Scientific Research Publishing

Open Journal of Statistics, 2017, 7, 240-253
http://www.scirp.org/journal/ojs
ISSN Online: 2161-7198
ISSN Print: 2161-718X

Estimation of Attributable Risk from Clustered Binary Data: The Case of Cross-Sectional and Cohort Studies

Mohamed Shoukri[1*], Allan Donner[2,3], Futwan Al-Mohanna[1]

[1]Department of Cell Biology, Research Center King Faisal Specialist Hospital & Research Center, Riyadh, KSA
[2]Department of Epidemiology and Biostatistics, Schulich School of Medicine and Dentistry, University of Western Ontario, London, Ontario, Canada
[3]Robarts Clinical Trials, Robarts Research Institute, London, Ontario, Canada
Email: *shoukri@kfshrc.edu.sa

How to cite this paper: Shoukri, M., Donner, A. and Al-Mohanna, F. (2017) Estimation of Attributable Risk from Clustered Binary Data: The Case of Cross-Sectional and Cohort Studies. *Open Journal of Statistics*, **7**, 240-253.
https://doi.org/10.4236/ojs.2017.72019

Received: March 2, 2017
Accepted: April 21, 2017
Published: April 24, 2017

Abstract

Effect sizes are estimated from several study designs when the subjects are individually sampled. When the samples are the aggregate cluster of individuals, the within cluster correlation must be accounted for to construct correct confidence intervals, and to conduct valid statistical inference. The purpose of this article is to propose and evaluate statistical procedures for the estimation of the variance of the estimated attributable risk in parallel groups of clusters, and in a design dividing each of k clusters into two segments creating multiple sub-clusters. The estimated variance is the first order approximation and is obtained by the delta method. We apply the methodology and propose a Wald type confidence interval on the difference between two correlated attributable risks. We also construct a test on the hypothesis of equality of two correlated attributable risks. We evaluate the power of the proposed test via Monte-Carlo simulations.

Keywords

Correlated Binary Responses, Effect Size, Split-Cluster Design, Correlated Attributable Risks, Confidence Intervals, Monte-Carlo Simulations

1. Introduction

In the epidemiological research, it is important that the collected data are translated into interpretable results which can be easily communicated to clinicians. The need for "translatable" evidence from research studies is of prime impor-

tance in the evaluation of clinical interventions, because they hold the potential to immediately influence the course of patient treatment. When evaluating these studies, the examination of "Effect Size" or (ES) can be a useful measure of the comparative efficacy of the treatment under investigation. In randomized clinical trials, an effect size estimate quantifies the direction and magnitude of an effect of an intervention.

When exposure and disease risk are measured on a binary scale, several measures of effect size are in current use [1]. The odds ratio (*OR*), the relative risk (*RR*), and the population attributable risk (*AR*) are the most commonly used measures of effect size in clinical as well as analytic epidemiology.

The concept of *AR* was introduced in [2] and is a widely used measure of the amount of disease that can be attributed to a specific risk factor. The *AR* combines the relative risk (*RR*) and the prevalence of exposure $P(E)$ to measure the public health burden of a risk factor by estimating the proportion of cases of a disease that would not have occurred if we remove the risk factor.

The concept of *AR* and its statistical characteristics have been reviewed in [3] and in several publications [4] [5]. Statistical inferences on *AR* require the availability of data from subjects randomly assigned to intervention groups. However, when the sampling strategy involves aggregate or clusters of individuals, adjusting for the effect of intracluster correlation is essential in order to conduct valid statistical inferences [6] and the references therein. However, the statistical properties of estimators of *AR* when clusters are sampled have not yet been fully explored. The fundamental objective of our work is to fill the gap of performing statistical inference on *AR* under the clustered binary data situation.

In this paper, we obtain the variance of the estimated *AR* under cluster sampling, focusing on cohort and cross-sectional designs. In Section 2, we construct an *AR* estimator, and in Section 3, we derive its large sample variance adjusted for the intracluster correlation (ICC). In Section 4, we consider the split cluster design, and describe situations where we compare two correlated *AR* parameters. In Section 5, we conduct a Monte-Carlo experiment to evaluate the empirical power of Wald's test on the null hypothesis of equality of two correlated attributable risk parameters. At the end of each section, we provide an example.

2. *AR* from Cluster Sampling

We start with a parallel group design where k clusters are exposed to a specified risk factor, and l clusters are not exposed, as in the data layout given in **Table 1**.

In **Table 1**, we assume that k clusters have been selected at random from a well-defined population of exposed individuals, where the j^{th} cluster has n_j units. All individuals in this sample are assumed to be exposed to the risk factor. Let $x_{ij}\left(i=1,2,\cdots k,\ j=1,2,\cdots n_i\right)$ with $x_{ij}=1$ and 0, denoting positive and negative responses corresponding to the presence of the exposure with $\pi_i = P_r\left[x_{ij}=1 \middle| \text{exposed cluster } i\right]$. Similarly we assume that l clusters have been selected from the population of unexposed individuals, where the l^{th} cluster has m_l units. All units in the clusters can serve as controls assuming the

Table 1. Typical data layout for clustered data in two groups: exposed and unexposed.

Exposed (E)				Non-Exposed $(\bar{E})$			
1	2	…	k	1	2	…	l
x_{11}	x_{21}	…	x_{k1}	y_{11}	y_{21}		y_{l1}
x_{12}	x_{22}		x_{k2}	y_{12}	y_{22}		y_{l2}
⋮	⋮		⋮	⋮	⋮		⋮
x_{1n1}	x_{2n2}	…	x_{knk}	y_{1m1}	y_{2m2}	…	y_{lml}

absence of exposure. In the unexposed clusters, let $y_{rs}\left(r=1,2,\cdots l, s=1,2,\cdots m_r\right)$ with $y_{rs}=1$ and 0 denote positive and negative responses with $Q_r=P_r\left[y_{rs}=1\middle|\text{unexposed cluster } r\right]$. Furthermore, let $X_i=\sum_{j=1}^{n_i}x_{ij}$ and $Y_r=\sum_{s=1}^{m_r}y_{ij}$ denote respectively the total number of events in the exposed and non-exposed groups; provided that the misclassification error is zero. Therefore, conditional on π_i, X_i has binomial distribution with parameters $\left(n_i,\pi_i\right)$. Similarly, conditional on Q_r, Y_r has binomial distribution with parameters $\left(m_r,Q_r\right)$. To introduce a within cluster correlation, we assume that π_i follows a beta distribution $B\left(a,b\right)$ with probability density function (pdf) given in (1).

$$f\left(\pi_i\middle|a,b\right)=\frac{\Gamma\left(a+b\right)}{\Gamma\left(a\right)\Gamma\left(b\right)}\pi_i^{a-1}\left(i-\pi_i\right)^{b-1} \tag{1}$$

and that Q_r follows a similar beta distribution where pdf is denoted $B\left(\alpha,\beta\right)$. The effect of the intracluster correlation among the responses may be accounted for as follows.

Under the transformations, $P=\frac{a}{a+b}$ and, $\rho_1=\left(1+a+b\right)^{-1}$, the mean and variance of π_i are given respectively by $E\left(\pi_i\right)=P$ and $\text{Var}\left(\pi_i\right)=P\left(1-P\right)\rho_1$.

Similarly, $=\alpha/\left(\alpha+\beta\right)$ and $\rho_2=\left(1+\alpha+\beta\right)^{-1}$. Therefore, we have $E\left(Q_i\right)=Q$, $\text{Var}\left(Q_i\right)=Q\left(1-Q\right)\rho_2$. Consequently, the unconditional distribution of x_i is beta binomial with, $E\left(x_i\right)=n_i\rho$, and $\text{Var}\left(X_i\right)=n_iP\left(1-P\right)\left[1+\left(n_i-1\right)\rho_1\right]$. Similarly; $E\left(Y_i\right)=m_iQ$, and

$$\text{Var}\left(Y_i\right)=m_iQ\left(1-Q\right)\left[1+\left(m_i-1\right)\rho_2\right].$$

It should be noted that the beta distribution assumptions imposed on the model parameters is not necessary, and one may adopt a quasi-likelihood set-up, by specifying the first two moments for π_i and Q_i as shown in [7] and [8]. This set-up, would lead to the same expressions for $\text{Var}\left(X_i\right)$ and $\text{Var}\left(Y_i\right)$. The reason for introducing the beta distribution here is that while it serves as mechanism to create within cluster correlation, it will form the basis for generating data from beta binomial distribution using Monte-Carlo simulations in Section 5.

The parameters ρ_1 and ρ_2 are respectively interpreted as the within cluster correlations among all pairs of scores in the group of exposed and unexposed. We may obtain consistent estimators of $\text{Var}\left(X_i\right)$ and $\text{Var}\left(Y_i\right)$ on replacing the parameters, Q, ρ_1, and ρ_2 with appropriate estimators from the data as

will be shown in the next sections. We shall now construct unbiased point estimators for the parameters P and Q.

From [8] and [6], we have $X=\sum_{i=1}^{k}X_i$ has $E(X)=NP$, $\mathrm{Var}(X)=NP(1-P)\left[1+(n_0-1)\rho_1\right]$. Similarly, $Y=\sum_{r=1}^{l}Y_i$ has $E(Y)=MQ$, $\mathrm{Var}(Y)=MQ(1-Q)\left[1+(m_o-1)\rho_2\right]$, where $N=\sum_{i=1}^{k}n_i$, $M=\sum_{r=1}^{l}m_i$, $n_0=\sum_{i=1}^{k}n_i^2/N$, and $m_o=\sum_{r=i}^{l}m_i^2/M$. Clearly X/N and Y/M are unbiased point estimators for P and Q respectively.

The data, under the above set up can then be summarized in a 2 × 2 table as shown in Table 2.

Formally, the AR is defined in [2] as:

$$AR=\left\{P(D)-P\left(D\middle|\bar{E}\right)\right\}\Big/P(D) \tag{2}$$

where $P(D)$ is the percentage of disease in the population, and $P\left(D\middle|\bar{E}\right)$ is the percentage of disease in the population in the absence of exposure to the risk factor. Levin [2] defines the Attributable Risk (AR) as "the amount of disease that can be attributed to a specific risk factor".

Using Bayes theorem, and from [[3]; page 73], Equation (2) may be written as:

$$AR=\frac{p(E)(RR-1)}{1+p(E)(RR-1)}. \tag{3}$$

Here; RR is the relative risk or the risk ratio, and $P(E)$ is the risk of exposure. The RR is defined by $RR=\dfrac{p\left(D\middle|E\right)}{P\left(D\middle|\bar{E}\right)}$. In terms of population parameters, the AR as defined in (3) is equivalent to:

$$AR=\frac{P(1-Q)-Q(1-P)}{P+Q}=\frac{P-Q}{P+Q}. \tag{4}$$

Under the transformation, $\Psi=\dfrac{1-AR}{1+AR}$, we get $Q=\Psi P$. We shall use this transformation to facilitate the derivation of the large sample variance of AR.

The sample estimator of AR, is obtained using the data in a 2 × 2 cross classification as given in Table 2.

Epidemiologists use this statistic quite frequently to assess the consequences of an association between a binary outcome of interest (D) and exposure to a risk factor (E). The total number of observations in the non-exposed and the exposed groups are given respectively by M and N, assumed fixed.

For a cross sectional or cohort study designs the AR estimator is from [3]

Table 2. Disease-exposure cross classification in a 2 × 2 table.

		Response (D)		
		D^+	$\bar{D}$	Total
Exposure	E^+	X	$N-X$	N
	$\bar{E}$	Y	$M-Y$	M
	Total	$X+Y$	$M+N-X-Y$	$M+N$

given by:

$$\widehat{AR} = \frac{X(M-Y) - Y(N-X)}{(X+Y)M}. \tag{5}$$

Following [9], we shall derive the asymptotic variance of $\hat{\theta} = \ln\left(1-\widehat{AR}\right)$.

We first write, $\mathrm{Var}(X) = NP(1-P)c_1$, and, $\mathrm{Var}(Y) = MQ(1-Q)c_2$, where $c_1 = 1+(n_0-1)\rho_1$, $c_2 = 1+(m_0-1)\rho_2$, $n_0 = \sum_{i=1}^{k} n_i^2 \,|N$, and $m_0 = \sum_{i=1}^{l} m_i^2 \,|M$.

Using the delta method [10] we can show to the first order of approximation that:

$$\mathrm{Var}(\hat{\theta}) = \frac{N(1-P)C_1}{P(N+M\Psi)^2} + \frac{N^2(1-P\Psi)C_2}{MP\Psi(N+M\Psi)^2}. \tag{6}$$

A consistent estimator of $\mathrm{Var}(\hat{\theta})$ may be obtained on replacing the parameters P, c_1, c_2 and Ψ by their moment estimators. An $(1-\alpha)100\%$ confidence interval on AR is thus given as:

$$\left(1-\exp\left(\hat{\theta} + z_{\alpha/2}\sqrt{\mathrm{var}}(\hat{\theta})\right), 1-\exp\left(\hat{\theta} - z_{\alpha/2}\sqrt{\mathrm{var}}(\hat{\theta})\right)\right).$$

The moment estimators of the intraclass correlations are obtained separately from the groups of exposed and unexposed clusters. The moment estimator of ρ_1 is given by:

$$\hat{\rho}_1 = \frac{MSB - MSW}{MSB + (n_0-1)MSW} \tag{7}$$

where *r*,

$$MSW = \frac{1}{n-k}\sum_{i=1}^{k}\sum_{j=1}^{n_i}\frac{x_{ij}(n_{ij}-x_{ij})}{n_{ij}} \tag{8}$$

$$MSB = \frac{1}{k-1}\sum_{i=1}^{k}\sum_{j=1}^{n_i}\frac{(x_{ij}-x_i)^2}{n_{ij}}. \tag{9}$$

Similar expressions for the (*MSW*, *MSB*) are obtained for the clusters of unexposed. The quantities (*MSW*, *MSB*) are estimated from the one-way ANOVA model when the responses are measured on the binary scale. For details the readers are referred to [6].

We now consider two examples, the first is from data arising from a cross sectional study and the second example is on data from a randomized prospective trial.

Example 1: Cross-Sectional Study: The effect of consanguinity on congenital heart defects (CHD).

The Saudi Arabian CHD registry [11] was established in 1998, and by 2003 the registry evolved into Multi-Institutional research collaboration. The prime aim of this institution is to develop a registry whereby data from major referral hospitals across the country can provide patient information.

The participating hospitals are from regions that cover the country making the registry a nationwide data repository for the Kingdom of Saudi Arabia [Congenital Heart Disease Registry 2013]. The present example uses data on a major congenital heart disease; Patent Ductus Arteriosus (PDA). The incidence of PDA has been reported to be approximately, 1 in 2000 births, which accounts for 5% to 10% of all congenital heart diseases with female to male ratio of almost 2:1 [12]. The PDA was found to occur with increased frequency in several genetic syndromes, with precise mechanisms resulting in persistent PDA not yet clear [13] [14].

Arab countries are notorious for consanguineous marriages, with first cousin types being the most common. For example in Jordan the prevalence of consanguinity was reported in [15] as 51.3%, Yemen, 40% as reported [16], and almost 57% in Saudi Arabia as reported [17] [18]. More recently, a survey of Saudi families conducted in [19], estimated the prevalence of consanguinity to be as high as 56%.

For illustrative purposes of the methodologies presented in this section, we sampled two children from the registry whose mother are non-diabetic, with maternal age less than 40 years. Each sampled child was classified according to the presence/absence of PDA, and the type of parental consanguinity (exposure variable). Therefore, for the exposed (children from consanguineous marriages restricted to first degree cousin) and non-exposed (children from non-consanguineous marriages) the cluster size is $n = m = 2$. The data are presented in Table 3.

Direct applications using Equations (5), (8), (9), and (10) we get:

$$\text{AR} = \frac{(53)(66)-(30)(99)}{(83)(96)} = 6.6\%, \quad P(E) = 0.61$$

$$\rho_1 = 0.325, \quad \rho_2 = 0.332, \quad P = P_r\left(D\middle|\text{Consanguineous}\right) = 0.39,$$

$$Q = P_r\left(D\middle|\text{Non}-\text{Consanguineous}\right) = 0.31, \text{ and } RR = \frac{0.349}{0.313} = 1.11.$$

The square root of Equation (6) gives $se\left(\hat{\theta}\right) = 0.085$, and the 95% confidence interval of AR is: $-0.104 < AR < 0.210$.

The AR estimate is interpreted as follow: if among infants born with CHD, given that PDA among infants with CHD is a preventable event, then prohibiting first degree relatives' marriages will reduce the chance of having PDA by 6%.

Example 2: Prospective Cohort study (Weil's data)

The data in this example was given first in [20] taken from [21], and gives the

Table 3. Consanguinity and PDA.

		PDA		
		Present	Absent	Total
Consanguinity	Yes	53	99	152
	No	30	66	96
	Total	83	165	248

results from an experiment comparing two treatments. One group of 16 pregnant female rats was fed a control diet during pregnancy and lactation, while the diet of a second group of 16 pregnant females was treated with a chemical. For each cluster (litter consisting of the pups born to a female rat), the number n of pups alive at 4 days and the number y of pups that survived at 31 day lactation period were recorded. The data are given as a fraction y/n in Table 4.

In Table 4, the numerator is the number of dead pups during 21 days lactation period, and the numerator is the number who survived past 4 days. The purpose of the experiment was to determine if the chemical treatment significantly affects the survival rate among the pups. That is, we need to test the null hypothesis $H_0 : P = Q$. The data are presented in a 2×2 format in Table 5.

$P = 0.24$ and $Q = 0.10$, giving relative risk $RR = 2.4$.

$$AR = \frac{(35)(142)-(16)(112)}{(51)(158)} = 39.4\%.$$

$\rho(\text{control}) = 0.029$, $\rho(\text{treated}) = 0.040$, $n_0 = 9.84$, $m_0 = 9.16$, and $se(\hat{\theta}) = 0.0555$, and the 95% confidence interval on AR is: $0.33 < AR < 0.45$.

3. Split Cluster Design

Split-cluster experiments are being used by investigators in health sciences when naturally occurring aggregates of individuals with nested subgroups may be assigned to different treatments. Cited examples include split mouth trials, in which a subject's mouth is divided into two segments that are randomly assigned to different treatment groups. In other situation, randomization to treatment conditions may be possible at the person level within the cluster. In this case, when the treatment conditions are available within each cluster, the design is referred to as a multisite or split cluster design (SCD). The major attractiveness of this design is that it removes a large portion of the inter-subject variation from the estimate of treatment effect; and hence has the potential to require a lesser number of subjects than a parallel arm design with the same power. When the response variable of interest is binary, statistical methods developed to evaluate the effect of intervention depends on non-parametric methods, as shown in [22].

In this section we present the data layout for the SCD (see Table 6) and derive the large sample variance of the AR as a measure of effect size.

Under a similar set up to that we presented in the previous section and with appropriate change in notations the random variables $X_i = \sum_{j=1}^{n_i} x_{ij}$ and $Y_i = \sum_{j=1}^{m_i} y_{ij}$ will have the same beta-binomial distributions, but they are no-longer independent.

Table 4. Weil's data: Mortality due to exposure is a two arms clinical trial.

Control	13/13	12/12	9/9	9/9	8/8	8/8	12/13	11/12
	9/10	9/10	8/9	11/13	4/5	5/7	7/10	7/10
Treatment	12/12	11/11	10/10	9/9	10/11	9/11	9/11	8/9
	8/9	4/5	7/9	4/7	5/10	3/6	3/10	0/7

Table 5. Weil's data collapsed in a 2 × 2 table.

Exposure	Status		Total
	Dead	Alive	
Treated	35	112	147
Control	16	142	158
Total	51	254	305

Table 6. Data layout for the split-cluster design.

	Clusters					
Sub-Clusters	1	2	…	j	…	k
1	x_{11}	x_{21}		x_{j1}		x_{k1}
(Exposed)	x_{12}	x_{22}		x_{j2}		x_{k2}
	⋮					
	x_{1n1}	x_{2n2}		x_{jnj}		x_{knk}
2	y_{11}	y_{21}		y_{j1}		y_{k1}
(Unexposed)	y_{12}	y_{22}		y_{j2}		y_{k2}
	⋮					
	y_{1m1}	y_{2m2}		y_{jmj}		y_{kmk}

$$\mathrm{Var}(X) = NP(1-P)\left[1+(u_1-1)\rho_1\right]$$

$$\mathrm{Var}(Y) = MQ(1-Q)\left[1+(u_2-1)\rho_2\right].$$

The correlation parameters ρ_1 and ρ_2 are estimated as shown in Equations (7)-(9).

Although the AR estimator maintains the same expression under split clusters, its variance is affected by the correlations within the sub-clusters, and between units in the exposed and the non-exposed sub-clusters.

Using the delta method, we can therefore show that

$$\mathrm{Var}(\hat{\theta}) = \frac{N(1-P)c_1}{P\left[N+M\psi\right]^2} + \frac{N^2(1-\psi P)c_2}{M\psi P\left[N+M\psi\right]^2} + \left\{\frac{-2NP\rho_{12}}{M\psi P\left[N+M\psi\right]^2}\left[NM\psi(1-P)(1-\psi P)c_1c_2\right]^{1/2}\right\} \tag{10}$$

where $c_1 = 1+(u_1-1)\rho_1$, $c_2 = 1+(u_2-1)\rho_2$, $u_1 = \sum_{i=1}^{k} n_i^2/N$ and $u_2 = \sum_{i=1}^{k} m_i^2/M$.

Here, ρ_1 is the intraclass correlation among the individuals in the sub-clusters of exposed, and ρ_2 is the intraclass correlation among the individuals in the sub-clusters of unexposed. Both correlations are estimated from the one-way ANOVA layout as explained in Equations (7)-(9). The cross-clusters correlation which is interpreted as an intercluster correlation denoted by ρ_{12} is similarly estimated from the data by first ignoring the splitting structure of the data, and

then use the one-way ANOVA to obtain the within and between mean squares. Substituting these quantities in (7) we obtain a moment estimator of ρ_{12}.

Example 3: Split-Mouth Trial

For illustrating the proposed methodology, as a third example, we consider data from a split-mouth trial on 23 patients evaluating the effect of chlorhexidine in the treatment of gingivitis [22]. The data are presented in Table 7. The chlorhexidine and control treatments were randomly applied to four sites located in the patient's left and right sides of the upper and lower jaws. We are interested here in testing the effect of treatment on the presence or absence of plaque, as based on the measurements taken two weeks after baseline and summarized in Table 8. The sample estimates and standard errors (SE) of P and Q, the proportion of patients having plaque in the chlorhexidine and control groups, are estimated at 0.89 (SE = 0.0343), and 0.77 (SE = 0.0491), respectively. The intra-class and inter-class correlation coefficients are estimated as $\hat{\rho}_1 = 0.0395$, $\hat{\rho}_2 = 0.087$, as shown in the previous section, pooled estimate $\hat{\rho} = 0.070$ The sample estimate of the relative risk RR is 1.155.

$$\rho_1 = 0.0395, \quad \rho_2 = 0.087, \quad \rho_{12} = 0.039.$$

$$AR = \frac{(82)(21)-(10)(71)}{(153)(92)} = \frac{1722-710}{14076} = 7.19\%.$$

$se(\hat{\theta}) = 0.092$, and the 95% CI on AR is (−0.112, 0.225).

4. Testing the Equality of Two Correlated *AR* Parameters

Interest is focused on studying the change in disease-exposure etiology under varying conditions. We illustrate this situation using the published data [23].

For example in the case of family data we may be interested in evaluating the effect of disease status of a parental exposure variable on their siblings, which can be divided into males and females within the same family. In this case, we

Table 7. Number of sites with plaque in four sites $(m_{ij} = 4)$ treated with chlorhexidine and control in 23. The data are adapted from [22] and is given in Table 7.

Treat	Affected (+)	Not Affected (−)	Total
Chloro. (1)	82	10	92
Control (2)	71	21	92
Total	153	31	184

Table 8. Disease distribution among males and females according to father (exposure variable) disease status.

Exposure	Males (b)			Females (g)			Total
	D^+	D^-	Subtotal	D^+	D^-	Subtotal	
Father+	43	144	187	61	134	195	382
Father−	21	107	128	22	94	116	244
Total	64	251	315	83	228	311	626

have two correlated attributable risk estimator, one describing the disease-expose etiology for males, and the other for females. The main interest here is to compare the AR of males to that of females from the same sib-ship.

Example 4: Correlated *AR*'s from Cross Sectional Study: Family Data

We now consider a highly structured clustered familial data that has a two level hierarchy with blood measurements taken on parents (level two) and their offspring (level one) together with other anthropometric features [23]. Familial data sets are known to have considerable "within-cluster" correlation due to the homogeneous nature of family members. The goal is to classify the offspring blood pressure status based on parents BP and other anthropometric features. The data set contains 223 families with a mean number of siblings equal to 3 siblings per family. The outcome variable in this data set is a binary variable defined as offspring blood pressure status. If simultaneously SBP > 130 and DBP > 80, then an offspring is considered diseased $\left(D^+\right)$ or otherwise normal $\left(D^-\right)$. The exposure variable in this example is whether a parent (here we select the father) has the condition (presence of exposure) or does not have the condition (absence of exposure). The data are presented in Table 8.

We present the general methodology as follows: Testing for gender difference in the population AR is formulated as testing the null hypothesis $H_0: AR_1 = AR_2$ against a general unspecified alternative $H_1: AR_2 = AR_1 + \Delta$. Note that testing this null hypothesis is equivalent to testing $H_0: \theta_1 = \theta_2$, or $H_0: \Delta = 0$.

Let the point estimators be denoted by $\hat{\theta}_1$ and $\hat{\theta}_2$. The difference $D = \hat{\theta}_1 - \hat{\theta}_2$ is asymptotically unbiased and has variance $\operatorname{var}(D) = \operatorname{var}\left(\hat{\theta}_1\right) + \operatorname{var}\left(\hat{\theta}_2\right) - 2\operatorname{cov}\left(\hat{\theta}_1, \hat{\theta}_2\right)$.

Hence the null hypothesis is rejected whenever $Z = D\Big/\sqrt{\operatorname{var}(D)}$ falls in the interval $Z > z_{\alpha/2}$ or $Z < -z_{\alpha/2}$, where $z_{\alpha/2}$ is the $(1-\alpha/2)100\%$ cut off point on the standard normal curve. With a slight difference in notation, $\operatorname{var}\left(\hat{\theta}_i\right)$ is similar to the expression in Equation (6). We derive $\operatorname{cov}\left(\hat{\theta}_1, \hat{\theta}_2\right)$ using the delta method. In general, the data will have a structure similar to that given in Table 9.

We define the moment estimator as before:

$$\widehat{AR}_j = \frac{x_j\left(M_j - Y_j\right) - Y_j\left(N_j - X_j\right)}{M_j\left(X_j + Y_j\right)} \quad j = 1, 2.$$

Let $\hat{\theta}_j = \ln\left(1 - \widehat{AR}_j\right)$, then similar to the first situation, we have:

$$\tau_j^2 = \operatorname{var}\left(\hat{\theta}_j\right) = \frac{N_j\left(1-P_j\right)c_{1j}}{P_j\left(N_j + M_j\psi_j\right)^2} + \frac{N_j^2\left(1-P_j\psi_j\right)c_{2j}}{M_jP_j\psi_j\left(N_j + M_j\psi_j\right)^2}.$$

Table 9. Collapsed data for the analysis of correlated AR parameters.

Exposure	Condition (1)		Condition (2)	
	D^+	D^-	D^+	D^-
E^+	X_1	$N_1 - X_1$	X_2	$N_2 - X_2$
E^-	Y_1	$M_1 - Y_1$	Y_2	$M_2 - Y_2$

Here;

$N_j = \sum_{i=1}^{k_j} n_{ji}$, $M_j = \sum_{i=1}^{l_j} m_{ji}$, $c_{1j} = 1 + \left(n_{0j} - 1\right)\rho_{1j}$, $c_{2j} = 1 + \left(m_{0j} - 1\right)\rho_{2j}$,

and $n_{0j} = \frac{1}{N_j}\sum_{i=1}^{k_j} n_{ji}^2$, $m_{0j} = \frac{1}{M_j}\sum_{i=1}^{l_j} m_{ji}^2$, $\psi_j = \left(1 - AR_j\right)\Big/\left(1 + AR_j\right)$, and

P_j = rate of exposure to the risk factor under the *jth* condition .

Moreover, ρ_{1j} is the intracluster correlation of the exposed clusters under *jth* condition, and ρ_{2j} is the intracluster correlation of the unexposed clusters under *jth* condition. They two parameters are estimated as described in (7).

For simplicity we assume that these correlations are constant among the exposed and unexposed.

Using the delta method we can show after some algebra that:

$$\operatorname{cov}\left(\hat{\theta}_1, \hat{\theta}_2\right) = \rho\left[\alpha_1\alpha_2\gamma_1\gamma_2 + \alpha_2\beta_1\gamma_2\delta_1 + \alpha_1\beta_2\gamma_1\delta_2 + \beta_1\beta_2\delta_1\delta_2\right]. \tag{11}$$

The correlation ρ which, under both conditions is the average correlation among the responses, is estimated as described in Section 3.

The values inside the square bracket are given by:

$$\alpha_j = \frac{\overline{\partial}\theta_j}{\partial x_j} = -\frac{1}{P_j\left(N_j + M_j\psi_j\right)}$$

$$\beta_j = \frac{\overline{\partial}\theta_j}{\partial y_j} = -\frac{N_j}{M_j\psi_j P_j\left(N_j + M_j\psi_j\right)}$$

$$\gamma_j^2 = \operatorname{var}\left(x_j\right) = N_j P_j\left(1 - P_j\right)c_{1j}$$

$$\delta_j^2 = \operatorname{var}\left(y_j\right) = M_j Q_j\left(1 - Q_j\right)c_{2j}$$

$$= M_j\psi_j P_j\left(1 - \psi_j P_j\right)c_{2j} \qquad j = 1, 2.$$

Therefore;

$$\operatorname{var}\left(\hat{\theta}_1 - \hat{\theta}_2\right) = \tau_1^2 + \tau_2^2 - 2\rho\left[\alpha_1\alpha_2\gamma_1\gamma_2 + \alpha_2\beta_1\gamma_2\delta_1 + \alpha_1\beta_2\gamma_1\delta_2 + \beta_1\beta_2\delta_1\delta_2\right]. \tag{12}$$

Using the data in Table 8 we get:

Males: $P_1 = .59$, $M_1 = 128, AR_1 = .19$, $\operatorname{var}\left(\hat{\theta}_1\right) = .0142$.

Females: $P_2 = .63$, $M_2 = 244, AR_2 = .29$, $\operatorname{var}\left(\hat{\theta}_2\right) = .00578$

$\operatorname{var}\left(\hat{\theta}_1 - \hat{\theta}_2\right) = .01987$, $z = \frac{-.211 + .342}{.141} = 0.929$, and $p - \text{value} = 0.353$.

Therefore there is not enough evidence in the data to support the hypothesis of presence of gender differences for the paternal effect on the siblings' hypertension status.

5. Simulations

We carried out a Monte-Carlo study generating the observations from bivariate beta binomial distribution. We restricted our simulations to the situation when the intracluster and the cross clusters correlation are equal. We also assumed a fixed number of observations within each cluster. The purpose was to limit the

number of scenarios under which we examine the properties of the proposed test statistic Z. The statistic $Z = D\Big/\sqrt{\text{var}(D)}$ is computed when both P_1 and P_2 are strictly positive with additional restriction, $\rho < \left(\tau_1^2 + \tau_2^2\right)\Big/2\left[\alpha_1\alpha_2\gamma_1\gamma_2 + \alpha_2\beta_1\gamma_2\delta_1 + \alpha_1\beta_2\gamma_1\delta_2 + \beta_1\beta_2\delta_1\delta_2\right]$. If these conditions are not satisfied, the sample is replaced until a total of 1000 iterations are obtained for each parameter combination. Table 10 shows the empirical levels and powers of the test assuming that the number of clusters is the same under both conditions and for the exposed and the non-exposed clusters. The main conclusions from Table 10, is that the proposed test statistic hold its empirical Type I error rate levels. There is an increase in the test power when the correlation increases, and when the prevalence of exposure parameters P_1 and P_2 are away from the boundaries of the interval (0, 1). We also note an appreciate increase in the power when the number of clusters is above 50, and naturally when the tested parameters are well separated.

6. Discussion

The population Attributable risk, like the odds ratio and relative risk is a measure of disease risk association. However it has a special appeal to public health epidemiologists as it measures the percent reduction in the chances of having the outcome among subjects who are exposed to the risk factor. Clearly, not everyone in the population is exposed to the risk factor. For example, in evaluating the relationship between consanguinity and the risk of PDA, not all parents are relatives. We assume say that 55% of women in the population (as in the Saudi traditional society) are married to a first cousin. To determine how much of a reduction there would be in PDA among CHD newborns we have 0.55 × 0.06 = 3.3%.

We have developed estimators of the variance and the confidence interval on AR when the units of sampling are aggregates of individuals under three study designs. In all situations the estimation of the intraclass correlation is crucial to

Table 10. Empirical type I error rates and powers based on 1000 replications from the bivariate beta binomial distribution. We set $AR_1 = 0.05$, and therefore, $AR_2 = AR_1 - \Delta$.

		$\rho = .1$, $P_1 = .1$, $P_2 = .2$				$\rho = .2$, $P_1 = .1$, $P_2 = .2$				$\rho = .2$, $P_1 = .6$, $P_2 = .7$			
$k = l$	$n = m$	$\Delta = 0$	0.05	0.10	0.25	$\Delta = 0$	0.05	0.10	0.25	$\Delta = 0$	0.05	0.10	0.25
5	2	0.049	0.058	0.078	0.101	0.049	0.060	0.084	0.110	0.049	0.083	0.180	0.296
	3	0.049	0.060	0.084	113	0.049	0.062	0.089	0.122	0.049	0.089	0.210	0.356
	5	0.050	0.062	0.092	0.132	0.050	0.063	0.097	0.136	0.049	0.096	0.247	0.436
50	2	0.049	0.081	0.183	0.350	0.049	0.090	0.22	0.42	0.049	0.218	1.00	1.00
	3	0.050	0.087	0.221	0.444	0.050	0.095	0.260	0.512	0.050	0.257	1.00	1.00
	5	0.051	0.097	0.281	0.595	0.050	0.104	0.312	0.635	0.050	0.309	1.00	1.00
100	2	0.051	0.097	0.281	0.604	0.051	0.111	0.357	0.740	0.050	0.353	1.00	1.00
	3	0.051	0.108	0.359	0.787	0.051	0.122	0.435	0.910	0.051	0.429	1.00	1.00
	5	0.050	0.124	0.472	0.99	0.050	0.140	0.537	1.00	0.051	0.530	1.00	1.00

conduct valid statistical inferences.

One of the objectives of this paper was to develop and evaluate simple test statistic that could be used to compare dependent attributable risks in the case of clustered dichotomous outcome variables.

7. Conclusions

1) Through simulations, a major finding of our work is that to test the equality of correlated attributable risks, either from cross sectional or cohort studies, one needs a much larger number of clusters than that expected to achieve high power.

2) An interesting extension of our study is to construct model-based inference on the *AR*. This would require the development of a semi parametric model similar to the generalized estimating equation, or a full probabilistic model such as generalized linear mixed model where the effect of multiple covariates may be accounted for.

3) A limitation of the simulation study is the restrictions that the number of observations in all clusters is held constant (balanced design) and that the within cluster and the cross cluster correlations are equal. The reason for this assumption is to limit the number of factors which affect the power so that reasonable conclusions can be made. But we believe that these restrictions should not affect the overall conclusions.

Conflict of Interest

The authors declare that there is no conflict of interest.

References

[1] Richardson, J. (1996) Measures of Effect Size. *Behavior Research Methods, Instruments and Computers*, **28**, 12-22. https://doi.org/10.3758/BF03203631

[2] Levin, M.L. (1953) The Occurrence of Lung Cancer In Man. *Acta Unio Internationalis Contra Cancrum*, **9**, 531-541.

[3] Fleiss, J. (1982) Statistical Methods for Rates and Proportions. 2nd Edition, John Wiley, New York.

[4] Fletcher, R.H., Fletcher, S.W. and Wagner, E.H. (1996) Clinical Epidemiology: The Essentials. Lippincott Williams & Wilkins, Philadelphia.

[5] Gordis, L. (1996) Epidemiology. WB Saunders Co., Philadelphia.

[6] Donner, A. and Klar, N. (2000) Design and Analysis of Cluster Randomized Trials in Health Research. Arnold, New York.

[7] McCullagh, P. and Nelder, J.A. (1989) Generalized Linear Models. 2nd Edition, Chapman & Hall/CRC.

[8] Cox, D.R. and Snell, E.J. (1989) Analysis of Binary Data. 2nd Edition, CRC Press, Boca Raton, FL.

[9] Walter, S.D. (1975) The Distribution of Levin's Measure of Attributable Risk from Case-Control Studies. *American Journal of Epidemiology*, **106**, 206.

[10] Kendall, M. and Stuart, A. (1987) Advanced Theory of Statistics. Vol. 1, 5th Edition, Griffin, London.

[11] Saudi Congenital Heart Disease Registry. http://rc.kfshrc.edu.sa/chd_program

[12] Mitchell, S.C., Korones, S.B. and Berendes, H.W. (1971) Congenital Heart Disease, in 56,109 Births. Incidence and Natural History. *Circulation*, **43**, 323-332. https://doi.org/10.1161/01.CIR.43.3.323

[13] Satoda, M., Pierpont, M.E., Diaz, G.A., Bornemeier, R.A. and Gelb, B.D. (1999) Char Syndrome, an Inherited Disorder with Patent Ductus Arteriosus, Maps to Chromosome 6p12-p21. *Circulation*, **99**, 3036-3042. https://doi.org/10.1161/01.CIR.99.23.3036

[14] Satoda, M., Zhao, F., Diaz, G.A., Burn, J., Goodship, J., Davidson, H.R., Pierpont, M.E. and Gelb, B.D. (2000) Mutations in TFAP2B Cause Char Syndrome, a Familial Form of Patent Ductus Arteriosus. *Nature Genetics*, **25**, 42-46. https://doi.org/10.1038/75578

[15] Khoury, S.A. and Massad, D. (1992) Consanguineous Marriages in Jordan. *American Journal of Medical Genetics*, **43**, 769-775. https://doi.org/10.1002/ajmg.1320430502

[16] Jurdi, R. and Saxena, P.C. (2003) The Prevalence and Correlates of Consanguineous Marriages in Yemen: Similarities and Correlates with Other Arab Countries. *Journal of Biosocial Sciences*, **35**, 1-13. https://doi.org/10.1017/S0021932003000014

[17] El-Hazmi, M.A., Al-Swailem, A.R. and Warsey, A.S. (1995) Consanguinity among the Saudi Arabian Population. *Journal of Medical Genetics*, **32**, 623-626. https://doi.org/10.1136/jmg.32.8.623

[18] Becker, S.M., Al Halees, Z., Molina, C. and Paterson, R.M. (2001) Consanguinity and Congenital Heart Disease in Saudi Arabia. *American Journal of Medical Genetics Part A*, **99**, 8-13. https://doi.org/10.1002/1096-8628(20010215)99:1<8::AID-AJMG1116>3.0.CO;2-U

[19] El-Mouzan, M., Al-Salloum, A., Al-Herbish, A., Qurashi, M. and Al-Omar, A. (2008) Consanguinity and Major Genetic Disorders in Saudi Children: A Community-Based Cross-Sectional Study. *Annals of Saudi Medicine*, **28**, 169-174. https://doi.org/10.4103/0256-4947.51726

[20] William, D.A. (1975) The Analysis of Binary Responses from Toxicological Experiments Involving Reproduction and Teratogenicity. *Biometrics*, **31**, 949-952. https://doi.org/10.2307/2529820

[21] Weil, C.S. (1971) Selection of the Valid Number of Sampling Units and a Consideration of Their Combination in Toxicological Studies Involving Reproduction, Teratogenesis or Carcinogenesis. *Food Cosmetics Toxicology*, **8**, 177-182.

[22] Donner, A., Klar, N. and Zou, G. (2004) Methods for the Statistical Analysis of Binary Data in Split-Cluster Designs. *Biometrics*, **60**, 919-925. https://doi.org/10.1111/j.0006-341X.2004.00247.x

[23] Miall, W.E. and Oldham, P.O. (1955) A Study of Arterial Blood Pressure and Its Inheritance in a Sample of the General Population. *Clinical Science*, **14**, 459-487.