Estimating Sibling Recurrence Risk In Population Sample Surveys

Barry I. Graubard[1], Monroe G. Sirken[2]

[1]National Cancer Institute, 6120 Executive Blvd, Bethesda, MD 20892

[2]CDC/National Center for Health Statistics, 3311 Toledo Road, Hyattsville, MD 20782

Abstract

Sibling recurrence risk (SRR) is a measure of familial aggregation of a disease and is often used in family-based studies in genetic epidemiology to indicate the existence of possible genes conferring susceptibility of disease. Estimating SRR requires information about the disease status of sibships of families with affected children. Since family-based studies are not usually random samples, estimates of SRR derived from these studies may be biased. Probability samples of individuals obtained in surveys such as the National Health Interview Survey (NHIS) can be used to obtained unbiased estimators of SRR and its related SRR ratio (SRR divided by the prevalence of disease). Two methods of ascertaining sibships of affected families are described and illustrated for estimating SRR and SRR ratio for diabetes from the NHIS. Estimators of standard errors of SRR and SRR ratio are provided along with consideration of reporting error to compare the ascertainment methods.

Key words: Counting rule; network estimation; sampling errors; reporting errors

1. Introduction

Sibling recurrence risk (SRR) is a measure of familial aggregation of a disease or an adverse health condition. SRR may be operationally defined as the proportion of siblings of affected individuals in a population that are affected. The measure is widely used in genetic epidemiological studies to provide evidence of possible genes conferring susceptibility to disease (Penrose, 1953; Risch, 1990a, 1990b, 1990c; Guo, 1998; Olsen and Cordell, 2000; Zou and Zhao, 2004) and has been used to aid in the counseling of patients and families about the risk of genetic diseases (Sutton, 2011). Also, it is potentially useful in social and ecological epidemiological studies assessing the extent that person-to-person spread of adverse health conditions (viz. smoking, obesity, alcoholism, etc.) contribute to epidemics (Christakis and Fowler, 2007; Christakis and Fowler, 2008; Rosenquist et al., 2010).

Typically, genetic studies that collect sibship data that are used to estimate SRR are relatively small nonrandom samples obtained from, for example, an ad hoc collection of affected individuals from clinics or physician practices. It is through the reporting by these sampled affected individuals (called probands) that their sibships are ascertained

along with disease statuses of the members of sibships. Because the probands are not a random sample the ascertainment of sibships can be a biased (called ascertainment bias) where estimates of SRR derived from these samples sibships can be substantially be biased (Guo, 1998). One approach to avoid ascertainment bias is to select a simple random sample or a census of sibships from the population with at least one affected in each sibship and estimate SRR as proposed by Olson and Cordell (2000). Ascertaining samples of sibships in this fashion is called complete ascertainment in the genetic literature (Olson and Cordell, 2000). However, population registries or sample frames of sibships from which simple random samples can be selected are often lacking. Another approach is to ascertain sibships from a simple random sample of affected individuals from the population. The sibships obtained in this fashion are assumed to be selected proportional to the number of affected individuals in each sibship. This type of sampling of sibships is called single ascertainment, and Olson and Cordell (2000) provide a consistent estimator of SRR under appropriate assumptions and variances for the estimator are given by Zou and Zhang (2004). However, because of limited population-based disease registries from which to obtain simple random samples of affected individual proposed estimators of SRR are problematic.

In this paper we propose an alternative approach that extends the single ascertainment approach by using large population-based surveys for randomly sampling individuals in the population (whether or not they are affected) who report on the disease status of themselves and their siblings. These sampled individuals will be referred to as probands throughout this paper. We allow the same sibships to be reported by potentially multiple sampled siblings, but will correct for multiple reporting by appropriately weighting reported sibships. We consider two estimators of SRR where each estimator is based on a different counting rule – one where only affected individuals who are sampled can report the disease status of their siblings and the other where individuals regardless of disease status can report the disease status of themselves and their siblings. The counting rule designates individuals through which each sibship is ascertained. In genetic studies affected probands are thought to be more accurate reporters of the disease status of their sibling than unaffected probands (Guo, 1998). We will examine this empirically in our diabetes example by comparing estimates of prevalence and SRR of diabetes under the two counting rules. We describe how to weight the observations to account for (1) the differential sampling fractions used to sample probands in the survey and (2) the multiple reporting defined by the counting rule in order to obtain unbiased estimators of SRR. Since a population survey allows also for the unbiased estimation of the population prevalence of the disease, we also consider estimation of the SRR ratio (also called sibling relative risk) which is the SRR divided by an estimate of the prevalence of the disease. The SRR ratio is a measure familial aggregation of disease relative to the prevalence the disease and is useful for planning linkage analyses of studies of affected relative pairs (Risch, 1990a, 1990b, 1990c). This paper presents a mathematical basis for estimating sibling recurrence risk and risk recurrence ratio and the precision in population surveys of large random samples of individuals.

The rest of paper is organized as follows: Section 2 describes the concept of counting rule used in surveys and proposes two counting rules that are used later to estimate SRR and SRR ratio of diabetes. In Section 3 our proposed population-based estimators of SRR and SRR ratio associated with each of our two counting rules are given, and Section 4 provides variance estimators for our proposed SRR and SRR ratios estimators. In Section 5 the 1976 U.S. National Health Interview Survey (NHIS) is used to illustrate our estimators for obtaining family aggregation diabetes. Finally in Section 6 we briefly discuss the results of the paper.

2. Counting Rules and Network Sample Estimation

In surveys *counting rules* specify conditions for linking population elements to selection units at which the population elements are eligible to be reported in the survey (Sirken, 1998). In our application the population elements are the sibships and the selection units are all individuals that are eligible to be selected in the survey sample. We consider only counting rules with the property that every sibship is linked to at least one individual eligible to be sampled in the survey. We will consider two counting rules for reporting of sibships: (1) an *All Sibling* counting rule (AllSCR) where any individual who is selected in the NHIS sample can report about their sibship; and (2) an *Affected Sibling* counting rule (AffSCR) where only an affected who is selected in the NHIS sample can report about their sibship. The AffSCR violates the property stated above because sibships without any affected siblings will not have an affected individual linked to it that can report those sibships. It will become evident later that only the ascertainment of sibships with at least one affected sibling will contribute to our estimators for SRR so this violation will not be an issue. The selection units that are linked to each population element according to the counting rule is called the network for the population element. For example, under the AllSCR all living brothers or sisters from the same family is the network for their sibship (throughout this paper we will only consider sibships consisting of only full siblings and not half or adopted siblings). Since under these counting rules each sibship can be reported multiple times, where the number of reports for a sibship is the network size, larger sibships under the AllSCR and sibships with more affected members under the AffSCR have greater probabilities of being reported in the survey. Biased estimation of prevalence and recurrence risk of disease can result if these differential probabilities of reporting sibships under the counting rule are not accounted for in the estimation. A way to take account of this multiple reporting in the estimation is to down weight the observations of the reports of each sibship by the inverse of the number of selection units that can report the sibship. (This is one example of a multiplicity weight (Birnbaum and Sirken, 1965)). In our example of family aggregation of diabetes, the NHIS collected reports of the number of living sibs and number of living diabetic sibs from each sampled individual so that these network weights can be computed for our counting rules.

In the next section, two consistent estimators of SRR and of prevalence of disease, p_a, are presented corresponding to each of the two counting rules. Because we are using counting rules with network sizes that can exceed one the estimators utilizing these reports under these types of count rules are called network estimators in the literature (Sirken, 1998). To our knowledge, the network sampling estimators of SRR has not been investigated previously. In the next section our estimators are based on the reporting of affected status for the living sibs in a family. In the Appendix we present estimators based on the reporting of the affected status of both living and deceased siblings.

3. Estimation of the Sibling Recurrent Risk and Recurrent Risk Ratio

Sibling recurrence risk, K_s, has been formally defined in two ways that are equivalent (Olson and Cordell, 2000):

(1) The probability that a sibling of an affected individual is also affected i.e., let A be the set of affected individuals in the population and for a pair of siblings i, j this probability is $P(i \in A | j \in A)$.

(2) The proportion of affected individuals among all siblings of affected individuals in a population.

The second definition is population-based and lends itself to being applied to population surveys. Following the notation of Zou and Zhao (2004) we can define K_s for a finite population as

$$K_s = \frac{\sum_{s=2}^{\mathcal{S}} \sum_{a=1}^{s} a(a-1)N_{s(a)}}{\sum_{s=2}^{\mathcal{S}} \sum_{a=1}^{s} a(s-1)N_{s(a)}},$$

where $N_{s(a)}$ is the number of sibships of size $s = 2, \ldots, \mathcal{S}$, with a affected siblings in the population. Without delineating the sibship sizes in the population of all sibships, K_s can be expressed as

$$K_s = \frac{\sum_{j=1}^{T} a_j(a_j - 1)}{\sum_{j=1}^{T} a_j(s_j - 1)}, \quad (1)$$

where T is number of sibships in the population, a_j and s_j are the number of affected individuals in sibship j and the size of sibship j, respectively.

Suppose we have a national survey of $i = 1, \ldots, n$ randomly sampled individuals, which are the selection units as described in the previous section. In addition to the network weight mentioned previously, each sampled individual i has sample weight w_i, which is the inverse of the probability of including individual i in the sample. A consistent network estimator of K_s under the AllSCR is given by

$$\widehat{K}_{s1} = \frac{\sum_{i=1}^{n} \frac{w_i}{s_i} a_i(a_i - 1)}{\sum_{i=1}^{n} \frac{w_i}{s_i} a_i(s_i - 1)}, \quad (2)$$

where for each sampled individual i the reported sibship has a network weight that is specific to the AllSCR, which is the inverse of the number of selection units (individuals) that can report the same sibship, $1/s_i$. The numerator and denominator are unbiased estimators of the numerator and denominator of K_s, respectively, is given by (1).

Another quantity that is of interest in genetics is the sibling recurrence risk ratio, which is the ratio of K_s to the population prevalence of the disease, i.e., the proportion of the population that is affected, p_a, i.e., $\lambda_s \stackrel{\text{def}}{=} K_s/p_a$. Unlike most genetic studies, which obtain an estimate of p_a from an external source, we can use the survey to obtain an estimate p_a. If use the survey to estimate p_a, then the population prevalence is usually the proportion of individuals eligible to sampled in the survey. Since our definition K_s involves only living siblings, we estimate p_a among living siblings using the AllCR as

$$\hat{p}_{a1} = \frac{\sum_{i=1}^{n} \frac{w_i}{s_i} a_i}{\sum_{i=1}^{n} w_i}, \quad (3)$$

and the estimator $\hat{p}_{a1}$ is a network estimator of the prevalence of disease. We combine (2) and (3) to obtain a network estimator of λ_s as

$$\hat{\lambda}_{s1} = \widehat{K}_{s1}/\hat{p}_{a1}$$

$$= \frac{\sum_{i=1}^{n} \frac{w_i}{s_i} a_i(a_i - 1) \sum_{i=1}^{n} w_i}{\sum_{i=1}^{n} \frac{w_i}{s_i} a_i(s_i - 1) \sum_{i=1}^{n} \frac{w_i}{s_i} a_i} . \quad (4)$$

Next we provide estimators for K_s and λ_s under AffSCR.

$$\widehat{K}_{s2} = \frac{\sum_{i=1}^{n} \frac{w_i}{a_i} a_i(a_i - 1)\delta_i}{\sum_{i=1}^{n} \frac{w_i}{a_i} a_i(s_i - 1)\delta_i} = \frac{\sum_{i=1}^{n} w_i (a_i - 1)\delta_i}{\sum_{i=1}^{n} w_i (s_i - 1)\delta_i} \quad (5)$$

where the summand are set equal to zero if a_i is zero, δ_i an indicator variable equal to one if the ith sampled individual is affected and zero otherwise. In (5) each observation *i* has a network weight of $1/a_i$ in accordance with the AffSCR. An estimator of the recurrence risk ratio, λ_s is

$$\hat{\lambda}_{s2} = \widehat{K}_{s2}/\hat{p}_{a2} = \frac{\sum_{i=1}^{n} w_i (a_i - 1)\delta_i \sum_{i=1}^{n} w_i}{\sum_{i=1}^{n} w_i (s_i - 1)\delta_i \sum_{i=1}^{n} w_i \delta_i} , \quad (6)$$

where $\hat{p}_{a2} = \frac{\sum_{i=1}^{n} \frac{w_i}{a_i} a_i \delta_i}{\sum_{i=1}^{n} w_i} = \frac{\sum_{i=1}^{n} w_i \delta_i}{\sum_{i=1}^{n} w_i}$. It is noteworthy that under the AffSCR the network prevalence estimator, $\hat{p}_{a2}$, reduces to the conventional sample weighted prevalence estimator where each selected individual reports only about their own disease status. In the Appendix we consider estimators of K_s and λ_s for each counting rule among alive and deceased siblings.

4. Variance Estimation of Sibling Recurrent Risk and Recurrent Risk Ratio

Many national household health interview surveys such as the NHIS are cross-sectional surveys that have stratified multistage cluster sample designs. In this section we describe this type of sample design and the notation that will be used to express the variance estimation for estimators of K_s and λ_s. The population of individuals to be surveyed is partitioned into a set of primary sampling units (PSUs) and the PSUs are divided into *L* sampling strata. For household surveys the PSUs are usually geographically defined, e.g., counties, and the strata defined by demographic characteristics, e.g., population size of the PSUs. At the first stage of sampling, t_h PSUs are randomly sampled from each stratum *h, h*=1, …, *L*. There can be additional stages of sampling nested within the sampled PSUs to obtain a random sample of t_{hi} individuals from the i^{th} sampled PSU in stratum *h* where each sampled individual has a sample weight w_{hij}, j=1 , …, t_{hi}.

The estimators of K_s and λ_s in (2) and (4) can be re-expressed in terms of the sample strata and PSUs:

$$\widehat{K}_{s1} = \frac{\sum_{h=1}^{L} \sum_{i=1}^{t_h} \sum_{j=1}^{t_{hi}} \frac{w_{hij}}{s_{hij}} a_{hij}(a_{hij} - 1)}{\sum_{h=1}^{L} \sum_{i=1}^{t_h} \sum_{j=1}^{t_{hij}} \frac{w_{hij}}{s_{hij}} a_{hij}(s_{hij} - 1)}$$

and

$$\hat{\lambda}_{s1} = \frac{\sum_{h=1}^{L}\sum_{i=1}^{t_h}\sum_{j=1}^{t_{hi}} \frac{w_{hij}}{s_{hij}} a_{hij}(a_{hij}-1) \times \sum_{h=1}^{L}\sum_{i=1}^{t_h}\sum_{j=1}^{t_{hi}} w_{hij}}{\sum_{h=1}^{L}\sum_{i=1}^{t_h}\sum_{j=1}^{t_{hij}} \frac{w_{hij}}{s_{hij}} a_{hij}(s_{hij}-1) \times \sum_{h=1}^{L}\sum_{i=1}^{t_h}\sum_{j=1}^{t_{hij}} \frac{w_{hij}}{s_{hij}} a_{hij}}$$

We apply the Taylor linearization method to estimate the variances of $\widehat{K}_{s1}$ and $\hat{\lambda}_{s1}$ (Korn and Graubard, 1999). This approach uses a first order Taylor expansion to approximate the variances of $\widehat{K}_{s1}$ and $\hat{\lambda}_{s1}$as
$Var(\widehat{K}_{s1}) \approx Var\left[\sum_{h=1}^{L}\sum_{i=1}^{t_h}\sum_{j=1}^{t_{hij}} z_{hij}^{(K)}\right]$, where

$$z_{hij}^{(K)} = w_{hij}\widehat{K}_{s1}\left(\frac{\frac{a_{hij}(a_{hij}-1)}{s_{hij}}}{\sum_{h=1}^{L}\sum_{i=1}^{t_h}\sum_{j=1}^{t_{hij}} \frac{w_{hij}}{s_{hij}} a_{hij}(a_{hij}-1)} - \frac{\frac{a_{hij}(s_{hij}-1)}{s_{hij}}}{\sum_{h=1}^{L}\sum_{i=1}^{t_h}\sum_{j=1}^{t_{hij}} \frac{w_{hij}}{s_{hij}} a_{hij}(s_{hij}-1)}\right),$$

and $Var(\hat{\lambda}_{s1}) \approx \sum_{h=1}^{L}\sum_{i=1}^{t_h}\sum_{j=1}^{t_{hij}} z_{hij}^{(\lambda)}$, where

$$z_{hij}^{(\lambda)} = w_{hij}\hat{\lambda}_{s1}\left(\frac{\frac{a_{hij}(a_{hij}-1)}{s_{hij}}}{\sum_{h=1}^{L}\sum_{i=1}^{t_h}\sum_{j=1}^{t_{hij}} \frac{w_{hij}}{s_{hij}} a_{hij}(a_{hij}-1)} + \frac{1}{\sum_{h=1}^{L}\sum_{i=1}^{t_h}\sum_{j=1}^{t_{hij}} w_{hij}} - \frac{\frac{a_{hij}(s_{hij}-1)}{s_{hij}}}{\sum_{h=1}^{L}\sum_{i=1}^{t_h}\sum_{j=1}^{t_{hij}} \frac{w_{hij}}{s_{hij}} a_{hij}(s_{hij}-1)} - \frac{\frac{a_{hij}}{s_{hij}}}{\sum_{h=1}^{L}\sum_{i=1}^{t_h}\sum_{j=1}^{t_{hij}} \frac{w_{hij}}{s_{hij}} a_{hij}}\right).$$

Under the assumption that the sample of PSUs is approximately a stratified sample PSUs with replacement, which is often accurate since the sampling fraction for sampling PSU's is usually small, the variance of $\widehat{K}_{s1}$ can be estimated as

$$\widehat{var}(\widehat{K}_{s1}) \approx \widehat{var}\left[\sum_{h=1}^{L}\sum_{i=1}^{t_h}\sum_{j=1}^{t_{hij}} z_{hij}^{(K)}\right] = \sum_{h=1}^{L}\sum_{i=1}^{t_h}\left[z_{hi}^{(K)} - \bar{z}_{h}^{(K)}\right]^2$$

where $z_{hi}^{(K)} = \sum_{j=1}^{t_{hi}} z_{hij}^{(K)}$ and $\bar{z}_h^{(K)} = \frac{1}{t_h}\sum_{i=1}^{t_h} z_{hi}^{(K)}$ and similarly the variance of $\hat{\lambda}_{s1}$ can be estimated replacing $z_{hij}^{(K)}$ by $z_{hij}^{(\lambda)}$ in the above expressions.

5. Illustration using the 1976 National Health Interview Survey Diabetes Supplement

The sample survey estimators of K_s under the AffSCR and AllSCR presented in this paper are illustrated with data from the 1976 NHIS and its Diabetes Supplement. The 1976 NHIS collected a US national cross-sectional random sample of civilian noninstitutionalized individuals of all ages, using a stratified multistage clustered probability sample design (Black, 1978). For purposes of variance estimation the sample design is approximated by the sampling of two (pseudo-)PSUs from 149 (pseudo-) sampling strata (NCHS, 2009). There is a sample weight for each of the 113,178 respondents, which reflects the probabilities for selecting each respondent and adjustments for differential nonresponse and poststratification to US population sizes.

In the Diabetes Supplement the following questions (paraphrased) were asked of each sampled respondent:

Q 1. Do you have diabetes?

Q 2. How many living brothers and sisters do you have? (Do not count adopted, step or half brothers and sisters)

Q 3. How many of these brothers and sisters have diabetes?

Q 4. How many of your brothers and sisters are no longer living?

Q 5. How many of these brothers and sisters had diabetes?

Q6. Does (did) your mother have diabetes?

Q.7 Does (did) your father have diabetes?

The responses to these questions were used to obtain the diabetes status of each responding individual and his/her siblings and the sibship size separately for the alive and deceased siblings, and to obtain the diabetes status of the parents.

Table 1 shows for each of the two counting rules the number enumerated individuals in the 1976 NHIS who have different diabetic statuses among their sibling plus themselves. The rows with ≥ 1 diabetics are the sample sizes of responding

Table 1. Number of enumerated individuals in the 1976 National Health Interview Survey reporting different types diabetic statuses of their sibships by counting rule.

Counting Rule	Diabetic Status of Proband Plus Sibship	Number[1]
Any Sibling	≥ 1 living diabetic	5,831
	≥ 2 living diabetics	1,112
Affected Sibling	≥ 1 living diabetic	2,005
	≥ 2 living diabetics	571
Any Sibling	≥ 1 living or deceased diabetic	6,914
	≥ 2 living or deceased diabetics	1,533
Affected Sibling	≥ 1 living or deceased diabetic	2,197
	≥ 2 living or deceased diabetics	712
Any Sibling	≥ 1 living diabetic, and ≥ 1 parental diabetics	1,976
	≥ 2 living diabetics, and ≥ 1 parental diabetics	536
Affected Sibling	≥ 1 living diabetic, and ≥ 1 parental diabetics	640
	≥ 2 living diabetics, and ≥ 1 parental diabetics	267
Any Sibling	≥ 1 living diabetic or deceased, and ≥ 1 parental diabetics	2,239
	≥ 2 living diabetics or deceased, and ≥ 1 parental diabetics	693
Affected Sibling	≥ 1 living diabetic or deceased, and ≥ 1 parental diabetics	688
	≥ 2 living diabetics or deceased, and ≥ 1 parental diabetics	317

[1] Among reported sibships with ≥ 2 living siblings for reported sibships involving only living diabetics and sibships with ≥ 2 living or deceased siblings for reported sibships involving living or deceased diabetics.

individuals with nonmissing data that are utilized for the denominators of the estimators for SRR and the rows with ≥ 2 diabetics are the sample sizes that are utilized in the numerators for SRR in Table 2. We can see the reduction in sample sizes when using the AffSCR compared to the AllSCR, and the increase in sample sizes when including deceased siblings in the computation of the SRR.

Table 2 presents the estimates of the prevalence, SRR, SRR ratio and their relative standard errors (i.e., the standard error for an estimate divided by corresponding estimate), RSE, under the AffCR and AllCR for reporting the diabetes status of living siblings and living or deceased siblings. Also, estimates are given for sibships when at

Table 2. Estimates of prevalence, sibling recurrence risk, and sibling recurrence risk ratio of diabetes by counting rule from the 1976 U.S. National Health Interview Survey.

Counting Rule	Prevalence % (RSE[1] %)	Sibling Recurrence Risk % (RSE %)	Sibling Recurrence Risk Ratio (RSE %)
Reporting Living Siblings			
Affected Sibling	2.16 (2.61)	12.51 (4.56)	5.79 (5.12)
All Sibling	1.97 (1.79)	15.10 (4.03)	7.66 (3.76)
Reporting Living or Deceased Siblings			
Affected Sibling	2.34 (2.43)	13.75 (4.43)	5.87 (4.38)
All Sibling	2.16 (1.72)	15.37 (3.89)	7.10 (3.46)
Reporting Living Siblings Among Sibships with Parental Diabetes			
Affected Sibling	6.13 (4.19)	18.08 (6.26)	8.37[2] (6.74)
All Sibling	6.05 (2.74)	20.93 (5.20)	10.62[2] (5.04)
Reporting Living Siblings Among Sibships without Parental Diabetes			
Affected Sibling	1.68 (2.89)	9.50 (5.99)	4.40[2] (6.37)
All Sibling	1.47 (2.05)	11.45 (5.27)	5.81[2] (5.62)
Reporting Living Siblings or Deceased Siblings Among Sibships with Parental Diabetes			
Affected Sibling	6.51 (3.86)	19.65 (5.60)	8.39[3] (5.74)
All Sibling	6.43 (2.83)	21.87 (4.33)	10.10[3] (4.05)
Reporting Living Siblings or Deceased Siblings Among Sibships without Parental Diabetes			
Affected Sibling	1.83 (2.69)	11.11 (6.40)	4.74[3] (6.30)
All Sibling	1.64 (1.96)	12.14 (6.18)	5.61[3] (5.89)

[1]Relative standard error

[2]Prevalences from the general population were used to compute the sibling recurrence risk ratio, 2.16 and 1.97 for the affect sibling counting rule and all sibling counting rule, respectively.

[3]Prevalences from the general population were used to compute the sibling recurrence risk ratio, 2.34 and 2.16 for the affect sibling counting rule and all sibling counting rule, respectively.

least one parent had diabetes. The estimates of the prevalence of diabetes were slightly higher under the AffSCR than under the AllSCR indicating possibly more accurate reporting from affected probands under the AffSCR than from a mix of affected and unaffected probands under the AllSCR, assuming there is no overreporting of diabetes. Also, the prevalence of diabetes was about three times larger for sibships where at least

one parent was reported to have diabetes, indicating the genetic association with development of diabetes. The estimates of SRR and SRR ratios were smaller for the AffSCR compared to the AllSCR. The smaller estimates for AffSCR could be due to more sibships reported with only one diabetic under the AffSCR than under the AllSCR since the probands are affected under this counting rule. In fact the data from the NHIS indicates this. A greater number of sibships with only one affected makes the estimate of SRR smaller. As expected these ratios increase considerably when at least one parent was reported to have diabetes. A comparison of the RSE's shows that they are smaller for the AllSCR than the AffSCR. This gain in efficiency is due to more individuals reporting diabetics under the AllSCR than the AffSCR, which is described in Sirken (1998).

6. Discussion

Population-based surveys that randomly sample individuals in the population can be used to unbiasedly ascertain the sibships of the surveyed individuals and obtain information about the disease status of each member of the ascertained sibships. Using appropriate counting rules that link individuals enumerated in the survey to their sibships, we showed that additional data about the number of individuals who can report the ascertained sibships can be collected during the survey interviews to appropriately down weight the disease status information of the sibships for possible multiple reporting. We provide design-unbiased estimators for SRR and SRR ratios based on two counting rules. A counting rule where either affected or unaffected individuals can report the disease statuses of themselves and their siblings and a counting rule where only affected individuals enumerated in the survey can report about the disease statuses of themselves and their siblings. Thus, population-based surveys can provide unbiased ascertainment of sibships and disease statuses of the individuals in the sibships that geneticists can use to accurately estimate SRR and SRR ratios to determine the extent of family aggregation of various disease. Using the Diabetes Supplement to the 1976 NHIS, we were able to illustrate our estimation methods for calculating SRR and SRR ratios for diabetes in the US population.

Our methods assume accurate reporting of disease statuses of siblings and reporting of data used for the down weighting in the estimation by enumerated individuals in the survey. As we observed from the diabetes illustration, the estimates of prevalence, SRR and SRR ratios differed by counting rule and also by whether or not the reported siblings were deceased. These differences could reflect the accuracy of the reported data, which may be a function of the counting rule that determines who is eligible to report the sibship information and the disease status of the siblings. Further theoretical and empirical work is needed to develop counting rules and estimators that are robust to reporting error while maintaining good statistical efficiency. For example hybrid counting rules that combine existing counting rules, e.g., the counting rule used for the AffCR for reporting alive and deceased siblings, may be one approach for obtaining improved estimation.

REFERENCES

Black, ER. 1978. Current estimates from the Health Interview Survey United States-1976. Vital Health Stat 10;(119):1-80.

Christakis NA, Fowler JH. 2007. The spread of obesity in a large social network over 32 years. N Engl J Med. 357(4):370-9

Christakis NA, Fowler JH. 2008. The collective dynamics of smoking in a large social network. N Engl J Med. 358(21):2249-58.

Guo SW. 1998. Inflation of sibling recurrence-risk ratio, due to ascertainment bias and/or overreporting. Am J Hum Genet. 63(1):252-8.

Korn EL, Graubard BI. 1999. *Analysis of Health Surveys*. John Wiley & Sons, New York.

NCHS. 2009. Variance estimation for the 1973-84 NHIS public use person data. http://www.cdc.gov/nchs/data/nhis/7384var.pdf.

Olsen JM, Cordell JM. 2000. Ascertainment bias in the estimation of sibling genetic risk parameters. Genetic Epidemiology 18: 217-235.

Risch N. 1990a. Linkage strategies for genetically complex traits. I. Multilocus models. Am J Hum Genet. 46(2):222-8.

Risch N. 1990b. Linkage strategies for genetically complex traits. II. The power of affected relative pairs. Am J Hum Genet. 46(2):229-41.

Risch N. 1990c. Linkage strategies for genetically complex traits. III. The effect of marker polymorphism on analysis of affected relative pairs. Am J Hum Genet. 46(2):242-53.

Rosenquist JN, Murabito J, Fowler JH, Christakis NA. 2010. Ann Intern Med. 152(7):426-33.

Sirken MG, Graubard BI, and McDaniel M. 1978. National network surveys of diabetes. In: 1978 Proceedings of the American Statistical Association of the Section on Survey Research Methods. American Statistical Association. 631-35.

Sirken, MG. 1997. Network sampling. In Encyc. of Biostat. 2977–2986. Wiley, New York.

Sutton R. 2011. Referring patients for a medical genetics consultation and genetic counseling. Adv Otorhinolaryngol. 70:25-7.

Zou G, Zhao H. 2004. The estimation of sibling genetic risk parameters revisited. Genetic Epidemiology 26: 286-293.

Appendix: Estimation that includes Deceased Siblings.

Two estimators of K_s using the AllSCR and AffSCR in which the affected status of both living and deceased sibs are reported are given by

$$\widehat{K}_{s1}^{*} = \frac{\sum_{i=1}^{n} \frac{w_i}{s_i} a_i^{*}(a_i^{*} - 1)}{\sum_{i=1}^{n} \frac{w_i}{s_i} a_i^{*}(s_i^{*} - 1)}$$

and

$$\widehat{K}_{s2}^{*} = \frac{\sum_{i=1}^{n} w_i \left[I_i \frac{\delta_i a_i^{*}(a_i^{*}-1)}{a_i} + (1-I_i)\frac{a_i^{*}(a_i^{*}-1)}{s_i} \right]}{\sum_{i=1}^{n} w_i \left[I_i \frac{\delta_i a_i^{*}(s_i^{*}-1)}{a_i} + (1-I_i)\frac{a_i^{*}(s_i^{*}-1)}{s_i} \right]},$$

respectively, where a_i^{*} are the number of affected siblings who are living or deceased, s_i^{*} are the sibship sizes for the living and deceased siblings for the sibship reported by the ith sampled individual, a_i are the number of affected individuals among the living sibs, I_i is an indicator variable that is equal 1 when $a_i \neq 0$ and equal to zero when $a_i = 0$. The purpose of the I_i in $\widehat{K}_{s2}^{*}$ is to permit the use of a hybrid counting rule where the AffSCR is used for enumerated individuals that report $a_i \neq 0$ and the AllSCR is used when for enumerated individuals that report $a_i = 0$. If we had used only the AffSCR rule for all enumerated individuals then the estimator would be biased because for sibships where all affected siblings are deceased these sibships would not be counted in $\widehat{K}_{s2}^{*}$. Even though the counting rule for $\widehat{K}_{s2}^{*}$ is a hybrid of the AllSCR and AffSCR , for simplicity we refer to this counting rule as the AffSCR.

Since our definition K_s includes reports of siblings alive or deceased, we estimate p_a using the AllSCR to obtain

$$\hat{p}_1^{*} = \frac{\sum_{i=1}^{n} \frac{w_i}{s_i} a_i^{*}}{\sum_{i=1}^{n} \frac{w_i}{s_i} s_i^{*}}$$

and using the (hybrid) AffSCR to obtain

$$\hat{p}_2^{*} = \frac{\sum_{i=1}^{n} w_i \delta_i \left[I_i \frac{a_i^{*}}{a_i} + (1-I_i)\frac{a_i^{*}}{s_i} \right]}{\sum_{i=1}^{n} \frac{w_i}{s_i} s_i^{*}}$$

The estimators for λ for each of the counting rules are

$$\hat{\lambda}_{s1}^{*} = \frac{\sum_{i=1}^{n} \frac{w_i}{s_i} a_i^{*}(a_i^{*}-1)}{\sum_{i=1}^{n} \frac{w_i}{s_i} a_i^{*}(s_i^{*}-1)} \times \frac{\sum_{i=1}^{n} \frac{w_i}{s_i} s_i^{*}}{\sum_{i=1}^{n} \frac{w_i}{s_i} a_i^{*}}$$

and

$$\hat{\lambda}_{s2}^{*} = \frac{\sum_{i=1}^{n} w_i \left[I_i \frac{\delta_i a_i^{*}(a_i^{*}-1)}{a_i} + (1-I_i)\frac{a_i^{*}(a_i^{*}-1)}{s_i} \right]}{\sum_{i=1}^{n} w_i \left[I_i \frac{\delta_i a_i^{*}(s_i^{*}-1)}{a_i} + (1-I_i)\frac{a_i^{*}(s_i^{*}-1)}{s_i} \right]} \times \frac{\sum_{i=1}^{n} \frac{w_i}{s_i} s_i^{*}}{\sum_{i=1}^{n} w_i \left[I_i \frac{w_i \delta_i a_i^{*}}{a_i} + (1-I_i)\frac{a_i^{*}}{s_i} \right]}.$$

The Taylor linearization variance estimators for $\hat{p}_1^{*}, \hat{p}_2^{*}, \widehat{K}_{s1}^{*}\ \widehat{K}_{s2}^{*}, \hat{\lambda}_{s1}^{*}$, and $\hat{\lambda}_{s2}^{*}$ are obtained in the same fashion as described in Section 4 for estimators for only living siblings.