Data mining and population genetics of birth defects: preliminary investigation

B. Little
Texas Data Mining Research Institute,
Tarleton State University, Stephenville, Texas

Abstract

Population level inbreeding can be estimated from mating patterns in marriage records. Birth defect registries separately collect information on the frequency and types of birth defects. The U.S. national census provides population structure data to the county or postal code levels in another separate database. State health departments maintain vital statistics information on marriage, reproduction, birth defects, and mortality in yet other databases. More than 6 million marriages in Texas were analysed to estimate inbreeding at county, regional, and state levels. Three types of birth defects were analysed: (1) one known to be associated with an autosomal recessive inheritance pattern (Ventricular Septal Defect – VSD); (2) one of unknown aetiology, speculated to be associated with a very rare autosomal recessive gene (Ebstein's Anomaly), and (3) one known to not be related to parental consanguinity (Fetal Alcohol Syndrome - FAS). A significant relationship between estimated local consanguinity and VSD was found. Ebstein's Anomaly, the birth defect of unknown aetiology but suspected to be recessively inherited, showed a strong relationship to estimated population level of inbreeding, suggesting a major recessive gene influence. In the case of a birth defect known to caused by environmental rather than genetic factors (FAS), no relation to estimated inbreeding was found. In conclusion, data mining population genetic data revealed patterns of birth defects in very large databases (VLDBs) when merged into a data structure suited to data mining. In this instance, a viable hypothesis was derived for the cause of an extremely rare birth defect of unknown aetiology. Hence, data mining of population genetic VLDBs can yield new information that may be useful to guide genomic and clinical research directions.



Data Mining V, A. Zanasi, N. F. F. Ebecken & C. A. Brebbia (Editors)
© 2004 WIT Press, www.witpress.com, ISBN 1-85312-729-9

1 Introduction

Very Large Databases (VLDBs) that include information relevant to the analysis of population genetics and birth defects exist. The information is in different VLDBs and before data mining may be done, it is necessary to join them at a level of aggregation and standardization that renders analysable data.

Data mining applications in population genetics are concentrated on microarray data analysis, which is a new area for population geneticists. However, the traditional interest of population geneticists such as genetic structure and the correlation of structure with disease states and anomalies have not been researched using data mining research.

Population genetics is concerned with the genetic structure of human populations, including but not confined to mating behaviours. In population genetics, variance is typically partitioned into three components: Phenotypic = Genetic + Environmental + Gene*Environment Interaction. The study of the patterns and possible causes of birth defects, or congenital anomalies, is sometimes ascribed to teratology (i.e., the study of monster babies). However, the precise mission of teratology is generally limited to the analysis of environmentally caused congenital anomalies.

A prevailing interest in teratology and genetics is differentiation of the heritable and environmental correlates of birth defects, and determination of causes. Data mining in population genetics and birth defects to analyse the association between inbreeding and congenital anomalies has not been reported previously. Heuristically, inbreeding is the probability that an individual will inherit a gene from a common ancestor through both parents. In the event that a recessive gene influences a birth defect, inbreeding would increase the probability of the occurrence of the birth defect.

In this preliminary investigation, the degree to which estimated inbreeding predicts the incidence of three select birth defects (one genetic, one unknown, one environmental) is analysed in Texas.

2 Materials

Approximately 6 million marriage records, including the bride's maiden name, were available for the period 1961 to 2001 for Texas. Texas began a birth defects registry in 1996 to monitor occurrence of congenital anomalies at the county level, and the database contains approximately 70,000 cases of birth defects. The most complete information was available for 1999-2001 (n~20,000). In addition, the US Census from the year 2000 was available at the county level for 20 million residents of Texas.

3 Methods

3.1 Databases used in the preliminary investigation

Surname isonomy is an established method to estimate population level of inbreeding [1-4]. Published methods for the estimated inbreeding coefficient from surname isonomy were employed to compute county level estimates.

Data Mining V, A. Zanasi, N. F. F. Ebecken & C. A. Brebbia (Editors)
© 2004 WIT Press, www.witpress.com, ISBN 1-85312-729-9

Birth defect surveillance data were available for Texas at the county level, and were joined into a VLDB that contained marriages at the county level. In addition, the Texas Department of Health Vital Statistics maintains a separate database with the number of live births that occur by county per year.

The US Census for Texas in the year 2000 was also available by age, gender, and ethnicity. These data were joined at the county level into a single database for data mining. US Census and Texas Department of Health Vital Statistics data were used to normalize the database. Normalization was necessary because of widely discrepant population sizes in Texas (e.g., counties range in size from > 1 Million to less than 5,000 people). The method of data mining chosen was regression because of its wide use and accepted interpretation in genetics [5].

3.2 Construction of hypotheses

Three types of birth defects were chosen for preliminary analysis: (1) one known to be associated with an autosomal recessive inheritance pattern (Ventricular Septal Defect – VSD) [6]; (2) one of unknown aetiology, speculated to be associated possibly autosomal recessive (Ebstein's Anomaly - EA) [6], and (3) one known to not be related to parental consanguinity (Fetal Alcohol Syndrome - FAS), i.e., environmentally caused [7].

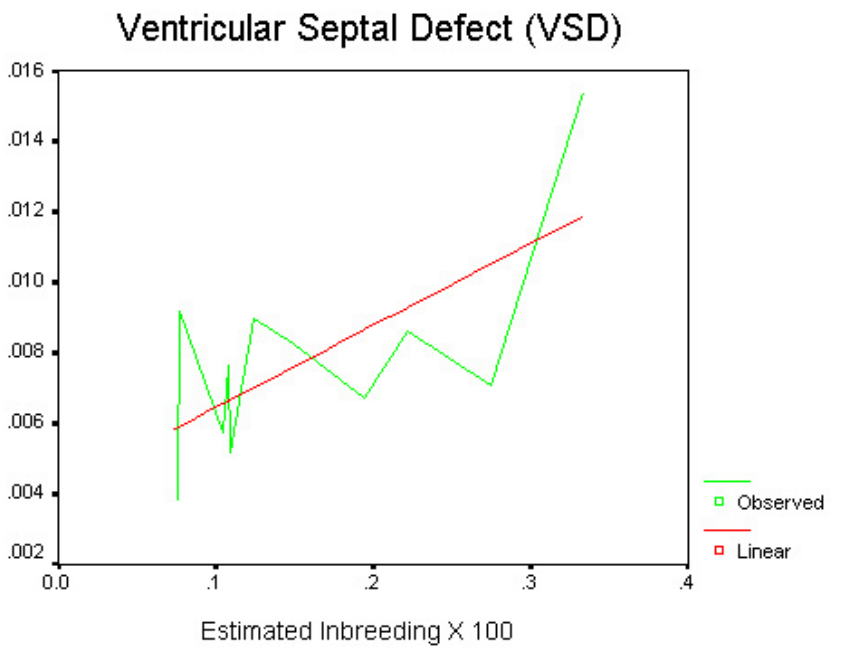


Figure 1: Incidence of ventricular septal defect regressed on estimated Inbreeding ($R2 = 0.43$, $\beta = 0.232$, $P < 0.03$).

Data Mining V, A. Zanasi, N. F. F. Ebecken & C. A. Brebbia (Editors)
© 2004 WIT Press, www.witpress.com, ISBN 1-85312-729-9

3.3 Data mining tools

Regression analysis is one of the most commonly employed tools in data mining. It is also the underlying technique for many of the other more complex data mining tools (e.g., neural networks, factor analysis, canonical correlation, MANOVA). In the current analysis, regression was used to assess the ability of estimated local inbreeding to predict the local incidence of a specific birth defect. The major weakness of regression analysis techniques is their sensitivity to large sample sizes.

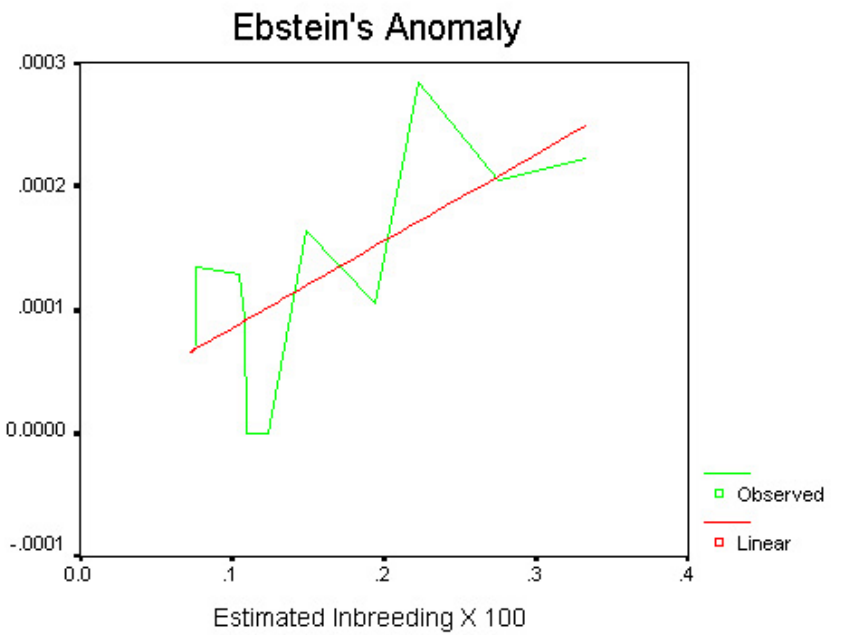


Figure 2: Incidence of Ebstein's anomaly regressed on estimated inbreeding ($R2 = 0.46$, $\beta = 0.007$, $P < 0.02$).

4 Results

Regression analysis of estimated inbreeding and birth defects was used to evaluate hypothesized genetic and environmental aetiologies. Because regression analysis is sometimes influenced by large sample sizes, the results were computed using a quasi-experimental design, as described above.

4.1 Hypothesis testing

In the case of the first birth defect analysed, it was *a priori* hypothesized that estimated inbreeding was strongly related to the incidence of VSD. Therefore, the slope of the regression line was expected to be positive.

© 2004 WIT Press, www.witpress.com, ISBN 1-85312-729-9

For the second birth defect analysed, it was unknown whether or not incidence of Ebstein's Anomaly was strongly influenced by estimated inbreeding. No *a priori* hypothesis existed, and the analysis was truly exploratory data mining.

For the third birth defect analysed, the *a priori* hypothesis was that the relationship between fetal alcohol syndrome and estimated inbreeding would not be different from a regression slope of zero.

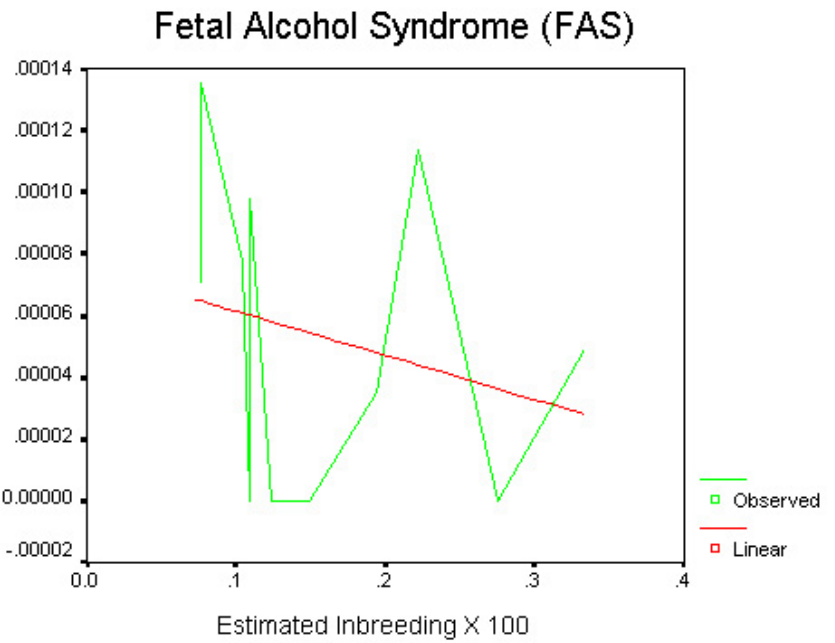


Figure 3: Incidence of fetal alcohol syndrome regressed on estimated inbreeding ($R2 = 0.06$, $\beta = -0.0001$, P = NS).

4.2 Ventricular septal defect (VSD) regressed on estimated inbreeding

For example, VSD was strongly related to estimated inbreeding in the regression analysis (Figure 1), and VSD is well known to be associated with an autosomal recessive pattern of inheritance [6]. Results of this analysis suggest that VSD is inherited in an autosomal recessive manner (Figure 1).

Numerous prior investigations have suggested that inbreeding may be associated with an increased risk for VSD, the analysis using VLDBs and data mining strengthen significantly that argument. Importantly, other causes of VSD such as teratogenic influences are known to exist. Therefore, the association between estimated inbreeding VSD is confounded by multiple causes.

Data Mining V, A. Zanasi, N. F. F. Ebecken & C. A. Brebbia (Editors)
© 2004 WIT Press, www.witpress.com, ISBN 1-85312-729-9

4.3 Ebstein's anomaly regressed on estimated inbreeding

Ebstein's Anomaly is a rare congenital heart defect, with no clear aetiology. Some investigators have proposed that the defect is related to environmental exposures during the first 6-8 weeks post conception (e.g., lithium) [8]. Other investigators have suggested that Ebstein's Anomaly is due to a rare recessive gene [6]. Findings in the present analysis suggested that Ebstein's Anomaly is due, at least in part, to recessive genetic inheritance because increased estimated inbreeding is positively related to the incidence of the birth defects. However, the greatest emphasis has been placed upon possible teratogenic causes of Ebstein's Anomaly [8]. The present study's findings are important to the aetiology of Ebstein's Anomaly because it provides strong evidence that inbreeding significantly increases the birth defect's incidence, indicating a major recessive gene effect.

4.4 Fetal alcohol syndrome regressed on estimated inbreeding

Fetal Alcohol Syndrome (FAS) is caused by maternal alcohol abuse during gestation [7]. It is not related to level of inbreeding, nor is it primarily caused by any known genetic mechanism. FAS was included in the present analysis as a "negative control." Some critics contend that just because data mining involves millions of cases and thousands of variables, any finding can be "manipulated" to be statistically significant.

The negative finding for FAS incidence regressed on estimated inbreeding was confirmatory because it was known a priori that no relationship should exist between the dependent and independent variables.

5 Discussion

These results confirmed that data mining is a powerful tool for traditional population genetic investigations, in addition to microarray data analysis. VSD was known to be associated with recessive genetics in prior investigations [6], and was used as a positive control in this preliminary investigation. Similarly, FAS was known from prior investigation to NOT be influenced by inbreeding levels [7], and was a negative control. Accordingly, FAS was not related to inbreeding in the present analysis. It may be noted that, in fact, the FAS line for regression on estimated inbreeding has a slope that is actually negative. Prior investigations have shown that diagnosis of FAS depends upon level of physician training and interest, and this may be related to the lower incidence in areas of higher estimated inbreeding [9].

5.1 Ventricular septal defect (VSD) regressed on estimated inbreeding

While VSD was chosen as a positive control in this preliminary analysis, new information was gained. It was not previously known that VSD incidence increased linearly with estimated inbreeding. This is an intriguing finding because it suggests that VSD may be primarily a recessive gene trait.

© 2004 WIT Press, www.witpress.com, ISBN 1-85312-729-9

Teratogenic causes of VSD cases may be less frequent than previously thought given the regression slope that is nearly 0.25, which would indicate a perfect fit for the recessive gene hypothesis [5].

5.2 Ebstein's anomaly regressed on estimated inbreeding

Prior investigations of Ebstein's Anomaly have focused primarily on the hypothesized teratogenic effect of lithium on the embryonic heart [8]. The present findings indicate that at least a certain proportion of Ebstein's Anomaly cases are probably associated with estimated inbreeding, perhaps caused by a recessive gene in the homozygous state. This is an intriguing finding for future researchers to analyse using genome scanning techniques, searching for this recessive gene. The research question arises "Could the manic-depression gene(s) be linked to the recessive gene(s) for Ebstein's Anomaly?" because of the putative association between lithium exposure during pregnancy for the treatment of manic-depression and the birth defect.

5.3 Fetal alcohol syndrome regressed on estimated inbreeding

This was simply a negative control in this preliminary investigation, and no new information was derived from its inclusion.

6 Conclusion: data mining studies in population genetics

Data mining techniques combined with the VLDBs contemporarily available are very powerful for the investigation of disease topics in population genetics important to human health. Discoveries of a substantive nature that can assist in directing genomic and clinical research await the construction of joined VLDBs and data mining analysis.

References

[1] Crow, J.F. and Mange, A.P. (1965) Measurement of inbreeding from the frequency of marriages between persons of the same surname. Eugenics Quarterly 12: 199-203.

[2] Crow, J.F. (1980) The estimation of inbreeding from isonomy. Human Biology 52: 1-14.

[3] Lasker, G.W. (1968) The occurrence of identical (isonymous) surnames in various relationships in pedigrees: a preliminary analysis of the relation of surname combinations to inbreeding. American Journal of Human Genetics 20: 250-257.

[4] Lasker, G.W. (1977) A coefficient of relationship by isonomy: a method for estimating the genetic relationship between populations. Human Biology 49: 489-493.

[5] Falconer, D.S., and T.F.C. Mackay. 1996. Introduction to Quantitative Genetics. 4th ed. Prentice Hall, Harlow, U.K.



Data Mining V, A. Zanasi, N. F. F. Ebecken & C. A. Brebbia (Editors)
© 2004 WIT Press, www.witpress.com, ISBN 1-85312-729-9

[6] McKusick, V.A. (1998) Mendelian Inheritance in Man: A Catalog of Human Genes and Genetic Disorders, 12th ed. Johns Hopkins University Press: Baltimore.
[7] Jones, K.L. (1997) Smith's Recognizable Patterns of Human Malformation. 5th ed. W.B. Saunders: Philadelphia.
[8] Yonkers, K.A., B.B. Little, D. March (1998) Lithium use during pregnancy: A controversy. CNS Drugs 9:261-269, 1998.
[9] Little, B.B., Snell, L.M., Gilstrap, L.C., Rosenfeld, C.R., Gant, N.F. (1990) Failure to recognize fetal alcohol syndrome in newborn infants. American Journal of Diseases in Children 144:1142-1146.



Data Mining V, A. Zanasi, N. F. F. Ebecken & C. A. Brebbia (Editors)
© 2004 WIT Press, www.witpress.com, ISBN 1-85312-729-9