

Nested Case-Control Studies With Multiple Measurements: Estimating Marginal Relationships

Joshua N. Sampson (✉ joshua.sampson@nih.gov)

National Cancer Institute

Paul S. Albert

National Cancer Institute

Mark P. Purdue

National Cancer Institute

Research Article

Keywords: conditional logistic regression (CLR), Nested Case-Control Studies, Multiple Measurements, Marginal Relationships

Posted Date: May 12th, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-482695/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Nested Case-Control Studies With Multiple Measurements: Estimating Marginal Relationships

Joshua N. Sampson* Paul S. Albert** Mark P. Purdue***

* Corresponding Author, Biostatistics Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, Rockville MD, USA

** Biostatistics Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, Rockville MD, USA

*** Occupational and Environmental Epidemiology Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, Rockville MD, USA

Abstract

Background: We consider the analysis of nested, matched, case-control studies that have multiple biomarker measurements per individual. We propose a simple approach for estimating the marginal relationship between a biomarker measured at a single time point and the risk of an event. We know of no other standard software package that can perform such analyses while explicitly accounting for the matching.

Results: We propose an application of conditional logistic regression (CLR) that can include all measurements and uses a robust variance estimator. We compare our approach to other methods such as performing CLR with only the first measurement, CLR with an average of all measurements, and Generalized Estimating Equations. In simulations, our approach is

significantly more powerful than CLR with one measurement or an average of all measurements, and has similar to power to GEE but correctly accounts for the matching. We then apply our approach to the CLUE cohort to show that an increased level of the immune marker sCD27 is associated with non-Hodgkin lymphoma (NHL) and, by evaluating the strength of the association as a function of time until diagnosis, that the an increased level is likely an effect of the disease as opposed to a cause of the disease. The approach can be implemented by the R function `clogitRV` available at <https://github.com/sampsonj74/clogitRV>.

Conclusion: We offered an approach and software for analyzing matched case-control studies with multiple measurements. We demonstrated that these methods are accurate, precise, and statistically powerful.

Introduction

We will propose a method to analyze data from a longitudinal, matched, nested case/control study that has multiple measurements per individual. Here, we are assuming that a study follows a cohort over time and observes if individuals have an event. Furthermore, throughout follow-up, the study is collecting biological specimens at multiple time points. At the end of the study, we select the cases (i.e. those with an event) and a set of individually matched controls for analysis. Then, for that selected case/control group, we use the stored biospecimens to measure our desired biomarker at multiple time-points. Our objective will be to evaluate the relationship between the biomarker and the risk of an event.

Our illustrative example is a study within the Campaign Against Cancer and Stroke (referred to as CLUE from its slogan, “Give us a Clue to Cancer”) (Purdue et al. 2019). CLUE has followed a cohort of US adults since 1974 and collected blood at multiple time-points. The illustrative study’s primary objective was to determine if two immune markers, sCD27 and sCD30, were associated with the risk of non-Hodgkin lymphoma (NHL). As high levels of the biomarkers could be either a cause (e.g. reflect chronic inflammation) or an effect of NHL,

the secondary aim would be to determine when a discovered association was first detectable (e.g. 2 years or 20 years prior to diagnosis). To address these objectives, the study measured sCD27 and sCD30 at two time-points in 83 matched case-control pairs.

We know there are multiple approaches for analyzing the described data (Chen et al. 2015) and discuss the various options in the Methods section. Here, for setting up the problem, we consider the overly-simplified approach of using only the first measurement from each individual. We could then perform standard conditional logistic regression to estimate the Relative Risk (RR) for a one unit increase in the biomarker level and its corresponding p-value; then we can repeat the analysis allowing the RR to vary by time to obtain a description similar to figure 1. Here, we have evaluated the association by describing the *marginal* relationship between a single measurement and the risk of disease. We acknowledge that we are not trying to formally perform causal inference (Robins 1997), but just want to illustrate the relationship over time and possibly test for an association.

In this paper, we propose a simple method to estimate the same marginal relationships, but using all available measurements. Therefore, the method is similar in spirit to Partly Conditional Survival Models (Zheng and Heagerty 2005) and General Estimating Equations (GEE) (Liang and Zeger 1986). In the Methods Section, we describe the various options for analyzing matched case/control studies with multiple measurements, introduce our approach for estimating marginal relationships, and discuss actual and simulated test data sets. In the brief Results Section, we apply the various methods to the data sets. Then, finally, in the Discussion we describe our approach in the context of other methods.

Methods

Notation

We describe the notation for a simplified study where all subjects have the the biomarker measured at J time points and there are no covariates. For each individual we have the following variables.

- 1) $Y_i(t)$: Disease Status at time t .
- 2) T_i : Time of disease or censoring
- 3) R_i : Time of the biomarker measurements ($R_i = \{R_{i1}, \dots, R_{iJ}\}$)
- 4) X_i : Biomarker levels ($X_i = \{X_{i1}, \dots, X_{iJ}\}$)

Furthermore, we let N_A be the total number of cases and $h_0(t)$ be the baseline hazard function. Finally, we assume there is 1:1 matching (i.e. one control per case).

Marginal Model (MM)

We describe our new approach. Our goal is to estimate the *marginal* relative risk (RR) relating a *single* measure of the biomarker and the risk of the disease. If the RR is constant over time, this relationship can be described by equation 1 and our goal can be restated as estimating $\beta = \log(RR)$.

$$\log(P(Y_i(t) = 1|T_i \geq t, X_{ij})) = h_0(t) + \beta X_{ij} \quad (1)$$

If we had a matched case/control study with one measurement per individual, we would perform conditional logistic regression with N_A case/control pairs. In our study, we will have $K = J \times J$ pairs of measurements per matched set (e.g. 1st measure in the case and 1st in the control, 1st in the case and 2nd in control, etc). Therefore, our suggested approach is to perform conditional logistic regression with $K \times N_A$ case/control pairs and use a robust sandwich estimator to obtain the standard errors (see Appendix A1).

Marginal Model (MM) with time-varying effects

We now describe our new approach when we allow for a time-varying effect. We allow the RR to be a function of the time between the biomarker measurement and the disease diagnosis. Instead of the constant β in equation 1, we would now consider $\beta(t - R_{ij})$ to be the function of the time until diagnosis, $t - R_{ij}$ in 2,

$$\log(P(Y_i(t) = 1|T_i \geq t, X_{ij})) = h_0(t) + \beta(t - R_{ij})X_{ij} \quad (2)$$

We easily can obtain our estimate, $\hat{\beta}(\cdot)$, of the time-varying coefficient if we assume that the function $\beta(t - R_{ij})$ has a pre-specified form. For example, letting $r_{ij} = t - R_{ij}$ be the time until diagnosis, we can assume $\beta(r_{ij}) = \beta_0 + \beta_1 r_{ij} + \beta_2 r_{ij}^2 + \beta_3 r_{ij}^3$. As an illustration, consider $\{\beta_0, \beta_1, \beta_2, \beta_3\} = \{0.16, -0.08, 0.01, -0.0001\}$; then the $\log(\text{RR})$ at a time one year after the measurement is $\beta(1) = 0.16 - 0.08 \times 1 + 0.01 \times 1^2 - 0.0001 \times 1^3 = 0.089$ while the $\log(\text{RR})$ at a time two years after the measurement is only $\beta(2) = 0.16 - 0.08 \times 2 + 0.01 \times 2^2 - 0.0001 \times 2^3 = 0.032$. Estimating the β s is straight-forward as shown in Appendix A2 and Appendix A4. Moreover, we can test the null hypothesis of no time-varying effect by testing $H_0 : \{\beta_1, \beta_2, \beta_3\} = \{0, 0, 0\}$.

Alternative Approaches

As we noted in the introduction, there are other available approaches for analyzing the described data.

First Measurement Only (FM)

We could use only the first measurement from each individual and assume the RR is constant over time. Therefore, we can perform standard conditional logistic regression. The disadvantage to this method is a clear loss in power and precision. We could also use only the first measurement and assume that the RR can vary over time. The disadvantages are,

again, loss in power and precision, especially when estimating the behavior of the biomarker near the time of diagnosis.

Average Measurement (AM)

We could average all measurements from an individual to get a single value for that person. Then, we could again perform standard conditional logistic regression assuming a constant RR. Here, we would likely lose very little power or precision, as compared with FM. Yet, it's not without problems. Most notably, there is no clear way to model a time-varying RR. However, as discussed next, even when considering a constant RR, there can still be complications when performing meta-analyses and interpreting results.

The problems can occur when it is the individual's "usual" biomarker level, X_{iu} , that is associated with the risk of disease and that any single measurement is the usual level plus some error. Therefore, a single measured level and an individual's average level are both error-prone measure of their "usual" level. We know that the coefficients (i.e. estimates of effects) from regressing an outcome on an error-prone measure are attenuated towards the null and, in this scenario, the extent of that attenuation depends on the number of measurements per individual (i.e. averaging more measurements reduces the measurement error). Therefore, it is difficult to meta-analyze results from studies with varying numbers of measurements. Moreover, even in a single study, it becomes difficult to understand the meaning of the coefficient if subjects have varying numbers of measurements. Although, in theory, both problems can be ameliorated by "Regression Calibration" (Thomas, Stram, and Dwyer 1993) where we replace the observed value (or average of observed values) by the expected value of the usual level conditioned on the observed values.

Joint Modeling (JM)

We could jointly model the change in the biomarker level over time and the association between the biomarker and the outcome (Tsiatis and Davidian 2004). For illustration, a

simple model might assume that the biomarker level changes linearly with time and that the risk of the outcome is associated with an individual’s usual level over that time period.

$$X_{ij} = \beta_{0i} + \beta_{1i}R_{ij} + \epsilon_{ij} \quad (3)$$

$$\log(P(Y_i(t) = 1|T_i \geq t, \beta_{0i}, \beta_{1i}, X_i)) = h_0(t) + \gamma\beta_{0i} \quad (4)$$

$$\beta_{0i} \sim N(\beta_0, \sigma_0^2); \beta_{1i} \sim N(\beta_1, \sigma_1^2)$$

We acknowledge that equations 3 and 4 represent one of the simpler possible models (e.g. biomarkers need not change linearly over time and the outcome can depend on the entire biomarker history). Overall, JMs are an appropriate tool for analyzing longitudinal case/control studies. However, we note three issues. First, these models do not allow us to easily describe how the RR changes over time (i.e. figure 1). Second, these models do not easily allow us to construct potentially useful clinical prediction models that relate a single measurement with the risk of the outcome. Third, and we state this is more as a comment than a limitation, estimating the coefficient from equation 3 requires appropriately weighting subjects.

GEE

We could use General Estimating Equations (GEE). Here, we have $n = 2n_A$ individuals in our study and $n \times J$ observations, $(\{X_{11}, Y_1\}, \dots, \{X_{1J}, Y_1\}), \dots, (\{X_{n1}, Y_n\}, \dots, \{X_{nJ}, Y_n\})$. If we cluster by individual, we have n clusters of J measurements. If we cluster by matched set, we have n_A clusters of $2 \times J$ measurements. We note, for the first option, all data-points within a cluster have the same value of Y , but this peculiarity does not invalidate the theory (Appendix A3) behind GEE. When we evaluate GEE, we assume the mean relationship can be described by a logistic link with a time-independent Odds Ratio (OR) and make the rare disease assumption that $RR \approx OR$.

$$\log\left(\frac{P(Y_i = 1)}{1 - P(Y_i = 1)}\right) = \beta_0(t) + \beta X_{ij}$$

We note that using a GEE here would be akin to using unconditional logistic regression and therefore we caution against this approach. However, we include this description because a GEE would be a logical choice had the individuals been a random subset of the cohort and we have only seen only a few papers showing that GEEs remain valid when all data-points in the cluster are tied to the same outcome (i.e. Y_i is repeated across J data-points).

Simulations

We first describe the scenario where the biomarker is a time-independent causal factor (e.g. environmental pollutant) for the disease. In this scenario, we have a cohort of $n = 100 \times n_A$ individuals enrolled in the study and blood is collected regularly. Each individual has a usual biomarker level, $X_{iu} \sim N(0, 1)$ and, at time j , has a measurable level of $X_{ij} \sim N(X_{iu}, \sigma_e^2)$. Note, the Intraclass Correlation Coefficient (ICC), a useful metric for describing the measurement error of the biomarker, would be $ICC = 1/(1 + \sigma_e^2)$. To generate the event time of an individual, we assume the log-risk of an event at a given time t is proportional to the usual level, $T_i = -\log(U_i)/\exp(\beta_u X_{iu})$ with U_i drawn for a uniform(0,1) distribution (i.e. the cox model holds). We then define our study population as either the first 83 or 500 cases and a set of individually matched controls. For each individual in this study population, we generate the biomarker at two randomly selected time points prior to the case diagnosis, $R_{ij} \sim \text{uniform}(0, T_i^{\text{case}})$. We estimate coverage and power of the four methods (FM, AM, MM, GEE) based on 10,000 simulated studies.

We next describe the scenario where the biomarker is affected by the disease. In this scenario, we have a cohort of $n = 100 \times n_A$ individuals enrolled in the study and blood is collected regularly. Each individual develops a disease at $T_i^1 \sim \text{Weibull}(\text{shape} = 3, \text{scale} = 100)$ years into the study, but will only be diagnosed with the disease $T_i^2 \sim \text{Weibull}(\text{shape} = 2, \text{scale} = 9)$

years after it's development ($T_i = T_i^1 + T_i^2$). In other words, the disease is initially undetectable and in a pre-clinical stage. The biomarker level only starts to increase after the subclinical disease starts; at time j , $X_{ij} \sim N(X_{iu} + \beta \times (R_{ij} - T_i^1) \times 1(R_{ij} > T_j^1), \sigma_e^2)$, $X_{iu} \sim N(0, 1)$. We then define our study population as either the first 83 or 500 cases and a set of individually matched controls. For each individual in this study population, we generate the biomarker at two randomly selected time points prior to the case diagnosis, $R_{ij} \sim \text{uniform}(0, T_i^{\text{case}})$. We estimate coverage of MM based on 10,000 simulated studies.

We also considered one modification to the scenario with the time-independent causal factor. We allowed both the biomarker and outcome to depend on a binary covariate $Z_{ij} \sim \text{bernoulli}(0.5)$ that was used for matching. In this example, $X_{ij} \sim N(0.2Z_{ij} - 0.1, 0.99)$ and $T_i = -\log(U_i) / \exp(\beta_u X_{iu} + \beta_u Z_{ij})$. However, as this modification had no effect, corresponding results were omitted.

Example

The immune biomarker study in CLUE has been previously described elsewhere (Purdue et al. 2019). We consider the 83 matched case/control pairs that had biomarker measurements at two time points (CLUE-I in 1974, CLUE-II in 1989) and the case was diagnosed after 1989. Controls were matched on ethnicity (white, other), gender, date of birth (within 1 year), date of blood sample donation, sample storage location, and number of previous specimen thaws. We performed regression using MM with NHL diagnosis (Y_i) as the dependent variable and each biomarker (X_{ij}) as the independent variable. Because of the limited number of cases, we considered a binary time variable ($r_{ij}^* = 1$ if measurement was from 1989, $r_{ij}^* = 0$ otherwise) so our model was $\log(P(Y_{ij} = 1)) \propto \beta_1(X_{ij}(1 - r_{ij}^*)) + \beta_2(X_{ij}r_{ij}^*)$. The two biomarkers were soluble CD27 (sCD27) and sCD30, cleaved fragments of the lymphocyte transmembrane proteins CD27 and CD30 that serve as markers of immune activation (Vermeulen et al. 2011).

Results

Simulations

We first consider testing the relationship between a causal biomarker and disease. In four illustrative examples, we compare four methods: FM (using the first measure), AM (using the average measure), GEE, and MM (our proposed marginal method). We note that these methods provide an estimate $\hat{\beta} = \log(\hat{RR})$ relating either the single measure or the average of two measures with the risk of disease and its standard error $\hat{\sigma}_\beta$. Because of measurement error, we need to adjust the model estimates by dividing by the attenuation factor λ (Thomas, Stram, and Dwyer 1993) (e.g. FM, MM, GEE: $\lambda = 1/(\sigma_e^2 + 1)$; AM: $\lambda = 1/(\sigma_e^2/2 + 1)$) to get an unbiased estimate of the true β_u : $\hat{\beta}_u = \hat{\beta}/\lambda$. In table 1, we show that after adjustment, the confidence intervals of FM, AM, GEE, and MM (i.e. $\hat{\beta}/\lambda \pm 1.96\hat{\sigma}_\beta/\lambda$) have 95% coverage and, when $\beta_u = 0$, the nominal type I error (i.e. proportion of confidence intervals that exclude 0) is around 0.025. In table 1, we further show that the power to reject the null hypothesis ($H_0 : \beta_u = 0$) is similar when using AM, GEE, and MM, and, so long as $\sigma_e > 0$, exceeds the power when using FM. We reported GEE results with individual as cluster, but results were similar when the matched set was the cluster.

We next consider testing the relation between an affected biomarker and disease. For four illustrative examples, we model the relationship assuming that $\beta(t - R_{ij}) = \sum_{k=0}^3 \beta_k(t - R_{ij})^k$ (i.e. third order polynomial spline). We show the estimated relationship from one simulation in Figure 1 (e.g. $\beta = 0.2$, $\sigma_e^2 = 1$ in the model described the Methods section).

More generally, we describe the coverage of the 95% confidence interval for 5 time points ranging from 5 to 25 years prior to diagnosis. Here, coverage is the proportion of times that the confidence band includes the true value, where the truth was estimated by a large simulation. Note, we use a simulation because it is difficult to derive the observed relationship between the RR and the time until diagnosis from the biological relationship defined in the

Methods. In table 2, we show that the nominal coverage of the confidence intervals is close to 95% in most scenarios, except far from the disease (> 20 years) when there are few subjects.

Example

Among the 83 matched pairs, the mean time between the CLUE-I and CLUE-II measurements and diagnosis are 25.7 years and 10.9 years respectively. We found a significant ($p = 0.006$) relationship between sCD27 and NHL, with a significantly stronger relationship in CLUE-II (RR = 1.8, 95% CI = 1.2 - 2.7) than CLUE-I (RR=1.1, 0.8-1.5). Importantly, this suggests an increased level of sCD27 is an effect of NHL, as opposed to suggesting that chronic inflammation is a cause of the disease. We acknowledge that had we included time as a linear term (i.e. $\log(P(Y_{ij} = 1)) \propto \beta_1 X_{ij} + \beta_2 X_{ij} r_{ij}$, $p = 0.05$) or not included any interaction term (i.e. $\log(P(Y_{ij} = 1)) \propto \beta_1 X_{ij}$, $p = 0.07$), the association would not have been significant at $\alpha = 0.05/2$. We found no statistically significant association between sCD30 and NHL regardless of the chosen model.

Discussion

We have proposed a method to analyze data from a matched, nested case control study that has multiple measurements per individual. Our proposed approach estimates the marginal relationship between a single measurement and the risk of an event by using a simple application of a robust variance estimator.

We demonstrated the method using an easily available data set. However, we note the method would remain appropriate when (i) The number of measurements and timing of measurements varies for case/control pairs (ii) There are a large number of biomarkers being evaluated (e.g. it's computationally efficient and resulting p-values provide a robust test for associations) (iii) The study is a meta-analysis of multiple cohorts, with some cohorts only having 1 measurement per person.

We have also demonstrated that GEEs continue to work when the outcomes are perfectly correlated (i.e. the J data-points for cluster i are $\{X_{i1}, Y_i\}, \dots, \{X_{iJ}, Y_i\}$; note the common Y_i in all data-points). However, we avoided the deeper discussion of whether GEEs (e.g. unconditional logistic regression) can be generally used in place of conditional logistic regression. Instead, we showed that the GEE worked in one set of simulations, but note that our simulations did not include confounders or competing events related to those confounders. Finally, we used the logistic-link (i.e. estimated OR's), as opposed to the log-link, because the relative risks estimated in a case-control study are not unbiased estimates of the relative risk in the overall population.

We note there are potential modifications and improvements for MM. First, we currently weight all of the pairs of measurements equally (i.e. a case/control pair with three measurements per person has $9\times$ the influence of a case/control pair with one measurement per person). We can therefore modify our approach by adding weights to our estimating equations or using a non-independent working correlation matrix. Second, we model the time-varying coefficient using a polynomial spine which works well when the relationship is simple (e.g. the effect is larger closer to diagnosis). However, there are other options such as Generalized Additive Models (GAM) (Wood 2017) or, as suggested for Partly Conditional Survival Models (Zheng and Heagerty 2005), a kernel function. For the latter, we would require a two dimensional kernel to accommodate case and control times. Third, we do not allow the coefficient to vary by age or time since enrollment. Fourth, we need not use all matched pairs of measurements. In the CLUE study, for each case/control set, we could have used only two matched pairs without matching samples from 1974 and 1989.

Declarations

Ethics approval and consent to participate

Not Applicable

Consent for publication

All authors consent for publication

Availability of data and materials

Computer code is available <https://github.com/sampsonj74/clogitRV>. The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

Competing interests

Authors have no competing interests

Funding

Authors are funded by the intramural program of the National Cancer Institute

Authors' contributions

JS, PA, and MP developed the method and wrote/reviewed the manuscript. JS performed the analyses.

Acknowledgements

We thank the participants of Clue I and Clue II and appreciate the continued efforts of the staff at the Johns Hopkins George W. Comstock Center for Public Health Research and Prevention in the conduct of the Clue Cohort Studies.

Cancer data was provided by the Maryland Cancer Registry, Center for Cancer Prevention and Control, Maryland Department of Health, with funding from the State of Maryland and the Maryland Cigarette Restitution Fund. The collection and availability of cancer registry

data is also supported by the Cooperative Agreement NU58DP006333, funded by the Centers for Disease Control and Prevention. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the Centers for Disease Control and Prevention or the Department of Health and Human Services.

This work utilized the computational resources of the NIH HPC Biowulf cluster. (<http://hpc.nih.gov>)

References

Chen, Y. H., K. K. Ferguson, J. D. Meeker, T. F. McElrath, and B. Mukherjee. 2015. “Statistical Methods for Modeling Repeated Measures of Maternal Environmental Exposure Biomarkers During Pregnancy in Association with Preterm Birth.” *Environ Health* 14: 9.

Liang, Kung-Yee, and Scott L. Zeger. 1986. “Longitudinal data analysis using generalized linear models.” *Biometrika* 73 (1): 13–22. <https://doi.org/10.1093/biomet/73.1.13>.

Purdue, Mark P., Qing Lan, Judith Hoffman-Bolton, Allan Hildesheim, Catherine L. Callahan, Paul Strickland, Kala Visvanathan, and Nathaniel Rothman. 2019. “Circulating sCD27 and sCD30 in Pre-Diagnostic Samples Collected Fifteen Years Apart and Future Non-Hodgkin Lymphoma Risk.” *International Journal of Cancer* 144 (8): 1780–5.

Robins, James M. 1997. “Causal Inference from Complex Longitudinal Data.” In *Latent Variable Modeling and Applications to Causality*, edited by Maia Berkane, 69–117. New York, NY: Springer New York.

Thomas, D, D Stram, and J Dwyer. 1993. “Exposure Measurement Error: Influence on Exposure-Disease Relationships and Methods of Correction.” *Annual Review of Public Health* 14 (1): 69–93.

Tsiatis, Anastasios A., and Marie Davidian. 2004. “Joint Modeling of Longitudinal and Time-to-Event Data: An Overview.” *Statistica Sinica* 14 (3). Institute of Statistical Science,

Academia Sinica: 809–34.

Vermeulen, Roel, Fatemeh Saberi Hosnijeh, Lützen Portengen, Vittorio Krogh, Domenico Palli, Salvatore Panico, Rosario Tumino, et al. 2011. “Circulating Soluble Cd30 and Future Risk of Lymphoma; Evidence from Two Prospective Studies in the General Population.” *Cancer Epidemiology and Prevention Biomarkers* 20 (9). American Association for Cancer Research: 1925–7.

Wood, S.N. 2017. *Generalized Additive Models: An Introduction with R*. 2nd ed. Chapman; Hall/CRC.

Zheng, Yingye, and Patrick J. Heagerty. 2005. “Partly Conditional Survival Models for Longitudinal Data.” *Biometrics* 61 (2): 379–91.

Tables and Figures

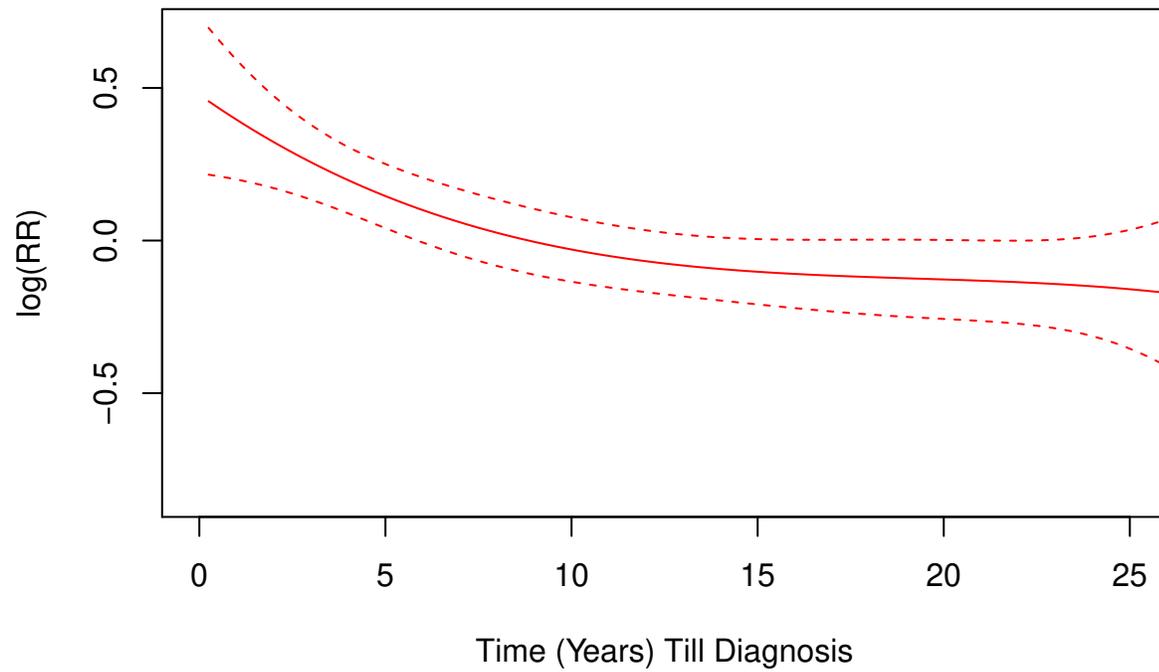


Figure 1: Effect size as a function of the time between measurement and diagnosis (simulated data)

Table 1: Coverage and Power as a function of n , β_U and ICC for the four methods.

N	β_U	ICC	Coverage				Power			
			FM	AM	MM	GEE	FM	AM	MM	GEE
83	0.0	1.0	0.96	0.96	0.96	0.95	0.021	0.021	0.023	0.025
83	0.0	0.5	0.96	0.95	0.95	0.95	0.023	0.027	0.028	0.028
83	0.2	1.0	0.95	0.95	0.95	0.95	0.231	0.234	0.238	0.247
83	0.2	0.5	0.95	0.96	0.95	0.95	0.126	0.162	0.173	0.175
500	0.0	1.0	0.95	0.95	0.95	0.95	0.026	0.025	0.026	0.026
500	0.0	0.5	0.95	0.95	0.95	0.95	0.026	0.023	0.023	0.023
500	0.2	1.0	0.95	0.95	0.95	0.95	0.881	0.881	0.880	0.882
500	0.2	0.5	0.95	0.95	0.95	0.95	0.601	0.730	0.732	0.735

Table 2: Coverage as a function of n , β , ICC, and time till diagnosis

n	β	ICC	Years					
			1	5	10	15	20	25
83	0.0	1.0	0.94	0.95	0.96	0.94	0.94	0.91
83	0.0	0.5	0.95	0.95	0.96	0.94	0.95	0.92
83	0.1	1.0	0.94	0.93	0.94	0.93	0.93	0.90
83	0.2	0.5	0.94	0.95	0.95	0.95	0.94	0.93
500	0.0	1.0	0.95	0.95	0.96	0.95	0.95	0.96
500	0.0	0.5	0.94	0.95	0.94	0.94	0.95	0.95
500	0.1	1.0	0.94	0.95	0.96	0.95	0.94	0.95
500	0.2	0.5	0.93	0.95	0.94	0.94	0.94	0.95

Figures

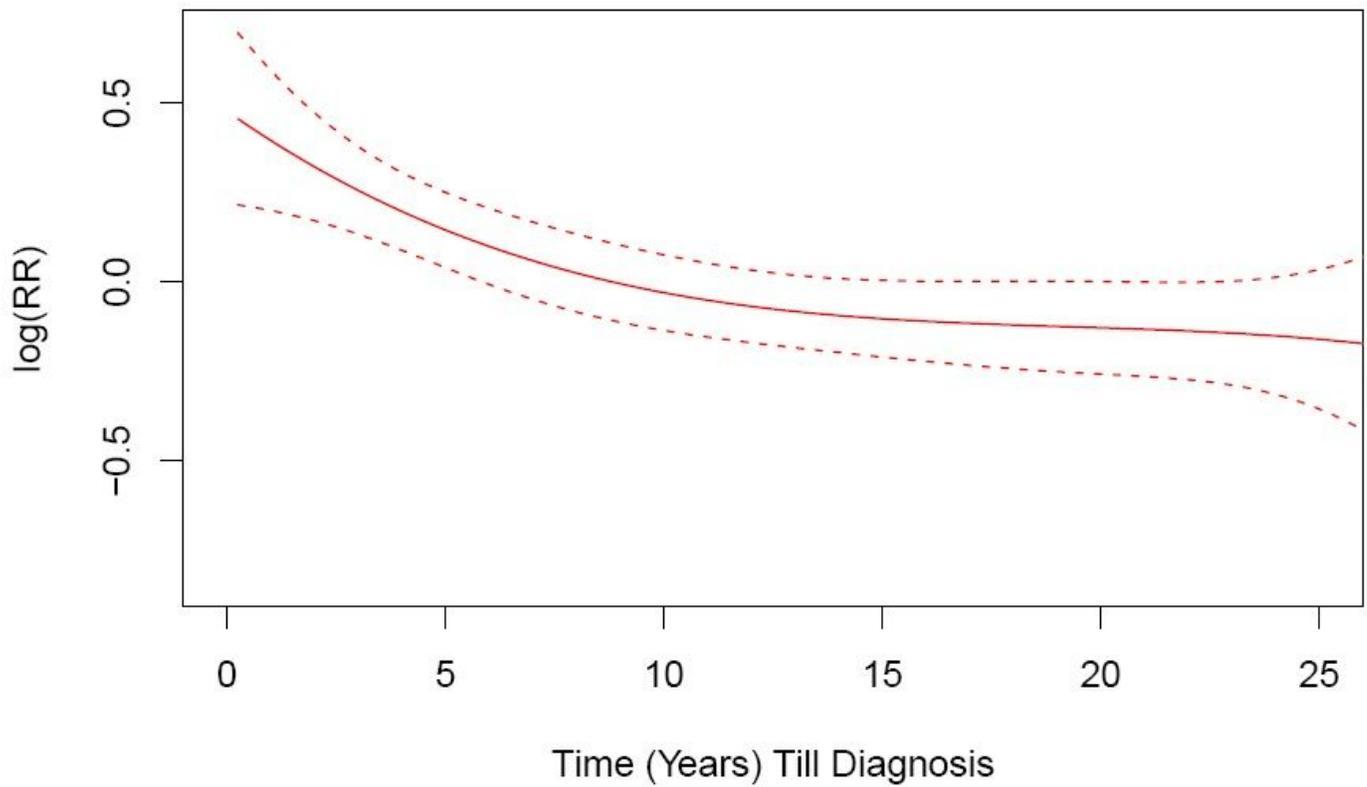


Figure 1

Effect size as a function of the time between measurement and diagnosis (simulated data)

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [GEE8appendix.pdf](#)