Twin Research and Human Genetics (2021), **24**, 168–175
doi:10.1017/thg.2021.24

CAMBRIDGE
UNIVERSITY PRESS

Article

Determining Zygosity in Infant Twins – Revisiting the Questionnaire Approach

Irzam Hardiansyah[1,*], Linnea Hamrefors[1,*], Monica Siqueiros[1,2], Terje Falck-Ytter[1,3] and Kristiina Tammimies[1]

[1]Centre of Neurodevelopmental Disorders (KIND), Division of Neuropsychiatry, Department of Women's and Children's Health, Karolinska Institutet, Stockholm, Sweden, [2]Division of Interdisciplinary Brain Sciences, Department of Psychiatry and Behavioural Sciences, Stanford University, Stanford, California, USA and [3]Developmental and Neurodiversity Lab (DIVE), Division of Developmental Psychology, Department of Psychology, Uppsala University, Uppsala, Sweden

Abstract

Accurate zygosity determination is a fundamental step in twin research. Although DNA-based testing is the gold standard for determining zygosity, collecting biological samples is not feasible in all research settings or all families. Previous work has demonstrated the feasibility of zygosity estimation based on questionnaire (physical similarity) data in older twins, but the extent to which this is also a reliable approach in infancy is less well established. Here, we report the accuracy of different questionnaire-based zygosity determination approaches (traditional and machine learning) in 5.5 month-old twins. The participant cohort comprised 284 infant twin pairs (128 dizygotic and 156 monozygotic) who participated in the Babytwins Study Sweden (BATSS). Manual scoring based on an established technique validated in older twins accurately predicted 90.49% of the zygosities with a sensitivity of 91.65% and specificity of 89.06%. The machine learning approach improved the prediction accuracy to 93.10%, with a sensitivity of 91.30% and specificity of 94.29%. Additionally, we quantified the systematic impact of zygosity misclassification on estimates of genetic and environmental influences using simulation-based sensitivity analysis on a separate data set to show the implication of our machine learning accuracy gain. In conclusion, our study demonstrates the feasibility of determining zygosity in very young infant twins using a questionnaire with four items and builds a scalable machine learning model with better metrics, thus a viable alternative to DNA tests in large-scale infant twin studies.

Keywords: Zygosity prediction; twins questionnaire method; ACE model; machine learning; bootstrapping; computational statistics

(Received 5 May 2021; accepted 8 June 2021; First Published online 12 July 2021)

Twin studies are valuable in determining the relative contributions of genetic and environmental factors to individual differences in human traits and diseases. In classic twin studies, the comparison of the degree of within-pair similarity of identical twins (monozygotic; MZ) to the degree of similarity within fraternal twin pairs (dizygotic; DZ) can infer these relative contributions of genes and environment to a trait. While twin studies have been around for decades, studies done on infant twins are scarce. Infant twin studies contribute to understanding key developmental processes and differences and to what extent these are driven by genetic factors.

To successfully obtain heritability estimates in the classic twin design, it is necessary to accurately measure their zygosity. The most accurate way of doing this is by extracting and comparing DNA from biological specimens such as saliva or blood samples. While zygosity determined by DNA testing has a reliability of nearly 100% (Hannelius et al., 2007), it is not always feasible, especially in large cohort studies (Jackson et al., 2001). Although genotyping is very accessible, there can still be obstacles in obtaining biological samples and genotyping from large infant cohorts, thus motivating the research of alternative methods to predict zygosity. There are also ethical implications associated with collecting DNA samples related to privacy, long-term storage and future analyses (Anderlik & Rothstein, 2001). Although parents or legal guardians may consent to collecting DNA from their infants, the information acquired from DNA could implicate relatives who have not given their consent. The infants themselves have not given their consent to collecting, using and storing their DNA and may oppose this as they get older (Botkin et al., 2015). It is also fundamental that parents are informed about the use of their infants' DNA in connection to their consent, but this cannot always be assumed. Samples are also usually stored for many years after collection, and consenting guardians may forget the initial reason for collection and what they consented to. Analyses that were not possible at the time for consent may be possible in the future, using the sample collected many years prior.

Previous studies have been able to predict zygosity using questionnaires, many using the items about highly heritable traits, such as overall appearance, eye color and hair color, as presented in the Goldsmith Child Zygosity Questionnaire (Goldsmith, 1991). Twin studies have used the questionnaire to determine zygosity in samples of 18-month-old twins with seemingly good results (Constantino et al., 2017; Hawks et al., 2019; Price et al., 2000). The largest disadvantage with questionnaire classification of zygosity is that there are always a number of twin pairs left

Author for correspondence: Irzam Hardiansyah, Email: irzam.hardiansyah@ki.se

*These authors contributed equally to the article.

Cite this article: Hardiansyah I, Hamrefors L, Siqueiros M, Falck-Ytter T, and Tammimies K. (2021) Determining Zygosity in Infant Twins – Revisiting the Questionnaire Approach. *Twin Research and Human Genetics* **24**: 168–175, https://doi.org/10.1017/thg.2021.24

unclassified, usually ranging from 5 to 10% (Goldsmith, 1991). For this reason, it is important to find a good analysis (classification) technique for processing questionnaire responses that would minimize the number of unclassified pairs while maintaining high accuracy.

To our knowledge, the only study to assess the accuracy of questionnaire method in predicting zygosity in 5-month-old twins is one by Forget-Dubois et al. (2003). They found that the questionnaire was able to correctly predict zygosity in 91.9% of 123 cases. As highlighted by Goldsmith (1991), it is important that zygosity questionnaires are constructed for the specific age group being studied. In this study, we aim to expand the existing literature of twins zygosity prediction by (1) replicating the use of questionnaire method for zygosity prediction in a 5-month-old infant population but now with a much larger sample size; (2) highlighting comparative techniques to process the questionnaire data, including the use of machine learning algorithms for calling the predictions; (3) providing a numeric generalization of the classification results utilizing computational simulations and (4) providing simple quantification on the impact of misclassification on estimates of heritability and shared environmental influence through numeric simulations. As there are inherently higher classification error rates with the questionnaire-based method compared with analysis of DNA samples, characterization of the amount of bias induced by different levels of zygosity misclassification on the two aforementioned parameters should provide valuable information for those contemplating the use of the questionnaire-based method.

Materials and Methods

Sample and Baseline Zygosity Estimated From DNA

The sample of our study included 284 twin pairs from the BabyTwins Study in Sweden (BATSS). The twin pairs participated in BATSS between 2016 and 2020. During the study visit, multiple questionnaires, experimental and biological data were collected (Falck-Ytter et al., 2021). The biological samples included saliva collected from all the twins using the DNA Genotek OG-575 collection kit during the study visit. The saliva samples were stored at the Karolinska Institutet biobank and processed for DNA extraction using a Chemagen kit based on magnetic bead separation in the Hamilton ChemagicSTAR® platform. The DNA samples were then used to perform zygosity analysis using two methods. The first 195 twin pairs were analyzed using selected single nucleotide polymorphisms (SNPs) based on the earlier reported protocol (Hannelius et al., 2007) at the Mutational Analysis Core Facility. An additional 89 twin pairs were genotyped using Infinium Global Screening Array-(Illumina, San Diego, CA, USA). The estimated identity by descent was analyzed using the PLINK software after quality control and removing SNPs with minor allele frequency less than 0.05 within the samples, deviation from Hardy–Weinberg equilibrium. All pairs of DNA samples showing $\pi \geq 0.99$ were considered as MZ pairs. In the sample, 128 twin pairs (45.1%) were dizygotic (DZ) and 156 (54.9%) were monozygotic (MZ). There were 137 (48.2%) female pairs and 147 (51.8%) male, all participants being of 168 ± 17.88 [135−324] days of age (mean ± *SD* [min−max]).

Zygosity Questionnaire

The eight-item zygosity questionnaire was administered online to the parents in connection to their visit at the Center of Neurodevelopmental Disorders at Karolinska Institutet. The questionnaire items are commonly used in zygosity questionnaires (Forget-Dubois et al., 2003; Jackson et al., 2001; Lichtenstein et al., 2002), originating from Goldsmith (1991). The first four items inquired how the parents perceive the physical similarity of their twins in terms of hair color, eye color and earlobe shape and if they thought their twins to be 'like two peas in a pod' (Cederlöf et al., 1961) or not more similar than siblings in general (Table S1 in Supplementary material). Another four items included questions about how often the twins get mixed up by strangers, if the parents thought their twins were MZ or DZ and if the twins share the same blood group and/or Rh factor. The questionnaire respondents were the parents of the twins (91.2% mothers and 8.8% fathers) who had no prior knowledge about their children's zygosity besides information obtained during pregnancy and at delivery.

Manual Classification of Zygosity

To classify the twin based on the questionnaire responses with a manual method, we used an algorithm validated in the Child and Adolescent Twin Study in Sweden data set (Lichtenstein et al., 2002). The manual method enabled us to analyze four out of the eight items in our questionnaire (hair color, eye color, two peas in a pod and mixed up by strangers). The responses to these items were given a score as to whether they indicated the twins being MZ (1), DZ (−1) or not valid (0) (Table S1). The scores of the individual items yielded a score sum for each twin pair, ranging from a maximum of 4 (MZ) to a minimum of −4 (DZ). These sums were then coded as zygosity estimations using different thresholds. In the study by Lichteinstein et al. (2002), the threshold 3/-3 was used, presenting that a pair needed a score equal to or larger than 3 to be estimated as MZ and a score smaller than or equal to -3 to be estimated as DZ. As our study sample is considerably younger, we decided to test the performance of different thresholds (1/-1, 2/-2, 3/-3 and 4/-4). The performances of these thresholds were compared by the proportion of accurate zygosity estimations and the proportion of unclassified pairs.

Machine Learning Algorithm Generation for Zygosity Determination

In addition to the manual approach described above, we employed machine learning algorithms trained as binary classifiers to predict the zygosity of the twins. Specifically, we used three different algorithms: Random Forest (RF; Ho, 1995), Support-Vector Machine (SVM; Cortes & Vapnik, 1995) and a simple feedforward artificial neural net (also called Multilayer Perceptron; MLP; Hastie et al., 2009) to learn from the items of our questionnaire to predict the zygosity outcome. Essentially, when trained as a binary classifier, a machine learning algorithm attempts to infer a decision boundary in the sample data space, which best separates data points belonging to different classes. The three algorithms we used thus differ in how they approach this inference problem. An RF classifier builds multitudes (hundreds or thousands) of simple decision trees simultaneously — thus, the notion of 'ensemble learning' — in a manner such that each tree makes predictions independently from one another and the whole 'forest' thus produces a distribution (a Gaussian in the case of a continuous-valued output) of predictions centered and peaked around the most likely outcome. 'Voting' of predictions among all the trees subsequently takes place to decide on the eventual class prediction of the entire forest (in practice, such 'voting' may amount to taking the modus,

https://doi.org/10.1017/thg.2021.24 Published online by Cambridge University Press

median, mean or some other summary statistical value of the prediction distribution).

In contrast, an SVM classifier looks for regions of widest gap ('margin') between data points of the two classes and then tries to fit a separating line (or a hyperplane in multidimensional case) which traces these regions, thus producing a boundary of 'maximum margin'. Finally, a neural network is a collection of many interconnected computational units that mimic a simplified biological neural system. Each of these computational units (called 'neurons') performs simple arithmetic or mathematical mapping operations, but when put together in such a configuration, they form a powerful inferential engine that can approximate any mathematical function: a 'universal function approximator'. Hence, when we use a sigmoid mapping operation in the output neuron, we can produce a logistic regression function with a potentially very complex and nonlinear combination of predictors, enabling the algorithm to draw a highly expressive decision boundary. Here, we used the linear variant of SVM (Lin-SVM) which draws a relatively simple linear boundary. Readers interested in reading more about these algorithms are advised to refer to, for example, Bishop (2006) or Hastie et al. (2009).

Unlike the manual algorithm, the ML algorithms could utilize all the eight items we included in our questionnaire as learning features (or predictors), thus making greater use of the available information. Nonetheless, while the ML algorithms were initially trained with all of these eight items, we eventually kept only the same four items used in the manual method to be used as ML features to maintain comparability between the two approaches and as the use of this smaller set of features had only negligible impact on the classification accuracy of the ML method. As the target variable for training the ML algorithms, we used the zygosity information produced by actual DNA analysis, containing binary outcomes of MZ or DZ. Here, MZ was taken as the reference (positive) class; hence, we used a binary coding: MZ = 1 and DZ = 0 to generate our target variable. Due to the relatively small sample employed in this study, a 10-fold cross-validation (10CV) technique (Kohavi, 1995) was used to train and test the three classifiers to maximize the amount of information available during training. This technique essentially reuses the whole data in a round-robin manner to train and test the model without setting aside a separate data set for testing the trained model. Finally, to ensure replicability of our results, the same random seed was used to fix the random initializations of various model parameters (i.e., interconnection weights in MLP, internal bootstrapping and predictor selections in RF and optimization steps in Lin-SVM) before training each algorithm. All implementation codes of our ML classifiers (in R language) and an anonymous sample data set are available in the supplementary materials.

Numeric Simulation Analyses

To obtain a more realistic performance measure for each algorithm (ML models and manual algorithm) under uncertain class distributional conditions (i.e., where the algorithms are faced with new data sets having not only a different zygosity prevalence but also a different combination of answers to questions by class, from what we have here), we performed a bootstrapping simulation with k = 50,000 resamplings with replacement (Efron & Tibshirani, 1993) to generate many possible different class distributions. For small data sets, simple cross-validation (without replication) was found to frequently result in overestimation of the predictive performance of binary diagnostic tests, and bootstrapping has been suggested as a reliable way to correct such bias (Smith et al., 2014). Therefore, we applied this technique to obtain realistic confidence intervals of the predictive performance of both our manual and ML-based methods for each of the three reported metrics, that is, sensitivity, specificity and total accuracy.

In the manual approach, resamplings were done by choosing randomly among the 284 twin pairs existing in the complete data set, where each pair could occur more than once in any single (re) sample, and recalculating the three metrics for each sample based on the scores assigned with the 1/-1 threshold. Each (re)sample had the same size (284 observations) as the original complete data set, but on average, only 63.2% of these were unique observations. In the ML-based approach, due to the need to set aside data for training and testing separately, the overall data were split into two based on the time upon which the DNA tests of the twin pairs were received. In the earlier (first) batch, we received DNA-based zygosity information for 195 twin pairs, and these data were used for algorithm training. In the later (second) batch, we received DNA-based zygosity data for the remaining 89 twin pairs, which was then used as a test set, on which we performed the bootstrapping simulation and tested the trained algorithms. Each (re)sample had the same size (89 observations) as the complete test set, but on average, only 63.2% of these were unique.

We observed that the class distribution (along the four features employed for each ML model) of the first-batch data set (of 195 pairs) was very similar to our complete data set. Thus, training the three algorithms using the earlier data set should somewhat mimic the training performed using the complete data set, while bootstrapping the latter data set could simulate how sensitive this particular training setting was to distributional variabilities unseen during training. From the bootstrapping, we subsequently obtained a confidence interval for each of the three algorithms (RF, MLP, Lin-SVM) and three performance metrics. Due to most of these score distributions being heavily right-skewed (with a long, thin left tail), we reported the median and interquartile range (IQR) to measure the central tendency and the spread, respectively, for all distributions obtained from the bootstrapping.

Sensitivity Analyses for Parameter Estimation in ACE Models

To perform sensitivity analysis of parameter estimation in structural equation model (SEM)-based twin modeling, also known as the ACE models (Knopik et al., 2017; Neale & Maes, 2004), we used example data sets that are widely available with the distribution of the free OpenMX package (Boker et al., 2011): one data set of body mass index (BMI) of 3808 twin pairs (2009 DZ, 1799 MZ) from the Australian social attitudes twin study of Martin et al. (1986)[1] and one synthetic data set of 400 pairs (200 DZ, 200 MZ) with two unnamed (simply 'X' and 'Y') phenotypes. To induce misclassification in the range of 0−10%, which includes the figures we observed in our results (see Results), for each increment of one percentage point, a corresponding number of randomly selected pairs (e.g., in the 'Y' phenotype data, for 1% error rate there were four pairs, for 2% there were 8 and so on) had their actual zygosity label flipped (i.e., from 'MZ' to 'DZ' and vice versa). MZ and DZ intraclass correlations (ICCs) along with their difference were calculated followed by an ACE model-fitting to this modified data set using the lightweight twinlm() function of the 'mets' package (Scheike et al., 2013) in R to obtain quick estimates of A (heritability) and C (shared environment) variance components. This process was repeated for 10,000x to obtain a confidence interval for each of the abovementioned parameters: r_{MZ}, r_{DZ}, Δr, A and C.

https://doi.org/10.1017/thg.2021.24 Published online by Cambridge University Press

In addition, the significance of the C estimate was also noted each time to see how the conclusion regarding the presence or absence of C could change due to zygosity misclassification. All the 906 opposite-sex DZ twins in the Australian data set were excluded to reduce the complexity of the twin analysis further, and the number of MZ pairs was downsampled to avoid having asymmetric simulated error rates between the two zygosity classes, bringing the final sample size to 2206 twin pairs (1103 DZ, 1103 MZ). Finally, we regressed the mean of the computed parameters on the error rates (from 0% to 10%, with 1% increment) to see how the estimates will change with increasing misclassification rate.

Statistical Analyses

The nonparametric Kruskal–Wallis tests were performed on the bootstrapped results (see 'Numerical simulation analyses') to identify differences in performance among the alternative methods (three ML-based and one manual), for each of the three metrics, followed by a pairwise posthoc analysis using the nonparametric Mann–Whitney–Wilcoxon (MWW) test for each significant difference detected.

Results

Manual Scoring Classification

Using the 'default' manual scoring techniques (Lichtenstein et al., 2002), the threshold of 4/-4 or 3/-3 yielded a perfect accuracy of 100% of the zygosity determination. However, due to a larger number of unclassified pairs in the higher threshold (i.e., 3/-3: 46.1% vs. 4/-4: 69.7%), we used the 3/-3 threshold for comparison with the lower thresholds (i.e., 1/-1, 2/-2; Table 1). The use of the lower thresholds reduced the accuracy marginally while noticeably increasing the number of classified pairs. A threshold of 2/-2 yielded 76.4% accurate zygosity predictions while leaving 24.3% of the twin pairs unclassified. The lowest and most generous threshold, 1/-1, accurately classified 90.5% of the sample while leaving 5.8% unclassified. Based on the DNA zygosity, 53.3% of these unclassified pairs were MZ and 46.7% were DZ.

The manual analysis also showed that the proportion of correctly estimated MZ pairs exceeded the proportion of DZ pairs in all thresholds (Table 1), although threshold 1/-1 produced the least difference. Next, using the 1/-1 threshold and MZ zygosity as the reference class, we performed the bootstrap simulation to obtain confidence intervals for total accuracy, specificity (DZ accuracy) and sensitivity (MZ accuracy). Summary statistics of the obtained distributional results are as follows: accuracy: 90.49% ± 1.33%; specificity: 89.06% ± 2.11% and sensitivity: 91.65% ± 1.69% (Figure 1, red boxplot).

Machine Learning Classification

We obtained a total of eight usable items from the questionnaire, of which only four were used in the manual approach to follow the reference study (see above, 'Manual Classification of Zygosity'). Incidentally, the variable importance ranking given by the RF algorithm after the initial training (with all eight items) showed that four out of the five top discriminative predictors were the same four items used in the manual approach (see Figure S1 in Supplementary Material). The complete descriptives (labels, frequencies and proportions) of each of the eight variables are presented in Table S2 in the Supplementary Material.

When trained and tested using the 10CV technique, out of the three algorithms we employed, the MLP classifier yielded the best performance with 14 out of 284 incorrect predictions (total accuracy of 95.1%), while the RF classifier yielded the worst with 29 out of 284 (total accuracy of 89.9%; Table 2, Figure 1). However, unlike the manual method, no twin pairs were left unclassified as the algorithms automatically set their binary decision thresholds to categorize each twin pair as either MZ or DZ. As a result, all the three classifiers made a larger, or at least equal, number of correct predictions compared with the manual approach in any threshold setting. We also performed subsequent bootstrapping simulation for each of the three algorithms and obtained the following results: RF accuracy: 93.10% ± 2.98%, RF specificity: 91.30% ± 6.11%, RF sensitivity: 94.29% ± 4.18%; SVM accuracy: 91.07% ± 3.44%, SVM specificity: 85.71% ± 6.28%, SVM sensitivity: 94.29 ± 4.09%; MLP accuracy: 93.10% ± 2.98%, MLP specificity: 91.30% ± 6.11%, MLP sensitivity: 94.29% ± 4.18% (Figure 1). The wider confidence intervals observable in these results were due to a much smaller resampling data set (89 obs.) than in the manual method (the whole 284 obs.).

Table 1. Performance of the manual algorithm, thresholds across twins and in the DZ/MZ group respectively

Threshold	3/-3	2/-2	1/-1
Correct	153/284 (53.9%)	212/284 (74.6%)	257/284 (90.5%)
Correct DZ	53/128 (41.4%)	87/128 (68.0%)	114/128 (89.1%)
Correct MZ	100/156 (64.1%)	125/156 (80.1%)	143/156 (91.7%)
Unclassified	131/284 (46.1%)	69/284 (24.3%)	15/284 (5.3%)
Unclassified DZ	75/128 (58.6%)	38/128 (29.7%)	7/128 (5.5%)
Unclassified MZ	56/156 (35.9%)	31/156 (19.9%)	8/156 (5.1%)

Note: 'Correct' shows the number of zygosity estimations that were correctly classified by the threshold, separated by (/) the number of responses that met the threshold criteria, as well as the number of correct classifications. 'Unclassified' shows the number of pairs that did not meet the threshold criteria. $N = 284$; $n_{(DZ)} = 128$; $n_{(MZ)} = 156$.

When comparing the performance of the manual algorithm with the three ML models using the nonparametric Kruskal–Wallis test, we show overall significant differences for all metrics (in all, $p < .001$). Subsequent post hoc analysis with the pairwise MWW tests revealed that all the three ML-based algorithms had significantly different (in all, $p < .001$) performance compared with the manual algorithm (Figure 1). Furthermore, the accuracy and specificity performance of the SVM classifier was significantly different (in both, $p < .001$) from those of RF and MLP, but no statistically significant difference was found in sensitivity among the three algorithms (results not shown). Finally, no statistically significant difference in the performance of RF and MLP was found for any of the three performance metrics.

Impact Analysis of Misclassification On Twin Modeling Parameter Estimations

As zygosity misclassifications occurred in all the algorithms, we wanted to investigate the impact of these misclassifications. Misclassifications of DZ twins as MZ tend to lower the r_{MZ} (MZ intraclass correlation), while misclassifications of MZ twins as DZ tend to increase the r_{DZ}. Hence, the general effect of zygosity misclassifications is to narrow the gap between the two ICCs (i.e., reducing the Δr). As expected, this pattern was observed with all the three phenotype data sets employed in our numerical simulations. We produced a distribution of values for each of the three ICC parameters (r_{MZ}, r_{DZ} and Δr) and the two ACE model parameters, A and C, which stands for additive genetics and shared

https://doi.org/10.1017/thg.2021.24 Published online by Cambridge University Press

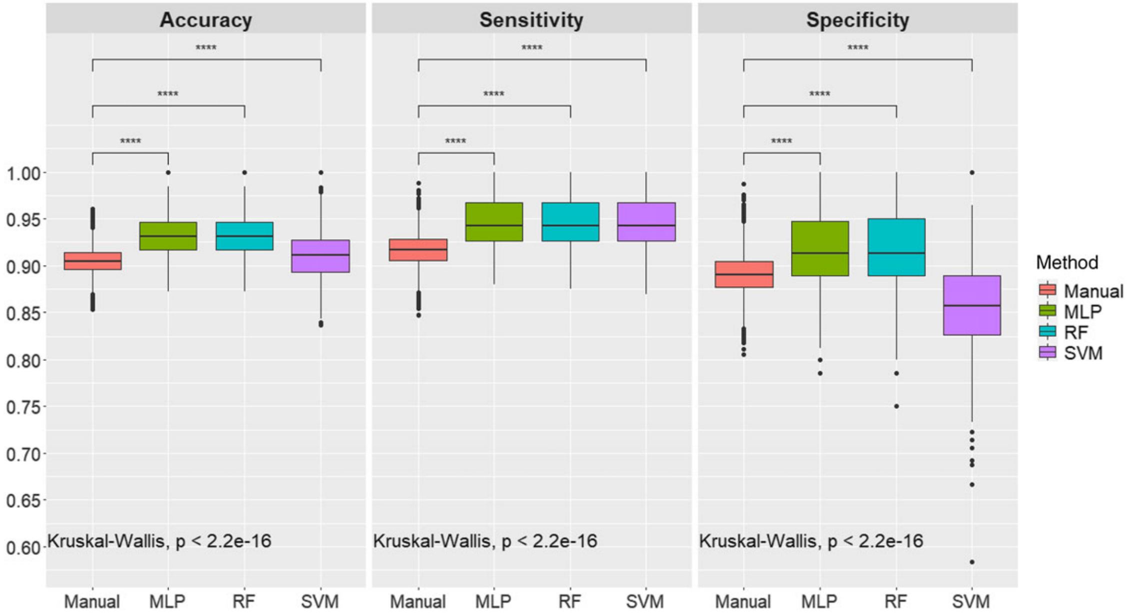

Fig. 1. Comparative bootstrapped performance of the three ML-based and manual approaches. Bootstrapped distributions for the manual approach are approximately Gaussian, while those for the three ML approaches were right-skewed. The manual method is the baseline for the comparisons. In all cases, MZ is the positive class; thus, sensitivity is the MZ accuracy, while specificity is the DZ accuracy.
Note: ML, machine learning; MZ, monozygotic; DZ, dizygotic.

environmental variance component, respectively. These distributions were invariably symmetrical and bell-shaped, and they quantify the uncertainty surrounding the parameter calculations in the presence of a certain amount of zygosity prediction error.

Regressing the *mean* of the above distributions on the error rates (i.e., from 0% to 10%, increment by 1%, where 0% corresponds to values calculated from the original data with no flipped zygosity label) for all the three ICC parameters, it was found that the mean varies linearly with an increased error rate, with an almost perfect fit ($R^2 \simeq 1$) and a highly significant linear slope effect ($p < .001$; Figure 2). The negative–positive slope of these regression lines depends mainly on the magnitude of the corresponding ICC itself, but invariably in all the three phenotype data sets: r_{MZ} decreases (a negative slope), r_{DZ} increases (a positive slope), and Δr decreases even more steeply than r_{MZ} (actually, the slope of Δr is the negative sum of the magnitudes of r_{MZ} and r_{DZ} slopes). Equally important, the standard deviation of the distributions (and hence confidence interval around the computed ICC values) also grows linearly with increasing error rate. Figure 2a shows the regression lines (along with each confidence band) of r_{MZ}, r_{DZ} and Δr for the Australian BMI data as an illustration. Similar figures (not shown here) were obtained for the other two data sets.

In contrast, for the A and C parameters of the ACE model, the mean varies *quadratically* with an increased error rate, again with an almost perfect regression fit ($R^2 \simeq 1$) where the quadratic effect was highly significant ($p < .001$). As with ICC, invariably for all three phenotype data sets: mean of A decreases (negative quadratic effect), mean of C increases (positive quadratic effect) and confidence interval of both parameters grows wider quadratically with increasing zygosity misclassification rate. As an illustration, Figure 2b shows the regression lines (along with each confidence band) of A and C for the Australian BMI data (https://rdrr.io/cran/OpenMx/man/twinData.html). Similar plots (not shown here) were obtained for the other two data sets. Finally, Figure 2c presents the probability of false detection of the presence of C influence (i.e., a significant non-zero estimate) for the three phenotypes — all of which were known to have strictly zero C variance component — as a function of increasing misclassification rate. The probability was calculated as the proportion of 10,000 simulation trials, for each 1% increment of the error rate, in which the lower-limit value of C's confidence interval was found to be positive (i.e., the whole CI was on the right of zero). As seen in the figure, the probability grows exponentially with an increased error rate, but the magnitude stays zero or very close to zero even with larger error rates near the end of the range.

Discussion

We found that even using a threshold of 1/-1 and allowing all participants with a score sum over zero to be coded as either MZ or DZ produced 90.5% accuracy in classifications (leaving 5.3% unclassified). Compared to the reports of Forget-Dubois et al. (2003), where a zygosity questionnaire predicted 91.9% of zygosities in infant twins, we suggest that a threshold of 1/-1 is sufficient for analyzing a four-item zygosity questionnaire for infants. Our bootstrap simulation showed the distribution of results to have an overall accuracy, sensitivity and specificity of approximately 90%.

To improve the use of the questionnaire-based zygosity determination, we built three different ML algorithms to help classify the twins based on the same answers. RF algorithm is known to be among those possessing the best capacity to generalize its learning to new unseen data due to its bootstrapping-based ensemble inference that makes its learning very robust against overfitting. Therefore, it is no surprise that its performance on our particular twin data set was very close to the generalized one obtained using the bootstrapping simulations. On the contrary, the much higher performance of Lin-SVM and MLP on our twin data set in contrast with their respective less stellar generalization performance indicates some extent of overfitting in both algorithms' learning to the particular class distribution observed in our twin data set. A

https://doi.org/10.1017/thg.2021.24 Published online by Cambridge University Press

Table 2. Performance of the three ML classifiers for binary classification of zygosity

Algorithm	RF	Lin-SVM	MLP
Correct total (Accuracy)	255/284 (89.9%)	269/284 (94.7%)	270/284 (95.1%)
Correct DZ (Specificity)	116/128 (90.6%)	122/128 (95.3%)	125/128 (97.7%)
Correct MZ (Sensitivity)	139/156 (89.1%)	147/156 (94.2%)	145/156 (93.0%)

Note: With the machine learning-based approach, the number of unclassified pairs is zero (no pair left unclassified). In all three algorithms, $N = 284$, $n_{(DZ)} = 128$, $n_{(MZ)} = 156$ and MZ is the ref. (positive) class.

MZ, monozygotic; DZ, dizygotic; RF, Random Forest, SVM, Support-Vector Machine; MLP, multilayer perceptron.

case in point is the specificity of linear SVM, which showed a substantial discrepancy between its generalized mean performance (85.7%) and performance on our twin data set (95.3%) — as indicated by the statistical test, its generalized specificity tended to be even inferior to that of the manual method. As the bootstrapping simulations for the ML were calculated with a much smaller bootstrap sample (89 pairs) than for the manual approach (284 pairs), it may have contributed to this and the skewed and more dispersed bootstrapped distributions for the ML approach. Based on both the performance of the ML models in our data and simulation bootstrapping, we propose that either an RF or MLP model be used to determine zygosity in further studies. However, the three algorithms' generalized sensitivity was shown to be no different statistically (Figure 1).

Both the machine learning and manual scoring approach showed that the zygosity of DZ twins is more challenging to predict than MZ twins. We hypothesize that this is due to the items of zygosity questionnaires, including ours, that generally focus on physical attributes (i.e., hair and eye color and generally perceived physical similarity). As DZ twins can share the same hair and eye color, they can be misclassified as MZ by a scoring system with binary outputs as in our manual scoring approach. DZ twins can also be mixed up by strangers based on them being twins and not necessarily because they are physically identical.

As shown by our numerical simulations, zygosity misclassification always underestimates heritability and may lead to overestimation of shared environmental (C) influence on phenotypic variance. However, the magnitude of such biases depends on both the strength of the underlying ICCs and the sample size (smaller samples tend to produce much more variable estimates and more significant biases). Furthermore, when the C effect in the ACE model is virtually zero, as is the case for many known phenotypes, the inflated estimation of C due to zygosity prediction error will not so much lead to an erroneous conclusion about the presence of shared environmental influence as merely inducing a doubtful C estimate due to a wide confidence interval that usually crosses zero, unless the misclassification is very severe. For the range of misclassification rate presented here, we showed that the probability of such erroneous conclusion occurring should be minimal: the worst case was ˜2% chance (i.e., 224 in 10,000 trials) for the Australian BMI twin data in the presence of 10% misclassification rate, while the other two (synthetic) phenotypes shown probabilities in mere tenths of a percentage point.

It is somewhat surprising that both the bias and uncertainty of ACE parameter estimation were magnified with a quadratic rate as the misclassification rate increases. As there is scarce referential literature investigating the issue, we could only speculatively think that such quadratic biases arise from the manner computation of model parameters is implemented in SEM-based ACE modeling software: The path coefficients a, c, e of the structural model might be directly estimated from the observed ICCs, while the variance components A, C, E were each calculated as the square of the corresponding path coefficient. Therefore, a linear change in a and/or c (due to a proportional change in ICCs) would produce a corresponding quadratic change in A and/or C respectively. Nevertheless, such a quadratic pattern has two important implications: (1) incremental worsening of zygosity prediction error would inflate the uncertainty in much larger jumps — for example, a 40% increase in error rate will cause the confidence interval to double and (2) the observed performance gain from using RF and MLP algorithms compared to the manual scoring technique would indeed meaningfully translate to much lower underlying uncertainty in ACE model parameter estimation; the confidence interval should be only half as wide in the former case due to a reduction of error rate from 10% to 7% (i.e., $7^2/10^2 \cong 0.5$). Combined with the simplicity of using a pretrained ML zygosity classifier to crunch questionnaire responses, the latter implication should make a good case for deploying our, or a similar ML-based approach, for this problem.

As another point of reflection, we saw a need to adapt the original questionnaire of Goldsmith (1991) to better suit infancy studies, at least in the way the questions and/or answer choices are formulated and phrased. As items in the questionnaire generally ask about how similar the twins' physical looks are, attributes that are still developing and have not reached stability in the pediatric population (especially at 5 months of age), strong ambiguities often arise in perceiving the similarity. Setting aside parents' subjective bias in viewing how similar their twin children are, objective differences in physical measures such as weight, head shape and body size certainly affect the similarity judgment, while such discrepancies usually are temporary and would later disappear. It is not unusual that physical differences manifest in a pair of MZ twins during their early months of life due to factors present during pregnancy, but as both infants grow older and physical features further develop, the resemblance of MZ twins becomes more apparent. In light of this, we suggest that the questionnaire be adapted by including items related to medical diagnoses, expert opinions and facts that are generally disclosed to parents during pregnancy and delivery; for example, shared or nonshared amniotic sac(s), chorionicity already mentioned above. The aforementioned study by Forget-Dubois et al. (2003), in which they included an item about chorionicity to boost their prediction accuracy to ˜96%, provides a case in point. Another case is Segal (1984), which reported a classification accuracy of ˜94% using expert's guesses of twins' zygosity, albeit with much older twins (children and adolescents) and thus may not readily translatable to our sample of very young infants. The choices of such questions will need to consider the issue of privacy.

Finally, our findings of the sensitivity of heritability estimate to zygosity mislabeling in twin modeling hint at a prospect of developing fully ML-based techniques for zygosity labeling as a viable alternative to the conventional DNA-based labeling. Indeed, without analyzing a sample of bodily substances (e.g., saliva, skin or blood) from the participants, zygosity predictions made by an ML-based questionnaire technique might never reach the level of accuracy comparable to that of a DNA-based technique. However, as our results showed that a false detection of shared environment component is very unlikely to occur in twin modeling

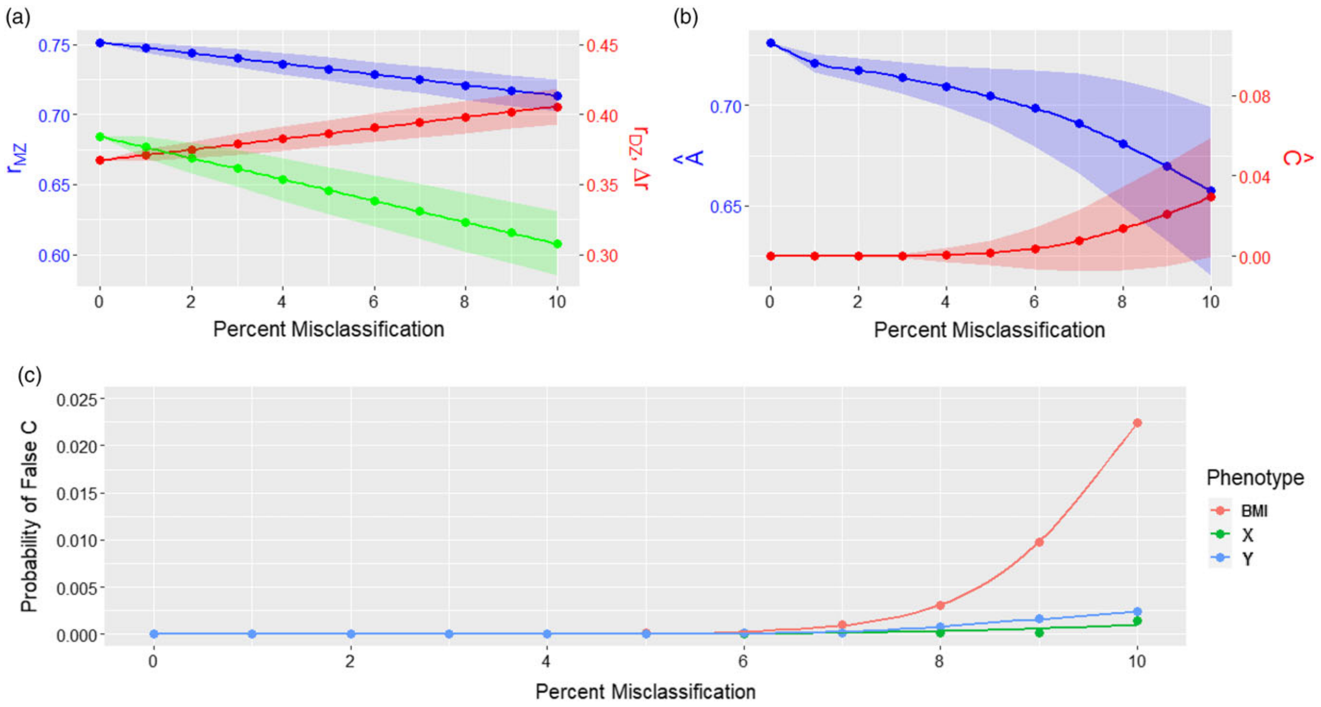

Fig. 2. Illustrations of misclassification impact on heritability model estimates. (a, top-left) Biaxis plot of r_{MZ} (blue line, left y-axis), r_{DZ} and Δr (red line and green line, respectively, both right y-axis) along with their respective confidence band as a linear function of zygosity prediction error; (b, top-right) Biaxis plot of A (blue line, left y-axis) and C (red line, right y-axis) along with their respective confidence band as a quadratic function of zygosity prediction error; (c, bottom) Probability of false detection of C in ACE model grows much faster than a linear rate with increasing zygosity prediction error, although the nominal probability remains very small in the shown error range. The first two plots are from the Australian twin BMI data set, the third from all three data sets.
Note: A, additive genetic variance; C, common (or shared) environmental factors; E, specific (or nonshared) environmental factors plus measurement error.

even at a 10% misclassification rate, devising an ML-based algorithm capable of churning out the accuracy of, say, ˜98% may already be good enough to get a reliable estimate of heritability. As shown in the study by Forget-Dubois et al. (2003), with a proper redesign of the questionnaire, such level of performance should not be a bridge too far. An important caveat to this would be, in very rare cases of non-twins (e.g. twins switched during birth or twins having different fathers), a ML binary zygosity classifier would not be able to reveal the true zygosity information, since parental responses would most likely show much, if not complete, overlap with those of DZ twins' parents (obviously, such twins will not look very similar physically). However, this is a limitation of the questionnaire method in general.

Conclusion

Here, we demonstrate the feasibility of using a questionnaire method to determine zygosity adopted from Goldsmith (1991) in infant twins with reasonably high accuracy, thus replicating the results of Forget-Dubois et al. (2003). We introduce an ML algorithm to replace this manual scoring process to improve scalability and eliminate unclassified pairs for large-scale use. Two of our ML models based on Random Forest and MLP yielded an even better result. Furthermore, the results that we reported in this study can be generalized to many other data sets having different distributions, and they showed the limits of prediction accuracy achievable with the questionnaire method in its current form. Perhaps more importantly, our results demonstrated that using only four items most commonly employed for adults in the original questionnaire were sufficient to reach the reported accuracy levels, thus pointing to an opportunity for simplifying the original questionnaire to include fewer, but the most zygosity-discriminating, questions.

Supplementary material. To view supplementary material for this article, please visit https://doi.org/10.1017/thg.2021.24.

Acknowledgments. We would like to thank PhD student Lynnea Myers and research assistants Joy Hättestrand, Johanna Kronqivst, Sofia Tahbaz, Anna Kernell and Carolin Schreiner for their great work with data collection for BabyTwins. We would also like to thank postdoc Ana Maria Portugal and research assistants Sophie Lingö and Angelinn Liljebäck for their assistance in the project.

Financial support. This study was supported by grants to IH from Marie S. Curie Actions SAPIENS Project (MSCA-ITN-2018 N. 814302), to MS from H2020 Marie Sklodowska-Curie Actions (642996), to TFY from Stiftelsen Riksbankens Jubileumsfond (NHS14-1802:1; Pro-Futura Grant in collaboration with the Swedish Collegium for Advanced Study) the Swedish Research Council (SRC; 2018-06232) and to KT from Swedish Research Council and Swedish Foundation for Strategic Research.

Conflict of interest. None.

Ethical standards. The authors assert that all procedures contributing to this work comply with the ethical standards of the relevant national and institutional committees on human experimentation and with the Helsinki Declaration of 1975, as revised in 2008.

Notes

[1] Also available at: https://rdrr.io/cran/OpenMx/man/twinData.html

References

Anderlik, M. R., & Rothstein, M. A. (2001). Privacy and Confidentiality of Genetic Information: What rules for the New Science? *Annual Review of Genomics and Human Genetics*, *2*, 401–433.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer-Verlag.

Boker, S., Neale, M., Maes, H., Wilde, M., Spiegel, M., Brick, T., . . . Fox, J. (2011). OpenMX: An open source extended structural equation modeling framework. *Psychometrika*, *76*, 306–317.

Botkin, J. R., Belmont, J. W., Berg, J. S., Berkman, B. E., Bombard, Y., Holm, I. A., . . . & McInerney, J. D. (2015). Points to consider: Ethical, legal, and psychosocial implications of genetic testing in children and adolescents. *The American Journal of Human Genetics*, *97*, 6–21.

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

Constantino, J. N., Kennon-McGill, S., Weichselbaum, C., Marrus, N., Haider, A., Glowinski, A. L., . . . Jones, W. (2017). Infant viewing of social scenes is under genetic control and is atypical in autism. *Nature*, *547*, 340–344.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*, 273–297.

Efron, B., & Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. Chapman and Hall.

Falck-Ytter, T., Hamrefors, L., Sanchez, M.S., Portugal, A.M., Taylor, M., Li, D., . . . Ronald, A. (2021). Babytwins Study Sweden (BATSS): A multi-method infant twin study of genetic and environmental factors influencing infant brain and behavioral development. *bioRxiv*.

Forget-Dubois, N., Perusse, D., Turecki, G., Girard, A., Billette, J-M., Rouleau, G., . . . Tremblay, R. E. (2003). Diagnosing zygosity in infant twins: Physical similarity, genotyping, and chorionicity. *Twin Research and Human Genetics*, *6*, 479–485.

Goldsmith, H. (1991). A zygosity questionnaire for young twins: A research note. *Behavior Genetics*, *21*, 257–269.

Hannelius, U., Gherman, L., Mäkelä, V.-V., Lindstedt, A., Zucchelli, M., Lagerberg, C., . . . Lindgren C. M. (2007). Large-scale zygosity testing using single nucleotide polymorphisms. *Twin Research and Human Genetics*, *10*, 604–625.

Hastie, R., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd ed.). Springer-Verlag.

Hawks, Z. W., Marrus, N., Glowinski, A. L., & Constantino, J. N. (2019). Early origins of autism comorbidity: Neuropsychiatric traits correlated in childhood are independent in infancy. *Journal of Abnormal Child Psychology*, *47*, 369–379.

Ho, T. K. (1995). Random decision forests. In *Proceedings of the 3rd International Conference on Document Analysis and Recognition* (vol. 1. pp. 278–282).

Jackson, R., Sneider, H., Davis, H., & Treiber, A. (2001). Determination of twin zygosity: A comparison of DNA with various questionnaire indices. *Twin Research and Human Genetics*, *4*, 12–18.

Knopik, V. S., Neiderhiser, J. M., DeFries, J. C., & Plomin, R. (2017). *Behavioral Gnetics* (7th ed.). Worth Publishers, Macmillan Learning.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence* (vol. 2, pp. 1137–1145).

Lichtenstein, P., De Faire, U., Floderus, B., Svartengren, M., Svedberg, P., & Pedersen N. L. (2002). The Swedish Twin Registry: A unique resource for clinical, epidemiological and genetic studies. *Journal of Internal Medicine*, *252*, 184–205.

Martin, N. G., Eaves, L. J., Heath, A. C., Jardine, R., Feingold, L. M., & Eysenck, H. J. (1986). Transmission of social attitudes. *Proceedings of the National Academy of Science*, *83*(12), 4364–4368.

Neale, M. C., & Maes, H. H. M. (2004). *Methodology for Genetic Studies of Twins and Families*. Kluwer Academic Publisher.

Price, T. S., Freeman, B., Craig, I., Petrill, S. A., Ebersole, L., & Plomin, R. (2000). Infant zygosity can be assigned by parental report questionnaire data. *Twin Research and Human Genetics*, *3*, 129–133.

Scheike, T. H., Holst, K. K., & Hjelmborg, J. B. (2013). Estimating heritability for cause specific mortality based on twin studies. *Lifetime Data Analysis*, *20*, 210–233.

Segal, N. (1984). Zygosity testing: Laboratory and investigator's judgment. *Acta Geneticae Medicae et Gemellologiae: Twin Research*, *33*, 515–520.

Smith, G. C. S., Seaman, S. R., Wood, A., & Royston, P. (2014). Correcting for optimistic prediction in small data sets. *American Journal of Epidemiology*, *180*, 318–324.

https://doi.org/10.1017/thg.2021.24 Published online by Cambridge University Press