analyses in health-related cluster randomized trials: a systematic review. PLoS One. 2019;14:e0219894.
14. Yang S, Li F, Starks MA, et al. Sample size requirements for detecting treatment effect heterogeneity in cluster randomized trials. Stat Med. 2020;39:4218–37.
15. Campbell MK, Piaggio G, Elbourne DR, et al. Consort 2010 statement: extension to cluster randomised trials. BMJ. 2012;345:e5661.
16. Ivers NM, Taljaard M, Dixon S, et al. Impact of CONSORT extension for cluster randomised trials on quality of reporting and study methodology: review of random sample of 300 trials, 2000–8. BMJ. 2011;343:d5886.
17. Ouyang Y, Hemming K, Li F, et al. Estimating intra-cluster correlation coefficients for planning longitudinal cluster randomized trials: a tutorial. Int J Epidemiol. 2023:dyad062.
18. Taljaard M, Donner A, Villar J, et al. Intracluster correlation coefficients from the 2005 WHO Global Survey on Maternal and Perinatal Health: implications for implementation research. Paediatr Perinat Epidemiol. 2008;22:117–25.
19. Thompson DM, Fernald DH, Mold JW. Intraclass correlation coefficients typical of cluster-randomized studies: estimates from the Robert Wood Johnson Prescription for Health projects. Ann Fam Med. 2012;10:235–40.
20. Gulliford MC, Adams G, Ukoumunne OC, et al. Intraclass correlation coefficient and outcome prevalence are associated in clustered binary data. J Clin Epidemiol. 2005;58:246–51.
21. Li F, Chen X, Tian Z, et al. Designing three-level cluster randomized trials to assess treatment effect heterogeneity. Biostatistics 2022:kxac026.
22. Tong G, Esserman D, Li F. Accounting for unequal cluster sizes in designing cluster randomized trials to detect treatment effect heterogeneity. Stat Med. 2022;41:1376–96.
23. Tong J, Li F, Harhay MO, et al. Accounting for expected attrition in the planning of cluster randomized trials for assessing treatment effect heterogeneity. BMC Med Res Methodol. 2023;23:85.
24. Ryan MM, Esserman D, Li F. Maximin optimal cluster randomized designs for assessing treatment effect heterogeneity. https://doi.org/10.48550/arXiv.2302.09138. Epub ahead of print 30 May 2023.
25. McCarthy EP, Chang C-H, Tilton N, et al. Validation of claims algorithms to identify Alzheimer's disease and related dementias. J Gerontol A Biol Sci Med Sci. 2022;77:1261–71.
26. Dartmouth Atlas of Health Care. Dartmouth Atlas of Health Care. https://www.dartmouthatlas.org/. Accessed 21 Mar 2023.
27. Hu Y, Wang F, Xierali IM. Automated delineation of hospital service areas and hospital referral regions by modularity optimization. Health Serv Res. 2018;53:236–55.
28. Wennberg JE, Cooper MM. Strategies and methods. In: The Dartmouth Atlas of Health Care in the United States: The Center for the Evaluative Clinical Science. Chicago: American Hospital Publishing, Inc.; 1996. https://www.ncbi.nlm.nih.gov/books/NBK585005/. Accessed 14 Jul 2023.
29. Ridout MS, Demétrio CG, Firth D. Estimating intraclass correlation for binary data. Biometrics. 1999;55:137–48.
30. Donner A. A review of inference procedures for the intraclass correlation coefficient in the one-way random effects model. Int Stat Rev. 1986;54:67–82.
31. Preisser JS, Young ML, Zaccaro DJ, et al. An integrated population-averaged approach to the design, analysis and sample size determination of cluster-unit trials. Stat Med. 2003;22:1235–54.
32. Zeger SL, Liang KY. Longitudinal data analysis for discrete and continuous outcomes. Biometrics. 1986;42:121–30.
33. Gravenstein S, Davidson HE, Taljaard M, et al. Comparative effectiveness of high-dose versus standard-dose influenza vaccination on numbers of US nursing home residents admitted to hospital: a cluster-randomised trial. Lancet Respir Med. 2017;5:738–46.
34. Murray DM, Blistein JL. Methods to reduce the impact of intraclass correlation in group-randomized trials. Eval Rev. 2003;27:79–103.
35. Rutterford C, Copas A, Eldridge S. Methods for sample size determination in cluster randomized trials. Int J Epidemiol. 2015;44:1051–67.
36. Brody AA, Barnes DE, Chodosh J, et al. Building a national program for pilot studies of embedded pragmatic clinical trials in dementia care. J Am Geriatr Soc. 2020;68(Suppl 2):S14–20.
37. Preisser JS, Lu B, Qaqish BF. Finite sample adjustments in estimating equations and covariance estimators for intracluster correlations. Stat Med. 2008;27:5764–85.
38. Nevins P, Ryan M, Davis-Plourde K, et al. Adherence to key recommendations for design and analysis of stepped-wedge cluster randomized trials: a review of trials published 2016–2022. Clin Trials. 2023:17407745231208397.
39. Wu S, Crespi CM, Wong WK. Comparison of methods for estimating the intraclass correlation coefficient for binary responses in cancer prevention cluster randomized trials. Contemp Clin Trials. 2012;33:869–80.
40. Turner RM, Omar RZ, Thompson SG. Constructing intervals for the intracluster correlation coefficient using Bayesian modelling, and application in cluster randomized trials. Stat Med. 2006;25:1443–56.
41. Chakraborty H, Moore J, Carlo WA, et al. A simulation based technique to estimate intracluster correlation for a binary variable. Contemp Clin Trials. 2009;30:71–80.
42. Zou G, Donner A. Confidence interval estimation of the intraclass correlation coefficient for binary outcome data. Biometrics. 2004;60:807–11.

Publisher's Note