# Identifying the Zygosity Status of Twins Using Bayes Network and Estimation-Maximization Methodology

*Yicun Ni (ID#: 9064804041),*

*Jin Ruan (ID#: 9070059457),*

*Ying Zhang (ID#: 9070063723)*

## Abstract

As the renaissance of family-based genomic research, identifying diseases and families for this kind of study becomes increasingly important. Recently, a large cohort of predicted twins, called the Marshfield Clinic Twin Cohort, has been generated by analyzing standard demographic information in an electronic medical record. However, one piece of critical information, the zygosity status (identical vs. fraternal) of each twin pair is missing in this data set. In this study, we used the Bayes network and estimation-maximization method to infer whether a certain pair of twins is monozygotic or dizygotic. By properly choosing the initial conditional probability table entries, we successfully reproduced the known priors of the probabilities for identical and fraternal twins. Armed with these imputed data, we also studied the p-values of most interesting diseases, and their relative risks in identical and fraternal twins. Our study demonstrates that such a methodology is very useful in filling the missing values in a family-based phenotype data set, and thus eliminates many difficulties in family-based research.

## I. Introduction

In the last decade, population-based approaches in human genomic research have become widespread and achieved great success, *e.g*. genome-wide association studies (GWASs). Since 2005, over 3000 genetic variants have been identified and linked to hundreds of diseases.[1] Unfortunately, population-based genomics has its own difficulties. First, many associations between genome and diseases have little significance in clinic. This is probably caused by the fact that population-based methods usually focus on genomes of unrelated individuals instead of, for instance, family members. Second, the functions of a large amount of genomes are difficult to understand. Although there are some improvements on the traditional population-based studies by considering the electronic medical record (EMR) systems of patients,[2-4] such as phenome-wide association studies (PheWASs), some of the difficulties said above still persist.[3]

Because of these difficulties, and the great development of the "next generation sequencing" (NGS) technologies in the recent a few years,[5-6] a previously popular approach, family-based genomics starts to revitalize. However, even with this method, associating the most interesting diseases with families for research is still challenging. Alternatively, the EMR systems may provide another kind of information to help understand the heritability of many diseases. Recently, a large cohort of predicted twins has been successfully generated, by using familial aggregation analysis on nothing more than standard demographic information in an EMR. This cohort is called the Marshfield Clinic Twin Cohort (MCTC).[7] Thousands of diseases are included in this cohort and can be studied simultaneously. Ye *et al*. then investigated the correlation between these diseases and the gender status (same-sex vs. opposite-sex) of twins, and led to the conclusion that heritable factors are more likely to contribute to the etiology of the diseases with greater relative risks.[8]

However, investigating gender status only focuses on those diseases that have inheritance on X or Y chromosomes. To better understand the family genomic conditions for thousands of disease captured by EMR systems, it is necessary to consider the zygosity status (fraternal vs. identical) of twins. Unfortunately, this information is usually missing in an EMR.[7-8] Therefore, in this study, we use Bayes network (BN) and estimation-maximization (EM) methodology[9-10] to infer the missing zygosity status of all twins, which provides the availability of using an EMR to better understand the inheritances of most interesting disease.

The paper is organizes as follows. In section II, the description of the methods used in this study is given. Section III presents the results and data analysis. And we conclude in Section IV.

## II. Methods

### A. Marshfield Clinic Twin Cohort

The MCTC was generated by exclusively analyzing the standard demographic information in EMR system. There are 9906 defined diseases and 15706 identified patients in this data set. Among these patients, most of them are twin pairs from the same families, and the rest few are individuals without other relatives in their own families. In this study, we are interested in the association between the family genomic conditions and diseases. Therefore, only families with twins are included in the present work, and those families without twins and the diseases only exist in them are excluded from further analysis.

The modified cohort data set contains 9817 defined diseases, and 15174 identified individuals or 7587 families. Among these individuals, 7952 patients are male and 7222 patients are female. With respect to the families, 5021 of them have twins with the same gender, where 2693 pairs of twins are male and 2328 pairs of twins are female. The other 2566 twin pairs are those with opposite genders.

### B. Bayes Network Structure

Bayes Network (BN) is a probabilistic graphical model that represents a set of random variables and their conditional dependencies via a directed acyclic graph (DAG). Each random variable is represented by a vertex on the DAG and has an associated conditional probability table (CPT). A directed edge in BN describes a direct correlation from the predecessor to the successor. The CPT of a vertex exhibits the conditional probabilities of the values of the random variable, given the values taken on all its predecessor nodes. It is widely used to model many real-world problems such as the probabilistic relationships between diseases and symptoms. Therefore, it makes BN a perfect tool to model the MCTC data set. Inference and learning can be performed in BN readily, and we mainly focus on the parameter learning here.

There are many efficient algorithms to do the structure learning in BN, however, we fix the structure of the BN (the DAG) in the present study. There are three kinds of vertices in the BN: the zygosity status (ZS), the gender agreement (GA), and the diseases (D). The ZS is the “class” feature which has no parent node, and it is also the missing data we are intent to infer. The ZS feature has two values: fraternal (F) if the twins are dizygotic or identical (I) if they are monozygotic. The GA vertex has ZS vertex as its parent, and can take one of the two values: yes (Y) if the twins have a same gender, or no (N) if they have opposite genders. For the diseases, each one of them is a separate vertex. Although some of the diseases studied here are gender related and the others are not, we treat all of them in a uniform manner by relating each disease to the zygosity status and gender, and leave the algorithm itself to decide which ones are more likely to occur in one gender. Therefore, for disease *i* ($D_i$), it has both ZS and GA as the parents, and has three possible values: 0 if none of the twins are affected by this disease, 1 if one of them is affected, and 2 if both of them have the disease. A scheme of the BN structure is shown in Figure 1.

### C. The Estimation-Maximization Algorithm

One of the goals of BN is to perform the parameter learning, which means to infer the parameters of the CPTs given the structure of a BN and a training set. If a fully observed training set is given, one can use methods such as maximum likelihood estimation (MLE) or maximum a posteriori (MAP) to calculate those parameters. However, it is common that in a machine learning task, some feature values are missing, in which case these standard methods fail to work. Therefore, in order to do parameter learning on partially observed data, one has to first impute these missing values, and one way to do it is the EM methodology. The main idea of the EM method is to compute the expectation of missing data, and to find the best model that maximizes the probability of the data given an initial model. There are many ways to do EM algorithm, in this study, we implement it through inference and parameter learning.

More specifically, the EM algorithm is seeking to estimate

$$\theta \leftarrow \underset{\theta}{\operatorname{argmax}} E_{Z|X,\theta}[\log P(X,Z|\theta)] \qquad (1)$$

where, $\theta$ is the parameters of the model, which in our case, is the CPTs, $X$ represents all observed variable values, and $Z$ represents all missing values. The algorithm consists of the iteration of the following two steps:[11]

E step: Use observed data X and current parameters $\theta$ to calculate $P(Z|X, \theta)$,

M step: Replace the current $\theta$ by $\theta \leftarrow \operatorname{argmax}_{\theta'} E_{Z|X,\theta}[log\ P(X,Z|\theta')]$.

The computation stops if the iteration of E-M steps converges. The stopping criterion employed in this study is that the change of total counts for fraternal (or equivalently, identical) twins in two successive iterations is less than one. Although the EM algorithm is guaranteed to converge, it can only find the local optimum, which means that choosing the initial parameters are important. We will discuss more about this in the next section.

### D. Relative Risk

Once the EM algorithm is converged, we obtain the CPTs interested in the BN model, and the prediction of probabilities of monozygotic and dizygotic for each family. Therefore, we can classify the zygostic status of each pair of twins by choosing the more probable one. Once this piece of information is available to us, we can investigate the relation between the zygostic status and the etiology of the diseases.

However, to do this, it would be better if we can first categorize all the diseases base on some proper quantitative metrics. Here, we choose to use the relative risk (RR) for each disease, which measures the relative ratio of the probability that an individual is affected by the disease given the other twin is already affected, to the probability that an individual is affected in general. The RR of the disease $i$ is therefore computed as follows,

$$RR(D_i) = \frac{\# \; of \; concordant \; affected \; families / \# \; of \; affected \; families}{\# \; of \; affecteds \; / total \; population} \qquad (2)$$

### E. P-values

The p-value[12] of each disease is also calculated in the present study. We conducted a two-sample t-test for the differences in concordance rates between identical and fraternal twins for each disease. First, for any disease, if both twins are affected, we treated it as "success", which has the value 1; and if only one twin is affected, it is a "failure", which has the value 0. Thus, the probability of "success" for each group of twins (identical/fraternal) equals the mean, which is exactly the concordance rate. Assuming that identical and fraternal twins having different variances, we calculate the t-statistic using

$$T = \frac{\mu_1 - \mu_2}{\sqrt{s_1^2/n_1 + s_2^2/n_2}} \qquad (3)$$

where $\mu_1, \mu_2$ are the sample means, $s_1^2, s_2^2$ are the sample variances, $n_1, n_2$ are the sample sizes for identical and fraternal twins.

The null hypothesis and alternative hypothesis are defined as $H_0: \mu_1 = \mu_1$ and $H_a: \mu_1 \neq \mu_1$ respectively. Thus, the t-test is two-tailed and the p-value is calculated as

$$p = Prob(|t_v| > |T|) \qquad (4)$$

where $T$ is calculated from Eqn (3) and $t_v$ is distributed according to t-distribution with degree ν, which is given by

$$v = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{(s_1^2/n_1)^2/(n_1 - 1) + (s_2^2/n_2)^2/(n_2 - 1)} \qquad (5)$$

where all the symbols have the same meaning as they are in Eqn (3).

## III. Analysis

### A. Choice of the initial CPTs

As mentioned earlier, the EM algorithm converges to a local maximum. Therefore, choosing reasonable initial CPTs is essential for this parameter learning task. Moreover, there are some constraints that the CPTs are subject to in this particular case. For example, given a pair of twins as identical, it is one hundred percent that they are the same gender. It is important that the CPTs satisfy these constraints in each step of iterations in the EM algorithm.

In the very first attempt, we use the same predefined CPT as the initial parameters for all the diseases. One of the examples is given in Table 1. Since MAP is employed in the maximization step, we also exhibit pseudocounts in parentheses in Table 1. The EM algorithm with this set of initial values leads to the priors of 53% of fraternal twins and 47% identical twins. However, it is known that in MCTC, 75% of twins are fraternal and 25% are identical.[13] Therefore, our estimation does not agree with the true data well, which suggests that the initial CPT entries are not appropriate. Changing those initial values in a certain range does not make the EM algorithm converge to another maximum thus shows no improvement. Therefore, it seems that using a same initial CPT for every disease is not proper.

To solve this problem, we adopt a "baseline" approach in this study to generate the initial CPT entries for each node in the BN. First we estimate the conditional probability for being dizygotic given the twins share the same gender. This is achieved by using Bayes rule as follows $P(ZS = F|GA = Y) = P(GA = Y|ZS = F)P(ZS = F)/P(GA = Y)$. $P(GA = Y|ZS = F)$ equals 0.5 which is a biological fact. $P(ZS = F)$ is known to be 0.75. And $P(GA = Y)$ can be calculated by MLE from the data set, which is 0.66. Therefore, $P(ZS = F|GA = Y)$ equals 0.57. Then, we use this value as a partial count to assign a pair of opposite-gender twins as 0.57 fraternal and 0.43 identical (same-gender twins are counted as 1 fraternal twin pair for certain). With this information, we compute a "baseline" CPT for each node and use them as the initial parameters for the EM algorithm. With the same pseudocounts shown in Table 1, this time, the EM converges to priors of 81% fraternal twins and 19% identical twins, which is much closer to the known probabilities (75% and 25% respectively). Tuning this partial count in a range from 0.52 to 0.75 does not introduce significant change to the final priors, although using values smaller than 0.5 may make the algorithm converge to other optima. Therefore, we decide that this approach is more appropriate to make initial guesses for parameters, and use its predictions for further analysis.

### B. Analysis for p-values

P-values of 5048 diseases are calculated, where 4769 diseases out of the total 9817 are excluded, since they satisfy the condition that $n_1 < 2$ or $n_2 < 2$. In general, 318 out of 5048 diseases have p-values smaller than 0.05, which means that there is significant difference in concordance rates between identical and fraternal twins for these diseases (significant level: 0.05). Figure 2 is the Manhattan plot showing the p-values of the t-tests for the differences between the concordance rates of identical and fraternal twins for all these diseases. The p-values of some diseases equal zero, which are replaced by a very smaller number (1E-20) in Figure 2 because of the easiness for plotting. These are the points that have the largest y-axis value (20) in Figure 2.

### C. Analysis for Relative Risk

We next calculate the relative risks (RRs) for all phenotypes, and classify them into 10 categories. More specifically, there are 7325 phenotypes having RR=0, 34 phenotypes having 0<RR<1, 254 phenotypes having 1<RR<2, 257 phenotypes having 2<RR<3, 168 phenotypes having 3<RR<4, 123 phenotypes having 4<RR<5, 363 phenotypes having 5<RR<10, 828 phenotypes having 10<RR<100, 343 phenotypes having 100<RR<1000, and 122 phenotypes having RR>1000. Among them, diseases whose RR values equal zero are mostly consisted of the rarest conditions. Therefore, these phenotypes are excluded in the following analysis.

To better assess if these diseases have any unappreciated genetic etiologies, the percentage of affected families that consisted of concordant pairs is measured in identical twins and compared to fraternal twins. The result is shown in Figure 3. Comparing to a similar study conducted by Ye *et al*, where the comparison is between same-sex twins and opposite-sex twins, our result exhibits surprisingly different trend. The category with the largest enrichment for identical twins comparing to fraternal twins is the one with RR<1. For the categories with 1<RR<2, 2<RR<3, and 3<RR<4, there is very little difference between identical and fraternal twins. Interestingly, the percentage of concordant fraternal twins is higher than that of identical twins in categories with 4<RR<5, 5<RR<10, 10<RR<100, 100<RR<1000, and RR>1000, where the largest difference exists in category 4<RR<5 and it slightly decreases as RR grows. This observation strikingly contrasts the result shown in Figure 2B in Ref. 8, where the percentage of concordant same-sex twins is always larger than that of opposite sex twins, and the enrichment of concordant families with same-sex twins increases as the RR value increases. The reason for this discrepancy is currently still unclear. However, one thing worth to notice is that by using the MCTC data set, we cannot quantitatively reproduce the result between same-sex twins and opposite-sex twins reported in Ref. 8. For example, in the category of RR>1000, the enrichment of same-sex twins is 4.2 fold comparing to opposite-sex twins in Ref. 8, but it is 1.8 fold in our reproduction. We suspect that this disagreement may be caused by the fact that only ~5000 diseases in MCTC are considered in Ref 8, but all the 9817 diseases are included in our study. Therefore, further clarification of data analysis is necessary before the result can be properly interpreted comparing identical and fraternal twins.

## IV. Conclusion

Family-based genome research has been starting to regain more and more attention in recent years due to the limitation of population-based approaches and the development of sequencing techniques. With large EMR systems quickly becoming popularized, thousands of pedigrees may be applied to phenome-wide research. However, some important information about the families and patients may be missing in an EMR. This study demonstrates an example of how to infer these missing values. By investigating the MCTC data set, we successfully impute the zygosity status of twin families by using Bayes network and estimation-maximization methodology. With properly chosen initial parameters, the resulting priors produced by the algorithm agree with the *a priori* knowledge of the data set. Furthermore, the association between the zygosity status and various diseases is also studied, which may provide invaluable information to the clinic.

Another large piece of useful information yet not analyzed here is the CPTs of the diseases in the BN. This kind of knowledge may help physicians diagnose a disease for a twin given whether the other twin (or another family member in general) is affected by the same disease. Other machine learning techniques, such as text mining, may also be applied to identify identical/fraternal twins, which can provide a comparison to the predictions from the BN&EM algorithm. Other directions of future work include altering the structure of the BN, or introducing new random variables to the BN.

## References

1. Welter, D. *et al*., The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic acids research*, **42**, D1001-1006 (2014).
2. Denny, J. C. *et al*., PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. *Bioinformatics*, **26**, 1205-1210 (2010).
3. Hebbring, S. J., The challenges, advantages and future of phenome-wide association studies. *Immunology*, **141**, 157-165 (2014).
4. Namjou, B. *et al*., Phenome-wide association study (PheWAS) in EMR-linked pediatric cohorts, genetically links PLCL1 to speech language development and IL5-IL13 to Eosinophilic Esophagitis. *Frontiers in genetics*, **5**, 401 (2014).
5. Green, R. C. *et al*., ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing. *Genetics in medicine: official journal of the American College of Medical Genetics*, **15**, 565-574 (2013).
6. Yang, Y. *et al*., Clinical whole-exome sequencing for the diagnosis of Mendelian disorders. *The New England journal of medicine*, **369**, 1502-1511 (2013).
7. Mayer, J. *et al*., Use of an electronic medical record to create the Marshfield clinic twin/multiple birth cohort. *Genetic epidemiology*, **38**, 692-698 (2014).
8. Ye, Z. *et al*., Large-scale phenome-wide scan in twins, *to be submitted*
9. Heckerman, D., A tutorial on learning with Bayesian network. Technical Report MSR-TR-95-06, Microsoft Research Advanced Technology Division (1996).
10. Page, C. D., <http://pages.cs.wisc.edu/~dpage/cs760/BNall.pdf> (2015).
11. Mitchell, T., <http://www.cs.cmu.edu/~tom/10701_sp11/slides/GrMod3_2_17_2011-ann.pdf> (2011).
12. NIST/SEMATECH e-Handbook of Statistical Methods, <http://www.itl.nist.gov/div898/handbook/eda/section3/eda353.htm>, (2013).
13. Hebbring, S. J., *private communication*, (2015).

a)

| Condition | | D | | |
|---|---|---|---|---|
| ZS | GA | 0 | 1 | 2 |
| F | Y | 0.9 (0.8) | 0.07 (0.1) | 0.03 (0.1) |
| F | N | 0.9 (0.8) | 0.08 (0.1) | 0.02 (0.1) |
| I | Y | 0.9 (0.8) | 0.05 (0.1) | 0.05 (0.1) |
| I | N | 0 (0) | 0 (0) | 0 (0) |

b)

| Condition | GA | |
|---|---|---|
| ZS | Y | N |
| F | 0.5 (0.5) | 0.5 (0.5) |
| I | 1 (1) | 0 (0) |

c)

| ZS | |
|---|---|
| F | I |
| 0.5 (0.5) | 0.5 (0.5) |

Table 1. An example of initial CPTs. In the parentheses are the pseudocounts. a) disease node. b) gender-agree node. c) zygosity-status node.

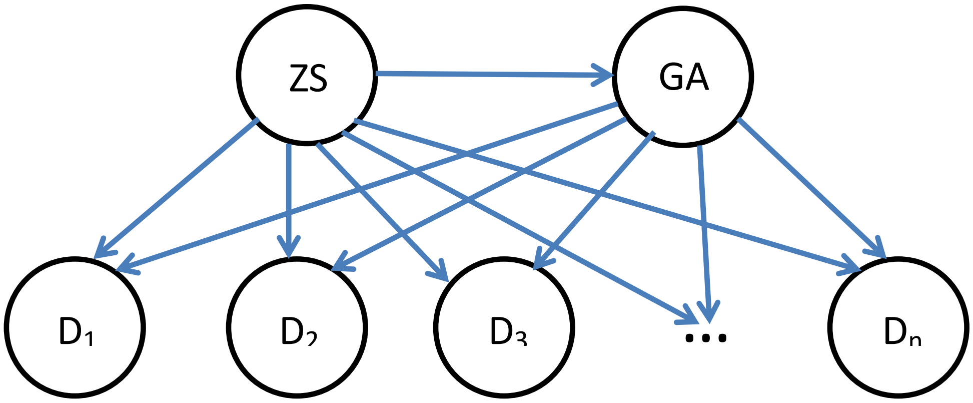

Figure 1. Structure of the Bayes network

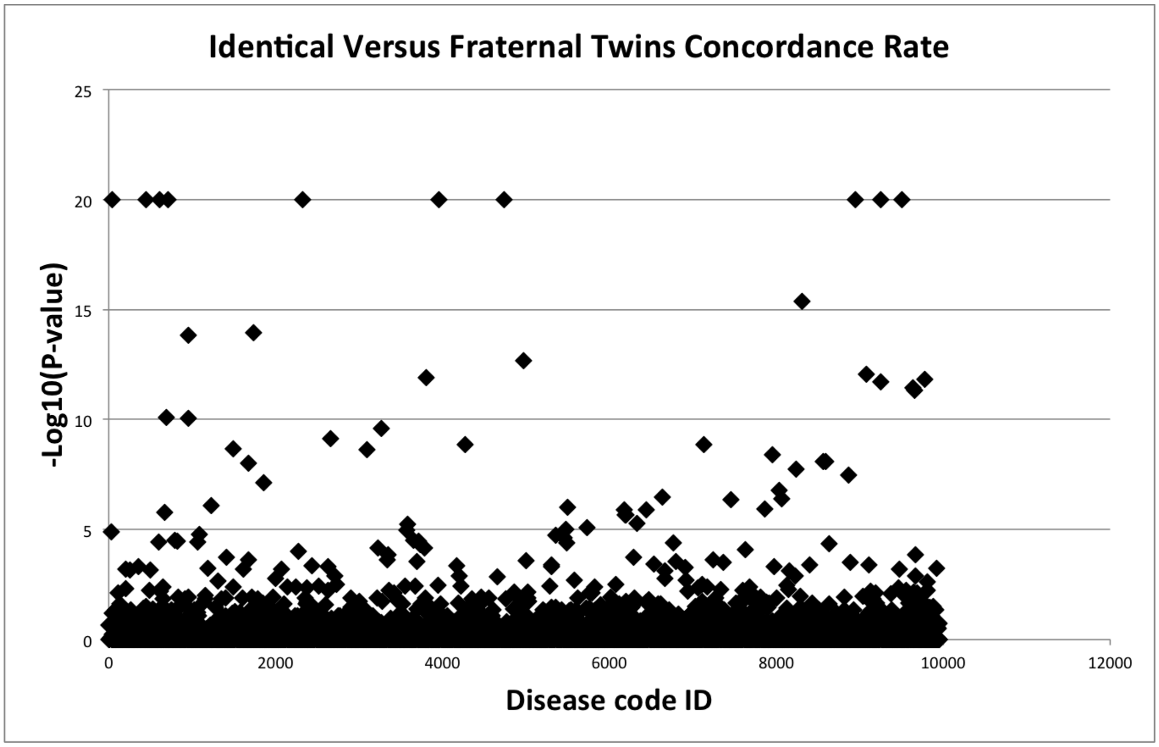

Figure 2. The Manhattan plot showing the differences of p-values between the concordance rates of identical and fraternal twins for 5048 diseases

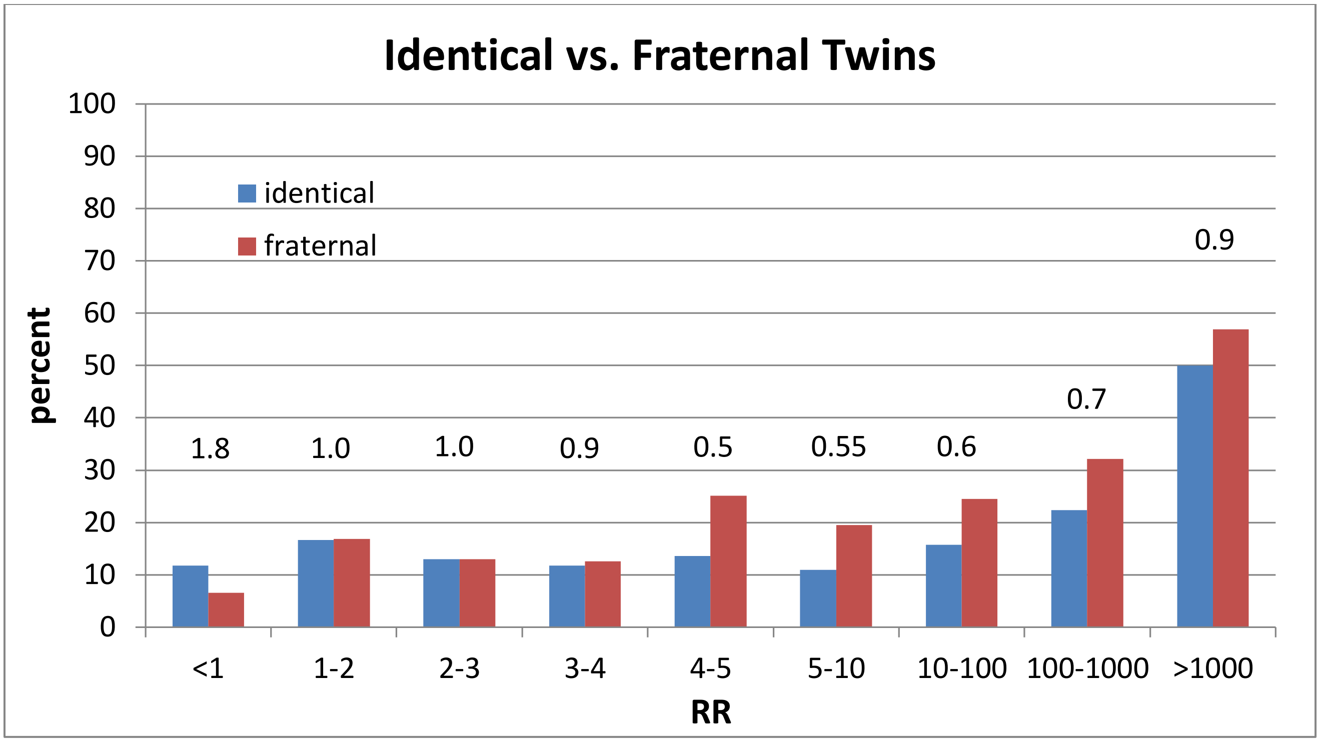

Figure 3. Relative risks (RRs) for diseases in the Marshfield Clinic Twin Cohort (MCTC). It illustrates the percent of concordant affected families in different RR categories. Numbers on the top of the histograms are the ratio of identical over fraternal values for each RR category.