The International Journal of Biostatistics

Volume 6, Issue 1

2010

Article 3

Estimation of Modified Concordance Ratio in Sib-Pairs: Effect of Consanguinity on the Risk of Congenital Heart Diseases

Mohamed M. Shoukri, *King Faisal Specialist Hospital and
Research Centre*

Allan Donner, *University of Western Ontario*

Nadia Abdalla Dessouky, *King Faisal Specialist Hospital
and Research Centre*

Shazia Subhani, *King Faisal Specialist Hospital and
Research Centre*

Mansour Al-Joufan, *King Faisal Specialist Hospital and
Research Centre*

Ahmed Al-Omrani, *King Faisal Specialist Hospital and
Research Centre*

Futwan Al-Mohanna, *King Faisal Specialist Hospital and
Research Centre*

Zohair Y. Al Halees, *King Faisal Specialist Hospital and
Research Centre*

Recommended Citation:

Shoukri, Mohamed M.; Donner, Allan; Dessouky, Nadia Abdalla; Subhani, Shazia; Al-Joufan, Mansour; Al-Omrani, Ahmed; Al-Mohanna, Futwan; and Al Halees, Zohair Y. (2010)
"Estimation of Modified Concordance Ratio in Sib-Pairs: Effect of Consanguinity on the Risk of
Congenital Heart Diseases," *The International Journal of Biostatistics*: Vol. 6: Iss. 1, Article 3.
DOI: 10.2202/1557-4679.1216

Estimation of Modified Concordance Ratio in Sib-Pairs: Effect of Consanguinity on the Risk of Congenital Heart Diseases

Mohamed M. Shoukri, Allan Donner, Nadia Abdalla Dessouky, Shazia Subhani, Mansour Al-Joufan, Ahmed Al-Omrani, Futwan Al-Mohanna, and Zohair Y. Al Halees

Abstract

Family studies are widely used for research into genetic and environmental influences on human traits. In this paper, we establish statistical methodology for the estimation of a new measure of sib similarity with respect to dichotomous traits measured on each member of within family sib-pair. We call this parameter "excess risk." For inference problems involving a single sample, we construct a large sample confidence interval on the concerned parameter. It has long been suspected that consanguinity is a risk factor for many genetic defects. Therefore, we establish a procedure to test the significance of the difference between excess risk parameters in a sample of consanguineous marriages and another sample of non-consanguineous marriages. We apply the methodology to data from a hospital-based congenital heart defects registry in Saudi Arabia, a population in which consanguinity is quite common.

KEYWORDS: intra-class correlation, maximum likelihood estimation, non-central chi-square, sample size requirements, Monte Carlo simulations

Author Notes: The authors appreciate constructive comments made by two anonymous referees.

1. Introduction

Evaluating the influence of genetic, environmental, and cultural factors on the variability of traits is a subject of great interest to genetic epidemiologists. The biometrical approach to assessing the degree of similarity of quantitative traits among family members has traditionally relied on measures of correlations. This approach has been quite satisfactory as it depends on the use of the well-developed multivariate normal theory, and is easily extended to investigate the similarity (resemblance) among family members with respect to multiple traits (e.g. obesity and hypertension). However, interest in studying the familial aggregation of traits measured on the dichotomous scale is long standing; that is when interest is focused on the presence or absence of a specific condition, is long standing. For example, in cancer epidemiology, the familial aggregation of ovarian and breast cancer was investigated by Betensky and Whittemore [1996]. Their study involved samples of cases with ovarian cancer and controls without ovarian cancer, all of whom were called probands. The data contained information on the breast cancer status of each proband, as well as the breast cancer and ovarian status of each proband's mother and sisters. The authors used the Quadratic Exponential Model (QEM), which was developed by Zhao and Prentice [1990], to investigate the influence of specified risk factors on the joint occurrence of breast and ovarian cancer. To achieve this objective, the odds ratios were selected as target population parameters to be estimated from the available data.

In genetic epidemiology, studies that investigate the potential familial aggregation of diseases focus on two possibilities, the first of which is that diseases may occur because some of the genes that influence the susceptibility to one disease also influence susceptibility of other family members to the same or other diseases. Second, one or more diseases or disorders may occur because familial-environmental factors may predispose to other disorders. One of the advantages of a sampling scheme that includes family members (clusters) is that related individuals living in the same household are uniquely matched for many cultural, environmental, and genetic factors. The fundamental objective is to compare similarities and differences, and to assess the excess in the risk of disease attributed to such factors. We focus in this paper on consanguinity as a primary risk factor.

This paper has four-fold objectives. Firstly; we identify two of the most important congenital heart defects that are believed to cluster within families. Second; we introduce in this paper a new parameter; the increase-in-risk, or the excess in risk of a specific disease for a sib in a randomly selected family when another sib within the same family has the same disease. This parameter is a simple modification of the well-known Twin Concordance Coefficient (TCC).

The rationale is that, the presence of a diseased family member should increase the risk of developing the disease for his/her sibling; see for example Smith, C. (1974), Allen and Hrubec [1979], Kendler [1989], Hannah, et al. [1993] and Shoukri & Donner (2007). The third objective is to develop probability models whereby the within family excess risk is interpreted as an average population parameter measuring the degree of clustering of the trait of interest. The fourth objective is to apply the methodology to congenital heart defects data and investigate its relation to parental consanguinity.

2- Background on Congenital Heart Defects

Congenital heart disease (CHD) is a heart condition resulting from an abnormality in heart structure or function that is present at birth, although it may not be diagnosed until later in life (Angela et al [2005]). We shall focus on two major heart congenital heart diseases; the Patent Ductus Arteriosus (PDA) and Tetralogy of Fallot (TOF). The PDA is essential for normal fetal development, permitting right ventricular cardiac output to be diverted away from the high-resistance pulmonary circulation. Epidemiologically, prematurity increases the incidence of PDA, mostly due to physiological factors rather than inherent abnormality of the ductus. Kitterman et al [1972], Tanner et al [2005]. 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, Mitchell et al [1971]. PDA was found to occur with increased frequency in several genetic syndromes, although precisely which precise mechanisms resulting in persistent PDA are not yet clear (Mani et al, [2002], Satoda et al, [1999], Satoda et al, [2000]).

PDA is one of the frequently associated lesions in Tetralogy of Fallot (TOF). The first complete description of TOF is credited to the French physician Etienne Fallot, who published his findings in 1888. The classic components that constitute TOF are:

- 1- Ventricular septal defect (single),
- 2- Right ventricular outflow tract obstruction,
- 3- Overriding of aorta, and
- 4- Right ventricular hypertrophy. Presence of PDA helps in maintaining pulmonary blood flow in more complicated cases of TOF (see; Hirsch et al., 2003).

The pathogenesis of congenital heart diseases is complex, with about one-quarter of cases are associated with chromosomal abnormalities. Early studies of familial clustering of congenital heart defects suggested either polygenic or multifactorial inheritance (Nora et al, 1978). Patterns of recurrence of congenital heart diseases in one or more affected first-degree relative were studied, where

exact concordance rate was seen in 37% and 55 % of cases (Gill et al., 2003). Studies to understand the different patterns of congenital heart defects among siblings and twins have been carried out widely (Digilio et al, 2001) supporting the monogenic or oligogenic inheritance of Transposition of Great Arteries (TGA). It has been reported that in a family having one sibling with PDA, there is almost 3% chance of having PDA occurrence in a subsequent offspring (Nora et al., 1968).

Morton [1958] suggested that consanguinity can be used as a tool to investigate a recessive component in the etiology of diseases whose inheritance is uncertain because it promotes homozygosity of genes. This issue can be investigated through case control or cross-sectional investigations. Arab countries in particular are notorious for consanguineous marriages, with first cousin types being the most common. For example in Jordan the prevalence of consanguinity was reported by Khoury and Massad as 51.3% [1992], Yemen, 40% as reported by Jurdi [2003] and almost 57% in Saudi Arabia as reported by El-Hazmi et al. [1995]. In one comparative study from Saudi Arabia, researchers found that first-cousin consanguinity is significantly associated with some congenital heart defects. They concluded that, in a population with a high degree of inbreeding, first-cousin consanguinity may exacerbate underlying genetic risk factors for some types of congenital heart diseases (Becker et al., 2001). More recently, a survey of Saudi families conducted by El Mouzan et al. [2008], estimated the prevalence of consanguinity as 56%.

To investigate the association between consanguinity and congenital heart defects, Gev et al [1986] found that the incidence of CHD among offspring of consanguineous marriages was considerably higher than that among the unrelated, and especially high among the children resulting from first-degree cousin marriages. The study was performed in five villages in the Western Galilee, the population of which is of Arab and Druze origin. This high incidence of CHD in children from closely related parents indicates that genetic influences are important in the etiology. Recently, Khalid et al. [2006] conducted a case-control study that included nine medical centers in Beirut. They confirmed the presence of a significant association between consanguinity and subgroup of CHD among newborns.

One of the major drawbacks of the above studies is that the variances of the estimated effect sizes, whether it is the risk difference or the odds ratio, are under estimated, because the correlation between relatives was ignored. Since it is well known that ignoring the correlation between responses of individuals from the same family, will inflate the Type I error rates leading to misleading conclusions. The models presented in this paper accounts for such correlations within family correlation in estimating disease prevalence and the excess risk parameters.

We focus in this paper on comparing the increase-in-risk parameter, between two groups of consanguineous and non-consanguineous marriages after accounting for the correlation between the pairs of sibs. In Section 3 we construct a probabilistic model characterized by two parameters; prevalence and excess risk (ER). Estimation, hypothesis testing and sample size requirements are discussed in subsequent sections.

3- Methods

Consider two independent strata involving n_j pairs in the j^{th} stratum. Each subject is cross classified by another subject from the same family with ratings denoted as either positive or negative. Denote the probability of a positive and negative rating conditional on family i by $P(+|i) = p_{ij}$ and $P(-|i) = q_{ij}$ respectively, where $p_{ij} + q_{ij} = 1$. Let x_{ij} denote the observations in the response categories as described in Table 1. We summarize the distribution of responses in Table 1(a):

Table 1(a): Conditional distribution of paired of binary responses for j^{th} stratum

Conditional Response category	Observation	Probability
$(++ i)$	$x_{2j}=2$	p_{ij}^2
$(+,- i)$ or $(-,+ i)$	$x_{1j}=1$	$2p_{ij}(1-p_{ij})$
$(-,- i)$	$x_{0j}=0$	$(1-p_{ij})^2$

When averaged over all pairs in the population we obtain:

$$E(p_{ij}) = \pi_j$$

$$E(p_{ij}^2) = \text{var}(p_{ij}) + \pi_j^2$$

$$\begin{aligned} E(2p_{ij}(1-p_{ij})) &= 2\pi_j - 2[\text{var}(p_{ij}) + \pi_j^2] \\ &= 2\pi_j(1-\pi_j) - 2\text{var}(p_{ij}) \end{aligned}$$

and

$$\begin{aligned} E((1-p_{ij})^2) &= 1 - 2E(p_{ij}) + E(p_{ij}^2) \\ &= 1 - 2\pi_j + \text{var}(p_{ij}) + \pi_j^2 \end{aligned}$$

Moreover, we assume that $\text{var}(p_{ij}) = \rho_j \pi_j (1 - \pi_j)$. Therefore we obtain:

$$\begin{aligned}
 E(p_{ij}^2) &= \pi_j^2 + \rho_j \pi_j (1 - \pi_j) \\
 E(2p_{ij}(1 - p_{ij})) &= 2\pi_j(1 - \pi_j)(1 - \rho_j) \\
 \text{and } E((1 - p_{ij})^2) &= (1 - \pi_j)^2 + \rho_j \pi_j (1 - \pi_j).
 \end{aligned}$$

Therefore, the unconditional distribution of responses for the j^{th} stratum can be summarized in Table 1(b).

Table 1(b): Distribution of pairs of correlated binary responses in the j^{th} stratum

Category	Observation	Probability	Observed Frequency
(+,+)	$x_{2j}=2$	$\pi_j^2 + \rho_j \pi_j (1 - \pi_j)$	n_{2j}
(+,-)or(-,+)	$x_{1j}=1$	$2\pi_j(1 - \pi_j)(1 - \rho_j)$	n_{1j}
(-,-)	$x_{0j}=0$	$(1 - \pi_j)^2 + \rho_j \pi_j (1 - \pi_j)$	n_{0j}
Total		1	n_j

Dropping the suffix j , for two individuals who belong to the same family, we define the excess risk Δ as the difference between two probabilities:

$$\begin{aligned}
 \Delta &= P_r[X_2 = + | X_1 = +] - P_r[X_2 = +] \\
 &= \frac{P_r[X_1 = +, X_2 = +]}{P_r[X_1 = +]} - P_r[X_2 = +] \\
 &= \frac{\pi^2 + \rho\pi(1 - \pi)}{\pi} - \pi \\
 &= \rho(1 - \pi)
 \end{aligned}$$

We interpret Δ as the increase in disease risk for a randomly selected sibling, who already has a diseased sibling in the same family. The parameter ρ is known as the intracluster correlation and measures the extent of clustering of a condition measured on the dichotomous scale (see; Rosner 1989). Clearly, when there is no clustering of disease (i.e. $\rho=0$) then there is no increase in the risk of disease that can be attributed to familial effect. Therefore the model in Table 2 can be re-parameterized as follows:

Table 2: Re-parameterized model in terms of prevalence and excess risk parameter

Category	Observation	Probability	Observed Frequency
(+,+)	$x_{2j}=2$	$\pi_j^2 + \pi_j \Delta_j$	n_{2j}
(+,-) or (-,+)	$x_{1j}=1$	$2\pi_j(1 - \pi_j - \Delta_j)$	n_{1j}
(-,-)	$x_{0j}=0$	$(1 - \pi_j)^2 + \pi_j \Delta_j$	n_{0j}
Total		1	n_j

3-1 Likelihood Estimation

Suppose that random samples of n_j pairs of sibs are available from the two strata.

The logarithm of the joint likelihood of the samples is:

$\ln L(\pi, \Delta) = \ln L_1(\pi_1, \Delta_1) + \ln L_2(\pi_2, \Delta_2)$, where

$$\ln L_j(\pi_j, \Delta_j) = n_{2j} \ln\{\pi_j^2 + \Delta_j \pi_j\} + n_{1j} \ln\{2\pi_j(1 - \pi_j - \Delta_j)\} + n_{0j} \ln\{(1 - \pi_j)^2 + \Delta_j \pi_j\}$$

Differentiating with respect to π_j and Δ_j , equating to zero and solving for π_j and Δ_j we get:

$$\hat{\pi}_j = \frac{n_{1j} + 2n_{2j}}{2n_j}, \quad \hat{\Delta}_j = 1 - \frac{n_{1j} + 2n_{2j}}{2n_j} - \frac{n_{1j}}{n_{1j} + 2n_{2j}}$$

Inverting the Fisher's information matrix, the elements of the asymptotic variance-covariance matrix of the maximum likelihood estimators are given by:

$$\text{var}(\hat{\pi}_j) = \frac{\pi_j(1 - \pi_j) + \pi_j \Delta_j}{2n_j} \quad (1)$$

$$\begin{aligned} \text{var}(\hat{\Delta}_j) &= \frac{(2\pi_j + \Delta_j)^2}{2n_j \pi_j} [(1 - \pi_j - \Delta_j)(1 - 2\pi_j(1 - \pi_j - \Delta_j))] \\ &+ \frac{(1 - 2\pi_j - \Delta_j)^2}{n_j \pi_j} [(\pi_j + \Delta_j)(1 - \pi_j^2 - \Delta_j \pi_j)] \\ &+ \frac{2(\pi_j + \Delta_j)(1 - \pi_j - \Delta_j)(2\pi_j + \Delta_j)(1 - 2\pi_j - \Delta_j)}{n_j} \end{aligned} \quad (2)$$

$$\text{Moreover, } \text{cov}(\hat{\pi}_j, \hat{\Delta}_j) = -\frac{2\pi_j + \Delta_j}{2n_j}.$$

From (1) and (2), an asymptotic $(1-\alpha)100\%$ confidence interval on Δ_j is given by: $\hat{\Delta}_j \pm z_{\alpha/2} \sqrt{\text{var}(\hat{\Delta}_j)}$, where $z_{\alpha/2}$ is the $(1-\alpha/2)$ percentile of the standard normal distribution.

Remarks:

It should be noted that admissibility and validity of asymptotic statistical inferences for the maximum likelihood estimators (MLE) depend on the data structure as shown in Table 2. For example, n_{0j} carries no information. Moreover, when $n_{1j} = 0$, the MLE's are admissible and $\hat{\pi}_j = 1 - \hat{\Delta}_j = n_{2j} / n_j$. On the other hand if $n_{1j} = n_{2j} = 0$, then $\hat{\pi}_j = 0$, $\hat{\Delta}_j = 1$ and the variance covariance matrix of the MLE estimators is zero, with no meaningful inference possible. Finally, if $n_{2j} = 0$ (the condition is absent from both siblings), then $\hat{\pi}_j = -\hat{\Delta}_j = n_{1j} / 2n_j$, which is an inadmissible estimator. However we note that $P_r[n_{2j} = 0] = (\pi_j^2 + \Delta_j \pi_j)^{n_j}$ which approaches zero when n_j approaches infinity. It is clear then that maximum likelihood based inference is valid when sufficiently large samples are available.

3-2 Homogeneity Testing

For genetic diseases, comparing the ER and prevalence parameters between two independent groups is an important step in establishing quantitative evidence to be used by genetic councilors and genetic epidemiologists. In this section we formulate the problem of testing the equality of (Δ_j, π_j) ($j = 1, 2$) in a general set up. We assume that a random sample of pairs of siblings is available from the respective group. It is of interest to test the simultaneous hypotheses given by:

$$H_0: \Delta_1 = \Delta_2 \cap \pi_1 = \pi_2 \text{ Versus } H_1: \text{all possible alternatives} \quad (3).$$

The hypothesis of interest here tests the equality of the excess risk and prevalence parameters in the two populations.

Let $\phi = (\Delta_1, \pi_1, \Delta_2, \pi_2)$ and let $\hat{\phi} = (\hat{\Delta}_1, \hat{\pi}_1, \hat{\Delta}_2, \hat{\pi}_2)$ denote its maximum likelihood estimator.

Under regularity conditions set by Bishop et al [1975] and asymptotic results due to Serfling [1983] we can establish, that $\hat{\phi}$ is asymptotically normally distributed with mean ϕ and asymptotic variance covariance matrix V , or equivalently:

$$\hat{\phi} \xrightarrow{p} \phi, \sqrt{n}(\hat{\phi} - \phi) \xrightarrow{d} N(0, V) \quad (4)$$

The matrix V is a 4 x 4 such that:

$$V = \begin{bmatrix} V_1 & O \\ O & V_2 \end{bmatrix}$$

Here V_j is a 2X2 square symmetric matrix, whose first diagonal element is the asymptotic variance of $\hat{\Delta}_j$, the second diagonal element is the asymptotic variance of $\hat{\pi}_j$ and the off-diagonal elements are the asymptotic covariance between $\hat{\Delta}_j$ and $\hat{\pi}_j$. Setting $\delta = L\phi$, and using results from Graybill [1976] the hypothesis in (3) is equivalent to testing $H_0 : \delta = 0$ versus $H_1 : \delta > 0$, where L is a suitably chosen linear transformation similar to that given in Shoukri and Donner [2007].

Following a similar procedure similar to that of Shoukri and Donner [2007] the quadratic form

$$Q(\xi) = n\hat{\delta}^T (LVL^T)^{-1} \hat{\delta} \quad (5)$$

has, asymptotically non-central chi-square distribution with two degrees of freedom and non-centrality parameter:

$$\xi = n\delta^T (LVL^T)^{-1} \delta = n\vartheta. \quad (6)$$

Moreover, $\xi = 0$ if and only if H_0 is true. Therefore, for given type I error rate α , by referring $Q(0)$ to the table of chi-square distribution with two degrees of freedom, $H_0 : \delta = 0$ is rejected if $Q(0)$ exceeds the tabulated value of a chi-square with two degrees of freedom.

3-3 Data Analyses

In this example, data are provided by the Congenital Heart Defects Registry, established in 1998 at the King Faisal Specialist Hospital and Research Center (KFSHRC), Riyadh region. It is an on-going research project with a centralized database, using a web-based registry interface. The registry has been extended on national level with collaborations from major hospitals. The data included in the study were obtained by a two-stage-sampling procedure. In the first stage families

were selected from the list of family identification numbers, and in the second stage two siblings were randomly selected from within the attained family. Total counts pertaining to the targeted defects were collected by querying the CHD database according to the ICD.9.CM [2005]. The data are tabulated in Tables 4 (a, b, c, d). In the following analysis, we focus on the estimation in excess risk attributed to family member history, and compare the ER between two independent samples of consanguineous and non-consanguineous marriages. The results are summarized in Table 3.

Table 3: Maximum likelihood estimation of model parameters for both diseases.

Stratum	Consanguineous marriages Stratum (1)	Consanguineous Marriages Stratum (1)	Non-consanguineous Marriages Stratum (2)	Non-consanguineous Marriages Stratum (2)
Parameter estimates	PDA $n = 152$	TOF $n = 152$	PDA $n = 96$	TOF $n = 96$
$\hat{\pi}$	0.346 (0.031)	0.069 (0.019)	0.281 (0.038)	0.073 (0.019)
$\hat{\Delta}$	0.213 (0.053)	0.596 (0.114)	0.275 (0.074)	0.071 (0.124)
$\text{cov}(\hat{\pi}, \hat{\Delta})$	-0.003	-0.0024	-0.0043	-0.0011

Bracketed numbers are the estimated standard errors of the estimates.

For PDA, $Q(0) = 1.713$, p-value = 0.421, and for TOF, $Q(0) = 9.866$, p-value = 0.007.

Based on the results of the data analyses we may conclude that excess risk and prevalence of the PDA are not affected by consanguinity. For TOF, excess risk and prevalence are significantly higher among consanguineous marriages as compared to non-consanguineous marriages.

Table 4 a: Non-related marriages

		TOF <i>sib_1</i>		Total
		Yes	No	
<i>sib_2</i>	Yes	1	5	6
	No	7	83	90
Total		8	88	

Table 4 b: Related Marriages

		Yes	No	Total
<i>sib_2</i>	Yes	7	2	9
	No	5	138	143
Total		12	140	

Table 4 c: Non-related marriages

		PDA <i>sib_1</i>		Total
		Yes	No	
<i>sib_2</i>	Yes	15	14	29
	No	10	57	67
Total		25	71	

Table 4 d: Related Marriages

		Yes	No	Total
<i>sib_2</i>	Yes	28	23	51
	No	23	78	101
Total		51	101	

3-4 Sample Size Requirements

The issue of sample size estimation is quite important as it has two objectives. First, based on the estimated sample size one should be able to decide whether the study is under powered or not. Second, and as we demonstrated, the efficiency of the maximum likelihood estimation requires sufficiently large sample, and therefore we should be able to decide on the magnitude of the sample needed to report meaningful results. For a given level of significance α , we have $\alpha = P_r[Q(0) > X_{2,\alpha}^2]$. To control the power at $1 - \beta$, one needs to choose $n_1 = n_2 = n$ so that for fixed α , $1 - \beta = P_r[X_{2,\alpha,\xi}^2 > X_{2,\alpha}^2]$. Here, $X_{2,\alpha,\xi}^2$ and $X_{2,\alpha}^2$ denote, respectively, non-central and central chi-square distributions with two degrees of freedom. Solving the power equation for n , we get the required sample size, which is assumed constant in the two strata. Table 5 gives the sample sizes for selected values of $(\pi_1, \pi_2, \Delta_1, \Delta_2)$ for Type I error rate $\alpha = 0.05$ and power, $1 - \beta = 0.80$.

Table 5: Sample sizes for parameter combinations, Type I error rate $\alpha = 0.05$, and power $1 - \beta = 0.80$.

π_1	π_2	Δ_1	Δ_2	θ	$n = n_1 = n_2$
.50	.55	.50	.55	.009	1119
.50	.55	.50	.60	.022	477
.50	.55	.50	.70	.664	163
.50	.55	.50	.90	.172	61
.20	.55	.50	.60	.303	35
.20	.60	.40	.55	.416	26
.50	.60	.50	.60	.034	308
.50	.70	.50	.70	.110	95
.34	.28	.21	.28	.014	720*
.070	.073	.60	.07	.080	132**

* The total sample size required for the PDA condition to detect a significant effect of consanguinity using the results of the data analyses.

** The total sample size required for the TOF condition to detect a significant effect of consanguinity using the results of the data analyses.

We note that the available sample size is much lower than the required sample to achieve the desired level of power. However, since the registry is an ongoing surveillance system, it is hoped that the sample size needed to produce efficient estimates may be attainable in the near future. We may then view the

present analysis a “pilot” study whose results may be used as baseline information to guide future research.

3-5 Homogeneity Testing with Common Prevalence

Under certain conditions, the disease prevalence may be homogeneous in the two examined strata, in which case the interest may focus on the homogeneity of the excess risk parameters, with the prevalence characterized as a nuisance parameter. The data in this example clearly indicate that the prevalence of the two conditions is almost the same in both strata. The maximum likelihood estimators under this simple model are then as shown in the previous section. However, the estimated common prevalence $\hat{\pi}$ is now the simple average of $\hat{\pi}_1$ and $\hat{\pi}_2$, and the null hypothesis to be tested in this case is $H_0 : \Delta_1 = \Delta_2$ versus $H_1 : \Delta_1 \neq \Delta_2$, with π being a nuisance parameter. This simple hypothesis can be tested using the Wald’s statistic is given by:

$$Z = \frac{\hat{\Delta}_1 - \hat{\Delta}_2}{\sqrt{\text{var}_c(\hat{\Delta}_1) + \text{var}_c(\hat{\Delta}_2)}} \quad (7)$$

Asymptotically, Z has a standard normal distribution. One can also use the pivot Z to construct $(1 - \alpha)100\%$ confidence interval on the difference $\Delta_1 - \Delta_2$ in the form $(\hat{\Delta}_1 - \hat{\Delta}_2) \pm z_{\alpha/2} \sqrt{\text{var}_c(\hat{\Delta}_1) + \text{var}_c(\hat{\Delta}_2)}$ where $z_{\alpha/2}$ is the $(1 - \alpha/2)100\%$ percentile of the standard normal distribution. Note that $\text{var}_c(\hat{\Delta}_j)$ is the variance of $\hat{\Delta}_j$ given in equation (2) when π_j ($j=1, 2$) is replaced with the estimate of the common π .

Table 6: Parameters estimation under the assumption of prevalence homogeneity

Stratum	Consanguineous Stratum (1)	Consanguineous Stratum (1)	Non-consanguineous Stratum (2)	Non-consanguineous Stratum (2)
Parameter estimates	PDA $n = 152$	TOF $n = 152$	PDA $n = 96$	TOF $n = 96$
$\hat{\pi}$	0.209	0.071	0.209	0.071
$\hat{\Delta}$	0.213 (0.042)	0.596 (0.112)	0.275 (0.089)	0.071 (0.125)
$\text{cov}(\hat{\pi}, \hat{\Delta})$	- 0.002	- 0.0024	-0.0036	-0.0011

It is possible to provide sample size estimate in close form, allowing for unequal sample sizes in each stratum. This can be done using a method proposed by Donner and Eliasziw [1992], assuming unequal sample sizes n_1 and $n_2 = rn_1$. It is straightforward to show that:

$$n_1 = (z_{\alpha/2} + z_\beta)^2 \frac{r}{1+r} \cdot \frac{(\theta_1^{-1} + \theta_2^{-1} + \theta_3^{-1})^{-1}}{\pi^2 (\Delta_1 - \Delta_2)^2} \quad (8)$$

where, $\theta_1 = \pi^2 + \pi\Delta_w$, $\theta_2 = 2\pi(1 - \pi - \Delta_w)$, $\theta_3 = (1 - \pi)^2 + \pi\Delta_w$, and $\Delta_w = \frac{\Delta_1 + r\Delta_2}{1+r}$.

As an example, for Type I error rate 5% (two sided) and power 80%, $r = 2$, $\pi = 0.1$, $\Delta_1 = 0.4$, $\Delta_2 = 0.2$, the estimated sample size for the first stratum is $n_1 = 489$. Therefore, $n_2 = 978$, and the total sample size would be $n_1 + n_2 = 1467$.

4. Simulations

The results we presented in the above section depend on the accuracy of the assumption of normality for the Wald statistic Z . The common approach for evaluating this assumption is to conduct Monte Carlo simulations. For sample sizes 50, 100, and 200 we simulated 5000 pairs of correlated binary observations with exchangeable correlation structure using the algorithm proposed by Lunn and Davies [1998]. The validity of the normality assumption was evaluated graphically using Q-Q plots. The Q-Q plots are given in Figures 1, 2, 3 and they seem to support the asymptotic normality assumption.

In table 7 we estimated the power for total sample sizes 50, 100, and 200, assuming balanced and unbalanced group sizes. The results are summarized as follows:

- 1- As we expect, the power increases with the increase in n_1 and n_2 as well as with the effect size ε .
- 2- Even for large samples, the absence of balance in the group sizes significantly reduces the power.
- 3- For small values of the binary response probability π , the power is poor. However the power increases when this probability is 0.5. The reason is that the binomial distribution is symmetric, and is closer to normality when $\pi = 0.5$. This observation has been confirmed in similar studies; see for example Donner et al. [1996].

Table 7: Empirical Type I error rates α , and powers for testing $\varepsilon = \Delta_1 - \Delta_2$, based on 5000 simulations, at $\Delta_1 = 0.1$.

n_1	n_2	ε	Δ_1	π_1	π_2	Power %
25	25	0.0	0.1	0.1	0.1	4.3
		0.1	0.1	0.1	0.1	8.6
		0.3	0.1	0.1	0.1	27.1
		0.3	0.1	0.3	0.3	56.0
		0.3	0.1	0.5	0.5	52.0
50	50	0.0	0.1	0.1	0.1	4.5
		0.1	0.1	0.1	0.1	11.6
		0.3	0.1	0.1	0.1	48.2
		0.3	0.1	0.3	0.3	52.0
		0.3	0.1	0.5	0.5	83.3
10	40	0.0	0.1	0.1	0.1	4.5
		0.1	0.1	0.1	0.1	7.1
		0.3	0.1	0.1	0.1	15.1
		0.3	0.1	0.3	0.3	25.5
		0.3	0.1	0.5	0.5	48.5
10	90	0.0	0.1	0.1	0.1	4.5
		0.1	0.1	0.1	0.1	7.4
		0.3	0.1	0.1	0.1	16.2
		0.3	0.1	0.3	0.3	28.9
		0.3	0.1	0.5	0.5	51.4
100	100	0.0	0.1	0.1	0.1	4.5
		0.1	0.1	0.1	0.1	16.9
		0.3	0.1	0.1	0.1	89.8
		0.3	0.1	0.3	0.3	82.5
		0.3	0.1	0.5	0.5	98.8

Normal Quantile—Quantile plot for 5000 replicates

Figure 1: Normal plot for: $n = 50, \pi_1 = \pi_2 = \Delta_1 = \Delta_2 = 0.3$

Figure 2: $n = 100, \pi_1 = 0.4, \pi_2 = 0.3, \Delta_1 = 0.5, \Delta_2 = 0.4$

Figure 3: $n = 200, \pi_1 = 0.1, \pi_2 = 0.3, \Delta_1 = 0.1, \Delta_2 = 0.2$

5- Discussion

Disease registries constitute large national observational data bases that collect data on patients classified with specific diseases. Disseminating information about disease distribution and response to treatment help improve understanding and disease management, and hence appropriate allocation of resources. The key objectives of the Congenital Heart Defects Registry (CHD) of Saudi Arabia are, to provide accurate information about families with inherited heart diseases, create awareness and provide a reliable source of information to policy makers, and establish data about treatment outcomes and create opportunities to participate in research programs. In this paper we used cross-sectional data from the Saudi CHD, and applied epidemiological concepts and statistical methodology to address the issue of clustering of CHD within families, and emphasize the role of consanguinity as an antecedent risk factor. The proposed index of disease clustering, which we named “excess risk” is a modification of the well-known concordance ratio. Two “excess risk” indices were compared and sample size requirements (the total number of pairs of siblings required each groups) to achieve a certain power levels were provided. We note that, depending on the magnitude of the effect size, relatively large sample sizes are needed, implying continued support for the registry is needed. This support comes in the form of coordination among cardiologists from the participating cardiac centers and hospitals, all of them are members of the CHD steering committee and the Registry Core Facility (RCF) of the Department of Biostatistics and Epidemiology of the King Faisal specialist Hospital and Research Center. The RFC plays a critical role in organizing and maintaining high data quality. Continued support is needed to guarantee availability of sufficient data to conduct registry-based research.

References

- Allen, G. and Hrubec, Z. (1979). Twin Concordance: A general model. *Acta geneticae medicae et gemellologiae*. 28; 3-13.
- Angela E Lin, Holly H Ardinger. (2005). Genetic epidemiology of cardiovascular malformations. *Progress of Pediatric Cardiology Journal*, 20: 113-126.
- Becker SM, Al Halees Z, Molina C, Paterson RM. (2001) Consanguinity and congenital heart disease in Saudi Arabia. *American Journal of Medical Genetics*, Feb 15; 99(1):8-13.
- Betensky, R.A., and Whitmore, A.S. (1996). An analysis of correlated Multivariate Binary Data: Application to Familial Cancers of the Ovary and Breast. *Applied Statistics*, 45; No.4, 411-429.

- Bishop, YM, Fienberg, SE, Holland PN. (1975) *Discrete Multivariate Analysis: Theory and Practice*. MIT Press: Cambridge, MA.
- Digilio MC, Casey B, Toscano A, Calabrò R, Pacileo G, Marasini M, Banaudi E, Giannotti A, Dallapiccola B, Marino B. (2001). Complete transposition of the great arteries: patterns of congenital heart disease in familial precurrence. *Circulation*, Dec 4; 104 (23):2809-14.
- Donner, A., and Eliasziw, M. (1992). A goodness-of-fit approach to inference procedures for the kappa statistics: Confidence interval construction, significance testing and sample size estimation. *Statistics in Medicine*, 11 1511-1519.
- Donner, A. Eliasziw, M. and Klar, N. (1996). A comparison of methods for testing homogeneity of proportions in teratologic studies. *Statistics in Medicine*, 3, 1253-1264.
- El-Hazmi, M.A., Al-Swailem, A. R., Warsey, A.S., (1995). Consanguinity among the Saudi Arabian population. *Journal of Medical Genetics*, 32:623-626.
- El-Mouzan, M., Al-Salloum, A., Al-Herbish, A., Qurashi, M., Al-Omar, A. (2008). Consanguinity and major genetic disorders in Saudi children: A community-based cross-sectional study. *Annals of Saudi Medicine*, 28:3,169-174.
- Gev, D., Roguin, N., Freundlich, E. (1986). Consanguinity and congenital heart diseases in the rural Arab population in Northern Israel. *Human Heredity*, 36: 213-217.
- Gill HK, Splitt M, Sharland GK, Simpson JM. (2003) Patterns of recurrence of congenital heart disease: an analysis of 6,640 consecutive pregnancies evaluated by detailed fetal echocardiography. *Journal of American College of Cardiology*. Sep 3, 42(5):923-9.
- Graybill F. (1976). *Theory and Application of the Linear Models*. Duxbury: North Scituate, MA
- Hannah, M.C., Hopper, J.L., Mathews, J.D. (1983). Twin concordance for a binary trait: Statistical models illustrated with data on drinking status. *Acta Geneticae Medicae et Gemellologiae*, 32: 127-137
- Hirsch JC and Bove EL. Tetralogy of Fallot in: Mavroudis C and Backer CL, eds. (2003) *Pediatric Cardiac Surgery*. 3rd ed. Philadelphia, Mosby,: p 383-385. International Classification of Diseases 9th Revision. Clinical Modification, sixth revision (2005). 4727 Wilshire Boulevard, Suite 300, Los Angeles, California.
- Jurdi, R., 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.
- Kendler K.S. (1989). Limitations of the ratio of concordance rates in monozygotic twins and dizygotic twins. *Archives of General Psychiatry*, 46: 477-478.

- Khalid, Y., Ghina, M., Fadi, B., Fadi, C., May, K., Joseph, R., Makhoul, G., Hala, T. (2006). Consanguineous marriages and congenital heart defects: A case-control study in the neonatal period. *American Journal of Medical Genetics*, part A, 140A: 1524-1530.
- Khoury, S A, Massad, D. (1992). Consanguineous marriages in Jordan. *American Journal of Medical Genetics*, 43: 769-775.
- Kitterman JA, Edmunds LH Jr, Gregory GA, Heyman MA, Tooley WH, Rudolph AM. (1972) Patent ductus arteriosus in premature infants: incidence, relation to pulmonary disease and management. *New England Journal of Medicine*, 287: 473–477.
- Lunn, A.D., and Davies, S.J. (1998). A note on generating correlated binary variables. *Biometrika*, 85, 2, 487-490.
- Mani A, Meraji S, Houshyar R, Radhakrishnan J, Mani A, Ahanger M, Rezaie TM, Taghabinejad MA, Broumand B, Zhao H, Nelson-Williams C, Lifton RP. (2002) Finding genetic contributions to sporadic disease: a recessive locus at 12q24 commonly contributes to patent ductus arteriosus. *Proceedings of the National Academy of Sciences*, USA. 99: 15054–15059
- Mitchell SC, Korones SB, Berendes HW. (1971) Congenital heart disease, in 56,109 births. Incidence and natural history. *Circulation*.; 43:323–332
- Morton, N.E., (1958). Empirical risks in consanguineous marriages: Birth weight, gestation time, and measurements of infants. *American Journal of Human Genetics*, 10: 344- 349.
- Nora JJ, Nora AH. (1978) The evolution of specific genetic and environmental counseling in congenital heart diseases. *Circulation*, Feb 57(2):205-13.
- Nora JJ. (1968) Multifactorial inheritance hypothesis for the etiology of congenital heart diseases: the genetic-environmental interaction. *Circulation*, 38(3): 604-617.
- Rosner, B. (1989) Multivariate methods for clustered binary data with more than one level of nesting. *Journal of the American Statistical Association*, 84: 373-380.
- Satoda M, Pierpont ME, Diaz GA, Bornemeier RA, Gelb BD. (1999) Char syndrome, an inherited disorder with patent ductus arteriosus, maps to chromosome 6p12-p21. *Circulation*, 99: 3036–3042
- Satoda M, Zhao F, Diaz GA, Burn J, Goodship J, Davidson HR, Pierpont ME, Gelb BD. (2000) Mutations in TFAP2B cause Char syndrome, a familial form of patent ductus arteriosus. *Nature Genetics*, 25: 42–46.
- Serfling RJ. (1983). *Approximation Theory of Mathematical Statistics*. Wiley: New York.
- Shoukri, M.M., and Donner, A. (2007). Bivariate models for co-aggregation of dichotomous traits in twins. *Statistics in Medicine*, 26: 337-351.

- Smith, C. (1974). Concordance in twins: methods and interpretation. *American Journal of Human Genetics*, 26: 454-466.
- Tanner K, Sabrine N, Wren C. (2005) Cardiovascular malformations among preterm infants. *Pediatrics*, Dec; 116(6):e833-8.
- Zhao, L.P., and Prentice, R.L. (1990) Correlated binary regression using a quadratic exponential model. *Biometrika*, 77, 642-648.