On the Analysis of Case–Control Studies in Cluster-correlated Data Settings

Sebastien Haneuse and Claudia Rivera-Rodriguez

Abstract: In resource-limited settings, long-term evaluation of national antiretroviral treatment (ART) programs often relies on aggregated data, the analysis of which may be subject to ecological bias. As researchers and policy makers consider evaluating individual-level outcomes such as treatment adherence or mortality, the well-known case–control design is appealing in that it provides efficiency gains over random sampling. In the context that motivates this article, valid estimation and inference requires acknowledging any clustering, although, to our knowledge, no statistical methods have been published for the analysis of case–control data for which the underlying population exhibits clustering. Furthermore, in the specific context of an ongoing collaboration in Malawi, rather than performing case–control sampling across all clinics, case–control sampling within clinics has been suggested as a more practical strategy. To our knowledge, although similar outcome-dependent sampling schemes have been described in the literature, a case–control design specific to correlated data settings is new. In this article, we describe this design, discuss balanced versus unbalanced sampling techniques, and provide a general approach to analyzing case–control studies in cluster-correlated settings based on inverse probability–weighted generalized estimating equations. Inference is based on a robust sandwich estimator with correlation parameters estimated to ensure appropriate accounting of the outcome-dependent sampling scheme. We conduct comprehensive simulations, based in part on real data on a sample of N = 78,155 program registrants in Malawi between 2005 and 2007, to evaluate small-sample operating characteristics and potential trade-offs associated with standard case–control sampling or when case–control sampling is performed within clusters.
(*Epidemiology* 2018;29: 50–57)

The case–control design is a mainstay of epidemiologic research. Central to its appeal is that rare binary outcomes can be investigated in a cost-effective manner without the need to sample a huge number of study units from the population of interest. Building on early seminal work,[1–3] the literature has expanded rapidly over the last 40 years to give researchers a wide variety study designs, including matched case–control,[3] two-phase,[4,5] and case-crossover[6] designs for binary outcomes; nested case–control[7,8] and case–cohort[9,10] designs for time-to-event outcomes; and outcome-dependent sampling schemes for continuous responses.[11]

Recently, a number of outcome-dependent sampling schemes have been proposed for longitudinal or cluster-correlated data settings. For the most part, this work has focused on correlated binary data, including designs for longitudinal data[12–14]; family-based case–control sampling in genetic studies[15–17]; and cluster-based sampling where a subset of clusters is chosen on the basis of the observed outcome rates and detailed information (retrospectively) collected on all study units within each of the chosen clusters.[18] As these schemes have been developed, care has been taken to ensure that valid analysis methods have also been developed, specifically to ensure that correlation attributable to clustering is appropriately taken into account.[19] To our knowledge, no one has considered the analysis of data arising from a standard case–control design applied to a population in which the study units are naturally cluster-correlated by, say, geographic location or hospital/physician. That is, no one has considered the scenario where a case–control study is conducted wherein cases and controls are selected without regard to cluster membership and yet cluster membership, which is retained nonetheless, must be accounted for post-hoc. Consider, for example, the evaluation of outcomes among patients registered in the national antiretroviral treatment (ART) program in Malawi.[20] As we elaborate below, current data collection efforts undertaken by the Malawian Ministry of Health focus on clinic-level aggregated counts. As interest increasingly turns to understanding risk factors for patient-specific outcomes such as program retention, treatment adherence, and mortality, the case–control design might be a viable option for cost-effective collection of patient-level data. The analysis of data obtained from such a study, however, would need to account for the fact that program registrants are naturally clustered by clinic. The first major goal of this article, therefore, is to present and evaluate a valid analysis strategy for data from a case–control study for which cluster membership must be accounted for post-hoc.

Editor's Note: A Commentary on this article appears on p.76.
Submitted July 20, 2016; accepted September 27, 2017.
From the Harvard T.H. Chan School of Public Health, Boston, MA.
Supported, in part, by Harvard University Center for AIDS Research Feasibility Project grant P03 A106054 and National Institutes of Health grant 5DP1 ES025459. All code for the simulations is available from the first author on request.
The authors report no conflicts of interest.
SDC Supplemental digital content is available through direct URL citations in the HTML and PDF versions of this article (www.epidem.com).
Correspondence: Sebastien Haneuse, Department of Biostatistics, Harvard T.H. Chan School of Public Health, 655 Huntington Ave, Building II, Boston, MA 02115. E-mail: shaneuse@hsph.harvard.edu.

ISSN: 1044-3983/18/2901-0050
DOI: 10.1097/EDE.0000000000000763

In parallel to the standard case–control design, a second design that could be employed in the Malawian context is one where case–control sampling is performed within each clinic. This design has appeal in that it would provide at least some patient-level data from each clinic, while requiring considerably less ongoing coordination between the Ministry of Health and the clinics. To the best of our knowledge, however, such a case–control design for cluster data has not been explicitly described. One related design is the hybrid design of Haneuse and Wakefield[21,22] in which an ecological study is supplemented with case–control data. Although potentially useful in the Malawian context, the practical utility of the design is limited by the severe computational burden associated with existing analysis methods.[23] The second major goal of this article, therefore, is to formally describe the proposed design and to develop and evaluate a valid analysis strategy that appropriately accounts for cluster-correlation.

METHODS

Monitoring and Evaluation of the National ART Program in Malawi

In the developing world, national antiretroviral treatment (ART) programs are often designed to be simple, standardized, and decentralized to ensure as broad and efficient coverage as possible.[24,25] One such program was initiated in 2004 in Malawi, a southern African nation for which HIV/AIDS is the leading cause of death among adults ages 15–49 years.[26] Overall, the goals of the Malawian ART program are to reduce population-level mortality and morbidity attributable to HIV/AIDS, as well as to increase the percent of HIV-positive adults physically capable of staying in the workforce.[20]

Although decentralized national ART programs have been shown to be effective in facilitating a rapid scale-up of treatment coverage, a drawback is that the quality of data available for monitoring and evaluation is typically limited.[27] In the current Malawian program, for example, although detailed patient-level information is recorded on paper-based "mastercards," routine data collection consists of aggregated admission counts, clinic-level covariate data, and outcome tallies, collected every 3 months.[20] As such, the data available for analyses constitute an ecologic study.[28]

In practice, analyses based on aggregated or cluster-level data are subject to a range of potential biases, an umbrella term for which is ecologic bias.[29,30] Although the literature is rich with methods for analyzing aggregated data, the only reliable approach to overcoming ecologic bias, if one is to avoid making untestable assumptions, is to collect and analyze patient-level data.[31] In the long-term, the Malawian Ministry of Health and Population is developing an electronic system for storing patient-level data.[32] In the meantime, however, detailed patient-level data are not available on a routine basis, which presents a significant dilemma. One solution is to focus data collection efforts on a judiciously chosen subsample of patient registrants. When the outcome of interest is binary and (relatively) rare, the case–control study design is well known to be highly efficient relative to random sampling.[33] Such a design could be readily implemented in the Malawian context, although, as already emphasized, valid inference would require acknowledging the clustering of program registrants within clinic.

The Complete Data Setting

Suppose the population of interest can be classified into one of K mutually exclusive groups or clusters. In practice, study units and clusters may correspond to patients within hospitals or clinics, as in the Malawi data, or children in schools, or individuals who reside in close proximity to each other. Let N_k denote the number of study units in the kth cluster. For the ith study unit in the kth cluster, let Y_{ki} denote the outcome of interest and X_{ki} a vector of risk factors/covariates, with the first element corresponding to the intercept.

From a scientific perspective, we assume that interest lies with the mean of the outcome given the covariates, denoted by $\mu_{\text{ki}} = E[Y_{ki}|X_{ki}]$ for the ith study unit in the kth cluster. Typically, μ_{ki} is assumed to be related to X_{ki} via some link function $g()$, specifically that:

$$g(\mu_{ki}) = X_{ki}^T \boldsymbol{\beta} \quad (1)$$

with $\boldsymbol{\beta}$ a vector of regression coefficients. If Y is a binary outcome, for example, a common choice for $g()$ is the logit function, in which case (1) would correspond to a logistic regression.

Estimation and Inference

In complete data settings, that is where (Y_{ki}, X_{ki}) is observed for all N_k study units in each of the K clusters, analyses are typically conducted using generalized linear mixed models[34] or generalized estimating equations[35] (GEE). Although much can be said of the relative merits of these two frameworks[36], moving forward, we assume that in the complete data setting, primary interest would lie in estimation/inference with respect to the marginal parameters in model (1) based on GEE. Practically, in conducting GEE analysis, one is required to specify a working correlation structure. For the most part, the choice of this structure will be dictated by efficiency considerations (i.e., the closer the choice of working correlation structure to the true correlation structure, the more efficient the GEE estimator), although in some settings, researchers may wish to adopt a working independence structure.[37] Regardless of the choice, however, the robust sandwich estimator can be used to obtain valid inference. eAppendix 1 (http://links.lww.com/EDE/B278) provides technical details.

Application to the Malawi Data

Between April 2008 and May 2009, the Malawian Ministry of Health conducted a one-time cross-sectional survey of their national ART program. To help illustrate the ideas of this article, we consider a hypothetical study of the relationship between select covariates measured at the time of registration with a binary outcome of "status at six months

post-registration." For this outcome, patients were considered to have a negative status ($Y = 1$) if they were recorded to have stopped treatment, been lost to follow-up, or had died; patients were considered to have a non-negative outcome status ($Y = 0$) if they had either transferred to another clinic or if they were alive and on treatment.

Measures are in place in all ART facilities to ensure patient confidentiality, consent for HIV testing, and counseling and support for those who receive a positive HIV test result. Studies using data collected routinely within the context of monitoring and evaluation, such as ART registers, do not require formal approval by the Malawi National Health Science Research Committee. Before data analysis, all individual-level data was completely deidentified.

Table 1 and Figure 1 provide information on select patient- and clinic-level covariates for N = 78,155 patients ages 18–65 years, who registered at one of $K = 189$ clinics between 2005 and 2007 with a WHO clinical stage of 3 or 4. The latter is a clinical classification of HIV/AIDS infection stage used in resource-limited settings in lieu of laboratory-based measurements such as CD4 counts.[38] From Table 1, most registrants were female (60.9%) and that registration increased over time between 2005 and 2007. We also see that the majority of patients registered at one of the 147 public clinics (97.3%). Figure 1 provides a graphical representation of the clinic sizes across the $K = 189$ clinics.

TABLE 1. Distributions of Patient- and Clinic-Specific Characteristics, Together with Negative Outcome Status Rates, Among N = 78,155 Patients Ages 18–65 Years Who Registered Between 2005 and 2007 with a WHO HIV/AIDS Clinical Stage of 3 or 4

	Overall N (%)	Negative Outcome Status, %
Total	78,155 (100.0)	20.3
Sex		
Male	30,588 (39.1)	23.0
Female	47,567 (60.9)	18.5
Age, years		
16–25	8,735 (11.2)	23.9
26–35	31,111 (39.8)	20.8
36–45	24,453 (31.3)	19.1
46–55	10,087 (12.9)	18.3
≥56	3,769 (4.8)	20.5
WHO stage		
3	59,132 (75.7)	17.5
4	19,023 (24.3)	28.7
Year of registration		
2005	15,256 (19.5)	22.9
2006	28,900 (37.0)	21.1
2007	33,999 (43.5)	18.4
Clinic type		
Public	76,011 (97.3)	20.4
Private	2,144 (2.7)	13.8

WHO, World Health Organization.

Returning to Table 1, 20.3% of the N = 78,155 program registrants had a negative outcome status at 6 months. We also see that female registrants had a lower rate of negative 6-month status than males (18.5% vs 23.0%), as did patients who presented with a WHO clinical state of 3 compared with those with a stage of 4 (17.5% vs 28.7%), as well as those who registered in a private clinic when compared with those who registered in a public clinic (13.8% vs 20.4%). As shown in Table 2, these raw rates are mirrored by the results from a GEE regression analysis based on the model:

$$\begin{aligned}\text{logit}(\mu_{ki}) = \beta_0 &+ \beta_1 \text{Female}_{ki} + \beta_2 \text{Age}_{ki} + \beta_3 \text{WHO4}_{ki} \\ &+ \beta_4 \text{Year2006}_{1,ki} + \beta_5 \text{Year2007}_{2,ki} + \beta_6 \text{Private}_k\end{aligned} \quad (2)$$

where "Female" = 0/1 = male/female, "Age" is age at registration standardized so that β_0 corresponds to a 45 year old and β_2 corresponds to a 10-year contrast, "WHO4" is an indicator of WHO clinical stage with 0/1 = stage3/4, "Year2006," and "Year2007" are indicators of whether the patient registered in 2006 or 2007, with 2005 as the referent, and "Private" is a clinic-specific binary indicator of whether the clinic was privately funded or publicly funded.

Outcome-Dependent Sampling

In standard settings, the theory that justifies estimation and inference from GEE relies on the observed K clusters being a random sample from some (possibly hypothetical) population of clusters and the N_k study units a random sample from the (possibly hypothetical) population of such study units in the cluster. Here, we consider estimation and inference for the parameters of model (1) when the data arise as the result of one of two outcome-dependent sampling schemes.

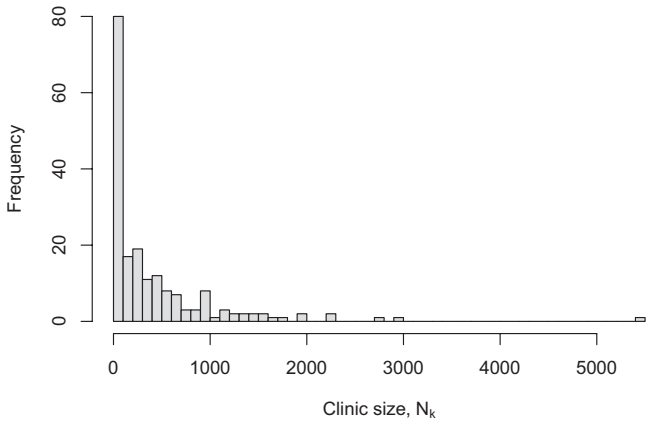

FIGURE 1. Distribution of clinic sizes, N_k, across the $K = 189$ clinics represented by $N = 78,155$ patients ages 18–65 years who registered between 2005 and 2007 with a World Health Organization (WHO) HIV/AIDS clinical stage of 3 or 4.

TABLE 2. Results From a Complete Data GEE Analysis of Model (2) Based on the N = 78,155 Registrants, Using a Working Exchangeable Correlation Structure

	Log Odds Ratio, β			Odds Ratio, Exp(β)		
		SE			95% CI	
	Est	Naive	Robust	Est	Naive	Robust
Sex						
Male	REF			REF		
Female	−0.296	0.020	0.034	0.74	(0.71, 0.77)	(0.70, 0.79)
Age, years						
10-year increment	−0.085	0.010	0.013	0.92	(0.90, 0.94)	(0.90, 0.94)
WHO stage[a]						
3	REF			REF		
4	0.658	0.022	0.032	1.93	(1.85, 2.02)	(1.81, 2.06)
Year of registration						
2005	REF			REF		
2006	−0.080	0.027	0.041	0.92	(0.88, 0.97)	(0.85, 1.00)
2007	−0.232	0.027	0.074	0.79	(0.75, 0.84)	(0.69, 0.92)
Clinic type						
Public	REF			REF		
Private	−0.500	0.110	0.130	0.61	(0.49, 0.75)	(0.47, 0.78)

[a]World Health Organization HIV/AIDS clinical stage.
CI indicates confidence interval, REF, reference category; SE, standard error; WHO, World Health Organization; Est, estimate; Exp, exponential.

The Standard Case–Control Design

In the standard case–control design, one can think of the data as arising from an initial stratification of the population by the binary outcome into cases and noncases (either implicitly or explicitly). In the cluster-correlated setting, there are two possible scenarios: (1) a case–control study was conducted, and study units were post-hoc identified as naturally clustered, and (2) the study units were a priori known to be naturally clustered and yet a case–control design ignoring cluster membership was employed.

Under either of these scenarios, let $N = N_1 + \ldots + N_K$ be the total number of study units across the K clusters. Furthermore, let N_{k1} and N_{k0} denote the total number of cases and noncases in the kth cluster, and N_{+1} and N_{+0} the corresponding totals. Under case–control sampling, a random subsample of n_1 cases is drawn from the N_{+1} observable cases and a random subsample of n_0 noncases is drawn from the N_{+0} observable noncases. Detailed covariate information is then retrospectively ascertained for each of the $n = n_0 + n_1$ individual in the two subsamples.[33]

The Case–Control Design for Clustered Data

Although appealing in the sense that it is well known, the application of a case–control design in the Malawian context may be challenging from a practical perspective. In particular, once the subsamples of cases and noncases were identified, representatives from the Ministry of Health would need to communicate with each clinic from which patients were selected so that the individual records could be extracted. Although feasible, it would be far simpler to ask each clinic to identify and extract the records for a prespecified number of cases and noncases. That is, a logistically simpler model would be to perform case–control sampling within each clinic. Notationally, this would require stratifying the N_k patients in the kth clinic into two outcome groups: N_{1k} cases and $N_{0k} = N_k - N_{1k}$ noncases. Given this stratification, $n_{1k} \leq N_{1k}$ cases and $n_{0k} \leq N_{0k}$ noncases are drawn at random and detailed covariate information retrospectively ascertained.

Estimation and Inference via Weighted GEE

As in the complete data setting, valid estimation and inference based on data obtained from a case–control design when the underlying population of study units exhibits clustering must account for potential correlation. As a general analysis strategy that ensures valid estimation in either of these settings, we consider inverse probability–weighted GEE.[39,40] Towards this, suppose that n_k of the n selected study units are identified as members of the kth cluster. Note, under the proposed design where case–control sampling is clinic-specific, n_k will be fixed and known to the research team a priori, whereas it will be random under the standard case–control design. Given these, each study unit that is selected by the outcome-dependent sampling scheme can be assigned a weight: under standard case–control sampling, the weight will be N_{+1}/n_1 if the study unit is a case and N_{+0}/n_0 if they are a noncase; under the cluster-specific case–control sampling, these values will be N_{k1}/n_{k1} and N_{k0}/n_{k0}, respectively. Technical details regarding estimation of $\boldsymbol{\beta}$ in model (1) under general working correlation structures, along with estimation of

robust standard errors, are provided in eAppendix 2 (http://links.lww.com/EDE/B278).

Simulation

To illustrate the proposed designs and inverse probability–weighted GEE analysis, we conducted a series of simulation studies. One set of simulation studies is presented in eAppendix 3 (http://links.lww.com/EDE/B278). Although not discussed in detail here, those results serve to illustrate that, in a broad range of data scenarios: (1) while point estimates obtained from a standard logistic regression analysis (i.e., one that ignore correlation attributable to clustering) are unbiased, inference is generally invalid (eTable 3; http://links.lww.com/EDE/B278) and (2) the proposed methods appropriately account for correlation to ensure that inference is valid (eTables 4 and 5; http://links.lww.com/EDE/B278).

Here, we present a second set of simulations based on the observed data from Malawi that focuses on the relative merits of unstratified sampling versus sampling within each cluster.

Set-up and Analyses

Mimicking the Malawi data summarized in Table 1 and Figure 1, we generated $R = 10{,}000$ simulated datasets each of size N = 78,155 and with the same covariate distribution as the original survey data. To induce correlation among the simulated outcomes based on model (2) for a given cluster, we used the marginalized model framework of Heagerty,[41] Heagerty and Zeger,[42] and Schildcrout and Heagerty.[43] Briefly, the strategy couples the marginal model given by expression (2) with a separately specified random effect. This induces a corresponding conditional mean model, which is then used to simulate the outcomes. As with standard random effects models, we generated the V_k as random draws from a Normal(0, σ_v^2) distribution. In the simulations presented here, we set $\sigma_v = 0.53$, the observed value from a generalized linear mixed model fit of model (2) to the Malawi data.

For each of the complete simulated datasets, we applied four balanced study designs: (1) simple random sampling; (2) random sampling within each clinic; (3) standard case–control sampling; and (4) case–control sampling within each clinic. Toward this, we first selected subsamples so that, on average, 20 patients would be selected from each clinic or, in the case of the designs that sample within clinics, all patients would be selected if N_k was <20. Under simple random sampling, therefore, 189*20 were selected at random; under random sampling within clinics, the minimum of 20 or N_k patients were selected at random from the kth clinic; under case–control sampling, 189*10 cases and 189*10 noncases were selected; finally, under case–control sampling within clinics, the minimum of 10 or N_{1k} cases were selected from the kth clinic together with the minimum of 10 or N_{0k} noncases. This process was repeated so that the cluster-specific averages were 40, 60, and 80, to give a total of four sample size scenarios.

In addition to the four balanced designs, we also considered two unbalanced case–control designs for correlated data. The rationale for doing so is that inverse probability weighting is well known to be inefficient when the weights are large. In the Malawian context, if balanced sampling is conducted (i.e. the same number of subsamples are drawn from each clinic or approximately so), patients from small clinics will have relatively small weights, whereas patients from large clinics will have potentially very large weights. One approach to mitigating this phenomenon is to oversample (relative to a balanced design) from the large clinics so that the resulting weights are less variable. Although the most effective way to reduce variability across the weights is to use simple random sampling (so that all patients will have the same weight) or case–control sampling (so that all cases have the same weight and all controls have the same weight), we sought balance reducing this variability while ensuring at least some cases and noncases were sampled from each clinic. Towards this, we initially stratified the $K = 189$ clinics by size: clinic with $N_k \leq 50$ were labeled as "small"; clinics with N_k between 51 and 500, inclusive, were labeled as "medium"; and clinics with $N_k > 500$ were labeled as "large." Based on this, the first unbalanced design selected subsamples using the same strategy as in the above balanced designs but with cluster-specific average totals of 10, 20, and 34 for the small, medium, and large clinics. These values were chosen so that the overall size of the subsample (i.e., n) was the same as when 20 was used across all clinics. To mimic the other three sample size scenarios, we scaled the cluster-specific averages for the small, medium, and large clinics accordingly. For the second of the unbalanced designs, the cluster-specific averages were set to 10, 16, and 40 for the small, medium, and large clinics for the sample size scenario when 20 was used across all clinics. Again, we scaled these accordingly to mimic the other three scenarios.

Finally, for each of the six designs, estimates of the parameters in model (2) and robust sandwich-based standard error estimates were obtained based on adopting a working independence correlation structure as well as a working exchangeable structure.

RESULTS

Table 3 and Figure 2 report select results from the simulation. Focusing on the scenarios where the average cluster-specific sample size is 40, Table 3 confirms that case–control sampling is generally superior to random sampling especially if sampling is performed within each clinic and that adopting a working exchangeable correlation structure improved efficiency (i.e., results in smaller standard errors) over adopting a working independence structure. Interestingly, comparing the third and fourth columns of Table 3, we find substantial trade-off in efficiency across the covariates between the standard case–control design and the balanced case–control design for correlated data. Specifically, although the efficiency for β_6, the coefficient associated with private/public clinic status, is improved by 41%, the efficiency for each of the other slope parameters is reduced with standard errors between 1.5 and

TABLE 3. Relative Efficiency, Defined as the Ratio of the Standard Error Under a Given Design/Working Correlation Structure to the Standard Error Under a Standard (Balanced) Case–Control Design, for the Simulation Setting Where the Cluster-Specific Average Case–Control Sample Size Was 40

	Random		Case–Control			
				Sampling Within Cluster		
	Simple	Sampling Within Cluster	Standard	Balanced	Unbalanced 1	Unbalanced 2
Working independence						
Intercept	118	356	100	167	132	123
Sex	150	694	100	339	205	193
Age	153	726	100	358	220	193
WHO stage 4[a]	138	489	100	283	182	167
Year 2006	136	664	100	318	206	181
Year 2007	130	672	100	263	179	160
Clinic type	148	125	100	59	58	58
Working exchangeable						
Intercept	75	331	55	156	109	96
Sex	135	657	86	330	193	181
Age	143	703	89	351	212	186
WHO stage 4[a]	110	468	71	272	162	143
Year 2006	112	598	75	317	194	168
Year 2007	98	535	64	265	163	140
Clinic type	137	121	87	57	54	54

[a]World Health Organization HIV/AIDS clinical stage.
WHO, World Health Organization.

3.5 times higher. These results are not surprising given that private/public clinic status is a clinic-level covariate (and hence exhibits no within-cluster variation), whereas each of the patient-specific covariates exhibit little between- versus within-cluster variation (eTable 1; http://links.lww.com/EDE/B278). We note, however, that some of the trade-off is reduced when an unbalanced design is adopted. From the fifth and sixth columns of Table 3, for example, although the efficiency gain for β_6 is retained, the increase in standard error for the other slope parameters is substantially reduced (to between 1.5 and 2.0 times higher), although clearly not eliminated.

From Figure 2, the patterns observed in Table 3 regarding relative efficiency are also observed when one considers statistical power as a function of subsample size. Specifically, across each of the five patient-specific covariates, the standard case–control design yields the highest power. If case–control sampling within clinics is adopted, however, statistical power for these covariates is dramatically improved by using an unbalanced design. Finally, each of three case–control design that sample within clinic has more than 90% power to detect the effect of private clinic status even when the cluster-specific average total sample sizes is as low as 20.

DISCUSSION

Owing to its practical utility, the case–control study remains a widely used design with >100 case–control studies having been published in each of *The Lancet* and *JAMA* in the last 5 years. Although it seems intuitive that the study units would have been naturally clustered in at least some of these studies, to the best our knowledge, no methods have been published that permit researchers to account for potential correlation in a standard case–control study. As such, although point estimates from standard analyses remain unbiased, correlation attributable to clustering may represent an unaccounted-for threat to the validity of conclusions drawn from case–control studies in the published literature, especially for covariates that exhibit large between-cluster variation. In this article, we have sought to resolve this by proposing an inverse probability–weighted GEE estimator that is intuitive and readily programmable, along with a robust sandwich variance estimator that ensures valid estimation/inference regardless of the choice of the working correlation structure.

In addition, as a logistically appealing alternative in resource-limited settings, we have proposed a novel design in which one performs case–control sampling within clusters. To the best of our knowledge, this design has not been previously described, although it does have a clear connection with the matched case–control design. Typically, analyses for the latter are conducted using conditional logistic regression.[3] Such an analysis could, in principle, be applied to the proposed design, with cluster membership being the "matching factor." In doing so, the resulting estimates must be interpreted conditionally with respect to cluster membership (as in generalized linear mixed models), so that the conclusions cannot correspond to

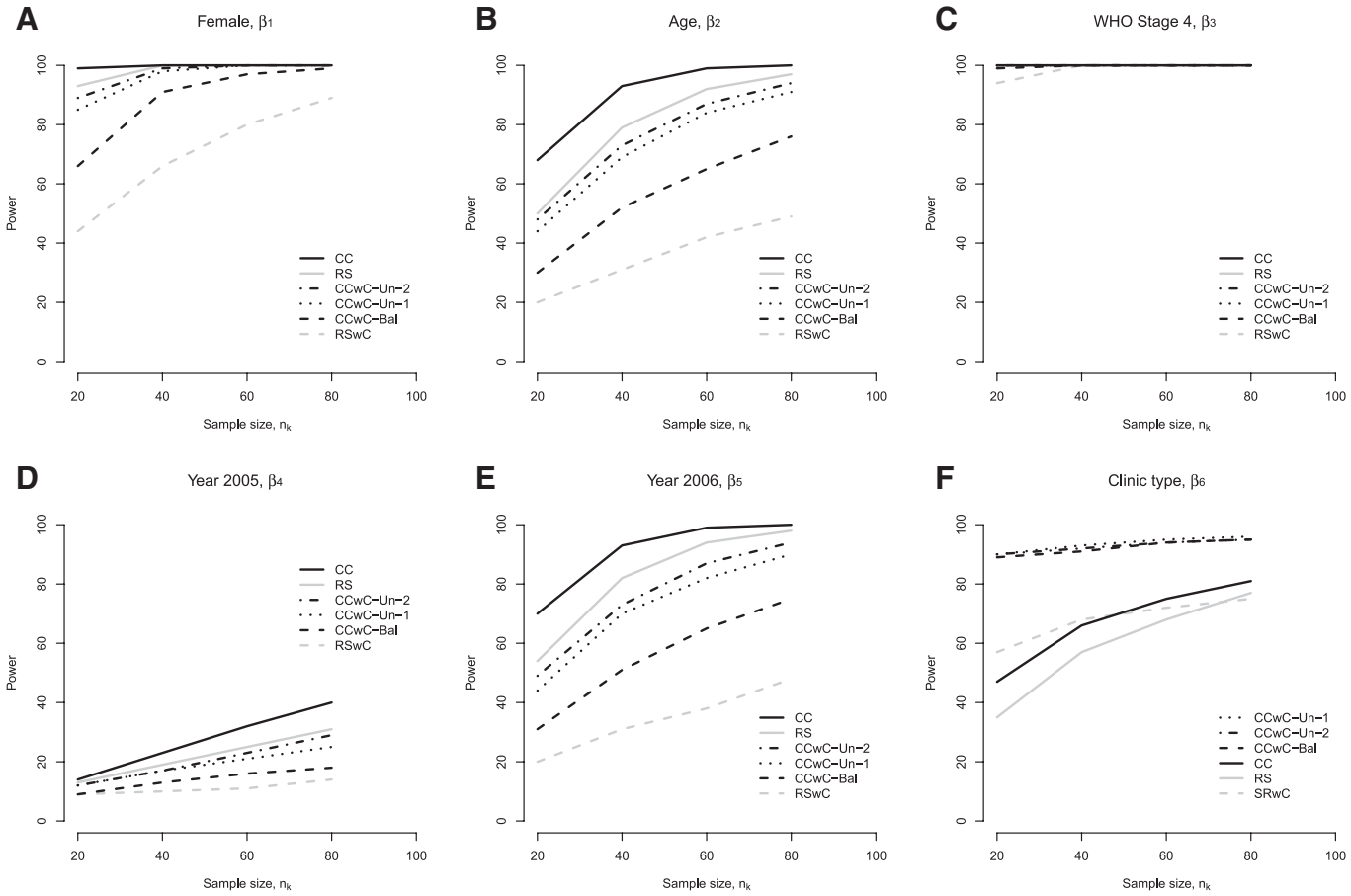

FIGURE 2. Estimated statistical power as a function of cluster-specific case–control sample size for the six slope parameters in model (2) across six designs: random sampling (RS); random sampling within clinic (RSwC); standard case–control sampling (CC); balanced case–control sampling within clinic (CCwC-Bal); and two unbalanced case–control sampling within clinic designs (CCwC-Un-1 and CCwC-Un-2).

those that would have been drawn on the basis of a complete data GEE analysis. In settings where an analyst would have fit a conditional generalized linear mixed model had complete data been available, conditional logistic regression could be applied, although effects for any covariate that varies purely by cluster (e.g., public/private clinic status in the Malawi data) could not be estimated. As part of our ongoing work, therefore, we are developing methods for estimation/inference for a generalized linear mixed model based on case–control and cluster-stratified case–control data.

As our simulations highlight, researchers considering sampling within clinics must contend with an efficiency trade-off: cluster-specific covariates and patient-specific covariates that exhibit large between- versus within-cluster variation will benefit from stratification, whereas patient-specific covariates that exhibit little between- versus within-cluster variation will suffer. In settings where the cluster sizes vary, however, adopting an unbalanced design that seeks to reduce variability in the weights while simultaneously ensuring at least some cases are selected from each cluster may mitigate this trade-off. How to optimize the choice of the unbalanced design, however, is an open avenue for research and one that we are actively pursuing.

Finally, beyond the optimal allocation of resources, there are a number of other open avenues for research. First, although inverse probability weighting techniques are generally known to be inefficient, it may be possible to adapt the inverse probability–weighted GEE to create more efficient estimators in the outcome-dependent sampling setting, along the lines of Robins and others.[44] Second, in some settings, the cluster-specific N_{0k} and N_{1k} totals may not be known or be subject to measurement error. In these settings, the weights used in the analysis to account for the design may be incorrect possibly leading to bias. Sensitivity analyses and/or the collection of validation data may provide a way forward. Finally, Haneuse and others[45] recently considered the two-phase study design as a framework for making use of the aggregated data currently collected in Malawi when selecting subsamples of patients on whom to collect detailed covariate data. Existing methods for the analysis of data from two-phase designs,

however, do not facilitate explicit accounting of potential correlation.[5] New methods are, therefore, needed if these useful designs are to be expanded to the cluster-data setting.

REFERENCES

1. Anderson JA. Separate sample logistic discrimination. *Biometrika*. 1972;59:19–35.
2. Prentice RL, Pyke R. Logistic disease incidence models and case–control studies. *Biometrika*. 1979;66:403–411.
3. Breslow NE, Day NE. *Statistical Methods in Cancer Research Vol. 1: The Analysis of Case–Control Studies*. Lyon: I.A.R.C.; 1980.
4. Scott AJ, Wild CJ. Fitting regression models to case–control data by maximum likelihood. *Biometrika*. 1997;84:57–71.
5. Breslow NE, Chatterjee N. Design and analysis of two-phase studies with binary outcome applied to Wilms tumour prognosis. *J R Stat Soc Ser C Appl Stat*. 1999;48:457–468.
6. Maclure M. The case-crossover design: a method for studying transient effects on the risk of acute events. *Am J Epidemiol*. 1991;133:144–153.
7. Wacholder S. Practical considerations in choosing between the case-cohort and nested case-control designs. *Epidemiology*. 1991;2:155–158.
8. Langholz B, Borgan O. Counter-matching: a stratified nested case–control sampling method. *Biometrika*. 1995;82:69–79.
9. Prentice R. A case–cohort design for epidemiologic cohort studies and disease prevention trials. *Biometrika*. 1986;73:1–11.
10. Barlow WE, Ichikawa L, Rosner D, Izumi S. Analysis of case-cohort designs. *J Clin Epidemiol*. 1999;52:1165–1172.
11. Zhou H, Chen J, Rissanen TH, et al. Outcome-dependent sampling: an efficient sampling and inference procedure for studies with a continuous outcome. *Epidemiology*. 2007;18:461–468.
12. Schildcrout JS, Heagerty PJ. On outcome-dependent sampling designs for longitudinal binary response data with time-varying covariates. *Biostatistics*. 2008;9:735–749.
13. Schildcrout JS, Mumford SL, Chen Z, Heagerty PJ, Rathouz PJ. Outcome-dependent sampling for longitudinal binary response data based on a time-varying auxiliary variable. *Stat Med*. 2012;31:2441–2456.
14. Schildcrout JS, Garbett SP, Heagerty PJ. Outcome vector dependent sampling with longitudinal continuous response data: stratified sampling based on summary statistics. *Biometrics*. 2013;69:405–416.
15. Neuhaus JM, Jewell NP. The effect of retrospective sampling on binary regression models for clustered data. *Biometrics*. 1990;46:977–990.
16. Neuhaus J. The analysis of retrospective family studies. *Biometrika*. 2002;89:23–37.
17. Neuhaus JM, Scott AJ, Wild CJ. Family-specific approaches to the analysis of case-control family data. *Biometrics*. 2006;62:488–494.
18. Cai J, Qaqish B, Zhou H. Marginal analysis for cluster-based case–control studies. *Sankhya Ser B*. 2001:326–337.
19. Diggle P, Heagerty P, Liang K-Y, Zeger S. *Analysis of Longitudinal Data*. Oxford: University Press; 2002.
20. Harries AD, Gomani P, Teck R, et al. Monitoring the response to antiretroviral therapy in resource-poor settings: the Malawi model. *Trans R Soc Trop Med Hyg*. 2004;98:695–701.
21. Haneuse SJ, Wakefield JC. Hierarchical models for combining ecological and case-control data. *Biometrics*. 2007;63:128–136.
22. Haneuse SJ, Wakefield JC. The combination of ecological and case-control data. *J R Stat Soc Series B Stat Methodol*. 2008;70:73–93.
23. Smoot E, Haneuse S. On the analysis of hybrid designs that combine group- and individual-level data. *Biometrics*. 2015;71:227–236.
24. Gilks CF, Crowley S, Ekpini R, et al. The WHO public-health approach to antiretroviral treatment against HIV in resource-limited settings. *Lancet*. 2006;368:505–510.
25. Harries AD, Nyangulu DS, Hargreaves NJ, Kaluwa O, Salaniponi FM. Preventing antiretroviral anarchy in sub-Saharan Africa. *Lancet*. 2001;358:410–414.
26. *Malawi: Summary Country Profile for HIV/AIDS Treatment Scale-Up*. World Health Organization. 2005. Available at: http://www.who.int/hiv/HIVCP_MWI.pdf. Accessed 23 November 2015.
27. Bemelmans M, van den Akker T, Ford N, et al. Providing universal access to antiretroviral therapy in Thyolo, Malawi through task shifting and decentralization of HIV/AIDS care. *Trop Med Int Health*. 2010;15:1413–1420.
28. Morgenstern H. Ecologic studies. *Mod Epidemiol*. 1998;2:459–480.
29. Greenland S, Morgenstern H. Ecological bias, confounding, and effect modification. *Int J Epidemiol*. 1989;18:269–274.
30. Wakefield J. Ecologic studies revisited. *Annu Rev Public Health*. 2008;29:75–90.
31. Haneuse S, Bartell S. Designs for the combination of group- and individual-level data. *Epidemiology*. 2011;22:382–389.
32. Douglas GP, Gadabu OJ, Joukes S, et al. Using touchscreen electronic medical record systems to support and monitor national scale-up of antiretroviral therapy in Malawi. *PLoS Medicine*. 2010;7:e1000319.
33. Breslow NE. Statistics in epidemiology: the case-control study. *J Am Stat Assoc*. 1996;91:14–28.
34. Breslow NE, Clayton DG. Approximate inference in generalized linear mixed models. *J Am Stat Assoc*. 1993;88:9–25.
35. Liang K-Y, Zeger S. Longitudinal data analysis using generalized linear models. *Biometrika*. 1986;73:13–22.
36. Heagerty PJ, Kurland BF. Misspecified maximum likelihood estimates and generalised linear mixed models. *Biometrika*. 2001;88:973–985.
37. Pepe MS, Anderson GL. A cautionary note on inference for marginal regression models with longitudinal data and general correlated response data. *Commun Stat Simul Comput*. 1994;23:939–951.
38. Interim WHO Clinical Staging of HIV/AIDS and HIV/AIDS Case Definitions for Surveillance. World Health Organization. 2005. http://www.who.int/hiv/pub/guidelines/clinicalstaging.pdf. Accessed 23 November 2015.
39. Robins JM, Rotnitzky A, Zhao LP. Estimation of regression coefficients when some regressors are not always observed. *J Am Stat Assoc*. 1994;89:846–866.
40. Robins JM, Rotnitzky A, Zhao LP. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *J Am Stat Assoc*. 1995;90:106–121.
41. Heagerty PJ. Marginally specified logistic-normal models for longitudinal binary data. *Biometrics*. 1999;55:688–698.
42. Heagerty PJ, Zeger SL. Marginalized multilevel models and likelihood inference. *Stat Sci*. 2000;15:1–26.
43. Schildcrout JS, Heagerty PJ. Marginalized models for moderate to long series of longitudinal binary response data. *Biometrics*. 2007;63:322–331.
44. Robins JM, Rotnitzky A. Semiparametric efficiency in multivariate regression. *J Am Stat Assoc*. 1995;90:122–129.
45. Haneuse S, Hedt-Gauthier B, Chimbwandira F, Makombe S, Tenthani L, Jahn A. Strategies for monitoring and evaluation of resource-limited national antiretroviral therapy programs: the two-phase design. *BMC Med Res Methodol*. 2015;15:31.