Universiteit Leiden
The Netherlands

Combining Family and Twin Data in Association Studies to Estimate the Noninherited Maternal Antigens Effect

Balliu, B.; Tsonaka, R.; Woude, D. van der; Boehringer, S.; Houwing-Duistermaat, J.J.

Citation
Balliu, B., Tsonaka, R., Woude, D. van der, Boehringer, S., & Houwing-Duistermaat, J. J. (2012). Combining Family and Twin Data in Association Studies to Estimate the Noninherited Maternal Antigens Effect. *Genetic Epidemiology*, 811-819. doi:10.1002/gepi.21667

Version: Not Applicable (or Unknown)
License: Leiden University Non-exclusive license
Downloaded from: https://hdl.handle.net/1887/116916

Note: To cite this publication please use the final published version (if applicable).

Combining Family and Twin Data in Association Studies to Estimate the Noninherited Maternal Antigens Effect

Brunilda Balliu,[1]* Roula Tsonaka,[1] Diane van der Woude,[2] Stefan Boehringer,[1] and Jeanine J. Houwing-Duistermaat[1]

[1]*Department of Medical Statistics and Bioinformatics, Leiden University Medical Center, Leiden, The Netherlands*
[2]*Department of Rheumatology, Leiden University Medical Center, Leiden, The Netherlands*

It is hypothesized that certain alleles can have a protective effect not only when inherited by the offspring but also as noninherited maternal antigens (NIMA). To estimate the NIMA effect, large samples of families are needed. When large samples are not available, we propose a combined approach to estimate the NIMA effect from ascertained nuclear families and twin pairs. We develop a likelihood-based approach allowing for several ascertainment schemes, to accommodate for the outcome-dependent sampling scheme, and a family-specific random term, to take into account the correlation between family members. We estimate the parameters using maximum likelihood based on the combined joint likelihood (CJL) approach. Simulations show that the CJL is more efficient for estimating the NIMA odds ratios as compared to a families-only approach. To illustrate our approach, we used data from a family and a twin study from the United Kingdom on rheumatoid arthritis, and confirmed the protective NIMA effect, with an odds ratio of 0.477 (95% CI 0.264–0.864). *Genet. Epidemiol.* 36:811–819, 2012. © 2012 Wiley Periodicals, Inc.

Key words: ascertainment; family studies; joint likelihood; mixed models; NIMA; twin studies

Supporting Information is available in the online issue at wileyonlinelibrary.com.
*Correspondence to: Brunilda Balliu, Department of Medical Statistics and Bioinformatics, Leiden University Medical Center, Post Zone S5-P, P.O. Box 9600, 2300 RC Leiden, The Netherlands. E-mail: b.balliu@lumc.nl
Received 19 March 2012; Revised 6 June 2012; Accepted 20 June 2012
Published online 31 July 2012 in Wiley Online Library (wileyonlinelibrary.com/journal/gepi).
DOI: 10.1002/gepi.21667

INTRODUCTION

Genetic studies typically focus on testing whether a genetic variant is associated with disease risk directly through the genotype of the offspring, offspring allelic effect, to identify susceptibility genes involved in complex disorders. However, many genes influence disease susceptibility through more complex biological mechanisms, such as conditions during embryonic or fetal life. One such mechanism, the noninherited maternal antigens (NIMA) effect, may be involved in the pathogenesis of certain autoimmune diseases, such as rheumatoid arthritis (RA) [Feitsma et al., 2007; Hsieh et al., 2007], renal graft survival [Smits et al., 1998], and scleroderma [Azzouz et al., 2011; Nelson et al., 1998]. The NIMA effect affects disease susceptibility through a specific maternal-offspring genotype combination, i.e., the mother carries the allele of interest but the offspring does not. When the NIMA effect is present and not correctly modeled it can result in biased estimates of the offspring allelic effect [Sinsheimer et al., 2003; Weinberg, 1999].

In order to investigate such mechanisms, ascertained multicase family designs are typically used. They are known to improve efficiency when studying the association of a rare disease and a rare mutation, as compared to case-control studies. To accommodate for potential residual correlation in disease risks among family members, due to shared but unmeasured genetic or environmental factors, mixed models with family-specific random terms are used. An ascertainment correction is needed to account for the outcome-dependent sampling schemes, often used to increase efficiency when studying a rare disease.

Several methods have been developed to model and/or test for the NIMA effect [Feitsma et al., 2007; Hsieh et al., 2006]. However, these methods are not appropriate for families that contain both multiple cases and healthy siblings. Feitsma et al. [2007] use information only from one affected offspring per family. Hsieh et al. [2006] take into account information from multiple affected siblings, but the correlation between disease outcomes among family members, is ignored. Ignoring this correlation may have an effect on the ascertainment correction, resulting in biased results for both standard errors and effect sizes. Both methods ignore the information available from healthy siblings by excluding them from the analysis.

Recruiting, genotyping, and interviewing members of multicase families can be difficult due to the lack of clear sampling definition and the high cost, resulting in data sets with small sample size, thus low power to detect the effect of interest. To enhance the statistical power to identify disease susceptibility genes, Pfeiffer et al. [2008] and Zheng et al. [2010] proposed to combine family-based studies with case-control studies using a prospective likelihood (PL) approach to model association between genotypes and phenotypes of family members. These methods focus on direct effects, and as expected, due to the larger sample size, they increase the power to detect the direct offspring allelic effect

© 2012 Wiley Periodicals, Inc.

[Pfeiffer et al., 2008; Zheng et al., 2010]. Typically, studies with multicase families lack power to estimate the effects of rare protective factors, such as the NIMA effect. Thus, we propose to combine the multicase family study with a twin-based study and use the joint likelihood (JL), which models the joint genotype and phenotype distribution, instead of the PL. The JL can be more efficient for estimating the genetic odds ratios [Kraft and Thomas, 2000] since it only conditions on the ascertainment event, and uses information from the modeling of genotype distribution of the parents.

The parental genotypes of twins are not at hand thus the twin likelihood itself contains no information about the NIMA effect. However, we can include the NIMA parameter in the model as a nuisance parameter and marginalize the likelihood by summing over all possible parental genotypes combinations. We can then estimate the direct protective effect from both family and twin likelihood and the indirect NIMA effect from the family likelihood. In a similar way, Chen et al. [2012] use a semiparametric likelihood where the environmental effect is treated as a nuisance parameter. By combining families with twins, as compared to case-controls, we have more information on familial genotypes distribution, by assuming Mendelian inheritance, random mating, and Hardy-Weinberg proportions (HWP).

The disease of interest in this article is RA, a genetic disorder in which alleles of the HLA-DRB1 gene contribute most to the genetic risk. A group of alleles in this gene, called DERAA alleles, are known to have a protective effect against RA, when present in the genotype of the offspring. Recent observations suggest that biologically relevant exposure to HLA-antigens may occur during fetal development and subsequently through the persistence of maternal cells in the offspring. This phenomenon is called microchimerism. It has been proposed that not only inherited but also noninherited maternal HLA-antigens can influence RA susceptibility [Feitsma et al., 2007]. This implies that the exposure of DERAA-negative offspring to maternal DERAA-positive HLA-DRB1 antigens during fetal development might have a protective effect on the offspring. We applied the combined joint likelihood (CJL) to 94 multicase RA nuclear families [Hay et al., 1993; Worthington et al., 1994] and 78 dizygotic twin pairs [Silman et al., 1993], both collected from the National Repository of Family Material of the Arthritis and Rheumatism Council's.

Our method is a general framework for family-based association analysis, incorporating the advantages of several previously proposed methods such as combining different data sets, likelihood-based modeling, ascertainment correction, and modeling correlation between disease outcome of siblings. This novel method models the joint genotype and phenotype distribution, taking into account the ascertainment and correlation present in the data, and combines families and twins studies to increase information to estimate the NIMA effect. We introduce the general idea of the CJL for family-based and twin-based studies. We provide detailed estimation procedures for the family study and generalize the method to the twin study. The performance of our proposed method is assessed numerically and different approaches are compared for several scenarios, on the efficiency to estimate genetic odds ratios. The proposed method is illustrated with an analysis of the Arthritis and Rheumatism Council data and we close with discussion.

DATA AND METHODS

DATA

Consider a study where information is available from two different data sets, a family-based and a twin-based study. For every family, genotype and phenotype information is available for the offspring, affected and/or healthy, and most of their parents. Families were ascertained on the event of at least two affected offspring per family. Genotypic and phenotypic information is also available for each twin, but not for their parents. Twin pairs were ascertained such that each pair contains at least one affected member.

STATISTICAL MODELS

A commonly used approach for family data is the conditional logistic regression [Breslow and Day, 1980]. It conditions on the number of observed cases in each family, to accommodate for the outcome-dependent sampling scheme, and uses a family-specific random term, to account for dependencies in disease risk among siblings. When twins are also available, we propose to estimate the genetic odds ratios by maximizing the combined likelihood for families and twins, instead of a families-only approach. Under the assumptions that the data sets are sampled separately from the same population, with no overlap between them and with comparable data collection methods, the combined likelihood can be obtained by the product of the likelihoods for each independent study.

Notation. Let $\mathbf{Y}_i = (Y_{i1}, Y_{i2}, \ldots, Y_{in_i})$ denote phenotypes or disease status of n_i offspring in family i, where $Y_{ij} = 1$ if offspring j is affected and $Y_{ij} = 0$ if j is unaffected, $i = 1, \ldots, N_f$ and $j = 1, \ldots, n_i$. Similarly, let $\mathbf{G}_i^c = (G_{i1}^c, G_{i2}^c, \ldots, G_{in_i}^c)$ denote the genotypes of the n_i offspring and $\mathbf{G}_i^p = (G_i^m, G_i^f)$ their maternal and paternal genotypes. We denote by N_f and N_t the total number of families and twin pairs, respectively. Last, let A_i be the ascertainment event for a family or twin pair.

Likelihood for family-based study. To model the association between genotypes and phenotypes of family members we use the JL. This approach is based on the joint probability of phenotypes and genotypes, that is $P(\mathbf{Y}_i, \mathbf{G}_i^c, \mathbf{G}_i^p \mid A_i)$ and is given by:

$$JL_f(\boldsymbol{\theta}) = \prod_{i=1}^{N_f} P\left(\mathbf{Y}_i, \mathbf{G}_i^c, \mathbf{G}_i^p \mid A_i\right), \tag{1}$$

where $\boldsymbol{\theta}$ is the parameter vector. $P(\mathbf{Y}_i, \mathbf{G}_i^c, \mathbf{G}_i^p \mid A_i)$ for family i is defined as follows:

$$P\left(\mathbf{Y}_i, \mathbf{G}_i^c, \mathbf{G}_i^p \,\middle|\, \sum_{j=1}^{n_i} Y_{ij} \geq 2\right) = \frac{P\left(\mathbf{Y}_i, \mathbf{G}_i^c, \mathbf{G}_i^p, \sum_{j=1}^{n_i} Y_{ij} \geq 2\right)}{P\left(\sum_{j=1}^{n_i} Y_{ij} \geq 2\right)}$$

$$= \frac{P\left(\mathbf{Y}_i \mid \mathbf{G}_i^c, \mathbf{G}_i^p\right) \times P\left(\mathbf{G}_i^c \mid \mathbf{G}_i^p\right) \times P\left(\mathbf{G}_i^p\right)}{P\left(\sum_{j=1}^{n_i} Y_{ij} \geq 2\right)}. \tag{2}$$

The second identity of (2) requires two assumptions. First, subjects selection should depend only upon potential subjects' disease status, not on their covariates, that is $P(\sum_{j=1}^{n_i} Y_{ij} \geq 2 \mid \mathbf{Y}_i, \mathbf{G}_i^c, \mathbf{G}_i^p) = P(\sum_{j=1}^{n_i} Y_{ij} \geq 2 \mid \mathbf{Y}_i)$. Second, families should be selected under complete ascertainment, that is $P(\sum_{j=1}^{n_i} Y_{ij} \geq 2 \mid \mathbf{Y}_i) = 1$ for a family with at least two affected offspring, and 0 otherwise.

The numerator of (2) is a product of the *disease penetrance function* $P(\mathbf{Y}_i \mid \mathbf{G}_i^c, \mathbf{G}_i^p)$, the *transmission probabilities* $P(\mathbf{G}_i^c \mid \mathbf{G}_i^p)$, and the *parental genotype probabilities* $P(\mathbf{G}_i^p)$. The disease penetrance function models the disease probability of n_i offspring given the genotypes of the family. We will explain how we model the penetrance function in the next section. We assume Mendelian inheritance for the transmission probability $P(\mathbf{G}_i^c \mid \mathbf{G}_i^p)$, random mating for the parents, and HWP for the genotype distribution. Thus, the parental genotype probability $P(\mathbf{G}_i^p)$ is characterized by a single parameter, the allele frequency q.

The denominator is the *ascertainment correction* and models the probability that at least two of the offspring in the family are affected $P(\sum_{j=1}^{n_i} Y_{ij} \geq 2)$. This probability can be expressed in terms of the marginal distribution by summing the joint distribution of phenotype and genotypes over all possible genotype combinations in a family, namely:

$$P\left(\sum_{j=1}^{n_i} Y_{ij} \geq 2\right) = 1 - \sum_{\mathbf{G}_*^c, \mathbf{G}_*^p} P\left(\mathbf{G}_*^c \mid \mathbf{G}_*^p\right) \times P\left(\mathbf{G}_*^p\right) \times \left\{ P\left(\sum_{j=1}^{n_i} Y_{ij} = 1 \middle| \mathbf{G}_*^c, \mathbf{G}_*^p\right) + P\left(\sum_{j=1}^{n_i} Y_{ij} = 0 \middle| \mathbf{G}_*^c, \mathbf{G}_*^p\right) \right\}. \tag{3}$$

Disease penetrance function. In this section, we present the penetrance function for a family in the data set. Given a set of family-specific random effects u_i, we assume that $(Y_{i1}, Y_{i2}, \ldots, Y_{in_i})$ are conditionally independent. Thus, the penetrance function for one family can be expressed as the product of the penetrance functions for each offspring in the family:

$$P\left(\mathbf{Y}_i \mid \mathbf{G}_i^c, \mathbf{G}_i^p, u_i\right) = \prod_{j=1}^{n_i} P\left(Y_{ij} = y_{ij} \mid G_{ij}^c, \mathbf{G}_i^p, u_i\right).$$

In order to estimate the parameters of interest, we use the marginal probability of the disease outcome of the ith family, given by:

$$P\left(\mathbf{Y}_i \mid \mathbf{G}_i^c, \mathbf{G}_i^p\right) = \int_{u_i} P\left(\mathbf{Y}_i \mid \mathbf{G}_i^c, \mathbf{G}_i^p, u_i\right) f(u_i) du_i. \tag{4}$$

We assume that the random intercept is normally distributed, $u_i \sim N(0, \tau_u^2)$. The integral is analytically intractable and we resort to numerical integration. To evaluate the integral, we used the Gauss-Hermite Quadrature rule.

Last, we specify the individual penetrance function. We consider here the case where a direct offspring allelic effect and an indirect NIMA effect affect the disease probability for each offspring. We assume no direct maternal or paternal allelic effect. The disease probability for each offspring is a function of offspring genotype, combination of maternal and offspring genotype, and the random effect u_i:

$$P\left(Y_{ij} = 1 \mid G_{ij}^c, \mathbf{G}_i^m, u_i\right) = \text{logit}^{-1}\left(\beta_0 + \beta_1 \times I[OAE_{ij}] + \beta_2 \times I[NIMA_{ij}] + u_i\right), \tag{5}$$

where logit^{-1} is the inverse logit function, $\text{logit}^{-1}(x) = \frac{\exp(x)}{1+\exp(x)}$. Parameter β_0 is the intercept of the logistic model. Let $I[\cdot]$ denote an indicator function. OAE_{ij} denotes an event of offspring allelic effect. We assume a dominant model, where $I[OAE_{ij}] = 1$ when one or two copies of the protective allele are present in the offspring's genotype and zero otherwise. Parameter β_1 represents the log odds ratio of disease probability for the offspring allelic effect. Let $NIMA_{ij}$ denote an event of NIMA, where $I[NIMA_{ij}] = 1$ if a copy of the protective allele is present in the maternal genotype but not present in the offspring's genotype and zero otherwise. Parameter β_2 represents the log odds ratio of the NIMA effect. The interpretation of parameters is conditional on the family-specific random effects. In Table I, all possible genotype combination of mother-offspring pair and resulting effects is reported.

Likelihood for twin-based study. In this section, we modify the JL presented in the previous section to model data from twin-based studies. Since no parental genotypes are available in the twin study, it is not possible to estimate the indirect NIMA effect. Namely, the twin likelihood contains no information about NIMA. However, we need to include the NIMA parameter in the twin likelihood to ensure that the parameters of the family and twin likelihood have the same interpretation. Missing data are dealt with by marginalizing over all possible parental genotypes combinations, treating β_2 as a nuisance parameter. Following the notation used in (1), the JL for the twin data set is given by:

$$JL_t(\boldsymbol{\theta}) = \prod_{i=1}^{N_t} P\left(\mathbf{Y}_i, \mathbf{G}_i^c, \mid A_i\right), \tag{6}$$

where $P(\mathbf{Y}_i, \mathbf{G}_i^c, \mid A_i)$ for twin pair i is given as follows:

$$P\left(\mathbf{Y}_i, \mathbf{G}_i^c \middle| \sum_{j=1}^{2} Y_{ij} \geq 1\right) = \sum_{\mathbf{G}_*^p} P\left(\mathbf{Y}_i, \mathbf{G}_i^c, \mathbf{G}_*^p \middle| \sum_{j=1}^{2} Y_{ij} \geq 1\right) = \sum_{\mathbf{G}_*^p} \frac{P(\mathbf{Y}_i \mid \mathbf{G}_i^c, \mathbf{G}_*^m) \times P(\mathbf{G}_i^c \mid \mathbf{G}_*^p) \times P(\mathbf{G}_*^p)}{1 - \sum_{\mathbf{G}_*^c, \mathbf{G}_*^p} P\left(\sum_{j=1}^{2} Y_{ij} = 0 \mid \mathbf{G}_*^c, \mathbf{G}_*^m\right) \times P(\mathbf{G}_*^c \mid \mathbf{G}_*^p) \times P(\mathbf{G}_*^p)}.$$

Combined likelihood for the family and twin studies. To obtain joint estimates for the NIMA and direct offspring allelic effect, we maximize the combined likelihood for both data sets, given by the product of the likelihood contribution from family study (1), and the likelihood contribution from twin study (6):

$$CJL(\tau_u, \beta_0, \beta_1, \beta_2) = JL_f(\tau_u, \beta_0, \beta_1, \beta_2) \times JL_t(\tau_u, \beta_0, \beta_1, \beta_2). \tag{7}$$

TABLE I. Possible genotype combination of mother-offspring pair and resulting protective effects

Offspring genotype	Maternal genotype	Resulting effect
No copy of protective allele	No copy of protective allele	Reference category
One/two copies of protective allele	No/one/two copies of protective allele	Offspring allelic effect
No copy of protective allele	One copy of protective allele	NIMA effect

Information to estimate the direct allelic effect, the baseline risk, and the variance of the random effect comes both from twins and families. On the other hand, the family likelihood allows us to estimate also the NIMA effect. By adding the twins to the families, we borrow information to better estimate the direct allelic effect, which will also improve the estimate of the NIMA parameter through the family likelihood.

SIMULATION STUDY

The primary goal of the simulation study was to test efficiency gain for estimating effects that depend on parental genotype, such as NIMA, when a twin data set, with missing parental information, is combined with a data set comprised of nuclear families. In addition, we wanted to study the finite sample properties of the JL itself and relative to the PL. In particular, we investigated the impact of family size, variance of random effects and ascertainment scheme on the parameter estimates, and compared our method with the PL used in previous studies, in terms of efficiency and bias of estimates of NIMA effect.

In each scenario, genotype frequencies were selected to mimic the frequency of DERAA alleles in the English population, that is 0.15 [Ann Morgan, personal communication]. To generate genotypes of family members, maternal and paternal genotypes were generated assuming random mating and HWP. Offspring genotypes were generated assuming Mendelian transmission. Disease outcomes of offspring were generated according to the random effects model (5). The family-specific random intercept was assumed to be normally distributed with mean zero and variance either 1.5 or 2.5, resembling results from previous literature on heritability of RA [van der Woude et al., 2009]. Two different ascertainment schemes were used, that is, families were included in the study if at least one or two offspring were affected. Twins were generated as families with two offspring, ascertained such that at least one twin per pair is affected. Parental genotype and phenotype information was ignored to mimic the real data set. We set β_0 to -3, representing a rare disease with marginal population prevalence approximately 5%. The true parameter values for offspring allelic and NIMA effect, β_1 and β_2, were fixed at -0.5 and -1, corresponding to an odds ratio of 0.6 and 0.4, respectively. In total, 16 scenarios were generated, each consisting of 10^3 simulated data sets, with corresponding family and sample size, ascertainment scheme, and variance of the random effect as indicated in Table II.

To study the finite sample properties of the JL, we applied the likelihood to all scenarios of Table II. Results are summarized in Table III. Effect of different family and sample size on the parameter estimates is reflected by comparing scenarios 1–4. When both sample and family size are small, e.g., scenario 1, τ_u^2 is overestimated resulting in an underestimated β_0 . However, estimates of the log odds ratios for the offspring allelic and NIMA effect are nearly unbiased, -2.3% and 3.4%, respectively. Increasing family size from 3 to 5, scenario 2, reduces the bias of both effects to 0.1% and 2.4% and their standard deviations by 8.5% and 11.43%, respectively. On the other hand, increasing the number of families from 100 to 500, scenario 3, reduces the bias of both effects to -1.4% and -1.0% and their standard deviations by 55.6% and 58.4%, respectively. To study the effect of different τ_u^2 on the parameter estimates, we compared scenarios 1–4 with scenarios 5–8 or/and scenarios 9–12 with scenarios 13–14. When τ_u^2 increases from 1.5 to 2.5, from scenario 1 to scenario 5, bias on the estimate of β_0 and τ_u^2 itself increases. However, this does not introduce much bias in the estimation of the offspring allelic and NIMA parameters. Different ascertainment schemes were compared by contrasting scenarios 1–4 with scenarios 9–12. Bias in τ_u^2 and β_0 estimates increases when ascertainment is $\sum_j Y_{ij} \geq 2$, as compared to $\sum_j Y_{ij} \geq 1$ while estimates of the offspring allelic and NIMA parameters remain unbiased, e.g., bias in scenario 9, for β_1 and β_2, is 1.9% and 5.7%, respectively.

TABLE II. Simulation scenarios with varying sample and family size, ascertainment scheme and variance of the random effects

Scenario	Number of families	Number of offspring	Ascertainment scheme	Variance of random effect
1	100	3	$\sum_j Y_{ij} \geq 1$	1.5
2	100	5	$\sum_j Y_{ij} \geq 1$	1.5
3	500	3	$\sum_j Y_{ij} \geq 1$	1.5
4	500	5	$\sum_j Y_{ij} \geq 1$	1.5
5	100	3	$\sum_j Y_{ij} \geq 1$	2.5
6	100	5	$\sum_j Y_{ij} \geq 1$	2.5
7	500	3	$\sum_j Y_{ij} \geq 1$	2.5
8	500	5	$\sum_j Y_{ij} \geq 1$	2.5
9	100	3	$\sum_j Y_{ij} \geq 2$	1.5
10	100	5	$\sum_j Y_{ij} \geq 2$	1.5
11	500	3	$\sum_j Y_{ij} \geq 2$	1.5
12	500	5	$\sum_j Y_{ij} \geq 2$	1.5
13	100	3	$\sum_j Y_{ij} \geq 2$	2.5
14	100	5	$\sum_j Y_{ij} \geq 2$	2.5
15	500	3	$\sum_j Y_{ij} \geq 2$	2.5
16	500	5	$\sum_j Y_{ij} \geq 2$	2.5

Next, we compare the two different likelihoods to model family/twin data in terms of efficiency, the PL used in existing methods, with the approach we use in this article, the JL. We define the percentage of efficiency improvement of likelihood A over B, for estimating a parameter β, as $\text{EI} = (1 - \frac{\text{Var}(\beta_A)}{\text{Var}(\beta_B)}) \times 100$. Positive values mean that likelihood A performs better. In Figure 1, we plot the EI of the JL over the PL, for estimating the log odds ratios of the offspring allelic and NIMA effect. All values are positive;

TABLE III. Summary statistics for parameter estimates of the JL (1) under the penetrance model $P(Y_{ij} = 1 \mid G_{ij}^{c}, G_{i}^{m}, u_i) = \text{logit}^{-1}(\beta_0 + \beta_1 \times I[OAE_{ij}] + \beta_2 \times I[NIMA_{ij}] + u_i)$. The results are based on 1,000 simulated data sets per scenario, each of sample size 100 or 500 and number of offspring per family three or five. The families were ascertained such that each family would have at least one or at least two affected offspring. The protective allele frequency was 0.15. All scenarios are described in detail in Table II. Each entry lists the mean estimates (standard deviation of estimates) over the 1,000 simulated data sets

Scenario	True values			
	$\tau_u^2 = 1.5$	$\beta_0 = -3$	$\beta_1 = -0.5$	$\beta_2 = -1$
1	2.148 (2.538)	−3.365 (1.460)	−0.477 (0.349)	−1.034 (0.507)
2	1.584 (1.038)	−3.049 (0.579)	−0.501 (0.319)	−1.024 (0.449)
3	1.571 (0.771)	−3.052 (0.466)	−0.486 (0.155)	−0.990 (0.211)
4	1.543 (0.411)	−3.022 (0.226)	−0.497 (0.140)	−1.000 (0.197)
	$\tau_u^2 = 2.5$	$\beta_0 = -3$	$\beta_1 = -0.5$	$\beta_2 = -1$
5	3.517 (3.382)	−3.476 (1.612)	−0.478 (0.397)	−1.016 (0.541)
6	2.724 (1.662)	−3.104 (0.766)	−0.503 (0.344)	−1.022 (0.469)
7	2.587 (1.152)	−3.045 (0.572)	−0.492 (0.169)	−1.001 (0.231)
8	2.577 (0.629)	−3.033 (0.290)	−0.504 (0.149)	−1.003 (0.196)
	$\tau_u^2 = 1.5$	$\beta_0 = -3$	$\beta_1 = -0.5$	$\beta_2 = -1$
9	2.827 (3.001)	−4.236 (2.864)	−0.519 (0.265)	−1.057 (0.407)
10	1.98 (1.765)	−3.408 (1.442)	−0.501 (0.258)	−1.020 (0.386)
11	2.472 (2.290)	−3.929 (2.194)	−0.499 (0.120)	−0.999 (0.173)
12	1.607 (0.672)	−3.091 (0.560)	−0.497 (0.112)	−0.994 (0.167)
	$\tau_u^2 = 2.5$	$\beta_0 = -3$	$\beta_1 = -0.5$	$\beta_2 = -1$
13	3.468 (3.347)	−3.673 (2.465)	−0.540 (0.316)	−1.056 (0.459)
14	2.944 (2.032)	−3.308 (1.341)	−0.501 (0.299)	−1.024 (0.423)
15	3.524 (2.852)	−3.778 (2.148)	−0.500 (0.144)	−1.016 (0.205)
16	2.716 (1.084)	−3.141 (.721)	−0.501 (0.129)	−0.999 (0.175)

thus, the JL is always more efficient. Improvement mainly depends on sample size and less on family size, e.g., EI is approximately the same in scenarios 1 and 3 as compared to scenario 2. Moreover, improvement, due to JL, is higher when information is limited, i.e., when families are small and ascertainment is $\sum_j Y_{ij} \geq 2$.

Last, we compared the performance of the JL when different data sources are available: ascertained families-only vs. ascertained families and twins. In terms of likelihoods, we compare the JL in (1) with the CJL in (7). Efficiency improvement of the families-only against the combined approach, with families and 100 twin pairs, is plotted in Figure 2. The CJL approach is more efficient under all scenarios studied. The percentage of improvement is similar across different values of variance of the random effects or ascertainment scheme. Nonetheless, improvement is noticeably high when the sample size of the nuclear family data is small. When the twin data set was added, we expected efficiency improvement for the offspring allelic effect, due to increased sample size. Interestingly, there was also efficiency improvement for the NIMA effect, which depends on the maternal genotype. The parameter estimates and their standard deviations, using the CJL, are listed in Supporting Information Table SVI.

In order to asses the performance of our method when both direct offspring and NIMA effects are under the null, $\beta_1 = \beta_2 = 0$, and cases in which there only exists a direct offspring, $\beta_2 = 0$, or only a NIMA effect, $\beta_1 = 0$, we simulated the scenarios presented in Table II with the corresponding effect sizes. We first estimated the effects optimizing the JL using only the families. Later, we added 100 twin pairs and optimized the CJL. The estimated effect sizes remain unbiased. The results are listed in Supporting Information Tables SI and SII for the JL and in the Supporting Information Tables SIII, SIV, and SV for the CJL.

The performance of our approach will vary across different frequencies of the protective allele. All the results presented above concern an allele frequency of 0.15, in order to mimic the allele frequency in the population we are studying. To study the performance of the method when allele frequency is lower, we also applied the CJL to samples generated with a protective allele frequency of 0.05. As expected, the parameter estimates are more biased for small sample sizes. Larger samples are needed to obtain unbiased estimates. Results are listed in Supporting Information Table SVII.

DATA EXAMPLE

This study was motivated by a data set consisting of 94 ascertained nuclear families, collected from the Arthritis and Rheumatism Council. Our goal is to study the effect of NIMA in RA susceptibility. In 51 families, the genotype of one of the parents, mainly the father, was missing. In 34 families, of which eight had a missing mother and 26 a missing father, we were able to construct the genotypes using the genotypes of the offspring and the genotype of the other parent. Namely, we reconstructed the missing genotype in accordance with Mendelian transmission law. For the remaining 17 families, of which nine were mothers and eight were fathers, we were able to reconstruct only one of the alleles using this approach. In order to impute the

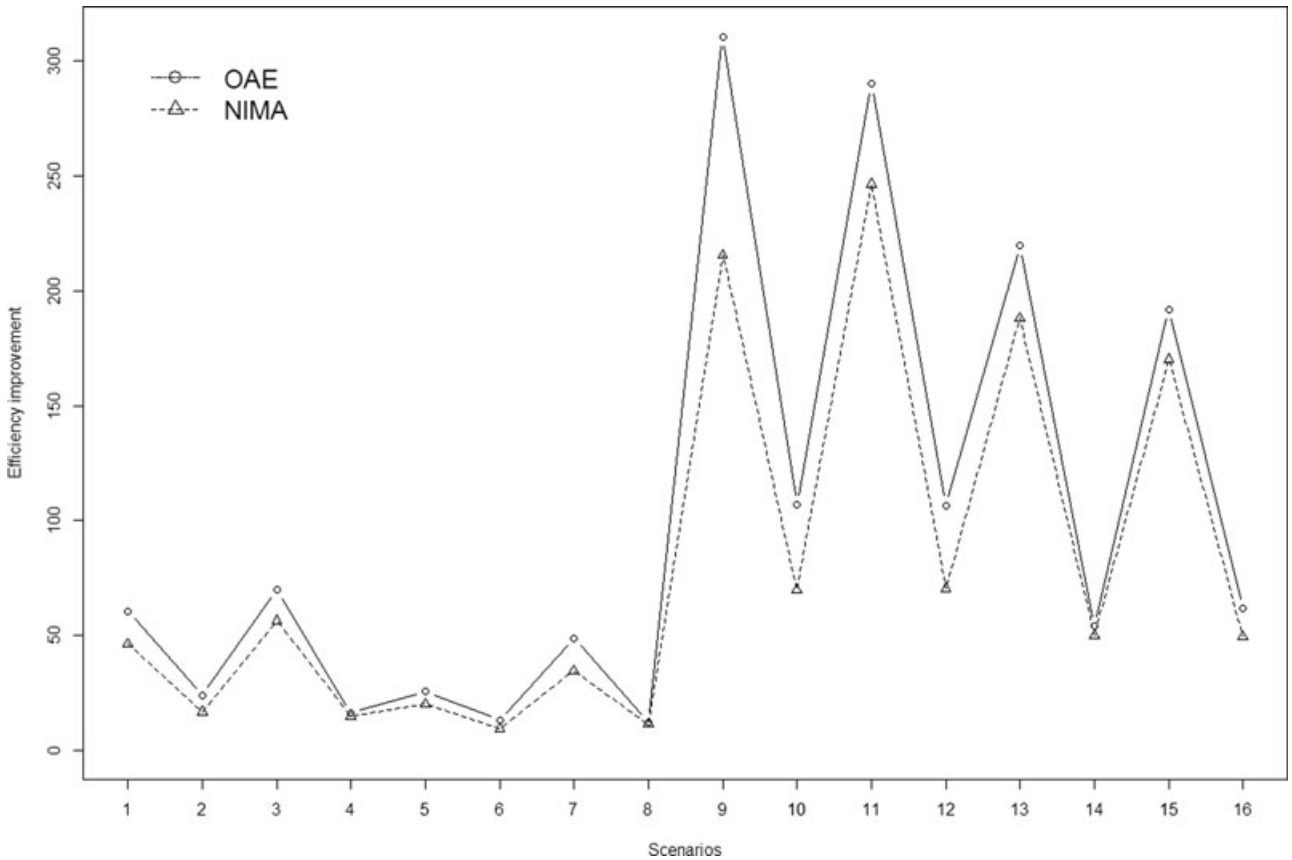

Fig. 1. Efficiency improvement of *JL* against *PL* compared for different family/sample size, ascertainment schemes, and variance of random effect. For both likelihoods, the disease penetrance function was $P(Y_{ij} = 1 \mid G^c_{ij}, G^m_i, u_i) = \text{logit}^{-1}(\beta_0 + \beta_1 \times I[OAE_{ij}] + \beta_2 \times I[NIMA_{ij}] + u_i)$. Values below zero represent no EI by using the *JL* and values above zero represent EI of the *JL* against the *PL*. Each point represents the efficiency improvement in each of the 16 scenarios presented in Table II .

second allele, we made use of the initial 4-digit allele coding of the HLA-DRB1 gene. There are 26 possible 4-digit sequences in the HLA-DRB1 gene, six of which express this DERAA allele, see van der Woude et al. [2010]. We imputed the second allele based on sampling from control 4-digit allele distribution. For six of nine mothers, we had only the first 2 digits of the 4-digit genotyping and for the rest three, we had no information about the second allele.

Families mainly contain two, three, and four offspring. There are also three large families with five, eight, and 10 offspring. A total of 86 families of 94 contain exactly two affected offspring and eight families contain three affected offspring. The maternal-offspring genotype combination that leads to the potential NIMA effect occurs only in eight families. In these eight families, four have one child, two have two children, and two have three children under potential NIMA effect. In addition, 20 offspring belonging to 13 families are under offspring allelic effect. Since there is so little information in the family data set, we decided to combine it with a data set of 78 ascertained twin pairs, also collected from the Arthritis and Rheumatism Council in the same period. Pairs mainly contain one affected member and only in three pairs both members are affected. In four pairs, both twins carry the DERAA allele, DERAA-concordant, while in 10 pairs, only one twin has the allele, DERAA-discordant. In total, 18 twins are under offspring allelic effect. Information on parental genotype of twins is not available, thus the exact number of twins under a possible NIMA effect cannot be determined.

Initially, we only analyzed the family data, using both the *JL* and the *PL* approach. Results are listed in the first two lines of Table IV. None of the likelihoods gave statistically significant results for the NIMA effect, estimated odds ratios 0.176 (95% CI 0.010–3.066) and 0.607 (95% CI 0.348–1.058) for the *PL* and the *JL* approach, respectively. Concerning the offspring allelic effect, only the *JL* resulted in a statistically significant result, odds ratios 0.194 (95% CI 0.023–1.622) for the prospective and 0.297 (95% CI 0.179–0.493) for the JL approach. Then we combined the families with the twins and applied the *CJL*. The odds ratio of the NIMA effect was statistically significant, 0.477 (95% CI 0.264–0.864) and the confidence intervals of the odds ratios of the offspring allelic effect became narrower; 0.241 (95% CI 0.152–0.380).

To conclude, we estimated a significant protective effect of the DERAA allele, coming directly from the genotype of the offspring and indirectly from the maternal genotype. That is, individuals carrying the DERAA allele have a decrease in risk of RA compared to individuals who do not carry it. Furthermore, individuals who do not carry the protective allele DERAA, but their mother does, have a decrease in risk

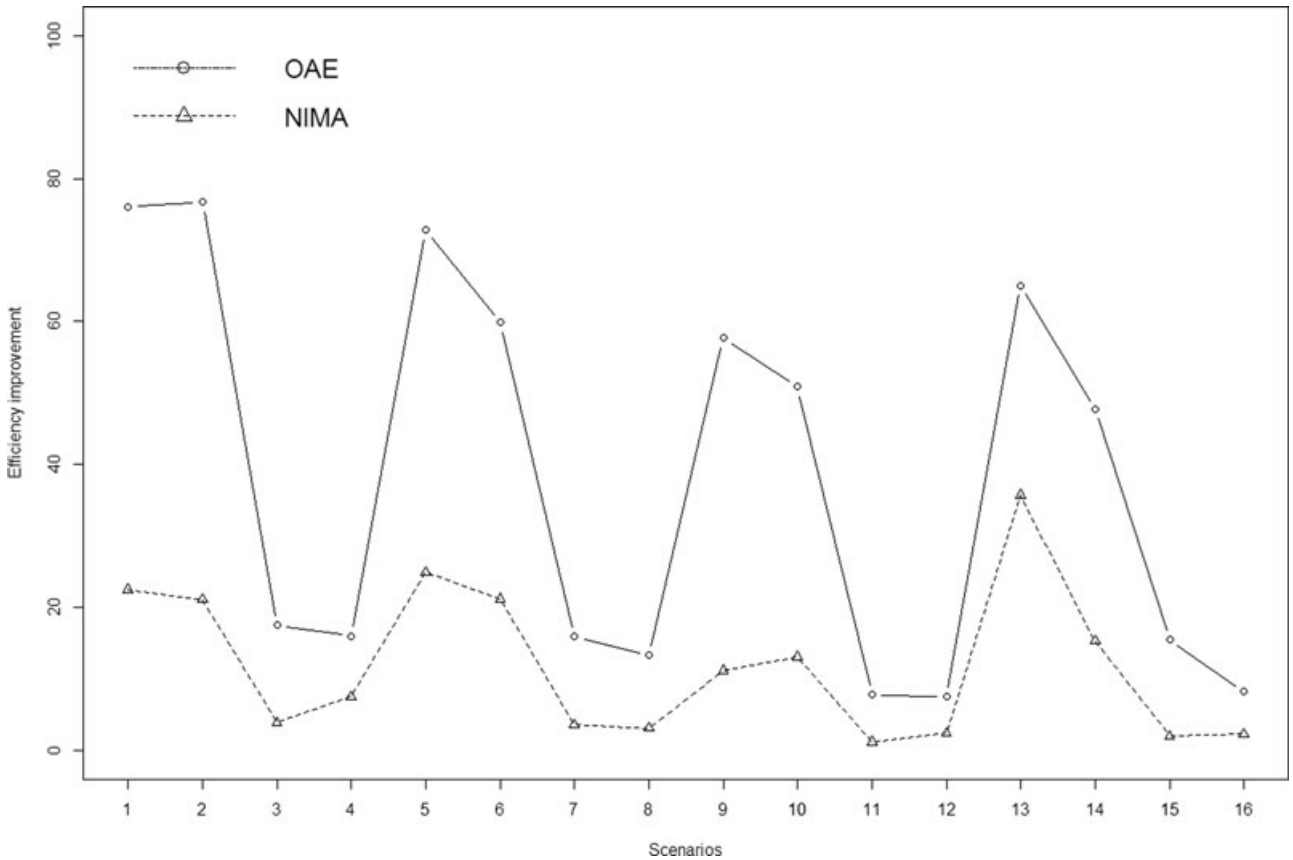

Fig. 2. Efficiency improvement of CJL approach of families and twins, against the JL approach for families-only. The efficiency improvement is compared for different family/sample size, ascertainment schemes, and variance of random effect. For both likelihoods, disease penetrance function was $P(Y_{ij} = 1 \mid G_{ij}^c, \mathbf{G}_i^m, u_i) = \text{logit}^{-1}(\beta_0 + \beta_1 \times I[OAE_{ij}] + \beta_2 \times I[NIMA_{ij}] + u_i)$. Values below zero represent no efficiency improvement by using the CJL and values above zero represent improvement of the CJL against the PL. Each point represents the efficiency improvement in each of the 16 scenarios presented in Table II.

to develop RA as compared to non-DERAA carriers whose mother also does not carry the protective allele.

DISCUSSION

In this article, we have presented a likelihood-based method for association studies combining family with twin data. Our method is appropriate for testing and estimating effects of genes that act directly through the individual's genotype but also for genes that act through complex biological mechanisms. We overcome the problem of small sample size by combining the family data set with a twin data set and using a JL approach to model the association between genotypes and phenotypes. By using a JL approach, we exploit the information coming from Mendelian transmission law, HWP, random mating and modeling of parental genotype distribution, to increase the efficiency to estimate the genetic odds ratios. The combined approach, not only enhances the statistical power to detect direct allelic effects, but also effects depending on maternal-offspring genotype combinations, such as NIMA

TABLE IV. Parameter estimates (95% CI) of the disease penetrance model $P(Y_{ij} = 1 \mid G_{ij}^c, \mathbf{G}_i^m, u_i) = \text{logit}^{-1}(\beta_0 + \beta_1 \times I[OAE_{ij}] + \beta_2 \times I[NIMA_{ij}] + u_i)$ by types of likelihood approaches used, prospective (PL), joint (JL), or combined joint likelihood (CJL), and type of data included, families-only or families and twins

Design	Variance	Intercept	OR_{OAE}	OR_{NIMA}
Families-only				
PL	1.573 (1.161–2.130)	0.005 (0.001–0.025)	0.194 (0.023–1.622)	0.176 (0.010–3.066)
JL	2.133 (1.633–2.786)	0.001 (0.000–0.006)	0.297 (0.179–0.493)	0.607 (0.348–1.058)
Families and twins				
CJL	2.416 (1.709–3.416)	0.002 (0.000–0.010)	0.241 (0.152–0.380)	0.477 (0.264–0.864)

effects. Namely, we use information from both data sets to better estimate the direct allelic effect, which gives us increased efficiency to estimate also the indirect NIMA effect. The method takes into account both the sampling scheme of the data and residual correlation between phenotype of siblings using an ascertainment correction and a family-specific random effects model.

Our approach extends existing methods for combining data sets [Pfeiffer et al., 2008; Zheng et al., 2010] to include indirect effects, using a JL, instead of a PL approach and adding twins, instead of a case-control data set. We compared the proposed JL method with the traditionally used PL approach and showed that our method is more efficient for estimating the genetic odds ratios, especially for small families with stringent selection schemes. For prospective or JL methods, including ours, ascertainment is essential to obtain unbiased parameter estimates. Here, we considered cases for which subjects' selection depends only upon potential subjects' disease status and not on their covariates. When ascertainment is also based on covariates, here genotypes, another model for ascertainment correction should be considered.

Using the JL, power can considerably increased, however at the cost of greater computational intensity, in the presence of large families. In our data set, the families where relatively small and numerical optimization of the JL was possible on a single computer. However, in the presence of large families, the computational burden rises exponentially with the family size. For given parameter values and allele frequency, the denominator (3) for family i sums over maximum 3^{n_i} possible familial genotype combinations. If all the families in the data set have a fixed size, the denominator needs to be calculated only p times for each maximization iteration, where p is the number of sample points to use for the Gauss-Hermite Quadrature approximation of the integral (4). Unfortunately, this is rarely the case in real data sets where the family size varies but the computation burden can be essentially reduced by using a grid search.

Here, we combine a family data set with a twin data set. However, the method can be extended to include other types of readily available data, such as sibling pairs, monozygotic twins, or case-parent trios data sets. Nowadays, the combination of already available data is facilitated from existing nationwide registries of families and twins at high risk for particular traits. Extension of the likelihood-based analysis described here, to accommodate multiallelic marker, is trivial, if HWP and random mating assumptions are made. Although we have focused on association of single single-nucleotide polymorphisms the approach can be extended to allow for the analysis of haplotypes . Since haplotypes combine linkage disequilibrium information from multiple markers simultaneously, this approach could be more powerful than our current approach. Direct extension to accommodate haplotypes is not straightforward, due to the increase in the number of parameters needed to model the haplotypes, and is beyond the scope of this article. The proposed method can be extended to other complex biological mechanisms, such as maternal effects or imprinting, by adding the appropriate covariates in the logistic regression (5). Last, by incorporating our method to methodology applied in Houwing-Duistermaat et al. [2000], we could study whether genetic NIMA effects of RA could create a protection for diseases associated with RA, such as cardiovascular disease or anemia.

We employed a fully parametric models for the random effects distribution. Since no straightforward diagnostics are available to evaluate the validity of the random effects model assumptions, there is a potential for model misspecification. Nevertheless, the estimates of the fixed effects are robust-to-moderate misspecifications of the underlying random effects distribution [Heagerty and Kurland, 2001; Pfeiffer et al., 2003]. One could also analyze the data simply by using a generalized estimating equations (GEE) approach [Liang and Zeger, 1986]. However, since the GEE estimates do not take into account the sampling design, the resulting covariate effect estimates might be biased, because the family and twin data sets are not a random sample of the families and twins in the population. While the random effects model allows one to accommodate ascertainment of the families as well as residual familial correlation, the interpretation of the parameters is conditional on the random effects [Fitzmaurice et al., 1993]. Marginal parameter estimates can be obtained using the approximate formula of Diggle et al. [1994]. This approximation uses the variance of the random effects. In the simulation study, we observed that the estimate of the variance, needed for the marginalization, might be biased when sample size is small. Thus, we recommend to use the approximation formula only when the sample size and/or family size are large, e.g., 500 families with three offspring when ascertainment is at least one affected offspring.

To conclude, we confirmed the protective effect of the inherited DERAA alleles, offspring allelic effect, and the noninherited maternal DERAA alleles, NIMA effect. The simulation study and the result of the real data analysis suggest that a combined approach can be more powerful, as compared to a families-only approach, when the information on the initial family data set is restricted.

ACKNOWLEDGMENTS

This work was supported by a Program Grant (917.66.344) from the Netherlands Organization for Scientific Research and from the Dutch Arthritis Foundation (Reumafonds).

REFERENCES

Azzouz D, Martin M, Roudier J, Lambert NC. 2011. Could microchimerism be a source of disease-associated HLA alleles in patients with scleroderma? Ann Rheum Dis 70: A34.

Breslow NE, Day NE. 1980. Statistical Methods in Cancer Research, 1: The Analysis of Case-Control Studies. Lyon: International Agency for Research on Cancer.

Chen J, Lin D, Hochner H. 2012. Semiparametric maximum likelihood methods for analyzing genetic and environmental effects with case-control mother child pair data. Biometrics doi:10.1111/j.1541-0420.2011.01728.x.

Diggle Pj, Liang KY, Zeger SL. 1994. Analysis of Longitudinal Data. Oxford University Press, USA.

Feitsma AL, Worthington J, van der Helm-van Mil AHM, Plant D, Thomson W, Ursum J, van Schaardenburg D, van der Horst-Bruinsma IE, van Rood J, Huizinga TWJ, Toes REM, de Vries RRP. 2007. Protective effect of noninherited maternal HLA-DR antigens on rheumatoid arthritis development. Proc Natl Acad Sci USA 104: 19966–19970.

Fitzmaurice BY, Garrett M, Laird N. 1993. A likelihood-based method for analysing longitudinal binary responses. Biometrika 80: 141–151.

Hay EM, Olliver WFR, Silman AJ. 1993. The arthritis and rheumatism council's national family material repository. Br J Rheumatol 32: 443–444.

Heagerty PJ, Kurland BF. 2001. Misspecified maximum likelihood estimates and generalised linear mixed models. Biometrika 88: 973–985.

Houwing-Duistermaat JJ, van Houwelingen HC, de Winter JP. 2000. Estimation of individual genetic effects from binary observations on relatives applied to a family history of respiratory illnesses and chronic lung disease of newborns. Biometrics 56: 808–814.

Hsieh H, Palmer CGS, Harney S, Newton JL, Wordsworth P, Brown MA, Sinsheimer JS. 2006. The v-MFG test: investigating maternal, offspring and maternal-fetal genetic incompatibility effects on disease and viability. Genet Epidemiol 30: 333–347.

Hsieh H, Palmer CGS, Harney S, Chen H, Bauman L, Brown MA, Sinsheimer JS. 2007. Using the maternal-fetal genotype incompatibility test to assess noninherited maternal HLA-DRB1 antigen coding alleles as rheumatoid arthritis risk factors. BMC Proc I(Suppl I): S124.

Kraft P, Thomas DC. 2000. Bias and efficiency in family-based gene-characterization studies: conditional, prospective, retrospective, and joint likelihoods. Am J Hum Genet 66: 1119–1131.

Liang KY, Zeger SL. 1986. Longitudinal data analysis using generalized linear models. Biometrika 73:13–22.

Nelson JL, Furst DE, Maloney S, Gooley T, Evans PC, Smith A, Bean MA, Ober C, Bianchi D. 1998. Microchimerism and HLA-compatible relationships of pregnancy in scleroderma. Lancet 351:559–562.

Pfeiffer RM, Hildesheim A, Gail MH, Pee D, Chen CJ, Goldstein AM, Diehl SR. 2003. Robustness of inference on measured covariates to misspecification of genetic random effects in family studies. Genet Epidemiol 24: 14–23.

Pfeiffer RM, Pee D, Landi MT. 2008. On combining family and case-control studies. Genet Epidemiol 32: 638–646.

Silman AJ, MacGregor AJ, Thomson W, Holligan S, Carthy D, Farhan A, Ollier WER. 1993. Twin concordance rates for rheumatoid arthritis: results from a nationwide study. Br J Rheumatol 32:903–907.

Sinsheimer JS, Palmer CGS, Woodward JA. 2003. Detecting genotype combinations that increase risk for disease: the maternal-fetal genotype incompatibility test. Genet Epidemiol 24: 1–13.

Smits JMA, Claas FHJ, van Houwelingen HC, Persijn GG. 1998. Do non-inherited maternal antigens (NIMA) enhance renal graft survival?. Transpl Int 11: 82–88.

van der Woude D, Houwing-Duistermaat JJ, Toes REM, Huizinga TWJ, Thomson W, Worthington J, van der Helm-van Mil AHM, de Vries RRP. 2009. Quantitative heritability of anti-citrullinated protein antibody-positive and anti-citrullinated protein antibody-negative rheumatoid arthritis. Arthritis Rheum 60: 916–923.

van der Woude D, Lie BA, Lundström E, Balsa A, Feitsma AL, Houwing-Duistermaat JJ, Verduijn W, Nordang GBN, Alfredsson L, Klareskog L, Pascual-Salcedo D, Gonzalez-Gay MA, Lopez-Nevot MA, Valero F, Roep BO, Huizinga TWJ, Kvien TK, Martín J, Padyukov L, de Vries RRP, Toes REM. 2010. Protection against anti-citrullinated protein antibody-positive rheumatoid arthritis is predominantly associated with HLA-DRB1*1301: a meta-analysis of HLA-DRB1 associations with anti-citrullinated protein antibody-positive and anti-citrullinated protein. Arthritis Rheum 62: 1236–1245.

Weinberg, CR. 1999. Methods for detection of parent-of-origin effects in genetic studies of case-parents triads. Am J Hum Genet 65: 229–235.

Worthington J, Olliver WFR, Leach MK, Smith I, Hay EM, Thomson W, Pepper L, Carthy D, Farhan A, Martin S, Dyer P, Davison J, Bamber S, Silman AJ. 1994. Research practice the arthritis and rheumatism council's national repository of family material: pedigrees from the first 100 rheumatoid arthritis families containing affected sibling pairs. Br J Rheumatol 33: 970–976.

Zheng Y, Heagerty PJ, Hsu L, Newcomb P. 2010. On combining family-based and population-based case-control data in association studies. Biometrics 66: 1024–1033.