Zygosity Diagnosis in the Absence of Genotypic Data: An Approach Using Latent Class Analysis

Andrew C. Heath[1], Dale R. Nyholt[2], Rosalind Neuman[1], Pamela A. F. Madden[1], Kathleen K. Bucholz[1], Richard D. Todd[1], Elliot C. Nelson[1], Grant W. Montgomery[2], and Nicholas G. Martin[2]

[1] *Missouri Alcoholism Research Center, Department of Psychiatry, Washington University School of Medicine, St Louis, Missouri, U.S.A.*
[2] *Genetic Epidemiology Laboratory, Queensland Institute of Medical Research, Brisbane, Australia*

For zygosity diagnosis in the absence of genotypic data, or in the recruitment phase of a twin study where only single twins from same-sex pairs are being screened, or to provide a test for sample duplication leading to the false identification of a dizygotic pair as monozygotic, the appropriate analysis of respondents' answers to questions about zygosity is critical. Using data from a young adult Australian twin cohort (N = 2094 complete pairs and 519 singleton twins from same-sex pairs with complete responses to all zygosity items), we show that application of latent class analysis (LCA), fitting a 2-class model, yields results that show good concordance with traditional methods of zygosity diagnosis, but with certain important advantages. These include the ability, in many cases, to assign zygosity with specified probability on the basis of responses of a single informant (advantageous when one zygosity type is being oversampled); and the ability to quantify the probability of misassignment of zygosity, allowing prioritization of cases for genotyping as well as identification of cases of probable laboratory error. Out of 242 twins (from 121 like-sex pairs) where genotypic data were available for zygosity confirmation, only a single case was identified of incorrect zygosity assignment by the latent class algorithm. Zygosity assignment for that single case was identified by the LCA as uncertain (probability of being a monozygotic twin only 76%), and the co-twin's responses clearly identified the pair as dizygotic (probability of being dizygotic 100%). In the absence of genotypic data, or as a safeguard against sample duplication, application of LCA for zygosity assignment or confirmation is strongly recommended.

Historically, early twin studies relied upon questionnaire-based assessment of twin pair zygosity, using questions about physical similarity and confusion in childhood, with typing of genetic markers in a subset of pairs establishing a high level of validity of such questionnaire assessments (e.g., Cederlof et al., 1961; Hauge et al., 1989; Kasriel & Eaves, 1976; Magnus et al., 1983; Nichols & Bilbro, 1966; Sarna et al., 1978). Typically, discriminant function or logistic regression analyses have been used in such studies to establish the accuracy of classification. While most such methods have relied upon combining information from both twins from a pair, even answers from a single twin were found to yield a misclassification rate below 5% (Magnus et al., 1983). Because of the high accuracy and low costs of questionnaire-based assessment of zygosity, and the far from negligible laboratory error rate associated with processing of samples for genotyping (e.g., accidental sample duplication leading to the misclassification of a DZ pair as monozygotic), accurate zygosity assessment by questionnaire remains an important component of contemporary twin studies.

Latent class analysis (LCA; e.g., McCutcheon, 1987; Eaves et al., 1993) provides an alternative statistical framework for the analysis of zygosity questionnaire items and derivation of predicted zygosity types. LCA may be viewed as a categorical variant of factor analysis, in that it assumes the existence of discrete mutually exclusive categories or "classes" (e.g., monozygotic versus dizygotic twins), rather than the continuously distributed latent variables assumed in a factor model. It is further assumed that within classes, item response probabilities are statistically independent, so that it is the existence of discrete classes which explains the clustering of responses to a set of items. Input data are the observed response profiles to the set of items. Parameters of a latent class model are class-specific item endorsement probabilities (e.g., the probabilities of endorsing particular zygosity questionnaire items for those assigned to MZ versus DZ classes), and class prevalence estimates. Of particular importance, LCA does not require a priori assignment of individuals as monozygotic or dizygotic. Rather, from the estimates of model parameters predicted probabilities of membership in each estimated class (e.g., of being monozygotic versus dizygotic) may be derived. Latent class analysis thus provides a natural framework for classifying twin pairs when genotypic data are unavailable (or need to be scanned for potential laboratory errors) and for quantifying the certainty or uncertainty of the zygosity classification of a twin pair.

Address for correspondence: Andrew C. Heath, D.Phil., Missouri Alcoholism Research Center, Department of Psychiatry, Washington University School of Medicine, 40 N. Kingshighway Suite One, St Louis, MO 63108, USA. Email: andrew@matlock.wustl.edu

https://doi.org/10.1375/twin.6.1.22 Published online by Cambridge University Press

Methods

Sample

Respondents were 2094 like-sex twin pairs from a young adult Australian twin cohort, and 519 singleton twins from like-sex pairs whose co-twin did not respond. This sample (including excluded unlike-sex pairs) are referred to as the Australian "1989 cohort" (because they were first surveyed as adults by mailed questionnaire in 1989: Heath et al., 2001). Twins were born 1964–1971. Eight cases with missing data for one or more zygosity items were deleted from analyses presented here and excluded from the reported sample sizes for this paper. The sample is described in greater detail elsewhere (Heath et al., 2001; Nelson et al., 2002). Although a volunteer cohort, recruited through appeals to their parents through Australian school systems and through media appeals, when the twins were children (1980–1982), respondents were drawn from a broad range of socioeconomic levels (Heath et al., 2001). Unlike-sex pairs were of course excluded from these analyses. Genotypic data were available for a subsample of 121 pairs (58 MZ, 63 DZ same-sex).

Assessment

Standard questions for zygosity diagnosis were used (Cederlof et al., 1961; Nichols & Bilbro, 1966; Magnus et al., 1983). Questions covering (a) how often parents had difficulty telling the respondent and his or her co-twin apart, (b) how often teachers had difficulty telling them apart, and (c) how often strangers had difficulty telling them apart, were referenced to the period when the twins were 6–13 years old, and were answered using a 6-point scale: (1) always, (2) usually, (3) sometimes, (4) rarely, (5) never, (6) don't know. Questions concerning physical similarity addressed whether the twins had the same eye color, same natural hair color, and same complexion. The standard question about "peas in a pod" was worded as "When you and your twin were children, were you as alike as "two peas in a pod", or only of normal family likeness — that is, no more alike physically than ordinary sisters or brothers". Finally, respondents were asked for their own assessment of their zygosity: "In your opinion, are you and your twin … (1) definitely identical, (2) probably identical, (3) probably fraternal or (4) definitely fraternal, with an additional response option of "Not Sure" used if the respondent volunteered this response. Questions were embedded in a telephone interview (Nelson et al., 2002).

Genotyping

Of the 121 like-sex twin pairs with zygosity confirmed via DNA, 116 pairs were confirmed using genotypes from genome scan data, 2 pairs were confirmed using genotypes from candidate gene data, and 3 pairs were confirmed using genotypes from the AmpF*l*STR–Profiler–Plus™ zygosity determination system (Applied Biosystems). Briefly, genome-scan and candidate-gene genotypes obtained by gel electrophoresis after PCR amplification were analysed using the graphical representation of relationship errors (GRR) program (Abecasis et al., 2001). GRR calculates the mean and variance of identical-by-state (IBS) allele sharing over a number of polymorphic loci for each twin pair. MZ and DZ twin pairs are determined due to their characteristic pattern of allele sharing; that is, MZ twin pairs have higher sharing on average (IBS≅1) and lower variance compared to DZ twin pairs (IBS≅0.5). Twin pairs with identical (9/9) AmpF*l*STR–Profiler–Plus™ genotypes indicate monozygosity, while different genotypes indicate dizygosity.

Analyses

Item endorsement probabilities were computed for the entire sample of same-sex twin respondents. Using data from complete pairs, twin pair concordances for item responses were also computed, using the weighted kappa statistic. A 2-class latent class model was then fitted to the 615 unique response profiles obtained for the 8 zygosity questions, using the Latent Class Analysis (LCA) program LCAP, which uses EM estimation of latent class parameters (class membership probabilities, and class-specific item endorsement probabilities) (see Neuman et al., 1999, for further details). Because of program limitations, item response categories were collapsed to a maximum of 5, including a "Don't Know" category, by combining "always" and "usually" categories for the questions about parents and teachers, and by combining "rarely" and "never" categories for the question concerning strangers. Only results under a 2-class model were considered: since the standard LCA model does not allow for correlated measurement errors, such as will arise with consecutive questions about topics such as confusion during childhood, we would expect to find additional "nuisance" classes by estimating 3 or more classes, but these would not be relevant to the task of classifying twin pairs as monozygotic versus dizygotic. Predicted probabilities of class membership associated with every observed profile of responses to the zygosity questions were output from LCAP. Most likely class membership was then compared to the best estimate zygosity assignment made by one of us (ACH) on the basis of a review of the same item responses. Also computed using LCAP were the conditional probabilities that a respondent was from Class I (the probable monozygotic pairs) associated with every response option.

Results

Zygosity item endorsement probabilities for the entire sample of twins from like-sex pairs, and weighted kappa estimates of twin pair agreement for these items, are shown in Table 1. Twin pair agreement was highest for the summary question about perceived zygosity (kappa = 0.80) and for binary items about physical resemblance and being as like as two peas in a pod (kappas = 0.77–0.81), lower for multiple category items about confusion by strangers and teachers (kappas = 0.6–-0.64), and lowest for confusion by parents (kappa = 0.44).

Fitting a 2-class latent class model identified classes that could be identified as probable MZ twins (57.3% of the sample: Class I in Table 1) and probable DZ twins (42.7% of the sample: Class II in Table 1). (Since we do not use genotypic data for the entire sample, we shall refer to probable MZ and probable DZ throughout, recognizing that questionnaire responses cannot definitively establish zygosity). Columns 2 and 3 in Table 1 summarize

https://doi.org/10.1375/twin.6.1.22 Published online by Cambridge University Press

Table 1

Response Frequencies, LCA Class-specific Item Endorsement Probabilities, and Associated Probabilities that a Respondent Is Identified as Being From Class I (i.e., Probably Monozygotic), for Zygosity Questionnaire Items. Also Shown Is Twin Pair Agreement for These Items (Weighted Kappa Statistic, and 95% Confidence Interval Shown in Parentheses)

Item		Response frequency (%) (N = 4707)	Class-specific endorsement probabilities		Conditional probability that twin is from Class I
			Class I	Class II	
Parents had difficulty telling apart	always/usually	4.6	.079	.001	.995
	sometimes	13.2	.223	.009	.971
	rarely	19.9	.315	.045	.904
	never	62.3	.382	.946	.351
	don't know (N = 1)	0.0	.000	.000	1.000
Kappa = 0.44 (0.41–0.48)					
Teachers had difficulty telling apart	always/usually	32.5	.554	.017	.977
	sometimes	26.4	.381	.109	.824
	rarely	10.9	.053	.185	.277
	never	30.0	.010	.688	.020
	don't know	0.1	.002	.001	.720
Kappa = 0.60 (0.58–0.64)					
Strangers had difficulty telling apart	always	28.2	.488	.005	.993
	usually	23.5	.388	.031	.944
	sometimes	14.1	.114	.178	.461
	rarely/never	34.1	.008	.786	.014
	don't know	0.1	.001	.000	.750
Kappa = 0.64 (0.62–0.67)					
As like as "two peas in a pod"		54.4	.924	.034	.973
	only family likeness	45.0	.071	.957	.091
	don't know	0.6	.005	.008	.436
Kappa = 0.81 (0.78–0.83)					
Same eye color	no	21.8	.013	.493	.034
	yes	75.2	.969	.461	.738
	don't know	3.0	.018	.046	.341
Kappa = 0.78 (0.74–0.81)					
Same complexion	no	24.8	.016	.558	.038
	yes	75.1	.984	.440	.749
	don't know	0.1	.000	.001	.000
Kappa = 0.63 (0.59–0.67)					
Same hair color	no	27.1	.010	.620	.022
	yes	72.9	.990	.379	.778
	don't know (N = 1)	0.0	.000	.000	.000
Kappa = 0.77 (0.74–0.80)					
Respondent believes…	definitely MZ	36.9	.636	.011	.988
	probably MZ	15.0	.251	.014	.959
	probably DZ	7.2	.054	.096	.432
	definitely DZ	40.8	.056	.879	.079
	don't know	0.1	.002	.000	.857
Kappa = 0.80 (0.78–0.82)					

maximum-likelihood class-specific item endorsement probability estimates obtained under this 2-class model. Those assigned as probable MZ twins had high probability of reporting that they and their co-twin were as like as "two peas in a pod" (.92), and had similar eye color, complexion and hair color (.97–.99). These latter physical resemblance items, however, were not estimated at 100% probability, suggesting that there may be some reporting error for these physical similarity questions. Those assigned as probable MZ twins also had a high probability of reporting that they were definitely or probably MZ (.89). Those assigned as probable DZ twins had high probability of reporting that their parents never had difficulty telling them apart (.95), that they were only of family likeness rather than like two peas in a pod (.96), and that they were definitely DZ (.88).

The fourth column in Table 1 shows the conditional probability for a given response option that the respondent is from Class I (i.e., probable monozygotic), derived from the two-class solution. Thus, while a relatively small proportion of the overall sample reported that their parents always, usually or sometimes had difficulty telling them apart, those who did endorse these items had high probability of being assigned as probable monozygotic (.97–.99). Other response categories associated with a better than

https://doi.org/10.1375/twin.6.1.22 Published online by Cambridge University Press

90% probability of being assigned as probable monozygotic were reporting that parents rarely (rather than never) had difficulty telling them apart; that teachers or strangers usually or always had difficulty telling them apart; that they were as like as "two peas in a pod"; or that they were definitely or probably MZ. Response categories associated with a high probability of being assigned as probable dizygotic (shown as a low probability of being assigned as monozygotic in column 4) were reporting that teachers never had difficulty telling them apart (.98), that strangers rarely or never had difficulty telling them apart (.99), that they were only of normal family likeness (.91), had differences in eye color, complexion or hair color (.96–.98), and were definitely DZ (.92). Respondents who responded "Don't Know" to questions about being mistaken for their co-twin, and about their zygosity, had relatively high probability of being assigned as monozygotic (.72–1.00). However, since these Don't Know responses were rare (1–6 cases), this cannot be considered a reliable finding.

A relatively high proportion of respondents were assigned to one class or the other with very high probability (results not shown). Thus 86.7% were assigned with probability greater than or equal to 0.9999 (36.7% as DZ, 50.2% as MZ); 96.7% with probability greater than or equal to .95 (40.9% as DZ, 55.8% as MZ); and 98.3% with probability greater than .80. Using the most stringent criterion of $p > 0.9999$, there were only 7 pairs who received inconsistent zygosity assignments (i.e. the two twins received different zygosity assignments) by latent class algorithm (0.5% of the total number of pairs); using the broader criterion of $p > 0.95$, this total increased to 40 pairs with inconsistent zygosity assignments (2.1%). Using the former criterion, if one twin was assigned as MZ with $p > 0.9999$, there was 99.0% probability that the co-twin would also be assigned as probable MZ with $p > 0.5$, while if the twin was assigned as DZ, there was a 97.1% probability that the co-twin would also be assigned as DZ with $p > 0.5$. Using the latter criterion, if one twin was assigned as MZ with $p > 0.95$, there was a 97.6% probability that the co-twin would also be assigned as probable MZ with $p > 0.5$, with a corresponding probability of 96.0% if the twin was assigned as DZ that the co-twin would also be assigned as probable DZ.

A very high level of agreement was obtained between best-estimate zygosity diagnosis, and LCA-assigned zygosity. Out of a total of 1127 complete pairs classified as MZ by best-estimate diagnosis, there were only 2 pairs (one male and one female like-sex: 0.2%) where the latent class analysis independently assigned the two twins from the pair as dizygotic; and, on review, it was seen that these 2 pairs indeed should have been classified as DZ. There were an additional 26 pairs (2.2%) where the two twins received discrepant zygosity assignments by LCA. Out of 906 pairs classified as DZ by best-estimate diagnosis, there were 10 (6 male like-sex, 4 female like-sex) that were assigned as MZ by LCA algorithm. In 3 of these cases (1 male like-sex, 2 female like-sex) the original best estimate diagnosis was found to be apparently erroneous, with the remaining 7 cases needing to be clarified by genotyping. There were an additional 63 pairs classified as DZ where the two twins received discrepant zygosity assignment by LCA.

There were 121 pairs (242 twins) where zygosity had been definitively established by genotyping. Only a single twin received a zygosity assignment by LCA algorithm that was inconsistent with the assignment made from the genotypic data. That individual was from a pair which had discrepant LCA assignments (which would therefore have been flagged for follow-up genotyping). The discrepant individual gave zygosity questionnaire responses that led to his assignment as monozygotic with probability 76%; the co-twin was assigned by LCA as dizygotic (the true zygosity of the twin pair) with probability 100%.

Discussion

In the absence of genotypic data for zygosity assessment, a zygosity diagnosis algorithm derived by fitting a 2-class latent class model has a number of attractions. Much discussion of zygosity diagnosis has considered data from two twin informants. For many profiles of responses to zygosity questions; however, zygosity can be assigned with relatively high probability (better than 95%, and in many cases better than 99%) based on the responses of a single twin. Thus, using data from the young adult cohort of the Australian twin panel, we have found that for response profiles associated with better than a .9999 probability of being a "probable MZ" twin, there is a 99% probability that the co-twin will give responses that confirm the assignment as MZ, and even for response profiles associated with higher than .95 probability of being from a probable MZ twin, there is a 97.6% probability that the co-twin responses would confirm this assignment. For studies which have a specific focus which requires oversampling one zygosity type (e.g., dizygotic pairs for an affected sib pair linkage study; or discordant MZ pairs for a risk-factor study), the ability to eliminate many newly ascertained pairs on the basis of the responses of a single twin will produce an important gain in efficiency.

In traditional studies comparing MZ and DZ pairs, the identification of pairs where zygosity assignment by LCA is uncertain (e.g., less than 95%), or discrepant between the two twins, will allow more efficient targeting of resources for detailed review of responses to zygosity questions and for genotyping where necessary of these latter pairs. Where such uncertainty cannot be resolved, weighted analyses that include some pairs as MZ with probability x and DZ with probability (1– x), where x is the estimated probability of being monozygotic, could also be conducted. Because the latent class model is probabilistic, it can also handle reporting errors (e.g., MZ pairs who report differences in physical appearance based on perceived slight differences in hair color or other features) and environmentally determined differences in appearance (e.g., the twin who reported that strangers never had difficulty distinguishing him from his MZ co-twin, but who only later clarified that he had a limb amputated in early childhood). Finally, quantification of the probability of monozygosity using questionnaire data will be of value even when genotypic data are available, because of the risk of laboratory errors. Whenever a twin pair is assigned as monozygotic based on genotypic data,

https://doi.org/10.1375/twin.6.1.22 Published online by Cambridge University Press

but is assigned with high probability as dizygotic by latent class algorithm, the possibility that a blood or other sample from a single respondent has been duplicated under the co-twin's respondent number must be considered, and a second set of samples obtained where possible.

The latent class analyses that we have described here are easily implemented using software that is freely available (e.g., in the case of LCAP (Neuman et al., 1999) from http://hardy.wustl.edu). Once a latent class model has been fitted, the class membership probabilities and class-specific item endorsement probabilities that have been estimated can be used to generate zygosity class assignment probabilities for new data, including new profiles of responses to zygosity questions. Programs for this purpose also already exist. The validity of LCA-based assignment of zygosity must ultimately be confirmed by genotypic data. However, the high degree of concordance that we observed with traditional methods suggests that such validity will indeed be confirmed. Detailed review of responses to zygosity questions, and in many cases genotypic data, will also be needed to clarify those relatively rare pairs receiving inconsistent zygosity assignments. (A constrained version of the model that we have fitted here, jointly analyzing data from both twin informants, can also be used for this purpose, but is beyond the scope of this paper). The LCA approach that we have reviewed here is not without limitations. For example, since it does not allow for correlated measurement errors, it may misclassify pairs in rare cases (e.g., a block of items have been accidentally reverse coded for a respondent) where an expert diagnostician would not. Nonetheless, the analyses that we have presented strongly suggest that the probabilistic approach to zygosity diagnosis advocated here will prove to be of greater utility than more traditional approaches using questionnaire data.

Acknowledgments

Supported by NIH grants AA07728, AA11998, AA13321 (to ACH), DA/CA12854 (to PAFM) and AA13326 (to NGM).

References

Abecasis, G. R., Cherny, S. S., Cookson, W. O., & Cardon, L. R. (2001). GRR: Graphical representation of relationship errors. *Bioinformatics, 17,* 742–743.

Cederlof, R., Friberg, L., Jonsson, E., & Kaij, L. (1961). Studies on similarity diagnosis in twins with the aid of a mailed questionnaire. *Acta Genetica et Statistica Medica, 11*, 338–362.

Eaves, L. J., Siberg, J. L., Hewitt, J. K., Rutter, M., Meyer, J. M., Neale, M. C., et al. (1993). Analyzing twin resemblance in multisymptom data: Genetic applications of a latent class model for symptoms of conduct disorder in juvenile boys. *Behavior Genetics, 23,* 5–20.

Hauge, M., Harwald, M., Holm, N., Kristofferson, K., & Gurtler, H. (1989). Evaluation of zygosity diagnosis in twin pairs below age seven by means of a mailed questionnaire. *Acta Geneticae Medicae et Gemellologiae, 38,* 305–313.

Heath, A. C., Howells, W., Madden, P. A. F., Bucholz, K. K., Nelson, E. C., Slutske, W. S., et al. (2001). Predictors of non-response to a questionnaire survey of a volunteer twin panel: Findings from the Australian 1989 twin cohort. *Twin Research, 4,* 73–80.

Kasriel, J., & Eaves, L. J. (1976). The zygosity of twins: Further evidence on the agreement between diagnosis by blood groups and written questionnaires. *Journal of Biosocial Science, 8,* 263–266.

McCucheon, A. L. (1987). *Latent class analysis*. Newbury Park, CA: Sage Publications.

Magnus, P., Berg, K., & Nance, W. E. (1983). Predicting zygosity in Norwegian twins born 1915–1960. *Clinical Genetics, 24,* 103–112.

Nelson, E. C., Heath, A. C., Madden, P. A. F., Cooper, L. C., Dinwiddie, S. H., Glowinski, A., et al. (2002). Association between self-reported childhood sexual abuse and adverse psychosocial outcomes: Results from a twin study. *Archives of General Psychiatry, 59,* 139–145.

Neuman, R. J., Todd, R. D., Heath, A. C., Reich, W., Hudziak, J. J., Bucholz, K. K., et al. (1999). Evaluation of ADHD typology in three contrasting samples: A latent class approach. *Journal of the American Academy of Child and Adolescent Psychiatry, 38,* 25–33.

Nichols, R. C., & Bilbro, W. C. (1966). The diagnosis of twin zygosity. *Acta Genetica et Statistica Medica, 16,* 265–275.

Peters, H., Van Gestel, S., Vlietinck, R., Derom, C., & Derom, R. (1998). Validation of a telephone zygosity questionnaire in twins of known zygosity. *Behavior Genetics, 28,* 159–163.

Sarna, S., Kaprio, J., Sistonen, P., & Koskenvuo, M. (1978). Diagnosis of twin zygosity by mailed questionnaire. *Human Heredity, 28,* 241–254.

https://doi.org/10.1375/twin.6.1.22 Published online by Cambridge University Press